A FREE LUNCH: GRADIENT AVERAGING ACHIEVES OPTIMAL STATISTICAL EFFICIENCY IN NEWTON SKETCH

Anonymous authors
Paper under double-blind review

ABSTRACT

We study statistical inference for online sketched Newton methods, in which each Newton system is approximately solved by an inner randomized sketching solver. Existing work reveals a computational–statistical trade-off: the sketching solver (and its accelerated variants) reduces the per-step cost of Newton methods from $O(d^3)$ to $O(d^2)$, but the randomness introduced by sketching also inflates the limiting covariance of both the last and the averaged Newton iterates. In this work, we address this trade-off via gradient averaging. We show that simply replacing the stochastic gradient with a moving average at each step can mitigate, and in some cases eliminate, the trade-off in Newton sketch. Importantly, gradient averaging requires no additional oracle evaluations and leaves the contraction bound of the inner sketching solver unchanged. In particular, we prove that the last sketched Newton iterate with gradient averaging attains a limiting covariance no larger than that without gradient averaging, while the averaged iterate even attains the *minimax optimal* limiting covariance, matching that of the exact Newton method. To our knowledge, this is the first analysis showing that randomness introduced to improve computational efficiency need not compromise statistical efficiency as long as proper gradient estimates are used. The key analytical challenge is that gradient averaging destroys the crucial martingale difference structure underlying central limit theorem arguments. We overcome this challenge by analyzing the coupled two-timescale recursions governing the iterate and the gradient estimator. Numerical experiments on linear and logistic regression support our theoretical covariance comparisons.

1 INTRODUCTION

We consider the stochastic optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{P}}[F(\boldsymbol{x};\xi)], \tag{1.1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective and $F(\cdot;\xi):\mathbb{R}^d\to\mathbb{R}$ is its realization with $\xi\sim\mathcal{P}$. Problem (1.1) is one of the most fundamental building blocks for many decision-making tasks in statistics and machine learning, such as online recommendation (Li et al., 2010), precision medicine (Kosorok & Laber, 2019), and portfolio allocation (Fan et al., 2012). In these applications, (1.1) is interpreted as a parameter estimation problem, where $\boldsymbol{x}$ denotes the model parameter, ξ denotes an observation, and the minimizer $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$ denotes the true model parameter.

The ubiquity of streaming data has made online methods increasingly attractive, where only one *single* sample is processed at each step. Stochastic gradient descent (SGD) and its many acceleration variates are prevalent examples (Robbins & Monro, 1951), with extensively studied convergence properties (Pelletier, 1998; Moulines & Bach, 2011). Although first-order methods enjoy inexpensive updates, their performance can be sensitive to ill-conditioning and heterogeneous gradient noise. Online Newton methods address these issues by preconditioning each stochastic gradient with curvature information (Bottou et al., 2018):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \varphi_t\Delta\boldsymbol{x}_t, \qquad \text{with} \qquad B_t\Delta\boldsymbol{x}_t = -\nabla F(\boldsymbol{x}_t;\xi_t), \tag{1.2}$$

where B_t approximates the Hessian $\nabla^2 f(\boldsymbol{x}_t)$. Beyond convergence, reliable decision-making also requires uncertainty quantification for the resulting estimator. A standard approach for SGD is Polyak–Ruppert averaging, which averages the iterates to yield an asymptotically normal estimator (Ruppert,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1988; Polyak & Juditsky, 1992). Building on these prior works, the asymptotic properties of both the last and averaged iterates of online Newton methods have also been established in unconstrained and constrained settings (Leluc & Portier, 2023; Gao et al., 2025). In particular, under suitable regularity conditions and $B_t \to B^\star = \nabla^2 f(\boldsymbol{x}^\star)$, we have

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0},\ \iota \cdot \Omega) \qquad \text{and} \qquad \sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0},\ \Omega), \tag{1.3}$$

where $\bar{\boldsymbol{x}}_t \coloneqq \frac{1}{t}\sum_{i=0}^{t-1} \boldsymbol{x}_i$, $\Omega = (B^\star)^{-1}\,\mathrm{Cov}(\nabla F(\boldsymbol{x}^\star;\xi))(B^\star)^{-1}$ (see (2.6)), and $\iota = C_\varphi/(2C_\varphi - 1)$ if $\varphi_t = C_\varphi/(t+1)^\varphi$ with $\varphi = 1$ while $\iota = 0.5$ if $\varphi \in (0.5, 1)$. The limiting covariance Ω is further known to be *asymptotic minimax optimal* in Hájek and Le Cam sense (Duchi & Ruan, 2021; Davis et al., 2024). Despite promising statistical properties, a key computational bottleneck of online Newton methods is solving the Newton system in (1.2), which requires $O(d^3)$ flops at each step.

Online sketched Newton methods approximately solve the Newton system using an inner randomized iterative solver that performs a fixed number of sketch-and-project updates, reducing the per-step cost to $O(d^2)$ (Na & Mahoney, 2025). A recent Nesterov-accelerated variant further improves the decay rate of the solver's approximation error while preserving the same order of computational cost (Dereziński et al., 2025). Despite these computational gains, the sketching solver degrades the favorable statistical efficiency of the outer Newton iterates. In particular, Wang et al. (2026) and Du et al. (2026) established asymptotic normality analogous to (1.3) for the last and averaged sketched Newton iterates, with limiting covariances $\Sigma^l_{\boldsymbol{g},\mathrm{as}}$ and $\Sigma^a_{\boldsymbol{g},\mathrm{as}}$, respectively. Their results reveal an interesting computational–statistical trade-off: randomized sketching avoids the exact Newton solve but introduces an additional covariance term that inflates the limiting covariance. Specifically, the last and averaged iterates have limiting covariances satisfying $\Omega \preceq \Sigma^a_{\boldsymbol{g},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}}$. Therefore, we ask:

Can sketching solver reduce the cost of online Newton without inflating the limiting covariance?

In this paper, we provide an affirmative answer to this question by replacing the single-sample gradient $\boldsymbol{g}_t \coloneqq \nabla F(\boldsymbol{x}_t;\xi_t)$ with its moving average $\boldsymbol{m}_t = (1-\eta_t)\boldsymbol{m}_{t-1} + \eta_t \boldsymbol{g}_t$. Gradient averaging has long been used to smooth stochastic estimates (Ruszczyński & Syski, 1983), and recent work studies its interaction with Polyak–Ruppert averaging in first-order methods (Tang et al., 2023). Here, however, we use gradient averaging for quite a different purpose: to weaken the interaction between sampling noise and the randomness introduced by the sketching solver in Newton methods. We establish convergence and asymptotic normality of both the last and averaged sketched Newton iterates with gradient averaging. In particular, under the same standard conditions, we show that **(i)** the last-iterate covariance satisfies $\Sigma^l_{\boldsymbol{m},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}}$; and **(ii)** the averaged-iterate covariance satisfies $\Sigma^a_{\boldsymbol{m},\mathrm{as}} = \Omega$. Thus, gradient averaging mitigates the computational–statistical trade-off for the last iterate, and even eliminates it for the averaged iterate.

The analysis requires addressing the following **challenges** introduced by gradient averaging.

(a) $\boldsymbol{m}_t$ averages stochastic gradients evaluated at different iterates and is therefore not conditionally unbiased for $\nabla f(\boldsymbol{x}_t)$, unlike $\boldsymbol{g}_t$, whose centered noise forms crucial martingale difference structure underlying central limit theorem arguments. More precisely, the tracking error $\boldsymbol{e}_t \coloneqq \boldsymbol{m}_t - \nabla f(\boldsymbol{x}_t)$ contains inherited error from earlier steps, gradient drift, and martingale noise, and must be analyzed jointly with the optimization error $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{x}_t - \boldsymbol{x}^\star$. As such, we derive coupled recursions for $(\boldsymbol{e}_t, \boldsymbol{\Delta}_t)$ and control their evolution on the two timescales governed by η_t and φ_t.

(b) To characterize how gradient averaging changes the leading sketching noise, we must separate sampling noise, sketching fluctuations, and variations in the Hessian approximation and sketching parameters across iterations. Since the resulting perturbation terms are sufficiently small only when $(\boldsymbol{\Delta}_t, B_t - B^\star, \boldsymbol{e}_t)$ remain in a local neighborhood of zero, we introduce a stopping time at the first exit from this neighborhood and localize the convergence rate analysis (see (C.25) in the proof of Lemma 4.2). The effect of the localization is then removed by our almost-sure convergence guarantee.

(c) To isolate $\nabla F(\boldsymbol{x}^\star;\xi_t)$ as the leading stochastic noise in the asymptotic linear representations established in Theorem 4.3, we rewrite the outer Newton recursion in terms of the stochastic gradient $\boldsymbol{g}_t$. This introduces weighted differences $\boldsymbol{m}_t - \boldsymbol{m}_{t-1}$ that do not appear in the analysis of vanilla online Newton methods. We use summation by parts to transfer the differences from the moving-average gradients to their weights, and then bound the resulting boundary and weighted-sum terms to show that they are asymptotically negligible under both the last- and averaged-iterate normalizations.

Our main **contributions** are summarized as follows.

(a) We establish global almost-sure convergence of online sketched Newton with gradient averaging. Under standard two-timescale conditions, we show that the iterate $\boldsymbol{x}_t$ converges to $\boldsymbol{x}^\star$ and the moving-average gradient $\boldsymbol{m}_t$ tracks $\nabla f(\boldsymbol{x}_t)$ almost surely (Theorem 3.5). Moreover, gradient averaging requires no additional oracle evaluations and preserves the inner-solver contraction bound.

(b) We establish last-iterate asymptotic normality, $1/\sqrt{\varphi_t}\cdot(\boldsymbol{x}_t-\boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^l_{\boldsymbol{m},\mathrm{as}})$ (Theorem 4.5), and prove $\Sigma^l_{\boldsymbol{m},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}}$ (Prop. 4.6). Thus, gradient averaging yields no larger asymptotic variance in any direction than vanilla online sketched Newton. Our local convergence rates and asymptotic linear representations reveal the mechanism behind this variance reduction: the random matrix $I-\widetilde{K}^\star_i$, which captures the limiting effect of the inner solver on gradient noise, is replaced by its mean $I-K^\star$ (Theorem 4.3). Consequently, sketching enters the limiting covariance only through its mean effect.

(c) We further establish averaged-iterate asymptotic normality, $\sqrt{t}\,(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^a_{\boldsymbol{m},\mathrm{as}})$, with $\Sigma^a_{\boldsymbol{m},\mathrm{as}} = \Omega \preceq \Sigma^a_{\boldsymbol{g},\mathrm{as}}$ (Theorem 4.7). Polyak–Ruppert averaging eliminates the remaining mean effect of sketching and recovers the covariance Ω of exact online Newton method, which is asymptotically *minimax optimal* (Duchi & Ruan, 2021). This result shows that randomized sketching solver can gain computational efficiency without sacrificing statistical efficiency.

Notation. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm for matrices. We write $O(\cdot)$ and $o(\cdot)$ ($O_p(\cdot)$ and $o_p(\cdot)$) for the usual big- and small-o notation (in probability). We use I and $\mathbf{0}$ to denote identity matrices and zero vectors/matrices, with dimensions clear from context. For symmetric A and B, $A \succeq B$ means $A-B$ is positive semidefinite. We abbreviate $\nabla f_t = \nabla f(\boldsymbol{x}_t)$ and $\nabla^2 f_t = \nabla^2 f(\boldsymbol{x}_t)$. We write $\Sigma^q_{u,s}$ for limiting covariances, where $u \in \{\boldsymbol{g}, \boldsymbol{m}\}$ denotes the stochastic gradient or its moving average; $s \in \{\mathrm{e}, \mathrm{as}\}$ denotes the solver (exact or accelerated sketching); and $q \in \{l, a\}$ denotes the iterate (last or averaged). In particular, we know from (1.3) that $\Omega = \Sigma^a_{\boldsymbol{g},\mathrm{e}}$.

2 Preliminaries

We review online sketched Newton methods, which replace the exact Newton solve with an inexact randomized sketching solver (Na & Mahoney, 2025; Kuang et al., 2026). These methods were subsequently improved by leveraging Nesterov's acceleration (Wang et al., 2026; Du et al., 2026). In what follows, we first introduce the vanilla online Newton, then describe the accelerated sketching solver (of which the unaccelerated solver is a special case), and finally present the preliminary guarantees.

• **Online Newton with Hessian averaging.** Online Newton methods precondition stochastic gradients using curvature information. Given $\boldsymbol{x}_t$, we draw $\xi_t \sim \mathcal{P}$ independently of the past and compute $\boldsymbol{g}_t := \nabla F(\boldsymbol{x}_t;\xi_t)$ and $H_t := \nabla^2 F(\boldsymbol{x}_t;\xi_t)$. As in standard online Newton methods, we assume access to H_t; if unavailable, it may be approximated by quasi-Newton or finite-difference schemes (Dennis & Moré, 1977; Spall, 2000). For $t \geq 1$, we use the Hessian average $B_t := \sum_{i=0}^{t-1} H_i/t = (1-1/t)B_{t-1}+H_{t-1}/t$, which is fixed before ξ_t is drawn. Hessian averaging stabilizes the curvature estimate and improves local convergence behavior (Na et al., 2022).

Given a prescribed stepsize sequence $\{\varphi_t\}_{t\geq 0}$, the vanilla online Newton method updates as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \varphi_t \Delta\boldsymbol{x}_t, \qquad \text{with} \qquad B_t\Delta\boldsymbol{x}_t = -\boldsymbol{g}_t \quad \text{and} \quad B_t := \frac{1}{t}\sum_{i=0}^{t-1} H_i. \tag{2.1}$$

Solving this Newton system costs $O(d^3)$ flops per step, motivating the sketching solver next.

• **Nesterov's accelerated sketch-and-project.** The inner solver for (2.1) combines randomized sketch-and-project updates with Nesterov's acceleration, yielding faster decay rate of the approximation error at the same order of per-step cost (Gower et al., 2018; Dereziński et al., 2025). We suppress the outer index t and present the accelerated solver $\texttt{NASketch}(B, -\boldsymbol{g}; \alpha, \beta, \gamma, \tau)$ in Algorithm 1, where (α, β, γ) are the acceleration parameters and τ is the number of sketching steps.

At each outer iteration, $(B, \boldsymbol{g})$ are held fixed, and we draw $S_j \overset{\text{iid}}{\sim} S$ for $0 \leq j \leq \tau-1$. Without acceleration, the sketch-and-project solver performs the update $\boldsymbol{z}_{j+1} = \boldsymbol{z}_j - BS_j(S_j^\top B^2 S_j)^\dagger S_j^\top(B\boldsymbol{z}_j+\boldsymbol{g})$, where $\dagger$ denotes the pseudoinverse. Nesterov's acceleration introduces an auxiliary momentum sequence $\boldsymbol{v}_j$. Given $(\boldsymbol{z}_j, \boldsymbol{v}_j)$, Algorithm 1 first forms the middle point $\boldsymbol{y}_j = \alpha\boldsymbol{v}_j+(1-\alpha)\boldsymbol{z}_j$, computes a sketching correction at $\boldsymbol{y}_j$, and updates $(\boldsymbol{z}_j, \boldsymbol{v}_j)$ jointly. By Algorithm 1, the setup $(\alpha, \beta, \gamma) = (0.5, 0, 1)$ leads to $\boldsymbol{v}_j = \boldsymbol{z}_j$ and reduces the method to unaccelerated sketch-and-project.

Algorithm 1 Sketch-and-Project with Nesterov's Acceleration: $\texttt{NASketch}(B, -\boldsymbol{g}; \alpha, \beta, \gamma, \tau)$

Initialize: $\boldsymbol{z}_0 = \boldsymbol{v}_0 = \mathbf{0}$, $\alpha \in (0,1)$, $\beta \in [0,1)$, $\gamma \geq 1$, $\tau \geq 1$.
for $j = 0, 1, \ldots, \tau - 1$ **do**
  $\boldsymbol{y}_j \leftarrow \alpha \boldsymbol{v}_j + (1-\alpha)\boldsymbol{z}_j$ and draw a sketching matrix $S_j \sim S$, where $S_j \in \mathbb{R}^{d \times s}$, $1 \leq s \leq d$.
  $\boldsymbol{w}_j \leftarrow B S_j (S_j^\top B^2 S_j)^\dagger S_j^\top (B\boldsymbol{y}_j + \boldsymbol{g})$.
  $(\boldsymbol{z}_{j+1}, \boldsymbol{v}_{j+1}) \leftarrow (\boldsymbol{y}_j - \boldsymbol{w}_j, \beta \boldsymbol{v}_j + (1-\beta)\boldsymbol{y}_j - \gamma \boldsymbol{w}_j)$.
end for
Output: $\boldsymbol{z}_\tau$.

Following Gower et al. (2018); Dereziński et al. (2025), we parameterize (α, β, γ) by (μ, ν) as

$$\alpha = 1/(1+\gamma\nu), \qquad \beta = 1 - \sqrt{\mu/\nu}, \qquad \gamma = 1/\sqrt{\mu\nu}. \tag{2.2}$$

For a generic sketching matrix S, define $\widetilde{Z} \coloneqq BS(S^\top B^2 S)^\dagger S^\top B$ and $Z \coloneqq \mathbb{E}[\widetilde{Z}]$, where the expectations below are taken over S. When $Z \succ \mathbf{0}$, define

$$\mu \coloneqq \inf_{\boldsymbol{a} \neq \mathbf{0}} \frac{\boldsymbol{a}^\top Z \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{a}}, \qquad \nu \coloneqq \sup_{\boldsymbol{a} \neq \mathbf{0}} \frac{\boldsymbol{a}^\top \mathbb{E}[\widetilde{Z} Z^{-1} \widetilde{Z}] \boldsymbol{a}}{\boldsymbol{a}^\top Z \boldsymbol{a}}. \tag{2.3}$$

At outer iteration t, evaluating (2.3) at B_t gives (μ_t, ν_t) and hence $(\alpha_t, \beta_t, \gamma_t)$. Since the exact computation of these parameters requires expectations over the sketching distribution, (2.3) primarily serves as a theoretical parameterization for analyzing the solver's convergence; implementations may approximate these quantities using periodically updated estimates or fixed parameter choices (Wang et al., 2026). With specified parameters, the online sketched Newton method replaces the exact Newton direction $\Delta \boldsymbol{x}_t$ in (2.1) with $\boldsymbol{z}_{t,\tau}$:

$$\boldsymbol{z}_{t,\tau} = \texttt{NASketch}(B_t, -\boldsymbol{g}_t; \alpha_t, \beta_t, \gamma_t, \tau), \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \varphi_t \boldsymbol{z}_{t,\tau}. \tag{2.4}$$

• **Theory of accelerated sketched Newton.** We next review the inner-solver contraction bound and existing asymptotic normality results for online sketched Newton methods. At any fixed outer iteration t, each inner update of the solution approximation in Algorithm 1 is linear in $\boldsymbol{g}_t$. More specifically, by Wang et al. (2026, Lemma 3.5), there exists a realized coefficient factor matrix $\widetilde{K}_t$ for the solution approximation error, independent of $\boldsymbol{g}_t$, such that (see (B.2) for the precise form of $\widetilde{K}_t$)

$$\boldsymbol{z}_{t,\tau} - \Delta \boldsymbol{x}_t = \widetilde{K}_t (\boldsymbol{z}_{t,0} - \Delta \boldsymbol{x}_t), \qquad \Delta \boldsymbol{x}_t = -B_t^{-1} \boldsymbol{g}_t. \tag{2.5}$$

With the above identity, the accelerated inner solver satisfies the following contraction bound.

Lemma 2.1 (Dereziński et al. (2025))**.** Suppose $Z_t = \mathbb{E}[\widetilde{Z}_t] = \mathbb{E}[B_t S (S^\top B_t^2 S)^\dagger S^\top B_t] \succ \mathbf{0}$, then

$$\mathbb{E}[\|\boldsymbol{z}_{t,j} - \Delta \boldsymbol{x}_t\|^2] \leq 2(1 - \sqrt{\mu_t/\nu_t})^j \|\boldsymbol{z}_{t,0} - \Delta \boldsymbol{x}_t\|^2, \qquad \text{for any } t \geq 0,\ j \geq 0,$$

where the expectation is taken over the inner sketches $\{S_j\}_j$.

By Gower et al. (2018, Lemma 2), $1 \leq \nu_t \leq 1/\mu_t$; thus, $1 - \sqrt{\mu_t/\nu_t} \leq 1 - \mu_t$, where the right-hand side is the contraction rate of vanilla unaccelerated sketch-and-project (Gower & Richtárik, 2015).

To introduce asymptotic normality, we need additional notation. Let $B^\star \coloneqq \nabla^2 f(\boldsymbol{x}^\star)$ be the Hessian at $\boldsymbol{x}^\star$ and $(\mu^\star, \nu^\star, \alpha^\star, \beta^\star, \gamma^\star)$ be parameters obtained from (2.2) and (2.3) with $B = B^\star$. We also define $\widetilde{K}^\star$ by evaluating (2.5) at these limiting quantities (cf. (B.2)), $K^\star \coloneqq \mathbb{E}[\widetilde{K}^\star]$, and two matrices

$$\Omega = (B^\star)^{-1} \mathbb{E}[\nabla F(\boldsymbol{x}^\star; \xi) \nabla F(\boldsymbol{x}^\star; \xi)^\top] (B^\star)^{-1} \quad \text{and} \quad \Gamma_{\boldsymbol{g},\mathrm{as}} \coloneqq \mathbb{E}[(I - \widetilde{K}^\star)\Omega(I - \widetilde{K}^\star)^\top]. \tag{2.6}$$

Let $\varphi_t \coloneqq C_\varphi/(t+1)^\varphi$, $\zeta_\varphi \coloneqq \mathbf{1}_{\{\varphi=1\}}/(2C_\varphi)$, and $A_\varphi^\star \coloneqq (I - K^\star) - \zeta_\varphi I$. The following theorem summarizes the existing normality results for both the last and the averaged iterates.

Theorem 2.2 (Wang et al. (2026); Du et al. (2026))**.** Under the assumptions in Theorem 4.3, we have

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^l_{\boldsymbol{g},\mathrm{as}}), \qquad \text{where} \quad A_\varphi^\star \Sigma^l_{\boldsymbol{g},\mathrm{as}} + \Sigma^l_{\boldsymbol{g},\mathrm{as}} A_\varphi^\star = \Gamma_{\boldsymbol{g},\mathrm{as}}. \tag{2.7}$$

Moreover, if $\varphi \in (0.5, 1)$, then the averaged iterate $\bar{\boldsymbol{x}}_t = \sum_{i=0}^{t-1} \boldsymbol{x}_i/t$ satisfies

$$\sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^a_{\boldsymbol{g},\mathrm{as}}), \qquad \text{where} \quad \Sigma^a_{\boldsymbol{g},\mathrm{as}} \coloneqq (I - K^\star)^{-1} \Gamma_{\boldsymbol{g},\mathrm{as}} (I - K^\star)^{-1}. \tag{2.8}$$

The last-iterate result recovers exact online Newton (1.3) by setting $\widetilde{K}^\star = K^\star = \mathbf{0}$ in (2.7). In general, however, when $\varphi = C_\varphi = 1$, it has been shown that $\Sigma^l_{\boldsymbol{g},\mathrm{as}} \succeq \Omega$. Since Ω is the classical Polyak–Ruppert covariance and the local asymptotic minimax bound (Duchi & Ruan, 2021), this ordering reveals a computational–statistical trade-off: randomized sketching reduces the cost of the Newton solve but inflates the limiting covariance.

Analogously, the averaged-iterate result (2.8) also recovers exact online Newton (1.3). However, it has also been shown that $\Sigma^l_{\boldsymbol{g},\mathrm{as}} \succeq \Sigma^a_{\boldsymbol{g},\mathrm{as}} \succeq \Omega$. Thus, while iterate averaging reduces limiting covariance relative to the last iterate, it does not fully eliminate the sketch-induced covariance inflation, and the trade-off persists. In this paper, we completely eliminate the trade-off by gradient averaging.

3 Global Convergence with Gradient Averaging

In this section, we incorporate a moving-average gradient into accelerated sketched Newton method, introduce assumptions, and establish almost-sure convergence of the method. Specifically, instead of using $\boldsymbol{g}_t$ directly, we initialize $\boldsymbol{m}_0 = \boldsymbol{g}_0$ and update

$$\boldsymbol{m}_t = (1-\eta_t)\boldsymbol{m}_{t-1} + \eta_t \boldsymbol{g}_t, \qquad \eta_t \in (0,1], \quad t \geq 1, \tag{3.1}$$

where $\{\eta_t\}_{t\geq 1}$ is a prescribed averaging sequence. We then replace $\boldsymbol{g}_t$ in (2.4) with $\boldsymbol{m}_t$ and obtain

$$\boldsymbol{z}_{t,\tau} = \texttt{NASketch}(B_t, -\boldsymbol{m}_t; \alpha_t, \beta_t, \gamma_t, \tau), \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \varphi_t \boldsymbol{z}_{t,\tau}. \tag{3.2}$$

This replacement changes only the right-hand side of (2.1); hence, B_t, $(\alpha_t, \beta_t, \gamma_t)$, and the inner sketching procedure remain unchanged, as does the contraction bound in Lemma 2.1. Moreover, updating $\boldsymbol{m}_t$ requires no additional oracle evaluations and incurs only $O(d)$ computation, preserving the per-step cost order. Since $\boldsymbol{m}_t$ averages gradients evaluated at different iterates, it is not conditionally unbiased for ∇f_t and destroys the crucial martingale difference structure. Thus, establishing global convergence requires joint control of the optimization error $\boldsymbol{\Delta}_t = \boldsymbol{x}_t - \boldsymbol{x}^\star$ and the gradient-tracking error $\boldsymbol{e}_t = \boldsymbol{m}_t - \nabla f_t$. We first specify the filtration and assumptions used here and in Section 4.

For $t \geq 0$, we define σ-algebras $\mathcal{F}_{t-0.5} = \sigma\left(\{\xi_i, \{S_{i,j}\}_j\}_{i=0}^{t-1} \cup \xi_t\right)$ and $\mathcal{F}_t = \sigma\left(\{\xi_i, \{S_{i,j}\}_j\}_{i=0}^{t}\right)$. For consistency, let $(\boldsymbol{x}_0, B_0)$ be deterministic and $\mathcal{F}_{-1} := \sigma(\boldsymbol{x}_0, B_0)$ be the trivial σ-algebra. Then, we know $(\boldsymbol{g}_t, H_t, \boldsymbol{m}_t)$ and $(\boldsymbol{z}_{t,\tau}, \boldsymbol{x}_{t+1}, B_{t+1})$ are $\mathcal{F}_{t-0.5}$- and $\mathcal{F}_t$-measurable, respectively.

The first assumption regards the strong convexity and Lipschitz smoothness of the objective.

Assumption 3.1. We assume that f and its realization $F(\cdot;\xi)$ are twice continuously differentiable, with $\gamma_H I \preceq \nabla^2 F(\boldsymbol{x};\xi) \preceq \Upsilon_H I$ for any $\boldsymbol{x}$ and some constants $\gamma_H, \Upsilon_H > 0$. We also assume that $\nabla^2 f$ is Υ_L-Lipschitz continuous.

We next impose growth conditions on the moments of the gradient and Hessian noises.

Assumption 3.2. For any $t \geq 0$, we assume the gradient estimate is unbiased $\mathbb{E}[\boldsymbol{g}_t \mid \mathcal{F}_{t-1}] = \nabla f_t$ and satisfies a growth condition on its q_g-th moment: for some constants $C_{g,1}, C_{g,2} > 0$,

$$\mathbb{E}[\|\boldsymbol{g}_t - \nabla f_t\|^{q_g} \mid \mathcal{F}_{t-1}] \leq C_{g,1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^{q_g} + C_{g,2}.$$

Assumption 3.3. For any $t \geq 0$, we assume the Hessian estimate is unbiased $\mathbb{E}[H_t \mid \mathcal{F}_{t-1}] = \nabla^2 f_t$ and satisfies a growth condition on its second moment: for some constants $C_{H,1}, C_{H,2} > 0$,

$$\mathbb{E}[\|H_t - \nabla^2 f_t\|^2 \mid \mathcal{F}_{t-1}] \leq C_{H,1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + C_{H,2}.$$

We will specify q_g in the theorem statements later. In particular, $q_g = 2$ suffices for global convergence and $q_g > 2$ for asymptotic normality; Assumption 3.3 will be used only for asymptotic normality in Section 4. Finally, we present an assumption on the sketching distribution.

Assumption 3.4. For any $t \geq 0$, we assume there exists $\gamma_S > 0$ such that $\{S_{t,j}\}_{j=0}^{\tau-1} \overset{iid}{\sim} S$ satisfying $Z_t := \mathbb{E}[B_t S(S^\top B_t^2 S)^\dagger S^\top B_t \mid \mathcal{F}_{t-1}] \succeq \gamma_S I$, and $\{S_{t,j}\}_{j=0}^{\tau-1}$ are independent of ξ_t. Moreover, the condition number of S has bounded expectation: $\mathbb{E}[\|S\|\|S^\dagger\|] \leq \Upsilon_S$ for $\Upsilon_S > 0$.

Assumptions 3.1–3.4 follow directly from the existing literature on online sketched Newton inference (Na & Mahoney, 2025; Wang et al., 2026; Du et al., 2026; Kuang et al., 2026), and the conditions on the gradient are also used in SGD inference (Chen et al., 2020; Zhu et al., 2021). For regression problems, they can be verified by the arguments in Chen et al. (2020, Appendix A). The lower bound

on expected projection matrix Z_t is required to ensure convergence of sketching solvers (Gower et al., 2018). These conditions hold for various sketching distributions, including Gaussian and Kaczmarz sketching (Strohmer & Vershynin, 2008).

