Asymptotic Convergence Rate and Statistical Inference for Stochastic Sequential Quadratic Programming

Sen Na, Michael W. Mahoney

International Computer Science Institute
Department of Statistics, University of California, Berkeley

Abstract

We apply a stochastic sequential quadratic programming (StoSQP) algorithm to solve constrained nonlinear optimization problems, where the objective is stochastic and the constraints are in equality and deterministic. We study a fully stochastic setup, where only a single sample is available in each iteration for estimating the gradient and Hessian of the objective. We allow StoSQP to select a *random* stepsize $\bar{\alpha}_t$ adaptively, such that $\beta_t \leq \bar{\alpha}_t \leq \beta_t + \chi_t$, where β_t, $\chi_t = o(\beta_t)$ are prespecified deterministic sequences. We also allow StoSQP to solve Newton system inexactly via randomized iterative solvers, e.g., with the *sketch-and-project* method; and we do not require the approximation error of inexact Newton direction to vanish (thus, the per-iteration computational cost does not blow up). For this general StoSQP framework, we establish the asymptotic convergence rate for its *last* iterate, with the worst-case iteration complexity as a byproduct; and we perform statistical inference. In particular, under mild assumptions and with proper decaying sequences β_t, χ_t, we show that: (i) the StoSQP scheme can take at most $O(1/\epsilon^4)$ iterations to achieve ϵ-stationarity; (ii) asymptotically and almost surely, $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = O(\sqrt{\beta_t \log(1/\beta_t)}) + O(\chi_t/\beta_t)$, where $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is the primal-dual StoSQP iterate; (iii) the sequence $1/\sqrt{\beta_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)$ converges to a mean zero Gaussian distribution with a nontrivial covariance matrix. Furthermore, we establish the Berry-Esseen bound for $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ to measure quantitatively the convergence of its distribution function. We also provide a practical estimator for the covariance matrix, from which the confidence intervals (or regions) of $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ can be constructed using the iterates $\{(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\}_t$. All our theorems are validated using nonlinear problems in CUTEst test set.

1 Introduction

We consider solving constrained stochastic nonlinear optimization problems of the form:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \; f(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x};\xi)], \qquad \text{s.t.} \; c(\boldsymbol{x}) = \boldsymbol{0}, \tag{1}$$

where $f: \mathbb{R}^d \to \mathbb{R}$ is a stochastic objective function that involves a random variable $\xi \sim \mathcal{P}$, following the distribution $\mathcal{P}$, and $c: \mathbb{R}^d \to \mathbb{R}^m$ provides deterministic equality constraints. Problems of this form appear widely in a variety of applications, including optimal control (Birge, 1997), multi-stage optimization (Pflug and Pichler, 2014), PDE-constrained optimization (Rees et al., 2010), constrained maximum likelihood estimation (Onuk et al., 2015), and constrained deep neural networks

(Chen et al., 2018). In practice, the variable ξ corresponds to a data point; $f(\boldsymbol{x};\xi)$ is the loss at data point ξ when using parameter $\boldsymbol{x}$ to fit the model; and $f(\boldsymbol{x})$ is the expected loss. Deterministic constraints are also common in many real examples. They can characterize the physics of systems, encode prior model knowledge, or reduce searching complexity.

Numerous methods have been proposed for solving constrained optimization problems (Bertsekas, 1982; Nocedal and Wright, 2006). Compared to most of the existing literature on solving constrained optimization by either penalty methods, augmented Lagrangian methods, or sequential quadratic programming (SQP), we consider here a stochastic objective in Problem (1), whose exact function evaluation, gradient, and Hessian are inaccessible due to the expensive calculation of the expectation. However, their stochastic estimates are accessible by drawing samples from $\mathcal{P}$. Recently, several algorithms built on stochastic SQP (StoSQP) have been proposed for solving (1). We point the reader to Na et al. (2021a,b); Berahas et al. (2021b,a, 2022); Curtis et al. (2021b), and will review these related literature in Section 1.3. Although these works all showed global convergence for a variety of StoSQP schemes (with or without line search, with or without inequality constraints etc.), a "finer" understanding of StoSQP is missing.

By "finer" understanding of StoSQP, we mean the (local) convergence rate, the (worst-case) iteration or sample complexity, and the stationary distribution of the iteration sequence. These aspects are important since, compared to global convergence, they characterize the behavior of the iterates more precisely and demonstrate the efficiency of the algorithm. Furthermore, the primal-dual solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, especially the primal solution $\boldsymbol{x}^\star$, is the optimal model parameter that minimizes the expected loss. Performing statistical inference on the model parameter, such as constructing a confidence region for $\boldsymbol{x}^\star$ or conducting a hypothesis testing $H_0 : \boldsymbol{w}^T\boldsymbol{x}^\star = 0$ v.s. $H_1 : \boldsymbol{w}^T\boldsymbol{x}^\star \neq 0$ for a direction $\boldsymbol{w} \in \mathbb{R}^d$, is necessary for drawing any *statistically significant* conclusions. Such a statistical inference task can be performed once we understand the stationary distribution of stochastic iterates.

In this paper, we make progress towards understanding the aforementioned aspects for StoSQP schemes. We complement the existing literature by establishing the asymptotic convergence rate as well as the asymptotic normality for the *last* iterate of an Adaptive Inexact StoSQP framework, shortened as `AI-StoSQP`. As a result, we can perform statistical inference on $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ based on the iterates $\{(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\}_t$ generated by AI-StoSQP. As a byproduct, we also show the worst-case iteration complexity. By the nature of the algorithm, the iteration complexity of AI-StoSQP is consistent with the sample complexity. In the remaining of this section, we first briefly introduce AI-StoSQP and connect it with the existing schemes in Section 1.1. Then, we summarize main results and contributions in Section 1.2, followed by reviewing related literature on both constrained optimization and unconstrained optimization in Section 1.3.

1.1 Algorithm sketch

This paper studies a StoSQP framework called AI-StoSQP. As suggested by the name, we allow the scheme to *adaptively* select random stepsize $\bar{\alpha}_t$, although we require $\bar{\alpha}_t$ to be controlled by $\beta_t \leq \bar{\alpha}_t \leq \beta_t + \chi_t =: \eta_t$ with β_t, χ_t being prespecified deterministic sequences. We also allow the scheme to solve Newton systems *inexactly* via randomized iterative solvers, e.g., with the *sketch-and-project* method.

In particular, given the current primal-dual iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, AI-StoSQP performs the following three steps (detailed descriptions are presented in Section 2):

• Step 1: We generate a *single* sample $\xi_t \sim \mathcal{P}$ to estimate the gradient $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$ and the Hessian $\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_t$ (with respect to $\boldsymbol{x}$) of the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$. Then, we construct a modified

Hessian B_t based on the *averaged Hessian* $\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i$, and use B_t to form the full Hessian K_t of the Lagrangian. See (2) and (4) for the expressions of B_t and K_t.

• Step 2: We solve Newton system $K_t\widetilde{\boldsymbol{z}}_t = -\bar{\nabla}\mathcal{L}_t$ *inexactly*, by performing a *fixed*, say τ, number of iterations of the sketch-and-project method. Then, we obtain the inexact Newton direction $\boldsymbol{z}_{t,\tau} = (\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$. See (6) for the updating rule of the sketch-and-project method.

• Step 3: We adaptively select a random stepsize $\bar{\alpha}_t$ within the interval $[\beta_t, \eta_t]$, where $\beta_t, \eta_t = \beta_t + \chi_t$ are deterministic prespecified sequences. Finally, we update the iterate by $(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\lambda}_t) + \bar{\alpha}_t(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$ and repeat from Step 1.

Connections to the existing schemes. The above StoSQP framework is related to the existing schemes (Na et al., 2021a,b; Berahas et al., 2021b,a, 2022; Curtis et al., 2021b,a), but it has several important enhancements.

First, different from Na et al. (2021a,b) but following from Berahas et al. (2021b,a); Curtis et al. (2021b,a), we study a fully stochastic setup where only a single sample is generated in each iteration for estimating the objective gradient. However, those fully stochastic schemes did not estimate the objective Hessian, and simply let the modified Hessian B_t be identity matrix in the experiments. Differently, we estimate the objective Hessian (with the same sample). This extra computation is necessary for local analysis even for deterministic problems (Nocedal and Wright, 2006). We emphasize that our modified Hessian B_t is constructed based on the averaged Hessian $\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i$ excluding the t-th iteration, i.e., not based on a single noisy Hessian estimate. This choice of Hessian is more effective in practice (e.g., a recent study on stochastic Newton method has demonstrated the effectiveness of such *Hessian averaging* (Na et al., 2022)), and it is critical for our analysis. In particular, B_t is deterministic conditional on $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, and converges to $\nabla_{\boldsymbol{x}}^2\mathcal{L}^\star$ if $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converges to $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$.

Second, we solve Newton systems inexactly and randomly via the sketch-and-project method. This method was initially proposed by Gower and Richtárik (2015) for solving general linear systems. We refer to that work for other equivalent interpretations of the method and for specific matrix-free examples (e.g., randomized Kaczmarz method). Berahas et al. (2021b,a); Curtis et al. (2021a) solved Newton systems exactly. Curtis et al. (2021b) solved Newton systems inexactly but deterministically (via conjugate gradient (CG) and minimum residual (MINRES) methods). In addition, we perform a fixed number of iterations of the solver, so that the per-iteration computational cost does not blow-up. In contrast, Curtis et al. (2021b) gradually vanished the approximation error. On the other hand, we should mention that Curtis et al. (2021b) is more adaptive than AI-StoSQP; the residual of their solver is controlled by either KKT residual or feasibility residual, both of which are computed from the iterates (although converge to zero), while our iteration budget τ is given and fixed.

Third, we allow for using any random stepsize $\bar{\alpha}_t$, as long as it ensures to lie in the interval $[\beta_t, \eta_t]$. For example, the designed stepsize selection schemes in Berahas et al. (2021b,a); Curtis et al. (2021b,a) all fit in our framework.

In this paper, we call AI-StoSQP a framework for two reasons. First, AI-StoSQP is trimmed from a complete StoSQP scheme. It preserves all the features that are essential for local asymptotic analysis (i.e. $t \to \infty$). However, for a complete StoSQP scheme, one may insert another step between Step 2 and Step 3 to select suitable penalty parameter for a certain merit function. The selected parameter may kick in β_t, η_t as a multiplier. See Berahas et al. (2021b) for the usage of the ℓ_1 merit function, $f(\boldsymbol{x}) + \mu\|c(\boldsymbol{x})\|_1$. The step of penalty parameter selection is important for global analysis, while is negligible for local analysis. This is because the penalty parameter is both upper and lower bounded by deterministic thresholds, and always stabilizes for large t (under suitable conditions). See, for example, (Berahas et al., 2021b, Section 3.2.2) and (Na et al., 2021a, Lemma 4.4). Thus, the local behavior of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is fully characterized by the direction $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$ and the controlled sequences $\{\beta_t, \eta_t\}$, which are preserved by AI-StoSQP. (Note that the deterministic SQP has the same trim, where a unit stepsize is employed and the local (quadratic/superlinear) behavior is established by purely investigating the nature of (modified) Newton system). Second, AI-StoSQP does not suggest particular designs in Step 2 and Step 3, thus providing much flexibility. One can adopt different sketching matrices in Step 2, driven by the structure of the objective; and more importantly, can adopt different procedures to select $\bar{\alpha}_t$ in Step 3. For example, Berahas et al. (2021b,a); Curtis et al. (2021b,a) utilized either ℓ_1 or ℓ_2 merit functions, and projected a random quantity into the interval $[\beta_t, \eta_t]$ to obtain $\bar{\alpha}_t$. One can design a similar procedure for augmented Lagrangian merit function, as used in Na et al. (2021a,b). However, such a design would also fit in the presented framework. Thus, our analysis on AI-StoSQP is broadly applicable.

1.2 Main results and contributions

We study AI-StoSQP with *decaying* β_t and χ_t. We establish three main results informally summarized as follows. The rigorous statements are stated in Sections 3 and 4.

(a) **Iteration complexity:** AI-StoSQP can take at most $O(\epsilon^{-4})$ iterations to achieve ϵ-stationarity for the expected KKT residual.

(b) **Asymptotic convergence rate:** we have $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = O(\sqrt{\beta_t \log(1/\beta_t)}) + O(\chi_t/\beta_t)$ almost surely. Thus, if $\chi_t = O(\beta_t^{3/2})$ and β_t decays polynomially in t (as commonly employed in practice), then the error of the last iterate vanishes sublinearly locally.

(c) **Asymptotic normality:** we have $1/\sqrt{\beta_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \xrightarrow{d} \mathcal{N}(0, \Xi^\star)$, where the covariance $\Xi^\star$ depends on the sketching distribution employed in the iterative solver. Furthermore, we establish the Berry-Esseen bound to measure quantitatively how quickly the distribution function of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converges, and we provide a practical estimator for $\Xi^\star$.

Our results contribute to the literature on StoSQP mentioned in Section 1.1. Na et al. (2021a,b); Berahas et al. (2021b,a, 2022); Curtis et al. (2021b) showed global convergence for various StoSQP schemes; Curtis et al. (2021a) showed iteration complexity for an *exact* StoSQP with *constant* β_t; none of them provided a local view of StoSQP. To be specific, our main results (a)-(c) lead to the following four-fold contributions.

(1) We show that the KKT residual $\|\nabla \mathcal{L}_t\|$ converges to zero almost surely, which differs from the convergence in expectation showed in prior works. Thus, our local rate also holds almost surely.

(2) Our result (a) relies on a *non-asymptotic* convergence rate of $\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}[\|\nabla \mathcal{L}_i\|]$ for decaying β_t, while Curtis et al. (2021a) showed a similar non-asymptotic result for constant β_t. By non-asymptotic we mean the result holds for any $t \geq 0$, which distinguishes from the asymptotic results in (b) and (c) that hold for sufficiently large t (i.e., (b) and (c) are local results).

(Curtis et al. (2021a) also established $O(\epsilon^{-4})$ iteration complexity, however, we would like to mention that two works are not fully comparable. That work involves the step of penalty parameter selection for the ℓ_1 merit function, which complicates the analysis; requires extra conditions; and is inessential for our analysis in (b) and (c). In this sense, (a) is only our by-product result, and is not as complete as the one in Curtis et al. (2021a). However, Curtis et al. (2021a) can only perform finite SQP iterations since β_t in their scheme is set as the reciprocal of the iteration length.)

(3) Our results (b) and (c) show the local behavior of the last StoSQP iterate, and we provide a statistical view of the scheme. To our knowledge, such a statistical view is missing in all existing literature on constrained optimization. However, it is important in real parameter estimation problems. When we apply StoSQP for estimating the true model parameter $\boldsymbol{x}^\star$, a natural task is to infer $\boldsymbol{x}^\star$ given stochastic iterates generated by StoSQP. The results (b) and (c) precisely characterize the uncertainty of the scheme, which in each iteration includes the randomness of sample, the randomness of stepsize, and the randomness of solver. Such characterization enables the inference of $\boldsymbol{x}^\star$ (and $\boldsymbol{\lambda}^\star$) based on StoSQP iterates, and is novel to the literature.

(4) For unconstrained stochastic optimization (we review in Section 1.3), the asymptotic convergence analysis involving a random stepsize is open even for first-order methods, and involving an inexact randomized Newton direction is open for second-order methods. This paper directly achieves both for constrained stochastic optimization.

1.3 Literature review

This paper relates to prior work both on constrained optimization and on unconstrained optimization.

Constrained optimization. As mentioned, there is a growing body of literature on designing various StoSQP schemes for solving Problem (1). Compared to penalty methods and augmented Lagrangian methods, SQP methods preserve the problem structure, are more robust to initialization, and do not suffer ill-conditioning issues. Berahas et al. (2021b) designed a very first StoSQP scheme. At each iteration, the scheme adaptively selects a penalty parameter of the ℓ_1 merit function to ensure the Newton direction generates a sufficient decrease on the merit function; and then selects a random stepsize $\bar{\alpha}_t$ based on sequences β_t and $\chi_t = O(\beta_t^2)$, such that $\beta_t \leq \bar{\alpha}_t \leq \eta_t = \beta_t + \chi_t$. An alternative StoSQP scheme is designed in Na et al. (2021a), where the authors embedded stochastic line search into StoSQP to get rid of the prespecified sequences β_t, χ_t. That algorithm is more adaptive than the algorithm of Berahas et al. (2021b), while requiring a more stringent setup for line search—one has to generate batch samples with an increasing batch size in each iteration. Subsequently, Berahas et al. (2021a) designed a StoSQP to remove constraint qualification condition; Curtis et al. (2021b) designed an inexact StoSQP to solve Newton systems approximately; Na et al. (2021b) designed an active-set StoSQP to enable inequality constraints; Berahas et al. (2022) designed an accelerated StoSQP by applying variance reduction technique on finite-sum objective. All these works showed global convergence of different StoSQP schemes. In addition to these works, Oztoprak et al. (2021); Sun and Nocedal (2022) considered optimization with noisy functions. Those analyses require known, deterministic, and bounded noise, and thus they are not suited for the considered problems in (1).

Unconstrained optimization. The asymptotic rate of convergence and asymptotic normality have been established for the averaged stochastic gradient descent (ASGD) (Polyak and Juditsky, 1992). Subsequently, results on the asymptotic normality of other first-order methods and on the construction of the covariance estimator have been reported (Chen et al., 2020, 2021; Zhu et al., 2021). The vast majority of the existing works on the asymptotic analysis of SGD focused on the averaged iterate, and excluded the stepsize $1/t$ due to some technical challenges. Recently, the asymptotic analysis of stochastic Newton methods has been proposed. Bercu et al. (2020) designed a Newton scheme for logistic objective, and Boyer and Godichon-Baggioni (2020) extended to general strongly convex objectives. Compared to the literature on first-order methods, both works showed the normality of the *last* iterate with $1/t$ stepsize. However, those analyses are not applicable for Problem (1) due to the following reasons.

First, Bercu et al. (2020) and Boyer and Godichon-Baggioni (2020) studied regression problems, where the Hessian is sum of rank one matrices so that the schemes directly update the Hessian inverse via Sherman–Morrison formula. This step is not suitable for our general objective. Second, those two papers computed the Hessian inverse to have exact Newton direction. Although the Hessian inverse requires less computation for regression task, this is not the case for Problem (1). Differently, we allow the scheme to solve Newton systems inexactly and randomly. Third, those two papers only studied $1/t$ deterministic stepsize, while we enable an adaptive *random* stepsize, and the controlled sequence β_t can have a general decay rate. Our analysis demonstrates that such extension is nontrivial since the covariance $\Xi^\star$ depends on the decay rate. Fourth, the Berry-Esseen bound is missing in these two papers, which provides a quantitative understanding on the convergence of the distribution function of the iterate. Fifth, our covariance $\Xi^\star$ is more complex than those in the two papers due to the additional sources of randomness (e.g., the randomness in the stepsize and the solver). Therefore, we have to provide a computable estimator of $\Xi^\star$ to make our theory practical.

We would also like to mention a different, but important, line of literature on solving stochastic objectives (with or without constraints) via sample-average approximation (SAA) methods. See, for example, Shapiro (1993); Kleywegt et al. (2002); Ahmed and Shapiro (2008) for applying SAA on different problems and Ruszczyński and Shapiro (2003); Shapiro et al. (2014) for surveys. The SAA methods approximate a stochastic objective with some sampling schemes (such as Monte Carlo), and apply deterministic solvers for solving the approximated objective. In contrast, we consider stochastic approximation setup, where we apply a stochastic solver (i.e. StoSQP) for solving the original (stochastic) objective. The solver utilizes stochastic gradient and Hessian of the objective that are estimated by sampling.

