OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Online Covariance Matrix Estimation in Stochastic Inexact Newton Methods

Wei Kuang WEIKUANG@UCHICAGO.EDU
Department of Statistics, The University of Chicago

Sen Na SENNA@BERKELEY.EDU
ICSI and Department of Statistics, University of California, Berkeley

Mihai Anitescu ANITESCU@MCS.ANL.GOV
Department of Statistics, The University of Chicago and Mathematics and Computer Science Division, Argonne National Laboratory

Abstract

We aim to study the practical statistical inference of the online second-order Newton method for general unconstrained stochastic optimization problems under the fixed dimension setting. We consider the adaptive inexact stochastic Newton method, which is reduced from the stochastic sequential programming (StoSQP) method in [18] to the unconstrained setting. Based on the asymptotic normality of the last iteration, we propose a weighted sample covariance matrix, which is a consistent covariance matrix estimator. With this estimator, we are able to conduct statistical inference on the solution of the stochastic optimization problem in practice. The update of the estimator is entirely online and efficient in computation and memory. We demonstrate the empirical performance through numerical experiments on linear regression models.

1. Introduction

In this paper, we focus on the following unconstrained stochastic optimization problems

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

where $f : \mathbb{R}^d \to \mathbb{R}$ is a strongly convex objective function, and the expectation is taken over the random variable $\xi \sim \mathcal{P}$. We consider the fixed dimension setting where d is a constant. Problem (1) appears in various domains, including deep learning [13] and empirical likelihood [8]. Specifically, (1) can be seen as a parameter estimation problem. We take the linear regression problem as an example: Suppose $\xi = (\xi_{\boldsymbol{a}}, \xi_b)$ is generated from the model $\xi_b = \xi_{\boldsymbol{a}}^T \boldsymbol{x}^\star + \epsilon$, where $\xi_{\boldsymbol{a}} \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$ and $\epsilon \sim \mathcal{N}(0,1)$. ξ_a and ϵ are independent. Let the loss function be $f(\boldsymbol{x};\xi) = (\xi_b - \xi_{\boldsymbol{a}}^T \boldsymbol{x})^2/2$, then we can easily verify that the true parameter $\boldsymbol{x}^\star$ is the minimum of (1). Therefore, the statistical inference of the solution $\boldsymbol{x}^\star$ is of high interest, which characterizes the variability of the estimation.

In today's era of exponential data growth, online algorithms have gained increasing prominence. One of the widely adopted algorithms is stochastic gradient descent (SGD). [22] and [20] develop an acceleration procedure by averaging the SGD iterates, and establishing the asymptotic normality for averaged SGD (ASGD). Moreover, [6] proposes two consistent estimators of the covariance matrix in the asymptotic Gaussian distribution: plug-in estimator, which achieves faster convergence but involves the inverse of the averaged Hessian; batch-means estimator, which converges more slowly

but relies solely on the iteration sequences. [27] further introduces a fully online version of the batch-means estimator, eliminating the need for a predefined sample size. Other recent progresses in SGD inference include: [15, 24, 25], which delve into the statistical inference aspects of implicit SGD. Also, [16, 17] consider SGD under constant stepsize setting. Finally, [11] and [10] build confidence intervals for SGD and averaged implicit SGD using bootstrap techniques.

The first-order method might not be efficient enough if the eigenvalues of the Hessian matrix have significant variations. Thus, the online second-order Newton's methods are important, the statistical inference of which has appeared recently. [3] proposes an efficient online stochastic Newton algorithm for the logistic regression problem. Furthermore, the authors establish the asymptotic normality and offer a consistent covariance matrix estimator. [5] and [4] consider more general regression problems and develop the inference for the averaged stochastic Newton method. The aforementioned three works have two limitations: They consider objective functions with specific forms that allow rank-one updates on the estimated Hessian. Thus, direct updates on the estimated Hessian inverse can be applied and the Newton systems are exactly solved. Besides, they build the asymptotic normality for last iterate only at stepsize $O(1/t)$. [18] considers general objective function and generalizes to the case with deterministic constraint. The authors develop a fully online StoSQP method for equality-constrained stochastic optimization problems and establish the asymptotic normality on the last iteration for a more general class of adaptive stepsize and inexact solves of the linear systems using sketching techniques [14] to avoid taking matrix inverse. For the readers interested in the stochastic SQP topic, we refer to [1, 2, 7, 12, 19] for such recent developments.

We aim to study the stochastic Newton method's statistical inference, a particular case of the stoSQP introduced in [18], adapted here for the unconstrained optimization setting. To facilitate practical inference, it is crucial to have an effective estimator of the limiting covariance matrix. [18] provides a plug-in estimator, but it has two drawbacks: First, it is not a consistent estimator as it fails to account for the randomness of sketching. Second, it requires the inverse of the estimated Hessian, preventing the aim of solving the linear systems inexactly. Motivated by the drawbacks, we propose a weighted sample covariance matrix and establish its consistency theoretically. The update of the estimator is fully online and does not involve matrix inversions. Therefore, we can conduct statistical inference on the minimum $\boldsymbol{x}^\star$ in practice. To our knowledge, we are the first to propose a consistent estimator of the limiting covariance matrix of a stochastic Newton's method for general strongly convex objective functions. We show the performance of our estimator through numerical experiments on linear regression models.

Notation. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. $\|\cdot\|_F$ denote the Frobenius norm for matrices. For scalars a, b, $a \wedge b = \min(a, b)$. $a_t = O(b_t)$ or $a_t \lesssim b_t$ means $a_t \leq c b_t$ for large enough t with a positive constant c. $a_t = o(b_t)$ means $\lim_{t\to\infty} a_t/b_t = 0$. I denotes identity matrix and $\mathbf{0}$ denotes the zero vector. $\boldsymbol{e}_i$ represents the vector with i-th entry as 1 and 0 for the rest coordinates. For a sequence of compatible matrices $\{A_i\}_i$, we let $\prod_{k=i}^{j} A_k = A_j A_{j-1} \cdots A_i$ if $j \geq i$ and I if $j < i$. For a symmetric matrix A, $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) represents the smallest (largest) eigenvalue of A. Denote the solution of (1) as $\boldsymbol{x}^\star$ and $B^\star = \nabla^2 f(\boldsymbol{x}^\star)$. Let $f_t = f(\boldsymbol{x}_t)$ and $f^\star = f(\boldsymbol{x}^\star)$ (similar notations apply to ∇f_t, etc.). $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

2. Recap: Adaptive Inexact Newton Method

In this section, we briefly review the adaptive inexact Newton method which is reduced from [18].

Step 1: Estimate the gradient and Hessian. We generate a sample $\xi_t \sim \mathcal{P}$ and estimate the gradient $\bar{g}_t = \nabla f(\boldsymbol{x}_t; \xi_t)$ and the Hessian $\bar{H}_t = \nabla^2 f(\boldsymbol{x}_t; \xi_t)$. We also compute $B_t = \frac{1}{t}\sum_{i=0}^{t-1} \bar{H}_i + \Delta_i$. The modification $\|\Delta_i\| = O(\omega_i)$ ensures B_t is invertible with a predeterministic sequence $\{\omega_i\}$.

Step 2: Approximately solve Newton system via sketching. Considering the Newton system $B_t \bar{\Delta}\boldsymbol{x}_t = -\bar{g}_t$, we approximate the solution by a randomized iterative solver using sketching techniques [14]. At iteration j, we generate sketching matrix $S_{t,j} \in \mathbb{R}^{d\times q} \overset{iid}{\sim} \mathcal{S}$ and solve the problem

$$\boldsymbol{z}_{t,j+1} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z} - \boldsymbol{z}_{t,j}\|^2 \quad \text{s.t.} \quad S_{t,j}^T B_t \boldsymbol{z} = -S_{t,j}^T \bar{g}_t. \quad (\text{ with } \boldsymbol{z}_{t,0} = \boldsymbol{0}) \tag{2}$$

The explicit solution is $\boldsymbol{z}_{t,j+1} = \boldsymbol{z}_{t,j} - B_t S_{t,j}(S_{t,j}^T B_t^2 S_{t,j})^\dagger S_{t,j}^T (B_t \boldsymbol{z}_{t,j} + \bar{g}_t)$. We perform the randomized solver for τ iterations, and let $\bar{\Delta}\boldsymbol{x}_t = \boldsymbol{z}_{t,\tau}$.

Step 3: Update the iterate with a random stepsize. With the direction $\bar{\Delta}\boldsymbol{x}_t$, we generate a randomized stepsize α_t satisfying $0 < \beta_t \le \bar{\alpha}_t \le \beta_t + \chi_t$ where $\{\beta_t, \chi_t\}$ are predetermined sequences. We update the iterate by $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \bar{\alpha}_t \bar{\Delta}\boldsymbol{x}_t$.

Denote the randomness in the randomized solver and the stepsize as ζ_t and ψ_t, respectively. We define the adapted filtration as $\mathcal{F}_{t-2/3} = \sigma(\{\xi_i, \zeta_i, \psi_i\}_{i=0}^{t-1} \cup \xi_t)$, $\mathcal{F}_{t-1/3} = \sigma(\{\xi_i, \zeta_i, \psi_i\}_{i=0}^{t-1} \cup \xi_t \cup \zeta_t)$, and $\mathcal{F}_t = \sigma(\{\xi_t, \zeta_t, \psi_t\}_{i=0}^{t})$.

3. Global convergence and asymptotic normality

Assumption 1 *We assume $f(\boldsymbol{x})$ is twice continuously differentiable over $\mathbb{R}^d$. For any ξ, we assume the stochastic estimate $f(\boldsymbol{x}; \xi)$ is twice continuously differentiable with respect to $\boldsymbol{x}$. Besides, the gradient $\nabla f(\boldsymbol{x})$ and the Hessian $\nabla^2 f(\boldsymbol{x})$ are Lipschitz continuous with parameters Υ_{Lg} and Υ_{LH} respectively. Furthermore, the function $f(\boldsymbol{x})$ is μ-strongly convex. Additionally, there exist constants γ_B and Υ_B such that $\gamma_B \le \lambda_{\min}(B_t) \le \lambda_{\max}(B_t) \le \Upsilon_B$ for any $t \ge 0$.*

Assumption 2 *(a) For any $\boldsymbol{x}_t$, we have $\mathbb{E}[\bar{g}_t \mid \mathcal{F}_{t-1}] = \nabla f_t$. There exist constants $C_{g,m}$ such that $\mathbb{E}[\|\bar{g}_t\|^m \| \mid \mathcal{F}_{t-1}] \le \|\nabla f_t\|^m + C_{g,m}$ for $m = 2, 3, 4$. (b) The assumption also holds at $\boldsymbol{x}^\star$.*

Assumption 3 *For $t \ge 0$, we assume that the sketching matrices $S_{t,j} \overset{iid}{\sim} S$. (a) There exists a constant $\gamma_S > 0$ such that $\mathbb{E}[B_t S(S^T B_t^2 S)^\dagger S^T B_t] \succeq \gamma_S I$. (b) There exists constants $\Upsilon_{S,m}$ such that $\mathbb{E}[\|S\|^m (\|S\|^\dagger)^m] \le \Upsilon_{S,m}$ for $m = 1, 2$.*

Assumption 1 is standard in literature [5, 6]. Assumption 2 is a growth condition on the moments of estimated gradient. Assumption 3 guarantees the performance of the sketching solver, which is easy to achieve by several choices of $\mathcal{S}$, like a uniform distribution on the canonical basis [23].

Theorem 4 *(global convergence) Under Assumptions 1, 2(a) with $m = 2$, and 3(a), we suppose $\beta_t = c_\beta/(t+1)^\beta$ and $\chi_t = c_\chi/(t+1)^\chi$ with constants $c_\beta, c_\chi > 0$, $1/2 < \beta < 1$ and $\chi > 1$. If τ, the number of inner sketching iterations, satisfies $\tau \ge \log(\gamma_B/4\Upsilon_B)/\log(1-\gamma_S)$, then the iterates $\boldsymbol{x}_t$ satisfy $\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^\star$ almost surely.*

Theorem 4 establishes the global almost sure convergence of the iterate sequence. Its proof can be found in Appendix A.2. Compared with [18], we remove the compactness on the iterate sequence, while we strengthen strong convexity to global, which is commonly seen in stochastic unconstrained optimization [3, 6]. Assumption 2 is weaker than the bounded moment condition in [18].

Assumption 5 *For any $\boldsymbol{x}_t$, we have $\mathbb{E}\left[\bar{H}_t \mid \mathcal{F}_{t-1}\right] = \nabla^2 f_t$. There exist constants $C_{H,m}$ such that $\mathbb{E}\left[\|\bar{H}_t - \nabla^2 f_t\|_F^m \mid \mathcal{F}_{t-1}\right] \leq C_{H,m}$ for $m = 2, 4$.*

Assumption 6 *There exists a function $H(\xi)$ such that $\|\nabla^2 f(\boldsymbol{x};\xi)\| \leq H(\xi)$ for all $\boldsymbol{x}$. Moreover, there exist constant $\Upsilon_{H,m} > 0$ such that $\mathbb{E}[H(\xi)^m] \leq \Upsilon_{H,m}$ for $m = 2, 4$.*

Assumption 5 is the moment condition on the estimated Hessian. Assumption 6 implies the Lipschitz continuity of $\nabla f(\boldsymbol{x};\xi)$ [6].

Let $\tilde{C}^\star = -\prod_{j=1}^{\tau}(I - B^\star S_j (S_j^T (B^\star)^2 S_j)^\dagger S_j^T B^\star)$, $C^\star = \mathbb{E}\tilde{C}^\star$, and $I + C^\star = U\Sigma U^T$ with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ be the eigenvalue decomposition. Denote $\Omega^\star = (B^\star)^{-1}\mathbb{E}[\nabla f(\boldsymbol{x};\xi)\nabla^T f(\boldsymbol{x};\xi)](B^\star)^{-1}$. The next theorem builds the asymptotic normality of the last iterate, with proof in Section A.2.2.

Theorem 7 *Under the conditions of Theorem 4, suppose Assumptions 2 with $m = 3$, 3(b) with $m = 1$, 5 with $m = 2$, 6 with $m = 2$ hold, and $\omega_t = c_\omega/(t+1)^\omega$ for a constant $c_\omega > 0$. If $\chi > \beta$ and $(1 - (1-\gamma_S)^\tau) - \beta/c_\beta \mathbf{1}_{\{\beta=1\}} > 0$, we have*

$$\sqrt{1/\beta_t}(\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Xi^\star), \tag{3}$$

where

$$\Xi^\star = U\big(\Theta \circ U^T \mathbb{E}[(I + \tilde{C})\Omega^\star(I + \tilde{C})^T]U\big)U^T \quad \textit{with} \quad [\Theta]_{k,l} = 1/(\sigma_k + \sigma_l - \beta/c_\beta \mathbf{1}_{\{\beta=1\}}). \tag{4}$$

4. Covariance matrix estimate

We propose a consistent estimate of the covariance matrix $\Xi^\star$ in Theorem 7. Define

$$\bar{\boldsymbol{x}}_M = \frac{1}{M}\sum_{t=1}^{M} \boldsymbol{x}_t \text{ and } \Xi_M = \frac{1}{M}\sum_{t=1}^{M} \frac{1}{\varphi_{t-1}}(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_M)(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_M)^T \text{ with } \varphi_i = \frac{\beta_i + \chi_i}{2}. \tag{5}$$

It is easy to see that the updating process of Ξ_M can be entirely online. The following theorem gives the convergence rate of Ξ_M. Proof can be found in Appendix A.5.