Under the assumptions shown above, the following theorem establishes joint global almost-sure convergence of the optimization error Δ_t and tracking error $\boldsymbol{e}_t$.

Theorem 3.5. Consider the Newton scheme (3.2) with $\varphi_t = C_\varphi/(t+1)^\varphi$, $\eta_t = C_\eta/(t+1)^\eta$ satisfying $C_\varphi > 0$, $C_\eta \in (0,1]$ and $0.5 < \eta < \varphi \leq 1$. Suppose Assumptions 3.1, 3.2, and 3.4 hold with $q_g = 2$ and τ satisfies $\tau(1-\sqrt{\mu_t/\nu_t})^{\tau-2} \leq \gamma_H/(4\Upsilon_H)$ for all $t \geq 0$. Then, we have $\boldsymbol{e}_t = \boldsymbol{m}_t - \nabla f_t \overset{a.s.}{\to} \mathbf{0}$ and $\boldsymbol{x}_t \overset{a.s.}{\to} \boldsymbol{x}^\star$ as $t \to \infty$.

Theorem 3.5 requires only the bounded variance condition on the stochastic gradient, which is standard in stochastic optimization (Ghadimi & Lan, 2013). The condition $\eta < \varphi$ ensures that $\boldsymbol{m}_t$ adopts a larger averaging sequence η_t than the outer Newton stepsize φ_t, so as to preserve as much gradient information from $\boldsymbol{g}_t$ as possible despite the continued evolution of $\boldsymbol{x}_t$. Moreover, Lemma 2.1 and the condition on τ ensure that the approximate Newton direction is sufficiently accurate for global convergence. By Wang et al. (2026, (3.5)), it is known that $1-\sqrt{\mu_t/\nu_t} \leq 1-\gamma_S$, so the condition on τ can indeed be satisfied uniformly across steps. Overall, Theorem 3.5 suggests that although $\boldsymbol{m}_t$ is generally not conditionally unbiased for ∇f_t, it tracks ∇f_t almost surely, and the outer iterate $\boldsymbol{x}_t$ also converges to $\boldsymbol{x}^\star$. This global result underlies the local analysis in Section 4.

4 Optimal Statistical Efficiency via Gradient Averaging

We study how moving-average gradient estimators in (3.1) change the statistical effect of sketching. We first establish local asymptotic rates and show that the centered sketching fluctuation is asymptotically negligible with gradient averaging. We then derive asymptotic normality of the last and averaged Newton iterates. The resulting limits reveal a sharp distinction (Theorem 4.3): gradient averaging reduces the random sketching factor in the leading stochastic noise to its mean, thereby reducing the sketching-induced fluctuation of the last iterate, while averaging the outer iterates further cancels the remaining mean factor and completely eliminates the inflation in asymptotic uncertainty.

4.1 Local rates and sketching noise reduction

We begin by showing the almost-sure convergence of the Hessian average B_t, the acceleration parameters $(\alpha_t, \beta_t, \gamma_t)$, and, more importantly, the evolution coefficient matrix of the sketching approximation error, $K_t := \mathbb{E}[\widetilde{K}_t \mid \mathcal{F}_{t-1}]$, from (2.5). Explicit form of $\widetilde{K}_t$ is given in (B.2) and not crucial here.

Lemma 4.1. Under the conditions of Theorem 3.5 and Assumption 3.3, we have $B_t \overset{a.s.}{\to} B^\star$, $(\alpha_t, \beta_t, \gamma_t) \overset{a.s.}{\to} (\alpha^\star, \beta^\star, \gamma^\star)$, $K_t \overset{a.s.}{\to} K^\star$; and $K^\star$ is symmetric with $\|K^\star\| \leq \gamma_H/(2\Upsilon_H) < 1$.

With the limiting quantities in place, we establish the local rates needed for asymptotic normality.

Lemma 4.2. Suppose Assumptions 3.1–3.4 hold with $q_g > 2$; τ satisfies $\tau(1-\sqrt{\mu_t/\nu_t})^{\tau-2} \leq \gamma_H/(4\Upsilon_H)$ for all $t \geq 0$; and the stepsize parameters satisfy $C_\eta \in (0,1]$ and $0.5 < \eta < \varphi \leq 1$. If $\varphi = 1$, we further suppose $C_\varphi > 1/\{2\lambda_{\min}(I-K^\star)\}$. Then, we have

$$\max\left\{\|\boldsymbol{m}_t - \nabla f_t\|,\ \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|,\ \|B_t - B^\star\|,\ \|(\alpha_t,\beta_t,\gamma_t) - (\alpha^\star,\beta^\star,\gamma^\star)\|,\ \|K_t - K^\star\|\right\} = O_p(\sqrt{\eta_t}).$$

The rates in Lemma 4.2 are obtained via a localization argument. Specifically, we introduce the first exit time in (C.25) to restrict the analysis to trajectories for which the iterate and gradient-tracking errors remain within a sufficiently small neighborhood of zero. This localization enables control of the resulting perturbation terms. The almost-sure convergence established in Theorem 3.5 and Lemma 4.1 ensures that the exit probability can be made arbitrarily small, thereby converting the localized bounds into the stated $O_p(\sqrt{\eta_t})$ rates. By Lemma 4.2, the centered sketching residual $(\widetilde{K}_t - K_t)B_t^{-1}\boldsymbol{m}_t$ forms a martingale difference sequence whose conditional second moment is of order $O_p(\eta_t)$. Thus, its cumulative contribution is negligible under the iterate normalization.

The following theorem states the resulting asymptotic linear representations.

Theorem 4.3. Suppose Assumptions 3.1–3.4 hold with $q_g > 2$ and τ satisfies $\tau(1-\sqrt{\mu_t/\nu_t})^{\tau-2} \leq \gamma_H/(4\Upsilon_H)$ for all $t \geq 0$. Let $C_\varphi > 0$, $C_\eta \in (0,1]$, $0.5 < \eta < \varphi \leq 1$, and $C_\varphi > 1/\{2\lambda_{\min}(I-K^\star)\}$

when $\varphi = 1$. Then, the last-iterate admits the linear representation

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) = -1/\sqrt{\varphi_t} \cdot \sum_{i=0}^{t-1} \Pi_{t,i+1} \varphi_i (I - K^\star)(B^\star)^{-1} \nabla F(\boldsymbol{x}^\star; \xi_i) + o_p(1), \tag{4.1}$$

where $\Pi_{n,i} \coloneqq \prod_{j=i}^{n-1} \{I - \varphi_j (I - K^\star)\}$ for $n > i$ with $\Pi_{n,n} \coloneqq I$. Moreover, if $\varphi < 1$, we have

$$\sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) = 1/\sqrt{t} \cdot \sum_{i=0}^{t-1} (\boldsymbol{x}_i - \boldsymbol{x}^\star) = -1/\sqrt{t} \cdot \sum_{i=0}^{t-1} (B^\star)^{-1} \nabla F(\boldsymbol{x}^\star; \xi_i) + o_p(1). \tag{4.2}$$

Remark 4.4 (Interaction between gradient averaging and randomized sketching). The conditions of Theorem 4.3 match those imposed for vanilla online sketched Newton methods (Wang et al., 2026; Du et al., 2026) (except that we have an additional averaging sequence η_t satisfying $\varphi_t = o(\eta_t)$). To isolate the effect of gradient averaging, consider vanilla Newton sketch that uses $\boldsymbol{g}_t$ in place of $\boldsymbol{m}_t$. For each $t \geq 0$, let $\widetilde{K}_t^\star$ denote the coefficient matrix in (2.5) constructed from the step-t sketches and evaluated at $B^\star$ and $(\alpha^\star, \beta^\star, \gamma^\star)$. Building upon Wang et al. (2026, Lemma B.3), we can derive the following linear representation for vanilla Newton sketch:

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) = -1/\sqrt{\varphi_t} \cdot \sum_{i=0}^{t-1} \Pi_{t,i+1} \varphi_i (I - \widetilde{K}_i^\star)(B^\star)^{-1} \nabla F(\boldsymbol{x}^\star; \xi_i) + o_p(1). \tag{4.3}$$

Comparing (4.1) and (4.3) reveals the effect of gradient averaging. For our method, the localized moment bounds in Lemma 4.2 imply that the centered sketching term $(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t$ has a second moment of order $O(\eta_t)$ and therefore does not contribute to the leading stochastic sum. In contrast, for vanilla Newton sketch, the corresponding centered sketching term $(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{g}_t$ has a constant order of second moment, which inevitably inflates the leading stochastic sum.

The results in (4.1)–(4.3) reveal two successive effects. First, replacing the stochastic gradient with its moving average turns the sketching factor $I - \widetilde{K}_i^\star$ into its mean $I - K^\star$ in the leading stochastic term of the last iterate. Second, averaging the Newton iterates further introduces the factor $(I - K^\star)^{-1}$, which cancels the remaining mean sketching factor. Consequently, (4.2) contains no sketch-dependent matrix and has the same leading stochastic term as exact online Newton. In contrast, the averaged iterate of vanilla sketched Newton still contains the centered sketching factor $(I - K^\star)^{-1}(\widetilde{K}_i^\star - K^\star)$ in its leading term, so the sketching randomness persists in its asymptotic distribution.

4.2 Asymptotic normality and covariance comparisons

In this section, we build upon Theorem 4.3 to establish asymptotic normality and to explore how gradient averaging and randomized sketching affect statistical efficiency. We begin with the last iterate.

Theorem 4.5. Under the conditions of Theorem 4.3, let $\Gamma_{\boldsymbol{m},\mathrm{as}} \coloneqq (I - K^\star)\Omega(I - K^\star)$ and we have

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Sigma^l_{\boldsymbol{m},\mathrm{as}}), \qquad \text{where} \quad A_\varphi^\star \Sigma^l_{\boldsymbol{m},\mathrm{as}} + \Sigma^l_{\boldsymbol{m},\mathrm{as}} A_\varphi^\star = \Gamma_{\boldsymbol{m},\mathrm{as}}. \tag{4.4}$$

Recall from Theorem 2.2 that $A_\varphi^\star = (I - K^\star) - \zeta_\varphi I$ and $\zeta_\varphi = \mathbf{1}_{\{\varphi=1\}}/(2C_\varphi)$. The result (4.4) characterizes $\Sigma^l_{\boldsymbol{m},\mathrm{as}}$ through a Lyapunov equation in which randomized sketching enters only through the mean factor matrix $K^\star$. When $\varphi = 1$, the condition $C_\varphi > 1/\{2\lambda_{\min}(I - K^\star)\}$ ensures that $A_\varphi^\star \succ \boldsymbol{0}$. Hence, the Lyapunov equation in (4.4) admits a unique positive semidefinite solution.

We next show the reduction in $\Sigma^l_{\boldsymbol{m},\mathrm{as}}$ compared to $\Sigma^l_{\boldsymbol{g},\mathrm{as}}$ (the asymptotic covariance of vanilla online Newton sketch), and we relate both to the asymptotic covariance of exact online Newton.

Proposition 4.6. Under the conditions of Theorem 4.3, the following statements hold.

(a) We have $\Gamma_{\boldsymbol{m},\mathrm{as}} \preceq \Gamma_{\boldsymbol{g},\mathrm{as}}$ and $\Sigma^l_{\boldsymbol{m},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}}$.

(b) If the Newton system (2.1) is solved exactly ($\widetilde{K}^\star = K^\star = \boldsymbol{0}$), the last-iterate covariances reduce to $\Sigma^l_{\boldsymbol{m},\mathrm{e}} = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = 0.5\Omega$ for $\varphi \in (0.5, 1)$ and $\Sigma^l_{\boldsymbol{m},\mathrm{e}} = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = C_\varphi/(2C_\varphi - 1) \cdot \Omega$ for $\varphi = 1$.

(c) ($\sqrt{t}$-consistency) Let $\varphi_t = 1/t$ (i.e., $\varphi = C_\varphi = 1$). Then, $\Omega = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = \Sigma^l_{\boldsymbol{m},\mathrm{e}} \preceq \Sigma^l_{\boldsymbol{m},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}}$.

Part **(a)** shows the covariance reduction achieved by gradient averaging. The gap $\Sigma^l_{\boldsymbol{g},\mathrm{as}} - \Sigma^l_{\boldsymbol{m},\mathrm{as}}$ is positive semidefinite and arises from the centered sketching fluctuation $K^\star - \widetilde{K}^\star$. Part **(b)** shows that gradient averaging leaves the last-iterate covariance of exact online Newton in (1.3) unchanged. Part **(c)** compares sketched and exact Newton with $\varphi_t = 1/t$, so that all four last-iterate covariances correspond to the same $\sqrt{t}$ normalization. In this regime, the remaining gap $\Sigma^l_{\boldsymbol{m},\mathrm{as}} - \Omega$ stems from the factor $I - K^\star$ in the Lyapunov equation (4.4). Thus, gradient averaging eliminates the covariance contribution from the centered sketching fluctuation $K^\star - \widetilde{K}^\star$, while the residual dependence on $I - K^\star$ in (4.1) prevents the covariance from further reducing to that of exact Newton method.

Lastly, we turn to the Polyak–Ruppert average $\bar{\boldsymbol{x}}_t$ studied in (4.2).

Theorem 4.7. Under the conditions of Theorem 4.3, suppose that $\varphi < 1$. Then, we have

$$\sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^a_{\boldsymbol{m},\mathrm{as}}), \qquad \text{where} \quad \Sigma^a_{\boldsymbol{m},\mathrm{as}} = \Omega. \tag{4.5}$$

The result (4.2) reduces the averaged iterate to a normalized sum of i.i.d. mean-zero random vectors up to an $o_p(1)$ term. Therefore, the multivariate central limit theorem and Slutsky's theorem yield (4.5). To quantify the effect of gradient averaging, consider vanilla online sketched Newton in (2.4). By Remark 4.4, $\Gamma_{\boldsymbol{g},\mathrm{as}}$ in (2.6), and Du et al. (2026, Theorem 4.2), its averaged-iterate covariance is

$$\begin{aligned}\Sigma^a_{\boldsymbol{g},\mathrm{as}} &\coloneqq (I - K^\star)^{-1}\Gamma_{\boldsymbol{g},\mathrm{as}}(I - K^\star)^{-1} \\ &= \Omega + (I - K^\star)^{-1}\mathbb{E}[(\widetilde{K}^\star - K^\star)\Omega(\widetilde{K}^\star - K^\star)^\top](I - K^\star)^{-1} \succeq \Omega. \end{aligned} \tag{4.6}$$

Comparing (4.5) with (4.6) reveals a sharp statistical distinction. Vanilla online sketched Newton suffers from a positive semidefinite covariance inflation induced by the centered sketching fluctuation, while $\Sigma^a_{\boldsymbol{m},\mathrm{as}} = \Omega$. The covariance Ω coincides with that of averaged exact online Newton in (1.3) and that of averaged SGD (Ruppert, 1988; Polyak & Juditsky, 1992), which is known to be asymptotically minimax optimal (Duchi & Ruan, 2021; Davis et al., 2024). Therefore, for every fixed τ satisfying the conditions of Theorem 3.5, accelerated sketching incurs **no computational–statistical trade-off for the averaged iterate**: it significantly *reduces the cost* of Newton system solve while preserving the asymptotically *optimal statistical efficiency* achieved by averaged exact Newton and SGD.

5 Numerical Experiments

We compare online Newton methods with and without gradient averaging, using either accelerated sketching or exact solves, and include SGD as a baseline. We consider linear and logistic regression with Gaussian designs, vary $d \in \{20, 40\}$, and vary the design covariance structure among identity, equicorrelation, or Toeplitz. Kaczmarz and Gaussian sketches with $\tau \in \{5, 10\}$ yield 48 matched settings. Each configuration uses $R = 200$ independent runs of $T = 10^5$ iterations. All methods share the same data streams, and matched sketching methods additionally share the same sketches. $\widehat{\Sigma}$ denotes the centered empirical covariances, computed across runs, of $(\boldsymbol{x}_t - \boldsymbol{x}^\star)/\sqrt{\varphi_t}$ and $\sqrt{t}(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star)$. Bias and mean squared error (MSE) are evaluated separately. Appendix D provides complete experimental setups and additional results.

• **Covariance reduction with gradient averaging.** Figure 1(a,c) shows that gradient averaging reduces the last-iterate covariance trace with the same data streams, sketches, inner sketching solve budgets, and outer stepsizes. In particular, each point on the plot corresponds to a specific configuration, and all points lie below the line $y = x$. Thus, the reduction holds across design covariance structures and sketching distributions, including higher-dimensional settings reported in Appendix D. By comparison, Table 2 shows that the covariance trace of exact Newton remains nearly unchanged under gradient averaging. These observations align with Proposition 4.6: gradient averaging reduces the last-iterate limiting covariance of sketched Newton while leaving that of exact Newton unchanged.

• **Comparison with exact online Newton.** Figure 1(b,d,e,f) shows that gradient averaging brings the averaged-iterate covariance trace closer to the exact Newton benchmark, since not only do all points lie below $y = x$, but more points also lie close to $y = 1$. Note that both vanilla sketched Newton and our method adopt iterate averaging, so the improvement is purely attributable to gradient averaging. Theorem 4.7 explains this comparison: gradient averaging along with iterate averaging recovers Ω asymptotically, while vanilla sketched Newton retains an additional sketch-dependent covariance contribution. The trajectories in panels (e,f) further support this conclusion: with gradient averaging, the covariance trace approaches $\mathrm{tr}(\Omega)$ while the number of sketching steps per iteration remains fixed.

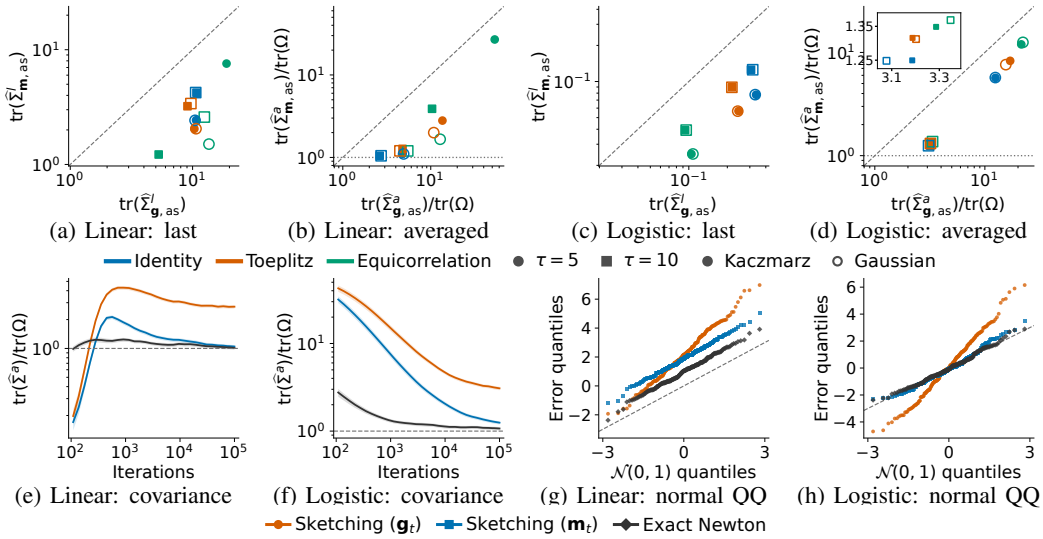


Figure 1: *Covariance comparisons ($d = 20$, $T = 10^5$, $R = 200$). Panels (a–d) compare covariance traces without (horizontal axis) and with (vertical axis) gradient averaging across 12 settings. All averaged-iterate covariance traces are normalized by $\mathrm{tr}(\Omega)$. For identity design covariance, Gaussian sketches, and $\tau = 10$, panels (e,f) show covariance trace trajectories with pointwise 95% bootstrap intervals, while panels (g,h) show standard normal QQ plots of $\sqrt{T}\,\mathbf{1}^\top(\bar{\boldsymbol{x}}_T - \boldsymbol{x}^\star)/\sqrt{\mathbf{1}^\top\Omega\mathbf{1}}$ without recentering. Diagonal lines indicate equality, and horizontal lines indicate one. The inset magnifies the six points corresponding to $\tau = 10$. Sketching uses $\boldsymbol{g}_t$ without gradient averaging and $\boldsymbol{m}_t$ with gradient averaging, while exact Newton uses $\boldsymbol{g}_t$.*

Table 1: *Comparisons at $T = 10^5$ with $R = 200$. Covariance-trace and MSE ratios compare sketched Newton with gradient averaging to its vanilla counterpart. The central columns report $\mathrm{tr}(\widehat{\Sigma}^a_{u,\mathrm{as}})/\,\mathrm{tr}(\Omega)$ for $u \in \{\boldsymbol{g}, \boldsymbol{m}\}$. Entries are medians of six settingwise ratios across three design covariances and two sketching distributions.*

			Covariance-trace ratio		Averaged trace/ $\mathrm{tr}(\Omega)$		MSE ratio	
Model	d	τ	Last	Averaged	$\boldsymbol{g}$	$\boldsymbol{m}$	Last	Averaged
Linear	20	5	0.213	0.215	11.808	1.830	0.213	0.249
	20	10	0.352	0.323	4.619	1.201	0.352	0.349
	40	5	0.164	0.232	66.756	19.108	0.234	0.301
	40	10	0.208	0.219	15.775	3.299	0.208	0.253
Logistic	20	5	0.233	0.490	16.234	7.953	0.233	0.483
	20	10	0.411	0.409	3.195	1.314	0.412	0.408
	40	5	0.686	0.854	751.846	641.821	0.691	0.814
	40	10	0.226	0.510	18.316	9.343	0.225	0.496

• **Estimation accuracy and normal approximation.** From Table 1, we observe that gradient averaging reduces the MSE of both the last and averaged iterates across all tested settings. The bias comparisons in Appendix D show similar improvements, indicating that the covariance reduction is accompanied by improved estimation accuracy. Increasing τ also brings the averaged-iterate covariance trace closer to $\mathrm{tr}(\Omega)$. Thus, although the limiting covariance is independent of sketching, more accurate inner solves remain beneficial at finite iterations. Figure 1(g,h) further complements these comparisons. In particular, with gradient averaging, the logistic regression QQ plot closely follows the standard normal reference. The linear regression QQ plot exhibits a similar pattern but is shifted due to finite-sample bias; this shift is still smaller with gradient averaging.

6 Conclusion

We studied statistical efficiency of online Newton methods with Nesterov accelerated sketching and gradient averaging. We established almost-sure convergence and asymptotic normality of the last and averaged iterates. We showed that gradient averaging yields a last-iterate covariance no larger than that of vanilla online sketched Newton, while subsequent averaging of the outer iterates further achieves the minimax optimal limiting covariance of exact online Newton. Thus, for averaged iterates, a fixed number of sketching steps can reduce the cost of an expensive Newton solve without inflating the limiting covariance. Gradient averaging requires no extra oracle evaluations and leaves the contraction bound of the inner sketching solver unchanged. Experiments on linear and logistic regression support our free-lunch finding and further demonstrate improvements in bias and mean squared error.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript. We have reviewed all AI-assisted text and take responsibility for the final manuscript, including text, claims, proofs, and experimental results.

REFERENCES

Rajendra Bhatia. *Matrix Analysis*. Springer New York, 1997. ISBN 9781461206538.

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

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

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in non-smooth stochastic approximation. *The Annals of Statistics*, 52(4), August 2024.

John E Dennis, Jr and Jorge J Moré. Quasi-newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

Michal Dereziński, Daniel LeJeune, Deanna Needell, and Elizaveta Rebrova. Fine-grained analysis and faster algorithms for iteratively solving linear systems. *Journal of Machine Learning Research*, 26(144):1–49, 2025.

Xinchen Du, Elizaveta Rebrova, Michał Dereziński, and Sen Na. Inference for newton methods with accelerated sketch-and-project via random scaling. *arXiv preprint arXiv:2609.12421*, 2026.

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

Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, Berlin New York, 2013. ISBN 9783540571001.

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

Yihang Gao, Michael K Ng, Michael W Mahoney, and Sen Na. Online inference of constrained optimization: Primal-dual optimality and sequential quadratic programming. *arXiv preprint arXiv:2512.08948*, 2025.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, January 2013. ISSN 1095-7189.

Robert Gower, Filip Hanzely, Peter Richtárik, and Sebastian U Stich. Accelerated stochastic matrix inversion: general theory and speeding up bfgs rules for faster second-order optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

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

Wei Kuang, Mihai Anitescu, and Sen Na. Online covariance matrix estimation in sketched newton methods. *Information and Inference: A Journal of the IMA*, 15(2):iaag012, 2026. doi: 10.1093/imaiai/iaag012.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Sen Na and Michael Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *Journal of Machine Learning Research*, 26(33):1–75, 2025.

Sen Na, Michał Dereziński, and Michael W. Mahoney. Hessian averaging in stochastic Newton methods achieves superlinear convergence. *Mathematical Programming*, dec 2022.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2018. ISBN 9783319915784.

Mariane Pelletier. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Processes and Their Applications*, 78(2):217–244, 1998.

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

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

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, 1988.

Andrzej Ruszczyński and Wojciech Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12):1097–1105, December 1983. doi: 10.1109/TAC.1983.1103184.

J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10):1839–1853, October 2000. ISSN 0018-9286.

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

Kejie Tang, Weidong Liu, Yichen Zhang, and Xi Chen. Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023.

Haoxuan Wang, Xinchen Du, and Sen Na. Inference of online newton methods with nesterov's accelerated sketching. *International Conference on Machine Learning*, 2026.

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, July 2021.

Appendix: "A Free Lunch: Gradient Averaging Achieves Optimal Statistical Efficiency in Newton Sketch"

CONTENTS

A Preparation Lemmas **13**

B Proofs of Section 3 **13**

B.1 Proof of Theorem 3.5 14

B.2 Proof of Lemma B.1 18

C Proofs of Section 4 **19**

C.1 Proof of Lemma 4.1 19

C.2 Proof of Lemma C.2 22

C.3 Proof of Lemma 4.2 23

C.4 Proof of Lemma C.3 28

C.5 Proof of Lemma C.4 30

C.6 Proof of Theorem 4.3 31

C.7 Proof of Theorem 4.5 41

C.8 Proof of Proposition 4.6 44

C.9 Proof of Theorem 4.7 45

D Experimental Details and Additional Results **45**

D.1 Detailed experimental setup 45

D.2 Covariance estimation and statistical comparisons 46

D.3 Additional covariance comparisons 47

D.4 Normal QQ plots with method-specific covariances 48

D.5 Estimation error and computational cost 48

A Preparation Lemmas

Lemma A.1 (Kronecker's lemma)**.** Suppose $\{A_t\}_{t\geq 0}$ satisfies

$$\sum_{t=0}^{\infty} A_t < \infty.$$

Then, for any positive, non-decreasing sequence $\{a_t\}_{t\geq 0}$ satisfying $a_t \to \infty$, we have

$$\frac{1}{a_T}\sum_{t=1}^{T} a_t A_t \to 0 \qquad \text{as} \qquad T \to \infty.$$

Lemma A.2 (Stolz–Cesàro theorem)**.** Let $\{a_t\}_{t\geq 1}$ and $\{b_t\}_{t\geq 1}$ be two sequences of real numbers. Suppose that $\{b_t\}_{t\geq 1}$ is strictly monotone and divergent, and that the following limit exists:

$$\lim_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t} = \ell.$$

Then,

$$\lim_{t\to\infty} \frac{a_t}{b_t} = \ell.$$

Lemma A.3. (Leluc & Portier, 2023, Lemma 3) Let $\{\beta_t\}_{t\geq 1}$ be a nonnegative sequence converging to zero, and let $\lambda > 0$, $m \geq 1$, and $q \geq 0$. Consider two nonnegative sequences $\{A_t\}_{t\geq 0}$ and $\{\varepsilon_t\}_{t\geq 1}$ such that, for all sufficiently large t,

$$A_t = (1-\lambda\beta_t)^m A_{t-1} + \beta_t^{q+1}\varepsilon_t.$$

If $\beta_t = \beta_0/(t+1)^a$ with $a \in (0,1)$, then

$$\limsup_{t\to\infty} \frac{A_t}{\beta_t^q} \leq \frac{1}{m\lambda}\limsup_{t\to\infty} \varepsilon_t.$$

If $\beta_t = \beta_0/(t+1)$ and $q < m\lambda\beta_0$, then

$$\limsup_{t\to\infty} \frac{A_t}{\beta_t^q} \leq \frac{\beta_0}{m\lambda\beta_0 - q}\limsup_{t\to\infty} \varepsilon_t.$$

Lemma A.4. (Hall & Heyde, 2014, Corollary 3.1) For every $t \geq 1$, let $\{W_{t,i}\}_{i=0}^{t-1}$ be a triangular array of $\mathbb{R}^d$-valued random vectors such that $W_{t,i}$ is $\mathcal{F}_i$-measurable and

$$\mathbb{E}[W_{t,i} \mid \mathcal{F}_{i-1}] = \mathbf{0}, \tag{A.1}$$

$$\sum_{i=0}^{t-1} \mathbb{E}[W_{t,i}W_{t,i}^\top \mid \mathcal{F}_{i-1}] \xrightarrow{p} V^\star \succeq \mathbf{0}, \tag{A.2}$$

$$\sum_{i=0}^{t-1} \mathbb{E}[\|W_{t,i}\|^2 \mathbf{1}_{\{\|W_{t,i}\|>\epsilon\}} \mid \mathcal{F}_{i-1}] \xrightarrow{p} 0, \qquad \forall\, \epsilon > 0. \tag{A.3}$$

Then, we have

$$\sum_{i=0}^{t-1} W_{t,i} \xrightarrow{d} \mathcal{N}(\mathbf{0}, V^\star).$$

B Proofs of Section 3

Throughout this section, we write $f_t := f(\boldsymbol{x}_t)$ and $f^\star := f(\boldsymbol{x}^\star)$.