Notation. We use boldface letters to denote column vectors, except that $\mathbf{0}$ may also denote the zero matrix. I denotes the identity matrix, whose dimension (and the dimension of $\mathbf{0}$) is clear from the context. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the spectral norm for matrices. For scalars a, b, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We use $O(\cdot)$ to denote big O notation in the usual sense. In particular, $f(x) = O(g(x))$ means $|f(x)|$ or $\|f(x)\| \leq Cg(x)$ for a constant C that is independent of x. When x is the iteration index t, $f_t = O(g_t)$ if $|f_t|$ or $\|f_t\| \leq Cg_t$ for sufficiently large t. In this case, we use $f_t = o(g_t)$ if $|f_t|/g_t \to 0$ as $t \to \infty$. For a sequence of compatible matrices $\{A_i\}_i$, we let $\prod_{k=i}^{j} A_k = A_j A_{j-1} \cdots A_i$ if $j \geq i$ and I if $j < i$ (similar for scalar sequence). We use $\bar{(\cdot)}$ to denote a random quantity that depends on a generated sample (except for the iterate). We reserve the notation $G(\boldsymbol{x})$ to denote the Jacobian matrix of constraints, that is $G(\boldsymbol{x}) = \nabla^T c(\boldsymbol{x}) = (\nabla c_1(\boldsymbol{x}), \ldots, \nabla c_m(\boldsymbol{x}))^T \in \mathbb{R}^{m \times d}$.

Structure of the paper. We introduce AI-StoSQP in Section 2; and we establish global almost sure convergence in Section 3. We establish the asymptotic convergence rate and asymptotic normality for AI-StoSQP iterates in Section 4. Experiments and conclusions are presented in Sections 5 and 6, respectively. We defer all proofs to the appendix.

2 An Adaptive Inexact StoSQP Framework

In this section, we present AI-StoSQP framework. Let $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$ be the Lagrangian function of Problem (1) where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the dual vector. Under certain constraint qualifications, a necessary condition for $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ being a local solution is the KKT conditions: $\nabla_{\boldsymbol{x}}\mathcal{L}^\star = \nabla f(\boldsymbol{x}^\star) + G^T(\boldsymbol{x}^\star)\boldsymbol{\lambda}^\star = \mathbf{0}$ and $\nabla_{\boldsymbol{\lambda}}\mathcal{L}^\star = c(\boldsymbol{x}^\star) = \mathbf{0}$.

We let $B_0 = I \in \mathbb{R}^{d\times d}$. Given the current iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, we let $c_t = c(\boldsymbol{x}_t)$ (similarly, $G_t = G(\boldsymbol{x}_t)$, $\nabla\mathcal{L}_t = \nabla\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, etc.) for shorthand. AI-StoSQP performs the following three steps.

Step 1: Estimate the gradient and Hessian. We generate a single sample $\xi_t \sim \mathcal{P}$ and compute

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

Based on this, we let $\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_t = \bar{H}_t + \sum_{i=1}^m (\boldsymbol{\lambda}_t)_i \nabla^2 c_i(\boldsymbol{x}_t)$ be the estimated Hessian of the Lagrangian. We also define the modified Hessian (used in the Newton system (3)) as

$$B_t = \frac{1}{t}\sum_{i=0}^{t-1} \bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i + \Delta_t, \tag{2}$$

where $\Delta_t = \Delta(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is a regularization term to let B_t be positive definite in the space $\{\boldsymbol{x} : G_t\boldsymbol{x} = \mathbf{0}\}$, which can simply be Levenberg-Marquardt type modification. Intuitively, we hope Δ_t vanishes when $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converges to a local solution. We emphasize that we do not use $\bar{B}_t$ notation since B_t (and Δ_t) is deterministic, conditioned on $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$. B_t does not utilize the sample ξ_t via $\bar{H}_t$. In other words, the calculation of $\bar{H}_t$ is only for preparation of the next iteration.

Step 2: Solve the Newton system. We let $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t = \bar{g}_t + G_t^T\boldsymbol{\lambda}_t$ and solve the Newton system

$$\begin{pmatrix} B_t & G_t^T \\ G_t & \mathbf{0} \end{pmatrix} \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} = -\begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}. \tag{3}$$

Instead of solving (3) exactly, we apply randomized iterative solvers introduced in Gower and Richtárik (2015), which are competitive or better than deterministic solvers in many cases (Strohmer and Vershynin, 2008). The core idea is the *sketch-and-project* step. In particular, we let

$$K_t = \begin{pmatrix} B_t & G_t^T \\ G_t & \mathbf{0} \end{pmatrix}, \quad \bar{\nabla}\mathcal{L}_t = \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}, \quad \widetilde{\boldsymbol{z}}_t = \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}, \tag{4}$$

and consider solving $K_t\boldsymbol{z}_t = -\bar{\nabla}\mathcal{L}_t$ with solution $\widetilde{\boldsymbol{z}}_t$. The j-th recursion has the form ($\boldsymbol{z}_{t,0} = \mathbf{0}$)

$$\boldsymbol{z}_{t,j+1} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z} - \boldsymbol{z}_{t,j}\|^2 \qquad \text{subject to} \quad S_{t,j}^T K_t \boldsymbol{z} = -S_{t,j}^T \bar{\nabla}\mathcal{L}_t, \tag{5}$$

where $S_{t,j} \in \mathbb{R}^{(d+m)\times q}$ is a random sketching matrix. Its dimension q can also be random. Let us denote its randomness by variable $\zeta_{t,j}$. We assume $\{\zeta_{t,j}\}_j$ are independent and identically distributed, and they are also independent of ξ_t. That is, how we generate the sketching matrix for solving the Newton system (3) is independent of how we generate samples for estimating the derivatives of f. (This is reasonable in practice.) By this setup, we have $S_{t,j} \overset{iid}{\sim} S$ and S denotes sketching distribution.

By an equivalent formula of (5) in (Gower and Richtárik, 2015, (2.7)), we derive a recursion

$$\boldsymbol{z}_{t,j+1} = \boldsymbol{z}_{t,j} - K_t S_{t,j}(S_{t,j}^T K_t^2 S_{t,j})^\dagger S_{t,j}^T (K_t\boldsymbol{z}_{t,j} + \bar{\nabla}\mathcal{L}_t), \tag{6}$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. Although (6) involves a pseudoinverse, when $q=1$ the quantity $S_{t,j}^T K_t^2 S_{t,j}$ reduces to a scalar and we then have a matrix-free update. See the randomized Kaczmarz method in Strohmer and Vershynin (2008), for example. We perform $\tau \geq 1$ iterations of (6) and let $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t) \coloneqq \boldsymbol{z}_{t,\tau}$ be the inexact Newton direction.

With (6) and defining $C_{t,j} = I - K_t S_{t,j}(S_{t,j}^T K_t^2 S_{t,j})^\dagger S_{t,j}^T K_t$, we have

$$\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t = C_{t,\tau-1}(\boldsymbol{z}_{t,\tau-1} - \widetilde{\boldsymbol{z}}_t) = \cdots = \left(\prod_{j=0}^{\tau-1} C_{t,j}\right)(\boldsymbol{z}_{t,0} - \widetilde{\boldsymbol{z}}_t) = -\left(\prod_{j=0}^{\tau-1} C_{t,j}\right)\widetilde{\boldsymbol{z}}_t \coloneqq \widetilde{C}_t \widetilde{\boldsymbol{z}}_t. \tag{7}$$

Theoretically, we can show that, if $\mathbb{E}[I - C_{t,j}]$ is invertible, then $\|\mathbb{E}[C_{t,j}]\| < 1$. This implies that $\|\mathbb{E}[\widetilde{C}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]\|$ vanishes to zero exponentially in τ, and further implies that $\|\mathbb{E}[\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t, \xi_t]\|$ vanishes to zero exponentially in τ due to (7). Moreover, if $\mathbb{E}[I - C_{t,j}]$ is positive definite, a similar guarantee holds for $\mathbb{E}[\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\| \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t, \xi_t]$. We refer to (Gower and Richtárik, 2015, Section 4) for more details, and we state necessary results later (cf. Lemma 3.4) to make our paper self-contained.

We also mention that we need τ to be large, but it is independent of t. In other words, we do not require a more and more precise Newton direction approximation as the iteration proceeds. This is in contrast to the inexact scheme in Curtis et al. (2021b), which controls the approximation error by the KKT residual and a stepsize related sequence β_t (cf. (9)). In other words, Curtis et al. (2021b) controls the inexactness more adaptively, while the approximation error also has to diminish to zero.

Step 3: Update the iterate with a random stepsize. With the inexact direction $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$, we update the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ with a random stepsize $\bar{\alpha}_t$

$$\begin{pmatrix}\boldsymbol{x}_{t+1}\\ \boldsymbol{\lambda}_{t+1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{x}_t\\ \boldsymbol{\lambda}_t\end{pmatrix} + \bar{\alpha}_t \begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}. \tag{8}$$

Certainly, $\bar{\alpha}_t$ depends on the randomness of ξ_t and $\{\zeta_{t,j}\}_j$. We allow for any adaptive scheme for selecting $\bar{\alpha}_t$, but we require $\bar{\alpha}_t$ to satisfy a sandwich condition:

$$0 < \beta_t \leq \bar{\alpha}_t \leq \eta_t \quad \text{with} \quad \eta_t = \beta_t + \chi_t. \tag{9}$$

Here, $\{\beta_t, \eta_t\}$ are deterministic prespecified upper and lower bound sequences, and χ_t is the gap. We will impose conditions on these sequences later. The particular adaptive schemes designed in Berahas et al. (2021b,a); Curtis et al. (2021b,a) all satisfy the above sandwich condition (e.g., (Berahas et al., 2021b, Lemma 3.6)); and our conditions imposed later on $\{\beta_t, \eta_t, \chi_t\}$ are satisfied by those studies as well.

It seems that upper and lower bounds in (9) reduce the difficulty for studying a random stepsize. However, as revealed by the series of works in Berahas et al. (2021b,a); Curtis et al. (2021b,a), the analysis involving a random stepsize is intrinsically different from the one for non-adaptive schemes; additional terms arise due to the adaptivity; and adaptive schemes have promising practical benefits. These differences inspire us to enable a random stepsize in the framework (under the restriction (9)). When $\chi_t = 0$, we arrive at a non-adaptive scheme.

We combine the above three steps and summarize the framework in Algorithm 1. We provide a remark to further discuss the condition (9).

Remark 2.1. We should point out that the condition (9) excludes the promising stochastic line search method, as studied in Na et al. (2021a,b). This is because there is no clear decaying trend

Algorithm 1 An Adaptive Inexact StoSQP (AI-StoSQP) Framework

1: **Input:** initial iterate $(\boldsymbol{x}_0, \boldsymbol{\lambda}_0)$, positive sequences $\{\beta_t, \eta_t = \beta_t + \chi_t\}$, integer $\tau > 0$, $B_0 = I$;
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Generate ξ_t and compute $\bar{g}_t = \nabla f(\boldsymbol{x}_t; \xi_t)$, $\bar{H}_t = \nabla^2 f(\boldsymbol{x}_t; \xi_t)$, $\bar{\nabla}^2_{\boldsymbol{x}} \mathcal{L}_t = \bar{H}_t + \sum_{i=1}^m (\boldsymbol{\lambda}_t)_i \nabla^2 c_i(\boldsymbol{x}_t)$;
4: Compute the modified Hessian $B_t = \frac{1}{t} \sum_{i=0}^{t-1} \bar{\nabla}^2_{\boldsymbol{x}} \mathcal{L}_i + \Delta_t$ to make it positive definite in the space $\{\boldsymbol{x} : G_t \boldsymbol{x} = \boldsymbol{0}\}$;
5: Generate $\{\zeta_{t,j}\}_{j=0}^{\tau-1}$ from certain distribution to formulate $\{S_{t,j}\}_{j=0}^{\tau-1}$, and apply (6) for τ times;
6: Select any random stepsize $\bar{\alpha}_t$ with $\beta_t \leq \bar{\alpha}_t \leq \eta_t$, and update the iterate as (8);
7: **end for**

for random stepsize that is selected by line search. On the other hand, line search step requires to generate more and more samples to have a precise estimation for the gradient and Hessian, which is inapplicable under our fully stochastic setup (i.e., we only generate a single sample per iteration).

To end this section, we introduce additional notation for Algorithm 1 that we will use later. As mentioned in Step 2, $\zeta_t = \{\zeta_{t,j}\}_{j=0}^{\tau-1}$ denote random variables for generating sketching matrices $\{S_{t,j}\}_{j=0}^{\tau-1}$ at the t-th iteration. We also allow the stepsize $\bar{\alpha}_t$ to depend on another random variable ψ_t in addition to ξ_t and ζ_t (cf. Line 6 in Algorithm 1). For the generated sequence $\{(\xi_t, \zeta_t, \psi_t)\}_t$, $\{\mathcal{F}_t\}_t$ is its adapted filtration; that is $\mathcal{F}_t = \sigma(\{\xi_i, \zeta_i, \psi_i\}_{i=0}^t)$, $\forall t \geq 0$, is the σ-algebra generated by the randomness of $\{\xi_i, \zeta_i, \psi_i\}_{i=0}^t$. We also let

$$\mathcal{F}_{t-2/3} = \sigma(\{\xi_i, \zeta_i, \psi_i\}_{i=0}^{t-1} \cup \xi_t), \qquad \mathcal{F}_{t-1/3} = \sigma(\{\xi_i, \zeta_i, \psi_i\}_{i=0}^{t-1} \cup \xi_t \cup \zeta_t),$$

and have $\mathcal{F}_{t-1} \subseteq \mathcal{F}_{t-2/3} \subseteq \mathcal{F}_{t-1/3} \subseteq \mathcal{F}_t$. For consistency, $\mathcal{F}_{-1}$ is the trivial σ-algebra. Algorithm 1 has a generating process as follows: given $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, we first generate ξ_t to estimate the gradient $\bar{g}_t$ and Hessian $\bar{H}_t$ and derive $\mathcal{F}_{t-2/3}$; then we generate ζ_t to obtain inexact Newton direction and derive $\mathcal{F}_{t-1/3}$; then we *may* generate ψ_t to select the stepsize $\bar{\alpha}_t$ and derive $\mathcal{F}_t$. For some stepsize selection procedures like Berahas et al. (2021b), $\bar{\alpha}_t$ is fully determined by ξ_t, ζ_t so that no random variable ψ_t has to be generated. In this case, we have $\mathcal{F}_{t-1/3} = \mathcal{F}_t$. By our setup, it is easy to see that the random quantities in Algorithm 1 have the following recursion

$$\begin{aligned} \sigma(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \cup \sigma(B_t) \subseteq \mathcal{F}_{t-1}, \qquad & \sigma(\bar{g}_t) \cup \sigma(\bar{H}_t) \cup \sigma(\widetilde{\Delta} \boldsymbol{x}_t, \widetilde{\Delta} \boldsymbol{\lambda}_t) \subseteq \mathcal{F}_{t-2/3}, \\ \sigma(\bar{\Delta} \boldsymbol{x}_t, \bar{\Delta} \boldsymbol{\lambda}_t) \subseteq \mathcal{F}_{t-1/3}, \qquad & \sigma(\bar{\alpha}_t) \cup \sigma(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1}) \cup \sigma(B_{t+1}) \subseteq \mathcal{F}_t. \end{aligned}$$

We also let $(\Delta \boldsymbol{x}_t, \Delta \boldsymbol{\lambda}_t)$ be the solution of (3) but replace $\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t$ with $\nabla_{\boldsymbol{x}} \mathcal{L}_t$.

3 Global Almost Sure Convergence

In this section, we establish an almost sure convergence for the KKT residual $\nabla \mathcal{L}_t$ of Algorithm 1, under standard assumptions. This type of convergence guarantee differs from the convergence in expectation, that is $\liminf_{t \to \infty} \mathbb{E}[\|\nabla \mathcal{L}_t\|] = 0$, established in Berahas et al. (2021b,a, 2022); Curtis et al. (2021b). On the other hand, those convergence in expectation results may be reformed to almost sure convergence by applying our following analysis on the ℓ_1 (or ℓ_2) merit function.

We utilize an exact augmented Lagrangian merit function to show the convergence. This function has the form

$$\mathcal{L}_{\mu,\nu}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) + \frac{\mu}{2} \|c(\boldsymbol{x})\|^2 + \frac{\nu}{2} \|\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\|^2 \qquad \text{for } \mu, \nu > 0. \tag{10}$$

The augmented Lagrangian (10) was initially proposed by Pillo and Grippo (1979), and it has been adopted in SQP schemes for different problems (Na et al., 2021c; Na, 2021). The advantage of this augmented Lagrangian is that it is differentiable, and the inner product between Newton direction and the gradient $\nabla\mathcal{L}_{\mu,\nu}$ with properly chosen μ, ν is sufficiently negative to endure inexactness. By a simple calculation, we have

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu} \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu} \end{pmatrix} = \begin{pmatrix} I + \nu\nabla^2_{\boldsymbol{x}}\mathcal{L} & \mu G^T \\ \nu G & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L} \\ c \end{pmatrix} \tag{11}$$

(the evaluation point has been suppressed). We will show that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ of Algorithm 1 decreases $\mathcal{L}_{\mu,\nu}$ in expectation in each step, and finally converges to a KKT point.

3.1 Assumptions and preliminary results

We begin by stating assumptions for showing global convergence.

Assumption 3.1. We assume f, c are twice continuously differentiable, and there exists a convex compact set $\mathcal{X} \times \Lambda$ that contains the iterates $\{(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\}_t$ generated by Algorithm 1. We also assume $\nabla^2\mathcal{L}$ and ∇f are Υ_L-Lipschitz continuous in $\mathcal{X} \times \Lambda$. That is, for any $(\boldsymbol{x}, \boldsymbol{\lambda}), (\boldsymbol{x}', \boldsymbol{\lambda}') \in \mathcal{X} \times \Lambda$,

$$\|\nabla^2\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) - \nabla^2\mathcal{L}(\boldsymbol{x}', \boldsymbol{\lambda}')\| \le \Upsilon_L\|(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{\lambda} - \boldsymbol{\lambda}')\|, \quad \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \le \Upsilon_L\|\boldsymbol{x} - \boldsymbol{x}'\|. \tag{12}$$

Furthermore, we assume that G_t has full row rank, and that Δ_t is chosen such that B_t satisfies $\|B_t\| \le \Upsilon_B$, and $\boldsymbol{x}^T B_t \boldsymbol{x} \ge \gamma_{RH}\|\boldsymbol{x}\|^2$, for any $\boldsymbol{x} \in \{\boldsymbol{x} : G_t\boldsymbol{x} = \boldsymbol{0}\}$, for some constants $0 < \gamma_{RH} \le 1 \le \Upsilon_B$.

Assumption 3.1 is a standard assumption in SQP analysis (Bertsekas, 1982; Nocedal and Wright, 2006). The compactness of $\mathcal{X} \times \Lambda$ ensures that $\mathcal{L}_{\mu,\nu}$ is lower bounded, and there exists a constant $\Upsilon_u \ge 1$ such that

$$\|\nabla^2\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\| \vee \|\nabla\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\| \vee \|\nabla f(\boldsymbol{x})\| \le \Upsilon_u, \qquad \text{for any } (\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{X} \times \Lambda. \tag{13}$$

The convexity of $\mathcal{X} \times \Lambda$ ensures the Taylor expansion of $\mathcal{L}_{\mu,\nu}$ at any iterate. The Lipschitz continuity of $\nabla^2\mathcal{L}$ relaxes the thrice continuous differentiability of f, c, as assumed in (Bertsekas, 1982, Chapter 4.3). Note that imposing conditions on $\nabla^2\mathcal{L}$ is equivalent to imposing the same conditions on its components $\nabla^2_{\boldsymbol{x}}\mathcal{L}$ and $G(\boldsymbol{x})$, although the Lipschitz constant may be different. The Lipschitz continuity of ∇f is implied by the compactness of the set $\mathcal{X}$, while we express it out explicitly with constant Υ_L. With the above setup and a simple calculation, we know from (11) that $\nabla\mathcal{L}_{\mu,\nu}$ is also Lipschitz continuous in $\mathcal{X} \times \Lambda$. That is, for any $(\boldsymbol{x}, \boldsymbol{\lambda}), (\boldsymbol{x}', \boldsymbol{\lambda}') \in \mathcal{X} \times \Lambda$,

$$\begin{aligned} &\|\nabla\mathcal{L}_{\mu,\nu}(\boldsymbol{x}, \boldsymbol{\lambda}) - \nabla\mathcal{L}_{\mu,\nu}(\boldsymbol{x}', \boldsymbol{\lambda}')\| \\ &\quad \le \{(1 + (2\nu + \mu)\Upsilon_u)\,\Upsilon_u + (2\nu + \mu)\Upsilon_u\Upsilon_L\}\,\|(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{\lambda} - \boldsymbol{\lambda}')\| =: \Upsilon_{\mu,\nu}\|(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{\lambda} - \boldsymbol{\lambda}')\|. \end{aligned} \tag{14}$$

Assumption 3.1 also assumes that G_t has full row rank, which is a common constraint qualification to ensure the uniqueness of the dual solution. By the compactness of $\mathcal{X}$, we have $G_tG_t^T \succeq \gamma_G I$ for some constant $0 < \gamma_G \le 1$. Together with the conditions on B_t, we know that the Newton system (3) has a unique solution (cf. (Nocedal and Wright, 2006, Lemma 16.1)), and that the KKT matrix inverse K_t^{-1} is uniformly bounded. We denote by $\|K_t^{-1}\| \le \Upsilon_K$ for $\Upsilon_K \ge 1$.