Theorem 8 *(consistency of covariance matrix estimate) Suppose Assumptions 1, 2, 3, 5, and 6 hold. Let $\beta_t = c_\beta/(t+1)^\beta$, $\chi_t = c_\chi/(t+1)^\chi$, and $\omega_t = c_\omega/(t+1)^\omega$ with constants $c_\beta, c_\chi, c_\omega > 0$. If $1/2 < \beta < 1$, $\chi > 2\beta$, $\omega > \beta/2$, and $c_\chi \leq c_\beta^2$, then we have*

$$\mathbb{E}\Big[\big\|\Xi_M - \Xi^\star\big\|\Big] \lesssim M^{-\frac{1-\beta}{2}}. \tag{6}$$

We note that the plug-in estimator in [18], defined as

$$\Xi_M^{PI} := B_M^{-1}\Big(\frac{1}{M}\sum_{t=1}^{M} \bar{g}_t \bar{g}_t^T\Big)B_M^{-1} \tag{7}$$

. The plug-in estimator has $O((1-\gamma_S)^\tau)$ gap from Ξ, and the gap does not vanish as the sample size goes to infinity. Besides, the plug-in estimator requires the inverse of B_t but Ξ_M avoids this. Theorem 8 allows us to practically perform statistical inference on $\boldsymbol{x}^\star$. For example, if the conditions in Theorems 4, 7 and 8 hold, we can construct the following α%-confidence interval for $\mu := \boldsymbol{c}^T\boldsymbol{x}^\star$, which is asymptotic exact by [6, Corollary 4.4].

$$\Pr\Big(\boldsymbol{c}^T\boldsymbol{x}_t - z_{1-\alpha/2}\sqrt{\varphi_t \cdot \boldsymbol{c}^T\Xi_t\boldsymbol{c}} \leq \mu \leq \boldsymbol{c}^T\boldsymbol{x}_t + z_{1-\alpha/2}\sqrt{\varphi_t \cdot \boldsymbol{c}^T\Xi_t\boldsymbol{c}}\Big) \to \alpha. \tag{8}$$

5. Numerical Experiments

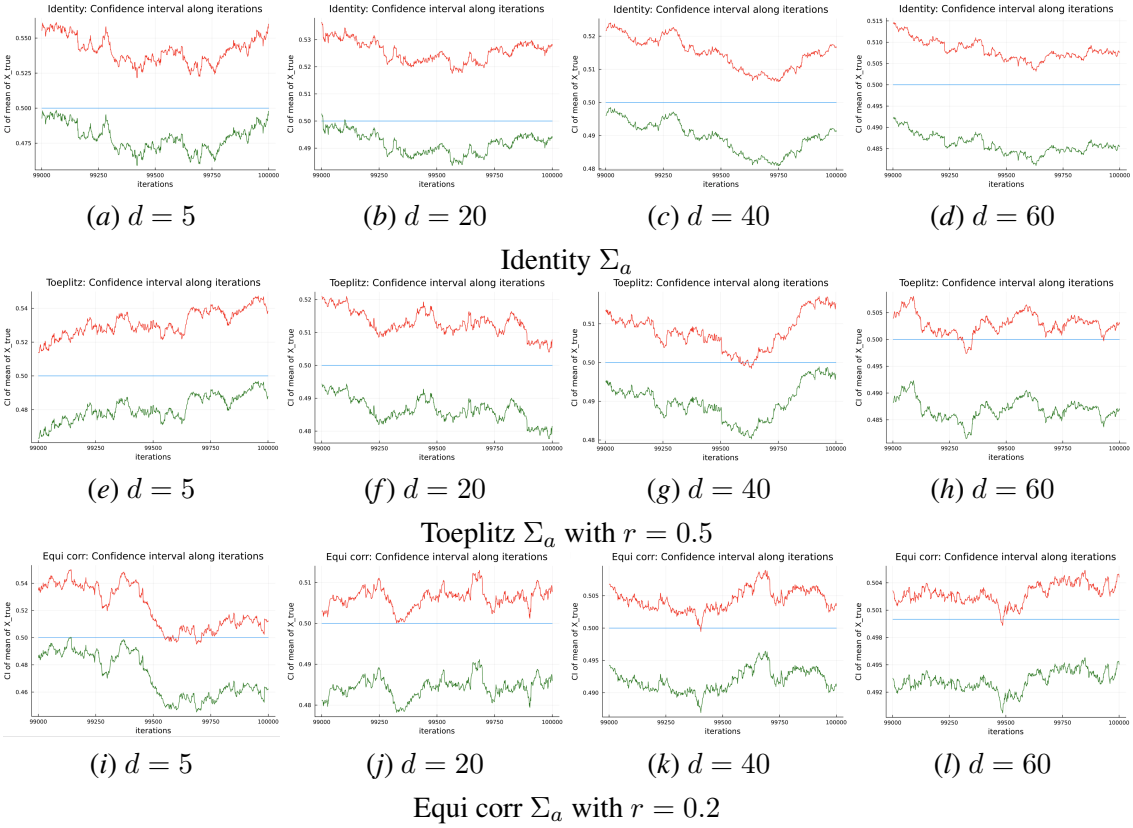
We demonstrate the empirical performance of the estimated covariance matrix in linear regression problems. Our experimental setup follows the settings in a previous study [6]. Specifically, we consider dimension $d \in \{5, 20, 40, 60\}$ and perform 10^5 iterations. The components for the true solution $\boldsymbol{x}^\star \in \mathbb{R}^d$ are set linearly spaced between 0 for the first component of $\boldsymbol{x}^\star$ and 1 for the d-th one. At each iteration, we independently generate $(\xi_{\boldsymbol{a}}, \xi_b)$ from the linear regression model defined in Section 1. With loss function $f(\boldsymbol{x}; \xi) = \frac{1}{2}(\xi_b - \xi_{\boldsymbol{a}}^T \boldsymbol{x})^2$, we can compute the explicit form of $\bar{g}_t$ and $\bar{H}_t$. We explore three different structures of covariance matrix Σ_a: 1) Identity matrix: $\Sigma_a = I$; 2) Toeplitz: $[\Sigma_a]_{i,j} = r^{|i-j|}$ with $r = 0.5$; 3) Equi-correlation: $[\Sigma_a]_{i,j} = r$ for $i \neq j$ and $[\Sigma_a]_{i,i} = 1$ for all i. We compare four solvers for the Newton system: the exact solver and the iterative randomized solver (2) with $\tau = 20, 40, 60$ respectively. For the randomized solver, the sketching matrices S are generated from the uniform distribution on $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$. Additionally, we choose $\beta_t = 1/(t+1)^{0.501}$ and $\chi_t = \beta_t^2$. The stepsize $\bar{\alpha}_t$ is randomly drawn from the uniform distribution on the interval $[\beta_t, \beta_t + \chi_t]$. For each run, we construct the 95% confidence interval for $\sum_{i=1}^d \boldsymbol{x}_i^\star / d$ following (8). Under each setting, we repeat the entire process 200 times and compute the average coverage rate and the average confidence interval length.

Table 1 shows the average coverage rate of the confidence intervals across 200 runs. The average length of 95% confidence intervals is left to Table 2 in Appendix B. The coverage rate of the plug-in estimator decreases as the dimension increases for randomized solvers. Also, the performance of Ξ_M^{PI} gets better for larger τ under the same problem setting. This is because more inner iterations in the randomized solver leads to more accurate approximate to the update direction, which is shown in Lemma 12. However, the average coverage rates stay around 95% for all 4 solvers for weighted sample covariance matrix Ξ_M. Ξ_M not only reduces the computation complexity by using sketching techniques, but also numerically performs better in statistical inference due to its consistency.

		exactly solved		$\tau = 10$		$\tau = 20$		$\tau = 40$	
Structure of Σ_a	Dim	PI	WSC	PI	WSC	PI	WSC	PI	WSC
Identity	5	94.50	94.50	97.00	96.50	95.50	95.50	95.00	94.00
	20	96.50	97.50	96.00	96.50	93.00	93.00	94.50	94.00
	40	95.50	93.50	95.50	92.00	95.00	93.50	95.50	95.50
	60	97.00	96.50	93.50	92.50	94.50	93.50	95.50	96.00
Toeplitz r = 0.5	5	94.00	94.00	90.00	94.00	94.00	94.50	95.00	95.00
	20	91.50	91.00	89.00	95.00	92.50	98.00	92.50	95.00
	40	95.00	94.00	91.50	95.50	85.50	93.50	87.00	96.00
	60	95.50	94.00	93.00	96.50	88.50	96.50	91.50	94.50
Equi-corr r = 0.2	5	95.50	97.50	93.00	94.00	94.50	93.50	96.00	95.50
	20	94.00	94.00	84.50	95.00	80.50	95.00	90.00	94.00
	40	95.00	94.00	73.00	95.00	77.50	92.00	80.50	94.50
	60	94.50	93.50	68.50	93.50	66.00	95.00	76.00	93.00

Table 1: The average coverage rate of the confidence interval for $\sum_{i=1}^d \boldsymbol{x}_i^\star / d$ for linear regression at the target level 95%. PI represents plug-in estimator Ξ_M^{PI} and WSC represents weighted sample covariance matrix Ξ_M.

Figure 1 plots the online construction of the confidence interval of $\sum_{i=1}^{d} \boldsymbol{x}_i^\star/d$ for the last 1000 iterations for one run of the experiment. We use the randomized solver with $\tau = 20$. We can see that the confidence intervals consistently cover the true value, which demonstrates the performance of the weighted sample covariance matrix.



(a) $d = 5$ (b) $d = 20$ (c) $d = 40$ (d) $d = 60$

Identity Σ_a

(e) $d = 5$ (f) $d = 20$ (g) $d = 40$ (h) $d = 60$

Toeplitz Σ_a with $r = 0.5$

(i) $d = 5$ (j) $d = 20$ (k) $d = 40$ (l) $d = 60$

Equi corr Σ_a with $r = 0.2$

Figure 1: 95% *confidence intervals of* $\sum_{i=1}^{d} \boldsymbol{x}_i^\star/d$ *for least squares. The inner iteration number of the randomized solver is chosen as* $\tau = 20$*. Each figure has three lines: the green and red lines correspond to the lower and upper interval boundaries, respectively; and the blue line corresponds to the true value.*

6. Conclusion

For the stochastic Newton method derived from [18], we proved the global almost sure convergence on the iterate sequence of the algorithm and proved asymptotic normality of the last iterate under weaker assumptions compared to [18]. We proposed a fully online consistent estimator of the covariance matrix in the asymptotic Gaussian distribution. The estimator construction is efficient in both storage and computation. Based on that, it became practical to conduct statistical inference on the minimum $\boldsymbol{x}^\star$. We showed the validity of the confidence interval by numerical experiments on a linear regression model.

Acknowledgement

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

[1] Albert S Berahas, Frank E Curtis, Michael J O'Neill, and Daniel P Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv preprint arXiv:2106.13015*, 2021.

[2] Albert S Berahas, Frank E Curtis, Daniel Robinson, and Baoyu Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

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

[4] Claire Boyer and Antoine 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, 2023.

[5] Peggy Cénac, Antoine Godichon-Baggioni, and Bruno Portier. An efficient averaged stochastic gauss-newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*, 2020.

[6] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. 2020.

[7] Frank E Curtis, Daniel P Robinson, and Baoyu Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*, 2021.

[8] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

[9] Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

[10] Yixin Fang. Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics*, 46(4):987–1002, 2019.

[11] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 2018.

[12] Yuchen Fang, Sen Na, Michael W Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *arXiv preprint arXiv:2211.15943*, 2022.

[13] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

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

[15] Yoonhyung Lee, Sungdong Lee, and Joong-Ho Won. Statistical inference with implicit sgd: proximal robbins-monro vs. polyak-ruppert. In *International Conference on Machine Learning*, pages 12423–12454. PMLR, 2022.

[16] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

[18] Sen Na and Michael W Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv preprint arXiv:2205.13687*, 2022.

[19] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199 (1-2):721–791, 2023.

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

[21] Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163, 2009.

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

[23] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[24] Panagiotis Toulis, Edoardo Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *International Conference on Machine Learning*, pages 667–675. PMLR, 2014.

[25] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. 2017.

[26] Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

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

Appendix A. Proofs

A.1. Preparation Lemmas

We first introduce the following lemmas on the series of stepsize, which are important tools for our proofs.

Lemma 9 *[18, Lemma A.1]Suppose $\{\varphi_i\}_i$ is a positive sequence that satisfies $\lim_{i\to\infty} i(1-\varphi_{i-1}/\varphi_i) = \varphi$. Then, for any $p \geq 0$, we have $\lim_{i\to\infty} i\left(1-\varphi_{i-1}^p/\varphi_i^p\right) = p\cdot\varphi$.*

Lemma 10 *[18, Lemma A.3] Let $\{\phi_i\}_i, \{\varphi_i\}_i, \{\sigma_i\}_i$ be three positive sequences. Suppose we have,*

$$\lim_{i\to\infty} i\left(1-\phi_{i-1}/\phi_i\right) = \phi, \qquad \lim_{i\to\infty}\varphi_i = 0, \qquad \lim_{i\to\infty} i\varphi_i = \tilde{\varphi} \tag{9}$$

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

$$\sum_{k=1}^{l}\sigma_k + \phi/\tilde{\varphi} > 0, \tag{10}$$

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

$$\frac{1}{\phi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i \longrightarrow \frac{1}{\sum_{k=1}^{l}\sigma_k + \phi/\tilde{\varphi}}, \tag{11}$$

$$\frac{1}{\phi_t}\left\{\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i a_i + b\cdot\prod_{j=0}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\right\} \longrightarrow 0, \tag{12}$$

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

In the next lemma, we characterize the convergence rate of (11) for a specific choice of $\{\varphi_i\}_i$.

Lemma 11 *Assume the conditions of Lemma 10 hold. Furthermore, let the sequence $\varphi_i = \eta/(i+1)^\alpha$ for constants $\eta > 0$ and $1/2 < \alpha < 1$. Define $\gamma_i = 1/(i+1)^{1-\alpha}$, then we have*

$$\frac{1}{\phi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i - \frac{1}{\sum_{k=1}^{l}\sigma_k} = O(\gamma_t).$$

Proof Since $\varphi_i = \eta/(i+1)^\alpha$, we have $\lim_{i\to\infty} i\varphi_i = \infty$. Thus, $\sum_{k=1}^{l}\sigma_k + \phi/\tilde{\varphi} = \sum_{k=1}^{l}\sigma_k$, i.e. condition (9) holds whatever the choice of ϕ. [18, Eq A.3] gives us the decomposition

$$\begin{aligned}
&\frac{1}{\phi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}\left(1-\varphi_j\sigma_k\right)\varphi_i\phi_i - \frac{1}{\sum_{k=1}^{l}\sigma_k} \\
&= \frac{1}{\phi_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\phi_i\left\{\varphi_i - \frac{1}{\sum_{k=1}^{l}\sigma_k}\left(1-\frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^{l}(1-\varphi_i\sigma_k)\right)\right\} \\
&\quad + \frac{1}{\phi_t}\prod_{j=1}^{t}\prod_{k=1}^{l}(1-\varphi_i\sigma_k)\cdot\phi_0\left(\varphi_0 - \frac{1}{\sum_{k=1}^{l}\sigma_k}\right).
\end{aligned} \tag{13}$$

Note that

$$\prod_{k=1}^{l}(1-\varphi_i\sigma_k) = 1-\sum_{k=1}^{l}\sigma_k\varphi_i + O(\varphi_i^2) \quad \text{and} \quad 1/i\varphi_i = (i+1)^{\alpha}/i = \gamma_i, \tag{14}$$

Here γ_i is defined to be $1/(i+1)^{1-\alpha}$. By (9) and (14), we have

$$\frac{\phi_{i-1}}{\phi_i} = 1-\phi\cdot\frac{1}{i}+o\left(\frac{1}{i}\right) = 1+O\left(\varphi_i\gamma_i\right).$$

Multiplying the two terms, we get

$$\frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^{l}(1-\varphi_i\sigma_k) = 1-\sum_{k=1}^{l}\sigma_k\varphi_i + O\left(\varphi_i\gamma_i\right),$$

where we also use the fact that $\varphi_i^2 = o\left(\varphi_i\gamma_i\right)$ for $\alpha\in(1/2,1)$. Thus, the first term in (13) can be simplified as

$$\begin{aligned}\frac{1}{\phi_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\phi_i\left\{\varphi_i-\frac{1}{\sum_{k=1}^{l}\sigma_k}\left(1-\frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^{l}(1-\varphi_i\sigma_k)\right)\right\}&\\ = \frac{1}{\phi_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i O\left(\phi_i\gamma_i\right).&\end{aligned} \tag{15}$$

By the definition of γ_i in (14), we know $\lim_{i\to\infty} i\left(1-\gamma_{i-1}/\gamma_i\right) = -\left(1-\alpha\right) =: \gamma$. Moreover,

$$\lim_{i\to\infty} i\left(1-\frac{\phi_{i-1}\gamma_{i-1}}{\phi_i\gamma_i}\right) = \lim_{i\to\infty} i\left(1-\frac{\phi_{i-1}}{\phi_i}-\frac{\phi_{i-1}}{\phi_i}\left(1-\frac{\gamma_{i-1}}{\gamma_i}\right)\right) = \phi+\gamma.$$