B.1 Proof of Theorem 3.5

We first introduce the inner-solver quantities used in the proof. For every $t \geq 0$, we define

$$\widetilde{Z}_{t,j} \coloneqq B_t S_{t,j} (S_{t,j}^\top B_t^2 S_{t,j})^\dagger S_{t,j}^\top B_t \in \mathbb{R}^{d\times d}$$

and

$$\widetilde{C}_{t,j} \coloneqq \begin{pmatrix} 1-\alpha_t & \alpha_t \\ (1-\alpha_t)(1-\beta_t) & \alpha_t+\beta_t-\alpha_t\beta_t \end{pmatrix} \otimes I_d - \begin{pmatrix} 1-\alpha_t & \alpha_t \\ (1-\alpha_t)\gamma_t & \alpha_t\gamma_t \end{pmatrix} \otimes \widetilde{Z}_{t,j}. \tag{B.1}$$

Let $\widetilde{C}_t \coloneqq \widetilde{C}_{t,\tau-1}\cdots\widetilde{C}_{t,0}$ and write $[\widetilde{C}_t]_{a,b}$ for its (a,b)-th $d\times d$ block. The realized residual matrix introduced in (2.5) has the explicit representation

$$\widetilde{K}_t \coloneqq [\widetilde{C}_t]_{1,1} + [\widetilde{C}_t]_{1,2} = (I \quad \mathbf{0})\, \widetilde{C}_t \begin{pmatrix} I \\ I \end{pmatrix}. \tag{B.2}$$

Moreover, we define

$$M_t \coloneqq \begin{pmatrix} 1-\alpha_t & \alpha_t \\ (1-\alpha_t)(1-\beta_t) & \alpha_t+\beta_t-\alpha_t\beta_t \end{pmatrix} \otimes I_d - \begin{pmatrix} 1-\alpha_t & \alpha_t \\ (1-\alpha_t)\gamma_t & \alpha_t\gamma_t \end{pmatrix} \otimes Z_t. \tag{B.3}$$

The conditional independence of the inner sketches gives

$$C_t \coloneqq \mathbb{E}[\widetilde{C}_t \mid \mathcal{F}_{t-1}] = M_t^\tau, \qquad K_t \coloneqq \mathbb{E}[\widetilde{K}_t \mid \mathcal{F}_{t-1}] = (I \quad \mathbf{0})\, C_t \begin{pmatrix} I \\ I \end{pmatrix}.$$

The matrices $\widetilde{K}_t$ and K_t are precisely the realized and mean residual matrices introduced in Section 2. Indeed, the block matrices in (B.1) and their conditional means depend on B_t, the acceleration parameters, and the current sketches or their distribution, but not on the right-hand side of the Newton system. Therefore, replacing $\boldsymbol{g}_t$ with $\boldsymbol{m}_t$ changes neither $\widetilde{K}_t$ nor K_t. Consequently, the bound $1-\sqrt{\mu_t/\nu_t} \leq 1-\gamma_S < 1$ from Wang et al. (2026, Lemma 3.6) and the mean-residual bound from Wang et al. (2026, Lemma 3.7) continue to hold without modification. For the moving-average gradient estimates, the exact direction $\Delta\boldsymbol{x}_t = -B_t^{-1}\boldsymbol{m}_t$, and the conditional-mean of the approximate direction $\boldsymbol{z}_{t,\tau}$ must be formulated with respect to $\mathcal{F}_{t-0.5}$.

The following auxiliary lemma records these two relations. Its proof is deferred to Appendix B.2.

Lemma B.1. Define $\Delta\boldsymbol{x}_t \coloneqq -B_t^{-1}\boldsymbol{m}_t$. Then, for every $t \geq 0$, we have

$$\boldsymbol{z}_{t,\tau} - \Delta\boldsymbol{x}_t = \widetilde{K}_t(\boldsymbol{z}_{t,0} - \Delta\boldsymbol{x}_t), \qquad \mathbb{E}[\boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-0.5}] = -(I-K_t)B_t^{-1}\boldsymbol{m}_t.$$

Consequently, $(\boldsymbol{z}_{t,\tau} - (I-K_t)\Delta\boldsymbol{x}_t, \mathcal{F}_t)_{t\geq 0}$ forms a martingale difference sequence. Furthermore, if $\gamma_t = 1$, then $\widetilde{K}_t$ has a simple form of $\widetilde{K}_t = (I-\widetilde{Z}_{t,\tau-1})\cdots(I-\widetilde{Z}_{t,0})$.

The realized block matrices in (B.1) and $\widetilde{K}_t$ depend on B_t, the acceleration parameters, and the inner sketches, while their conditional means depend on B_t, the acceleration parameters, and the sketch distribution. None of these matrices depends on the right-hand side of the Newton system. Therefore, replacing $\boldsymbol{g}_t$ with $\boldsymbol{m}_t$ does not affect the bound $1-\sqrt{\mu_t/\nu_t} \leq 1-\gamma_S < 1$ from Wang et al. (2026, Lemma 3.6) or the mean-residual bound from Wang et al. (2026, Lemma 3.7); only the conditional-mean relation in Lemma B.1 must be stated with respect to $\mathcal{F}_{t-0.5}$.

By Assumption 3.1, $f(\boldsymbol{x})$ is strongly convex. By the basic property of strong convexity (Nesterov, 2018), we have the following:

$$\frac{\gamma_H}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \leq f_t - f^\star \leq \frac{1}{2\gamma_H}\|\nabla f_t\|^2. \tag{B.4}$$

Still, by the Υ_H-Lipschitz continuity of $\nabla f(\boldsymbol{x})$ implied by Assumption 3.1, we have

$$\frac{1}{2\Upsilon_H}\|\nabla f_t\|^2 \leq f_t - f^\star \leq \frac{\Upsilon_H}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2. \tag{B.5}$$

Furthermore, the Υ_H-Lipschitz continuity of $\nabla f(\boldsymbol{x})$ yields

$$f_{t+1} - f^\star \leq f_t - f^\star + \varphi_t \nabla f_t^\top \boldsymbol{z}_{t,\tau} + \frac{\Upsilon_H}{2}\varphi_t^2\|\boldsymbol{z}_{t,\tau}\|^2.$$

Taking the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_{t-0.5}]$ on both sides yields

$$\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-0.5}] \le f_t - f^\star + \varphi_t \mathbb{E}[\nabla f_t^\top \boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-0.5}] + \frac{\Upsilon_H}{2}\varphi_t^2 \mathbb{E}[\|\boldsymbol{z}_{t,\tau}\|^2 \mid \mathcal{F}_{t-0.5}]. \quad \text{(B.6)}$$

Also, by Assumption 3.1 and the construction of B_t, we have $\gamma_H I \preceq B_t \preceq \Upsilon_H I$.

We first control the conditional mean of the search direction. If $\mu_t = 1$, the eigendirection representation in the proof of Wang et al. (2026, Lemma 3.7) gives $K_t = \mathbf{0}$. If $\mu_t < 1$, Wang et al. (2026, Lemma 3.7) and the condition on τ give

$$\|K_t\| \le 2\tau \left(1 - \sqrt{\frac{\mu_t}{\nu_t}}\right)^{\tau-2} \le 2 \cdot \frac{\gamma_H}{4\Upsilon_H} \le \frac{\gamma_H}{2\Upsilon_H}.$$

Therefore, in either case, $\|K_t\| \le \gamma_H/(2\Upsilon_H)$ for every $t \ge 0$. By Lemma B.1, we obtain

$$\begin{aligned}
\varphi_t \mathbb{E}[\nabla f_t^\top \boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-0.5}] &= -\varphi_t \nabla f_t^\top (I - K_t) B_t^{-1} (\nabla f_t + \boldsymbol{e}_t) \\
&\le -\frac{\varphi_t}{\Upsilon_H}\|\nabla f_t\|^2 + \frac{\varphi_t}{\gamma_H}\|K_t\|\|\nabla f_t\|^2 + \frac{\varphi_t}{\gamma_H}(1 + \|K_t\|)\|\nabla f_t\|\|\boldsymbol{e}_t\| \\
&\le -\frac{\varphi_t}{2\Upsilon_H}\|\nabla f_t\|^2 + \frac{\varphi_t}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\|\nabla f_t\|\|\boldsymbol{e}_t\| \\
&\le -\frac{\varphi_t}{4\Upsilon_H}\|\nabla f_t\|^2 + \frac{\Upsilon_H \varphi_t}{\gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 \|\boldsymbol{e}_t\|^2, \qquad \text{(B.7)}
\end{aligned}$$

where the last inequality follows from the Young's inequality

$$\frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) ab \le \frac{1}{4\Upsilon_H}a^2 + \frac{\Upsilon_H}{\gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 b^2.$$

To control the second moment of the search direction, we first establish a uniform bound for $\|\widetilde{K}_t\|$. Each $\widetilde{Z}_{t,j}$ is an orthogonal projection, and the definitions of the acceleration parameters together with $1 \le \nu_t \le 1/\mu_t$ and $\mu_t \ge \gamma_S$, imply

$$\|\widetilde{Z}_{t,j}\| \le 1, \qquad \|I - \widetilde{Z}_{t,j}\| \le 1, \qquad 1 \le \gamma_t = \frac{1}{\sqrt{\mu_t \nu_t}} \le \frac{1}{\sqrt{\mu_t}} \le \frac{1}{\sqrt{\gamma_S}}.$$

Consequently,

$$\begin{aligned}
\|\widetilde{K}_t\| &= \left\| (I \quad \mathbf{0})\, \widetilde{C}_t \begin{pmatrix} I \\ I \end{pmatrix} \right\| \le \sqrt{2}\|\widetilde{C}_t\| \le \sqrt{2}\prod_{j=0}^{\tau-1}\|\widetilde{C}_{t,j}\| \\
&\le \sqrt{2}\prod_{j=0}^{\tau-1}\left\{\|I - \widetilde{Z}_{t,j}\| + (1-\alpha_t)(1-\beta_t) + (1-\alpha_t)\gamma_t\|\widetilde{Z}_{t,j}\| + (\alpha_t + \beta_t - \alpha_t\beta_t) + \alpha_t\gamma_t\|\widetilde{Z}_{t,j}\|\right\} \\
&\le \sqrt{2}\prod_{j=0}^{\tau-1}\left(\|I - \widetilde{Z}_{t,j}\| + 1 + \gamma_t\|\widetilde{Z}_{t,j}\|\right) \\
&\le \sqrt{2}\left(2 + \frac{1}{\sqrt{\gamma_S}}\right)^\tau =: C_K. \qquad \text{(B.8)}
\end{aligned}$$

By Lemma B.1 and (B.8), we immediately have

$$\mathbb{E}[\|\boldsymbol{z}_{t,\tau}\|^2 \mid \mathcal{F}_{t-0.5}] \le \frac{2(1 + C_K)^2}{\gamma_H^2}\left(\|\nabla f_t\|^2 + \|\boldsymbol{e}_t\|^2\right). \qquad \text{(B.9)}$$

Substituting (B.7) and (B.9) into (B.6) yields

$$\begin{aligned}
\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-0.5}] \le f_t - f^\star &- \left\{\frac{1}{4\Upsilon_H} - \frac{\Upsilon_H(1 + C_K)^2}{\gamma_H^2}\varphi_t\right\}\varphi_t\|\nabla f_t\|^2 \\
&+ \frac{\Upsilon_H}{\gamma_H^2}\left\{\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 + (1 + C_K)^2\varphi_t\right\}\varphi_t\|\boldsymbol{e}_t\|^2.
\end{aligned}$$

Since $\varphi_t \to 0$, there exists a deterministic index after which $\varphi_t \le \min\left\{1, \gamma_H^2/[8\Upsilon_H^2(1+C_K)^2]\right\}$. For such sufficiently large t, we have

$$\mathbb{E}[f_{t+1} - f^\star \mid \mathcal{F}_{t-0.5}] \le f_t - f^\star - \frac{\varphi_t}{8\Upsilon_H}\|\nabla f_t\|^2 + \frac{\Upsilon_H}{\gamma_H^2}\left\{\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 + (1+C_K)^2\right\}\varphi_t\|\boldsymbol{e}_t\|^2. \quad \text{(B.10)}$$

We next control the moving-average error $\boldsymbol{e}_t$, which admits the recursion

$$\boldsymbol{e}_{t+1} = (1-\eta_{t+1})(\boldsymbol{e}_t + \nabla f_t - \nabla f_{t+1}) + \eta_{t+1}(\boldsymbol{g}_{t+1} - \nabla f_{t+1}). \quad \text{(B.11)}$$

Conditional on $\mathcal{F}_t$, the first term on the right-hand side is fixed. By Assumption 3.2, we have

$$\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_t] = (1-\eta_{t+1})^2\|\boldsymbol{e}_t + \nabla f_t - \nabla f_{t+1}\|^2 + \eta_{t+1}^2\mathbb{E}[\|\boldsymbol{g}_{t+1} - \nabla f_{t+1}\|^2 \mid \mathcal{F}_t]. \quad \text{(B.12)}$$

For $\eta_{t+1} \in (0,1)$, applying Young's inequality to the cross term gives

$$\begin{aligned}
&(1-\eta_{t+1})^2\|\boldsymbol{e}_t + \nabla f_t - \nabla f_{t+1}\|^2 \\
&= (1-\eta_{t+1})^2\left\{\|\boldsymbol{e}_t\|^2 + 2\boldsymbol{e}_t^\top(\nabla f_t - \nabla f_{t+1}) + \|\nabla f_t - \nabla f_{t+1}\|^2\right\} \\
&\le (1-\eta_{t+1})^2\left\{\left(1 + \frac{\eta_{t+1}}{1-\eta_{t+1}}\right)\|\boldsymbol{e}_t\|^2 + \left(1 + \frac{1-\eta_{t+1}}{\eta_{t+1}}\right)\|\nabla f_t - \nabla f_{t+1}\|^2\right\} \\
&= (1-\eta_{t+1})\|\boldsymbol{e}_t\|^2 + \frac{(1-\eta_{t+1})^2}{\eta_{t+1}}\|\nabla f_t - \nabla f_{t+1}\|^2 \\
&\le (1-\eta_{t+1})\|\boldsymbol{e}_t\|^2 + \eta_{t+1}^{-1}\|\nabla f_t - \nabla f_{t+1}\|^2.
\end{aligned}$$

When $\eta_{t+1} = 1$, the left-hand side is zero, so the same inequality holds trivially. Moreover, the Υ_H-Lipschitz continuity of $\nabla f(\boldsymbol{x})$ and the outer update imply

$$\|\nabla f_{t+1} - \nabla f_t\| \le \Upsilon_H\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \Upsilon_H\varphi_t\|\boldsymbol{z}_{t,\tau}\|.$$

Applying these inequalities to (B.12), taking the conditional expectation with respect to $\mathcal{F}_{t-0.5}$, and using the tower property and (B.9), we obtain

$$\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] \le (1-\eta_{t+1})\|\boldsymbol{e}_t\|^2 + \frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\frac{\varphi_t^2}{\eta_{t+1}}\left(\|\nabla f_t\|^2 + \|\boldsymbol{e}_t\|^2\right) + \eta_{t+1}^2\mathbb{E}[\|\boldsymbol{g}_{t+1} - \nabla f_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}].$$

By Assumption 3.2 with $q_g = 2$, we have

$$\mathbb{E}[\|\boldsymbol{g}_{t+1} - \nabla f_{t+1}\|^2 \mid \mathcal{F}_t] \le C_{g,1}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + C_{g,2}.$$

For all sufficiently large t, (B.4), (B.5), and (B.9) yield

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \mid \mathcal{F}_{t-0.5}] &\le 2\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + 2\varphi_t^2\mathbb{E}[\|\boldsymbol{z}_{t,\tau}\|^2 \mid \mathcal{F}_{t-0.5}] \\
&\le \left\{\frac{4}{\gamma_H} + \frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}\right\}(f_t - f^\star) + \frac{4(1+C_K)^2}{\gamma_H^2}\|\boldsymbol{e}_t\|^2.
\end{aligned}$$

Therefore, by the tower property,

$$\mathbb{E}[\|\boldsymbol{g}_{t+1} - \nabla f_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] \le C_{g,1}\left\{\frac{4}{\gamma_H} + \frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}\right\}(f_t - f^\star) + \frac{4C_{g,1}(1+C_K)^2}{\gamma_H^2}\|\boldsymbol{e}_t\|^2 + C_{g,2}.$$

Substituting this bound into the preliminary tracking recursion yields

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] \le (1-\eta_{t+1})\|\boldsymbol{e}_t\|^2 + \frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\frac{\varphi_t^2}{\eta_{t+1}}\left(\|\nabla f_t\|^2 + \|\boldsymbol{e}_t\|^2\right) \\
&+ C_{g,1}\left\{\frac{4}{\gamma_H} + \frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}\right\}\eta_{t+1}^2(f_t - f^\star) + \frac{4C_{g,1}(1+C_K)^2}{\gamma_H^2}\eta_{t+1}^2\|\boldsymbol{e}_t\|^2 + C_{g,2}\eta_{t+1}^2.
\end{aligned} \quad \text{(B.13)}$$

Since the prescribed schedules satisfy $\varphi_t/\eta_{t+1} \to 0$, there exists a deterministic index after which

$$\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\frac{\varphi_t}{\eta_{t+1}} \le \frac{1}{16\Upsilon_H},$$
$$\frac{\Upsilon_H}{\gamma_H^2}\left\{\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)^2+(1+C_K)^2\right\}\frac{\varphi_t}{\eta_{t+1}} \le \frac{1}{4},$$
$$\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\left(\frac{\varphi_t}{\eta_{t+1}}\right)^2 \le \frac{1}{4}.$$

Adding (B.10) and (B.13) and using these three inequalities gives us

$$\begin{aligned}&\mathbb{E}[f_{t+1}-f^\star+\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}]\\ &\quad\le\left[1+C_{g,1}\left\{\frac{4}{\gamma_H}+\frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}+\frac{4(1+C_K)^2}{\gamma_H^2}\right\}\eta_{t+1}^2\right](f_t-f^\star+\|\boldsymbol{e}_t\|^2)\\ &\qquad-\frac{\varphi_t}{16\Upsilon_H}\|\nabla f_t\|^2-\frac{\eta_{t+1}}{2}\|\boldsymbol{e}_t\|^2+C_{g,2}\eta_{t+1}^2. \quad \text{(B.14)}\end{aligned}$$

Since $2\eta > 1$, we have $\sum_{t=0}^\infty \eta_{t+1}^2 < \infty$. Applying the Robbins–Siegmund theorem (Duflo, 2013, Theorem 1.3.12) with the filtration $(\mathcal{F}_{t-0.5})_{t\ge 0}$, we obtain that $f_t - f^\star + \|\boldsymbol{e}_t\|^2$ converges to some finite random variable and

$$\sum_{t=0}^\infty \varphi_t\|\nabla f_t\|^2 < \infty, \qquad \sum_{t=0}^\infty \eta_{t+1}\|\boldsymbol{e}_t\|^2 < \infty \quad \text{(B.15)}$$

almost surely. In particular, $f_t - f^\star$ and $\|\boldsymbol{e}_t\|^2$ are bounded almost surely. Moreover,

$$\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\left(\frac{\varphi_t}{\eta_{t+1}}\right)^2+\frac{4C_{g,1}(1+C_K)^2}{\gamma_H^2}\eta_{t+1}\to 0.$$

Hence, for all sufficiently large t, (B.13) implies

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] &\le \left(1-\frac{\eta_{t+1}}{2}\right)\|\boldsymbol{e}_t\|^2+\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\frac{\varphi_t^2}{\eta_{t+1}}\|\nabla f_t\|^2\\ &\quad+C_{g,1}\left\{\frac{4}{\gamma_H}+\frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}\right\}\eta_{t+1}^2(f_t-f^\star)+C_{g,2}\eta_{t+1}^2. \quad \text{(B.16)}\end{aligned}$$

The positive terms on the right-hand side of (B.16) are almost surely summable. Indeed, by (B.15),

$$\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\sum_{t=0}^\infty\frac{\varphi_t^2}{\eta_{t+1}}\|\nabla f_t\|^2=\frac{2\Upsilon_H^2(1+C_K)^2}{\gamma_H^2}\sum_{t=0}^\infty\frac{\varphi_t}{\eta_{t+1}}\left(\varphi_t\|\nabla f_t\|^2\right)<\infty$$

almost surely. Furthermore, the almost-sure boundedness of $f_t - f^\star$ and $\sum_t \eta_{t+1}^2 < \infty$ also imply

$$\sum_{t=0}^\infty \eta_{t+1}^2\left[C_{g,1}\left\{\frac{4}{\gamma_H}+\frac{8\Upsilon_H(1+C_K)^2}{\gamma_H^2}\right\}(f_t-f^\star)+C_{g,2}\right]<\infty$$

almost surely. Again we apply the Robbins–Siegmund theorem to (B.16) and conclude that $\|\boldsymbol{e}_t\|^2$ converges almost surely. Since $\sum_{t=0}^\infty \eta_{t+1} = \infty$ and $\sum_{t=0}^\infty \eta_{t+1}\|\boldsymbol{e}_t\|^2 < \infty$, we can conclude

$$\boldsymbol{e}_t = \boldsymbol{m}_t - \nabla f_t \overset{a.s.}{\to} \mathbf{0}.$$

Finally, the almost-sure convergence of $f_t - f^\star + \|\boldsymbol{e}_t\|^2$ and $\|\boldsymbol{e}_t\|^2 \to 0$ imply that $f_t - f^\star$ also converges almost surely. Since $\sum_{t=0}^\infty \varphi_t = \infty$ and $\sum_{t=0}^\infty \varphi_t\|\nabla f_t\|^2 < \infty$, we have

$$\liminf_{t\to\infty}\|\nabla f_t\| = 0 \qquad \text{almost surely.}$$

It follows from (B.4) that $\liminf_{t\to\infty}(f_t - f^\star) = 0$ almost surely. Given that $f_t - f^\star$ converges almost surely, we conclude that $f_t - f^\star \to 0$ almost surely. Again, we can apply the lower bound in (B.4) and obtain

$$\boldsymbol{x}_t \overset{a.s.}{\to} \boldsymbol{x}^\star,$$

which completes the proof.

B.2 Proof of Lemma B.1

Fix an outer iteration t. In Section 2, we temporarily suppress the outer iteration index. Below, we write $\{\boldsymbol{z}_{t,j}, \boldsymbol{v}_{t,j}, \boldsymbol{y}_{t,j}, \boldsymbol{w}_{t,j}\}_{j=0}^{\tau}$ to make the dependence on t explicit. Recall that $B_t \Delta \boldsymbol{x}_t = -\boldsymbol{m}_t$. It follows from the definition of $\boldsymbol{w}_{t,j}$ in Algorithm 1 that

$$\boldsymbol{w}_{t,j} = B_t S_{t,j} (S_{t,j}^\top B_t^2 S_{t,j})^\dagger S_{t,j}^\top B_t (\boldsymbol{y}_{t,j} - \Delta \boldsymbol{x}_t) = \widetilde{Z}_{t,j} (\boldsymbol{y}_{t,j} - \Delta \boldsymbol{x}_t). \tag{B.17}$$

Consequently, by Algorithm 1,

$$\begin{aligned} \boldsymbol{z}_{t,j+1} - \Delta \boldsymbol{x}_t &= \boldsymbol{y}_{t,j} - \Delta \boldsymbol{x}_t - \boldsymbol{w}_{t,j} \\ &= (I - \widetilde{Z}_{t,j}) \left\{ (1-\alpha_t)(\boldsymbol{z}_{t,j} - \Delta \boldsymbol{x}_t) + \alpha_t (\boldsymbol{v}_{t,j} - \Delta \boldsymbol{x}_t) \right\}, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{v}_{t,j+1} - \Delta \boldsymbol{x}_t &= \beta_t (\boldsymbol{v}_{t,j} - \Delta \boldsymbol{x}_t) + (1-\beta_t)(\boldsymbol{y}_{t,j} - \Delta \boldsymbol{x}_t) - \gamma_t \boldsymbol{w}_{t,j} \\ &= \left\{ (1-\alpha_t)(1-\beta_t) I - (1-\alpha_t)\gamma_t \widetilde{Z}_{t,j} \right\} (\boldsymbol{z}_{t,j} - \Delta \boldsymbol{x}_t) \\ &\quad + \left\{ (\alpha_t + \beta_t - \alpha_t \beta_t) I - \alpha_t \gamma_t \widetilde{Z}_{t,j} \right\} (\boldsymbol{v}_{t,j} - \Delta \boldsymbol{x}_t). \end{aligned}$$

Combining the two displays, we obtain the block recursion

$$\begin{pmatrix} \boldsymbol{z}_{t,j+1} - \Delta \boldsymbol{x}_t \\ \boldsymbol{v}_{t,j+1} - \Delta \boldsymbol{x}_t \end{pmatrix} = \widetilde{C}_{t,j} \begin{pmatrix} \boldsymbol{z}_{t,j} - \Delta \boldsymbol{x}_t \\ \boldsymbol{v}_{t,j} - \Delta \boldsymbol{x}_t \end{pmatrix}, \tag{B.18}$$

where $\widetilde{C}_{t,j}$ is defined in (B.1). Iterating (B.18) yields

$$\begin{pmatrix} \boldsymbol{z}_{t,\tau} - \Delta \boldsymbol{x}_t \\ \boldsymbol{v}_{t,\tau} - \Delta \boldsymbol{x}_t \end{pmatrix} = \widetilde{C}_{t,\tau-1} \cdots \widetilde{C}_{t,0} \begin{pmatrix} \boldsymbol{z}_{t,0} - \Delta \boldsymbol{x}_t \\ \boldsymbol{v}_{t,0} - \Delta \boldsymbol{x}_t \end{pmatrix}. \tag{B.19}$$

Since $\boldsymbol{z}_{t,0} = \boldsymbol{v}_{t,0} = \boldsymbol{0}$, $\widetilde{C}_t = \prod_{j=0}^{\tau-1} \widetilde{C}_{t,j}$, taking the first block row in (B.19) and using (B.2) gives

$$\boldsymbol{z}_{t,\tau} - \Delta \boldsymbol{x}_t = -\widetilde{K}_t \Delta \boldsymbol{x}_t = \widetilde{K}_t (\boldsymbol{z}_{t,0} - \Delta \boldsymbol{x}_t),$$

or, equivalently,

$$\boldsymbol{z}_{t,\tau} = (I - \widetilde{K}_t) \Delta \boldsymbol{x}_t = -(I - \widetilde{K}_t) B_t^{-1} \boldsymbol{m}_t. \tag{B.20}$$

Conditional on $\mathcal{F}_{t-0.5}$, the quantities $(B_t, \boldsymbol{m}_t, \alpha_t, \beta_t, \gamma_t)$ are fixed, and the only randomness in (B.20) comes from the sketching matrices. By Assumption 3.4, we have

$$\mathbb{E}[\widetilde{K}_t \mid \mathcal{F}_{t-0.5}] = \mathbb{E}[\widetilde{K}_t \mid \mathcal{F}_{t-1}] = K_t.$$

Consequently,

$$\mathbb{E}[\boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-0.5}] = -(I - K_t) B_t^{-1} \boldsymbol{m}_t. \tag{B.21}$$

Moreover,

$$\mathbb{E}[\boldsymbol{z}_{t,\tau} - (I - K_t) \Delta \boldsymbol{x}_t \mid \mathcal{F}_{t-0.5}] = \boldsymbol{0}.$$

By the tower property,

$$\mathbb{E}[\boldsymbol{z}_{t,\tau} - (I - K_t) \Delta \boldsymbol{x}_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[\boldsymbol{z}_{t,\tau} - (I - K_t) \Delta \boldsymbol{x}_t \mid \mathcal{F}_{t-0.5}] \mid \mathcal{F}_{t-1}] = \boldsymbol{0},$$

which proves the martingale-difference statement with respect to $\{\mathcal{F}_t\}_{t \geq 0}$.

Finally, suppose that $\gamma_t = 1$. For any $Y \in \mathbb{R}^{d \times d}$, direct multiplication in (B.1) gives

$$\widetilde{C}_{t,j} \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes Y \right\} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes \{(I - \widetilde{Z}_{t,j}) Y\}. \tag{B.22}$$

By the above property, we then obtain

$$\widetilde{C}_t \begin{pmatrix} I_d \\ I_d \end{pmatrix} = \widetilde{C}_{t,\tau-1} \widetilde{C}_{t,\tau-2} \cdots \widetilde{C}_{t,0} \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes I_d \right) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes \left(\prod_{j=0}^{\tau-1} (I_d - \widetilde{Z}_{t,j}) \right).$$

Applying (B.22) successively and then using (B.2), we obtain

$$\widetilde{K}_t = (I \quad \boldsymbol{0}) \, \widetilde{C}_t \begin{pmatrix} I \\ I \end{pmatrix} = (I - \widetilde{Z}_{t,\tau-1}) \cdots (I - \widetilde{Z}_{t,0}).$$

This completes the proof.