We also impose the bounded moment condition on stochastic gradient $\bar{g}_t$ and Hessian $\bar{H}_t$.

Assumption 3.2. We assume $\mathbb{E}[\bar{g}_t \mid \boldsymbol{x}_t] = \nabla f_t$, $\mathbb{E}[\bar{H}_t \mid \boldsymbol{x}_t] = \nabla^2 f_t$, and assume following moment conditions *when needed*: for a constant $\Upsilon_m \geq 1$,

$$\text{gradient} \quad (\text{bounded 2nd moment}): \quad \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^2 \mid \boldsymbol{x}_t] \leq \Upsilon_m, \tag{15a}$$

$$(\text{bounded 3rd moment}): \quad \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^3 \mid \boldsymbol{x}_t] \leq \Upsilon_m, \tag{15b}$$

$$(\text{bounded 4th moment}): \quad \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^4 \mid \boldsymbol{x}_t] \leq \Upsilon_m, \tag{15c}$$

and

$$\text{Hessian} \quad (\text{bounded 2nd moment}): \quad \mathbb{E}[\|\bar{H}_t - \nabla^2 f_t\|^2 \mid \boldsymbol{x}_t] \leq \Upsilon_m, \tag{15d}$$

$$(\text{bounded 2nd moment}): \quad \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla^2 f(\boldsymbol{x};\xi)\|^2] \leq \Upsilon_m. \tag{15e}$$

We write $\mathbb{E}[\cdot \mid \boldsymbol{x}_t]$ to express the conditional variable clearly. It can also be written as $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, which means that the expectation is taken over randomness of $\xi_t \sim \mathcal{P}$. For conditions (15), we do not require all of them at once, but we impose them step by step. In fact, (15c) implies (15b), which implies (15a). By the compactness of $\mathcal{X}$ (which implies the boundedness of $\|\nabla^2 f_t\|$), (15e) implies (15d) although for a different constant.

We mention that (15e) is imposed even for asymptotic analysis of ASGD (Chen et al., 2020). It ensures the Lipschitz continuity of the mapping $\boldsymbol{x} \to \mathbb{E}[\nabla f(\boldsymbol{x};\xi)\nabla^T f(\boldsymbol{x};\xi)]$ (as proved in (C.9)). See (Chen et al., 2020, Assumption 3.2(2) and Lemma 3.1) for the discussion. We note that (Boyer and Godichon-Baggioni, 2020, Assumption (A1c)) directly assumed the mapping is continuous. We prefer to adopt (15e) due to two reasons. First, (15e) has a clear connection to (15d). It is satisfied by a variety of objective functions, such as least squares regression $f(\boldsymbol{x};\xi) = (\xi_y - \boldsymbol{x}^T\xi_{\boldsymbol{x}})^2/2$ and logistic regression $f(\boldsymbol{x};\xi) = \log\left(1 + \exp\left(\boldsymbol{x}^T\xi_{\boldsymbol{x}}\right)\right) - \xi_y \cdot \boldsymbol{x}^T\xi_{\boldsymbol{x}}$, where $\xi = (\xi_y, \xi_{\boldsymbol{x}})$ is the response-feature pair with $\xi_y \in \mathbb{R}, \xi_{\boldsymbol{x}} \in \mathbb{R}^d$ for least squares regression and $\xi_y \in \{0,1\}, \xi_{\boldsymbol{x}} \in \mathbb{R}^d$ for logistic regression, as long as $\xi_{\boldsymbol{x}}$ has bounded 4-th moment. This condition is more intuitive and checkable than the continuity of the mapping. Second, although our asymptotic normality and Berry-Esseen bound also hold under the continuity condition on the mapping, (15e) further enables us to design a practical covariance estimator with an explicit convergence rate (cf. Lemma 4.12). The covariance matrix estimation is not considered in Boyer and Godichon-Baggioni (2020).

In this section, we only require (15a) for studying the convergence of $\nabla\mathcal{L}_t$ and require (15d) for studying the convergence of K_t. In the next section, we require higher order moment condition for conducting asymptotic analysis.

In terms of the distribution of sketching matrices of randomized iterative solvers, we need the following assumption.

Assumption 3.3. For any $t \geq 0$, we assume that the sketching matrices $S_{t,j} \overset{iid}{\sim} S$ with distribution S satisfies

$$\mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \succeq \gamma_S I \qquad \text{for } \gamma_S > 0.$$

Assumption 3.3 is required by (Gower and Richtárik, 2015, Theorem 4.6) to ensure that the iterates generated by randomized solvers converge in expectation. It can be easily verified for some sketching matrices. For example, in randomized Kaczmarz method, $S = \boldsymbol{e}_i$ with equal probability where $\boldsymbol{e}_i$ is the i-th canonical basis of $\mathbb{R}^{d+m}$. Then, we have

$$\mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \succeq \frac{\mathbb{E}[K_t S S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]}{\max_j [K_t^2]_{j,j}} = \frac{K_t^2}{(d+m)\cdot \max_j [K_t^2]_{j,j}} \succeq \frac{I}{(d+m)\kappa(K_t^2)},$$

where $[K_t^2]_{j,j}$ denotes the (j,j)-entry of K_t^2 and $\kappa(K_t^2)$ denotes the condition number of K_t^2 (it is independent of t due to the compactness of $\mathcal{X} \times \Lambda$). In this case, Assumption 3.3 is implied by Assumption 3.1. Assumption 3.3 directly leads to the following result.

Lemma 3.4 (Guarantees of randomized solvers)**.** Under Assumption 3.3, the following statements hold for all $t \geq 0$.
(a): $0 < \gamma_S \leq 1$.
(b): $\mathbb{E}[\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t, \xi_t] = -(I - \mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t])^\tau \widetilde{\boldsymbol{z}}_t =: C_t \widetilde{\boldsymbol{z}}_t$, and $\|C_t\| \leq \rho^\tau$ with $\rho = 1 - \gamma_S < 1$.
(c): $\mathbb{E}[\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\|^2 \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t, \xi_t] \leq \rho^\tau \|\widetilde{\boldsymbol{z}}_t\|^2$.

Proof. We know that

$$\gamma_S \leq \|\mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]\| \leq \mathbb{E}[\|K_t S(S^T K_t^2 S)^\dagger S^T K_t\| \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \leq 1,$$

where the second inequality is by Jensen's inequality; the third inequality is since $K_t S(S^T K_t^2 S)^\dagger S^T K_t$ is a projection matrix. This shows (a). (b) follows from (7) and the independence among $\{\zeta_{t,j}\}_j$. (c) is proved by (Gower and Richtárik, 2015, Theorem 4.6). ∎

3.2 Almost sure convergence

We now set the stage to establish the global almost sure convergence. The first result shows that $(\Delta\boldsymbol{x}_t, \Delta\boldsymbol{\lambda}_t)$ is a descent direction of $\mathcal{L}_{\mu,\nu}^t = \mathcal{L}_{\mu,\nu}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ if μ is sufficiently large and ν is sufficiently small. A similar result has been shown in (Na et al., 2021c, (5.15)), while we provide an alternative proof that is more straightforward than there.

Lemma 3.5. Under Assumption 3.1, we have

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \leq -\frac{\nu\gamma_G}{8} \left\{ \left\| \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_t \\ c_t \end{pmatrix} \right\|^2 \right\},$$

provided that

$$\nu \leq \frac{\gamma_{RH}}{5(\Upsilon_B + \Upsilon_u)^2} \quad \text{and} \quad \mu\nu \geq \frac{9}{\gamma_G}. \tag{16}$$

Proof. See Appendix B.1. ∎

Combining Lemmas 3.4 and 3.5, we are able to show the following recursion.

Lemma 3.6. Under Assumptions 3.1, 3.2(15a), 3.3, and suppose that (μ, ν) satisfies (16) and $\rho^\tau \leq \nu\gamma_G/(16\mu\Upsilon_u)$, we have

$$\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}_{\mu,\nu}^t - \frac{\nu\gamma_G\beta_t}{16}\|\nabla\mathcal{L}_t\|^2 + \widetilde{\Upsilon}_{\mu,\nu}(\chi_t + \eta_t^2)$$

where $\widetilde{\Upsilon}_{\mu,\nu} = 4\mu\Upsilon_K\Upsilon_u^2(\sqrt{\Upsilon_m} \vee \Upsilon_u) \vee 8\Upsilon_{\mu,\nu}\Upsilon_K^2(\Upsilon_u^2 \vee \Upsilon_m)$.

Proof. See Appendix B.2. ∎

A similar recursion to Lemma 3.6 for the ℓ_1 merit function was established in Berahas et al. (2021b). By Lemma 3.6, we are able to show an almost sure convergence of $\|\nabla\mathcal{L}_t\|$.

Theorem 3.7 (Convergence of the KKT residual)**.** Consider Algorithm 1 under Assumptions 3.1, 3.2(15a), 3.3. Suppose τ satisfies

$$\tau \geq \frac{12\log 2 + 5\log(\Upsilon_B + \Upsilon_u) + 2\log(1/(\gamma_{RH}\gamma_G))}{\log(1/(1-\gamma_S))}, \tag{17}$$

and $\{\beta_t, \eta_t = \beta_t + \chi_t\}$ satisfies

$$\sum_{t=0}^{\infty} \beta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \qquad \sum_{t=0}^{\infty} \chi_t < \infty, \tag{18}$$

then we have $\|(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)\| \to 0$ and $\|\nabla \mathcal{L}_t\| \to 0$ as $t \to \infty$ almost surely.

Proof. See Appendix B.3. ∎

Based on Theorem 3.7, we provide a corollary about the worst-case iteration complexity of Algorithm 1, which recovers the result (a) introduced in Section 1.2. We do not need the condition (18) since we provide the convergence rate of averaged expected KKT residual, $\sum_{j=0}^{t-1} \mathbb{E}[\|\nabla \mathcal{L}_j\|^2]/t$, instead of providing the convergence rate of $\|\nabla \mathcal{L}_t\|$.

Corollary 3.8. Consider Algorithm 1 under Assumptions 3.1, 3.2(15a), 3.3. Suppose τ satisfies (17), $\beta_t = (t+1)^{-a}$, $\chi_t = (t+1)^{-b}$ where $a \in (0,1)$ and $a < b$, we define $\mathcal{T}_\epsilon = \inf_t \{t \geq 1 : \mathbb{E}[\|\nabla \mathcal{L}_t\|] \leq \epsilon\}$. Then, we have

$$\mathcal{T}_\epsilon = O\left(\epsilon^{-\frac{2}{a \wedge (1-a) \wedge (b-a)}}\right).$$

In particular, if $b = 2a$ (as used in Berahas et al. (2021b)), then $\mathcal{T}_\epsilon = O(\epsilon^{-2/\{a \wedge (1-a)\}})$. In this case, the best complexity is achieved by $a = 1/2$ and we have $\mathcal{T}_\epsilon = O(\epsilon^{-4})$.

Proof. See Appendix B.4. ∎

Before studying the convergence of the Hessian matrix K_t, we provide two remarks for Theorem 3.7 and Corollary 3.8.

Remark 3.9. Theorem 3.7 shows that all limiting points of the iteration sequence are stationary, while it does not suggest that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ always converges to a stationary point or a local solution. This is similar to deterministic SQP guarantees (Nocedal and Wright, 2006, Theorem 18.3). The convergence of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is equivalent to the convergence of $\nabla \mathcal{L}_t$ for problems where the objective is strongly convex and constraints are affine. For nonlinear problems, if $\mathcal{X} \times \Lambda$ has a single stationary point inside, then $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ also converges to that point. Furthermore, if that point is a strict local solution (by chance), then $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converges to that strict local solution. To facilitate our analysis on the convergence rate of StoSQP, we assume in Assumption 3.11 that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ for a *strict* local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, whose Hessian $K^\star = \nabla^2 \mathcal{L}^\star$ is invertible (as implied by (i) $\nabla_{\boldsymbol{x}}^2 \mathcal{L}^\star$ is positive definite in $\{\boldsymbol{x} : G^\star \boldsymbol{x} = \boldsymbol{0}\}$; (ii) $G^\star$ has full row rank). Note that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ is equivalent to assuming that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ jumps into a small neighborhood of $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ for a sufficiently large t. This is because $\|(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)\| \vee \|\nabla \mathcal{L}_t\| \to 0$ implies that there exists an attraction neighborhood around $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$—once $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ lies in the neighborhood, then all the following iterates stay in the neighborhood. We emphasize that targeting on an iteration sequence that converges to a *strict* local solution is a routine setup for convergence rate analysis of deterministic SQP (e.g., Bertsekas (1982); Lucidi (1990); Sun and Yuan (2006); Liu and Yuan (2011)); and it is reasonable for StoSQP as well, given the convergence guarantee in Theorem 3.7 matches the one of deterministic SQP, up to a a zero-probability event.

Remark 3.10. Corollary 3.8 complements the existing iteration complexity result in Curtis et al. (2021a), in the sense that it shows the complexity with decaying sequences $\{\beta_t, \eta_t\}$ while (Curtis et al., 2021a, Theorem 2) showed the complexity with constant sequences $\beta_t = \beta$ and $\eta_t = \beta + \beta^2$. Our analysis allows one to set χ_t to be different higher orders of β_t, instead of always setting $\chi_t = O(\beta_t^2)$ as did in that work. Our result indicates that, to achieve ϵ stationarity, $O(\epsilon^{-4})$ iterations are required at most with $\beta_t = 1/\sqrt{t+1}$ and $\chi_t = 1/(t+1)$. The complexity of $O(\epsilon^{-4})$ matches that work as well. We should mention that Curtis et al. (2021a) studied a more practical algorithm that involves the step of merit parameter selection. We remove such a step since we focus on (local) asymptotic convergence rate. In addition, both Corollary 3.8 and Curtis et al. (2021a) differ from our main analyses in Section 4. Here, we provide a *non-asymptotic* convergence rate for the *averaged expected* KKT residual (i.e., the result holds for any t); while in Section 4 we will provide an *asymptotic almost sure* convergence rate for the *last* iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ (i.e., the result holds for sufficiently large t).

Finally, we study the convergence of K_t defined in (4). We need the following (local) assumption.

Assumption 3.11. We assume $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ almost surely to a strict local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ that satisfies (i) $G^\star$ has full row rank, (ii) $\nabla_{\boldsymbol{x}}^2 \mathcal{L}^\star$ is positive definite in the space $\{\boldsymbol{x} : G^\star \boldsymbol{x} = \boldsymbol{0}\}$. Furthermore, we assume $\|\Delta_t\| \leq \omega_t$ with a deterministic sequence $w_t \to 0$.

Assumption 3.11 assumes that the iteration sequence converges to a local solution. By Remark 3.9, it enables us to study the local rate of the iterates. Assumption 3.11 also assumes that the Hessian modification Δ_t vanishes, which is standard in deterministic SQP analysis. In fact, a naive way to choose Δ_t is based on checking the sign of index of inertia of the KKT matrix K_t. When the inertia has correct sign, i.e., d positive eigenvalues and m negative eigenvalues, we do not need any modification. By this way, we have $\omega_t = 0$ for all sufficiently large t.

We now show that K_t converges to $K^\star = \nabla^2 \mathcal{L}^\star$ in the next theorem.

Theorem 3.12 (Convergence of the Hessian matrix)**.** Suppose Assumptions 3.2(15d) and 3.11 hold, and $\nabla^2 \mathcal{L}$ is Υ_L-Lipschitz continuous (cf. (12)). We have $K_t \to K^\star$ where K_t is defined in (4).

Proof. See Appendix B.5. ∎

4 Asymptotic Convergence Rate and Normality

In this section, we provide an explicit convergence rate for the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ when $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converges to a strict local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$. Our main results are presented in Theorems 4.8 and 4.11, which certify the results (b) and (c) introduced in Section 1.2. These results extend the results of Section 3, where we showed that the KKT residual $\|\nabla \mathcal{L}_t\| \to 0$ and the Hessian $K_t \to K^\star$. We summarize all analyses in Sections 3 and 4, and present a complete result of AI-StoSQP in Theorem 4.14.

To study the convergence rate, we first establish an iteration recursion in Section 4.1. The recursion consists of three terms, which we analyze in Section 4.2. Then, we establish the asymptotic convergence rate and asymptotic normality in Section 4.3. To preform statistical inference of $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ in practice, we propose a covariance estimator in Section 4.4. All global and local results are then summarized in Section 4.5 with a discussion on the conditions on the sequences $\{\beta_t, \chi_t\}_t$.

4.1 Iteration recursion

From a high-level view, we can show that Algorithm 1 generates a complicated stochastic process

$$\begin{pmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star \end{pmatrix} = (1-\bar{\alpha}_t) \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \boldsymbol{\theta}_{\boldsymbol{x}}^t \\ \boldsymbol{\theta}_{\boldsymbol{\lambda}}^t \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \boldsymbol{\delta}_{\boldsymbol{x}}^t \\ \boldsymbol{\delta}_{\boldsymbol{\lambda}}^t \end{pmatrix}, \tag{19}$$

where $(\boldsymbol{\theta}_{\boldsymbol{x}}^t, \boldsymbol{\theta}_{\boldsymbol{\lambda}}^t)$ is a martingale difference with $\mathbb{E}[(\boldsymbol{\theta}_{\boldsymbol{x}}^t, \boldsymbol{\theta}_{\boldsymbol{\lambda}}^t) \mid \mathcal{F}_{t-1}] = \mathbf{0}$, brought by generating a sample $\xi_t \sim \mathcal{P}$ and solving Newton systems via randomized solvers; and $(\boldsymbol{\delta}_{\boldsymbol{x}}^t, \boldsymbol{\delta}_{\boldsymbol{\lambda}}^t)$ is the remaining error term. Compared to the studies on unconstrained ASGD and stochastic Newton (Polyak and Juditsky, 1992; Chen et al., 2020; Bercu et al., 2020; Boyer and Godichon-Baggioni, 2020), the two main components of AI-StoSQP—adaptivity and inexactness—kick in all terms in the recursion (19), and they lead to a much more challenging analysis. In short, we have to deal with a random stepsize $\bar{\alpha}_t$, and $\boldsymbol{\theta}^t = (\boldsymbol{\theta}_{\boldsymbol{x}}^t, \boldsymbol{\theta}_{\boldsymbol{\lambda}}^t)$ and $\boldsymbol{\delta}^t = (\boldsymbol{\delta}_{\boldsymbol{x}}^t, \boldsymbol{\delta}_{\boldsymbol{\lambda}}^t)$ also contain the approximation error from randomized solvers.

We formalize the recursion (19) in the following lemma. To ease notation, we let $\varphi_t = (\beta_t + \eta_t)/2$.