Thus, we are able to apply Lemma 10. (11) and (12) give us

$$\begin{aligned}&\frac{1}{\phi_t\gamma_t}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\cdot\phi_i\gamma_i \longrightarrow \frac{1}{\sum_{k=1}^{l}\sigma_k},\\ &\frac{1}{\phi_t\gamma_t}\prod_{j=0}^{t}\prod_{k=1}^{l}(1-\varphi_i\sigma_k) \longrightarrow 0.\end{aligned}$$

Combining the above display with (13) and (15), we complete the proof. ■

We now discuss the sketching randomized solver. We subtract $\tilde{\Delta}\boldsymbol{x}_t$ from both sides of the explicit solution of (2) and plug in $\bar{g}_t = -B_t\tilde{\Delta}\boldsymbol{x}_t$, then

$$\boldsymbol{z}_{t,j+1}-\tilde{\Delta}\boldsymbol{x}_t = z_{t,j}-\tilde{\Delta}\boldsymbol{x}_t-\left(B_tS_{t,j}(S_{t,j}^TB_t^2S_{t,j})^{\dagger}S_{t,j}^TB_t\right)(z_{t,j}-\tilde{\Delta}\boldsymbol{x}_t).$$

We define

$$\tilde{C}_{t,j} = I-\left(B_tS_{t,j}(S_{t,j}^TB_t^2S_{t,j})^{\dagger}S_{t,j}^TB_t\right) \quad \text{and} \quad \tilde{C}_t = -\prod_{j=1}^{\tau}\tilde{C}_{t,j}. \tag{16}$$

Note that $\tilde{C}_{t,j}$ is a projection matrix, thus $\|\tilde{C}_{t,j}\| \leq 1$ and $\|\tilde{C}_t\| \leq 1$. Given $\boldsymbol{z}_0 = \mathbf{0}$, we know $\bar{\Delta}\boldsymbol{x}_t = (I + \tilde{C}_t)\tilde{\Delta}\boldsymbol{x}_t = -(I + \tilde{C}_t)B_t^{-1}\bar{g}_t$. We let $C_t = \mathbb{E}[\tilde{C}_t \mid \mathcal{F}_{t-1}]$. Moreover, since $\bar{g}_t$ is conditionally unbiased, it easy to verify that

$$\mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-2/3}] = (I + C_t)\tilde{\Delta}\boldsymbol{x}_t \quad \text{and} \quad \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] = (I + C_t)\Delta\boldsymbol{x}_t, \tag{17}$$

where $\Delta\boldsymbol{x}_t = B_t^{-1}\nabla f_t$. The following lemma guarantees the performance of the randomized solver.

Lemma 12 *[18, Lemma 3.4] Under Assumption 3 (a), for all $t \geq 0$:*

1. *Let $\rho = 1 - \gamma_S$. We have $0 \leq \rho < 1$.*
2. *$\mathbb{E}[\bar{\Delta}\boldsymbol{x}_t - \tilde{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-2/3}] = C_t\tilde{\Delta}\boldsymbol{x}_t$, and $\|C_t\| \leq \rho^\tau$.*
3. *$\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t - \tilde{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-2/3}\big] \leq \rho^\tau\|\tilde{\Delta}\boldsymbol{x}_t\|^2$.*

A.2. Proofs of Section 2

A.2.1. PROOF OF THEOREM 4

The main outline is similar to [18]. With the compactness removed, we do not have the bound on ∇f anymore. Besides, we relax the bounded covariance condition to Assumption 2. We have to deal with these two points.

Proof Due to the Lipschitz continuity of ∇f, we have

$$f_{t+1} - f^\star \leq f_t - f^\star + \bar{\alpha}_t\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t + \frac{\bar{\alpha}_t^2}{2}\Upsilon_{Lg}\|\bar{\Delta}\boldsymbol{x}_t\|^2. \tag{18}$$

We subtract $f^\star$ from both sides and take the conditional expectation.

$$\begin{aligned}\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-1}] \leq (f_t - f^\star) + \mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \\ + \mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] + \frac{\Upsilon_{Lg}}{2}\mathbb{E}\big[\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big].\end{aligned} \tag{19}$$

We first deal with the second term on the right hand side. By (17), we know

$$\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] = \nabla f_t^T(I+C_t)\Delta\boldsymbol{x}_t = -\nabla f_t^T(I+C_t)B_t^{-1}\nabla f_t \leq -\frac{1}{\Upsilon_B}\|\nabla f_t\|^2 + \frac{\rho^\tau}{\gamma_B}\|\nabla f_t\|^2,$$

where the last inequality is due to $\|C_t\| \leq \rho^\tau$ and $\gamma_B \leq \lambda_{\min}(B_t) \leq \lambda_{\max}(B_t) \leq \Upsilon_B$. Furthermore, since $\beta_t \leq \bar{\alpha}_t$ and $\rho^\tau \leq \gamma_B/4\Upsilon_B$, we have

$$\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \leq -\frac{3}{4\Upsilon_B}\beta_t\|\nabla f_t\|^2. \tag{20}$$

Noting that $\mathbb{E}\big[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\big] = 0$, we have

$$\begin{aligned}\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[(\bar{\alpha}_t - \beta_t)\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] \\ &\leq \frac{\chi_t}{2}\|\nabla f_t\|\mathbb{E}\Big[\Big\|\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\| \mid \mathcal{F}_{t-1}\Big],\end{aligned} \tag{21}$$

where the inequality uses $|\bar{\alpha}_t - \beta_t| \leq \chi_t/2$. (17) leads to

$$\begin{aligned}\left\|\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\right\| &= \left\|(I+\tilde{C}_t)B_t^{-1}\bar{g}_t - (I+C_t)B_t^{-1}\nabla f_t\right\| \\ &\leq \|B_t^{-1}\|\|\nabla f_t\|\|C_t - \tilde{C}_t\| + \|I+\tilde{C}_t\|\|B_t^{-1}\|\|\bar{g}_t - \nabla f_t\| \leq \frac{2}{\gamma_B}\|\nabla f_t\| + \frac{2}{\gamma_B}\|\bar{g}_t - \nabla f_t\|,\end{aligned}$$

where the last inequality holds due to $\|B_t^{-1}\| \leq 1/\gamma_B$, $\|C_t\| \leq 1$, and $\|\tilde{C}_t\| \leq 1$. Besides, Assumption 2(a) gives us

$$\begin{aligned}\mathbb{E}\big[\|\bar{g}_t - \nabla f_t\| \mid \mathcal{F}_{t-1}\big] &\leq \sqrt{\mathbb{E}\big[\|\bar{g}_t - \nabla f_t\|^2 \mid \mathcal{F}_{t-1}\big]} \leq \sqrt{2\mathbb{E}\big[\|\bar{g}_t\|^2 \mid \mathcal{F}_{t-1}\big] + 2\|\nabla f_t\|^2} \\ &\leq \sqrt{2}\sqrt{2\|\nabla f_t\|^2 + C_{g,2}} \leq 2\|\nabla f_t\| + \sqrt{2}\sqrt{C_{g,2}}.\end{aligned}$$

Plugging the above two displays in (21), we have

$$\begin{aligned}\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big] &\leq \chi_t\Big(\frac{3}{\gamma_B}\|\nabla f_t\|^2 + \frac{\sqrt{2}}{\gamma_B}\sqrt{C_{g,2}}\|\nabla f_t\|\Big) \\ &\leq \chi_t\Big(\frac{4}{\gamma_B}\|\nabla f_t\|^2 + \frac{1}{2\gamma_B}C_{g,2}\Big). \quad \text{Young's inequality} \quad (22)\end{aligned}$$

Finally, using $\bar{\alpha}_t \leq \eta_t$ with $\eta_t := \beta_t + \chi_t$, we have $\mathbb{E}\big[\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] \leq \eta_t^2\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big]$. Since $\bar{\Delta}\boldsymbol{x}_t = -(I+\tilde{C}_t)B_t^{-1}\bar{g}_t$, we have

$$\mathbb{E}\big[\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \mid \mathcal{F}_{t-1}\big] \leq \eta_t^2\mathbb{E}\big[\|(I+\tilde{C}_t)\|^2\|B_t^{-1}\|^2\|\bar{g}_t\|^2 \mid \mathcal{F}_{t-1}\big] \leq \frac{2^2\eta_t^2}{\gamma_B^2}\big(\|\nabla f_t\|^2 + C_{g,2}\big). \quad (23)$$

The last inequality follows from $\lambda_{\min}(B_t) \geq \gamma_B$, $\|\tilde{C}_t\| \leq 1$, and Assumption 3 (a) with $m = 2$. We plug (20), (22), and (23) back into (19). We rearrange the items and obtain

$$\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \leq (f_t - f^\star) - \Big(\frac{3}{4\Upsilon_B}\beta_t - \frac{4}{\gamma_B}\chi_t - \frac{4}{\gamma_B^2}\eta_t^2\Big)\|\nabla f_t\|^2 + \frac{4C_{g,2}}{\gamma_B}(\chi_t + \eta_t^2).$$

By $\beta_t = c_\beta/(t+1)^\beta$ and $\chi_t = o(\beta_t)$, there exists a fixed integer t_0 such that $\frac{4}{\gamma_B}\chi_t - \frac{2^2}{\gamma_B^2}\eta_t^2 \leq \frac{2}{\Upsilon_B}\beta_t$ for all $t \geq t_0$. Thus, for $t \geq t_0$, we have the recursion

$$\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-1}] \leq (f_t - f^\star) - \frac{1}{4\Upsilon_B}\beta_t\|\nabla f_t\|^2 + \frac{2^2}{\gamma_B^2}C_{g,2}\cdot(\chi_t + \eta_t^2).$$

We note that $\sum_{t=t_0}^\infty \chi_t < \infty$ and $\sum_{t=t_0}^\infty \eta_t^2 \lesssim \sum_{t=t_0}^\infty \beta_t^2 + \sum_{t=t_0}^\infty \chi_t^2 < \infty$ due to $\beta > 1/2$ and $\chi > 1$. Therefore, the Robbins-Siegmund theorem [9, Theorem 1.3.12] leads to

$$f_t - f^\star \text{converges to a finite random variable, and } \sum_{t=t_0}^\infty \beta_t\|\nabla f_t\|^2 < \infty \text{ a.s.}$$

Furthermore, since $\sum_{t=t_0}^\infty \beta_t = \infty$, we can conclude that $\liminf_{t\to\infty}\|\nabla f_t\| = 0$ almost surely. The strong convexity[26] of f gives us

$$\frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \leq f(x_t) - f^* \leq \frac{1}{2\mu}\|\nabla f_t\|^2. \quad (24)$$

By the second inequality in (24) and $\liminf_{t\to\infty} \|\nabla f_t\| = 0$, we get $\liminf_{t\to\infty} f_t - f^\star = 0$. Since $f_t - f^\star$ converges a.s., the conclusion can be strengthened to $\lim_{t\to\infty} f_t - f^\star = 0$. We use the first inequality in (24) and obtain

$$\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^\star \quad \text{almost surely.}$$

This completes the proof. ■

A.2.2. PROOF OF THEOREM 7

The proof is almost the same as Theorem 4.6 in [18]. The only difference is the removal of compactness and the relaxation of the bounded variance on $\bar{g}_t$. The techniques of dealing with the difference is similar to what we do in the proof of Theorem 4. Although we cannot bound ∇f_t, we know that $\nabla f_t \to 0$ as $t \to \infty$. Thus, it does not affect the almost sure convergence in the proof. We will skip the details and refer readers to [18].

A.3. Convergence rate of iterates and Hessian estimates

In this section, we introduce two lemmas bounding $\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4\big]$ and $\mathbb{E}\big[\|B_t - B^\star\|^4\big]$.

Lemma 13 *Under assumptions 1, 2(a) with $m = 4$, 3(a), we suppose $\beta_t = c_\beta/(t+1)^\beta$ and $\chi_t = c_\chi/(t+1)^\chi$ with constants $0 < c_\chi < c_\beta^2$, $0 < \beta < 1$, and $\chi > 2\beta$. If $\tau \geq \log(\gamma_B/4\Upsilon_B)/\log\rho$, then we have*

$$\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^4\big] \lesssim D\beta_t^2,$$

where

$$D = \frac{\Upsilon_{Lg}^2}{\mu^2} \max\Big(\frac{\Upsilon_B}{\mu\gamma_B^4} \max(C_{g,4}, C_{g,2}^2), \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^4\Big). \tag{25}$$

Proof Taking squares on both sides of (18), we have

$$\begin{aligned}(f_{t+1} - f^\star)^2 \leq (f_t - f^\star)^2 &+ 2\bar{\alpha}_t(f_t - f^\star)\nabla f_t^T \bar{\Delta}\boldsymbol{x}_t + \bar{\alpha}_t^2\Upsilon_{Lg}(f_t - f^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2 \\ &+ \bar{\alpha}_t^2(\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t)^2 + \bar{\alpha}_t^3\Upsilon_{Lg}\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t\|\bar{\Delta}\boldsymbol{x}_t\|^2 + \frac{1}{4}\bar{\alpha}_t^4\Upsilon_{Lg}^2\|\bar{\Delta}\boldsymbol{x}_t\|^4. \end{aligned} \tag{26}$$

We take the conditional expectation given $\mathcal{F}_{t-1}$, then classify these terms by the order of stepsize and analyze them one by one.

Part 1. $\mathbb{E}[\, 2\bar{\alpha}_t(f_t - f^\star)\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]$

Same as the proof of Theorem 4, we have the decomposition

$$\begin{aligned}\mathbb{E}[\, 2\bar{\alpha}_t(f_t - f^\star)\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] &= 2(f_t - f^\star)\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \\ &\quad + 2(f_t - f^\star)\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t - \mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}]\Big\} \mid \mathcal{F}_{t-1}\Big].\end{aligned}$$

By (20) and (24), the following holds with $\rho^\tau \leq \gamma_B/4\Upsilon_B$.

$$2(f_t - f^\star)\mathbb{E}\Big[\bar{\alpha}_t\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}\Big] \leq -\frac{3}{2\Upsilon_B}\beta_t(f_t - f^\star)\|\nabla f_t\|^2 \leq -\frac{3\mu}{\Upsilon_B}\beta_t(f_t - f^\star)^2.$$

By (22), we know

$$2(f_t-f^\star)\mathbb{E}\Big[\bar{\alpha}_t\Big\{\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t-\mathbb{E}[\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t\mid\mathcal{F}_{t-1}]\Big\}\mid\mathcal{F}_{t-1}\Big]\le\chi_t(f_t-f^\star)\Big(\frac{8}{\gamma_B}\|\nabla f_t\|^2+\frac{1}{\gamma_B}C_{g,2}\Big).$$

The Lipschitz continuity [26] of ∇f leads to

$$\frac{1}{2\Upsilon_{Lg}}\|\nabla f_t\|^2\le f_t-f^*\le\frac{\Upsilon_{Lg}}{2}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2. \tag{27}$$

By (27) and Young's inequality, we have

$$\begin{aligned}\chi_t(f_t-f^\star)\Big(\frac{8}{\gamma_B}\|\nabla f_t\|^2+\frac{1}{\gamma_B}C_{g,2}\Big)&\le\frac{16\Upsilon_{Lg}}{\gamma_B}\chi_t(f_t-f^\star)^2+\frac{\mu}{2\Upsilon_B}\chi_t^{1/2}(f_t-f^\star)^2+\frac{C_{g,2}^2\Upsilon_B}{2\mu\gamma_B^2}\chi_t^{3/2}\\&\le\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_t^2(f_t-f^\star)^2+\frac{\mu}{2\Upsilon_B}\beta_t(f_t-f^\star)^2+\frac{C_{g,2}^2\Upsilon_B}{2\mu\gamma_B^2}\eta_t^3.\end{aligned}$$

In the last inequality, we use $\chi_t\le\eta_t$ and $\chi_t\le\beta_t^2\le\eta_t^2$, which is easy to check according to $0<c_\chi\le c_\beta^2$ and $\chi>2\alpha$.