C Proofs of Section 4

Throughout this section, let $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{x}_t - \boldsymbol{x}^\star$, $\mathcal{R} \coloneqq I - K^\star$ and $\lambda_m \coloneqq \lambda_{\min}(\mathcal{R})$. We use $a_t = O(b_t)$ and $a_t \lesssim b_t$ interchangeably to simplify notation, where the implicit constants are deterministic and independent of t. We also use $\mathbf{1}_{\{\cdot\}}$ to denote the indicator function.

C.1 Proof of Lemma 4.1

• **Convergence of B_t.** By the construction of B_t, we have

$$\begin{aligned}\|B_t - B^\star\| &= \left\| \frac{1}{t}\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i) + \frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - B^\star) \right\| \\ &\le \left\| \frac{1}{t}\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i) \right\| + \left\| \frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - B^\star) \right\| \\ &\le \left\| \frac{1}{t}\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i) \right\| + \frac{\Upsilon_L}{t}\sum_{i=0}^{t-1}\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|. \end{aligned} \tag{C.1}$$

By Theorem 3.5, $\boldsymbol{x}_i \overset{a.s.}{\to} \boldsymbol{x}^\star$. Hence, the second term on the right-hand side of (C.1) converges to zero almost surely by Lemma A.2. For the first term, $\{H_i - \nabla^2 f_i\}_{i\ge 0}$ is a matrix-valued martingale difference sequence since $\mathbb{E}[H_i \mid \mathcal{F}_{i-1}] = \nabla^2 f_i$. Moreover, Assumption 3.1 implies that $\|H_i - \nabla^2 f_i\|_{\mathrm{F}} \le 2\sqrt{d}\,\Upsilon_H$. Therefore, we have

$$\mathbb{E}\left[\left\|\sum_{i=0}^{t-1}\frac{H_i - \nabla^2 f_i}{i+1}\right\|_{\mathrm{F}}^2\right] = \sum_{i=0}^{t-1}\frac{\mathbb{E}[\|H_i - \nabla^2 f_i\|_{\mathrm{F}}^2]}{(i+1)^2} \le 4d\Upsilon_H^2\sum_{i=0}^{\infty}\frac{1}{(i+1)^2} < \infty,$$

where the equality follows from the orthogonality of martingale differences under the Frobenius inner product. Hence, $\{\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i)/(i+1)\}_{t\ge 1}$ is an L^2-bounded martingale. By Doob's martingale convergence theorem (Hall & Heyde, 2014, Corollary 2.2), $\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i)/(i+1)$ converges almost surely. Applying Lemma A.1 yields

$$\frac{1}{t}\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i) \overset{a.s.}{\to} \mathbf{0}.$$

Combining the preceding conclusion with (C.1), we obtain

$$B_t \overset{a.s.}{\to} B^\star. \tag{C.2}$$

• **Convergence of $(\alpha_t, \beta_t, \gamma_t)$.** Recall from (2.2) that the acceleration parameters $(\alpha_t, \beta_t, \gamma_t)$ (the distinction between $(\alpha_t, \beta_t, \gamma_t)$ and $(\alpha^\star, \beta^\star, \gamma^\star)$ is solely due to the evaluation at (μ_t, ν_t) v.s. $(\mu^\star, \nu^\star)$) can be regarded as functions of (μ_t, ν_t) as follows:

$$\alpha(\mu_t, \nu_t) = \frac{1}{1 + \sqrt{\nu_t/\mu_t}}, \qquad \beta(\mu_t, \nu_t) = 1 - \sqrt{\frac{\mu_t}{\nu_t}}, \qquad \gamma(\mu_t, \nu_t) = \frac{1}{\sqrt{\mu_t \nu_t}}.$$

Each of these functions is continuously differentiable on $(0, \infty)^2$. Since $(\mu_t, \nu_t) \in [\gamma_S, 1] \times [1, 1/\gamma_S]$ for all $t \ge 0$, their gradients are uniformly bounded on this compact and convex set. Therefore, by the mean-value theorem, we further have

$$|\alpha_t - \alpha^\star| + |\beta_t - \beta^\star| + |\gamma_t - \gamma^\star| \lesssim |\mu_t - \mu^\star| + |\nu_t - \nu^\star|. \tag{C.3}$$

Thus, it suffices to study the convergence of (μ_t, ν_t). For a generic sketching matrix S, define

$$\begin{aligned} \widetilde{Z}_t(S) &\coloneqq B_t S(S^\top B_t^2 S)^\dagger S^\top B_t, & Z_t &\coloneqq \mathbb{E}[\widetilde{Z}_t(S) \mid \mathcal{F}_{t-1}], \\ \widetilde{Z}^\star(S) &\coloneqq B^\star S(S^\top (B^\star)^2 S)^\dagger S^\top B^\star, & Z^\star &\coloneqq \mathbb{E}[\widetilde{Z}^\star(S)]. \end{aligned} \tag{C.4}$$

We first recall the following perturbation result for the projection matrix.

Lemma C.1. (Na & Mahoney, 2025, Lemma 5.2) Suppose $B_t, B^\star \in \mathbb{R}^{d\times d}$ are both non-singular. Then, for any sketching matrix $S \in \mathbb{R}^{d\times s}$, we have

$$\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \le \frac{2\|B_t - B^\star\|}{\sigma_{\min}(B^\star)}\|S\|\|S^\dagger\|,$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

By Lemma C.1, Assumption 3.4, and the independence between S and $\mathcal{F}_{t-1}$, we have

$$\mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}] \le \frac{2\|B_t - B^\star\|}{\gamma_H}\mathbb{E}[\|S\|\|S^\dagger\|] \lesssim \|B_t - B^\star\|. \tag{C.5}$$

•• Convergence of μ_t. By the definitions of μ_t and $\mu^\star$, we have $\mu_t = \lambda_{\min}(Z_t)$ and $\mu^\star = \lambda_{\min}(Z^\star)$. Applying Weyl's perturbation theorem (Bhatia, 1997, Corollary III.2.6), we obtain

$$\begin{aligned}|\mu_t - \mu^\star| \le \|Z_t - Z^\star\| &= \Big\|\mathbb{E}[\widetilde{Z}_t(S) - \widetilde{Z}^\star(S) \mid \mathcal{F}_{t-1}]\Big\| \\ &\le \mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}] \lesssim \|B_t - B^\star\|. \end{aligned}\tag{C.6}$$

Since we have shown $B_t \overset{a.s.}{\to} B^\star$, it follows that $\mu_t \overset{a.s.}{\to} \mu^\star$.

•• Convergence of ν_t. Define

$$A_t \coloneqq \mathbb{E}[\widetilde{Z}_t(S)Z_t^{-1}\widetilde{Z}_t(S) \mid \mathcal{F}_{t-1}], \qquad A^\star \coloneqq \mathbb{E}[\widetilde{Z}^\star(S)(Z^\star)^{-1}\widetilde{Z}^\star(S)].$$

The matrices A_t and $A^\star$ are symmetric positive semidefinite. Letting $\boldsymbol{b} = Z_t^{1/2}\boldsymbol{a}$ in (2.3), we obtain

$$\begin{aligned}\nu_t = \sup_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^\top A_t \boldsymbol{a}}{\boldsymbol{a}^\top Z_t \boldsymbol{a}} = \sup_{\boldsymbol{b}\neq\boldsymbol{0}} \frac{\boldsymbol{b}^\top Z_t^{-1/2} A_t Z_t^{-1/2}\boldsymbol{b}}{\boldsymbol{b}^\top \boldsymbol{b}} \\ = \lambda_{\max}(Z_t^{-1/2}A_tZ_t^{-1/2}) = \|Z_t^{-1/2}A_tZ_t^{-1/2}\|. \end{aligned}\tag{C.7}$$

Similarly, we have $\nu^\star = \|(Z^\star)^{-1/2}A^\star(Z^\star)^{-1/2}\|$. Consequently,

$$\begin{aligned}|\nu_t - \nu^\star| &= \Big|\|Z_t^{-1/2}A_tZ_t^{-1/2}\| - \|(Z^\star)^{-1/2}A^\star(Z^\star)^{-1/2}\|\Big| \\ &\le \|Z_t^{-1/2}A_tZ_t^{-1/2} - (Z^\star)^{-1/2}A^\star(Z^\star)^{-1/2}\| \le \|(Z_t^{-1/2} - (Z^\star)^{-1/2})A_tZ_t^{-1/2}\| \\ &\quad + \|(Z^\star)^{-1/2}(A_t - A^\star)Z_t^{-1/2}\| + \|(Z^\star)^{-1/2}A^\star(Z_t^{-1/2} - (Z^\star)^{-1/2})\|. \end{aligned}\tag{C.8}$$

To control the three terms on the right-hand side, we use the following preparation lemma.

Lemma C.2. Under Assumption 3.4, we have

$$\|Z_t^{-1} - (Z^\star)^{-1}\| \le \frac{1}{\gamma_S^2}\|Z_t - Z^\star\|, \qquad \|Z_t^{-1/2} - (Z^\star)^{-1/2}\| \le \frac{1}{2\gamma_S^{3/2}}\|Z_t - Z^\star\|.$$

• • • The first term in (C.8). Since $\widetilde{Z}_t(S)$ is an orthogonal projection and $Z_t \succeq \gamma_S I$, we have $\|A_t\| \le 1/\gamma_S$ and $\|Z_t^{-1/2}\| \le 1/\sqrt{\gamma_S}$. Therefore, Lemma C.2 implies that

$$\|(Z_t^{-1/2} - (Z^\star)^{-1/2})A_tZ_t^{-1/2}\| \lesssim \|Z_t - Z^\star\|. \tag{C.9}$$

• • • The third term in (C.8). Similarly, $\|A^\star\| \le 1/\gamma_S$ and $\|(Z^\star)^{-1/2}\| \le 1/\sqrt{\gamma_S}$. Therefore,

$$\|(Z^\star)^{-1/2}A^\star(Z_t^{-1/2} - (Z^\star)^{-1/2})\| \lesssim \|Z_t - Z^\star\|. \tag{C.10}$$

• • • **The second term in (C.8).** Using $\|(Z^\star)^{-1/2}\| \le \gamma_S^{-1/2}$ and $\|Z_t^{-1/2}\| \le \gamma_S^{-1/2}$, we have

$$\begin{aligned}
\|(Z^\star)^{-1/2}(A_t - A^\star)Z_t^{-1/2}\| &\le \frac{1}{\gamma_S}\|A_t - A^\star\| \\
&= \frac{1}{\gamma_S}\left\|\mathbb{E}[\widetilde{Z}_t(S)Z_t^{-1}\widetilde{Z}_t(S) \mid \mathcal{F}_{t-1}] - \mathbb{E}[\widetilde{Z}^\star(S)(Z^\star)^{-1}\widetilde{Z}^\star(S)]\right\| \\
&= \frac{1}{\gamma_S}\left\|\mathbb{E}[\widetilde{Z}_t(S)Z_t^{-1}\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)(Z^\star)^{-1}\widetilde{Z}^\star(S) \mid \mathcal{F}_{t-1}]\right\| \\
&= \frac{1}{\gamma_S}\Big\|\mathbb{E}\Big[(\widetilde{Z}_t(S) - \widetilde{Z}^\star(S))Z_t^{-1}\widetilde{Z}_t(S) + \widetilde{Z}^\star(S)\{Z_t^{-1} - (Z^\star)^{-1}\}\widetilde{Z}_t(S) \\
&\qquad + \widetilde{Z}^\star(S)(Z^\star)^{-1}(\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)) \mid \mathcal{F}_{t-1}\Big]\Big\| \\
&\le \frac{1}{\gamma_S}\left\{\frac{1}{\gamma_S}\mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}] + \|Z_t^{-1} - (Z^\star)^{-1}\|\right. \\
&\qquad \left. + \frac{1}{\gamma_S}\mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}]\right\} \\
&\le \frac{1}{\gamma_S^3}\|Z_t - Z^\star\| + \frac{2}{\gamma_S^2}\mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}]. \qquad \text{(C.11)}
\end{aligned}$$

Substituting (C.9), (C.10), and (C.11) into (C.8), we obtain

$$|\nu_t - \nu^\star| \lesssim \|Z_t - Z^\star\| + \mathbb{E}[\|\widetilde{Z}_t(S) - \widetilde{Z}^\star(S)\| \mid \mathcal{F}_{t-1}] \lesssim \|B_t - B^\star\|, \qquad \text{(C.12)}$$

where the last inequality follows from (C.5). Since $B_t \overset{a.s.}{\to} B^\star$, we can conclude that $\nu_t \overset{a.s.}{\to} \nu^\star$.

Combining (C.3), (C.6), and (C.12), we obtain

$$|\alpha_t - \alpha^\star| + |\beta_t - \beta^\star| + |\gamma_t - \gamma^\star| \lesssim \|B_t - B^\star\|. \qquad \text{(C.13)}$$

Together with (C.2), this further implies

$$(\alpha_t, \beta_t, \gamma_t) \overset{a.s.}{\to} (\alpha^\star, \beta^\star, \gamma^\star).$$

• **Convergence of K_t.** Recall the constructions of K_t and $K^\star$ from C_t and $C^\star$ in (B.2). By the conditional independence of the sketching matrices, we have

$$C_t = M_t^\tau, \qquad C^\star = (M^\star)^\tau,$$

where M_t is defined in (B.3), and $M^\star$ is defined analogously by replacing $(\alpha_t, \beta_t, \gamma_t, Z_t)$ with $(\alpha^\star, \beta^\star, \gamma^\star, Z^\star)$. From the block representation of M_t, we have

$$\begin{aligned}
\|M_t - M^\star\| &\le \|(1-\alpha_t)(I - Z_t) - (1-\alpha^\star)(I - Z^\star)\| + \|\alpha_t(I - Z_t) - \alpha^\star(I - Z^\star)\| \\
&\quad + \|(1-\alpha_t)(1-\beta_t)I - (1-\alpha_t)\gamma_t Z_t - (1-\alpha^\star)(1-\beta^\star)I + (1-\alpha^\star)\gamma^\star Z^\star\| \\
&\quad + \|(\alpha_t + \beta_t - \alpha_t\beta_t)I - \alpha_t\gamma_t Z_t - (\alpha^\star + \beta^\star - \alpha^\star\beta^\star)I + \alpha^\star\gamma^\star Z^\star\| \\
&\lesssim |\alpha_t - \alpha^\star| + |\beta_t - \beta^\star| + |\gamma_t - \gamma^\star| + \|Z_t - Z^\star\| \lesssim \|B_t - B^\star\|, \qquad \text{(C.14)}
\end{aligned}$$

where the last inequality follows from (C.5) and (C.13). Furthermore, using $\|I - Z_t\| \le 1$, $\|Z_t\| \le 1$, $0 < \alpha_t < 1$, $0 \le \beta_t < 1$, and $\gamma_t \le 1/\sqrt{\gamma_S}$, we obtain

$$\begin{aligned}
\|M_t\| &\le \|I - Z_t\| + (1-\alpha_t)(1-\beta_t) + (1-\alpha_t)\gamma_t\|Z_t\| \\
&\quad + (\alpha_t + \beta_t - \alpha_t\beta_t) + \alpha_t\gamma_t\|Z_t\| \le 2 + \frac{1}{\sqrt{\gamma_S}}.
\end{aligned}$$

The same bound also holds for $\|M^\star\|$. Therefore, a telescoping expansion gives us

$$\begin{aligned}
\|C_t - C^\star\| &= \|M_t^\tau - (M^\star)^\tau\| \\
&= \left\|\sum_{k=0}^{\tau-1} M_t^{\tau-1-k}(M_t - M^\star)(M^\star)^k\right\| \\
&\le \left(\sum_{k=0}^{\tau-1}\|M_t\|^{\tau-1-k}\|M^\star\|^k\right)\|M_t - M^\star\| \\
&\lesssim \|M_t - M^\star\| \lesssim \|B_t - B^\star\|. \qquad \text{(C.15)}
\end{aligned}$$

Finally, by the definitions of K_t and $K^\star$, there exists a deterministic constant $L_K > 0$ such that

$$\|K_t - K^\star\| = \left\| (I \quad \mathbf{0}) (C_t - C^\star) \begin{pmatrix} I \\ I \end{pmatrix} \right\| \le \sqrt{2}\|C_t - C^\star\| \le L_K \|B_t - B^\star\|. \tag{C.16}$$

Since $B_t \overset{a.s.}{\to} B^\star$, we conclude that

$$K_t \overset{a.s.}{\to} K^\star.$$

• **Symmetry and bound of $K^\star$.** Recall from the proof of Wang et al. (2026, Lemma 3.7) that $U_t^\top K_t U_t = \mathrm{diag}\,(p_{t,\tau}(z_{t,1}), \ldots, p_{t,\tau}(z_{t,d}))$, implying K_t is symmetric. Since $K_t \overset{a.s.}{\to} K^\star$, $K^\star$ is symmetric.

If $\mu_t = 1$, Wang et al. (2026, Lemma 3.7) gives $K_t = \mathbf{0}$. If $\mu_t < 1$, Wang et al. (2026, Lemma 3.7) and the condition on τ give

$$\|K_t\| \le 2\tau \left(1 - \sqrt{\frac{\mu_t}{\nu_t}}\right)^{\tau-2} \le \frac{\gamma_H}{2\Upsilon_H}.$$

Therefore, $\|K_t\| \le \gamma_H/(2\Upsilon_H)$ for every $t \ge 0$. Taking the almost-sure limit yields

$$\|K^\star\| \le \frac{\gamma_H}{2\Upsilon_H} < 1.$$

This completes the proof.

C.2 Proof of Lemma C.2

For the first result, we have

$$\|Z_t^{-1} - (Z^\star)^{-1}\| = \|Z_t^{-1}(Z^\star - Z_t)(Z^\star)^{-1}\| \le \|Z_t^{-1}\|\|Z_t - Z^\star\|\|(Z^\star)^{-1}\| \le \frac{1}{\gamma_S^2}\|Z_t - Z^\star\|,$$

where the last inequality follows from $Z_t \succeq \gamma_S I$ in Assumption 3.4 and $Z^\star \succeq \gamma_S I$, which is due to $Z_t \overset{a.s.}{\to} Z^\star$ established above.

For the second result, we proceed in two steps.

• **Step 1.** We show that, for any positive definite matrix X,

$$X^{-1/2} = \frac{1}{\pi}\int_0^\infty s^{-1/2}(X + sI)^{-1}\, ds. \tag{C.17}$$

Indeed, for any scalar $x > 0$, the following identity holds:

$$x^{-1/2} = \frac{1}{\pi}\int_0^\infty s^{-1/2}(x + s)^{-1}\, ds.$$

To verify this identity, let $s = xu^2$, so that $ds = 2xu\, du$. Then,

$$\begin{aligned} \frac{1}{\pi}\int_0^\infty s^{-1/2}(x+s)^{-1}\, ds &= \frac{1}{\pi}\int_0^\infty (xu^2)^{-1/2}(x + xu^2)^{-1} 2xu\, du \\ &= \frac{1}{\pi}\int_0^\infty \frac{2}{x^{1/2}(1+u^2)}\, du \\ &= \frac{2}{\pi x^{1/2}} \arctan(u)\Big|_0^\infty = x^{-1/2}. \end{aligned} \tag{C.18}$$

We now extend (C.18) to matrices. Since X is positive definite, it admits an eigendecomposition $X = UDU^\top$, where D is diagonal and U is orthogonal. Then, $X^{-1/2} = UD^{-1/2}U^\top$, and

$$\begin{aligned} \frac{1}{\pi}\int_0^\infty s^{-1/2}(X+sI)^{-1}\, ds &= \frac{1}{\pi}\int_0^\infty s^{-1/2}(UDU^\top + sUIU^\top)^{-1}\, ds \\ &= \frac{1}{\pi}\int_0^\infty s^{-1/2}\{U(D+sI)U^\top\}^{-1}\, ds \\ &= \frac{1}{\pi}\int_0^\infty s^{-1/2}U(D+sI)^{-1}U^\top\, ds \\ &= U\left\{\frac{1}{\pi}\int_0^\infty s^{-1/2}(D+sI)^{-1}\, ds\right\}U^\top. \end{aligned} \tag{C.19}$$

Since $(D+sI)^{-1}$ is diagonal, the integral is evaluated entry-wise. By (C.18), its ith diagonal entry equals $\lambda_i^{-1/2}$, where λ_i is the ith eigenvalue of X. Therefore, applying (C.19) gives us

$$\frac{1}{\pi}\int_0^\infty s^{-1/2}(X+sI)^{-1}\,ds = U\left\{\frac{1}{\pi}\int_0^\infty s^{-1/2}(D+sI)^{-1}\,ds\right\}U^\top = UD^{-1/2}U^\top = X^{-1/2}.$$

• **Step 2.** By the representation (C.17) established in **Step 1**, we can write

$$\begin{aligned}
\|Z_t^{-1/2} - (Z^\star)^{-1/2}\| &= \left\|\frac{1}{\pi}\int_0^\infty y^{-1/2}\{(Z_t+yI)^{-1} - (Z^\star+yI)^{-1}\}\,dy\right\| \\
&= \left\|\frac{1}{\pi}\int_0^\infty y^{-1/2}(Z_t+yI)^{-1}(Z^\star - Z_t)(Z^\star+yI)^{-1}\,dy\right\| \\
&\le \frac{\|Z^\star - Z_t\|}{\pi}\int_0^\infty y^{-1/2}\|(Z_t+yI)^{-1}\|\|(Z^\star+yI)^{-1}\|\,dy \\
&\le \frac{\|Z^\star - Z_t\|}{\pi}\int_0^\infty \frac{y^{-1/2}}{(\gamma_S+y)^2}\,dy,
\end{aligned}$$

where the last inequality follows from $Z_t \succeq \gamma_S I$ and $Z^\star \succeq \gamma_S I$. To evaluate the integral, we let $y = \gamma_S u^2$ and hence obtain the following:

$$\int_0^\infty \frac{y^{-1/2}}{(\gamma_S+y)^2}\,dy = \int_0^\infty \frac{(\gamma_S u^2)^{-1/2}}{(\gamma_S+\gamma_S u^2)^2}(2\gamma_S u)\,du = \frac{2}{\gamma_S^{3/2}}\int_0^\infty \frac{1}{(1+u^2)^2}\,du.$$

Letting $u = \tan\theta$ further yields

$$\begin{aligned}
\int_0^\infty \frac{1}{(1+u^2)^2}\,du &= \int_0^{\pi/2}\frac{\sec^2\theta}{(1+\tan^2\theta)^2}\,d\theta = \int_0^{\pi/2}\frac{1}{\sec^2\theta}\,d\theta \\
&= \int_0^{\pi/2}\cos^2\theta\,d\theta = \int_0^{\pi/2}\frac{1+\cos(2\theta)}{2}\,d\theta = \frac{\pi}{4}.
\end{aligned}$$

Combining the preceding displays, we conclude that

$$\|Z_t^{-1/2} - (Z^\star)^{-1/2}\| \le \frac{\|Z^\star - Z_t\|}{\pi}\int_0^\infty \frac{y^{-1/2}}{(\gamma_S+y)^2}\,dy = \frac{1}{2\gamma_S^{3/2}}\|Z^\star - Z_t\|.$$

This completes the proof.

C.3 Proof of Lemma 4.2

Recall that $\boldsymbol{\Delta}_t = \boldsymbol{x}_t - \boldsymbol{x}^\star$, $\mathcal{R} = I - K^\star$ and $\lambda_m = \lambda_{\min}(\mathcal{R})$. By Lemma 4.1, $K^\star$ is symmetric and $\|K^\star\| < 1$. Hence, $\mathcal{R}$ is symmetric positive definite and $\lambda_m > 0$.

Also, define the realized and conditional-mean inner-solver operators as

$$\widetilde{Q}_t := (I - \widetilde{K}_t)B_t^{-1}, \qquad Q_t := (I - K_t)B_t^{-1}.$$

Conditional on $\mathcal{F}_{t-0.5}$, B_t is fixed. Therefore, by Lemma B.1, we obtain

$$\begin{aligned}
\mathbb{E}[\widetilde{Q}_t \mid \mathcal{F}_{t-0.5}] &= \mathbb{E}[(I - \widetilde{K}_t)B_t^{-1} \mid \mathcal{F}_{t-0.5}] \\
&= \left\{I - \mathbb{E}[\widetilde{K}_t \mid \mathcal{F}_{t-0.5}]\right\}B_t^{-1} = (I - K_t)B_t^{-1} = Q_t. \quad \text{(C.20)}
\end{aligned}$$

We next establish uniform bounds for these two operators. By (B.8), we have $\|\widetilde{K}_t\| \le C_K$. Moreover, Wang et al. (2026, Lemma 3.7) and the condition on τ ensure that $\|K_t\| \le 2\tau\left(1 - \sqrt{\mu_t/\nu_t}\right)^{\tau-2} \le \gamma_H/(2\Upsilon_H)$. Since Assumption 3.1 and the construction of B_t ensure that $B_t \succeq \gamma_H I$, it follows that

$$\begin{aligned}
\|\widetilde{Q}_t\| &= \|(I - \widetilde{K}_t)B_t^{-1}\| \le \big(1 + \|\widetilde{K}_t\|\big)\|B_t^{-1}\| \le \frac{1 + C_K}{\gamma_H} \lesssim 1, \\
\|Q_t\| &= \|(I - K_t)B_t^{-1}\| \le \big(1 + \|K_t\|\big)\|B_t^{-1}\| \le \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) \lesssim 1. \quad \text{(C.21)}
\end{aligned}$$

We further bound the deviation of the conditional-mean operator from its limit as follows:

$$\begin{aligned}\|Q_t B^\star - \mathcal{R}\| &= \left\|(I-K_t)B_t^{-1}B^\star - (I-K^\star)\right\| \\ &= \left\|(I-K_t)B_t^{-1}(B^\star - B_t) + (K^\star - K_t)\right\| \\ &\le \|I-K_t\|\|B_t^{-1}\|\|B^\star - B_t\| + \|K_t - K^\star\| \\ &\le \left\{\frac{1+\|K_t\|}{\gamma_H} + L_K\right\}\|B_t - B^\star\| \\ &\le \left\{\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right) + L_K\right\}\|B_t - B^\star\|. \end{aligned} \tag{C.22}$$

Here, the first equality follows from the definitions of Q_t and $\mathcal{R}$; the fourth inequality uses $\|B_t^{-1}\| \le 1/\gamma_H$ and (C.16); and the second-to-last inequality follows from $\|K_t\| \le \gamma_H/(2\Upsilon_H)$.

Also, by the Υ_L-Lipschitz continuity of $\nabla^2 f$, we obtain

$$\begin{aligned}\|\nabla f_t - B^\star \boldsymbol{\Delta}_t\| &= \left\|\int_0^1 \left\{\nabla^2 f(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t) - B^\star\right\}\boldsymbol{\Delta}_t\, ds\right\| \\ &\le \int_0^1 \left\|\nabla^2 f(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t) - \nabla^2 f(\boldsymbol{x}^\star)\right\|\|\boldsymbol{\Delta}_t\|\, ds \\ &\le \Upsilon_L \int_0^1 s\|\boldsymbol{\Delta}_t\|^2\, ds = \frac{\Upsilon_L}{2}\|\boldsymbol{\Delta}_t\|^2. \end{aligned} \tag{C.23}$$

Combining (C.21), (C.22), and (C.23), we further obtain

$$\begin{aligned}&\|Q_t\nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\| = \|Q_t(\nabla f_t - B^\star\boldsymbol{\Delta}_t) + (Q_t B^\star - \mathcal{R})\boldsymbol{\Delta}_t\| \\ &\le \|Q_t\|\|\nabla f_t - B^\star\boldsymbol{\Delta}_t\| + \|Q_t B^\star - \mathcal{R}\|\|\boldsymbol{\Delta}_t\| \\ &\le \frac{\Upsilon_L}{2\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)\|\boldsymbol{\Delta}_t\|^2 + \left\{\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right) + L_K\right\}\|B_t - B^\star\|\|\boldsymbol{\Delta}_t\| \\ &\lesssim \left\{\|\boldsymbol{\Delta}_t\| + \|B_t - B^\star\|\right\}\|\boldsymbol{\Delta}_t\|. \end{aligned} \tag{C.24}$$

Now we localize the analysis by introducing the stopping time: for any given $k \ge 0$, $r > 0$,

$$\tau_{k,r} := \inf\left\{t \ge k : \|\boldsymbol{\Delta}_t\| > r \ \text{ OR } \ \|B_t - B^\star\| > r \ \text{ OR } \ \|\boldsymbol{e}_{t-1}\| > r\right\}. \tag{C.25}$$

Since $(\boldsymbol{\Delta}_t, B_t, \boldsymbol{e}_{t-1})$ is $\mathcal{F}_{t-1}$-measurable, $\tau_{k,r}$ is a stopping time and $\{\tau_{k,r} > t\} \in \mathcal{F}_{t-1}$. Moreover, Theorem 3.5 and Lemma 4.1 together ensure

$$\boldsymbol{\Delta}_t \overset{a.s.}{\to} \mathbf{0}, \qquad B_t \overset{a.s.}{\to} B^\star, \qquad \boldsymbol{e}_t \overset{a.s.}{\to} \mathbf{0},$$

Consequently, for every fixed $r > 0$, we can conclude that

$$\mathbb{P}(\tau_{k,r} < \infty) \to 0 \qquad \text{as} \qquad k \to \infty. \tag{C.26}$$

This suggests that for any given $\epsilon, r > 0$, we can choose $k = k(\epsilon, r)$, which depends on the values of ϵ, r, such that $\mathbb{P}(\tau_{k,r} < \infty) \le 0.5\epsilon$. Furthermore, for any $t \ge k$, we have

$$\mathbb{P}(\tau_{k,r} \le t) \le \mathbb{P}(\tau_{k,r} < \infty) < \frac{\epsilon}{2}. \tag{C.27}$$

Next we determine the radius r and index k step by step. We first choose a margin $\varepsilon_0 > 0$. If $\varphi < 1$, we could choose any $\varepsilon_0 \in (0,1)$. If $\varphi = 1$, we choose $\varepsilon_0 \in (0,1)$ sufficiently small such that

$$2C_\varphi(1-\varepsilon_0)\lambda_m > \eta, \tag{C.28}$$

which is possible since $C_\varphi > \eta/(2\lambda_m)$. The precise choice of the localization radius r will be specified after deriving the preliminary recursion for the iterate error.