Lemma 4.1. Algorithm 1 generates a sequence of iterates $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ that can be expressed as

$$\begin{pmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star \end{pmatrix} = \mathcal{I}_{1,t} + \mathcal{I}_{2,t} + \mathcal{I}_{3,t}$$

where

$$\mathcal{I}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j (I + C^\star)\} \varphi_i \begin{pmatrix} \boldsymbol{\theta}_{\boldsymbol{x}}^i \\ \boldsymbol{\theta}_{\boldsymbol{\lambda}}^i \end{pmatrix}, \tag{20a}$$

$$\mathcal{I}_{2,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j (I + C^\star)\} (\bar{\alpha}_i - \varphi_i) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_i \\ \bar{\Delta}\boldsymbol{\lambda}_i \end{pmatrix}, \tag{20b}$$

$$\mathcal{I}_{3,t} = \prod_{i=0}^{t} \{I - \varphi_i (I + C^\star)\} \begin{pmatrix} \boldsymbol{x}_0 - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^\star \end{pmatrix} + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j (I + C^\star)\} \varphi_i \begin{pmatrix} \boldsymbol{\delta}_{\boldsymbol{x}}^i \\ \boldsymbol{\delta}_{\boldsymbol{\lambda}}^i \end{pmatrix}, \tag{20c}$$

and (see Lemma 3.4(b) for the definition of C_i)

$$C^\star = -(I - \mathbb{E}[K^\star S (S^T (K^\star)^2 S)^\dagger S^T K^\star])^\tau, \tag{21a}$$

$$\begin{pmatrix} \boldsymbol{\theta}_{\boldsymbol{x}}^i \\ \boldsymbol{\theta}_{\boldsymbol{\lambda}}^i \end{pmatrix} = -(I + C_i) K_i^{-1} \begin{pmatrix} \bar{g}_i - \nabla f_i \\ \mathbf{0} \end{pmatrix} + \left\{ \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_i \\ \bar{\Delta}\boldsymbol{\lambda}_i \end{pmatrix} - (I + C_i) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_i \\ \widetilde{\Delta}\boldsymbol{\lambda}_i \end{pmatrix} \right\}, \tag{21b}$$

$$\begin{pmatrix} \boldsymbol{\delta}_{\boldsymbol{x}}^i \\ \boldsymbol{\delta}_{\boldsymbol{\lambda}}^i \end{pmatrix} = -(I + C_i) \left\{ (K^\star)^{-1} \begin{pmatrix} \boldsymbol{\psi}_{\boldsymbol{x}}^i \\ \boldsymbol{\psi}_{\boldsymbol{\lambda}}^i \end{pmatrix} + \left\{ K_i^{-1} - (K^\star)^{-1} \right\} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_i \\ c_i \end{pmatrix} \right\} - (C_i - C^\star) \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star \end{pmatrix}, \tag{21c}$$

$$\begin{pmatrix} \boldsymbol{\psi}_{\boldsymbol{x}}^i \\ \boldsymbol{\psi}_{\boldsymbol{\lambda}}^i \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_i \\ c_i \end{pmatrix} - K^\star \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star \end{pmatrix}. \tag{21d}$$

Further, under Assumptions 3.2, 3.3, $\boldsymbol{\theta}^i = (\boldsymbol{\theta}_{\boldsymbol{x}}^i, \boldsymbol{\theta}_{\boldsymbol{\lambda}}^i)$ is a martingale difference with $\mathbb{E}[\boldsymbol{\theta}^i \mid \mathcal{F}_{i-1}] = \mathbf{0}$.

Proof. See Appendix C.1. ∎

From Lemma 4.1, we see that the iteration recursion consists of three terms. $\mathcal{I}_{1,t}$ is a martingale, which further consists of the randomness of generating sample ξ_t to estimate ∇f_t and the randomness of solving Newton system (3). $\mathcal{I}_{2,t}$ characterizes the randomness of the stepsize $\bar{\alpha}_t$. $\mathcal{I}_{3,t}$ contains all the remaining terms. We will show that $\mathcal{I}_{1,t}$ term provides the asymptotic convergence rate and asymptotic normality, ensured by the central limit theorem for martingales; and that $\mathcal{I}_{2,t}$ and $\mathcal{I}_{3,t}$ terms only contribute higher order errors as long as we set $\chi_t = \eta_t - \beta_t$ properly.

To facilitate our analysis, we state here a preliminary continuity property for the projection matrix $K_t S(S^T K_t^2 S)^\dagger S^T K_t$, appearing in quantities $C_{t,j}, C_t$ etc. (cf. (7) and Lemma 3.4(b)). It is fairly easy to see that if $S^T K_t^2 S$ is invertible (i.e., S has full column rank), then the projection matrix is continuous in K_t (because $(\cdot)^{-1}$ is a continuous function). However, to enable more general cases, we do not require S to have full column rank. The continuity property of the projection matrix still holds, ensured by Wedin's $\sin(\Theta)$ theorem (Wedin, 1972). [1]

Lemma 4.2. Suppose $K_t, K^\star \in \mathbb{R}^{(d+m)\times(d+m)}$ have full rank, for any $S \in \mathbb{R}^{(d+m)\times q}$, we have

$$\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star\| \leq \frac{2\|K_t - K^\star\|}{\sigma_{\min}(K^\star)} \cdot \frac{\|S\|}{\sigma_{\min}^+(S)},$$

where $\sigma_{\min}(\cdot)$ denotes the least singular value and $\sigma_{\min}^+(\cdot)$ denotes the least nonzero singular value.

Proof. See Appendix C.2. ∎

Lemma 4.2 not only suggests that the projection matrix $K_t S(S^T K_t^2 S)^\dagger S^T K_t$ is continuous in K_t for any S (given the full rankness of K_t, cf. Assumption 3.1), but also provides an explicit bound on characterizing the difference of two projection matrices. In particular, the difference of two projection matrices is proportional to the difference of two Hessians. We introduce the following condition on the sketching distribution S and present a corollary.

Assumption 4.3. The sketching distribution S satisfies $\mathbb{E}[\|S\|/\sigma_{\min}^+(S)] \leq \Upsilon_S$ for some $\Upsilon_S \geq 1$.

Assumption 4.3 is equivalent to assuming that the condition number of S, that is $\|S\|\|S^\dagger\|$, has bounded expectation. This assumption is mild, and is satisfied with $\Upsilon_S = 1$ for any sketching distribution if S is a vector, such as $S = \boldsymbol{e}_i \in \mathbb{R}^{d+m}$ for $i = 1, \ldots, d+m$ with equal probability as used in randomized Kaczmarz method.

Corollary 4.4. Suppose $K_t, K^\star$ have full rank and $K_t \to K^\star$ almost surely, then $C_t \to C^\star$ almost surely. Furthermore, if Assumption 4.3 holds, then we quantitatively have

$$\|C_t - C^\star\| \leq \frac{2\tau\Upsilon_S}{\sigma_{\min}(K^\star)}\|K_t - K^\star\|.$$

Proof. See Appendix C.3. ∎

By Corollary 4.4, we know that $\boldsymbol{\delta}^t$ in (21c) satisfies $\boldsymbol{\delta}^t = o(\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|)$. This is because $\boldsymbol{\psi}^t = O(\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|^2)$, $K_t \to K^\star$, and $C_t \to C^\star$. See Lemma 4.7 for details.

[1] In fact, the pseudo-inverse is continuous when rank is not perturbed (which is the case in our paper). In particular, $\|A^\dagger - B^\dagger\|$ is small if $\text{rank}(A) = \text{rank}(B)$ and $\|A - B\|$ is small. See (Wedin, 1973, Theorem 4.1) for details.

4.2 Bounding $\mathcal{I}_{1,t}$, $\mathcal{I}_{2,t}$, $\mathcal{I}_{3,t}$ terms

We bound $\mathcal{I}_{1,t}, \mathcal{I}_{2,t}, \mathcal{I}_{3,t}$ in (20a), (20b), (20c), respectively. To simplify the presentation, we introduce additional notation. We let $I + C^\star$ have an eigenvalue decomposition expressed as

$$I + C^\star = U\Sigma U^T \quad \text{with} \quad \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d+m}). \tag{22}$$

Under Assumption 3.3, full rankness of K_t and $K^\star$, and $K_t \to K^\star$, Lemma 3.4(b) and Corollary 4.4 imply that $\|C^\star\| \leq \rho^\tau$. Also, we note from (21a) that $C^\star \preceq \mathbf{0}$. Thus, for any $i = 1, \ldots, d+m$,

$$0 < 1 - \rho^\tau \leq \sigma_i \leq 1 \qquad \text{with } \rho = 1 - \gamma_S. \tag{23}$$

We generate sketch matrices $S_1, \ldots, S_\tau \overset{iid}{\sim} S$ and define a random matrix

$$\widetilde{C}^\star = -\prod_{j=1}^{\tau}(I - K^\star S_j (S_j^T (K^\star)^2 S_j) S_j^T K^\star). \tag{24}$$

Obviously, we have $\mathbb{E}[\widetilde{C}^\star] = C^\star$. We also define

$$\Omega^\star = (K^\star)^{-1} \begin{pmatrix} \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (K^\star)^{-1}. \tag{25}$$

With the above definitions, the next lemma establishes the asymptotic convergence rate, asymptotic normality, and Berry-Esseen bound for the martingale $\mathcal{I}_{1,t}$. Recall that $\varphi_t = (\beta_t + \eta_t)/2$.

Lemma 4.5. Under Assumptions 3.1, 3.2(15a, 15e), 3.3, 3.11 and suppose

$$\lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}}{\varphi_t}\right) = \varphi < 0, \qquad \lim_{t\to\infty} t\varphi_t = \widetilde{\varphi} \in (0, \infty], \qquad 1 - \rho^\tau + \frac{\varphi}{\widetilde{\varphi}} > 0. \tag{26}$$

Then, for any $\upsilon > 0$, we have

$$\mathcal{I}_{1,t} = o\left(\sqrt{\varphi_t \{\log(1/\varphi_t)\}^{1+\upsilon}}\right) \qquad \text{almost surely}. \tag{27}$$

Furthermore, if (15b) holds, then we have
(a) (asymptotic rate) $\mathcal{I}_{1,t} = O(\sqrt{\varphi_t \log(1/\varphi_t)})$ almost surely.
(b) (asymptotic normality) $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{1,t} \xrightarrow{d} \mathcal{N}(0, \Xi^\star)$ where

$$\Xi^\star = U\left(\Theta \circ U^T \mathbb{E}\left[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T U\right]\right) U^T \quad \text{with} \quad [\Theta]_{k,l} = \frac{1}{\sigma_k + \sigma_l + \frac{\varphi}{\widetilde{\varphi}}}. \tag{28}$$

Here, $\circ$ denotes the matrix Hadamard product.
(c) (Berry-Esseen bound) For any vector $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T \Xi^\star \boldsymbol{w} \neq 0$, we have

$$\sup_{z\in\mathbb{R}} \left| P\left(\frac{\sqrt{1/\varphi_t} \cdot \boldsymbol{w}^T \mathcal{I}_{1,t}}{\sqrt{\boldsymbol{w}^T \Xi^\star \boldsymbol{w}}} \leq z\right) - P\left(\mathcal{N}(0,1) \leq z\right)\right| = O\left(\sqrt{\varphi_t \log(1/\varphi_t)}\right).$$

Proof. See Appendix C.4. ∎

As we mentioned after Assumption 3.2 in Section 3, the condition (15e) in Lemma 4.5 can be weakened to (15d) if we additionally assume the mapping, $\boldsymbol{x} \rightarrow \mathbb{E}[\nabla f(\boldsymbol{x};\xi)\nabla^T f(\boldsymbol{x};\xi)]$, is continuous. (This is because (C.10) holds immediately by the continuity of the mapping, and all the remaining proofs still apply.)

Next, we characterize $\mathcal{I}_{2,t}$ defined in (20b). Recall that $\chi_t = \eta_t - \beta_t$ is the gap of upper and lower bounds for the stepsize $\bar{\alpha}_t$.

Lemma 4.6. Under Assumptions 3.1, 3.2(15a), 3.3, 3.11 and suppose (26) holds. Furthermore, we suppose χ_t satisfies for a constant χ, and a positive constant $\upsilon' > 0$ that

$$\lim_{t\rightarrow\infty} t\left(1 - \frac{\chi_{t-1}}{\chi_t}\right) = \chi, \quad 2(1-\rho^\tau) + \frac{2\chi - \varphi}{\widetilde{\varphi}} > 0, \quad \{\log(1/\chi_t)\}^{1+\upsilon'} = O(1/\varphi_t). \tag{29}$$

Then, we have

$$\mathcal{I}_{2,t} = O\left(\frac{\chi_t}{\varphi_t}\right) \quad \text{almost surely.}$$

Proof. See Appendix C.5. ■

Next, we characterize $\mathcal{I}_{3,t}$ defined in (20c).

Lemma 4.7. Under Assumptions 3.1, 3.2(15a, 15e), 3.3, 3.11 and suppose (26) and (29) hold. Then, for any $\upsilon > 0$,

$$\mathcal{I}_{3,t} = o\left(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\upsilon}}\right) + o(\chi_t/\varphi_t) = o(\mathcal{I}_{1,t} + \mathcal{I}_{2,t}) \quad \text{almost surely.}$$

Furthermore, if (15b) holds, the above result holds with $\upsilon = 0$.

Proof. See Appendix C.6. ■

4.3 Asymptotic rate and normality

We combine Lemmas 4.1, 4.5, 4.6, 4.7, and derive the following asymptotic convergence rate for the iterates. The proof is omitted.

Theorem 4.8 (Asymptotic convergence rate)**.** Under Assumptions 3.1, 3.2(15a, 15e), 3.3, 3.11 and suppose (26) and (29) hold. Then, for any $\upsilon > 0$,

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = o\left(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\upsilon}}\right) + O(\chi_t/\varphi_t) \quad \text{almost surely.}$$

Furthermore, if (15b) holds, we have

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = O\left(\sqrt{\varphi_t\log(1/\varphi_t)}\right) + O(\chi_t/\varphi_t) \quad \text{almost surely.}$$

From Theorem 4.8, we see that the asymptotic convergence rate consists of two terms. The first term comes from the strong law of large number for a martingale that characterizes the random sampling in each iteration (cf. $\mathcal{I}_{1,t}$ in (20a)). The second term comes from the adaptivity of random stepsizes (cf. $\mathcal{I}_{2,t}$ in (20b)). It disappears when $\chi_t = 0$, i.e., setting $\bar{\alpha}_t = \beta_t$ deterministically. From

Lemma 4.7, we see that all the remaining terms contained in $\mathcal{I}_{3,t}$ only contribute higher order errors. We note that Berahas et al. (2021b,a); Curtis et al. (2021b) set $\chi_t = O(\beta_t^2)$. In this case, $\chi_t = O(\varphi_t^2)$ and, hence, the first term dominates the convergence rate. In fact, our theorem suggests that $\chi_t = O(\beta_t^{3/2})$ is sufficient to have the first term dominate, which allows a wider interval for selecting the stepsize $\bar{\alpha}_t$.

We will transfer conditions (26) and (29) to some conditions on the sequence $\{\beta_t, \chi_t\}_t$ in Theorem 4.14, and justify them in Lemma 4.15. Before that, we establish the asymptotic normality and Berry-Esseen bound for the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$. To show these results, we need an explicit convergence rate for K_t, which further allows us to provide a better rate for the term $\mathcal{I}_{3,t}$.

Lemma 4.9. Under Assumptions 3.1, 3.2(15a, 15e), 3.3, 3.11 and suppose (26) and (29) hold. Then, for any $\upsilon > 0$, we have

$$\|K_t - K^\star\| = o\left(\sqrt{\varphi_t \{\log(1/\varphi_t)\}^{1+\upsilon}}\right) + O\left(\frac{\chi_t}{\varphi_t}\right) + \omega_t \quad \text{almost surely.}$$

Furthermore, if (15b) holds, we have

$$\|K_t - K^\star\| = O\left(\sqrt{\varphi_t \log(1/\varphi_t)}\right) + O\left(\frac{\chi_t}{\varphi_t}\right) + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t \quad \text{almost surely.}$$

Proof. See Appendix C.7. ∎

We provide a better rate for $\mathcal{I}_{3,t}$. The following lemma differs from Lemma 4.7 in the bound of $\boldsymbol{\delta}^t$. Using Lemma 4.9, we can have a more precise bound on $\boldsymbol{\delta}^t$ than (C.25) in the proof of Lemma 4.7. We need to impose Assumption 4.3 and make use of Corollary 4.4. We only present the result under (15b) and $\chi_t = o(\varphi_t^{3/2})$. This is because the asymptotic normality and Berry-Esseen bound, where we will apply the result, require both of them. In particular, (15b) is required for the asymptotic normality of $\mathcal{I}_{1,t}$ in Lemma 4.5. $\chi_t = o(\varphi_t^{3/2})$ is required to ensure that, compared to $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{1,t}$, $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{2,t} = o(1)$ contributes higher order errors (cf. Lemma 4.6) and does not affect the normality.

Lemma 4.10. Under Assumptions 3.1, 3.2(15b, 15e), 3.3, 3.11, 4.3, and suppose (26) and (29) hold. Suppose also that

$$\lim_{t\to\infty} t\left(1 - \frac{\omega_{t-1}}{\omega_t}\right) = \omega < 0, \quad 1 - \rho^\tau + \frac{\varphi + (-1 \wedge 2\omega)}{2\widetilde{\varphi}} > 0, \quad \chi < \frac{3\varphi}{2}. \tag{30}$$

Then, for any $\upsilon > 0$,

$$\mathcal{I}_{3,t} = O\left(\varphi_t \log(1/\varphi_t)\right) + o\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + O\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \omega_t\right) \quad \text{almost surely.}$$

The above result also holds if $\omega_t = 0$ for all sufficiently large t.

Proof. See Appendix C.8. ∎

Combining Lemmas 4.1, 4.5, 4.6, and 4.10, we derive the following theorem.

Theorem 4.11 (Asymptotic normality and Berry-Esseen bound)**.** Under Assumptions 3.1, 3.2(15b, 15e), 3.3, 3.11, 4.3, and suppose (26), (29), and (30) hold. Then,

$$\sqrt{1/\varphi_t} \cdot \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Xi^\star),$$

where $\Xi^\star$ is in (28). Further, for any vector $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T \Xi^\star \boldsymbol{w} \neq 0$ and any $\upsilon > 0$, we have

$$\sup_{z\in\mathbb{R}} \left| P\left(\frac{\sqrt{1/\varphi_t} \cdot \boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T \Xi^\star \boldsymbol{w}}} \leq z \right) - P\left(\mathcal{N}(0,1) \leq z\right) \right|$$
$$= O\left(\sqrt{\varphi_t} \log\left(1/\varphi_t\right)\right) + O\left(\frac{\chi_t}{\varphi_t^{3/2}}\right) + o\left(\frac{\sqrt{\log(1/\varphi_t)(\log t)^{1+\upsilon}}}{\sqrt{t}}\right) + O\left(\sqrt{\log(1/\varphi_t)} \cdot \omega_t\right).$$

The above results also hold if $\omega_t = 0$ for all sufficiently large t.

Proof. See Appendix C.9. ■

4.4 An estimator of the covariance matrix

Theorem 4.11 shows the asymptotic normality of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$. However, the covariance matrix $\Xi^\star$ (28) has a complex form. It depends on $\Omega^\star$ in (25) and an expectation of a random matrix $\widetilde{C}^\star$ in (24), both of which are inaccessible in practice. To perform statistical inference for $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, we propose a simple estimator for $\Xi^\star$. Our estimator is independent of the sketching distribution employed in randomized solvers.

Lemma 4.12. We let

$$\Omega_t = K_t^{-1} \begin{pmatrix} \frac{1}{t}\sum_{i=0}^{t-1} \bar{g}_i \bar{g}_i^T - \left(\frac{1}{t}\sum_{i=0}^{t-1} \bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1} \bar{g}_i\right)^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} K_t^{-1} \quad \text{and} \quad \Xi_t = \frac{\Omega_t}{2 + \varphi/\widetilde{\varphi}}. \tag{31}$$

Under Assumptions 3.1, 3.2(15c, 15e), 3.3, 3.11 and suppose (26) and (29) hold. Then, for any $\upsilon > 0$, we have

$$\|\Xi^\star - \Xi_t\| = O(\rho^\tau) + O\left(\sqrt{\varphi_t \log(1/\varphi_t)}\right) + O\left(\frac{\chi_t}{\varphi_t}\right) + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t.$$

Proof. See Appendix C.10. ■

Lemma 4.12 leads to the following corollary, which shows Berry-Esseen bound with true covariance being replaced by Ξ_t.