Part 2. $\mathbb{E}[\,\bar{\alpha}_t^2\Upsilon_{Lg}(f_t-f^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2+\bar{\alpha}_t^2(\nabla f_t^T\Delta\bar{\boldsymbol{x}}_t)^2\mid\mathcal{F}_{t-1}]$
Thus, the second term can be upper bounded by

$$(\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t)^2\le\|\nabla f_t\|^2\|\bar{\Delta}\boldsymbol{x}_t\|^2\le2\Upsilon_{Lg}(f_t-f^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2.$$

Now we use $\bar{\alpha}_t\le\eta_t:=\beta_t+\chi_t$ and merge two terms.

$$\begin{aligned}\mathbb{E}[\,\bar{\alpha}_t^2\Upsilon_{Lg}(f_t-f^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2+\bar{\alpha}_t^2(\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t)^2\mid\mathcal{F}_{t-1}]&\le3\Upsilon_{Lg}\eta_t^2\mathbb{E}\big[(f_t-f^\star)\|\bar{\Delta}\boldsymbol{x}_t\|^2\mid\mathcal{F}_{t-1}\big]\\&\le\frac{\mu}{2\Upsilon_B}\eta_t(f_t-f^\star)^2+\frac{9\Upsilon_B\Upsilon_{Lg}}{2\mu}\eta_t^3\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4\mid\mathcal{F}_{t-1}\big].\quad\text{(Young's inequality)}\end{aligned}$$

Part 3. $\mathbb{E}[\,\bar{\alpha}_t^3\Upsilon_{Lg}\nabla f_t^T\bar{\Delta}\boldsymbol{x}_t\|\bar{\Delta}\boldsymbol{x}_t\|^2\mid\mathcal{F}_{t-1}]$
Similarly, we use Young's inequality and apply (27) to bound this term.

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

Combining the analyses of all parts and (26), we have

$$\begin{aligned}\mathbb{E}[(f_{t+1}-f^\star)^2\mid\mathcal{F}_{t-1}]\le&\left(1-\frac{5\mu}{2\Upsilon_B}\beta_t+\frac{\mu}{2\Upsilon_B}\eta_t+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_t^2+\Upsilon_{Lg}^2\eta_t^3\right)(f_t-f^\star)^2\\&+\frac{C_{g,2}^2\Upsilon_B}{2\mu\gamma_B^2}\eta_t^3+3\frac{\Upsilon_B\Upsilon_{Lg}^2}{\mu}\eta_t^3(1+\eta_t)\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4\mid\mathcal{F}_{t-1}\big].\end{aligned}$$

Next, we bound $\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4\mid\mathcal{F}_{t-1}\big]$. Under the Assumption 3 with m=4, by the fact $\|\tilde{C}_t\|\le1$ and $\|B_t^{-1}\|\le1/\gamma_B$, we have

$$\begin{aligned}\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_t\|^4\mid\mathcal{F}_{t-1}\big]&=\mathbb{E}\big[\|(I+\tilde{C}_t)\|^4\|B_t^{-1}\|^4\|\bar{g}_t\|^4\mid\mathcal{F}_{t-1}\big]\\&\le\frac{2^4}{\gamma_B^4}(\|\nabla f_t\|^4+C_{g,4})\overset{(27)}{\le}\frac{2^6\Upsilon_{Lg}^2}{\gamma_B^4}(f_t-f^\star)^2+\frac{2^4}{\gamma_B^4}C_{g,4}.\end{aligned}$$

Therefore, we rearrange the terms and get the recursion

$$\mathbb{E}[(f_{t+1}-f^\star)^2] \le \Big(1-\frac{5\mu}{2\Upsilon_B}\beta_t+\frac{\mu}{2\Upsilon_B}\eta_t+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_t^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_t^3(1+\eta_t)\Big)\mathbb{E}[(f_t-f^\star)^2] + \frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4},C_{g,2}^2)\eta_t^3(1+\eta_t).$$

Applying the inequality above recursively, then we get

$$\begin{aligned}\mathbb{E}[(f_{t+1}-f^\star)^2] &\le \prod_{i=0}^{t}\left|1-\frac{5\mu}{2\Upsilon_B}\beta_i+\frac{\mu}{2\Upsilon_B}\eta_i+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_i^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_i^3(1+\eta_i)\right|(f_0-f^\star)^2\\ &+\frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4},C_{g,2}^2)\sum_{i=0}^{t}\prod_{j=i+1}^{t}\left|1-\frac{5\mu}{2\Upsilon_B}\beta_j+\frac{\mu}{2\Upsilon_B}\eta_j+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_j^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_j^3(1+\eta_j)\right|\eta_i^3(1+\eta_i)\\ &=: V_t+W_t. \end{aligned} \tag{28}$$

Since $\beta_t = c_\beta/(t+1)^\beta$ and $\chi_t = o(\beta_t)$, there exists a deterministic integer $\tilde{t}$ such that for all $t \ge \tilde{t}$

$$\frac{\mu}{2\Upsilon_B}\eta_t+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_t^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_t^3(1+\eta_t) \le \frac{3\mu}{2\Upsilon_B}\beta_t \quad \text{and} \quad 1-\frac{\mu}{\Upsilon_B}\beta_t > 0.$$

Then, we have

$$\begin{aligned}W_t &\le \frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4},C_{g,2}^2)\sum_{\tilde{t}-1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\mu}{\Upsilon_B}\beta_j\right)\eta_i^3(1+\eta_i)\\ &+\frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4},C_{g,2}^2)\sum_{i=0}^{\tilde{t}-2}\prod_{j=i+1}^{t}\left|1-\frac{5\mu}{2\Upsilon_B}\beta_j+\frac{\mu}{2\Upsilon_B}\eta_j+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_j^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_j^3(1+\eta_j)\right|\eta_i^3(1+\eta_i).\end{aligned}$$

As explained in the proof of [18, Lemma D.1], the first few terms do not affect the rate of series. When $t \ge \tilde{t}$, let $\tilde{\eta}_t = \eta_t$. When $t = \tilde{t}-1$, let $\tilde{\eta}_t$ satisfy

$$\tilde{\eta}_t^3(1+\tilde{\eta}_t) = \sum_{i=0}^{\tilde{t}-2}\prod_{j=i+1}^{\tilde{t}-1}\left|1-\frac{5\mu}{2\Upsilon_B}\beta_j+\frac{\mu}{2\Upsilon_B}\eta_j+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_j^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_j^3(1+\eta_j)\right|\eta_i^3(1+\eta_i)+\eta_t^3(1+\eta_t).$$

Thus, we can rewrite W_t as

$$W_t \le \frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4},C_{g,2}^2)\sum_{\tilde{t}-1}^{t}\prod_{j=i+1}^{t}\left(1-\frac{\mu}{\Upsilon_B}\beta_j\right)\tilde{\eta}_i^3(1+\tilde{\eta}_i).$$

We aim to use Lemma 10 to obtain the rate of W_t, hence we first verify the conditions (9) and (10). With the choice $\beta_t = c_\beta/(t+1)^\beta$ ($\beta < 1$), we have $\lim_{t\to\infty} t\beta_t = \infty$, thus condition (10) always hold. In the remainder of this paper, we will omit this point and only verify condition (9)

when applying Lemma 10. Since $\lim_{t\to\infty} t(1-\beta_{t-1}/\beta_t) = -\beta$ We have the limit, $\lim_{t\to\infty} t(1-\chi_{t-1}/\chi_t) = -\chi$, $\eta_t = \beta_t + \chi_t$ with $\chi_t = o(\beta_t)$, we have

$$\lim_{t\to\infty} t\left(1-\frac{\eta_{t-1}/\beta_{t-1}}{\eta_t/\beta_t}\right) = \lim_{t\to\infty} t\left(1-\frac{1+\chi_{t-1}/\beta_{t-1}}{1+\chi_t/\beta_t}\right) = \lim_{t\to\infty} t\left(\frac{\chi_t}{\beta_t}-\frac{\chi_{t-1}}{\beta_{t-1}}\right)$$
$$= \lim_{t\to\infty} t\left(1-\frac{\chi_{t-1}/\beta_{t-1}}{\chi_t/\beta_t}\right)\frac{\chi_t}{\beta_t} = \lim_{t\to\infty}\left(1-\frac{\chi_{t-1}}{\chi_t}+\frac{\chi_{t-1}}{\chi_t}\left\{1-\frac{\beta_{t-1}}{\beta_t}\right\}\right) = (\beta-\chi)\lim_{t\to\infty}\frac{\chi_t}{\beta_t} = 0.$$

Furthermore,

$$\lim_{t\to\infty} t\left(1-\frac{\eta_{t-1}}{\eta_t}\right) = \lim_{t\to\infty} t\left(1-\frac{\beta_{t-1}}{\beta_t}+\frac{\beta_{t-1}}{\beta_t}\left\{1-\frac{\eta_{t-1}/\beta_{t-1}}{\eta_t/\beta_t}\right\}\right) = -\beta,$$
$$\lim_{t\to\infty} t\left(1-\frac{1+\eta_{t-1}}{1+\eta_t}\right) = \lim_{t\to\infty} t\,\frac{\eta_t-\eta_{t-1}}{1+\eta_t} = \lim_{t\to\infty} t\left(1-\frac{\eta_{t-1}}{\eta_t}\right)\eta_t = -\beta\lim_{t\to\infty}\eta_t = 0.$$

Combining the three displays above and using Lemma 9, we know

$$\lim_{t\to\infty} t\left(1-\frac{\tilde\eta_{t-1}^3(1+\tilde\eta_{t-1})/\beta_{t-1}}{\tilde\eta_t^3(1+\tilde\eta_t)/\beta_t}\right)$$
$$= \lim_{t\to\infty} t\left\{\left(1-\frac{\eta_{t-1}^2}{\eta_t^2}\right)+\frac{\eta_{t-1}^2}{\eta_t^2}\left(1-\frac{\eta_{t-1}/\beta_{t-1}}{\eta_t/\beta_t}\right)+\frac{\eta_{t-1}^3/\beta_{t-1}}{\eta_t^3/\beta_t^3}\left(1-\frac{1+\eta_{t-1}}{1+\eta_t}\right)\right\} = -2\beta.$$

Therefore, we can apply Lemma 10. By (11), we have

$$W_t \lesssim \frac{50\Upsilon_B\Upsilon_{Lg}^2}{\mu\gamma_B^4}\max(C_{g,4}, C_{g,2}^2)\beta_t^2. \tag{29}$$

Similarly, (12) leads to

$$V_t \le \prod_{i=\tilde t}^{t}\left(1-\frac{\mu}{\Upsilon_B}\beta_i\right)\prod_{i=0}^{\tilde t-1}\left|1-\frac{5\mu}{2\Upsilon_B}\beta_i+\frac{\mu}{2\Upsilon_B}\eta_i+\frac{16\Upsilon_{Lg}}{\gamma_B}\eta_i^2+\frac{193\Upsilon_B\Upsilon_{Lg}^4}{\mu\gamma_B^4}\eta_i^3(1+\eta_i)\right|(f_0-f^\star)^2$$
$$= (f_0-f^\star)^2\cdot o(\beta_t^2). \tag{30}$$

Plugging (29), (30) into (28), and using (24), we have

$$\mathbb{E}\big[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^4\big] \le \frac{4}{\mu^2}\mathbb{E}\big[(f_t-f^\star)^2\big] \lesssim D\beta_t^2,$$

where D is defined in (25). We complete the proof. ■

Lemma 14 *Under the conditions in Lemma 13, we suppose Assumption 5 holds with $m=2$. Then we have*

$$\mathbb{E}\big[\|B_t-B^\star\|^2\big] \lesssim \max\big(C_{H,2}, \sqrt{D}\Upsilon_{LH}^2\big)(\beta_t+\omega_t^2), \tag{31}$$

where D is defined in (25). If Assumption 5 is strengthen to $m=4$, then we have

$$\mathbb{E}\big[\|B_t-B^\star\|^4\big] \lesssim \max\big(C_{H,4}, D\Upsilon_{LH}^4\big)(\beta_t^2+\omega_t^4). \tag{32}$$

Proof We first prove (32). By the construction of B_t in section 2, $B_t - B^\star$ can be decomposed into

$$B_t - B^* = \frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i) + \frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - \nabla^2 f^\star) + \Delta_t.$$

Thus, we have

$$\mathbb{E}\big[\|B_t - B^\star\|^4\big] \lesssim \mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|^4\bigg] + \mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - \nabla^2 f^\star)\Big\|^4\bigg] + \omega_t^4. \tag{33}$$

By Assumption 5, we know $\bar{H}_i - \nabla^2 f_i$ is a martingale difference sequence and $\mathbb{E}\big[\,\big\|\bar{H}_i - \nabla^2 f_i\big\|_F^4\,\big]$ is bounded, which allows us to apply [21, Theorem 2.1] and get

$$\mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|^4\bigg] \le \mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|_F^4\bigg] \lesssim \frac{1}{t^4}\bigg[\sum_{i=0}^{t-1}\Big(\mathbb{E}\big[\|\bar{H}_i - \nabla^2 f_i\|_F^4\big]\Big)^{1/2}\bigg]^2 \le \frac{C_{H,4}}{t^2}.$$

Regarding the second term in (33), we have

$$\begin{aligned}\mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - \nabla^2 f^\star)\Big\|^4\bigg] &\le \mathbb{E}\bigg[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\|\nabla^2 f_i - \nabla^2 f^\star\|\Big)^4\bigg] \le \frac{\Upsilon_{Lh}^4}{t^4}\mathbb{E}\bigg[\Big(\sum_{i=0}^{t-1}\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|\Big)^4\bigg] \\ &\le \Upsilon_{Lh}^4\bigg[\Big(\frac{1}{t}\sum_{i=0}^{t-1}\big(\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^4\big)^{1/4}\Big)^4\bigg] \lesssim D\Upsilon_{Lh}^4\Big(\frac{1}{t}\sum_{i=0}^{t-1}\beta_i^{1/2}\Big)^4.\end{aligned}$$

The second inequality holds since $\nabla^2 f$ is Υ_{Lh} Lipschitz continuous; the third inequality is due to Hölder's inequality; the last inequality comes from Lemma 13. Since $\beta < 1$, we thus know that $\frac{1}{t}\sum_{i=0}^{t-1}\beta_i^{1/2} = \frac{1}{t}\sum_{i=1}^{t} c_\beta/t^\beta \lesssim \beta_t^{1/2}$. We make a quick remark that the above technique computing the average rate is frequently used in our proofs. We will omit the derivation details in the remainder of this paper. Combining all the derivations, we get

$$\mathbb{E}\|B_t - B^\star\|^4 \lesssim \frac{C_{H,4}}{t^2} + D\Upsilon_{Lh}^4\beta_t^2 + \omega_t^4 \lesssim \max\big(C_{H,4}, D\Upsilon_{Lh}^4\big)(\beta_t^2 + \omega_t^4).$$

The proof for (31) is almost the same. We point out one significant difference, which lies in the martingale difference sequence. Note that [21, Theorem 2.1] requires the order of the moment higher than 2, thus we cannot apply this theorem. Instead, we have

$$\mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|^4\bigg] \le \mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|_F^4\bigg] = \frac{1}{t^2}\mathrm{Tr}\bigg(\mathbb{E}\Big[\big(\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\big)^2\Big]\bigg).$$

Since $\bar{H}_i - \nabla^2 f_i$ is martingale, we cancel the interaction terms and obtain

$$\mathbb{E}\bigg[\Big\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{H}_i - \nabla^2 f_i)\Big\|^4\bigg] = \frac{1}{t^2}\bigg(\sum_{i=0}^{t-1}\mathbb{E}\Big[\mathrm{Tr}\big((\bar{H}_i - \nabla^2 f_i)^2\big)\Big]\bigg) = \frac{1}{t^2}\bigg(\sum_{i=0}^{t-1}\mathbb{E}\big[\|\bar{H}_i - \nabla^2 f_i\|_F^2\big]\bigg) \le \frac{C_{H,2}}{t},$$

where the last inequality is due to Assumption 5. This completes the proof. ■

A.4. The linear case

We adapt [18, Lemma 4.1] to our unconstrained setting.