By the outer update, Lemma B.1, and the definition of $\widetilde{Q}_t$, we have

$$\boldsymbol{\Delta}_{t+1} = \boldsymbol{\Delta}_t - \varphi_t \widetilde{Q}_t(\nabla f_t + \boldsymbol{e}_t).$$

Conditional on $\mathcal{F}_{t-0.5}$, the quantities $(\boldsymbol{\Delta}_t, \nabla f_t, \boldsymbol{e}_t, Q_t)$ are fixed. Therefore, we obtain

$$\mathbb{E}[\|\boldsymbol{\Delta}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] = \|\boldsymbol{\Delta}_t - \varphi_t Q_t(\nabla f_t + \boldsymbol{e}_t)\|^2 + \varphi_t^2 \mathbb{E}[\|(\widetilde{Q}_t - Q_t)(\nabla f_t + \boldsymbol{e}_t)\|^2 \mid \mathcal{F}_{t-0.5}]. \quad \text{(C.29)}$$

Here, the cross term vanishes since $\mathbb{E}[\widetilde{Q}_t - Q_t \mid \mathcal{F}_{t-0.5}] = \mathbf{0}$ given in (C.20). Moreover,

$$\boldsymbol{\Delta}_t - \varphi_t Q_t(\nabla f_t + \boldsymbol{e}_t) = (I - \varphi_t \mathcal{R})\boldsymbol{\Delta}_t - \varphi_t(Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t) - \varphi_t Q_t \boldsymbol{e}_t.$$

Since $\mathcal{R}$ is symmetric positive definite, for all sufficiently large t such that $\|I - \varphi_t \mathcal{R}\| \le 1$, we have

$$\begin{aligned}
&\|\boldsymbol{\Delta}_t - \varphi_t Q_t(\nabla f_t + \boldsymbol{e}_t)\|^2 = \|(I - \varphi_t \mathcal{R})\boldsymbol{\Delta}_t - \varphi_t(Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t) - \varphi_t Q_t \boldsymbol{e}_t\|^2 \\
&\le \|(I - \varphi_t \mathcal{R})\boldsymbol{\Delta}_t\|^2 + 2\varphi_t \|\boldsymbol{\Delta}_t\| \|Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\| \\
&\quad + \frac{2\varphi_t}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\|\boldsymbol{\Delta}_t\|\|\boldsymbol{e}_t\| + 2\varphi_t^2 \|Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\|^2 + \frac{2\varphi_t^2}{\gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 \|\boldsymbol{e}_t\|^2 \\
&\le \left(1 - 2\lambda_m \varphi_t + \|\mathcal{R}\|^2 \varphi_t^2\right)\|\boldsymbol{\Delta}_t\|^2 + 2\varphi_t \|\boldsymbol{\Delta}_t\| \|Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\| \\
&\quad + \frac{2\varphi_t}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\|\boldsymbol{\Delta}_t\|\|\boldsymbol{e}_t\| + 2\varphi_t^2 \|Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\|^2 + \frac{2\varphi_t^2}{\gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 \|\boldsymbol{e}_t\|^2,
\end{aligned} \quad \text{(C.30)}$$

where the second inequality follows from the Cauchy–Schwarz inequality, $\|\boldsymbol{a} + \boldsymbol{b}\|^2 \le 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$, and the explicit uniform bound $\|Q_t\| \le 1/\gamma_H\,(1 + \gamma_H/(2\Upsilon_H))$ derived in (C.21), and the last inequality holds due to the fact that

$$\|(I - \varphi_t \mathcal{R})\boldsymbol{a}\|^2 \le \left(1 - 2\lambda_m \varphi_t + \|\mathcal{R}\|^2 \varphi_t^2\right)\|\boldsymbol{a}\|^2.$$

On $\{\tau_{k,r} > t\}$, by the definition of $\tau_{k,r}$ and (C.24), we have

$$\|Q_t \nabla f_t - \mathcal{R}\boldsymbol{\Delta}_t\| \le \left\{\frac{\Upsilon_L}{2\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + L_K\right\} r\|\boldsymbol{\Delta}_t\|.$$

Moreover, Young's inequality directly implies the following:

$$\frac{2\varphi_t}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\|\boldsymbol{\Delta}_t\|\|\boldsymbol{e}_t\| \le \varepsilon_0 \lambda_m \varphi_t \|\boldsymbol{\Delta}_t\|^2 + \frac{\varphi_t}{\varepsilon_0 \lambda_m \gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 \|\boldsymbol{e}_t\|^2.$$

On $\{\tau_{k,r} > t\}$, substituting these two bounds into (C.30) yields

$$\begin{aligned}
&\|\boldsymbol{\Delta}_t - \varphi_t Q_t(\nabla f_t + \boldsymbol{e}_t)\|^2 \\
&\le \left[1 - (2 - \varepsilon_0)\lambda_m \varphi_t + 2\left\{\frac{\Upsilon_L}{2\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + L_K\right\} r\varphi_t \right. \\
&\quad \left. + \left(\|\mathcal{R}\|^2 + 2\left\{\frac{\Upsilon_L}{2\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right) + L_K\right\}^2 r^2\right)\varphi_t^2\right]\|\boldsymbol{\Delta}_t\|^2 \\
&\quad + \frac{1}{\gamma_H^2}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)^2 \left(\frac{\varphi_t}{\varepsilon_0 \lambda_m} + 2\varphi_t^2\right)\|\boldsymbol{e}_t\|^2.
\end{aligned} \quad \text{(C.31)}$$

We next control the second term in (C.29). Since $(\nabla f_t, \boldsymbol{e}_t)$ is $\mathcal{F}_{t-0.5}$-measurable, (C.21) gives

$$\begin{aligned}
&\mathbb{E}[\|(\widetilde{Q}_t - Q_t)(\nabla f_t + \boldsymbol{e}_t)\|^2 \mid \mathcal{F}_{t-0.5}] \le \mathbb{E}[\|\widetilde{Q}_t - Q_t\|^2 \mid \mathcal{F}_{t-0.5}]\|\nabla f_t + \boldsymbol{e}_t\|^2 \\
&\le 2\left\{\frac{1 + C_K}{\gamma_H} + \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\right\}^2 \left(\|\nabla f_t\|^2 + \|\boldsymbol{e}_t\|^2\right) \\
&\le 2\left\{\frac{1 + C_K}{\gamma_H} + \frac{1}{\gamma_H}\left(1 + \frac{\gamma_H}{2\Upsilon_H}\right)\right\}^2 \left(\Upsilon_H^2 \|\boldsymbol{\Delta}_t\|^2 + \|\boldsymbol{e}_t\|^2\right).
\end{aligned} \quad \text{(C.32)}$$

Combining (C.29), (C.31) and (C.32), we obtain

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\Delta}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}]\mathbf{1}_{\{\tau_{k,r}>t\}} \\
&\le \left[1-(2-\varepsilon_0)\lambda_m\varphi_t + 2\left\{\frac{\Upsilon_L}{2\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+L_K\right\} r\varphi_t\right. \\
&\quad + \left(\|\mathcal{R}\|^2 + 2\left\{\frac{\Upsilon_L}{2\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+L_K\right\}^2 r^2\right. \\
&\quad\quad \left.\left. + 2\Upsilon_H^2\left\{\frac{1+C_K}{\gamma_H}+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)\right\}^2\right)\varphi_t^2\right]\|\boldsymbol{\Delta}_t\|^2\mathbf{1}_{\{\tau_{k,r}>t\}} \\
&\quad + \left[\frac{1}{\varepsilon_0\lambda_m\gamma_H^2}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)^2\varphi_t\right. \\
&\quad\quad \left. + 2\left\{\frac{1}{\gamma_H^2}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)^2 + \left\{\frac{1+C_K}{\gamma_H}+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)\right\}^2\right\}\varphi_t^2\right]\|\boldsymbol{e}_t\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}.
\end{aligned} \tag{C.33}$$

For notational conciseness, define the following deterministic constants:

$$\begin{aligned}
L_\Delta &:= \frac{\Upsilon_L}{2\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)+L_K, \\
L_S &:= \frac{1+C_K}{\gamma_H}+\frac{1}{\gamma_H}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right), \qquad D_\Delta(r) := \|\mathcal{R}\|^2 + 2L_\Delta^2 r^2 + 2\Upsilon_H^2 L_S^2, \\
D_e &:= \frac{1}{\varepsilon_0\lambda_m\gamma_H^2}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)^2 + 2\left\{\frac{1}{\gamma_H^2}\left(1+\frac{\gamma_H}{2\Upsilon_H}\right)^2 + L_S^2\right\}.
\end{aligned}$$

Therefore, for all sufficiently large t such that $\varphi_t \le 1$, (C.33) implies

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\Delta}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}]\mathbf{1}_{\{\tau_{k,r}>t\}} \\
&\quad \le \left\{\left[1-(2-\varepsilon_0)\lambda_m\varphi_t + 2L_\Delta r\varphi_t + D_\Delta(r)\varphi_t^2\right]\|\boldsymbol{\Delta}_t\|^2 + D_e\varphi_t\|\boldsymbol{e}_t\|^2\right\}\mathbf{1}_{\{\tau_{k,r}>t\}}.
\end{aligned} \tag{C.34}$$

We now choose $r \in (0,1]$ sufficiently small such that

$$2L_\Delta r \le \frac{\varepsilon_0\lambda_m}{2}. \tag{C.35}$$

For this fixed r and any given $\epsilon > 0$, we first choose $k = k(\epsilon, r)$ sufficiently large so that $\mathbb{P}(\tau_{k,r} < \infty) < \epsilon/2$, which guarantees (C.27) for every $t \ge k$. Additionally, for every $t \ge k$, we require

$$\varphi_t \le 1, \qquad \varphi_t\|\mathcal{R}\| \le 1, \qquad D_\Delta(r)\varphi_t \le \frac{\varepsilon_0\lambda_m}{2}, \qquad 0 < 2(1-\varepsilon_0)\lambda_m\varphi_t < 1. \tag{C.36}$$

Specifically, the third condition allows the second-order term $D_\Delta(r)\varphi_t^2$ in (C.34) to be absorbed into the first-order contraction margin, thereby preserving the desired contraction coefficient.

For these fixed (k, r), we define the following localized second moments

$$u_t := \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}], \quad v_t := \mathbb{E}[\|\boldsymbol{e}_t\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}], \quad w_t := \mathbb{E}[\|B_t - B^\star\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}]. \tag{C.37}$$

• **Convergence rate of B_t.** For any $M > 0$ and $t \ge k$, by Markov's inequality and (C.27), we have

$$\begin{aligned}
&\mathbb{P}(\|B_t - B^\star\| > M\sqrt{\eta_t}) = \mathbb{P}(\|B_t - B^\star\|^2 > M^2\eta_t) \\
&= \mathbb{P}\big(\{\|B_t - B^\star\|^2 > M^2\varphi_t\} \cap \{\tau_{k,r} > t\}\big) + \mathbb{P}\big(\{\|B_t - B^\star\|^2 > M^2\varphi_t\} \cap \{\tau_{k,r} \le t\}\big) \\
&\le \mathbb{P}(\tau_{k,r} \le t) + \mathbb{P}(\|B_t - B^\star\|^2\mathbf{1}_{\{\tau_{k,r}>t\}} > M^2\eta_t) \le \frac{\epsilon}{2} + \frac{w_t}{M^2\eta_t}.
\end{aligned} \tag{C.38}$$

Therefore, in order to obtain $O_p(\sqrt{\eta_t})$, it suffices to establish $w_t = O(\eta_t)$. Recall that $k = k(\epsilon, r)$ is fixed. Hence, the finitely many indices preceding k contribute $O(t^{-2})$ below, with a hidden constant

that may depend on the fixed k but not on t. By the construction of B_t, the Υ_L-Lipschitz continuity of $\nabla^2 f$, and the event inclusion $\{\tau_{k,r} > t\} \subseteq \{\tau_{k,r} > i\}$ for any $i \le t$, we obtain

$$\begin{aligned} w_t &= \mathbb{E}\left[\left\|\frac{1}{t}\sum_{i=0}^{t-1}(H_i - \nabla^2 f_i) + \frac{1}{t}\sum_{i=0}^{t-1}(\nabla^2 f_i - B^\star)\right\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}\right] \\ &\lesssim \frac{1}{t^2} + \frac{1}{t^2}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}(H_i - \nabla^2 f_i)\mathbf{1}_{\{\tau_{k,r}>i\}}\right\|_{\mathrm{F}}^2\right] + \frac{\Upsilon_L^2}{t^2}\mathbb{E}\left[\left(\sum_{i=k}^{t-1}\|\boldsymbol{\Delta}_i\|\mathbf{1}_{\{\tau_{k,r}>i\}}\right)^2\right] \\ &= \frac{1}{t^2} + \frac{1}{t^2}\sum_{i=k}^{t-1}\mathbb{E}[\|H_i - \nabla^2 f_i\|_{\mathrm{F}}^2 \mathbf{1}_{\{\tau_{k,r}>i\}}] + \frac{\Upsilon_L^2}{t^2}\mathbb{E}\left[\left(\sum_{i=k}^{t-1}\|\boldsymbol{\Delta}_i\|\mathbf{1}_{\{\tau_{k,r}>i\}}\right)^2\right] \\ &\lesssim \frac{1}{t} + \frac{1}{t^2}\left(\sum_{i=k}^{t-1}\left\{\mathbb{E}[\|\boldsymbol{\Delta}_i\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}^{1/2}\right)^2 = \frac{1}{t} + \frac{1}{t^2}\left(\sum_{i=k}^{t-1}\sqrt{u_i}\right)^2. \end{aligned} \tag{C.39}$$

Here, $(H_i - \nabla^2 f_i)\mathbf{1}_{\{\tau_{k,r}>i\}}$ is a matrix-valued martingale difference sequence because $\mathbf{1}_{\{\tau_{k,r}>i\}}$ is $\mathcal{F}_{i-1}$-measurable. Therefore, the third equality follows from martingale orthogonality under the Frobenius inner product. The second-to-last inequality follows from Assumption 3.3 and Minkowski's inequality. Then, it remains to establish the rates of u_t and v_t.

To complete the proof, we present the following two auxiliary lemmas. And their proofs will be deferred to Appendices C.4 and C.5, respectively.

Lemma C.3. For (k, r) chosen above and all sufficiently large t, we have

$$u_{t+1} \le \{1 - 2(1-\varepsilon_0)\lambda_m\varphi_t\}\, u_t + D_e\varphi_t v_t. \tag{C.40}$$

and

$$\begin{aligned} v_t \lesssim \left\{\prod_{j=k+1}^{t}(1-\eta_j)\right\}^2 \mathbb{E}[\|\boldsymbol{e}_k\|^2] + \sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)^2\right\}\eta_i^2 \\ + \left\{\sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)\right\}\varphi_{i-1}\left(\sqrt{u_{i-1}} + \sqrt{v_{i-1}}\right)\right\}^2. \end{aligned} \tag{C.41}$$

Note that $\prod_{j=m}^{n}(1-\eta_j) \coloneqq 1$ when $m > n$.

Lemma C.4. Under the conditions of Lemma 4.2, for every integer $q \ge 0$, there exists a deterministic constant $C_q(k, r) > 0$ such that

$$\max\{u_t, v_t\} \le C_q(k,r)\left\{\eta_t + \left(\frac{\varphi_t}{\eta_t}\right)^{2q}\right\} \tag{C.42}$$

for all sufficiently large t.

Since $\varphi > \eta$, we can choose the finite integer $q_\star \coloneqq \lfloor \eta/[2(\varphi-\eta)] \rfloor + 1$, which satisfies $2q_\star(\varphi-\eta) > \eta$. Consequently, $(\varphi_t/\eta_t)^{2q_\star} = o(\eta_t)$. Applying Lemma C.4 with $q = q_\star$ gives

$$u_t = O(\eta_t), \qquad v_t = O(\eta_t). \tag{C.43}$$

Returning to (C.39) and using (C.43), we obtain

$$w_t \lesssim \frac{1}{t} + \frac{1}{t^2}\left(\sum_{i=k}^{t-1}\sqrt{\eta_i}\right)^2 \lesssim \frac{1}{t} + \eta_t \lesssim \eta_t, \tag{C.44}$$

Therefore, for all sufficiently large t, there exists $C_{B,k,r} > 0$ such that

$$\mathbb{P}(\|B_t - B^\star\| > M\sqrt{\eta_t}) \le \frac{\epsilon}{2} + \frac{C_{B,k,r\varphi_t}}{M^2\varphi_t} = \frac{\epsilon}{2} + \frac{C_{B,k,r}}{M^2}.$$

Choosing $M \geq \sqrt{2C_{B,k,r}/\epsilon}$ ensures $\mathbb{P}(\|B_t - B^\star\| > M\sqrt{\eta_t}) < \epsilon$ for t large enough. Therefore,

$$\|B_t - B^\star\| = O_p(\sqrt{\eta_t}). \tag{C.45}$$

• **Convergence rates of $\boldsymbol{m}_t$ and $\boldsymbol{\Delta}_t$.** By (C.43) and the same stopping-time decomposition as above, for the fixed (k, r) and t large enough, there exist constants $C_{\Delta,k,r}, C_{e,k,r} > 0$ such that

$$\mathbb{P}(\|\boldsymbol{\Delta}_t\| > M\sqrt{\eta_t}) \leq \frac{\epsilon}{2} + \frac{C_{\Delta,k,r}}{M^2}, \qquad \mathbb{P}(\|\boldsymbol{e}_t\| > M\sqrt{\eta_t}) \leq \frac{\epsilon}{2} + \frac{C_{e,k,r}}{M^2}.$$

Again, choosing $M \geq \sqrt{2\max\{C_{\Delta,k,r}, C_{e,k,r}\}/\epsilon}$ ensures that

$$\|\boldsymbol{\Delta}_t\| = \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| = O_p(\sqrt{\eta_t}), \qquad \|\boldsymbol{e}_t\| = O_p(\sqrt{\eta_t}). \tag{C.46}$$

Since $\boldsymbol{m}_t = \boldsymbol{e}_t + \nabla f_t$ and $\|\nabla f_t\| \leq \Upsilon_H \|\boldsymbol{\Delta}_t\|$, (C.46) further yields

$$\|\boldsymbol{m}_t\| \leq \|\boldsymbol{e}_t\| + \Upsilon_H \|\boldsymbol{\Delta}_t\| = O_p(\sqrt{\eta_t}).$$

• **Convergence rates of $(\alpha_t, \beta_t, \gamma_t)$.** Combining (C.13) and (C.45), we obtain

$$\max\{|\alpha_t - \alpha^\star|,\ |\beta_t - \beta^\star|,\ |\gamma_t - \gamma^\star|\} = O_p(\sqrt{\eta_t}).$$

• **Convergence rate of K_t.** Combining (C.16) and (C.45) gives

$$\|K_t - K^\star\| \leq L_K \|B_t - B^\star\| = O_p(\sqrt{\eta_t}).$$

Together with (C.45) and (C.46), these bounds yield

$$\max\big\{\|\boldsymbol{m}_t - \nabla f_t\|,\ \|\boldsymbol{m}_t\|,\ \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|,\ \|B_t - B^\star\|,\ |\alpha_t - \alpha^\star|,\ |\beta_t - \beta^\star|,\ |\gamma_t - \gamma^\star|,\ \|K_t - K^\star\|\big\} = O_p(\sqrt{\eta_t}),$$

which completes the proof of Lemma 4.2.

C.4 Proof of Lemma C.3

Since $\mathbf{1}_{\{\tau_{k,r} > t+1\}} \leq \mathbf{1}_{\{\tau_{k,r} > t\}}$ and $\mathbf{1}_{\{\tau_{k,r} > t\}}$ is $\mathcal{F}_{t-0.5}$-measurable, (C.34)–(C.36) yield

$$\begin{aligned} u_{t+1} = \mathbb{E}[\|\boldsymbol{\Delta}_{t+1}\|^2 \mathbf{1}_{\{\tau_{k,r} > t+1\}}] &\leq \mathbb{E}\left[\mathbb{E}[\|\boldsymbol{\Delta}_{t+1}\|^2 \mid \mathcal{F}_{t-0.5}] \mathbf{1}_{\{\tau_{k,r} > t\}}\right] \\ &\leq \{1 - 2(1-\varepsilon_0)\lambda_m \varphi_t\}\, u_t + D_e \varphi_t v_t, \qquad t \geq k. \end{aligned} \tag{C.47}$$

This proves the first result. It remains to establish (C.41). By the moving-average update,

$$\boldsymbol{e}_t = (1-\eta_t)\boldsymbol{e}_{t-1} + \eta_t(\boldsymbol{g}_t - \nabla f_t) + (1-\eta_t)(\nabla f_{t-1} - \nabla f_t).$$

Iterating this identity from the fixed index k gives

$$\begin{aligned} \boldsymbol{e}_t = \left\{\prod_{j=k+1}^{t}(1-\eta_j)\right\}\boldsymbol{e}_k + \sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)\right\}\eta_i(\boldsymbol{g}_i - \nabla f_i) \\ + \sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)\right\}(1-\eta_i)(\nabla f_{i-1} - \nabla f_i). \end{aligned} \tag{C.48}$$

Since $\mathbf{1}_{\{\tau_{k,r} > i\}}$ is $\mathcal{F}_{i-1}$-measurable and $\mathbb{E}[\boldsymbol{g}_i - \nabla f_i \mid \mathcal{F}_{i-1}] = \mathbf{0}$, $\big\{(\boldsymbol{g}_i - \nabla f_i)\mathbf{1}_{\{\tau_{k,r} > i\}}\big\}_{i \geq k+1}$ forms a martingale difference sequence. Moreover, $\mathbf{1}_{\{\tau_{k,r} > t\}} \leq \mathbf{1}_{\{\tau_{k,r} > i\}}$ for every $i \leq t$. As a result, by Jensen's inequality and Assumption 3.2, we obtain

$$\begin{aligned} &\mathbb{E}\left[\left\|\sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)\right\}\eta_i(\boldsymbol{g}_i - \nabla f_i)\right\|^2 \mathbf{1}_{\{\tau_{k,r} > t\}}\right] \\ &\leq \mathbb{E}\left[\left\|\sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)\right\}\eta_i(\boldsymbol{g}_i - \nabla f_i)\mathbf{1}_{\{\tau_{k,r} > i\}}\right\|^2\right] \\ &= \sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)^2\right\}\eta_i^2 \mathbb{E}\left[\|\boldsymbol{g}_i - \nabla f_i\|^2 \mathbf{1}_{\{\tau_{k,r} > i\}}\right] \lesssim \sum_{i=k+1}^{t}\left\{\prod_{j=i+1}^{t}(1-\eta_j)^2\right\}\eta_i^2. \end{aligned} \tag{C.49}$$

For the last term in (C.48), by the Υ_H-Lipschitz continuity of ∇f and Lemma B.1, we have

$$\|\nabla f_i - \nabla f_{i-1}\| \le \Upsilon_H \varphi_{i-1} \|\boldsymbol{z}_{i-1,\tau}\| = \Upsilon_H \varphi_{i-1} \|\widetilde{Q}_{i-1}(\nabla f_{i-1} + \boldsymbol{e}_{i-1})\|.$$

Then, by (C.21) and $\mathbf{1}_{\{\tau_{k,r}>i\}} \le \mathbf{1}_{\{\tau_{k,r}>i-1\}}$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\nabla f_i - \nabla f_{i-1}\|^2 \mathbf{1}_{\{\tau_{k,r}>i\}}] \le \Upsilon_H^2 \varphi_{i-1}^2 \mathbb{E}[\|\boldsymbol{z}_{i-1,\tau}\|^2 \mathbf{1}_{\{\tau_{k,r}>i-1\}}] \\
&= \Upsilon_H^2 \varphi_{i-1}^2 \mathbb{E}[\|\widetilde{Q}_{i-1}(\nabla f_{i-1} + \boldsymbol{e}_{i-1})\|^2 \mathbf{1}_{\{\tau_{k,r}>i-1\}}] \\
&\le \frac{\Upsilon_H^2 (1+C_K)^2}{\gamma_H^2} \varphi_{i-1}^2 \mathbb{E}[\|\nabla f_{i-1} + \boldsymbol{e}_{i-1}\|^2 \mathbf{1}_{\{\tau_{k,r}>i-1\}}] \\
&\le \frac{2\Upsilon_H^2 (1+C_K)^2}{\gamma_H^2} \varphi_{i-1}^2 \left\{\mathbb{E}[\|\nabla f_{i-1}\|^2 \mathbf{1}_{\{\tau_{k,r}>i-1\}}] + v_{i-1}\right\} \\
&\le \frac{2\Upsilon_H^2 (1+C_K)^2}{\gamma_H^2} \varphi_{i-1}^2 \left(\Upsilon_H^2 u_{i-1} + v_{i-1}\right) \lesssim \varphi_{i-1}^2 (u_{i-1} + v_{i-1}),
\end{aligned} \tag{C.50}$$

where the third inequality follows from $\|\widetilde{Q}_{i-1}\| \le (1+C_K)/\gamma_H$ in (C.21), and the last inequality uses $\|\nabla f_{i-1}\| \le \Upsilon_H \|\boldsymbol{\Delta}_{i-1}\|$ together with the definitions of u_{i-1} and v_{i-1}.

Applying Minkowski's inequality to (C.48), we obtain

$$\begin{aligned}
\sqrt{v_t} &= \left\{\mathbb{E}[\|\boldsymbol{e}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}]\right\}^{1/2} \\
&\le \left\{\prod_{j=k+1}^{t} (1-\eta_j)\right\} \{\mathbb{E}[\|\boldsymbol{e}_k\|^2]\}^{1/2} \\
&\quad + \left\{\mathbb{E}\left[\left\|\sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} \eta_i (\boldsymbol{g}_i - \nabla f_i)\right\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}\right]\right\}^{1/2} \\
&\quad + \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} (1-\eta_i) \left\{\mathbb{E}[\|\nabla f_{i-1} - \nabla f_i\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}]\right\}^{1/2} \\
&\le \left\{\prod_{j=k+1}^{t} (1-\eta_j)\right\} \{\mathbb{E}[\|\boldsymbol{e}_k\|^2]\}^{1/2} + \left\{\sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)^2\right\} \eta_i^2\right\}^{1/2} \\
&\quad + \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} (1-\eta_i) \left\{\mathbb{E}[\|\nabla f_{i-1} - \nabla f_i\|^2 \mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}^{1/2} \\
&\lesssim \left\{\prod_{j=k+1}^{t} (1-\eta_j)\right\} \{\mathbb{E}[\|\boldsymbol{e}_k\|^2]\}^{1/2} + \left\{\sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)^2\right\} \eta_i^2\right\}^{1/2} \\
&\quad + \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} \varphi_{i-1}\left(\sqrt{u_{i-1}} + \sqrt{v_{i-1}}\right),
\end{aligned}$$

where the second inequality follows from (C.49) and $\mathbf{1}_{\{\tau_{k,r}>t\}} \le \mathbf{1}_{\{\tau_{k,r}>i\}}$, and the last inequality follows from (C.50). Squaring both sides and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ yields

$$\begin{aligned}
v_t \lesssim \left\{\prod_{j=k+1}^{t} (1-\eta_j)^2\right\} \mathbb{E}[\|\boldsymbol{e}_k\|^2] + \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)^2\right\} \eta_i^2 \\
+ \left\{\sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} \varphi_{i-1}\left(\sqrt{u_{i-1}} + \sqrt{v_{i-1}}\right)\right\}^2,
\end{aligned}$$

which proves (C.41) and completes the proof.

C.5 Proof of Lemma C.4

We proceed by induction on q. For $q = 0$, the right-hand side of (C.42) is of constant order. By the definition of the stopping time, $u_t = \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r} > t\}}] \le r^2$. Moreover, on $\{\tau_{k,r} > t\}$,

$$\|\boldsymbol{e}_{t-1}\| \le r, \qquad \|\boldsymbol{\Delta}_{t-1}\| \le r, \qquad \|\boldsymbol{\Delta}_t\| \le r.$$

Using the recursion for $\boldsymbol{e}_t$, Assumption 3.2 and the uniform bound (C.21), we obtain

$$\begin{aligned} v_t &= \mathbb{E}[\|\boldsymbol{e}_t\|^2 \mathbf{1}_{\{\tau_{k,r} > t\}}] \\ &\lesssim \mathbb{E}\left[\left\{\|\boldsymbol{e}_{t-1}\|^2 + \eta_t^2 \|\boldsymbol{g}_t - \nabla f_t\|^2 + \|\nabla f_{t-1} - \nabla f_t\|^2\right\} \mathbf{1}_{\{\tau_{k,r} > t\}}\right] \\ &\lesssim r^2 + \eta_t^2 + \varphi_{t-1}^2 r^2 \lesssim 1. \end{aligned}$$

Therefore, (C.42) holds for $q = 0$.