Corollary 4.13. Under Assumptions 3.1, 3.2(15c, 15e), 3.3, 3.11, 4.3, and suppose (26), (29), and (30) hold. Then, for any vector $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T \Xi_t \boldsymbol{w} \neq 0$, for any $\upsilon > 0$ and all sufficiently large τ, we have

$$\sup_{z\in\mathbb{R}} \left| P\left(\frac{\sqrt{1/\varphi_t} \cdot \boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T \Xi_t \boldsymbol{w}}} \leq z \right) - P\left(\mathcal{N}(0,1) \leq z\right) \right|$$
$$= O\left(\sqrt{\varphi_t} \log\left(\frac{1}{\varphi_t}\right)\right) + O\left(\frac{\chi_t}{\varphi_t^{3/2}}\right) + o\left(\frac{\sqrt{\log(1/\varphi_t)(\log t)^{1+\upsilon}}}{\sqrt{t}}\right) + O\left(\sqrt{\log(1/\varphi_t)} \cdot \omega_t\right) + O(\rho^\tau).$$

The above result also holds if $\omega_t = 0$ for all sufficiently large t.

Proof. See Appendix C.11. ■

Combining Corollary 4.13 with Theorem 4.11, we see that Berry-Esseen bound of using the covariance estimator Ξ_t only leads to an additional $O(\rho^\tau)$ term. This is negligible for a large τ.

4.5 A complete understanding of AI-StoSQP

We summarize all results in Sections 3 and 4, and rephrase all conditions (26), (29), and (30) using the input sequences $\{\beta_t, \chi_t = \eta_t - \beta_t\}$.

Theorem 4.14 (Global and local convergence of AI-StoSQP)**.** The following results of AI-StoSQP in Algorithm 1 hold. The assumptions are gradually strengthened.
(a) Global convergence. Suppose Assumptions 3.1, 3.2(15a), 3.3 hold; τ satisfies (17); and the input sequence $\{\beta_t, \chi_t\}_t$ satisfies

$$\lim_{t\to\infty} t\left(1-\frac{\beta_{t-1}}{\beta_t}\right) = \beta \in [-1, -\frac{1}{2}), \quad \lim_{t\to\infty} t\beta_t = \widetilde{\beta} \in (0,\infty], \quad \lim_{t\to\infty} t\left(1-\frac{\chi_{t-1}}{\chi_t}\right) = \chi < -1, \tag{32}$$

then $\|\nabla\mathcal{L}_t\| \to 0$ as $t\to\infty$ almost surely. Furthermore, if (15d) and Assumption 3.11 hold, then $K_t \to K^\star$ as $t\to\infty$ almost surely.
(b) Local asymptotic convergence rate. If (15d) is strengthen to (15e), and for a constant $\upsilon' > 0$,

$$1-\rho^\tau + \frac{\beta}{\widetilde{\beta}} > 0, \quad 2(1-\rho^\tau) + \frac{2\chi - \beta}{\widetilde{\beta}} > 0, \quad \{\log(1/\chi_t)\}^{1+\upsilon'} = O(1/\beta_t), \tag{33}$$

then, for any $\upsilon > 0$,

$$\left.\begin{aligned} \left\|\begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix}\right\| &= o\left(\sqrt{\beta_t \{\log(1/\beta_t)\}^{1+\upsilon}}\right) + O(\chi_t/\beta_t) \\ \|K_t - K^\star\| &= o\left(\sqrt{\beta_t \{\log(1/\beta_t)\}^{1+\upsilon}}\right) + O(\chi_t/\beta_t) + \omega_t \end{aligned}\right\} \quad \text{almost surely.}$$

Furthermore, if (15a) is also strengthen to (15b), then

$$\left.\begin{aligned} \left\|\begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix}\right\| &= O\left(\sqrt{\beta_t \log(1/\beta_t)}\right) + O(\chi_t/\beta_t) \\ \|K_t - K^\star\| &= O\left(\sqrt{\beta_t \log(1/\beta_t)}\right) + O(\chi_t/\beta_t) + o\left(\sqrt{(\log t)^{1+\upsilon}/t}\right) + \omega_t \end{aligned}\right\} \quad \text{almost surely.}$$

(c) Asymptotic normality and Berry-Esseen bound. If Assumption 4.3 holds and

$$\lim_{t\to\infty} t\left(1-\frac{\omega_{t-1}}{\omega_t}\right) = \omega < 0, \qquad 1-\rho^\tau + \frac{\beta+2\omega}{2\widetilde{\beta}} > 0, \qquad \chi < \frac{3\beta}{2} \tag{34}$$

(all following results also hold if $\omega_t = 0$ for large t; in this case, conditions that involve ω_t can be removed), then

$$\sqrt{1/\beta_t}\cdot\begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Xi^\star)$$

where $\Xi^\star$ is defined in (28), and for any vector $\boldsymbol{w}$ such that $\boldsymbol{w}^T\Xi^\star\boldsymbol{w} \neq 0$,

$$\sup_{z\in\mathbb{R}}\left|P\left(\frac{\sqrt{1/\beta_t}\cdot\boldsymbol{w}^T(\boldsymbol{x}_t-\boldsymbol{x}^\star,\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}}\leq z\right)-P\left(\mathcal{N}(0,1)\leq z\right)\right|$$
$$=O\left(\sqrt{\beta_t}\log\left(\frac{1}{\beta_t}\right)\right)+O\left(\frac{\chi_t}{\beta_t^{3/2}}\right)+o\left(\frac{\sqrt{\log(1/\beta_t)(\log t)^{1+\upsilon}}}{\sqrt{t}}\right)+O\left(\sqrt{\log(1/\beta_t)}\cdot\omega_t\right). \quad (35)$$

(d) Practical covariance estimator. If (15b) is further strengthen to (15c), then we can employ the estimator Ξ_t in (31), and a similar Berry-Esseen bound holds if we replace $\boldsymbol{w}^T\Xi^\star\boldsymbol{w}$ by $\boldsymbol{w}^T\Xi_t\boldsymbol{w}$—only an additional $O(\rho^\tau)$ term appears on the right hand side of (35).

Proof. See Appendix C.12. ■

We emphasize that the asymptotic convergence rate requires Assumptions 3.1, 3.2(15a, 15e), 3.3, 3.11, while the asymptotic normality and Berry-Esseen bound require us to impose Assumption 4.3 and strengthen (15a) to (15b), although the asymptotic rate is also improved under stronger conditions. To adopt the covariance estimator Ξ_t, (15b) need to be further strengthen to (15c). However, even for the most stringent conditions (15c) and (15e), they are quite common in the literature (Chen et al., 2020), and they enable wide applications including constrained least squares and constrained logistic regression.

We let β_t, χ_t be polynomial in t (which is convenient for implementation), and discuss the conditions (32), (33), (34) on the input sequence $\{\beta_t, \chi_t\}_t$.

Lemma 4.15. Suppose $\beta_t = c_1/t^{c_2}$ and $\chi_t = \beta_t^{c_3}$. Then,
(a) (32) is satisfied if $c_1 > 0$, $c_2 \in (0.5, 1]$, $c_3 > 1/c_2$.
(b) (32), (33) are satisfied if

$$c_2 = 1, c_3 > 1, c_1 > \frac{1\vee(c_3-0.5)}{1-\rho^\tau} \qquad \text{OR} \qquad c_2\in(0.5,1), c_1>0, c_3>1/c_2.$$

(c) (32), (33), (34) are satisfied if

$$c_2 = 1, c_3 > 1.5, c_1 > \frac{(0.5-\omega)\vee(c_3-0.5)}{1-\rho^\tau} \qquad \text{OR} \qquad c_2\in(0.5,1), c_1>0, c_3>1.5\vee 1/c_2.$$

Proof. The result holds immediately by noting that $\beta = -c_2$, $\widetilde{\beta} = c_1$ if $c_2 = 1$ and ∞ if $c_2 < 1$, and $\chi = -c_2c_3$. ■

With the setup in Lemma 4.15, the result in Theorem 4.14 can be further simplified. For instance, we let $c_2 \in (0.5, 1)$, $c_1 > 0$, $c_3 = 2$. Then, (32), (33), (34) are satisfied. Thus, under Assumptions 3.1, 3.2(15b, 15e), 3.3, 3.11, 4.3, we have

$$\left\|\begin{pmatrix}\boldsymbol{x}_t-\boldsymbol{x}^\star\\ \boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star\end{pmatrix}\right\| = O\left(\sqrt{\frac{\log t}{t^{c_2}}}\right)$$

and

$$\sup_{z\in\mathbb{R}}\left|P\left(\frac{\sqrt{t^{c_2}}\cdot\boldsymbol{w}^T(\boldsymbol{x}_t-\boldsymbol{x}^\star,\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star)}{\sqrt{c_1\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}}\leq z\right)-P\left(\mathcal{N}(0,1)\leq z\right)\right| = O\left(\frac{\log t}{\sqrt{t^{c_2}}}\right)+O\left(\sqrt{\log t}\cdot\omega_t\right).$$

The 95% confidence interval of $\boldsymbol{w}^T(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ is given by

$$\boldsymbol{w}^T(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \pm \frac{1.96}{t^{c_2/2}} \cdot \sqrt{c_1 \boldsymbol{w}^T \Xi_t \boldsymbol{w}}. \tag{36}$$

We emphasize that our asymptotic results also hold (and the convergence rate does not change) if we solve Newton systems exactly, and/or adopt deterministic stepsizes. The covariance matrix for solving Newton systems exactly is $\Xi^\star = \Omega^\star/(2+\beta/\widetilde{\beta})$. However, it is not clear to us whether the same results hold when we approximately solve Newton systems via deterministic solvers, such as with CG and MINRES methods as employed in Curtis et al. (2021b). In that case, Lemma 3.4 does not hold and significant adjustments of the analysis are required.

5 Numerical Experiment

In this section, we provide experimental results of Algorithm 1. We develop a Julia implementation (Siqueira and Orban, 2020); and we consider solving problems in CUTEst test set (Gould et al., 2014), in particular using problems `BYRDSPHR, HS7, HS48` as examples to validate our main theoretical results. For each problem, we perform 10^5 SQP iterations and, in each iteration, we perform 50 randomized Kaczmarz iterations for solving the Newton system. That is, for each t, we let $S_{t,j} \overset{iid}{\sim} \text{Uniform}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m+d}\}$ for $j = 1, \ldots, 50$, and perform the iteration (6) with $\boldsymbol{z}_{t,0} = \boldsymbol{0}$. The true solution of each problem is solved by applying IPOPT solver (Wächter and Biegler, 2006). See a Julia implementation of the solver at `https://github.com/jump-dev/Ipopt.jl`.

Given an iterate $\boldsymbol{x}_t$, we generate $\bar{g}_t \sim \mathcal{N}(\nabla f_t + \sigma^2(I + \boldsymbol{1}\boldsymbol{1}^T))$ where $\boldsymbol{1} \in \mathbb{R}^d$ is an all-one vector. We also generate the (i,j) and (j,i) entries of $\bar{H}_t$ from $\mathcal{N}((\nabla^2 f_t)_{i,j}, \sigma^2)$. Here, ∇f_t and $\nabla^2 f_t$ are provided by the package of CUTEst. Among the considered problems, we vary $\sigma^2 \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}, 1\}$, and vary $\beta_t \in \{2/t^{0.5}, 2/t^{0.6}\}$. We also let $\chi_t = \beta_t^2$ and $\eta_t = \beta_t + \chi_t$, and randomly pick the stepsize $\bar{\alpha}_t \sim \text{Uniform}([\beta_t, \eta_t])$. Here, we let β_t decay slower to favor larger stepsizes. We note that, although $\beta_t = 2/t^{0.5}$ is not allowed in our asymptotic analysis (see Lemma 4.15), we still implement it to investigate the performance of this setup. In addition, for the Hessian modification, we let

$$\Delta_t = (-\lambda_{\min}(Z_t^T \sum_{i=0}^{t-1} \bar{\nabla}_{\boldsymbol{x}}^2 \mathcal{L}_i Z_t)/t + 0.1) \cdot I \quad \text{whenever} \quad \lambda_{\min}(Z_t^T \sum_{i=0}^{t-1} \bar{\nabla}_{\boldsymbol{x}}^2 \mathcal{L}_i Z_t) < 0.$$

Here, $Z_t \in \mathbb{R}^{d\times(d-m)}$ has orthogonal columns that span the space $\{\boldsymbol{x} \in \mathbb{R}^d : G_t \boldsymbol{x} = \boldsymbol{0}\}$, which is obtained by QR decomposition in our implementation. Our code for implementing other problems in CUTEst test set is publicly available at `https://github.com/senna1128/Inf-StoSQP`.

Convergence behavior. We first study the convergence behavior of Algorithm 1. Figures 1 and 2 show the convergence plots of the KKT residual $\|\nabla \mathcal{L}_t\|$, the iteration error $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|$, and the Hessian error $\|K_t - K^\star\|$. We note that, by the Lipschitz continuity of the Hessian, Theorem 4.14 implies that the KKT residual $\|\nabla \mathcal{L}_t\|$ has (at least) the same asymptotic convergence rate as the iterate, that is $O(\sqrt{\beta_t \log(1/\beta_t)})$ in our experiments.

From Figure 1, we observe that the setting $\beta_t = 2/t^{0.5}$ works well on problems `HS7, HS48`, while it does not work well on `BYRDSPHR` in terms of the iteration and the Hessian errors (see Figures 1(b) and 1(c)). However, the KKT residual still converges to a level below the theoretical prediction (see Figure 1(a)). Thus, under this setup, Algorithm 1 converges to a stationary point that is different from the one returned by IPOPT. The sequence $\beta_t = 2/t^{0.6}$ enhances the performance of Algorithm 1 on BYRDSPHR for both the iteration and the Hessian errors. From all figures in Figures 1 and 2 (we should focus on the KKT residual plots in Figures 1(a) and 2(a) for BYRDSPHR), we see that the theoretical convergence rate precisely catches the asymptotic behavior of Algorithm 1.

Asymptotic normality. We next study the asymptotic normality of the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$. To visualize the results, we only plot the histograms for $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$; that is we only focus on the first coordinate of the primal and dual solutions. We regard the first half of the iterations as burn-in period; thus we only plot the errors for the second half of the iterations. For each histogram, we fit a normal density by matching the mean and variance. Figures 3, 4, 5 show the histograms of three problems BYRDSPHR, HS7, HS48, respectively.

From Figure 5 (i.e. histograms of problem HS48), we see that all iteration errors with both setups of $\beta_t = 2/t^{0.5}$ and $\beta_t = 2/t^{0.6}$ follow a normal distribution with mean around zero. This is consistent with our theorem. From Figure 4, we see that the histograms of $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ under the setup of $\beta_t = 2/t^{0.6}$ and $\sigma^2 = 10^{-1}$ do not fit a normal density (see Figures 4(n) and 4(s)). In fact, from Figures 2(d) and 2(e), we know that $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \nrightarrow (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ but $\|\nabla \mathcal{L}_t\| \to 0$ under this setup. Thus, the scheme converges to a stationary point that differs from the one given by IPOPT. Similar situations happen for few cases in Figure 3. In particular, Figure 3(r) suggests that $[\boldsymbol{\lambda}_t]_1 \nrightarrow \boldsymbol{\lambda}_1^\star$ with $\beta_t = 2/t^{0.6}$ and $\sigma^2 = 10^{-2}$, which is consistent with Figure 2(b). However, Figure 2(a) suggests that $\|\nabla \mathcal{L}_t\| \to 0$ in this case. Overall, the histograms in Figures 3, 4, 5 suggest that the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ follows a normal distribution with mean $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ when it converges to a local solution.

Confidence interval. We finally test the effectiveness of the proposed covariance estimator. We construct 95% confidence interval for the solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$. We take $\boldsymbol{x}_1^\star + \boldsymbol{\lambda}_1^\star$ as an example. The formula of 95% confidence interval is given by (36), where Ξ_t is estimated by Lemma 4.12, and $\boldsymbol{w} = (1, 0, \ldots, 0, 1, 0 \ldots, 0)$, i.e., $\boldsymbol{w}_1 = \boldsymbol{w}_{d+1} = 1$ and all other entries are zero.

Figure 6 shows the confidence intervals generated by the last 100 iterations. We observe from Figures 6(u)-6(ad) that the confidence intervals cover the true value of problem HS48 for all σ^2 varying from 10^{-8} to 1, under both setups of $\beta_t = 2/t^{0.5}$ and $2/t^{0.6}$. For problem HS7, we observe from Figures 6(k)-6(t) that the confidence intervals also cover the true value, except for one setup where $\beta_t = 2/t^{0.6}$ and $\sigma^2 = 10^{-1}$. As explained, the failure of this setup is because $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ does not converge to the solution returned by IPOPT. For problem BYRDSPHR, we observe from Figures 6(a)-6(j) that the confidence intervals are improved significantly from $\beta_t = 2/t^{0.5}$ to $\beta_t = 2/t^{0.6}$. The confidence intervals with the former sequence only cover the true value when $\sigma^2 = 10^{-2}$, while the confidence intervals with the latter sequence cover the true value except when $\sigma^2 = 10^{-2}$.

We also observe from Figure 6 that the power of our confidence interval construction is significant. In particular, the length of intervals varies from 10^{-3} to 10^{-1} when σ^2 varies from 10^{-8} to 1; and such a small length implies a high statistical power, which enables a precise inference for the solution in practice (the solution of the optimization problem is often the true model parameter, and the inference of the model parameter is an important statistical problem).

6 Conclusion

We investigated a stochastic sequential quadratic programming framework called AI-StoSQP. In particular, we allow the scheme to employ any method to select random stepsizes within an interval $[\beta_t, \eta_t]$; and we allow the scheme to solve Newton systems via randomized iterative solvers. For such a framework, we established the asymptotic convergence rate and asymptotic normality for the iterates. Under suitable conditions and with proper decaying β_t and η_t, we showed that the KKT residual $\|\nabla\mathcal{L}_t\|$ converges to zero *almost surely.* Moreover, if the sequence converges to a strict local minimum satisfying constraint qualifications, then the *last* iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ has a local rate $O(\sqrt{\beta_t \log(1/\beta_t)}) + O(\chi_t/\beta_t)$; and $\sqrt{1/\beta_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)$ converges to a Gaussian distribution with a nontrivial covariance matrix. The covariance matrix depends on the decay rate of β_t and the sketching distribution in the randomized solvers. An explicit form of the covariance matrix was provided. For practical purpose, we also provided a practical estimator for the covariance matrix.

Our analysis answered an important open problem—what is the (local) asymptotic behavior of StoSQP iterates under the fully stochastic setup. Our analysis precisely characterized the uncertainty of the Markov process generated by StoSQP schemes, which consists of the randomness of sample, randomness of solver, and randomness of stepsize. Since our analysis is independent of the choice of merit functions, we know that designing a StoSQP scheme with differentiable merit functions under the fully stochastic setup *does not* bring advantages over non-differentiable merit functions (such as the ℓ_1 merit function), in terms of the local rate. However, our analysis does not allow one to select the stepsize by doing stochastic line search. The question still remains open whether adopting differentiable merit functions in stochastic line search can overcome the Maratos effect, lead to a faster local rate, and bring advantages over non-differentiable merit functions (as they can in deterministic line search). Therefore, the local analysis of StoSQP under the stochastic line search setup deserves studying in the future work.

References

S. Ahmed and A. Shapiro. Solving chance-constrained stochastic programs via sampling and integer programming. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, pages 261–269. INFORMS, 2008.

A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv preprint arXiv:2106.13015*, 2021a.

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

A. S. Berahas, J. Shi, Z. Yi, and B. Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *arXiv preprint arXiv:2204.04161*, 2022.

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

D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Elsevier, Belmont, Mass, 1982.