Lemma 15 *[18, Lemma 4.1] Let $\varphi_t = (\beta_t + \eta_t)/2$. The iterate sequence can be expressed as*

$$\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star = \mathcal{I}_{1,t} + \mathcal{I}_{2,t} + \mathcal{I}_{3,t}, \tag{34}$$

where

$$\mathcal{I}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\, \varphi_i \boldsymbol{\theta}^i \tag{35}$$

$$\mathcal{I}_{2,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\, (\bar{\alpha}_i - \varphi_i)\, \bar{\Delta}\boldsymbol{x}_i \tag{36}$$

$$\mathcal{I}_{3,t} = \prod_{i=0}^{t} \{I - \varphi_i(I + C^\star)\}\, (\boldsymbol{x}_0 - \boldsymbol{x}^\star) + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\, \varphi_i \boldsymbol{\delta}^i \tag{37}$$

and

$$C^\star = -(I - \mathbb{E}[B^\star S (S^T (B^\star)^2 S)^\dagger S^T B^\star])^\tau, \tag{38}$$

$$\boldsymbol{\theta}^i = \bar{\Delta}\boldsymbol{x}_i - \mathbb{E}[\bar{\Delta}\boldsymbol{x}_i \mid \mathcal{F}_{i-1}] = -(I + C_i) B_i^{-1} (\bar{g}_i - \nabla f_i) + \{\bar{\boldsymbol{x}}_i - (I + C_i)\tilde{\Delta}\boldsymbol{x}_i\}, \tag{39}$$

$$\boldsymbol{\delta}^i = -(I + C_i)\big\{(B^\star)^{-1}\boldsymbol{\psi}^i + \{B_i^{-1} - (B^\star)^{-1}\}\nabla f_i\big\} - \big(C_i - C^\star\big)(\boldsymbol{x}_i - \boldsymbol{x}^\star), \tag{40}$$

$$\boldsymbol{\psi}^i = \nabla f_i - B^\star(\boldsymbol{x}_i - \boldsymbol{x}^\star). \tag{41}$$

$\mathcal{I}_{1,t}$ is a martingale; $\mathcal{I}_{2,t}$ characterizes the influence by the randomized stepsize; $\mathcal{I}_{3,t}$ contains all the remaining errors. We point the readers to [18] for details. With the decomposition above, we have

$$\frac{1}{M}\sum_{t=1}^{M} \frac{1}{\varphi_{t-1}} (\boldsymbol{x}_t - \boldsymbol{x}^\star)(\boldsymbol{x}_t - \boldsymbol{x}^\star)^T = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \mathcal{I}_{i,t}\mathcal{I}_{j,t}. \tag{42}$$

Among these terms, $\frac{1}{M}\sum_{t=0}^{M-1} \frac{1}{\varphi_t}\mathcal{I}_{1,t}\mathcal{I}_{1,t}^T$ approximates $\Xi^\star$, which will be shown in Lemma 18. In the next section, we will see that the remaining are higher-order terms.

Next, we give a finer decomposition on $\mathcal{I}_{1,t}$. Let

$$\tilde{C}_{t,j}^\star = I - \big(B^\star S_{t,j} (S_{t,j}^T (B^\star)^2 S_{t,j})^\dagger S_{t,j}^T B^\star\big) \quad \text{and} \quad \tilde{C}_t^\star = -\prod_{j=1}^{\tau} \tilde{C}_{t,j}^\star, \tag{43}$$

where $S_{t,j}$ is the same matrix in generating $\tilde{C}_{t,j}$ in (16). Thus, the randomness of $\tilde{C}_t^\star$ comes from ζ_t. We point out that $\mathbb{E}\tilde{C}_t^\star = C^\star$ by definition. Define

$$\tilde{\boldsymbol{\theta}}^i = -(I + \tilde{C}_i^\star)(B^\star)^{-1}\nabla f(\boldsymbol{x}^\star; \xi_i), \quad \text{and} \quad \hat{\boldsymbol{\theta}}^i = \boldsymbol{\theta}^i - \tilde{\boldsymbol{\theta}}^i. \tag{44}$$

It is easy to check that both $\tilde{\boldsymbol{\theta}}^i$ and $\hat{\boldsymbol{\theta}}^i$ are martingale difference sequences. If we compare the definition of $\tilde{\boldsymbol{\theta}}^i$ and $\boldsymbol{\theta}^i$, we will find that $\tilde{\boldsymbol{\theta}}^i$ share the same randomness as $\boldsymbol{\theta}^i$ but constructed at

$\boldsymbol{x}^\star$ instead of $\boldsymbol{x}_i$. We note that $\{\boldsymbol{\theta}^i\}_i$ are correlated since $\boldsymbol{x}_t$ is dependent on the previous iterates. However, $\{\tilde{\boldsymbol{\theta}}^i\}_i$ are independent due to the independence among $\{\xi_i\}_i$, which is easier to deal with. Now we decompose $\boldsymbol{\mathcal{I}}_{1,t}$ as

$$\mathcal{I}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\varphi_i \tilde{\boldsymbol{\theta}}^i + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\varphi_i \hat{\boldsymbol{\theta}}^i =: \tilde{\mathcal{I}}_{1,t} + \hat{\mathcal{I}}_{1,t}. \quad (45)$$

Intuitively, if $\boldsymbol{x}_i$ converges to $\boldsymbol{x}^\star$ as i goes to ∞, $\tilde{C}_i^\star$ should get closer to $\tilde{C}_i$ as well. Thus, $\tilde{\mathcal{I}}_{1,t}$ should be a good approximation to $\mathcal{I}_{1,t}$ and $\hat{\mathcal{I}}_{1,t}$ is negligible. The following two lemmas quantify the performance of $\tilde{\mathcal{I}}_{1,t}$ and $\hat{\mathcal{I}}_{1,t}$, with proof lying in Section A.4.1 and A.4.2.

Lemma 16 *Under Assumptions 1, 2(b) with $m = 4$, and 3(a), we suppose $\beta_t = c_\beta/(t+1)^\beta$, $\chi_t = c_\beta/(t+1)^\chi$ with $1/2 < \beta < 1$ and $\chi < \beta$. Define $\gamma_t = 1/(t+1)^{1-\beta}$, then we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\tilde{\mathcal{I}}_{1,t}\tilde{\mathcal{I}}_{1,t}^T - \Xi^\star\right\|\right] \lesssim \max\left(\|\Gamma\|_F, \frac{\sqrt{C_{g,4}}}{\mu^2}\right)\sqrt{\gamma_M}.$$

Lemma 17 *Suppose the conditions in Lemma 13, 14, and 19 hold. Let $\omega_t = c_\omega/(t+1)^\omega$ and $\omega > \alpha/2$, then we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\hat{\mathcal{I}}_{1,t}\hat{\mathcal{I}}_{1,t}^T\right\|^2\right] \le \frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\hat{\mathcal{I}}_{1,t}\|^2\right] \lesssim C_{\hat{\boldsymbol{\theta}}}\beta_M,$$

where

$$C_{\hat{\boldsymbol{\theta}}} = \max\left\{\frac{C_{g,2}}{\mu^2}\left(\frac{1}{\gamma_B^2} + \frac{\tau^2\Upsilon_{S,2}}{\mu^2}\right)\max\left(C_{H,2}, \sqrt{D}\Upsilon_{LH}^2\right), \frac{\sqrt{D}}{\gamma_B^2}(\Upsilon_{m,2} + \Upsilon_{Lg}^2)\right\}. \quad (46)$$

With the help of two lemmas above, we are able to show the convergence rate of linear case.

Lemma 18 *Assume the conditions in Lemma 16 and 17 hold. We have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{1,t}\mathcal{I}_{1,t}^T - \Xi^\star\right\|\right] \lesssim \max\left(\|\Gamma\|_F, \frac{\sqrt{C_{g,4}}}{\mu^2}, C_{\hat{\boldsymbol{\theta}}}\right)\sqrt{\gamma_M},$$

where $C_{\hat{\boldsymbol{\theta}}}$ is defined in (46).

Proof By (45), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{1,t}\mathcal{I}_{1,t}^T - \Xi^\star\right\|\right] &\le \mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\tilde{\mathcal{I}}_{1,t}\tilde{\mathcal{I}}_{1,t}^T - \Xi^\star\right\|\right] \\ &\quad + \mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\hat{\mathcal{I}}_{1,t}\hat{\mathcal{I}}_{1,t}^T\right\|\right] + 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\tilde{\mathcal{I}}_{1,t}\hat{\mathcal{I}}_{1,t}^T\right\|\right].\end{aligned}$$

The rates of the first two terms are given by Lemma 16 and 17. Regarding the last term, we use Hölder inequality twice and obtain

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\tilde{\mathcal{I}}_{1,t}\hat{\mathcal{I}}_{1,t}^T\right\|\right] \le \mathbb{E}\left[\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\|\tilde{\mathcal{I}}_{1,t}\|\|\hat{\mathcal{I}}_{1,t}\|\right]$$
$$\le \sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\|\tilde{\mathcal{I}}_{1,t}\|^2}\sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\|\hat{\mathcal{I}}_{1,t}\|^2}. \quad (47)$$

Thus, in order to complete the proof, we only need to bound $\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\|\tilde{\mathcal{I}}_{1,t}\|^2$.

$$\mathbb{E}\big[\|\tilde{\mathcal{I}}_{1,t}\|^2\big] = \sum_{i_1,i_2=0}^{t}\varphi_{i_1}\varphi_{i_2}\mathbb{E}\Big[\tilde{\boldsymbol{\theta}}^{i_1 T}\Big(\prod_{j_1=i_1+1}^{t}\{I-\varphi_{j_1}(I+C^\star)\}\Big)^T\prod_{j_2=i_2+1}^{t}\{I-\varphi_{j_2}(I+C^\star)\}\varphi_{i_2}\tilde{\boldsymbol{\theta}}^{i_2}\Big]$$
$$= \sum_{i=0}^{t}\varphi_i^2\mathbb{E}\left[\Big\|\prod_{j=i+1}^{t}\{I-\varphi_j(I+C^\star)\}\tilde{\boldsymbol{\theta}}^i\Big\|^2\right],$$

where the second equality is because $\{\tilde{\boldsymbol{\theta}}_i\}_i$ are mean zero and independent. Let $I + C^\star = U\Sigma U^T$ with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ be the eigenvalue decomposition. [18] pointed out that under Assumption 3 (a) the eigenvalues are uniformly bounded, i.e. $1 - \rho^\tau \le \sigma_k \le 1$ for $1 \le k \le d$. Thus, we can further bound $\mathbb{E}\big[\|\mathcal{I}_{1,t}\|^2\big]$ by

$$\mathbb{E}\big[\|\tilde{\mathcal{I}}_{1,t}\|^2\big] \le \sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))^2\varphi_i^2\mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}^i\|^2\big]. \quad (48)$$

Recall the definition of $\tilde{\boldsymbol{\theta}}^i$ in (44). For $m = 2, 4$, We have

$$\mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}^i\|^m\big] \le \mathbb{E}\big[\|I+\tilde{C}^\star\|^m\|(B^\star)^{-1}\|^m\|\nabla f(\boldsymbol{x}^\star;\xi_i)\|^m\big] \le \frac{2^m C_{g,m}}{\mu^m}, \quad (49)$$

The inequality holds due to $\|\tilde{C}^\star\| \le 1$, strong convexity, and Assumption 3(a). Combining the two displays above, we have

$$\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\tilde{\mathcal{I}}_{1,t}\|^2\big] \le \frac{4C_{g,2}}{\mu^2}\frac{1}{M}\sum_{t=0}^{M-1}\underbrace{\frac{1}{\varphi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))^2\varphi_i^2}_{\longrightarrow 1/2(1-\rho^\tau)\quad\text{by Lemma 10}} \lesssim \frac{C_{g,2}}{\mu^2}.$$

In last inequality, we use the fact that $\lim_{t\to\infty}\sum_{t=0}^{\infty}a_t = a$ if $\lim_{t\to\infty}a_t = a$. This completes the proof. ■

A.4.1. PROOF OF LEMMA 16

Proof With the eigenvalue decomposition $I + C^\star = U\Sigma U^T$ where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$, we have

$$\tilde{\mathcal{I}}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\}\varphi_i \tilde{\boldsymbol{\theta}}^i = U \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j \Sigma\}\varphi_t U^T \tilde{\boldsymbol{\theta}}^i,$$

Let $\tilde{\mathcal{Q}}_t = U^T \tilde{\mathcal{I}}_{1,t}$. By the definition of $\Xi^\star$, we have

$$\mathbb{E}\left[\left\| \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{I}}_{1,t} \tilde{\mathcal{I}}_{1,t}^T - \Xi^\star \right\|\right] = \mathbb{E}\left[\left\| \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{1,t} \tilde{\mathcal{Q}}_{1,t}^T - \Theta \circ \Gamma \right\|\right]$$
$$\leq \sqrt{\mathbb{E}\left[\left\| \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{1,t} \tilde{\mathcal{Q}}_{1,t}^T - \Theta \circ \Gamma \right\|_F^2\right]}.$$

where we apply Hölder's inequality to the last inequality. We decompose the expectation term as follows.

$$\mathbb{E}\left[\left\| \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{1,t} \tilde{\mathcal{Q}}_{1,t}^T - \Theta \circ \Gamma \right\|_F^2\right] = \sum_{p,q=1}^{d} \mathbb{E}\left[\left(\frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{t,p} \tilde{\mathcal{Q}}_{t,q} - \Theta_{p,q}\Gamma_{p,q} \right)^2\right] =: I + II,$$

where

$$I = \sum_{p,q} \mathbb{E}\left[\left(\frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{t,p} \tilde{\mathcal{Q}}_{t,q} \right)^2\right] - \Theta_{p,q}^2 \Gamma_{p,q}^2,$$
$$II = -2 \sum_{p,q} \left(\mathbb{E}\left[\frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{t,p} \tilde{\mathcal{Q}}_{t,q} \right] - \Theta_{p,q}\Gamma_{p,q} \right) \Theta_{p,q}\Gamma_{p,q}.$$

We first look at II. Expanding $\tilde{\mathcal{Q}}_t$, we have

$$\mathbb{E} \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \tilde{\mathcal{Q}}_{t,p} \tilde{\mathcal{Q}}_{t,q} = \frac{1}{M} \sum_{t=0}^{M-1} \frac{1}{\varphi_t} \sum_{i_1=0}^{t} \sum_{i_2=0}^{t} \prod_{j_1=i_1+1}^{t} (1 - \varphi_{j_1}\sigma_p) \cdot$$
$$\prod_{j_2=i_2+1}^{t} (1 - \varphi_{j_2}\sigma_q) \varphi_{i_1} \varphi_{i_2} \mathbb{E}\left[\left(U^T \tilde{\boldsymbol{\theta}}^{i_1} \tilde{\boldsymbol{\theta}}^{i_2^T} U \right)_{p,q} \right].$$

It is easy to verify that

$$\mathbb{E}[U^T \tilde{\boldsymbol{\theta}}^{i_1} \tilde{\boldsymbol{\theta}}^{i_2^T} U] = 0 \text{ for } i_1 \neq i_2, \text{ and } \mathbb{E}[U^T \tilde{\boldsymbol{\theta}}^i \tilde{\boldsymbol{\theta}}^{iT} U] = \Gamma,$$

where the first equality is because $\{\boldsymbol{\theta}^i\}_i$ is mean-zero and independent; the second inequality is from the definition of $\tilde{\boldsymbol{\theta}}^i$ in (44). Hence, we can bound $|II|$ by

$$|II| \leq 2 \sum_{p,q} \left\{ \frac{1}{M} \sum_{t=0}^{M-1} \left| \frac{1}{\varphi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} (1 - \varphi_j\sigma_p)(1 - \varphi_j\sigma_q)\varphi_i^2 - \Theta_{p,q} \right| \right\} \Theta_{p,q}\Gamma_{p,q}^2. \tag{50}$$

We aim to use Lemma 11 to give a uniform bound on the absolute value for any p, q. Going through the proof of Lemma 11, it is easy to verify that we are able to uniformly bound the absolute value term by $O(\gamma_t)$ with $\gamma_t = 1/(t+1)^{1-\alpha}$ due to the fact that $1 - \rho^\tau \leq \sigma_k \leq 1$. Therefore, we have

$$|II| \lesssim \frac{1}{M} \sum_{t=0}^{M-1} \gamma_t \sum_{p,q} (\Gamma_{p,q})^2 \lesssim \|\Gamma\|_F^2 \gamma_M.$$

Now we deal with I. Plugging in the definition of $\tilde{\mathcal{Q}}$, we have

$$\mathbb{E}\Big[\Big(\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\tilde{\mathcal{Q}}_{t,p}\tilde{\mathcal{Q}}_{t,q}\Big)^2\Big] = \frac{1}{M^2}\sum_{t_1,t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i_1,i_1'=0}^{t_1}\sum_{i_2,i_2'=0}^{t_2}\prod_{j_1=i_1+1}^{t_1}(1-\varphi_{j_1}\sigma_p)\prod_{j_1'=i_1'+1}^{t_1}(1-\varphi_{j_1'}\sigma_q)$$
$$\prod_{j_2'=i_2'+1}^{t_2}(1-\varphi_{j_2}\sigma_p)\prod_{j_2=i_2+1}^{t_2}(1-\varphi_{j_2'}\sigma_q)\varphi_{i_1}\varphi_{i_1'}\varphi_{i_2}\varphi_{i_2'}\mathbb{E}\Big[\big(U^T\tilde{\boldsymbol{\theta}}^{i_1}\tilde{\boldsymbol{\theta}}^{i_1'^T}U\big)_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^{i_2}\tilde{\boldsymbol{\theta}}^{i_2'^T}U\big)_{p,q}\Big].$$

We note that the expectation is nonzero only when the indices i_1, i_1', i_2, i_2' are pairwise identical. We divide the summation I into the four parts I_1, I_2, I_3, I_4 corresponding to the pairwise identical configurations.