Suppose that (C.42) holds for some integer $q \ge 0$. Then, we have

$$\sqrt{u_i} + \sqrt{v_i} \lesssim \sqrt{\eta_i} + \left(\frac{\varphi_i}{\eta_i}\right)^q. \tag{C.51}$$

Since $\eta < \varphi \le 1$, by Lemma A.3, we obtain

$$\begin{aligned} &\left\{\prod_{j=k+1}^{t} (1-\eta_j)\right\}^2 = o(\eta_t), \qquad \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} \varphi_{i-1}\sqrt{\eta_{i-1}} \lesssim \frac{\varphi_t}{\eta_t}\sqrt{\eta_t}, \\ &\sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)^2\right\} \eta_i^2 \lesssim \eta_t, \qquad \sum_{i=k+1}^{t} \left\{\prod_{j=i+1}^{t} (1-\eta_j)\right\} \varphi_{i-1} \left(\frac{\varphi_{i-1}}{\eta_{i-1}}\right)^q \lesssim \left(\frac{\varphi_t}{\eta_t}\right)^{q+1}. \end{aligned} \tag{C.52}$$

Substituting (C.51) and (C.52) into (C.41), we obtain

$$v_t \lesssim \eta_t + \eta_t \left(\frac{\varphi_t}{\eta_t}\right)^2 + \left(\frac{\varphi_t}{\eta_t}\right)^{2(q+1)} \lesssim \eta_t + \left(\frac{\varphi_t}{\eta_t}\right)^{2(q+1)}, \tag{C.53}$$

where the last inequality uses the fact that $\varphi_t/\eta_t \to 0$. Consequently, there exists a deterministic constant $C_{v,q+1}(k,r) > 0$ such that

$$v_t \le C_{v,q+1}(k,r) \left\{\eta_t + \left(\frac{\varphi_t}{\eta_t}\right)^{2(q+1)}\right\} \tag{C.54}$$

for all sufficiently large t.

Combining (C.54) with (C.40), we obtain a deterministic constant $C_{u,q+1}(k,r) > 0$ such that

$$u_{t+1} \le \left\{1 - \frac{2C_\varphi(1-\varepsilon_0)\lambda_m}{(t+1)^\varphi}\right\} u_t + \frac{C_{u,q+1}(k,r)}{(t+1)^{\varphi + \min\{\eta,\, 2(q+1)(\varphi-\eta)\}}}. \tag{C.55}$$

By Lemma A.3, we obtain for $\varphi \in (0.5, 1]$

$$\begin{aligned} &\limsup_{t\to\infty} \frac{u_t}{\eta_t + (\varphi_t/\eta_t)^{2(q+1)}} \lesssim \begin{cases} \dfrac{C_{u,q+1}(k,r)}{2C_\varphi(1-\varepsilon_0)\lambda_m}, & \varphi \in (0.5, 1), \\ \dfrac{C_{u,q+1}(k,r)}{2C_\varphi(1-\varepsilon_0)\lambda_m - \min\{\eta,\, 2(q+1)(\varphi-\eta)\}}, & \varphi = 1, \end{cases} \\ &\implies \qquad u_t \lesssim \eta_t + \left(\frac{\varphi_t}{\eta_t}\right)^{2(q+1)}. \end{aligned} \tag{C.56}$$

When $\varphi = 1$, the denominator is positive since

$$\min\{\eta,\, 2(q+1)(\varphi-\eta)\} \le \eta < 2C_\varphi(1-\varepsilon_0)\lambda_m,$$

which is ensured by the choice of ε_0 in (C.28).

Combining the above bound with (C.53), we obtain (C.42) with q replaced by $q+1$. This completes the induction and hence the proof.

C.6 Proof of Theorem 4.3

Recall that $\mathcal{R} := I - K^\star$, which is symmetric positive definite. Fix $r > 0$ and choose $k = k(\epsilon, r)$ as in the proof of Lemma 4.2. The localized rates established in (C.43) and (C.44) give

$$\max\{\mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}],\ \mathbb{E}[\|\boldsymbol{e}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}],\ \mathbb{E}[\|B_t - B^\star\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}]\} \lesssim \eta_t. \tag{C.57}$$

Since $\boldsymbol{m}_t = \boldsymbol{e}_t + \nabla f_t$ and $\|\nabla f_t\| \le \Upsilon_H \|\boldsymbol{\Delta}_t\|$, it follows that

$$\mathbb{E}[\|\boldsymbol{m}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] \le 2\mathbb{E}[\|\boldsymbol{e}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] + 2\Upsilon_H^2 \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] \lesssim \eta_t. \tag{C.58}$$

By Assumptions 3.1 and 3.2 as well as the fact that $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$, we directly obtain

$$\mathbb{E}[\nabla F(\boldsymbol{x}^\star; \xi_t) \mid \mathcal{F}_{t-1}] = \mathbf{0}, \qquad \mathbb{E}[(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t) \mid \mathcal{F}_{t-1}] = \mathbf{0}. \tag{C.59}$$

Moreover, Assumption 3.1 gives

$$\begin{aligned}
\|(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t)\| &= \|\{\nabla F(\boldsymbol{x}_t; \xi_t) - \nabla F(\boldsymbol{x}^\star; \xi_t)\} - \{\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}^\star)\}\| \\
&= \left\| \int_0^1 \left\{ \nabla^2 F(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t; \xi_t) - \nabla^2 f(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t) \right\} \boldsymbol{\Delta}_t \, ds \right\| \\
&\le \int_0^1 \left\{ \|\nabla^2 F(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t; \xi_t)\| + \|\nabla^2 f(\boldsymbol{x}^\star + s\boldsymbol{\Delta}_t)\| \right\} \|\boldsymbol{\Delta}_t\| \, ds \\
&\le 2\Upsilon_H \|\boldsymbol{\Delta}_t\|.
\end{aligned} \tag{C.60}$$

By (C.20), $B_t^{-1}\boldsymbol{m}_t$ is $\mathcal{F}_{t-0.5}$-measurable and hence

$$\mathbb{E}[(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t \mid \mathcal{F}_{t-0.5}] = \mathbb{E}[\widetilde{K}_t - K_t \mid \mathcal{F}_{t-0.5}] B_t^{-1} \boldsymbol{m}_t = \mathbf{0}.$$

Then, by the tower property and (C.59), we have

$$\begin{aligned}
&\mathbb{E}\left[(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t - \mathcal{R}(B^\star)^{-1} \{(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t)\} \mid \mathcal{F}_{t-1} \right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}[(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t \mid \mathcal{F}_{t-0.5}] \mid \mathcal{F}_{t-1} \right] \\
&\qquad\quad - \mathcal{R}(B^\star)^{-1} \mathbb{E}[(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t) \mid \mathcal{F}_{t-1}] = \mathbf{0}.
\end{aligned} \tag{C.61}$$

As a result, by (C.61), we could conclude that the expression inside the conditional expectation forms a martingale difference sequence.

We next control its localized second moment. By (B.8), (C.60), the Cauchy–Schwarz inequality as well as uniform bounds that $\|K_t\| \le \gamma_H/(2\Upsilon_H)$ and $\|(B^\star)^{-1}\|, \|B_t^{-1}\| \le 1/\gamma_H$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\| (\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t - \mathcal{R}(B^\star)^{-1} \{(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t)\} \right\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}} \right] \\
&\le 2\mathbb{E}[\|(\widetilde{K}_t - K_t) B_t^{-1} \boldsymbol{m}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] + 2\mathbb{E}[\|\mathcal{R}(B^\star)^{-1} \{(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star; \xi_t)\}\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\le \frac{2}{\gamma_H^2} \left(C_K + \frac{\gamma_H}{2\Upsilon_H} \right)^2 \mathbb{E}[\|\boldsymbol{m}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] + \frac{8\Upsilon_H^2 \|\mathcal{R}\|^2}{\gamma_H^2} \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] \lesssim \eta_t,
\end{aligned} \tag{C.62}$$

where the last inequality directly follows from (C.57).

Moreover, by the definition of Q_t, (C.22), and $\|(B^\star)^{-1}\| \le 1/\gamma_H$, we obtain

$$\begin{aligned}
\left\| (I - K_t) B_t^{-1} - \mathcal{R}(B^\star)^{-1} \right\| &= \left\| \left\{ (I - K_t) B_t^{-1} B^\star - \mathcal{R} \right\} (B^\star)^{-1} \right\| \\
&= \left\| (Q_t B^\star - \mathcal{R})(B^\star)^{-1} \right\| \le \|Q_t B^\star - \mathcal{R}\| \|(B^\star)^{-1}\| \\
&\le \frac{1}{\gamma_H} \left\{ \frac{1}{\gamma_H} \left(1 + \frac{\gamma_H}{2\Upsilon_H} \right) + L_K \right\} \|B_t - B^\star\| \lesssim \|B_t - B^\star\|.
\end{aligned} \tag{C.63}$$

Additionally, by the triangle inequality, (C.23), and (C.63), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\| \mathcal{R}(B^\star)^{-1} (\nabla f_t - B^\star \boldsymbol{\Delta}_t) + \left\{ (I - K_t) B_t^{-1} - \mathcal{R}(B^\star)^{-1} \right\} \boldsymbol{m}_t \right\| \mathbf{1}_{\{\tau_{k,r}>t\}} \right] \\
&\le \|\mathcal{R}\| \|(B^\star)^{-1}\| \mathbb{E}[\|\nabla f_t - B^\star \boldsymbol{\Delta}_t\| \mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\quad + \mathbb{E}\left[\left\| (I - K_t) B_t^{-1} - \mathcal{R}(B^\star)^{-1} \right\| \|\boldsymbol{m}_t\| \mathbf{1}_{\{\tau_{k,r}>t\}} \right] \\
&\lesssim \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] + \mathbb{E}[\|B_t - B^\star\| \|\boldsymbol{m}_t\| \mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\le \mathbb{E}[\|\boldsymbol{\Delta}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}] + \{\mathbb{E}[\|B_t - B^\star\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}]\}^{1/2} \{\mathbb{E}[\|\boldsymbol{m}_t\|^2 \mathbf{1}_{\{\tau_{k,r}>t\}}]\}^{1/2} \lesssim \eta_t,
\end{aligned} \tag{C.64}$$

where the second-to-last inequality is due to the Cauchy–Schwarz inequality, and the last inequality directly follows from (C.57) and (C.58).

• **Proof of (4.1).** The moving-average update implies the following identity

$$\boldsymbol{m}_t = \boldsymbol{g}_t - \frac{1-\eta_t}{\eta_t}(\boldsymbol{m}_t - \boldsymbol{m}_{t-1}).$$

Moreover, we have

$$\boldsymbol{g}_t = B^\star \boldsymbol{\Delta}_t + \nabla F(\boldsymbol{x}^\star;\xi_t) + \{(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star;\xi_t)\} + (\nabla f_t - B^\star \boldsymbol{\Delta}_t).$$

Substituting the above two identities into the outer update and rearranging terms, we obtain

$$\begin{aligned}
\boldsymbol{\Delta}_{t+1} &= \boldsymbol{\Delta}_t - \varphi_t(I - \widetilde{K}_t)B_t^{-1}\boldsymbol{m}_t \\
&= \boldsymbol{\Delta}_t - \varphi_t\mathcal{R}(B^\star)^{-1}\boldsymbol{m}_t + \varphi_t(\widetilde{K}_t - K_t)B_t^{-1}\boldsymbol{m}_t - \varphi_t\left\{(I-K_t)B_t^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_t \\
&= \boldsymbol{\Delta}_t - \varphi_t\mathcal{R}(B^\star)^{-1}\boldsymbol{g}_t + \varphi_t\frac{1-\eta_t}{\eta_t}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_t - \boldsymbol{m}_{t-1}) + \varphi_t(\widetilde{K}_t - K_t)B_t^{-1}\boldsymbol{m}_t \\
&\quad - \varphi_t\left\{(I-K_t)B_t^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_t \\
&= (I - \varphi_t\mathcal{R})\boldsymbol{\Delta}_t - \varphi_t\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_t) + \varphi_t\frac{1-\eta_t}{\eta_t}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_t - \boldsymbol{m}_{t-1}) \\
&\quad + \varphi_t\left[(\widetilde{K}_t - K_t)B_t^{-1}\boldsymbol{m}_t - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_t - \nabla f_t) - \nabla F(\boldsymbol{x}^\star;\xi_t)\}\right] \\
&\quad - \varphi_t\left[\mathcal{R}(B^\star)^{-1}(\nabla f_t - B^\star\boldsymbol{\Delta}_t) + \left\{(I-K_t)B_t^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_t\right].
\end{aligned} \tag{C.65}$$

Iterating (C.65) from the fixed index k yields

$$\begin{aligned}
\boldsymbol{\Delta}_t &= \Pi_{t,k}\boldsymbol{\Delta}_k - \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) \\
&\quad + \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i - \boldsymbol{m}_{i-1}) \\
&\quad + \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i - K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i - \nabla f_i) - \nabla F(\boldsymbol{x}^\star;\xi_i)\}\right] \\
&\quad - \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right].
\end{aligned} \tag{C.66}$$

We next derive bounds for $\Pi_{t,i}$ and the weighted sums appearing in (C.66).

Without loss of generality, we may assume that $\varphi_j\|\mathcal{R}\| \le 1$ for every $j \ge k$. Since $\mathcal{R}$ is symmetric positive definite, for every $i \ge k$, we obtain

$$\|I - \varphi_j\mathcal{R}\| = 1 - \lambda_m\varphi_j, \qquad \|\Pi_{n,i}\| \le \prod_{j=i}^{n-1}(1 - \lambda_m\varphi_j). \tag{C.67}$$

We first consider the case where $\varphi < 1$. For every fixed $i_0 \ge k$, by (C.67), we obtain

$$\|\Pi_{t,i_0}\| \le \exp\left\{-\lambda_m\sum_{j=i_0}^{t-1}\varphi_j\right\} \le \exp\left[-C\left\{(t+1)^{1-\varphi} - (i_0+1)^{1-\varphi}\right\}\right] = o(\sqrt{\varphi_t}),$$

where $C > 0$ is some constant independent of t. Since $\eta > \varphi/2$, we have $\eta_t = o(\sqrt{\varphi_t})$. In addition, $\varphi_t/\eta_t \to 0$ implies that

$$\frac{\varphi_t}{\sqrt{\eta_t}} = \sqrt{\varphi_t}\sqrt{\frac{\varphi_t}{\eta_t}} = o(\sqrt{\varphi_t}).$$

Therefore, by Lemma A.3, we have

$$\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2 \lesssim \varphi_t, \qquad \sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2\eta_i \lesssim \varphi_t\eta_t = o(\varphi_t),$$
$$\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|\varphi_i\eta_i \lesssim \eta_t = o(\sqrt{\varphi_t}), \qquad \sum_{i=k}^{t-2}\|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\sqrt{\eta_i}} \lesssim \frac{\varphi_t}{\sqrt{\eta_t}} = o(\sqrt{\varphi_t}). \tag{C.68}$$

Then we consider the case where $\varphi = 1$. For every fixed $i_0 \geq k$, by (C.67), we obtain

$$\|\Pi_{t,i_0}\| \leq \exp\left\{-\lambda_m\sum_{j=i_0}^{t-1}\varphi_j\right\} \lesssim \left(\frac{i_0+1}{t+1}\right)^{C_\varphi\lambda_m} \lesssim (t+1)^{-C_\varphi\lambda_m} = o(t^{-1/2}) = o(\sqrt{\varphi_t}),$$

where the last two relations follow from $C_\varphi\lambda_m > 1/2$. Moreover, since $\varphi_t \asymp t^{-1}$, we obtain

$$\begin{aligned}
\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2 &\lesssim (t+1)^{-2C_\varphi\lambda_m}\sum_{i=k}^{t-1}(i+1)^{2C_\varphi\lambda_m-2} \lesssim t^{-1} \lesssim \varphi_t,\\
\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2\eta_i &\lesssim (t+1)^{-2C_\varphi\lambda_m}\sum_{i=k}^{t-1}(i+1)^{2C_\varphi\lambda_m-2-\eta}\\
&\lesssim t^{-\min\{2C_\varphi\lambda_m,\,1+\eta\}}(1+\log t) = o(t^{-1}) = o(\varphi_t),\\
\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|\varphi_i\eta_i &\lesssim (t+1)^{-C_\varphi\lambda_m}\sum_{i=k}^{t-1}(i+1)^{C_\varphi\lambda_m-1-\eta}\\
&\lesssim t^{-\min\{C_\varphi\lambda_m,\,\eta\}}(1+\log t) = o(t^{-1/2}) = o(\sqrt{\varphi_t}),\\
\sum_{i=k}^{t-2}\|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\sqrt{\eta_i}} &\lesssim (t+1)^{-C_\varphi\lambda_m}\sum_{i=k}^{t-2}(i+1)^{C_\varphi\lambda_m-2+\eta/2}\\
&\lesssim t^{-\min\{C_\varphi\lambda_m,\,1-\eta/2\}}(1+\log t) = o(t^{-1/2}) = o(\sqrt{\varphi_t}).
\end{aligned} \tag{C.69}$$

Here, $1+\log t$ is included to cover the situation in which the exponent of a corresponding power sum equals -1; the little-o relations follow from $C_\varphi\lambda_m > 1/2$, $\eta > 1/2$ and $1-\eta/2 > 1/2$.

Combining the cases $\varphi < 1$ and $\varphi = 1$, for every fixed $i_0 \geq k$, we obtain $\|\Pi_{n,i_0}\| = o(\sqrt{\varphi_n})$ and

$$\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2 \lesssim \varphi_t, \qquad \sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2\eta_i = o(\varphi_t),$$
$$\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|\varphi_i\eta_i = o(\sqrt{\varphi_t}), \qquad \sum_{i=k}^{t-2}\|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\sqrt{\eta_i}} = o(\sqrt{\varphi_t}). \tag{C.70}$$

In (C.66), the second term on the right-hand side is the leading stochastic term. It remains to show the other three sums are $o_p(\sqrt{\varphi_t})$. We control these sums in the same order as they appear in (C.66).

We first consider the moving-average difference term. For the prescribed sequences, we have

$$\varphi_t\frac{1-\eta_t}{\eta_t} = \frac{C_\varphi}{C_\eta}(t+1)^{-(\varphi-\eta)} - C_\varphi(t+1)^{-\varphi} \lesssim \frac{\varphi_t}{\eta_t}.$$

Furthermore, by the mean-value theorem and $\varphi \leq 1$, we have

$$\left|\varphi_{t+1}\frac{1-\eta_{t+1}}{\eta_{t+1}} - \varphi_t\frac{1-\eta_t}{\eta_t}\right| \lesssim (t+1)^{-(\varphi-\eta)-1} + (t+1)^{-\varphi-1} \lesssim \frac{\varphi_t^2}{\eta_t}. \tag{C.71}$$

Applying discrete summation by parts to the third term on the right-hand side of (C.66), we obtain

$$\begin{aligned}
&\sum_{i=k}^{t-1} \Pi_{t,i+1}\varphi_i \frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i - \boldsymbol{m}_{i-1}) \\
&\qquad = \frac{\varphi_{t-1}(1-\eta_{t-1})}{\eta_{t-1}}\mathcal{R}(B^\star)^{-1}\boldsymbol{m}_{t-1} - \Pi_{t,k+1}\frac{\varphi_k(1-\eta_k)}{\eta_k}\mathcal{R}(B^\star)^{-1}\boldsymbol{m}_{k-1} \\
&\qquad\quad + \sum_{i=k}^{t-2}\left\{\Pi_{t,i+1}\frac{\varphi_i(1-\eta_i)}{\eta_i} - \Pi_{t,i+2}\frac{\varphi_{i+1}(1-\eta_{i+1})}{\eta_{i+1}}\right\}\mathcal{R}(B^\star)^{-1}\boldsymbol{m}_i. \qquad \text{(C.72)}
\end{aligned}$$

Since $\Pi_{t,i+1} = \Pi_{t,i+2}(I - \varphi_{i+1}\mathcal{R})$, for the last line of (C.72), we have

$$\begin{aligned}
&\left\|\Pi_{t,i+1}\frac{\varphi_i(1-\eta_i)}{\eta_i} - \Pi_{t,i+2}\frac{\varphi_{i+1}(1-\eta_{i+1})}{\eta_{i+1}}\right\| \\
&\le \|\Pi_{t,i+2}\|\left\{\left|\frac{\varphi_i(1-\eta_i)}{\eta_i} - \frac{\varphi_{i+1}(1-\eta_{i+1})}{\eta_{i+1}}\right| + \varphi_{i+1}\frac{\varphi_i(1-\eta_i)}{\eta_i}\|\mathcal{R}\|\right\} \lesssim \|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\eta_i}. \qquad \text{(C.73)}
\end{aligned}$$

where the last inequality follows from (C.71).

Since $\mathbf{1}_{\{\tau_{k,r}>t\}} \le \mathbf{1}_{\{\tau_{k,r}>i\}}$ for every $i \le t$, Jensen's inequality direclty yields

$$\mathbb{E}[\|\boldsymbol{m}_i\|\mathbf{1}_{\{\tau_{k,r}>t\}}] \le \left\{\mathbb{E}[\|\boldsymbol{m}_i\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}^{1/2}.$$

Therefore, applying the triangle inequality to (C.72) and then using (C.58) and (C.73), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] \\
&\le \frac{\varphi_{t-1}(1-\eta_{t-1})}{\eta_{t-1}}\|\mathcal{R}(B^\star)^{-1}\|\mathbb{E}[\|\boldsymbol{m}_{t-1}\|\mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\quad + \|\Pi_{t,k+1}\|\frac{\varphi_k(1-\eta_k)}{\eta_k}\|\mathcal{R}(B^\star)^{-1}\|\mathbb{E}[\|\boldsymbol{m}_{k-1}\|\mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\quad + \sum_{i=k}^{t-2}\left\|\Pi_{t,i+1}\frac{\varphi_i(1-\eta_i)}{\eta_i} - \Pi_{t,i+2}\frac{\varphi_{i+1}(1-\eta_{i+1})}{\eta_{i+1}}\right\|\|\mathcal{R}(B^\star)^{-1}\|\mathbb{E}[\|\boldsymbol{m}_i\|\mathbf{1}_{\{\tau_{k,r}>t\}}] \\
&\lesssim \frac{\varphi_{t-1}}{\eta_{t-1}}\left\{\mathbb{E}[\|\boldsymbol{m}_{t-1}\|^2\mathbf{1}_{\{\tau_{k,r}>t-1\}}]\right\}^{1/2} + \|\Pi_{t,k+1}\|\mathbb{E}[\|\boldsymbol{m}_{k-1}\|] \\
&\quad + \sum_{i=k}^{t-2}\|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\eta_i}\left\{\mathbb{E}[\|\boldsymbol{m}_i\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}^{1/2} \\
&\lesssim \frac{\varphi_{t-1}}{\sqrt{\eta_{t-1}}} + \|\Pi_{t,k+1}\| + \sum_{i=k}^{t-2}\|\Pi_{t,i+2}\|\frac{\varphi_i^2}{\sqrt{\eta_i}} = o(\sqrt{\varphi_t}). \qquad \text{(C.74)}
\end{aligned}$$

where the last two steps follow from (C.58), (C.70) and $\varphi_t/\eta_t \to 0$.

We next control the centered sum in (C.66). By (C.61), the bracketed expressions in this sum form a martingale difference sequence. Since $\mathbf{1}_{\{\tau_{k,r}>i\}}$ is $\mathcal{F}_{i-1}$-measurable, multiplying each expression

by $\mathbf{1}_{\{\tau_{k,r}>i\}}$ preserves the martingale difference property. Therefore, by (C.62) and (C.70), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\
&\le\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\mathbf{1}_{\{\tau_{k,r}>i\}}\right\|^2\right]\\
&=\sum_{i=k}^{t-1}\mathbb{E}\left[\left\|\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}\right]\\
&\le\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2\mathbb{E}\left[\left\|(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}\right]\\
&\lesssim\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2\eta_i=o(\varphi_t).
\end{aligned}\tag{C.75}$$

For the last sum in (C.66), by the triangle inequality, (C.64) and (C.70), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i-B^\star\boldsymbol{\Delta}_i)+\left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\
&\le\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|\varphi_i\mathbb{E}\left[\left\|\mathcal{R}(B^\star)^{-1}(\nabla f_i-B^\star\boldsymbol{\Delta}_i)+\left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right\|\mathbf{1}_{\{\tau_{k,r}>i\}}\right]\\
&\lesssim\sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|\varphi_i\eta_i=o(\sqrt{\varphi_t}).
\end{aligned}\tag{C.76}$$

Equations (C.74), (C.75), and (C.76) control the three remainder sums in (C.66) on $\{\tau_{k,r}>t\}$. We next remove the stopping time. For every $\zeta>0$, (C.66) gives us the following:

$$\begin{aligned}
&\mathbb{P}\left(\left\|\boldsymbol{\Delta}_t-\Pi_{t,k}\boldsymbol{\Delta}_k+\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|>\frac{\zeta\sqrt{\varphi_t}}{2}\right)\\
&\le\mathbb{P}(\tau_{k,r}\le t)+\mathbb{P}\left(\left\|\boldsymbol{\Delta}_t-\Pi_{t,k}\boldsymbol{\Delta}_k+\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}>\frac{\zeta\sqrt{\varphi_t}}{2}\right).
\end{aligned}$$

For the fixed (k,r), the vector inside the second probability is the sum of the three remainder terms in (C.66). By the triangle inequality and the union bound, it suffices to show that the probability associated with each remainder term, with threshold $\zeta\sqrt{\varphi_t}/6$, converges to zero.

For the moving-average difference term, by Markov's inequality and (C.74), we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}>\frac{\zeta\sqrt{\varphi_t}}{6}\right)\\
&\le\frac{6}{\zeta\sqrt{\varphi_t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right]=o(1).
\end{aligned}$$

For the centered term, by Chebyshev's inequality and (C.75), we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}>\frac{\zeta\sqrt{\varphi_t}}{6}\right)\\
&\le\frac{36}{\zeta^2\varphi_t}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}\right]=o(1).
\end{aligned}$$

For the noncentered term, by Markov's inequality and (C.76), we have

$$\mathbb{P}\left(\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta\sqrt{\varphi_t}}{6}\right)$$
$$\leq \frac{6}{\zeta\sqrt{\varphi_t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1} - \mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] = o(1).$$

Since second probability in the stopping-time decomposition converges to zero, we have

$$\limsup_{t\to\infty}\mathbb{P}\left(\left\|\boldsymbol{\Delta}_t - \Pi_{t,k}\boldsymbol{\Delta}_k + \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| > \frac{\zeta\sqrt{\varphi_t}}{2}\right) \leq \mathbb{P}(\tau_{k,r}<\infty) < \frac{\epsilon}{2}.$$

Lastly, we extend the stochastic sum from $i=k$ to $i=0$ and we have

$$\begin{aligned}
&\boldsymbol{\Delta}_t + \sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\\
&\qquad= \boldsymbol{\Delta}_t - \Pi_{t,k}\boldsymbol{\Delta}_k + \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\\
&\qquad\quad+ \Pi_{t,k}\left\{\boldsymbol{\Delta}_k + \sum_{i=0}^{k-1}\Pi_{k,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\}.
\end{aligned}$$

For fixed k, the quantity inside the braces does not depend on t and is finite almost surely. Moreover, (C.70) gives $\|\Pi_{t,k}\| = o(\sqrt{\varphi_t})$. Therefore, we obtain

$$\frac{1}{\sqrt{\varphi_t}}\left\|\Pi_{t,k}\left\{\boldsymbol{\Delta}_k + \sum_{i=0}^{k-1}\Pi_{k,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\}\right\| \overset{a.s.}{\to} 0.$$

Hence, the second term on the right-hand side of the preceding decomposition is $o_p(\sqrt{\varphi_t})$.

For every $\zeta > 0$, the triangle inequality further yields

$$\begin{aligned}
&\mathbb{P}\left(\left\|\boldsymbol{\Delta}_t + \sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| > \zeta\sqrt{\varphi_t}\right)\\
&\qquad\leq \mathbb{P}\left(\left\|\boldsymbol{\Delta}_t - \Pi_{t,k}\boldsymbol{\Delta}_k + \sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| > \frac{\zeta\sqrt{\varphi_t}}{2}\right)\\
&\qquad\quad+ \mathbb{P}\left(\left\|\Pi_{t,k}\left\{\boldsymbol{\Delta}_k + \sum_{i=0}^{k-1}\Pi_{k,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\}\right\| > \frac{\zeta\sqrt{\varphi_t}}{2}\right).
\end{aligned}$$

For the fixed $k = k(\epsilon, r)$, the preceding bound controls the first probability, whereas the second probability converges to zero by the almost-sure convergence established above. It follows that

$$\limsup_{t\to\infty}\mathbb{P}\left(\left\|\boldsymbol{\Delta}_t + \sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| > \zeta\sqrt{\varphi_t}\right) \leq \mathbb{P}(\tau_{k,r}<\infty) < \frac{\epsilon}{2}.$$

Since $\epsilon > 0$ is arbitrary and the term inside the probability is independent of k, we conclude that

$$\boldsymbol{\Delta}_t + \sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) = o_p(\sqrt{\varphi_t}).$$

Recalling that $\mathcal{R} = I - K^\star$ and dividing both sides by $\sqrt{\varphi_t}$ yield

$$\frac{1}{\sqrt{\varphi_t}}\boldsymbol{\Delta}_t = -\frac{1}{\sqrt{\varphi_t}}\sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i(I-K^\star)(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) + o_p(1).$$

Then we complete the proof of (4.1).