J. R. Birge. State-of-the-art-survey—stochastic programming: Computation and applications. *INFORMS Journal on Computing*, 9(2):111–133, 1997.

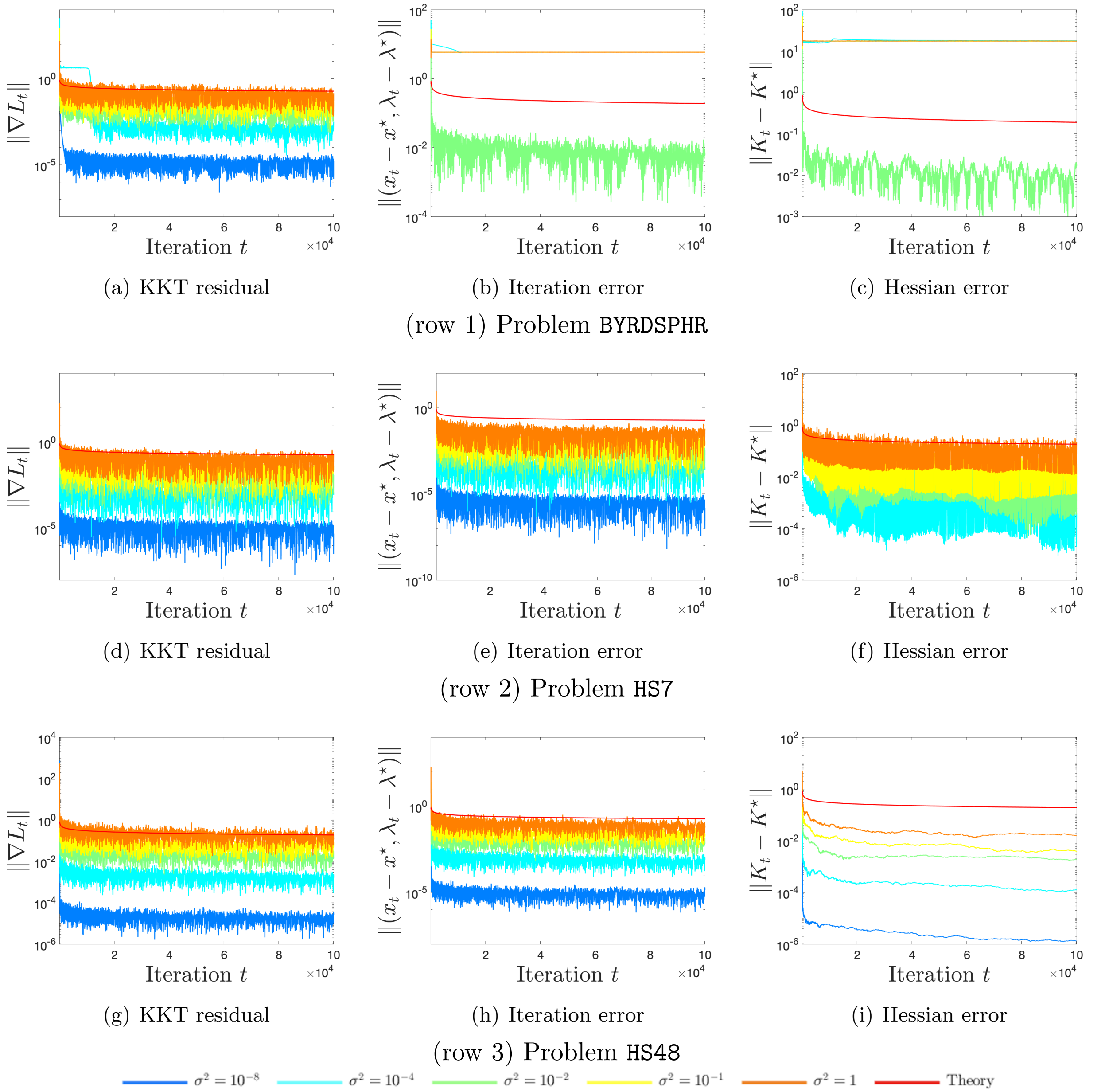


Figure 1: Convergence plots with $\beta_t = 2/t^{0.5}$. Each row corresponds to a problem, and has three figures in a log scale; from the left to the right, they correspond to $\|\nabla \mathcal{L}_t\|$ versus t, $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|$ versus t, and $\|K_t - K^\star\|$ versus t. Each figure has 6 lines—5 lines correspond to 5 setups of σ^2, and the red line shows $\sqrt{\beta_t \log(t)}$ versus t, which is the theoretical asymptotic rate (cf. Theorem 4.14).

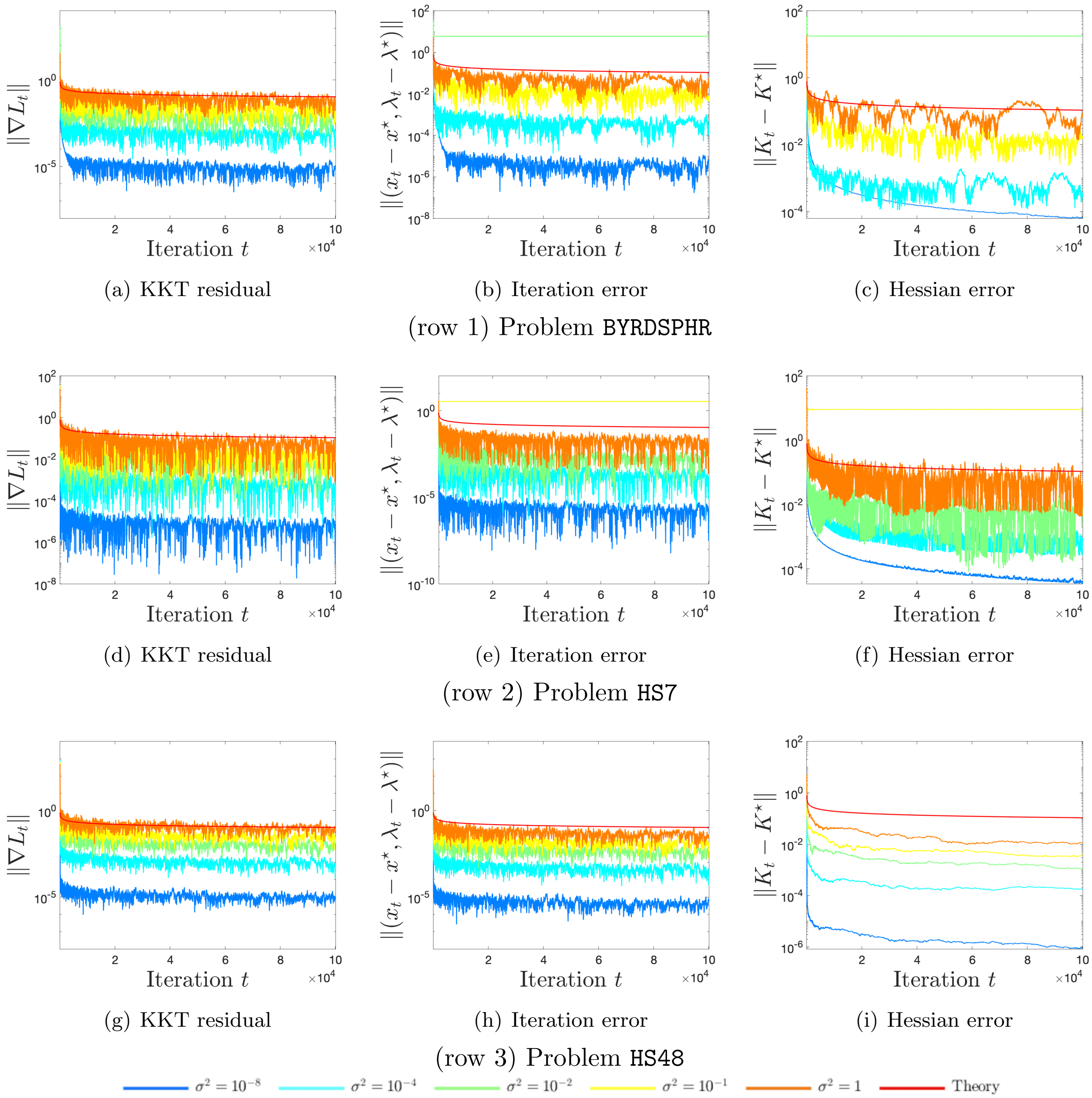
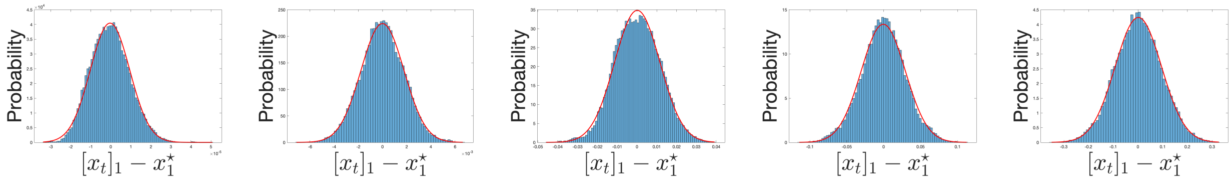
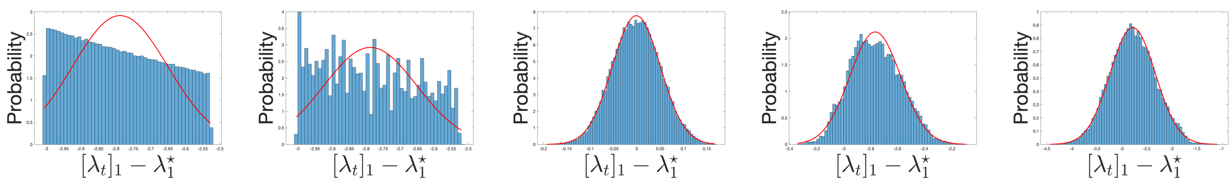


Figure 2: Convergence plots with $\beta_t = 2/t^{0.6}$. See Figure 1 for the interpretation.



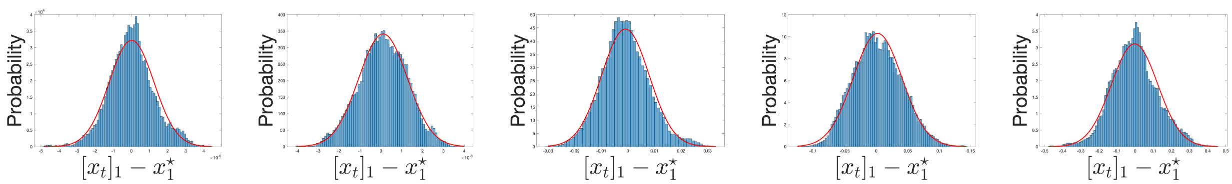
(a) $\sigma^2 = 10^{-8}$ (b) $\sigma^2 = 10^{-4}$ (c) $\sigma^2 = 10^{-2}$ (d) $\sigma^2 = 10^{-1}$ (e) $\sigma^2 = 1$

(row 1) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.5}$



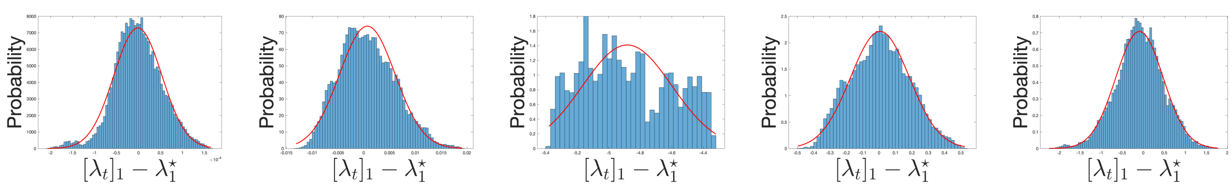
(f) $\sigma^2 = 10^{-8}$ (g) $\sigma^2 = 10^{-4}$ (h) $\sigma^2 = 10^{-2}$ (i) $\sigma^2 = 10^{-1}$ (j) $\sigma^2 = 1$

(row 2) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.5}$



(k) $\sigma^2 = 10^{-8}$ (l) $\sigma^2 = 10^{-4}$ (m) $\sigma^2 = 10^{-2}$ (n) $\sigma^2 = 10^{-1}$ (o) $\sigma^2 = 1$

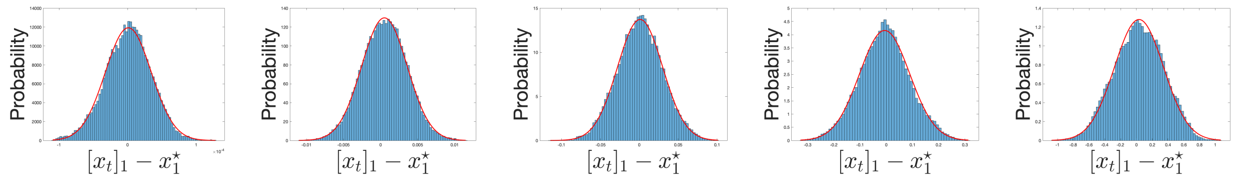
(row 3) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.6}$



(p) $\sigma^2 = 10^{-8}$ (q) $\sigma^2 = 10^{-4}$ (r) $\sigma^2 = 10^{-2}$ (s) $\sigma^2 = 10^{-1}$ (t) $\sigma^2 = 1$

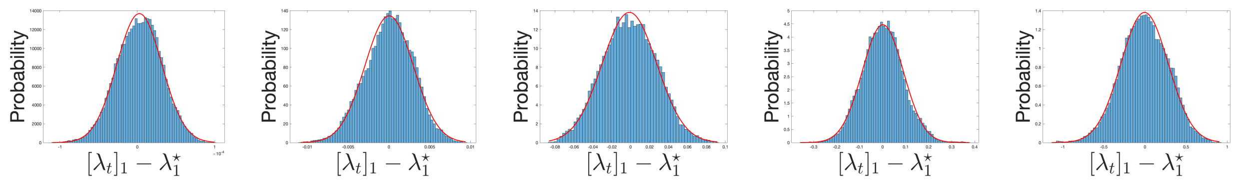
(row 4) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.6}$

Figure 3: Histograms of Problem `BYRDSPHR` (scaled to the probability density). Each row corresponds to the histogram of the combination of $\{[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star, [\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star\}$ and $\beta_t \in \{2/t^{0.5}, 2/t^{0.6}\}$. Each row has five figures, from the left to the right, corresponding to $\sigma^2 \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}, 1\}$. The red line on each figure is the fitted normal density.



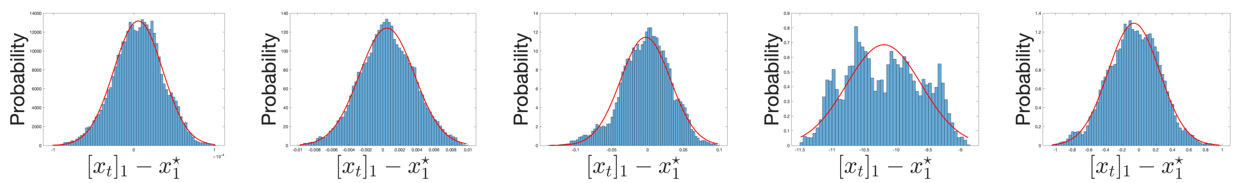
(a) $\sigma^2 = 10^{-8}$ (b) $\sigma^2 = 10^{-4}$ (c) $\sigma^2 = 10^{-2}$ (d) $\sigma^2 = 10^{-1}$ (e) $\sigma^2 = 1$

(row 1) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.5}$



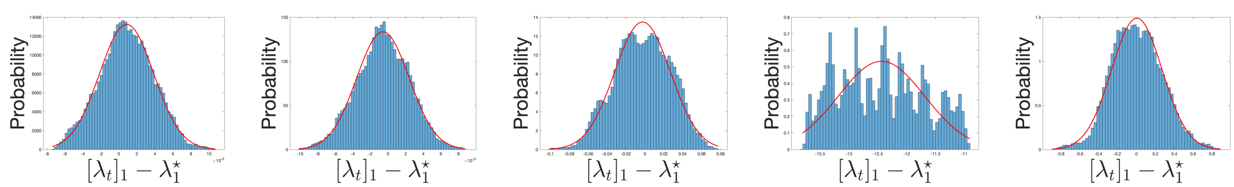
(f) $\sigma^2 = 10^{-8}$ (g) $\sigma^2 = 10^{-4}$ (h) $\sigma^2 = 10^{-2}$ (i) $\sigma^2 = 10^{-1}$ (j) $\sigma^2 = 1$

(row 2) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.5}$



(k) $\sigma^2 = 10^{-8}$ (l) $\sigma^2 = 10^{-4}$ (m) $\sigma^2 = 10^{-2}$ (n) $\sigma^2 = 10^{-1}$ (o) $\sigma^2 = 1$

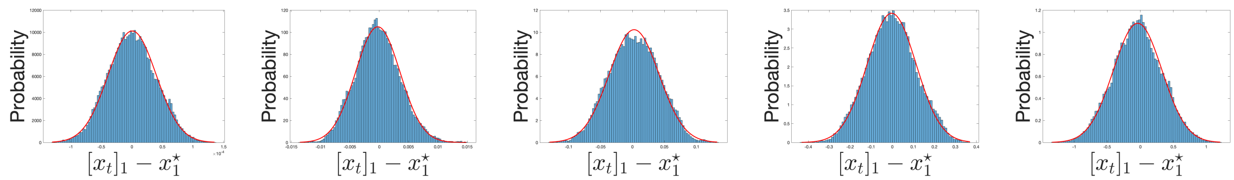
(row 3) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.6}$



(p) $\sigma^2 = 10^{-8}$ (q) $\sigma^2 = 10^{-4}$ (r) $\sigma^2 = 10^{-2}$ (s) $\sigma^2 = 10^{-1}$ (t) $\sigma^2 = 1$

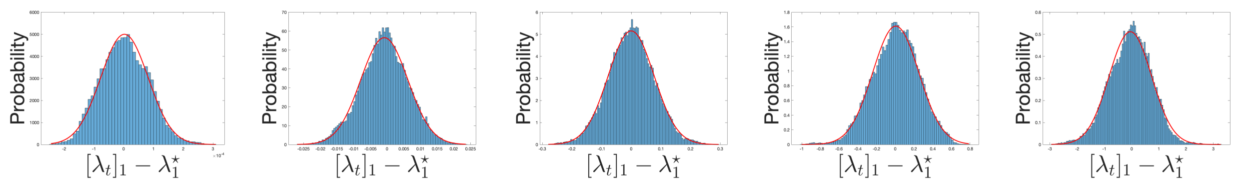
(row 4) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.6}$

Figure 4: Histograms of Problem HS7 (scaled to the probability density). See Figure 3 for the interpretation.