Case 1. $i_1 = i_1' = i_2 = i_2'$

Summing over all the indices under this case, we have

$$I_1 = \sum_{p,q}\frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i=0}^{t_1\wedge t_2}\prod_{j_1=i+1}^{t_1}(1-\varphi_{j_1}\sigma_p)(1-\varphi_{j_1}\sigma_q)\cdot$$
$$\prod_{j_2=i+1}^{t_2}(1-\varphi_{j_2}\sigma_p)(1-\varphi_{j_2}\sigma_q)\varphi_i^4\mathbb{E}\Big[\big(U^T\tilde{\boldsymbol{\theta}}^i\tilde{\boldsymbol{\theta}}^{i^T}U\big)_{p,q}^2\Big].$$

Since $\lim_{t\to\infty}\varphi_t = 0$, we know $1 - \varphi_t\sigma_k > 0$ for large enough t. Similar to the analysis of bounding W_t in the proof of Lemma 13, the sign of the first few terms does not affect the rate of the series. Thus, without loss of generality, we assume $1 - \varphi_t\sigma_k > 0$ for all $t \geq 1$ and $1 \leq k \leq d$ for all the remainder of this paper. Furthermore, by the uniform bound on σ_k, we have

$$|I_1| \leq \frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i=0}^{t_1\wedge t_2}\prod_{j_1=i+1}^{t_1}(1-\varphi_{j_1}(1-\rho^\tau))^2\cdot$$
$$\prod_{j_2=i+1}^{t_2}(1-\varphi_{j_2}(1-\rho^\tau))^2\varphi_i^4\mathbb{E}\Big[\sum_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^i\tilde{\boldsymbol{\theta}}^{i^T}U\big)_{p,q}^2\Big].$$

By (49) , we know

$$\mathbb{E}\Big[\sum_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^i\tilde{\boldsymbol{\theta}}^{i^T}U\big)_{p,q}^2\Big] = \mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}^i\|^4\big] \leq \frac{16C_{g,4}}{\mu^4}.$$

Due to the symmetry of indices t_1 and t_2, we have

$$|I_1| \le \frac{2}{M^2}\sum_{t_1=0}^{M-1}\frac{1}{\varphi_{t_1}}\sum_{t_2=0}^{t_1}\frac{1}{\varphi_{t_2}}\prod_{j_2=t_2+1}^{t_1}(1-\varphi_{j_2}(1-\rho^\tau))^2\underbrace{\sum_{i=0}^{t_2}\prod_{j_1=i+1}^{t_2}(1-\varphi_{j_1}(1-\rho^\tau))^4\varphi_i^4}_{=O(\varphi_{t_2}^3)\quad\text{by Lemma 10}}\frac{16C_{g,4}}{\mu^4}$$

$$\lesssim \frac{2}{M^2}\sum_{t_1=0}^{M-1}\underbrace{\frac{1}{\varphi_{t_1}}\sum_{t_2=0}^{t_1}\prod_{j_2=t_2+1}^{t_1}(1-\varphi_{j_2}(1-\rho^\tau))^2\varphi_{t_2}^2}_{\longrightarrow 1/2(1-\rho^\tau)\quad\text{by Lemma 10}}\frac{C_{g,4}}{\mu^4}\lesssim\frac{C_{g,4}}{\mu^4}\frac{1}{M}. \tag{51}$$

Case 2. $i_1=i_1', i_2=i_2', i_1\neq i_2$
By (44) and the independence among $\{\xi_i\}_i$, we know

$$\mathbb{E}\left[\left(U^\top\tilde{\boldsymbol{\theta}}^{i_1}\tilde{\boldsymbol{\theta}}^{i_1'T}U\right)_{p,q}\left(U^\top\tilde{\boldsymbol{\theta}}^{i_2}\tilde{\boldsymbol{\theta}}^{i_2'T}U\right)_{p,q}\right]=\Gamma_{p,q}^2.$$

Noting that the indices set $\{i_1=i_1', i_2=i_2', i_1\neq i_2\}=\{i_1=i_1', i_2=i_2'\}\backslash\{i_1=i_1'=i_2=i_2'\}$, we have

$$\begin{aligned}I_2=&-\sum_{p,q}\frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i=0}^{t_1\wedge t_2}\prod_{j_1=i+1}^{t_1}(1-\varphi_{j_1}\sigma_p)(1-\varphi_{j_1}\sigma_q)\prod_{j_2=i+1}^{t_2}(1-\varphi_{j_2}\sigma_p)(1-\varphi_{j_2}\sigma_q)\varphi_i^4\Gamma_{p,q}^2\\&+\sum_{p,q}\Big\{\frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i_1=0}^{t_1}\sum_{i_2=0}^{t_2}\prod_{j_1=i_1+1}^{t_1}(1-\sigma_p\varphi_{j_1})(1-\sigma_q\varphi_{j_1})\\&\qquad\prod_{j_2=i_2+1}^{t_2}(1-\sigma_p\varphi_{j_2})(1-\sigma_q\varphi_{j_2})\varphi_{i_1}^2\varphi_{i_2}^-\Theta_{p,q}^2\Big\}\Gamma_{p,q}^2=:I_{2,a}+I_{2,b}.\end{aligned}$$

Actually, $I_{2,a}$ shares the same series as (51), thus $I_{2,a}$ by $|I_{2,a}|\lesssim\frac{1}{M}\sum_{p,q}\Gamma_{p,q}^2=\frac{1}{M}\|\Gamma\|_F^2$. On the other hand, $I_{2,b}$ can be rewrite as

$$\begin{aligned}I_{2,b}=&\sum_{p,q}\left(\frac{1}{M}\sum_{t=0}^{M-1}\left|\frac{1}{\varphi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\left(1-\sigma_p\varphi_j\right)\left(1-\sigma_q\varphi_j\right)\varphi_i^2-\Theta_{p,q}\right|\right)^2\Gamma_{p,q}^2\\&+2\sum_{p,q}\left(\frac{1}{M}\sum_{t=0}^{M-1}\left|\frac{1}{\varphi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\left(1-\varphi_j\sigma_p\right)\left(1-\varphi_j\sigma_q\right)\varphi_i^2-\Theta_{pq}\right|\right)\Theta_{p,q}\Gamma_{p,q}^2\end{aligned}$$

From the analysis of (50), we get $|I_{2,b}|\lesssim\gamma_M\|\Gamma\|_F^2$. To sum up, we have

$$|I_2|\lesssim\|\Gamma\|_F^2\gamma_M.$$

Case 3. $i_1=i_2, i_1'=i_2', i_1\neq i_1'$

With the bound on the eigenvalues, we have

$$|I_3| \leq \frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\frac{1}{\varphi_{t_1}}\frac{1}{\varphi_{t_2}}\sum_{i_1=0}^{t_1\wedge t_2}\prod_{j_1=i_1+1}^{t_1}(1-\varphi_{j_1}(1-\rho^\tau))\prod_{j_2=i_1+1}^{t_2}(1-\varphi_{j_2}(1-\rho^\tau))\varphi_{i_1}^2\cdot$$
$$\sum_{i_1'=0,i_1\neq i_1'}^{t_1\wedge t_2}\prod_{j_1'=i_1'+1}^{t_1}(1-\varphi_{j_1'}(1-\rho^\tau))\prod_{j_2'=i_1'+1}^{t_2}(1-\varphi_{j_2'}(1-\rho^\tau))\varphi_{i_1'}^2\mathbb{E}\Big[\sum_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^{i_1}\big)_p^2\big(U^T\tilde{\boldsymbol{\theta}}^{i_1'}\big)_q^2\Big].$$

When $i_1 \neq i_1'$, by (44) and (49), we have

$$\mathbb{E}\Big[\sum_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^{i_1}\big)_p^2\big(U^T\tilde{\boldsymbol{\theta}}^{i_1'}\big)_q^2\Big] = \mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}^{i_1}\|^2\big]\mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}^{i_1'}\|^2\big] \overset{Hölder's}{\leq} \mathbb{E}\|\tilde{\theta}^{i_1}\|^4 \leq \frac{16C_{g,4}}{\mu^4}.$$

Due to the symmetry of indices t_1 and t_2, we have

$$|I_3| \lesssim \frac{C_{g,4}}{\mu^4}\cdot\frac{1}{M^2}\sum_{t_1=0}^{M-1}\frac{1}{\varphi_{t_1}}\sum_{t_2=0}^{t_1}\frac{1}{\varphi_{t_2}}\prod_{j=t_2+1}^{t_1}(1-\varphi_j(1-\rho^\tau))^2\Big\{\underbrace{\sum_{i_1=0}^{t_2}\prod_{j_1=i_1+1}^{t_2}(1-\varphi_{j_1}(1-\rho^\tau))^2\varphi_{i_1}^2}_{=O(\varphi_{t_2})\quad\text{by Lemma 10}}\Big\}^2$$
$$\lesssim \frac{C_{g,4}}{\mu^4}\cdot\frac{1}{M^2}\sum_{t_1=0}^{M-1}\underbrace{\frac{1}{\varphi_{t_1}}\sum_{t_2=0}^{t_1}\prod_{j=t_2+1}^{t_1}(1-(1-\rho^\tau)\varphi_j)^2\varphi_{t_2}}_{\longrightarrow 1/2(1-\rho^\tau)\quad\text{by Lemma 10}} \lesssim \frac{G_4}{\mu^4}\cdot\frac{1}{M^2}\sum_{t_1=0}^{M-1}\frac{1}{\varphi_{t_1}} \lesssim \frac{C_{g,4}}{\mu^4}\cdot\frac{1}{M^{1-\alpha}} = \frac{C_{g,4}}{\mu^4}\gamma_M.$$

Case 4. $i_1 = i_1', i_2 = i_2', i_1 \neq i_2$
I_4 is the same as (A.4.1) except that the expectation term is replaced by

$$\sum_{p,q}\mathbb{E}\Big[\big(U^T\tilde{\boldsymbol{\theta}}^{i_1}\tilde{\boldsymbol{\theta}}^{i_1^T}U\big)_{p,q}\big(U^T\tilde{\boldsymbol{\theta}}^{i_2}\tilde{\boldsymbol{\theta}}^{i_2^T}U\big)_{p,q}\Big] = \sum_{p,q}\Gamma_{p,q}^2 = \|\Gamma\|_F^2.$$

Therefore, we have $|I_4| \lesssim \|\Gamma\|_F^2\gamma_M$.

Therefore, we combine the four cases above and get $|I| \lesssim \max(\|\Gamma\|_F^2, C_{g,4}/\mu^4)\gamma_M$. Combining with the analysis on II, we complete the proof. ■

A.4.2. PROOF OF LEMMA 17

We first introduce the following auxiliary lemma with the proof in Section A.4.3.

Lemma 19 *Under Assumptions 1, 2 (b) with $m = 2$, 3 (b) with $m = 2$, and 6 with $m = 2$, we have*

$$\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}^i\|^2\big] \lesssim \frac{1}{\gamma_B^2}(\Upsilon_{Lg}^2 + \Upsilon_{m,2})\mathbb{E}\big[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2\big] + \frac{C_{g,2}}{\mu^2}\big(\frac{1}{\gamma_B^2} + \frac{\tau^2\Upsilon_{S,2}}{\mu^2}\big)\mathbb{E}\big[\|B_i - B^\star\|^2\big].$$

Proof Recall the definition of $\hat{\mathcal{I}}_{1,t}$

$$\hat{\mathcal{I}}_{1,t} = \sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I - \varphi_j(I + C^\star)\}\varphi_i\hat{\boldsymbol{\theta}}^i = U\sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I - \varphi_j\Sigma\}\varphi_t U^T\hat{\boldsymbol{\theta}}^i,$$

where $I + C^\star = U\Sigma U^T$ is the eigenvalue decomposition and $\lambda(\Sigma) \geq 1 - \rho^\tau$. Since $\hat{\boldsymbol{\theta}}^i$ is a martingale difference sequence, we follow the same analysis as (48) and get

$$\mathbb{E}\big[\|\hat{\mathcal{I}}_{1,t}\|^2\big] \leq \sum_{i=0}^{t} \prod_{j=i+1}^{t} (1 - \varphi_j(1-\rho^\tau))^2 \varphi_i^2 \mathbb{E}\big[\|\hat{\boldsymbol{\theta}}^i\|^2\big].$$

Combining Lemma 19, Lemma 13, and Lemma 14, we know

$$\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}^i\|^2\big] \lesssim C_{\hat{\boldsymbol{\theta}}}(\beta_i + \omega_i^2),$$

where $C_{\hat{\boldsymbol{\theta}}}$ is defined in (46). Therefore, we have

$$\frac{1}{M}\sum_{t=0}^{M-1} \frac{1}{\varphi_t}\mathbb{E}\|\hat{\mathcal{I}}_{1,t}\|^2 \lesssim C_{\hat{\boldsymbol{\theta}}} \frac{1}{M}\sum_{t=0}^{M-1} \frac{1}{\varphi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} (1 - \varphi_j(1-\rho^\tau))^2 \varphi_i \cdot \varphi_i(\beta_i + \omega_i^2).$$

Since $\lim_{t\to\infty} t(1 - \varphi_{t-1}\beta_{t-1}/\varphi_t\beta_t) = -2\beta)$ and $\lim_{t\to\infty} t(1 - \varphi_{t-1}\omega_{t-1}/\varphi_t\omega_t) = -\beta - 2\omega$, we apply Lemma 10 and get

$$\frac{1}{M}\sum_{t=0}^{M-1} \frac{1}{\varphi_t}\mathbb{E}\|\hat{\mathcal{I}}_{1,t}\|^2 \lesssim C_{\hat{\boldsymbol{\theta}}} \frac{1}{M}\sum_{t=0}^{M-1} \big(\beta_t + \omega_t^2\big) \lesssim C_{\hat{\boldsymbol{\theta}}}(\beta_M + \omega_M^2) \lesssim C_{\hat{\boldsymbol{\theta}}}\beta_M.$$

The last two inequalities are due to $\beta < 1$ and $\beta < 2\omega$. This completes the proof. ■

A.4.3. PROOF OF LEMMA 19

Proof Recall the definition of $\hat{\boldsymbol{\theta}}^i$ in (44), we have

$$\hat{\boldsymbol{\theta}}^i = \boldsymbol{\theta}^i - \tilde{\boldsymbol{\theta}}^i = -(I + \tilde{C}_i)B_i^{-1}\nabla f(\boldsymbol{x}_i;\xi_i) + (I + \tilde{C}_i^\star)(B^\star)^{-1}\nabla f(\boldsymbol{x}^\star;\xi_i) + (I + C_i)B_i^{-1}\nabla f_i.$$