• **Proof of (4.2).** Suppose throughout this part that $\varphi < 1$. We return to the fixed localized system (k, r) used above. Before averaging (C.65), we first derive a localized first-moment bound for $\boldsymbol{\Delta}_t$.

By Assumptions 3.1 and 3.2, $(\nabla F(\boldsymbol{x}^\star;\xi_i), \mathcal{F}_i)_{i\geq 0}$ forms a martingale difference sequence and $\mathbb{E}[\|\nabla F(\boldsymbol{x}^\star;\xi_i)\|^2] \lesssim 1$. Therefore, by (C.70), we have

$$\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|^2\right] = \sum_{i=k}^{t-1}\mathbb{E}\left[\left\|\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|^2\right] \lesssim \sum_{i=k}^{t-1}\|\Pi_{t,i+1}\|^2\varphi_i^2 \lesssim \varphi_t. \quad \text{(C.77)}$$

Taking norms in (C.66), multiplying by $\mathbf{1}_{\{\tau_{k,r}>t\}}$, and using $\mathbf{1}_{\{\tau_{k,r}>t\}} \leq \mathbf{1}_{\{\tau_{k,r}>i\}}$ for every $i \leq t$, the triangle inequality and the Cauchy–Schwarz inequality yield

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\Delta}_t\|\mathbf{1}_{\{\tau_{k,r}>t\}}] &\leq r\|\Pi_{t,k}\| + \left\{\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|^2\right]\right\}^{1/2} \\
&+ \mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\frac{1-\eta_i}{\eta_i}\mathcal{R}(B^\star)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] \\
&+ \left\{\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\mathbf{1}_{\{\tau_{k,r}>i\}}\right\|^2\right]\right\}^{1/2} \\
&+ \mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\Pi_{t,i+1}\varphi_i\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] \\
&\lesssim r\|\Pi_{t,k}\| + \sqrt{\varphi_t} + o(\sqrt{\varphi_t}) + \{o(\varphi_t)\}^{1/2} + o(\sqrt{\varphi_t}) \lesssim \sqrt{\varphi_t}. \quad \text{(C.78)}
\end{aligned}$$

where the first inequality uses $\|\boldsymbol{\Delta}_k\| \leq r$ on $\{\tau_{k,r} > t\}$, the second inequality follows from (C.74), (C.75), (C.76) and (C.77), while the last inequality uses $\|\Pi_{t,k}\| = o(\sqrt{\varphi_t})$ from (C.70).

We now rearrange (C.65) as follows,

$$\begin{aligned}
\boldsymbol{\Delta}_i = &-(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) - \mathcal{R}^{-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i} + \frac{1-\eta_i}{\eta_i}(B^\star)^{-1}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1}) \\
&+ \mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right] \\
&- \mathcal{R}^{-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]. \quad \text{(C.79)}
\end{aligned}$$

Since k is some fixed index, the averaged iterate can be decomposed as

$$\sqrt{t}(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) = \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\boldsymbol{\Delta}_i = \frac{1}{\sqrt{t}}\sum_{i=0}^{k-1}\boldsymbol{\Delta}_i + \frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i.$$

We first analyze the second term and restore the fixed finite prefix at the end. Summing (C.79) from $i = k$ to $t-1$ and dividing both sides by $\sqrt{t}$, we obtain

$$\begin{aligned}\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i = &-\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) - \frac{1}{\sqrt{t}}\mathcal{R}^{-1}\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i}\\ &+\frac{1}{\sqrt{t}}(B^\star)^{-1}\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\\ &+\frac{1}{\sqrt{t}}\mathcal{R}^{-1}\sum_{i=k}^{t-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\\ &-\frac{1}{\sqrt{t}}\mathcal{R}^{-1}\sum_{i=k}^{t-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i)+\left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right].\end{aligned} \tag{C.80}$$

The first term on the right-hand side is the leading stochastic term, and it remains to show that each of the other four terms is $o_p(1)$.

For the second term on the right-hand side of (C.80), summation by parts gives us

$$\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i} = \frac{\boldsymbol{\Delta}_t}{\varphi_{t-1}} - \frac{\boldsymbol{\Delta}_k}{\varphi_k} + \sum_{i=k+1}^{t-1}\left(\frac{1}{\varphi_{i-1}}-\frac{1}{\varphi_i}\right)\boldsymbol{\Delta}_i. \tag{C.81}$$

Since $\varphi_i^{-1} = (i+1)^\varphi / C_\varphi$, by the mean-value theorem, we have

$$\left|\frac{1}{\varphi_i}-\frac{1}{\varphi_{i-1}}\right| = \frac{1}{C_\varphi}\left|(i+1)^\varphi - i^\varphi\right| \lesssim (i+1)^{\varphi-1}.$$

Therefore, by (C.78) and $\mathbf{1}_{\{\tau_{k,r}>t\}} \le \mathbf{1}_{\{\tau_{k,r}>i\}}$, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i}\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\ &\le \frac{1}{\sqrt{t}}\left\{\frac{\mathbb{E}[\|\boldsymbol{\Delta}_t\|\mathbf{1}_{\{\tau_{k,r}>t\}}]}{\varphi_{t-1}} + \frac{r}{\varphi_k} + \sum_{i=k+1}^{t-1}\left|\frac{1}{\varphi_{i-1}}-\frac{1}{\varphi_i}\right|\mathbb{E}[\|\boldsymbol{\Delta}_i\|\mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}\\ &\lesssim \frac{1}{\sqrt{t}}\left\{\frac{1}{\sqrt{\varphi_t}}+1+\sum_{i=k+1}^{t-1}(i+1)^{\varphi-1}\sqrt{\varphi_i}\right\} \lesssim \frac{1}{t^{(1-\varphi)/2}} = o(1).\end{aligned} \tag{C.82}$$

Here, r/φ_k follows from $\|\boldsymbol{\Delta}_k\| \le r$ on $\{\tau_{k,r} > t\}$.

For the third term on the right-hand side of (C.80), again we apply summation by parts and obtain

$$\begin{aligned}\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1}) = &\frac{1-\eta_{t-1}}{\eta_{t-1}}\boldsymbol{m}_{t-1} - \frac{1-\eta_k}{\eta_k}\boldsymbol{m}_{k-1}\\ &+\sum_{i=k}^{t-2}\left(\frac{1-\eta_i}{\eta_i}-\frac{1-\eta_{i+1}}{\eta_{i+1}}\right)\boldsymbol{m}_i.\end{aligned} \tag{C.83}$$

In addition, the mean-value theorem gives us

$$\left|\frac{1-\eta_i}{\eta_i}-\frac{1-\eta_{i+1}}{\eta_{i+1}}\right| \lesssim (i+1)^{\eta-1} \lesssim \frac{1}{(i+1)\eta_i}.$$

Using (C.58), $\mathbf{1}_{\{\tau_{k,r}>t\}} \le \mathbf{1}_{\{\tau_{k,r}>i\}}$, and Jensen's inequality, we further obtain

$$
\begin{aligned}
&\frac{1}{\sqrt{t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\
&\le \frac{1}{\sqrt{t}}\left\{\frac{1}{\eta_{t-1}}\mathbb{E}[\|\boldsymbol{m}_{t-1}\|\mathbf{1}_{\{\tau_{k,r}>t\}}]+\frac{1-\eta_k}{\eta_k}\mathbb{E}[\|\boldsymbol{m}_{k-1}\|]+\sum_{i=k}^{t-2}\frac{1}{(i+1)\eta_i}\mathbb{E}[\|\boldsymbol{m}_i\|\mathbf{1}_{\{\tau_{k,r}>t\}}]\right\}\\
&\le \frac{1}{\sqrt{t}}\left\{\frac{1}{\eta_{t-1}}\left\{\mathbb{E}[\|\boldsymbol{m}_{t-1}\|^2\mathbf{1}_{\{\tau_{k,r}>t-1\}}]\right\}^{1/2}+\frac{1-\eta_k}{\eta_k}\mathbb{E}[\|\boldsymbol{m}_{k-1}\|]\right.\\
&\qquad\left.+\sum_{i=k}^{t-2}\frac{1}{(i+1)\eta_i}\left\{\mathbb{E}[\|\boldsymbol{m}_i\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}]\right\}^{1/2}\right\}\\
&\lesssim \frac{1}{\sqrt{t}}\left\{\frac{\sqrt{\eta_{t-1}}}{\eta_{t-1}}+1+\sum_{i=k}^{t-2}\frac{\sqrt{\eta_i}}{(i+1)\eta_i}\right\}=\frac{1}{\sqrt{t}}\left\{\frac{1}{\sqrt{\eta_{t-1}}}+1+\sum_{i=k}^{t-2}\frac{1}{(i+1)\sqrt{\eta_i}}\right\}\\
&\lesssim \frac{1}{\sqrt{t}}\left\{t^{\eta/2}+1+\sum_{i=k}^{t-2}(i+1)^{-1+\eta/2}\right\}\lesssim\frac{1}{\sqrt{t}}\left\{t^{\eta/2}+1+(t+1)^{\eta/2}\right\}\\
&\lesssim \frac{1}{t^{(1-\eta)/2}}=o(1).
\end{aligned}
\tag{C.84}
$$

Here, the fixed boundary term involving $\boldsymbol{m}_{k-1}$ has a finite first moment and is absorbed into the hidden constant; the last equality directly follows from $\eta<\varphi<1$.

For the fourth term on the right-hand side of (C.80), by (C.61) and (C.62), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\
&\le \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\mathbf{1}_{\{\tau_{k,r}>i\}}\right\|^2\right]\\
&=\frac{1}{t}\sum_{i=k}^{t-1}\mathbb{E}\left[\left\|\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i-\mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>i\}}\right]\\
&\lesssim\frac{1}{t}\sum_{i=k}^{t-1}\eta_i\lesssim t^{-\eta}=o(1).
\end{aligned}
\tag{C.85}
$$

For the final term on the right-hand side of (C.80), by the triangle inequality and (C.64), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i-B^\star\boldsymbol{\Delta}_i)+\left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right]\\
&\le\frac{\|\mathcal{R}^{-1}\|}{\sqrt{t}}\sum_{i=k}^{t-1}\mathbb{E}\left[\left\|\mathcal{R}(B^\star)^{-1}(\nabla f_i-B^\star\boldsymbol{\Delta}_i)+\left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right\|\mathbf{1}_{\{\tau_{k,r}>i\}}\right]\\
&\lesssim\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\eta_i\lesssim\frac{1}{t^{\eta-1/2}}=o(1),
\end{aligned}
\tag{C.86}
$$

where the last equality follows from $\eta>1/2$.

Similarly to the proof of (4.1), it remains to remove the stopping time and restore the finitely many indices preceding k. For every $\zeta > 0$, the stopping-time decomposition gives

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i + \frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| > \frac{\zeta}{2}\right)$$
$$\leq \mathbb{P}(\tau_{k,r} \leq t) + \mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i + \frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| \mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta}{2}\right).$$

On $\{\tau_{k,r} > t\}$, by (C.80) and the triangle inequality, we have

$$\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i + \frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\| \mathbf{1}_{\{\tau_{k,r}>t\}}$$
$$\leq \frac{\|\mathcal{R}^{-1}\|}{\sqrt{t}}\left\|\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i}\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} + \frac{\|(B^\star)^{-1}\|}{\sqrt{t}}\left\|\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}$$
$$+\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}$$
$$+\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}.$$

Therefore, if the left-hand side is greater than $\zeta/2$, then at least one term on the right-hand side must be greater than $\zeta/8$. By the union bound, it remains to control the corresponding four probabilities.

For the iterate-difference term, by Markov's inequality and (C.82), we have

$$\mathbb{P}\left(\frac{\|\mathcal{R}^{-1}\|}{\sqrt{t}}\left\|\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i}\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta}{8}\right)$$
$$\leq \frac{8\|\mathcal{R}^{-1}\|}{\zeta\sqrt{t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\frac{\boldsymbol{\Delta}_{i+1}-\boldsymbol{\Delta}_i}{\varphi_i}\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] = o(1).$$

For the moving-average difference term, by Markov's inequality and (C.84), we have

$$\mathbb{P}\left(\frac{\|(B^\star)^{-1}\|}{\sqrt{t}}\left\|\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta}{8}\right)$$
$$\leq \frac{8\|(B^\star)^{-1}\|}{\zeta\sqrt{t}}\mathbb{E}\left[\left\|\sum_{i=k}^{t-1}\frac{1-\eta_i}{\eta_i}(\boldsymbol{m}_i-\boldsymbol{m}_{i-1})\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] = o(1).$$

For the centered term, by Chebyshev's inequality and (C.85), we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta}{8}\right)$$
$$\leq \frac{64}{\zeta^2}\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[(\widetilde{K}_i-K_i)B_i^{-1}\boldsymbol{m}_i - \mathcal{R}(B^\star)^{-1}\{(\boldsymbol{g}_i-\nabla f_i)-\nabla F(\boldsymbol{x}^\star;\xi_i)\}\right]\right\|^2\mathbf{1}_{\{\tau_{k,r}>t\}}\right] = o(1).$$

For the noncentered term, by Markov's inequality and (C.86), we have

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}} > \frac{\zeta}{8}\right)$$
$$\leq \frac{8}{\zeta}\mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\mathcal{R}^{-1}\left[\mathcal{R}(B^\star)^{-1}(\nabla f_i - B^\star\boldsymbol{\Delta}_i) + \left\{(I-K_i)B_i^{-1}-\mathcal{R}(B^\star)^{-1}\right\}\boldsymbol{m}_i\right]\right\|\mathbf{1}_{\{\tau_{k,r}>t\}}\right] = o(1).$$

Since all four probabilities converge to zero, the second probability in the stopping-time decomposition converges to zero. As a result, we obtain

$$\limsup_{t\to\infty}\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i+\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|>\frac{\zeta}{2}\right)\le\mathbb{P}(\tau_{k,r}<\infty)<\frac{\epsilon}{2}.$$

Lastly, we extend both sums from $i=k$ to $i=0$. For the fixed k, we have

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{k-1}\left\{\boldsymbol{\Delta}_i+(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\}\overset{a.s.}{\to}\mathbf{0}.$$

Therefore, by the triangle inequality and the union bound,

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\boldsymbol{\Delta}_i+\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|>\zeta\right)\\&\qquad\le\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}\boldsymbol{\Delta}_i+\frac{1}{\sqrt{t}}\sum_{i=k}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|>\frac{\zeta}{2}\right)\\&\qquad\qquad+\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=0}^{k-1}\left\{\boldsymbol{\Delta}_i+(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\}\right\|>\frac{\zeta}{2}\right).\end{aligned}$$

Since the second probability on the right-hand side converges to zero by the preceding almost-sure convergence, then it follows that

$$\limsup_{t\to\infty}\mathbb{P}\left(\left\|\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\boldsymbol{\Delta}_i+\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\right\|>\zeta\right)\le\mathbb{P}(\tau_{k,r}<\infty)<\frac{\epsilon}{2}.$$

Since $\epsilon>0$ is arbitrary, we conclude that

$$\sqrt{t}(\bar{\boldsymbol{x}}_t-\boldsymbol{x}^\star)=\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(\boldsymbol{x}_i-\boldsymbol{x}^\star)=-\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)+o_p(1),$$

which proves (4.2) and completes the proof.

C.7 Proof of Theorem 4.5

By Lemma 4.1, $\mathcal{R}$ is symmetric positive definite. By (4.1) and Slutsky's theorem, it suffices to establish the asymptotic normality of its leading stochastic term. For every $0\le i\le t-1$, define

$$W_{t,i}:=-\frac{\varphi_i}{\sqrt{\varphi_t}}\Pi_{t,i+1}\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i).\tag{C.87}$$

- **Verification of (A.1).** By the unbiasedness in Assumption 3.2 and $\nabla f(\boldsymbol{x}^\star)=\mathbf{0}$, we have

$$\mathbb{E}[\nabla F(\boldsymbol{x}^\star;\xi_i)\mid\mathcal{F}_{i-1}]=\nabla f(\boldsymbol{x}^\star)=\mathbf{0}.$$

Since $\Pi_{t,i+1}$ is deterministic, it follows that

$$\mathbb{E}[W_{t,i}\mid\mathcal{F}_{i-1}]=-\frac{\varphi_i}{\sqrt{\varphi_t}}\Pi_{t,i+1}\mathcal{R}(B^\star)^{-1}\mathbb{E}[\nabla F(\boldsymbol{x}^\star;\xi_i)\mid\mathcal{F}_{i-1}]=\mathbf{0}.$$

- **Verification of (A.2).** By the definitions of Ω and $\Gamma_{\boldsymbol{m},\mathrm{as}}$, we have

$$\mathbb{E}\left[\mathcal{R}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\nabla F(\boldsymbol{x}^\star;\xi_i)^\top(B^\star)^{-1}\mathcal{R}^\top\mid\mathcal{F}_{i-1}\right]=\mathcal{R}\Omega\mathcal{R}^\top=\Gamma_{\boldsymbol{m},\mathrm{as}}.$$

Therefore, the conditional covariance sum of the triangular array is deterministic and satisfies

$$\mathcal{V}_t:=\sum_{i=0}^{t-1}\mathbb{E}[W_{t,i}W_{t,i}^\top\mid\mathcal{F}_{i-1}]=\frac{1}{\varphi_t}\sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i^2\Gamma_{\boldsymbol{m},\mathrm{as}}\Pi_{t,i+1}^\top.\tag{C.88}$$

Since $\Pi_{t+1,i+1} = (I - \varphi_t \mathcal{R})\Pi_{t,i+1}$ for every $i \le t-1$, we obtain the following recursion:

$$\mathcal{V}_{t+1} = \frac{\varphi_t}{\varphi_{t+1}}(I - \varphi_t \mathcal{R})\mathcal{V}_t(I - \varphi_t \mathcal{R})^\top + \frac{\varphi_t^2}{\varphi_{t+1}}\Gamma_{\boldsymbol{m},\mathrm{as}}. \tag{C.89}$$

If $\varphi < 1$, the mean-value theorem gives $\varphi_t/\varphi_{t+1} - 1 = O((t+1)^{-1}) = o(\varphi_t)$. If $\varphi = 1$, we have $\varphi_t/\varphi_{t+1} = 1 + \varphi_t/C_\varphi$. Hence, in both cases, the following relations hold:

$$\frac{\varphi_t}{\varphi_{t+1}} = 1 + 2\zeta_\varphi \varphi_t + o(\varphi_t), \qquad \frac{\varphi_t^2}{\varphi_{t+1}} = \varphi_t + o(\varphi_t). \tag{C.90}$$

Moreover, the condition on C_φ when $\varphi = 1$ ensures that $\lambda_m - \zeta_\varphi > 0$. For all sufficiently large t, we also have $\|I - \varphi_t \mathcal{R}\| = 1 - \lambda_m \varphi_t$. Consequently,

$$\frac{\varphi_t}{\varphi_{t+1}}\|I - \varphi_t \mathcal{R}\|^2 = 1 - 2(\lambda_m - \zeta_\varphi)\varphi_t + o(\varphi_t).$$

Therefore, there exists a sufficiently large fixed integer T such that, for every $t \ge T$,

$$\frac{\varphi_t}{\varphi_{t+1}}\|I - \varphi_t \mathcal{R}\|^2 \le 1 - (\lambda_m - \zeta_\varphi)\varphi_t, \qquad \frac{\varphi_t^2}{\varphi_{t+1}} \le 2\varphi_t, \qquad 0 < (\lambda_m - \zeta_\varphi)\varphi_t \le 1.$$

Then, we take norms in (C.89) and obtain

$$\begin{aligned}\|\mathcal{V}_{t+1}\| \le \frac{\varphi_t}{\varphi_{t+1}}\|I - \varphi_t \mathcal{R}\|^2\|\mathcal{V}_t\| + \frac{\varphi_t^2}{\varphi_{t+1}}\|\Gamma_{\boldsymbol{m},\mathrm{as}}\| &\le \{1 - (\lambda_m - \zeta_\varphi)\varphi_t\}\|\mathcal{V}_t\| + 2\|\Gamma_{\boldsymbol{m},\mathrm{as}}\|\varphi_t \\ &\le (1 - c\varphi_t)\|\mathcal{V}_t\| + C\varphi_t,\end{aligned}$$

where $c \coloneqq \lambda_m - \zeta_\varphi > 0$ and $C \coloneqq 2(\|\Gamma_{\boldsymbol{m},\mathrm{as}}\| \vee 1) > 0$. Now, we define

$$M \coloneqq \max\left\{\max_{1 \le s \le T}\|\mathcal{V}_s\|, \frac{C}{c}\right\}.$$

Suppose that $\|\mathcal{V}_t\| \le M$ for some $t \ge T$. Then, we have

$$\|\mathcal{V}_{t+1}\| \le (1 - c\varphi_t)M + C\varphi_t = M - c\varphi_t\left(M - \frac{C}{c}\right) \le M.$$

Since $\|\mathcal{V}_T\| \le M$, we can conclude that $\|\mathcal{V}_t\| \le M$ for every $t \ge T$ by induction. Together with the finitely many indices preceding T, this proves that $\{\mathcal{V}_t\}_{t \ge 1}$ is uniformly bounded.

Expanding (C.89) and using (C.90) together with the uniform boundedness of $\mathcal{V}_t$, we further obtain

$$\mathcal{V}_{t+1} = \mathcal{V}_t - \varphi_t\left[\{\mathcal{R} - \zeta_\varphi I\}\mathcal{V}_t + \mathcal{V}_t\{\mathcal{R} - \zeta_\varphi I\} - \Gamma_{\boldsymbol{m},\mathrm{as}}\right] + o(\varphi_t). \tag{C.91}$$

Let $\mathcal{R}_\varphi \coloneqq \mathcal{R} - \zeta_\varphi I$ and $\mathcal{Q}_\varphi \coloneqq I \otimes \mathcal{R}_\varphi + \mathcal{R}_\varphi \otimes I$. Since $\mathcal{R}_\varphi$ is symmetric positive definite, $\mathcal{Q}_\varphi$ is symmetric positive definite with $\lambda_{\min}(\mathcal{Q}_\varphi) = 2(\lambda_m - \zeta_\varphi) > 0$. Define

$$s_t \coloneqq \mathrm{vec}(\mathcal{V}_t), \qquad s^\star \coloneqq \mathrm{vec}(\Sigma^l_{\boldsymbol{m},\mathrm{as}}) = \mathcal{Q}_\varphi^{-1}\,\mathrm{vec}(\Gamma_{\boldsymbol{m},\mathrm{as}}).$$

By vectorizing the recursion (C.91), we obtain

$$s_{t+1} - s^\star = (I - \varphi_t \mathcal{Q}_\varphi)(s_t - s^\star) + \varphi_t \varepsilon_t,$$

where $\varepsilon_t \to \mathbf{0}$. Therefore, for all sufficiently large t,

$$\|s_{t+1} - s^\star\| \le \{1 - 2(\lambda_m - \zeta_\varphi)\varphi_t\}\|s_t - s^\star\| + \varphi_t\|\varepsilon_t\|.$$

Applying Lemma A.3 to the preceding recursion, for every $\varphi \in (0.5, 1]$, we obtain

$$\limsup_{t \to \infty}\|s_t - s^\star\| \le \frac{1}{2(\lambda_m - \zeta_\varphi)}\limsup_{t \to \infty}\|\varepsilon_t\| = 0.$$

Therefore, $s_t \to s^\star$, which is equivalent to $\mathcal{V}_t \to \Sigma^l_{\boldsymbol{m},\mathrm{as}}$.

Moreover, since $s^\star = \mathrm{vec}(\Sigma^l_{\boldsymbol{m},\mathrm{as}})$, the equation $\mathcal{Q}_\varphi s^\star = \mathrm{vec}(\Gamma_{\boldsymbol{m},\mathrm{as}})$ is equivalent to

$$(\mathcal{R} - \zeta_\varphi I)\Sigma^l_{\boldsymbol{m},\mathrm{as}} + \Sigma^l_{\boldsymbol{m},\mathrm{as}}(\mathcal{R} - \zeta_\varphi I) = \Gamma_{\boldsymbol{m},\mathrm{as}},$$

which is precisely (4.4). Since $\mathcal{Q}_\varphi$ is positive definite, the equation $\mathcal{Q}_\varphi s = \mathrm{vec}(\Gamma_{\boldsymbol{m},\mathrm{as}})$ admits a unique solution, and hence so does (4.4). Finally, each $\mathcal{V}_t$ is positive semidefinite because it is a sum of conditional covariance matrices. Therefore, the convergence of $\mathcal{V}_t$ implies that $\Sigma^l_{\boldsymbol{m},\mathrm{as}}$ is also positive semidefinite, and condition (A.2) in Lemma A.4 is satisfied.

• **Verification of (A.3).** Let $\delta := q_g - 2 > 0$. By Assumption 3.1, we have

$$\|\nabla F(\boldsymbol{x}^\star;\xi)\| \le \|\nabla F(\boldsymbol{x}_0;\xi) - \nabla f(\boldsymbol{x}_0)\| + \|\nabla f(\boldsymbol{x}_0)\| + \Upsilon_H \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|.$$

Then by Assumption 3.2 with $q_g = 2+\delta$, we have

$$\mathbb{E}[\|\nabla F(\boldsymbol{x}_0;\xi) - \nabla f(\boldsymbol{x}_0)\|^{2+\delta}] \le C_{g,1}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^{2+\delta} + C_{g,2}.$$

Therefore, by the preceding bound and Minkowski's inequality, we obtain

$$\begin{aligned}\mathbb{E}[\|\nabla F(\boldsymbol{x}^\star;\xi)\|^{2+\delta}] &\le \left[\left\{\mathbb{E}[\|\nabla F(\boldsymbol{x}_0;\xi) - \nabla f(\boldsymbol{x}_0)\|^{2+\delta}]\right\}^{1/(2+\delta)} + \|\nabla f(\boldsymbol{x}_0)\| + \Upsilon_H\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\right]^{2+\delta} \\ &\le \left[\left\{C_{g,1}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^{2+\delta} + C_{g,2}\right\}^{1/(2+\delta)} + \|\nabla f(\boldsymbol{x}_0)\| + \Upsilon_H\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\right]^{2+\delta} < \infty. \qquad \text{(C.92)}\end{aligned}$$

Define $A_t := \sum_{i=0}^{t-1} \varphi_i^{2+\delta}\|\Pi_{t,i+1}\|^{2+\delta}$. Then, for all sufficiently large t, we have

$$A_{t+1} \le (1-\lambda_m\varphi_t)^{2+\delta}A_t + \varphi_t^{2+\delta} = (1-\lambda_m\varphi_t)^{2+\delta}A_t + \varphi_t^{2+\delta/2}\varphi_t^{\delta/2}.$$

By Lemma A.3, for $\varphi \in (0.5, 1]$, we obtain

$$\limsup_{t\to\infty} \frac{A_t}{\varphi_t^{1+\delta/2}} \le \begin{cases} \dfrac{1}{(2+\delta)\lambda_m}\limsup_{t\to\infty}\varphi_t^{\delta/2}, & \varphi \in (0.5,1), \\ \dfrac{C_\varphi}{(2+\delta)C_\varphi\lambda_m - (1+\delta/2)}\limsup_{t\to\infty}\varphi_t^{\delta/2}, & \varphi = 1, \end{cases} = 0. \qquad \text{(C.93)}$$

Here, the denominator in the second case is positive because

$$1 + \frac{\delta}{2} = \frac{2+\delta}{2} < (2+\delta)C_\varphi\lambda_m,$$

where the inequality follows from $C_\varphi\lambda_m > 1/2$. Consequently, we obtain $A_t = o(\varphi_t^{1+\delta/2})$.

For every $\epsilon > 0$, by the inequality $a^2\mathbf{1}_{\{a>\epsilon\}} \le a^{2+\delta}/\epsilon^\delta$, (C.92) and (C.93), we obtain

$$\begin{aligned}\sum_{i=0}^{t-1}\mathbb{E}[\|W_{t,i}\|^2\mathbf{1}_{\{\|W_{t,i}\|>\epsilon\}} \mid \mathcal{F}_{i-1}] &\le \frac{1}{\epsilon^\delta}\sum_{i=0}^{t-1}\mathbb{E}[\|W_{t,i}\|^{2+\delta} \mid \mathcal{F}_{i-1}] \\ &\lesssim \frac{1}{\epsilon^\delta\varphi_t^{1+\delta/2}}\sum_{i=0}^{t-1}\varphi_i^{2+\delta}\|\Pi_{t,i+1}\|^{2+\delta} = \frac{A_t}{\epsilon^\delta\varphi_t^{1+\delta/2}} \to 0. \qquad \text{(C.94)}\end{aligned}$$

Therefore, condition (A.3) is satisfied.

Combining (A.1)–(A.3), Lemma A.4 yields

$$\sum_{i=0}^{t-1} W_{t,i} = -\frac{1}{\sqrt{\varphi_t}}\sum_{i=0}^{t-1}\Pi_{t,i+1}\varphi_i(I-K^\star)(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^l_{\boldsymbol{m},\mathrm{as}}). \qquad \text{(C.95)}$$

Finally, combining (4.1) and (C.95) and applying Slutsky's theorem, we obtain

$$\frac{1}{\sqrt{\varphi_t}}(\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^l_{\boldsymbol{m},\mathrm{as}}),$$

And we complete the proof.