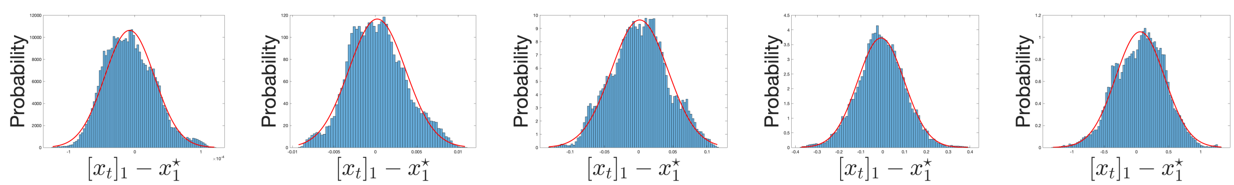
(a) $\sigma^2 = 10^{-8}$ (b) $\sigma^2 = 10^{-4}$ (c) $\sigma^2 = 10^{-2}$ (d) $\sigma^2 = 10^{-1}$ (e) $\sigma^2 = 1$

(row 1) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.5}$



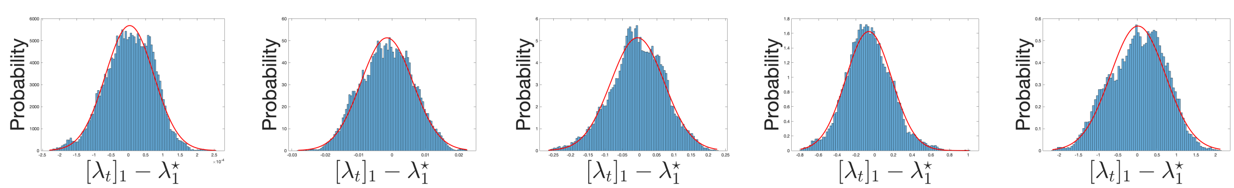
(f) $\sigma^2 = 10^{-8}$ (g) $\sigma^2 = 10^{-4}$ (h) $\sigma^2 = 10^{-2}$ (i) $\sigma^2 = 10^{-1}$ (j) $\sigma^2 = 1$

(row 2) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.5}$



(k) $\sigma^2 = 10^{-8}$ (l) $\sigma^2 = 10^{-4}$ (m) $\sigma^2 = 10^{-2}$ (n) $\sigma^2 = 10^{-1}$ (o) $\sigma^2 = 1$

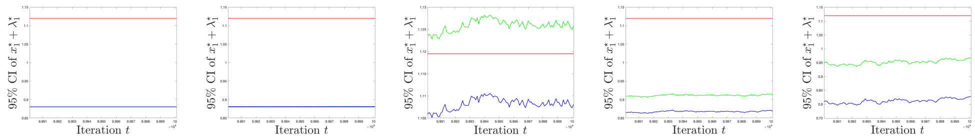
(row 3) $[\boldsymbol{x}_t]_1 - \boldsymbol{x}_1^\star$ and $\beta_t = 2/t^{0.6}$



(p) $\sigma^2 = 10^{-8}$ (q) $\sigma^2 = 10^{-4}$ (r) $\sigma^2 = 10^{-2}$ (s) $\sigma^2 = 10^{-1}$ (t) $\sigma^2 = 1$

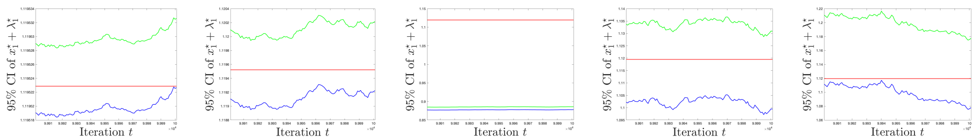
(row 4) $[\boldsymbol{\lambda}_t]_1 - \boldsymbol{\lambda}_1^\star$ and $\beta_t = 2/t^{0.6}$

Figure 5: Histograms of Problem `HS48` (scaled to the probability density). See Figure 3 for the interpretation.



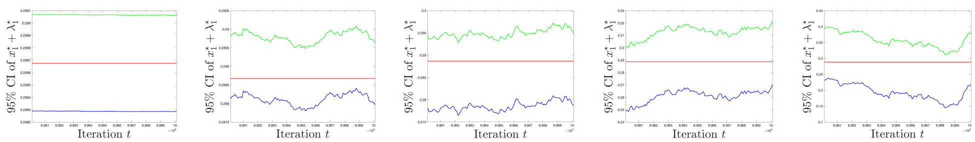
(a) $\sigma^2 = 10^{-8}$ (b) $\sigma^2 = 10^{-4}$ (c) $\sigma^2 = 10^{-2}$ (d) $\sigma^2 = 10^{-1}$ (e) $\sigma^2 = 1$

(row 1) Problem `BYRDSPHR` with $\beta_t = 2/t^{0.5}$



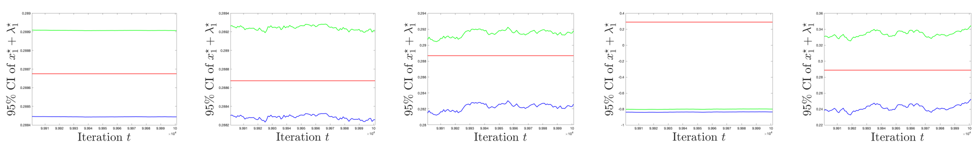
(f) $\sigma^2 = 10^{-8}$ (g) $\sigma^2 = 10^{-4}$ (h) $\sigma^2 = 10^{-2}$ (i) $\sigma^2 = 10^{-1}$ (j) $\sigma^2 = 1$

(row 2) Problem `BYRDSPHR` with $\beta_t = 2/t^{0.6}$



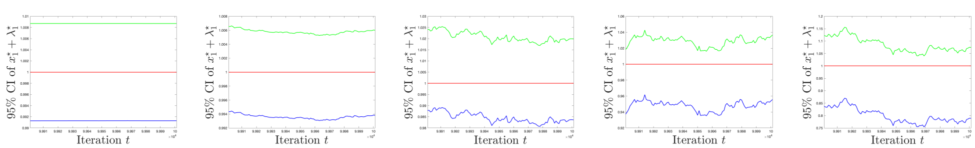
(k) $\sigma^2 = 10^{-8}$ (l) $\sigma^2 = 10^{-4}$ (m) $\sigma^2 = 10^{-2}$ (n) $\sigma^2 = 10^{-1}$ (o) $\sigma^2 = 1$

(row 3) Problem `HS7` with $\beta_t = 2/t^{0.5}$



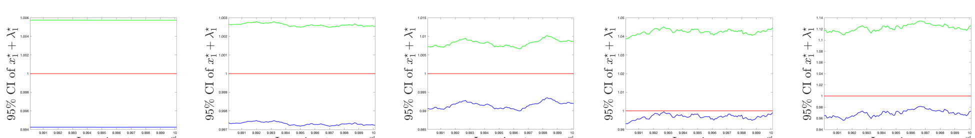
(p) $\sigma^2 = 10^{-8}$ (q) $\sigma^2 = 10^{-4}$ (r) $\sigma^2 = 10^{-2}$ (s) $\sigma^2 = 10^{-1}$ (t) $\sigma^2 = 1$

(row 4) Problem `HS7` with $\beta_t = 2/t^{0.6}$



(u) $\sigma^2 = 10^{-8}$ (v) $\sigma^2 = 10^{-4}$ (w) $\sigma^2 = 10^{-2}$ (x) $\sigma^2 = 10^{-1}$ (y) $\sigma^2 = 1$

(row 5) Problem `HS48` with $\beta_t = 2/t^{0.5}$



(z) $\sigma^2 = 10^{-8}$ (aa) $\sigma^2 = 10^{-4}$ (ab) $\sigma^2 = 10^{-2}$ (ac) $\sigma^2 = 10^{-1}$ (ad) $\sigma^2 = 1$

(row 6) Problem `HS48` with $\beta_t = 2/t^{0.6}$

Figure 6: 95% confidence intervals of $\boldsymbol{x}_1^\star + \boldsymbol{\lambda}_1^\star$. The first two rows correspond to `BYRDSPHR`; the middle two rows correspond to `HS7`; the last two rows correspond to `HS48`. Each row has five figures, from the left to the right, corresponding to $\sigma^2 \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}, 1\}$. Each figure has three lines. The blue and green lines correspond to the lower and upper interval boundaries, and the red line corresponds to the true value $\boldsymbol{x}_1^\star + \boldsymbol{\lambda}_1^\star$.

C. Boyer and A. Godichon-Baggioni. On the asymptotic rate of convergence of stochastic Newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*, 2020.

C. Chen, F. Tung, N. Vedula, and G. Mori. Constraint-aware deep neural network compression. In *Computer Vision – ECCV 2018*, pages 409–424. Springer International Publishing, 2018.

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

X. Chen, W. Liu, and Y. Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17, 2021.

F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2112.14799*, 2021a.

F. E. Curtis, D. P. Robinson, and B. Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*, 2021b.

M. Duflo. *Random iterative models*, volume 34. Springer, Berlin New York, 1997.

R. Durrett. *Probability*, volume 49. Cambridge University Press, 2019.

X. Fan. Exact rates of convergence in some martingale central limit theorems. *Journal of Mathematical Analysis and Applications*, 469(2):1028–1044, 2019.

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

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

A. J. Kleywegt, A. Shapiro, and T. H. de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

X. Liu and Y. Yuan. A sequential quadratic programming method without a penalty function or a filter for nonlinear equality constrained optimization. *SIAM Journal on Optimization*, 21(2): 545–571, 2011.

S. Lucidi. Recursive quadratic programming algorithm that uses an exact augmented lagrangian function. *Journal of Optimization Theory and Applications*, 67(2):227–245, 1990.

S. Na. Global convergence of online optimization for nonlinear model predictive control. *Advances in Neural Information Processing Systems*, 34, 2021.

S. Na, M. Anitescu, and M. Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *arXiv preprint arXiv:2102.05320*, 2021a.

S. Na, M. Anitescu, and M. Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *arXiv preprint arXiv:2109.11502*, 2021b.

S. Na, M. Anitescu, and M. Kolar. A fast temporal decomposition procedure for long-horizon nonlinear dynamic programming. *arXiv preprint arXiv:2107.11560*, 2021c.

S. Na, M. Dereziński, and M. W. Mahoney. Hessian averaging in stochastic Newton methods achieves superlinear convergence. *arXiv preprint arXiv:2204.09266*, 2022.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2nd edition, 2006.

A. E. Onuk, M. Akcakaya, J. P. Bardhan, D. Erdogmus, D. H. Brooks, and L. Makowski. Constrained maximum likelihood estimation of relative abundances of protein conformation in a heterogeneous mixture from small angle x-ray scattering intensity measurements. *IEEE Transactions on Signal Processing*, 63(20):5383–5394, 2015.

F. Oztoprak, R. Byrd, and J. Nocedal. Constrained optimization in the presence of noise. *arXiv preprint arXiv:2110.04355*, 2021.

G. C. Pflug and A. Pichler. *Multistage Stochastic Optimization*, volume 1104. Springer International Publishing, 2014.

G. D. Pillo and L. Grippo. A new class of augmented lagrangians in nonlinear programming. *SIAM Journal on Control and Optimization*, 17(5):618–628, 1979.

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

T. Rees, H. S. Dollar, and A. J. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010.

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

A. Ruszczyński and A. Shapiro. Stochastic programming models. *Handbooks in operations research and management science*, 10:1–64, 2003.

A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, 2014.

A. S. Siqueira and D. Orban. Cutest.jl. https://github.com/JuliaSmoothOptimizers/CUTEst.jl, 2020.

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

S. Sun and J. Nocedal. A trust region method for the optimization of noisy functions. *arXiv preprint arXiv:2201.00973*, 2022.

W. Sun and Y.-X. Yuan. *Optimization Theory and Methods*, volume 1. Kluwer Academic Publishers, 2006.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1, Ser. A):25–57, 2006.

J.-G. Wang. The asymptotic behavior of locally square integrable martingales. *The Annals of Probability*, 23(2):552–585, 1995.

P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1): 99–111, 1972.

P.-Å. Wedin. Perturbation theory for pseudo-inverses. *BIT*, 13(2):217–232, 1973.

W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–30, 2021.

A Auxiliary Lemmas

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

$$\lim_{i\to\infty} i\left(1-\frac{\varphi_{i-1}^p}{\varphi_i^p}\right) = p\cdot\varphi.$$

Proof. By the condition, we know

$$\frac{\varphi_{i-1}}{\varphi_i} = 1-\frac{\varphi}{i}+o\left(\frac{1}{i}\right).$$

Thus, we have

$$i\left(1-\frac{\varphi_{i-1}^p}{\varphi_i^p}\right) = i\left(1-\left\{1-\frac{\varphi}{i}+o\left(\frac{1}{i}\right)\right\}^p\right) = p\varphi+o(1).$$

This completes the proof. ∎

Lemma A.2. Let $\{\varphi_i\}_i$ be a positive sequence. If $\lim_{i\to\infty} i(1-\varphi_{i-1}/\varphi_i) = \varphi < 0$, then $\varphi_i \to 0$ as $i\to\infty$.

Proof. By Lemma A.1, we know for any positive constant p,

$$\lim_{i\to\infty} i\left(1-\frac{\varphi_{i-1}^p}{\varphi_i^p}\right) = p\cdot\varphi.$$

Choosing p large enough such that $p\varphi < -1$, the Raabe's test indicates that $\sum_{i=0}^{\infty}\varphi_i^p < \infty$. This implies $\varphi_i \to 0$ and we complete the proof. ∎

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

$$\lim_{i\to\infty} i\left(1-\frac{\phi_{i-1}}{\phi_i}\right) = \phi, \qquad \lim_{i\to\infty}\varphi_i = 0, \qquad \lim_{i\to\infty} i\varphi_i = \widetilde{\varphi} \tag{A.1}$$

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

$$\sum_{k=1}^{l} \sigma_k + \frac{\phi}{\widetilde{\varphi}} > 0,$$

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

$$\begin{aligned} &\frac{1}{\phi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i \longrightarrow \frac{1}{\sum_{k=1}^{l} \sigma_k + \phi/\widetilde{\varphi}}, \\ &\frac{1}{\phi_t} \left\{ \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i a_i + b \cdot \prod_{j=0}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \right\} \longrightarrow 0, \end{aligned} \tag{A.2}$$

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

Proof. For any scalar A, we have

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

For the last term, we have

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

Combining the above two displays, we obtain

$$\begin{aligned} &\frac{1}{\phi_t} \sum_{i=0}^{t} \prod_{j=i+1}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \varphi_i \phi_i - A \\ &= \frac{1}{\phi_t} \prod_{j=0}^{t} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k) \Bigg\{ \sum_{i=1}^{t} \prod_{j=0}^{i} \prod_{k=1}^{l} (1 - \varphi_j \sigma_k)^{-1} \phi_i \left\{ \varphi_i - A \left(1 - \frac{\phi_{i-1}}{\phi_i} \prod_{k=1}^{l} (1 - \varphi_i \sigma_k) \right) \right\} \\ &\quad + \phi_0 \prod_{k=1}^{l} (1 - \varphi_0 \sigma_k)^{-1} (\varphi_0 - A) \Bigg\}. \end{aligned} \tag{A.3}$$

We aim to select A such that the middle term in (A.3) is small. By (A.1), we know

$$\frac{\phi_{i-1}}{\phi_i} = 1 - \frac{\phi}{i} + o\left(\frac{1}{i}\right) = 1 - \frac{\phi}{\widetilde{\varphi}} \cdot \varphi_i + o(\varphi_i),$$

where the second equality is due to $1/(i\varphi_i) = 1/\widetilde{\varphi} + o(1)$ (which is true even if $\widetilde{\varphi} = \infty$). Furthermore,

$$\prod_{k=1}^{l}(1 - \varphi_i\sigma_k) = 1 - \varphi_i \sum_{k=1}^{l} \sigma_k + o(\varphi_i).$$

Combining the above two displays, we have

$$\begin{aligned}
\varphi_i - A\left(1 - \frac{\phi_{i-1}}{\phi_i}\prod_{k=1}^{l}(1 - \varphi_i\sigma_k)\right) &= \varphi_i - A\left\{1 - \left(1 - \frac{\phi}{\widetilde{\varphi}} \cdot \varphi_i + o(\varphi_i)\right)\left(1 - \varphi_i\sum_{k=1}^{l}\sigma_k + o(\varphi_i)\right)\right\} \\
&= \varphi_i - A\left(\frac{\phi}{\widetilde{\varphi}} + \sum_{k=1}^{l}\sigma_k\right)\varphi_i + o(\varphi_i). \qquad \text{(A.4)}
\end{aligned}$$

Thus, we let $A = 1/(\sum_{k=1}^{l}\sigma_k + \phi/\widetilde{\varphi})$ and (A.3) leads to

$$\begin{aligned}
&\frac{1}{\phi_t}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\varphi_i\phi_i - \frac{1}{\sum_{k=1}^{l}\sigma_k + \phi/\widetilde{\varphi}} \\
&\qquad = \frac{1}{\phi_t}\prod_{j=0}^{t}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\Bigg\{\sum_{i=1}^{t}\prod_{j=0}^{i}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)^{-1}\phi_i \cdot o(\varphi_i) \\
&\qquad\qquad + \phi_0\prod_{k=1}^{l}(1 - \varphi_0\sigma_k)^{-1}\left(\varphi_0 - \frac{1}{\sum_{k=1}^{l}\sigma_k + \phi/\widetilde{\varphi}}\right)\Bigg\}.
\end{aligned}$$

Comparing the above display with (A.2), we notice that the first result in (A.2) is implied by the second result. Thus, it suffices to prove the second result. We define

$$\Psi_t = \frac{1}{\phi_t}\left\{\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\varphi_i\phi_i a_i + b \cdot \prod_{j=0}^{t}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\right\}, \qquad \text{(A.5)}$$

then

$$\begin{aligned}
\Psi_t &= \frac{1}{\phi_t}\left\{\varphi_t\phi_t a_t + \prod_{k=1}^{l}(1 - \varphi_t\sigma_k)\left(\sum_{i=0}^{t-1}\prod_{j=i+1}^{t-1}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\varphi_i\phi_i a_i + b \cdot \prod_{j=0}^{t-1}\prod_{k=1}^{l}(1 - \varphi_j\sigma_k)\right)\right\} \\
&\overset{\text{(A.5)}}{=} \frac{\phi_{t-1}}{\phi_t}\prod_{k=1}^{l}(1 - \varphi_t\sigma_k)\Psi_{t-1} + \varphi_t a_t.
\end{aligned}$$

By (A.4), we know that

$$\frac{\phi_{t-1}}{\phi_t}\prod_{k=1}^{l}(1 - \varphi_t\sigma_k) = 1 - \left(\frac{\phi}{\widetilde{\varphi}} + \sum_{k=1}^{l}\sigma_k\right) \cdot \varphi_t + o(\varphi_t).$$

Since $\sum_{k=1}^{l} \sigma_k + \phi/\widetilde{\varphi} > 0$, we immediately know that for a constant $c > 0$ and for all large enough t,

$$|\Psi_t| \le (1 - c\varphi_t)|\Psi_{t-1}| + \varphi_t |a_t|.$$

Let t_1 be a fixed integer. We apply the above inequality recursively and have for any $t \ge t_1 + 1$,

$$|\Psi_t| \le \prod_{i=t_1+1}^{t} (1 - c\varphi_i)|\Psi_{t_1}| + \sum_{i=t_1+1}^{t} \prod_{j=i+1}^{t} (1 - c\varphi_j)\varphi_i |a_i|.$$

For any $\epsilon > 0$, since $a_i \to 0$, we select t_1 such that $|a_i| \le \epsilon$, for all $i \ge t_1$. Then, the above inequality leads to

$$\begin{aligned}|\Psi_t| \le \prod_{i=t_1+1}^{t} (1 - c\varphi_i)|\Psi_{t_1}| + \epsilon \sum_{i=t_1+1}^{t} \prod_{j=i+1}^{t} (1 - c\varphi_j)\varphi_i = \prod_{i=t_1+1}^{t} (1 - c\varphi_i)|\Psi_{t_1}| + \frac{\epsilon}{c}\{1 - \prod_{j=t_1+1}^{t} (1 - c\varphi_j)\} \\ \le |\Psi_{t_1}| \exp\left(-c \sum_{i=t_1+1}^{t} \varphi_i\right) + \frac{\epsilon}{c}.\end{aligned}$$

Since $n\varphi_i \to \widetilde{\varphi} \in (0, \infty]$, $\sum_t \varphi_t \to \infty$. Thus, for the above $\epsilon > 0$, there exists $t_2 > t_1$ such that $|\Psi_{t_1}| \exp\left(-c \sum_{i=t_1+1}^{t} \varphi_i\right) \le \epsilon/c$, $\forall t \ge t_2$, which implies $|\Psi_t| \le 2\epsilon/c$. This means $|\Psi_t| \to 0$ and we complete the proof. ■

Lemma A.4. For any scalars a, b, we have $P(a < \mathcal{N}(0,1) \le b) \le b - a$. Furthermore, if $0 < a \le b$, then $P(a < \mathcal{N}(0,1) \le b) \le b/a - 1$.