Thus, we have

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}^i\|^2 \lesssim\ & \|(I + C_i)B_i^{-1}\|^2\|\nabla f_i\|^2 + \|(I + \tilde{C}_i)B_i^{-1}\|^2\|\nabla f(\boldsymbol{x}_i;\xi_i) - \nabla f(\boldsymbol{x}^\star;\xi_i)\|^2 \\ & + \|(I + \tilde{C}_i)B_i^{-1} - (1 + \tilde{C}_i^\star)(B^\star)^{-1}\|^2\|\nabla f(\boldsymbol{x}^\star;\xi_i)\|^2 =: I + II + III.\end{aligned}$$

By the Lipschitz continuity of ∇f, $\lambda(B_i) \geq \gamma_B$, and $\|C_i\| \leq 1$, we get

$$\mathbb{E}I \leq 2^2 \frac{\Upsilon_{Lg}^2}{\gamma_B^2}\mathbb{E}\big[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2\big]. \tag{52}$$

By Assumption 6 with $m = 2$ and the bound on $\|B_i^{-1}\|$, we use the law of total expectation and have

$$\mathbb{E}II \leq \frac{2^2}{\gamma_B^2}\mathbb{E}\Big[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2\mathbb{E}\left[H(\xi_i)^2 \mid \mathcal{F}_{i-1}\right]\Big] \lesssim \frac{\Upsilon_{m,2}}{\gamma_B^2}\mathbb{E}\big[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2\big]. \tag{53}$$

Now we deal with $\mathbb{E}III$. According to the strong convexity and lower bound on $\lambda(B_i)$, we have

$$\|(I+\tilde{C}_i)B_i^{-1}-(1+\tilde{C}_i^\star)(B^\star)^{-1}\|^2 \le \|I+\tilde{C}_i\|^2\|B_i^{-1}\|^2\|B^{\star-1}\|^2\|B_i-B^\star\|^2 \\ +\|B^{\star-1}\|^2\|\tilde{C}_i-\tilde{C}_i^\star\|^2 \le \frac{4}{\gamma_B^2\mu^2}\|B_i-B^\star\|^2+\frac{1}{\mu^2}\|\tilde{C}_i-\tilde{C}_i^\star\|^2.$$

We compute $\mathbb{E}III$ by first conditioning on $\mathcal{F}_{i-1}$.

$$\mathbb{E}III \le \frac{4}{\gamma_B^2\mu^2}\mathbb{E}\Big[\|B_i-B^\star\|^2\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_i)\|^2 \mid \mathcal{F}_{i-1}\big]\Big]+\frac{1}{\mu^2}\mathbb{E}\Big[\mathbb{E}\big[\|\tilde{C}_i-\tilde{C}_i^\star\|^2 \mid \mathcal{F}_{i-1}\big]\mathbb{E}\big[\|\nabla f(\boldsymbol{x}^\star;\xi_i)\|^2 \mid \mathcal{F}_{i-1}\big]\Big] \\ \le \frac{4C_{g,2}}{\gamma_B^2\mu^2}\mathbb{E}\big[\|B_i-B^\star\|^2\big]+\frac{C_{g,2}}{\mu^2}\mathbb{E}\big[\|\tilde{C}_i-\tilde{C}_i^\star\|^2\big], \quad (54)$$

where the first inequality also uses the independence between ξ_i and ζ_i; the second inequality comes from the Assumption 6. By the definition of $\tilde{C}_i^\star$ in (43), we have

$$\|\tilde{C}_i-\tilde{C}_i^\star\| = \Big\|\prod_{j=0}^{\tau-1}\tilde{C}_{i,j}-\prod_{j=0}^{\tau-1}\tilde{C}_{i,j}^\star\Big\| \le \Big\|\prod_{j=0}^{\tau-2}\tilde{C}_{i,j}-\prod_{j=0}^{\tau-2}\tilde{C}_{i,j}^\star\Big\|\|C_{i,\tau-1}^\star\| \\ +\Big\|\prod_{j=0}^{\tau-2}\tilde{C}_{i,j}\Big\|\|\tilde{C}_{i,\tau-1}-\tilde{C}_{i,\tau-1}^\star\| \le \cdots \le \sum_{j=0}^{\tau-1}\Big\|\tilde{C}_{i,j}-\tilde{C}_{i,j}^\star\Big\|.$$

By [18, Lemma 4.2] and Assumption 3(b), we get

$$\mathbb{E}\big[\|\tilde{C}_i-\tilde{C}_i^\star\|^2 \mid \mathcal{F}_{i-1}\big] \le \frac{4\|B_i-B^\star\|^2}{\mu^2}\mathbb{E}\left[\left(\sum_{j=0}^{\tau-1}\frac{\|S_{i,j}\|}{\sigma_{\min}^+(S_{i,j})}\right)^2\right] \lesssim \frac{\tau^2\Upsilon_{S,2}}{\mu^2}\|B_i-B^\star\|^2.$$

Going back to (54), we get

$$\mathbb{E}III \lesssim \frac{C_{g,2}}{\mu^2}\Big(\frac{1}{\gamma_B^2}+\frac{\tau^2\Upsilon_{S,2}}{\mu^2}\Big)\mathbb{E}\big[\|B_t-B^\star\|^2\big]. \quad (55)$$

Combining (52), (53), and (55) completes the proof. ■

A.5. Proof of Theorem 8

In order to apply the decomposition in Lemma 15, we rewrite the sample covariance matrix Ξ_M as

$$\Xi_M = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T+\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T \\ +\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T+\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T. \quad (56)$$

Intuitively, $\bar{\boldsymbol{x}}$ is a good estimation of $\Xi^\star$. Thus, $\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T$ should be negligible while $\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T$ approximates the true covariance matrix $\Xi^\star$.

Lemma 20 *Under the conditions of Lemma 13 and 14, we suppose Assumption 3(b) with $m = 1$ holds. Let $\omega_t = c_\omega/(t+1)^\omega$ for constants $c_\omega > 0$ and $\omega > \beta/2$. Then we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T-\Xi^\star\right\|\right] \lesssim \max\left(\|\Gamma\|_F, \frac{\sqrt{C_{g,4}}}{\mu^2}, C_{\hat{\boldsymbol{\theta}}}, C_{\boldsymbol{\delta}}\right)\sqrt{\gamma_M},$$

where

$$C_{\boldsymbol{\delta}} := \max\left\{\left(\frac{\Upsilon_{S,1}^2\tau^2}{\mu^2}+\frac{\Upsilon_{Lg}^2}{\mu^2\gamma_B^2}\right)\sqrt{D}\cdot\max\left(\sqrt{C_{H,4}}, \sqrt{D}\Upsilon_{LH}^2\right), \|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2\right\}. \tag{57}$$

Lemma 21 *Under the conditions of Lemma 13, Lemma 14, we further suppose $\beta > 1/2$. Let $\omega_t = c_\omega/(t+1)^\omega$ with $\omega > \beta/2$. We have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T\right\|\right] \le \frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star\|^2\right] \lesssim \max\left(\frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}, C_{\boldsymbol{\delta}}\right)\gamma_M.$$

The proofs of these two lemmas are left to Section A.5.1 and A.5.2. Now we are prepared to prove Theorem 8.

Proof By the decomposition (56), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\Xi_M-\Xi^\star\right\|\right] &\le \mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T-\Xi^\star\right\|\right]\\ &+\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T\right\|\right]+2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star)^T\right\|\right]. \end{aligned}\tag{58}$$

The first two terms are analyzed in Lemma 20 and 21. Similar to (47), we apply Cauchy's inequality and have

$$\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|\|\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star\|\right] \le \sqrt{\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\right]}\sqrt{\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\bar{\boldsymbol{x}}_M-\boldsymbol{x}^\star\|^2\right]}. \tag{59}$$

In order to complete the proof, it is sufficient to bound $\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\right]$. By Lemma 13, we have

$$\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\right] \lesssim \sqrt{D}\frac{1}{M}\sum_{t=1}^{M}\frac{\beta_t}{\varphi_t} \lesssim \sqrt{D}. \quad \left(\text{since } \lim_{t\to\infty}\beta_t/\varphi_t = 1 \text{ by } \chi_t = o(\beta_t)\right). \tag{60}$$

Combining Lemma 20, Lemma 21 and the derivations above, we conclude that

$$\mathbb{E}\left[\|\Xi_M-\Xi^\star\|\right] \lesssim \max\left(\|\Gamma\|_F, \frac{\sqrt{C_{g,4}}}{\mu^2}, C_{\hat{\boldsymbol{\theta}}}, C_{\boldsymbol{\delta}}\right)\sqrt{\gamma_M}. \tag{61}$$

This completes the proof. ■

A.5.1. PROOF OF LEMMA 20

Recall the decomposition (42), we know the performance of $\frac{1}{M}\sum_{t=1}^{M}\frac{1}{\varphi_t}(\boldsymbol{x}_t-\boldsymbol{x}^\star)(\boldsymbol{x}_t-\boldsymbol{x}^\star)^T$ relies on $\mathcal{I}_{1,t}$, $\mathcal{I}2,t$, and $\mathcal{I}_{3,t}$, The linear case is analyzed in Section A.4. Here we introduce two lemmas regarding $\mathcal{I}_{2,t}$ and $\mathcal{I}_{3,t}$, the proofs of which are given in sections A.5.3 and A.5.4.

Lemma 22 *Under the conditions of Lemma 13, we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{2,t}\mathcal{I}_{2,t}^T\right\|\right]\leq\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\mathcal{I}_{2,t}\|^2\right]\lesssim\frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\beta_M.$$

Lemma 23 *Under the conditions of Lemma 13 and 14, we suppose Assumption 3(b) with $m=1$ holds. Let $\omega_t=c_\omega/(t+1)^\omega$ for constants $c_\omega>0$ and $\omega>\beta/2$. Then we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{3,t}\mathcal{I}_{3,t}^T\right\|\right]\leq\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\mathcal{I}_{3,t}\|^2\right]\lesssim C_{\boldsymbol{\delta}}\beta_M,$$

with $C_{\boldsymbol{\delta}}$ given in (57).

Proof According to the decomposition Lemma 15, we have

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\frac{1}{\varphi_i}(\boldsymbol{x}_i-\boldsymbol{x}^\star)(\boldsymbol{x}_i-\boldsymbol{x}^\star)^T-\Xi^\star\right\|\right]\leq\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{1,t}\mathcal{I}_{1,t}^T-\Xi^\star\right\|\right]+\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{2,t}\mathcal{I}_{2,t}^T\right\|\right]$$
$$+\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{3,t}\mathcal{I}_{3,t}^T\right\|\right]+2\sum_{1\leq i<j\leq 3}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{i,t}\mathcal{I}_{j,t}^T\right\|\right].$$

The rate of the first three terms are shown in Lemma 18, 22, and 23, respectively. For $1\leq i<j\leq 3$, we use Cauchy's inequality twice and have

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{i,t}\mathcal{I}_{j,t}^T\right\|\right]\leq\mathbb{E}\left[\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\|\mathcal{I}_{i,t}\|\|\mathcal{I}_{j,t}\|\right]$$
$$\leq\mathbb{E}\left[\sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\|\mathcal{I}_{i,t}\|^2}\sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\|\mathcal{I}_{j,t}\|^2}\right]\leq\sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\mathcal{I}_{i,t}\|^2\right]}\sqrt{\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\mathcal{I}_{j,t}\|^2\right]}.$$

Thus, it suffices to bound $\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\left[\|\mathcal{I}_{1,t}\|^2\right]$. Recall the expression of $\mathcal{I}_{1,t}$ (35). Since $\boldsymbol{\theta}^i$ is a martingale difference sequence, we can use techniques similar to (48) and get

$$\mathbb{E}\left[\|\mathcal{I}_{1,t}\|^2\right]\leq\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))^2\varphi_i^2\mathbb{E}\left[\|\boldsymbol{\theta}^i\|^2\right].$$

By the definition of $\boldsymbol{\theta}^i$ in (39), we have

$$\mathbb{E}\left[\|\boldsymbol{\theta}^i\|^2\right]\lesssim\frac{1}{\gamma_B^2}\left(\mathbb{E}\left[\|\bar{g}_i\|^2\right]+\mathbb{E}\left[\|\nabla f_i\|^2\right]\right)\lesssim\frac{1}{\gamma_B^2}\left(C_{g,2}+\Upsilon_{Lg}^2\left[\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2\right]\right)\lesssim\frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}(1+\beta_i).\tag{62}$$

The inequality holds due to $\|\tilde{C}^\star\| \leq 1$, strong convexity, Assumption 2(b), and Lemma 13. Therefore, we have

$$\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\mathcal{I}_{1,t}\|^2\big] \lesssim \frac{1}{\gamma_B^2}\max\big(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2\big)\frac{1}{M}\sum_{t=0}^{M-1}\underbrace{\frac{1}{\varphi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))^2\varphi_i^2(1+\beta_i)}_{\longrightarrow 1/2(1-\rho^\tau) \quad \text{by Lemma 13}}$$
$$\lesssim \frac{1}{\gamma_B^2}\max\big(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2\big). \quad (63)$$

Combining Lemma 18, 22, 23, and (63) completes the proof. ■

A.5.2. PROOF OF LEMMA 21

Proof By the decomposition of $\boldsymbol{x}_t - \boldsymbol{x}^\star$, we have

$$\bar{\boldsymbol{x}}_M - \boldsymbol{x}^\star = \frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{1,t} + \frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{2,t} + \frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathcal{I}_{3,t} =: \bar{\mathcal{I}}_{1,M} + \bar{\mathcal{I}}_{2,M} + \bar{\mathcal{I}}_{3,M}.$$

Expanding $\bar{\mathcal{I}}_{1,M}$ with the $\mathcal{I}_{1,t}$ expression (35) and exchanging the indices, we have

$$\bar{\mathcal{I}}_{1,M} = \frac{1}{M}\sum_{t=0}^{M-1}\sum_{i=0}^{t}\prod_{j=i+1}^{t}U\{I-\varphi_j\Sigma\}\varphi_i(U^T\boldsymbol{\theta}^i) = \frac{1}{M}\sum_{i=0}^{M-1}\sum_{t=i}^{M-1}\prod_{j=i+1}^{t}U\{I-\varphi_j\Sigma\}\varphi_i(U^T\boldsymbol{\theta}^i). \quad (64)$$

$\boldsymbol{\theta}^i$ is a martingale difference sequence, which helps us cancel the interaction terms in $\mathbb{E}\big[\|\bar{\mathcal{I}}_{1,M}\|^2\big]$ and returns

$$\mathbb{E}\big[\|\bar{\mathcal{I}}_{1,M}\|^2\big] = \frac{1}{M^2}\sum_{i=0}^{M-1}\varphi_i^2\mathbb{E}\bigg[\Big\|\sum_{t=i}^{M-1}\prod_{j=i+1}^{t}\{I-\varphi_j\Sigma\}(U^T\boldsymbol{\theta}^i)\Big\|^2\bigg]$$
$$\leq \frac{1}{M^2}\sum_{i=0}^{M-1}\Big(\sum_{t=i}^{M-1}\prod_{j=i+1}^{t}\big(1-\varphi_j(1-\rho^\tau)\big)\Big)^2\varphi_i^2\mathbb{E}\big[\|\boldsymbol{\theta}^i\|^2\big] =: (\#).$$