C.8 Proof of Proposition 4.6

Recall that $A_\varphi^\star = (I - K^\star) - \zeta_\varphi I$. By Lemma 4.1, $K^\star$ is symmetric and $I - K^\star \succ \mathbf{0}$. If $\varphi < 1$, then $\zeta_\varphi = 0$. If $\varphi = 1$, the condition on C_φ in Theorem 4.3 gives $\zeta_\varphi = 1/(2C_\varphi) < \lambda_{\min}(I - K^\star)$. Therefore, $A_\varphi^\star \succ \mathbf{0}$ under both stepsize regimes.

• **Proof of part (a).** Since $K^\star = \mathbb{E}[\widetilde{K}^\star]$, we have $\mathbb{E}[\widetilde{K}^\star - K^\star] = \mathbf{0}$. Also, we have

$$\begin{aligned}\Gamma_{\boldsymbol{g},\mathrm{as}} &= \Gamma_{\boldsymbol{m},\mathrm{as}} - (I - K^\star)\Omega\mathbb{E}[(\widetilde{K}^\star - K^\star)^\top] \\ &\quad - \mathbb{E}[\widetilde{K}^\star - K^\star]\Omega(I - K^\star)^\top + \mathbb{E}\left[(\widetilde{K}^\star - K^\star)\Omega(\widetilde{K}^\star - K^\star)^\top\right] \\ &= \Gamma_{\boldsymbol{m},\mathrm{as}} + \mathbb{E}\left[(\widetilde{K}^\star - K^\star)\Omega(\widetilde{K}^\star - K^\star)^\top\right].\end{aligned}$$

Since $\Omega \succeq \mathbf{0}$, this identity implies $\Gamma_{\boldsymbol{m},\mathrm{as}} \preceq \Gamma_{\boldsymbol{g},\mathrm{as}}$.

Subtracting (4.4) from the defining Lyapunov equation for $\Sigma^l_{\boldsymbol{g},\mathrm{as}}$, we obtain

$$\begin{aligned}&(\mathcal{R} - \zeta_\varphi I)\,(\Sigma^l_{\boldsymbol{g},\mathrm{as}} - \Sigma^l_{\boldsymbol{m},\mathrm{as}}) + (\Sigma^l_{\boldsymbol{g},\mathrm{as}} - \Sigma^l_{\boldsymbol{m},\mathrm{as}})\,(\mathcal{R} - \zeta_\varphi I) \\ &\qquad = \mathbb{E}\left[(\widetilde{K}^\star - K^\star)\Omega(\widetilde{K}^\star - K^\star)^\top\right] \succeq \mathbf{0}.\end{aligned}$$

Since $\mathcal{R} - \zeta_\varphi I \succ \mathbf{0}$, the unique solution of this Lyapunov equation admits the representation

$$\begin{aligned}&\Sigma^l_{\boldsymbol{g},\mathrm{as}} - \Sigma^l_{\boldsymbol{m},\mathrm{as}} \\ &\quad = \int_0^\infty \exp\{-(\mathcal{R} - \zeta_\varphi I)\,s\}\,\mathbb{E}\left[(\widetilde{K}^\star - K^\star)\Omega(\widetilde{K}^\star - K^\star)^\top\right]\exp\{-(\mathcal{R} - \zeta_\varphi I)\,s\}\,ds \succeq \mathbf{0},\end{aligned}$$

which completes the proof of part **(a)**.

• **Proof of part (b).** Suppose that the Newton system is solved exactly. Then, we have $K^\star = \widetilde{K}^\star = \mathbf{0}$ and $\Gamma_{\boldsymbol{m},\mathrm{as}} = \Gamma_{\boldsymbol{g},\mathrm{as}} = \Omega$. In this regime, the Lyapunov equations in (4.4) and (2.7) reduce to

$$2(1 - \zeta_\varphi)\Sigma^l_{\boldsymbol{m},\mathrm{e}} = 2(1 - \zeta_\varphi)\Sigma^l_{\boldsymbol{g},\mathrm{e}} = \Omega.$$

If $\varphi < 1$, then $\zeta_\varphi = 0$, which gives $\Sigma^l_{\boldsymbol{m},\mathrm{e}} = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = 0.5\Omega$. If $\varphi = 1$, then $\zeta_\varphi = 1/(2C_\varphi)$ and hence

$$\Sigma^l_{\boldsymbol{m},\mathrm{e}} = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = \frac{1}{2\{1 - 1/(2C_\varphi)\}}\Omega = \frac{C_\varphi}{2C_\varphi - 1}\Omega.$$

This completes the proof of part **(b)**.

• **Proof of part (c).** When $\varphi = C_\varphi = 1$, part **(b)** gives $\Sigma^l_{\boldsymbol{m},\mathrm{e}} = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = \Omega$, and $t\varphi_t \to 1$. Therefore, by Theorems 4.5 and 2.2 and Slutsky's theorem, exact online Newton, with or without gradient averaging, satisfies

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

We next establish the covariance ordering for sketched Newton with the same stepsizes. In this case, $\zeta_\varphi = 1/2$, and the condition in Theorem 4.3 implies that

$$\lambda_m = \lambda_{\min}(I - K^\star) > \frac{1}{2}, \qquad \frac{1}{2}I - K^\star \succ \mathbf{0}.$$

Moreover, by the definition of $\Gamma_{\boldsymbol{m},\mathrm{as}}$, we have

$$\begin{aligned}&\Gamma_{\boldsymbol{m},\mathrm{as}} - \left\{\left(\frac{1}{2}I - K^\star\right)\Omega + \Omega\left(\frac{1}{2}I - (K^\star)^\top\right)\right\} \\ &\quad = (I - K^\star)\Omega(I - K^\star)^\top - \left\{\left(\frac{1}{2}I - K^\star\right)\Omega + \Omega\left(\frac{1}{2}I - (K^\star)^\top\right)\right\} = K^\star\Omega(K^\star)^\top.\end{aligned}$$

Subtracting the expression inside the braces from (4.4) gives

$$\left(\frac{1}{2}I - K^\star\right)(\Sigma^l_{\boldsymbol{m},\mathrm{as}} - \Omega) + (\Sigma^l_{\boldsymbol{m},\mathrm{as}} - \Omega)\left(\frac{1}{2}I - (K^\star)^\top\right) = K^\star\Omega(K^\star)^\top \succeq \mathbf{0}.$$

Since $\frac{1}{2}I - K^\star \succ \mathbf{0}$, it follows that

$$\Sigma^l_{\boldsymbol{m},\mathrm{as}} - \Omega = \int_0^\infty \exp\left\{-\left(\frac{1}{2}I - K^\star\right)s\right\} K^\star \Omega (K^\star)^\top \exp\left\{-\left(\frac{1}{2}I - (K^\star)^\top\right)s\right\} ds \succeq \mathbf{0}.$$

Therefore, $\Omega \preceq \Sigma^l_{\boldsymbol{m},\mathrm{as}}$. Combining this inequality with parts **(a)** and **(b)** gives

$$\Omega = \Sigma^l_{\boldsymbol{g},\mathrm{e}} = \Sigma^l_{\boldsymbol{m},\mathrm{e}} \preceq \Sigma^l_{\boldsymbol{m},\mathrm{as}} \preceq \Sigma^l_{\boldsymbol{g},\mathrm{as}},$$

which completes the proof of part **(c)**.

C.9 Proof of Theorem 4.7

Since $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$, we have

$$\mathbb{E}[(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)] = (B^\star)^{-1}\nabla f(\boldsymbol{x}^\star) = \mathbf{0},$$
$$\mathbb{E}\left[(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\nabla F(\boldsymbol{x}^\star;\xi_i)^\top (B^\star)^{-1}\right] = \Omega.$$

Moreover, by (C.92) and Jensen's inequality, we further obtain

$$\mathbb{E}[\|(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i)\|^2] \le \frac{1}{\gamma_H^2}\mathbb{E}[\|\nabla F(\boldsymbol{x}^\star;\xi_i)\|^2] \le \frac{1}{\gamma_H^2}\left\{\mathbb{E}[\|\nabla F(\boldsymbol{x}^\star;\xi_i)\|^{2+\delta}]\right\}^{2/(2+\delta)} < \infty.$$

Therefore, the central limit theorem for i.i.d. random variables gives us

$$-\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) \xrightarrow{d} \mathcal{N}(\mathbf{0},\Omega).$$

Finally, by (4.2), we have

$$\sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) = -\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}(B^\star)^{-1}\nabla F(\boldsymbol{x}^\star;\xi_i) + o_p(1).$$

Combining the preceding two equations and applying Slutsky's theorem, we conclude that

$$\sqrt{t}\,(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0},\Sigma^a_{\boldsymbol{m},\mathrm{as}}),$$

where $\Sigma^a_{\boldsymbol{m},\mathrm{as}} = \Omega$. This proves (4.5) and completes the proof.

D Experimental Details and Additional Results

We provide the experimental details and additional results for Section 5. The comparisons below complement the main paper with higher-dimensional problems, full-matrix covariance errors, method-specific normal QQ plots, and estimation-error trajectories.

D.1 Detailed experimental setup

• **Regression models.** For both models, we set $\boldsymbol{x}^\star = (0, 1/(d-1), \ldots, 1)^\top$ and generate $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$. We consider $d \in \{20, 40\}$ and three covariate covariances: $\Sigma_a = I$, the equicorrelation matrix $(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$, and the Toeplitz matrix with entries $(\Sigma_a)_{jk} = \rho^{|j-k|}$, where $\rho = 0.4$. For linear regression, $y = \boldsymbol{a}^\top\boldsymbol{x}^\star + \epsilon$ with independent $\epsilon \sim \mathcal{N}(0,1)$, and

$$F(\boldsymbol{x};(\boldsymbol{a},y)) = \tfrac{1}{2}(\boldsymbol{a}^\top\boldsymbol{x} - y)^2.$$

For logistic regression, $y \mid \boldsymbol{a} \sim \mathrm{Bernoulli}\{\sigma(\boldsymbol{a}^\top\boldsymbol{x}^\star)\}$, where $\sigma(z) = (1+e^{-z})^{-1}$, and

$$F(\boldsymbol{x};(\boldsymbol{a},y)) = \log\{1+\exp(\boldsymbol{a}^\top\boldsymbol{x})\} - y\boldsymbol{a}^\top\boldsymbol{x} + \tfrac{c}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2, \qquad c = 2.$$

The penalty is centered at $\boldsymbol{x}^\star$ so that the population minimizer remains unchanged. We use this penalty only in simulations, since $\boldsymbol{x}^\star$ is unknown in practice. In both models, observations are drawn from a fixed distribution, and the gradient and Hessian are evaluated at the current iterate using the same observation. The Gaussian covariates are unbounded, and the linear sample Hessian is rank one. Thus, these experiments also examine the covariance comparisons when the uniform sample-Hessian bounds in Assumption 3.1 do not hold.

• **Methods and sketching parameters.** We compare online Newton with and without gradient averaging, using either accelerated sketching or exact solves, and include SGD. All methods start from $\boldsymbol{x}_0 = \mathbf{1}$. Newton methods use the Hessian average $B_t = \{I + \sum_{i=0}^{t-1} H_i\}/(t+1)$, which is fixed before ξ_t is drawn. The term $I/(t+1)$ keeps B_t nonsingular and vanishes as t increases. We use Kaczmarz sketches $S = \boldsymbol{e}_j$ with j uniform on $\{1, \dots, d\}$, or Gaussian sketches $S \sim \mathcal{N}(\mathbf{0}, I)$, with $\tau \in \{5, 10\}$. Each inner solve starts from $\boldsymbol{z}_{t,0} = \boldsymbol{v}_{t,0} = \mathbf{0}$. We compute the acceleration parameters from the current B_t at $t = 0, 500, 1000, \dots$ and keep them fixed between updates. The expectations in (2.3) are computed exactly by averaging over all d coordinates for Kaczmarz sketches. For Gaussian sketches, we use the same $M_{\mathrm{acc}} = 16{,}384$ Gaussian vectors at each update, generated independently of the training data and inner sketches.

• **Stepsizes.** We set $\varphi = 0.8$ for all methods and use $\eta_t = (t+1)^{-0.6}$ with $\boldsymbol{m}_0 = \boldsymbol{g}_0$ for gradient averaging. For logistic regression, the Newton stepsize is $\varphi_t = (t+1)^{-0.8}$. For linear regression, all four Newton methods use the same initial stepsizes, obtained by interpolating $\log \varphi_t$ linearly in $\log(t+1)$ through

$$(t, \varphi_t) \in \big\{(0, 0.02), (39, 0.0005), (99, 0.01), (199, 0.014427), (499, 0.020794)\big\},$$

where the final two stepsizes are rounded. For $t \geq 500$, we use $\varphi_t = 3/(t+1)^{0.8}$. Linear SGD uses $\min\{(t+1)^{-0.8}, 0.02\}$ for the first 500 updates and $(t+1)^{-0.8}$ thereafter, while logistic SGD uses $0.4/(t+1)^{0.8}$. These choices are fixed across runs. Within each problem setting, the Newton methods use the same stepsizes with and without gradient averaging, for both sketching distributions and both values of τ.

• **Runs and paired comparisons.** The two models, two dimensions, and three covariate covariance matrices give 12 problem settings. For each setting, we compare eight accelerated-sketching configurations, two exact-Newton configurations, and SGD, giving 132 configurations in total. For each configuration, we conduct $R = 200$ independent runs of $T = 10^5$ iterations. Within each setting, all methods use the same observations in each run. Sketched Newton with and without gradient averaging also uses the same sketches for the same sketching distribution and τ. The observations and sketches are generated independently.

D.2 Covariance estimation and statistical comparisons

We use the averaged iterate $\bar{\boldsymbol{x}}_t = t^{-1} \sum_{i=0}^{t-1} \boldsymbol{x}_i$. At iteration t, we compute centered sample covariances across the R runs of

$$Z^l_{r,t} = \frac{\boldsymbol{x}_t^{(r)} - \boldsymbol{x}^\star}{\sqrt{\varphi_t}}, \qquad Z^a_{r,t} = \sqrt{t}\,(\bar{\boldsymbol{x}}_t^{(r)} - \boldsymbol{x}^\star),$$

with denominator $R-1$. For the last iterate, normalization uses $C/(t+1)^{0.8}$ with the corresponding constant C for each method, rather than the modified initial stepsizes. We measure directional variance along $\mathbf{1}$. To compare estimation accuracy, we also report the empirical bias norm, MSE, and mean population objective gap using unnormalized errors. The bias norm is the norm of the mean error across runs.

For the last iterate, we compare each empirical covariance with the corresponding theoretical covariance, which generally differs from Ω. For the averaged iterate, online sketched Newton with gradient averaging and exact online Newton have limiting covariance Ω, whereas vanilla sketched Newton has an additional covariance term. We therefore compare averaged-iterate covariances both with their respective theoretical values and with Ω. We use covariance traces to compare total variance and Frobenius errors to measure differences between covariance matrices.

We construct pointwise 95% confidence intervals using 1,000 bootstrap resamples of the independent runs. Within each problem setting, we use the same resampled run indices for all methods and iterations. These intervals measure uncertainty across runs, with the theoretical covariances held fixed.

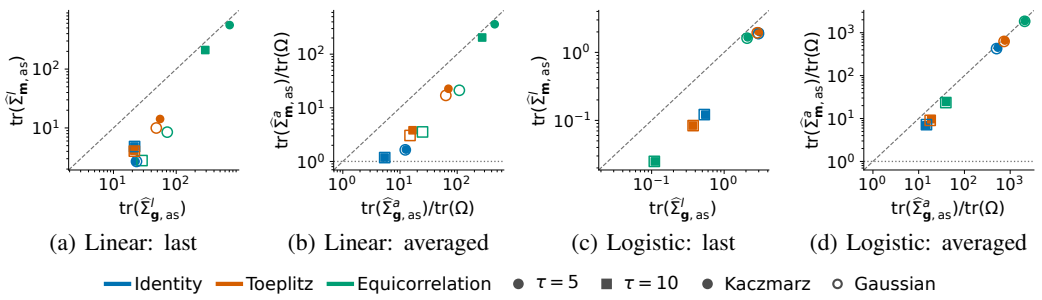


Figure 2: *Covariance comparisons for $d = 40$, $T = 10^5$, and $R = 200$. Horizontal and vertical axes show traces without and with gradient averaging, respectively; averaged-iterate traces are divided by $\mathrm{tr}(\Omega)$. Each panel contains 12 settings. Colors identify covariate covariances, circles/squares distinguish $\tau = 5/10$, and filled/open markers denote Kaczmarz/Gaussian sketches. Diagonals indicate equality; dotted horizontal lines indicate one. Open markers are enlarged to distinguish overlaps without changing their positions.*

Table 2: *Moving-average/vanilla ratios at $T = 10^5$ over $R = 200$ runs per configuration. Entries give median [minimum, maximum] over 48 sketching settings or 12 exact-solve settings. All sketching ratios are below one. Directional variance uses $\mathbf{1}$; bias, MSE, and objective gap use unnormalized errors.*

Solver	Quantity	Last iterate	Averaged iterate
Accelerated sketching	Covariance trace	0.232 [0.098, 0.787]	0.409 [0.130, 0.896]
	$\mathbf{1}$-direction variance	0.322 [0.110, 0.814]	0.427 [0.148, 0.906]
	Empirical bias norm	0.525 [0.290, 0.985]	0.818 [0.270, 0.986]
	MSE	0.234 [0.099, 0.935]	0.428 [0.192, 0.939]
	Population objective gap	0.241 [0.116, 0.935]	0.428 [0.220, 0.939]
Exact solve	Covariance trace	1.003 [0.981, 1.041]	1.003 [1.001, 1.009]

For the ratio tables, we compute each ratio within a matched setting and then summarize these point estimates across settings. For the QQ plots, we use theoretical rather than empirical standard deviations and do not subtract the empirical means.

We compute population moments analytically for linear regression and by numerical Gaussian integration for logistic regression. The acceleration parameters used in the theoretical covariances are computed at $B^\star$ using the same procedure as in the experiments. We estimate the sketch-dependent covariance terms from $M_K = 131{,}072$ independent realizations of the inner solver at $B^\star$, using sketches generated independently of the training and parameter-estimation samples. For these Monte Carlo estimates, we require both the relative Frobenius standard error and the relative standard error of the variance along $\mathbf{1}$ to be below 2%. The theoretical covariances are computed using the known data distributions and the true parameter $\boldsymbol{x}^\star$.

D.3 Additional covariance comparisons

From Figure 2, we observe smaller covariance traces with gradient averaging for both the last and averaged iterates in all 24 matched settings with $d = 40$. As in the $d = 20$ results in Figure 1(a–d), this reduction holds for both models, both sketching distributions, and all three covariate covariance matrices.

Table 2 reports additional comparisons of directional variance, bias, and objective gap, together with the ranges of the ratios across settings. We observe smaller values with gradient averaging for both the last and averaged iterates in all 48 sketching settings. Therefore, gradient averaging improves estimation accuracy as well as reducing covariance. In contrast, the covariance-trace ratios for exact online Newton remain close to one, consistent with Proposition 4.6.

We also compare the averaged-iterate covariance matrices with Ω using the relative Frobenius error $E_u = \|\widehat{\Sigma}^a_{u,\mathrm{as}} - \Omega\|_F / \|\Omega\|_F$ for $u \in \{\boldsymbol{g}, \boldsymbol{m}\}$, where $\|\cdot\|_F$ denotes the Frobenius norm. From Table 3, we observe smaller errors with gradient averaging in all 48 matched settings. Thus, gradient averaging

Table 3: *Relative Frobenius errors of averaged-iterate covariances against* Ω *at* $T = 10^5$. *Entries give median [minimum, maximum] over 12 dimension–covariate-covariance–sketch combinations. The final column aggregates settingwise error ratios, not ratios of the two median errors.*

Model	τ	E_g	E_m	E_m/E_g
Linear	5	14.979 [4.125, 604.798]	2.240 [0.343, 514.822]	0.148 [0.069, 0.851]
Linear	10	5.034 [1.824, 344.949]	0.556 [0.312, 275.337]	0.140 [0.090, 0.798]
Logistic	5	289.971 [11.986, 2867.951]	241.862 [5.051, 2590.490]	0.717 [0.421, 0.903]
Logistic	10	8.937 [2.150, 52.009]	3.813 [0.464, 32.736]	0.344 [0.203, 0.629]

Table 4: *Averaged-iterate trace divided by* $\mathrm{tr}(\Omega)$*, with pointwise 95% bootstrap intervals, for* $d = 20$*, identity covariate covariance, Gaussian sketches,* $\tau = 10$*,* $T = 10^5$*, and* $R = 200$*. Exact online Newton uses* $\boldsymbol{g}_t$.

Method	Linear regression	Logistic regression
Vanilla sketched Newton	2.725 [2.604, 2.829]	3.078 [2.925, 3.201]
Sketched Newton with gradient averaging	1.047 [1.002, 1.085]	1.248 [1.188, 1.292]
Exact online Newton	1.012 [0.968, 1.046]	1.071 [1.019, 1.112]

brings the covariance matrices closer to Ω, not only their traces. Increasing τ further reduces these errors, although some settings still have large errors at $T = 10^5$.

Table 4 reports the final covariance-trace ratios and 95% bootstrap intervals for Figure 1(e,f). We observe that gradient averaging brings the covariance traces closer to those of exact online Newton. The bootstrap intervals for sketched Newton with and without gradient averaging do not overlap.

D.4 Normal QQ plots with method-specific covariances

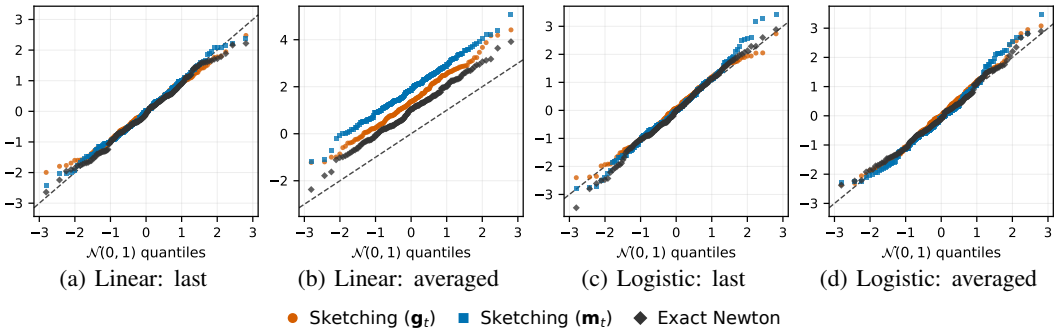
Figure 3 shows normal QQ plots using each method's theoretical covariance. We consider $d = 20$, identity covariate covariance, Gaussian sketches, and $\tau = 10$, as in Figure 1(e–h). For each method and iterate, we project $Z^q_{r,T}$ onto $\mathbf{1}$ and divide by the theoretical standard deviation $\sqrt{\mathbf{1}^\top \Sigma^q_{u,s} \mathbf{1}}$. Unlike the averaged-iterate QQ plots in the main paper, vanilla sketched Newton is normalized by its own larger covariance rather than by Ω.

From panels (a,c,d), we observe that the empirical quantiles closely follow the standard normal reference over most quantiles. Panel (b) shows a mean shift for the linear averaged iterate, which remains visible because we do not subtract the empirical means. These plots examine the normal approximation for each method using its corresponding limiting covariance.

D.5 Estimation error and computational cost

Figures 4 and 5 show bias, MSE, and population objective gap over iterations for the same representative problems. We include SGD and exact online Newton as first- and second-order baselines, respectively. Sketched Newton with and without gradient averaging uses the same observations, sketches, and outer stepsizes. From these figures, we observe smaller MSE and objective gaps with gradient averaging toward the end of the runs. Table 2 also shows smaller empirical bias in every matched sketching setting. Therefore, although the moving-average gradient is generally conditionally biased, the resulting parameter estimates have smaller bias in these experiments.

Gradient averaging requires no additional derivative evaluations and only $O(d)$ extra operations per iteration, preserving the $O(d^2)$ cost for a fixed sketching budget. Across the 48 matched sketching settings, the per-iteration time ratio with versus without gradient averaging has median 1.038 and range $[0.877, 1.564]$. Thus, the improvements in covariance and estimation accuracy require neither additional observations nor more sketching steps, with a median running-time increase of only 3.8%.



(a) Linear: last (b) Linear: averaged (c) Logistic: last (d) Logistic: averaged

Figure 3: *Method-specific normal QQ plots for $d = 20$, identity covariate covariance, Gaussian sketches, $\tau = 10$, $T = 10^5$, and $R = 200$. Horizontal axes show standard normal quantiles; vertical axes show quantiles of $\mathbf{1}^\top Z^q_{r,T}/\sqrt{\mathbf{1}^\top \Sigma^q_{u,s}\mathbf{1}}$, using each method's own theoretical covariance. Sketching uses $\boldsymbol{g}_t/\boldsymbol{m}_t$ without/with gradient averaging; exact Newton uses $\boldsymbol{g}_t$. Dashed lines indicate equality. Empirical means are retained.*

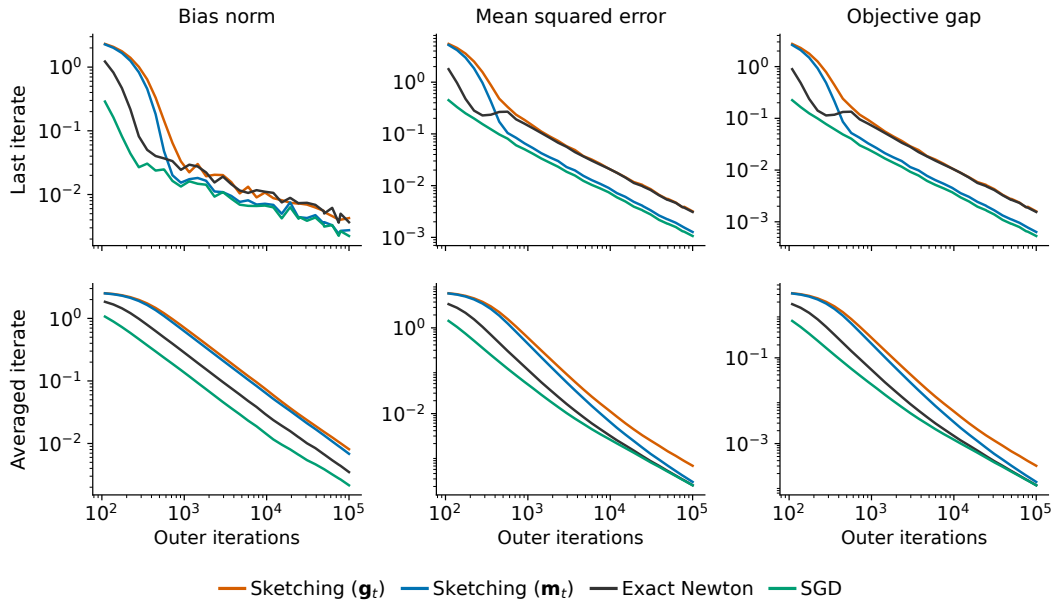


Figure 4: *Estimation errors for linear regression with $d = 20$, identity covariate covariance, Gaussian sketches, $\tau = 10$, and $R = 200$. Columns show empirical bias norm, MSE, and population objective gap; rows show last and averaged iterates. Sketching uses $\boldsymbol{g}_t/\boldsymbol{m}_t$ without/with gradient averaging; exact Newton uses $\boldsymbol{g}_t$, and SGD is the first-order baseline. All quantities use unnormalized errors.*

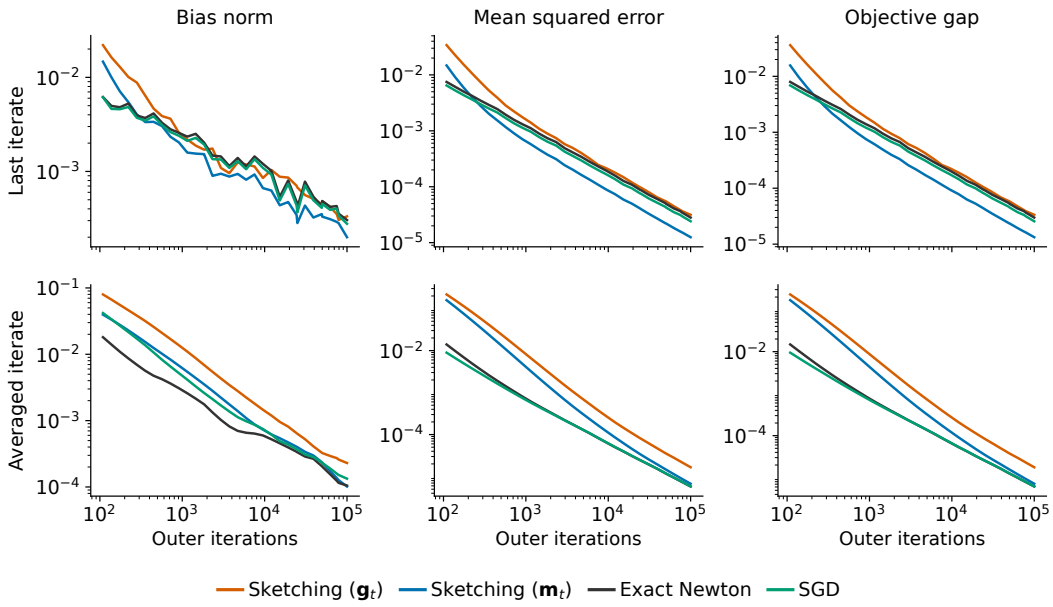


Figure 5: *Estimation errors for logistic regression with* $d = 20$*, identity covariate covariance, Gaussian sketches,* $\tau = 10$*, and* $R = 200$*. Methods, quantities, and layout follow Figure 4.*