Again, we exchange the indices and obtain

$$(\#) = \frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{t_1=i}^{M-1}\sum_{t_2=i}^{M-1}\prod_{j_1=i_1+1}^{t}(1-\varphi_{j_1}(1-\rho^\tau))\prod_{j_2=i_2+1}^{t}(1-\varphi_{j_2}(1-\rho^\tau))\varphi_i^2\mathbb{E}\big[\|\boldsymbol{\theta}^i\|^2\big]$$
$$= \frac{1}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{M-1}\sum_{i=0}^{t_1\wedge t_2}\prod_{j_1=i+1}^{t_1}(1-\varphi_{j_1}(1-\rho^\tau))\prod_{j_2=i+1}^{t_2}(1-\varphi_{j_2}(1-\rho^\tau))\varphi_i^2\mathbb{E}\big[\|\boldsymbol{\theta}^i\|^2\big]$$
$$\leq \frac{2}{M^2}\sum_{t_1=0}^{M-1}\sum_{t_2=0}^{t_1}\prod_{j_1=t_2+1}^{t_1}(1-\varphi_{j_1}(1-\rho^\tau))\sum_{i=0}^{t_2}\prod_{j_2=i+1}^{t_2}(1-\varphi_{j_2}(1-\rho^\tau))^2\varphi_i^2\mathbb{E}\big[\|\boldsymbol{\theta}^i\|^2\big],$$

where the inequality comes from the symmetry of indices i_1 and i_2. By (62), we know

$$(\#) \lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2} \frac{1}{M^2} \sum_{t_1=0}^{M-1} \sum_{t_2=0}^{t_1} \prod_{j_1=t_2+1}^{t_1} (1-\varphi_{j_1}(1-\rho^\tau)) \underbrace{\sum_{i=0}^{t_2} \prod_{j_2=i+1}^{t_2} (1-\varphi_{j_2}(1-\rho^\tau))^2 \varphi_i^2 (1+\beta_i)}_{=O(\varphi_{t_2}) \quad \text{by Lemma 13}}$$

$$\lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2} \frac{1}{M^2} \sum_{t_1=0}^{M-1} \underbrace{\sum_{t_2=0}^{t_1} \prod_{j=t_2+1}^{t_1} (1-(1-\rho^\tau)\varphi_j)\varphi_{t_2}}_{\longrightarrow \frac{1}{1-\rho^\tau} \quad \text{by Lemma 13}} \lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2} \frac{1}{M}. \tag{65}$$

Secondly, we consider $\bar{\mathcal{I}}_{2,M}$. Similar to (64), we have

$$\bar{\mathcal{I}}_{2,M} = \frac{1}{M} \sum_{t=0}^{M-1} \sum_{i=0}^{t} \prod_{j=i+1}^{t} U\{I-\varphi_j\Sigma\}(\bar{\alpha}_i - \varphi_i) U^T \bar{\Delta}\boldsymbol{x}_i$$

Furthermore, we use Cauchy's inequality and get

$$\mathbb{E}\|\bar{\mathcal{I}}_{2,M}\|^2 \le \frac{1}{4}\mathbb{E}\Big[\Big(\frac{1}{M}\sum_{t=0}^{M-1}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\chi_i\|\bar{\Delta}\boldsymbol{x}_i\|\Big)^2\Big]$$
$$\le \frac{1}{4}\Big(\frac{1}{M}\sum_{t=0}^{M-1}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\chi_i\sqrt{\|\bar{\Delta}\boldsymbol{x}_i\|^2}\Big)^2. \tag{66}$$

By (23), Υ_{Lg}-Lipschitz continuity of ∇f, and Lemma 13, we bound $\mathbb{E}\|\bar{\Delta}\boldsymbol{x}_i\|^2$ as follows.

$$\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_i\|^2\big] \le \frac{2^2}{\gamma_B^2}\big(C_{g,2} + \Upsilon_{Lg}^2\mathbb{E}\big[\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2\big]\big) \lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}(1+\beta_i). \tag{67}$$

Therefore, we have

$$\mathbb{E}\|\bar{\mathcal{I}}_{2,M}\|^2 \lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\Big(\frac{1}{M}\sum_{t=0}^{M-1}\underbrace{\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-(1-\rho^\tau)\varphi_j)\chi_i\sqrt{1+\beta_i}}_{=O(\chi_t/\beta_t)}\Big)^2$$
$$\lesssim \frac{\max(C_{g,2}, \sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\frac{\chi_M}{\beta_M} \quad (\text{since} \chi > 2\beta). \tag{68}$$

Finally, we have

$$\bar{\mathcal{I}}_{3,M} = \frac{1}{M}\sum_{t=0}^{M-1}\prod_{i=0}^{t} U\{I-\varphi_i\Sigma\}U^T(\boldsymbol{x}_0 - \boldsymbol{x}^\star) + \frac{1}{M}\sum_{t=0}^{M-1}\sum_{i=0}^{t}\sum_{j=1+1}^{t} U\{I-\varphi_j\Sigma\}\varphi_i U^T\boldsymbol{\delta}^i,$$

thus

$$\mathbb{E}\big[\|\mathcal{I}_{3,M}\|^2\big] \lesssim \Big(\frac{1}{M}\sum_{t=0}^{M-1}\prod_{i=0}^{t}\big(1-\varphi_i(1-\rho^\tau)\big)\Big)^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2+\Big(\frac{1}{M}\sum_{t=0}^{M-1}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\big(1-\varphi_j(1-\rho^\tau)\big)\varphi_i\sqrt{\mathbb{E}\big[\|\boldsymbol{\delta}^i\|^2\big]}\Big)^2. \tag{69}$$

Now we compute the rate of $\mathbb{E}\big[\|\boldsymbol{\delta}^i\|^2\big]$. By the definition of $\boldsymbol{\delta}^i$ in (40), we know

$$\begin{aligned}\|\boldsymbol{\delta}^i\|^2 &\le \|C_i - C^\star\|^2\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2 + 2\|(B^\star)^{-1}\|^2\|\boldsymbol{\psi}^i\|^2 + 2\|B_i^{-1}\|^2\|B^{\star-1}\|^2\|B_t-B^\star\|^2\|\nabla f_i\|^2\\ &\lesssim \frac{\tau^2\Upsilon_{S,1}^2}{\mu^2}\|B_i-B^\star\|^2\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2 + \frac{\Upsilon_{Lh}^2}{\mu^2}\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^4 + \frac{\Upsilon_{Lg}^2}{\mu^2\gamma_B^2}\|B_i-B^\star\|^2\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2,\end{aligned}$$

where the second inequality holds due to the following reasons: 1) [18, Lemma 4.2] introduces a bound on $\|C_t - C^\star\|$; 2)$\lambda_{\min}(B_t) \ge \gamma_B$; 3) f is μ strongly convex; 4) ∇f is Υ_{Lg} Lipschitz continuous; 5) $\|\boldsymbol{\psi}^i\| \le \frac{1}{2}\Upsilon_{Lh}\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2$ due to the Υ_{Lh} Lipschitz continuity of $\nabla^2 f$. Using Cauchy's inequality and plugging in the rate obtained in Lemma 13 and 14, we have

$$\mathbb{E}\big[\|\boldsymbol{\delta}^i\|^2\big] \lesssim \Big(\frac{\tau^2\Upsilon_{S,1}^2}{\mu^2}+\frac{\Upsilon_{Lg}^2}{\mu^2\gamma_B^2}\Big)\sqrt{\mathbb{E}\|B_i-B^\star\|^4}\sqrt{\mathbb{E}\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^4}+\frac{\Upsilon_{Lh}^2}{\mu^2}\mathbb{E}\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^4 \lesssim C_{\boldsymbol{\delta}}\beta_i(\beta_i+\omega_i^2), \tag{70}$$

where $C_{\boldsymbol{\delta}}$ is defines in (57). We apply Lemma 10 and have

$$\begin{aligned}\mathbb{E}\|\bar{\mathcal{I}}_{3,M}\|^2 &\lesssim C_\delta\Big(\frac{1}{M}\sum_{t=0}^{M-1}\underbrace{\prod_{i=0}^{t}\big(1-\varphi_i(1-\rho^\tau)\big)}_{=o\left(\sqrt{\beta_t(\beta_t+\omega_t^2)}\right)}\Big)^2+C_{\boldsymbol{\delta}}\Big(\frac{1}{M}\sum_{t=0}^{M-1}\underbrace{\sum_{i=0}^{t}\prod_{j=i+1}^{t}\big(1-\varphi_j(1-\rho^\tau)\big)\varphi_i\sqrt{\beta_i(\beta_i+\omega_i^2)}}_{=O\left(\sqrt{\beta_t(\beta_t+\omega_t^2)}\right)}\Big)^2\\ &\lesssim C_{\boldsymbol{\delta}}\beta_M(\beta_M+\omega_M^2) \lesssim C_{\boldsymbol{\delta}}\beta_M^2.\end{aligned} \tag{71}$$

The last two inequalities are due to $\beta < 1$ and $2\omega > \beta$. Now we combine. (65),(68), and (71). With the fact $\frac{1}{M}\sum_{t=0}^{M-1} 1/\varphi_t \lesssim \frac{1}{M}\sum_{t=1}^{M} t^\beta \lesssim M^\beta$, we know

$$\begin{aligned}\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\bar{\boldsymbol{x}}-\boldsymbol{x}^\star\|^2\big] &\lesssim \max\Big(\frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}, C_{\boldsymbol{\delta}}\Big)\frac{1}{\beta_M}\Big(\frac{1}{M}+\frac{\chi_M^2}{\beta_M^2}+\beta_M^2\Big)\\ &\lesssim \max\Big(\frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}, C_{\boldsymbol{\delta}}\Big)\gamma_M,\end{aligned}$$

where the last inequality follows from $1/2 < \beta < 1$ and $\chi > 2\beta$. We complete the proof. ■

A.5.3. PROOF OF LEMMA 22

Proof Recall the definition of $\mathcal{I}_{2,t}$ in (36), we have

$$\|\mathcal{I}_{2,t}\| \le \Big\|\sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j\Sigma\}(\alpha_i-\varphi_i)U^T\bar{\Delta}\boldsymbol{x}_i\Big\| \le \sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\frac{\chi_i}{2}\|\bar{\Delta}\boldsymbol{x}_i\|$$

where the last inequality is due to the uniform bound on σ_k and $|\alpha_i - \varphi_i| \leq \chi_i/2$. Furthermore, we apply Cauchy's inequality and obtain

$$\mathbb{E}\|\mathcal{I}_{2,t}\|^2 \leq \mathbb{E}\Big[\Big(\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\frac{\chi_i}{2}\|\bar{\Delta}\boldsymbol{x}_i\|\Big)^2\Big] \leq \Big(\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\frac{\chi_i}{2}\sqrt{\mathbb{E}\|\bar{\Delta}\boldsymbol{x}_i\|^2}\Big)^2.$$

We plug in the rate $\mathbb{E}\big[\|\bar{\Delta}\boldsymbol{x}_i\|^2\big]$ in (67). Moreover, since $\lim_{t\to\infty} t\Big(1-\frac{\chi_{t-1}(1+\beta_{t-1}^{1/2})/\varphi_{t-1}}{\chi_t(1+\beta_t^{1/2})/\varphi_t}\Big) = \beta-\chi$, we can apply Lemma 10 and obtain

$$\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\hat{\mathcal{I}}_{2,t}\|^2\big] \lesssim \frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\Big(\underbrace{\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j(1-\rho^\tau))\chi_i(1+\beta_i^{1/2})}_{=O(\chi_t/\varphi_t)\quad\text{by Lemma 10}}\Big)^2$$
$$\lesssim \frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\frac{1}{M}\sum_{t=0}^{M-1}\frac{\chi_t^2}{\varphi_t^3} \lesssim \frac{\max(C_{g,2},\sqrt{D}\Upsilon_{Lg}^2)}{\gamma_B^2}\beta_M.$$

The last inequality holds because $\chi_t^2/\varphi_t^3 \lesssim \beta_t$ due to the condition $\chi > 2\beta$ in the assumption of Lemma 13. This completes the proof. ■

A.5.4. PROOF OF LEMMA 23

Proof By the expression of $\mathcal{I}_{3,t}$ in (37), we have

$$\|\mathcal{I}_{3,t}\|^2 \lesssim \Big\|\prod_{i=0}^{t}\{I-\varphi_i\Sigma\}U^T(\boldsymbol{x}_0-\boldsymbol{x}^\star)\Big\|^2 + \Big\|\sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j\Sigma\}\varphi_i U^T\boldsymbol{\delta}^i\Big\|^2$$
$$\leq \prod_{i=0}^{t}(1-\varphi_i(1-\rho^\tau))^2\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2 + \Big(\sum_{i=0}^{t}(1-\varphi_j(1-\rho^\tau))\varphi_i\|\boldsymbol{\delta}^i\|\Big)^2.$$

By Cauchy's inequality, we have

$$\mathbb{E}\Big[\Big(\sum_{i=0}^{t}(1-\varphi_j(1-\rho^\tau))\varphi_i\|\boldsymbol{\delta}^i\|\Big)^2\Big] \leq \Big(\sum_{i=0}^{t}(1-\varphi_j(1-\rho^\tau))\varphi_i\sqrt{\mathbb{E}\big[\|\boldsymbol{\delta}^i\|^2\big]}\Big)^2.$$

Plugging in the rate of $\mathbb{E}\big[\|\boldsymbol{\delta}^i\|^2\big]$ in (70), we come to

$$\frac{1}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\mathbb{E}\big[\|\mathcal{I}_{3,t}\|^2\big] \lesssim \frac{C_\delta}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\prod_{i=0}^{t}(1-\varphi_i(1-\rho^\tau))^2$$
$$+\frac{C_\delta}{M}\sum_{t=0}^{M-1}\frac{1}{\varphi_t}\Big(\underbrace{\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-(1-\rho^\tau)\varphi_j)\varphi_i\sqrt{\beta_i}(\sqrt{\beta_i}+\omega_i)}_{=O(\sqrt{\beta_t}(\sqrt{\beta_t}+\omega_t))\quad\text{by Lemma 10}}\Big)^2 \lesssim C_\delta\beta_M.$$

The last two inequalities are due to $\beta < 1$ and $2\omega > \beta$. This completes the proof. ■

Appendix B. More Numerical Experiments

Table 2 lists the average length of the 95% confidence intervals for $\sum_{i=1}^{d} \boldsymbol{x}_i^\star/d$ along with the average rate shown in Table 1.

Dim	τ	Estimator	Id Σ_a	Toeplitz $\Sigma_a(r)$	Equi-corr $\Sigma_a(r)$
				0.5	0.2
5	∞	plug-in	7.00(0.05)	4.76(0.04)	5.20(0.04)
		sample cov	6.89(0.45)	4.74(0.30)	5.16(0.32)
	10	plug-in	6.98(0.05)	4.76(0.04)	5.20(0.04)
		sample cov	6.86(0.47)	5.16(0.34)	5.56(0.37)
	20	plug-in	6.98(0.05)	4.76(0.03)	5.20(0.04)
		sample cov	6.91(0.48)	4.95(0.33)	5.25(0.36)
	40	plug-in	6.97(0.06)	4.76(0.04)	5.20(0.04)
		sample cov	6.95(0.47)	4.83(0.34)	5.25(0.36)
20	∞	plug-in	3.54(0.03)	2.14(0.02)	1.61(0.01)
		sample cov	3.50(0.22)	2.11(0.14)	1.60(0.10)
	10	plug-in	3.53(0.03)	2.13(0.02)	1.61(0.01)
		sample cov	3.42(0.38)	2.61(0.18)	2.36(0.15)
	20	plug-in	3.53(0.03)	2.13(0.02)	1.61(0.01)
		sample cov	3.49(0.29)	2.52(0.18)	2.19(0.16)
	40	plug-in	3.53(0.03)	2.13(0.02)	1.61(0.01)
		sample cov	3.47(0.26)	2.38(0.15)	1.99(0.15)
40	∞	plug-in	2.54(0.02)	1.51(0.01)	0.86(0.01)
		sample cov	2.52(0.15)	1.49(0.10)	0.85(0.06)
	10	plug-in	2.54(0.02)	1.49(0.01)	0.84(0.01)
		sample cov	2.37(0.31)	1.88(0.14)	1.44(0.10)
	20	plug-in	2.54(0.02)	1.49(0.01)	0.84(0.01)
		sample cov	2.45(0.25)	1.83(0.13)	1.39(0.10)
	40	plug-in	2.54(0.05)	1.49(0.01)	0.85(0.01)
		sample cov	2.51(0.21)	1.78(0.13)	1.31(0.08)
60	∞	plug-in	2.12(0.02)	1.24(0.01)	0.59(0.00)
		sample cov	2.10(0.15)	1.23(0.07)	0.58(0.04)
	10	plug-in	2.12(0.02)	1.86(1.54)	0.57(0.00)
		sample cov	2.00(0.28)	2.40(1.93)	1.05(0.07)
	20	plug-in	2.12(0.02)	1.22(0.05)	0.58(0.00)
		sample cov	2.03(0.24)	1.55(0.13)	1.03(0.06)
	40	plug-in	2.12(0.02)	1.22(0.01)	0.58(0.00)
		sample cov	2.09(0.19)	1.48(0.10)	0.98(0.06)

Table 2: The average length ($\times 10^{-2}$) of the confidence interval for $\sum_{i=1}^{d} \boldsymbol{x}_i^\star/d$ for linear regression at the target level 95%.