Proof. The first part of statement holds naturally due to the fact that the density function of the standard Gaussian $\exp(-t^2/2)/\sqrt{2\pi} \le 1$ for any $t \in \mathbb{R}$. Moreover, for $0 < a \le b$, we have

$$\begin{aligned}P(a < \mathcal{N}(0,1) \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\, dt \le \frac{b-a}{\sqrt{2\pi}} \exp(-a^2/2) \\ = \left(\frac{b}{a} - 1\right) \frac{a}{\sqrt{2\pi}} \exp(-a^2/2) \le \frac{b}{a} - 1,\end{aligned}$$

where the last inequality is due to the fact that $a \exp(-a^2/2) \le 1$ for all a; the first inequality uses $0 < a \le b$. This completes the proof. ■

Lemma A.5. Let A_t, B_t, C_t be three variables depending on the index t; let $\Phi(z) = P(\mathcal{N}(0,1) \le z)$ be the cumulative distribution function of standard Gaussian variable. Suppose for the index t,

$$\sup_{z \in \mathbb{R}} |P(A_t \le z) - \Phi(z)| \le a_t, \quad |B_t| \le b_t, \quad |C_t| \le c_t \quad \text{almost surely} \tag{A.6}$$

where $a_t, b_t \ge 0$ and $0 \le c_t < 1$. Then, we have

$$\sup_{z \in \mathbb{R}} \left| P\left(\frac{A_t + B_t}{\sqrt{1 + C_t}} \le z\right) - \Phi(z) \right| \le a_t + b_t + \frac{c_t}{\sqrt{1 - c_t}}.$$

Proof. We only prove the result for any $z > 0$. The result of $z \le 0$ can be shown in the same way. We know from (A.6) that

$$\frac{A_t - b_t}{\sqrt{1+c_t}} \le \frac{A_t + B_t}{\sqrt{1+C_t}} \le \frac{A_t + b_t}{\sqrt{1-c_t}} \qquad \text{almost surely.}$$

Therefore, we have

$$\begin{aligned}
&P\left(\frac{A_t + B_t}{\sqrt{1+C_t}} \le z\right) \ge P\left(\frac{A_t + b_t}{\sqrt{1-c_t}} \le z\right) = P\left(A_t \le z(1-c_t)^{1/2} - b_t\right) \overset{(A.6)}{\ge} \Phi(z(1-c_t)^{1/2} - b_t) - a_t \\
&= \Phi(z) - P\left(z(1-c_t)^{1/2} - b_t < \mathcal{N}(0,1) \le z(1-c_t)^{1/2}\right) - P\left(z(1-c_t)^{1/2} < \mathcal{N}(0,1) \le z\right) - a_t \quad (z \ge 0) \\
&\ge \Phi(z) - b_t - \left(\frac{1}{\sqrt{1-c_t}} - 1\right) - a_t \quad \text{(by Lemma A.4)} \\
&\ge \Phi(z) - b_t - \frac{c_t}{\sqrt{1-c_t}} - a_t.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&P\left(\frac{A_t + B_t}{\sqrt{1+C_t}} \le z\right) \le P\left(\frac{A_t - b_t}{\sqrt{1+c_t}} \le z\right) = P\left(A_t \le z(1+c_t)^{1/2} + b_t\right) \overset{(A.6)}{\le} \Phi(z(1+c_t)^{1/2} + b_t) + a_t \\
&= \Phi(z) + P\left(z < \mathcal{N}(0,1) \le z(1+c_t)^{1/2}\right) + P\left(z(1+c_t)^{1/2} < \mathcal{N}(0,1) \le z(1+c_t)^{1/2} + b_t\right) + a_t \\
&\le \Phi(z) + \left((1+c_t)^{1/2} - 1\right) + b_t + a_t \quad \text{(by Lemma A.4)} \\
&\le \Phi(z) + c_t + b_t + a_t.
\end{aligned}$$

Combining the above two displays completes the proof. ∎

B Proofs of Section 3

B.1 Proof of Lemma 3.5

By (11) and the definition of $(\Delta\boldsymbol{x}_t, \Delta\boldsymbol{\lambda}_t)$, we have

$$\begin{aligned}
&\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(11)}{=} \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} I + \nu\nabla_{\boldsymbol{x}}^2\mathcal{L}_t & \mu G_t^T \\ \nu G_t & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix} \\
&\overset{(3)}{=} -\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} I + \nu\nabla_{\boldsymbol{x}}^2\mathcal{L}_t & \mu G_t^T \\ \nu G_t & I \end{pmatrix} \begin{pmatrix} B_t & G_t^T \\ G_t & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \\
&= -\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} B_t + \nu\nabla_{\boldsymbol{x}}^2\mathcal{L}_t B_t + \mu G_t^T G_t & G_t^T + \nu\nabla_{\boldsymbol{x}}^2\mathcal{L}_t G_t^T \\ G_t + \nu G_t B_t & \nu G_t G_t^T \end{pmatrix} \begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(13)}{\le} -\Delta\boldsymbol{x}_t^T B_t \Delta\boldsymbol{x}_t + \nu\Upsilon_B\Upsilon_u\|\Delta\boldsymbol{x}_t\|^2 - \mu\|G_t\Delta\boldsymbol{x}_t\|^2 - 2\Delta\boldsymbol{\lambda}_t^T G_t\Delta\boldsymbol{x}_t \\
&\quad + \nu(\Upsilon_u + \Upsilon_B)\|\Delta\boldsymbol{x}_t\|\|G_t^T\Delta\boldsymbol{\lambda}_t\| - \nu\|G_t^T\Delta\boldsymbol{\lambda}_t\|^2 \quad \text{(also use Assumption 3.1)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(3)}{\leq} -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu \Upsilon_B \Upsilon_u \|\Delta \boldsymbol{x}_t\|^2 - \mu\|c_t\|^2 + 2c_t^T \Delta \boldsymbol{\lambda}_t + \frac{\nu(\Upsilon_u + \Upsilon_B)^2}{2}\|\Delta \boldsymbol{x}_t\|^2 \\
&\quad - \frac{\nu}{2}\|G_t^T \Delta \boldsymbol{\lambda}_t\|^2 \quad \text{(also use Young's inequality)} \\
&\leq -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 - \mu\|c_t\|^2 + \frac{8}{\nu\gamma_G}\|c_t\|^2 + \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 \\
&\quad - \frac{\nu\gamma_G}{4}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{4}\|G_t^T \Delta \boldsymbol{\lambda}_t\|^2 \quad \text{(Young's inequality and Assumption 3.1)} \\
&\overset{(3)}{=} -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu\gamma_G}\right)\|c_t\|^2 - \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{4}\|B_t\Delta \boldsymbol{x}_t + \nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2 \\
&\leq -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu\gamma_G}\right)\|c_t\|^2 - \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{8}\|\nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2 + \frac{\nu\Upsilon_B^2}{4}\|\Delta \boldsymbol{x}_t\|^2 \\
&\leq -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu\gamma_G}\right)\|c_t\|^2 - \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{8}\|\nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2, \qquad \text{(B.1)}
\end{aligned}
$$

where the second last inequality uses $\|B_t\Delta \boldsymbol{x}_t + \nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2 \geq \|\nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2/2 - \|B_t\Delta \boldsymbol{x}_t\|^2 \geq \|\nabla_{\boldsymbol{x}}\mathcal{L}_t\|^2/2 - \Upsilon_B^2\|\Delta \boldsymbol{x}_t\|^2$. To further simplify (B.1), we decompose the step $\Delta \boldsymbol{x}_t$ as

$$
\Delta \boldsymbol{x}_t = \Delta \boldsymbol{u}_t + \Delta \boldsymbol{v}_t, \qquad \text{where } \Delta \boldsymbol{u}_t \in \text{span}(G_t^T) \text{ and } G_t\Delta \boldsymbol{v}_t = \boldsymbol{0}.
$$

Then, we have

$$
\begin{aligned}
&-\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 \\
&= -\Delta \boldsymbol{u}_t^T B_t \Delta \boldsymbol{u}_t - 2\Delta \boldsymbol{u}_t^T B_t \Delta \boldsymbol{v}_t - \Delta \boldsymbol{v}_t^T B_t \Delta \boldsymbol{v}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 \\
&\leq \Upsilon_B\|\Delta \boldsymbol{u}_t\|^2 + 2\Upsilon_B\|\Delta \boldsymbol{u}_t\|\|\Delta \boldsymbol{v}_t\| - \gamma_{RH}\|\Delta \boldsymbol{v}_t\|^2 + 2\nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 \quad \text{(Assumption 3.1)} \\
&\leq \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}}\right)\|\Delta \boldsymbol{u}_t\|^2 - \frac{\gamma_{RH}}{2}\|\Delta \boldsymbol{v}_t\|^2 + 2\nu(\Upsilon_B + \Upsilon_u)^2\|\Delta \boldsymbol{x}_t\|^2 \quad \text{(Young's inequality)} \\
&= \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}} + \frac{\gamma_{RH}}{2}\right)\|\Delta \boldsymbol{u}_t\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2\right)\|\Delta \boldsymbol{x}_t\|^2 \\
&\leq \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}} + \frac{\gamma_{RH}}{2}\right)\frac{1}{\gamma_G}\|c_t\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2\right)\|\Delta \boldsymbol{x}_t\|^2 \\
&\leq \frac{4\Upsilon_B^2}{\gamma_{RH}\gamma_G}\|c_t\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2\right)\|\Delta \boldsymbol{x}_t\|^2 \qquad \text{(use } \gamma_{RH} \leq 1 \leq \Upsilon_B\text{)},
\end{aligned}
$$

where the second last inequality uses the fact that $\Delta \boldsymbol{u}_t = G_t^T \Delta \bar{\boldsymbol{u}}_t$ for some vector $\Delta \bar{\boldsymbol{u}}_t \in \mathbb{R}^m$, and $\|c_t\|^2 \overset{(3)}{=} \|G_t\Delta \boldsymbol{x}_t\|^2 = \|G_t\Delta \boldsymbol{u}_t\|^2 = \|G_tG_t^T\Delta \bar{\boldsymbol{u}}_t\|^2 \geq \gamma_G\|G_t^T\bar{\boldsymbol{u}}_t\|^2 = \gamma_G\|\Delta \boldsymbol{u}_t\|^2$, and the inequality is due to Assumption 3.1. Combining the above display with (B.1), we have

$$
\begin{aligned}
\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \leq -\frac{\nu\gamma_G}{8}\left\|\begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 - \frac{\nu}{8}\left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2 - \frac{\nu\gamma_G}{8}\right)\|\Delta \boldsymbol{x}_t\|^2 \\
- \left(\mu - \frac{8}{\nu\gamma_G} - \frac{4\Upsilon_B^2}{\gamma_{RH}\gamma_G} - \frac{\nu}{8}\right)\|c_t\|^2.
\end{aligned}
$$

Thus, the statements hold provided

$$
\frac{\gamma_{RH}}{2} \geq 2\nu(\Upsilon_B + \Upsilon_u)^2 + \frac{\nu\gamma_G}{8} \quad \text{and} \quad \mu \geq \frac{8}{\nu\gamma_G} + \frac{4\Upsilon_B^2}{\gamma_{RH}\gamma_G} + \frac{\nu}{8},
$$

which is implied by (16) using $(\gamma_G \vee \gamma_{RH}) \leq 1 \leq (\Upsilon_B \wedge \Upsilon_u)$. This completes the proof.

B.2 Proof of Lemma 3.6

By (8) and (14), we have

$$\begin{aligned}
\mathcal{L}_{\mu,\nu}^{t+1} &\leq \mathcal{L}_{\mu,\nu}^{t} + \bar{\alpha}_t \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} + \frac{\Upsilon_{\mu,\nu}\bar{\alpha}_t^2}{2}\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 \\
&= \mathcal{L}_{\mu,\nu}^{t} + \bar{\alpha}_t \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T (I+C_t)\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \left\{\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} - (I+C_t)\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\} \\
&\quad + \frac{\Upsilon_{\mu,\nu}\bar{\alpha}_t^2}{2}\left\|\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2, \qquad \text{(B.2)}
\end{aligned}$$

where C_t is defined in Lemma 3.4(b). Furthermore, by Lemma 3.5 and Lemma 3.4, we have

$$\begin{aligned}
&\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T (I+C_t)\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \\
&\leq -\frac{\nu\gamma_G}{8}\left\{\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2\right\} + \|C_t\|\left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}\right\|\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\| \quad \text{(Lemma 3.5)} \\
&\overset{(11),(13)}{\leq} -\frac{\nu\gamma_G}{8}\left\{\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2\right\} + \rho^\tau\left(1+(2\nu+\mu)\Upsilon_u\right)\left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\| \quad \text{(Lemma 3.4)} \\
&\leq -\frac{\nu\gamma_G}{8}\left\{\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2\right\} + 2\rho^\tau\mu\Upsilon_u\left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\| \quad (1\leq\Upsilon_u \text{ and } 1+2\nu\leq\mu) \\
&\leq -\left(\frac{\nu\gamma_G}{8} - \rho^\tau\mu\Upsilon_u\right)\left\{\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2\right\}.
\end{aligned}$$

Thus, if $\rho^\tau \leq \nu\gamma_G/(16\mu\Upsilon_u)$, we have

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T (I+C_t)\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix} \leq -\frac{\nu\gamma_G}{16}\left\{\left\|\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 + \left\|\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2\right\}. \qquad \text{(B.3)}$$

Moreover, we deal with the last two terms on the right hand side of (B.2). We note that

$$\begin{aligned}
&\mathbb{E}\left[\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \mid \mathcal{F}_{t-2/3}\right] \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[(I+C_t)\begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \mid \mathcal{F}_{t-1}\right] \quad \text{(Lemma 3.4(b))} \\
&\overset{(3)}{=} -(I+C_t)K_t^{-1}\mathbb{E}\left[\begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix} \mid \mathcal{F}_{t-1}\right] = -(I+C_t)K_t^{-1}\begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix} \quad \text{(Assumption 3.2)} \\
&\overset{(3)}{=} (I+C_t)\begin{pmatrix} \Delta\boldsymbol{x}_t \\ \Delta\boldsymbol{\lambda}_t \end{pmatrix}, \qquad \text{(B.4)}
\end{aligned}$$

and

$$\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]$$
$$\le 3\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t-\widetilde{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t-\widetilde{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]+3\mathbb{E}\left[\left\|\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_t-\Delta\boldsymbol{x}_t\\ \widetilde{\Delta}\boldsymbol{\lambda}_t-\Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]+3\|C_t\|^2\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2$$
$$\le 3\rho^\tau\mathbb{E}\left[\left\|\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_t\\ \widetilde{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]+3\mathbb{E}\left[\left\|\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_t-\Delta\boldsymbol{x}_t\\ \widetilde{\Delta}\boldsymbol{\lambda}_t-\Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]+3\rho^{2\tau}\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\quad\text{(Lemma 3.4(b,c))}$$
$$=3(\rho^\tau+\rho^{2\tau})\left\|\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2+3(1+\rho^\tau)\mathbb{E}\left[\left\|\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_t-\Delta\boldsymbol{x}_t\\ \widetilde{\Delta}\boldsymbol{\lambda}_t-\Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]\quad\text{(bias-variance decomposition)}$$
$$\overset{(3),(13)}{\le}3(\rho^\tau+\rho^{2\tau})\Upsilon_K^2\Upsilon_u^2+3(1+\rho^\tau)\Upsilon_K^2\mathbb{E}[\|\bar{g}_t-\nabla f_t\|^2\mid\mathcal{F}_{t-1}]\quad\text{(also use }\|K_t^{-1}\|\le\Upsilon_K)$$
$$\le 3(1+\rho^\tau)\Upsilon_K^2(\rho^\tau\Upsilon_u^2+\Upsilon_m)\quad\text{(Assumption 3.2(15a)).}\tag{B.5}$$

Thus, using (B.4) and (B.5), we know

$$\mathbb{E}\left[\bar{\alpha}_t\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^t_{\mu,\nu}\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}^t_{\mu,\nu}\end{pmatrix}^T\left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\}\mid\mathcal{F}_{t-1}\right]$$
$$\overset{(B.4)}{=}\mathbb{E}\left[\left(\bar{\alpha}_t-\frac{\beta_t+\eta_t}{2}\right)\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^t_{\mu,\nu}\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}^t_{\mu,\nu}\end{pmatrix}^T\left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\}\mid\mathcal{F}_{t-1}\right]$$
$$\overset{(9)}{\le}\frac{\eta_t-\beta_t}{2}\mathbb{E}\left[\left\|\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}^t_{\mu,\nu}\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}^t_{\mu,\nu}\end{pmatrix}\right\|\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|\mid\mathcal{F}_{t-1}\right]$$
$$\overset{(13)}{\le}\frac{\eta_t-\beta_t}{2}(1+(2\nu+\mu)\Upsilon_u)\Upsilon_u\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|\mid\mathcal{F}_{t-1}\right]$$
$$\le(\eta_t-\beta_t)\mu\Upsilon_u^2\sqrt{\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]}\quad(1\le\Upsilon_u\text{ and }1+2\nu\le\mu)$$
$$\overset{(B.5)}{\le}2\mu\Upsilon_K\Upsilon_u^2(1+\rho^\tau)(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t)\le 4\mu\Upsilon_K\Upsilon_u^2(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t),\tag{B.6}$$

and

$$\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]$$
$$\overset{(B.4)}{=}\left\|(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2+\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]$$
$$\overset{(3),(13)}{\le}(1+\rho^\tau)^2\Upsilon_K^2\Upsilon_u^2+\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}-(I+C_t)\begin{pmatrix}\Delta\boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\mid\mathcal{F}_{t-1}\right]\quad\text{(also use Lemma 3.4(b))}$$
$$\overset{(B.5)}{\le}(1+\rho^\tau)^2\Upsilon_K^2\Upsilon_u^2+3(1+\rho^\tau)\Upsilon_K^2(\rho^\tau\Upsilon_u^2+\Upsilon_m)\le 16\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m).\tag{B.7}$$

Combining (B.7) with (B.6) and (B.3), plugging into (B.2), and using (9), we obtain

$$\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}_{\mu,\nu}^{t} - \frac{\nu\gamma_G\beta_t}{16}\left\{\left\|\begin{pmatrix}\Delta \boldsymbol{x}_t\\ \Delta\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2 + \left\|\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_t\\ c_t\end{pmatrix}\right\|^2\right\}$$
$$+ 4\mu\Upsilon_K\Upsilon_u^2(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t) + 8\Upsilon_{\mu,\nu}\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m)\eta_t^2.$$

This completes the proof.

B.3 Proof of Theorem 3.7

Note that the condition of τ in (17) implies that we can select (μ,ν) to satisfy (16) and have $\rho^\tau \leq \nu\gamma_G/(16\mu\Upsilon_u)$ with $\rho = 1-\gamma_S$. Thus, by Lemma 3.6, we have

$$\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1} - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}_{\mu,\nu}^{t} - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} - \frac{\nu\gamma_G\beta_t}{16}\|\nabla\mathcal{L}_t\|^2 + \widetilde{\Upsilon}_{\mu,\nu}(\chi_t+\eta_t^2).$$

By Robbins-Siegmund theorem (see Robbins and Siegmund (1971) or (Duflo, 1997, Theorem 1.3.12)), we immediately obtain $\sum_t \beta_t\|\nabla\mathcal{L}_t\|^2 < \infty$. Since $\sum_t\beta_t = \infty$, we have $\liminf_{t\to\infty}\|\nabla\mathcal{L}_t\| = 0$. Furthermore, we observe that

$$\mathbb{E}\left[\left\|\begin{pmatrix}\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\\ \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\begin{pmatrix}\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\\ \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2 \mid \mathcal{F}_{t-1}\right]\right]$$
$$\overset{(8)}{\leq} \eta_t^2\mathbb{E}\left[\mathbb{E}\left[\left\|\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_t\\ \bar{\Delta}\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2 \mid \mathcal{F}_{t-1}\right]\right] \overset{(B.7)}{\leq} 16\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m)\cdot\eta_t^2. \quad \text{(B.8)}$$

Summing over $t=1$ to ∞, exchanging the expectation and sum by applying Fubini's theorem (Durrett, 2019, Theorem 1.7.2), and noting that $\sum_t\eta_t^2 < \infty$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{\infty}\left\|\begin{pmatrix}\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\\ \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\end{pmatrix}\right\|^2\right] < \infty.$$

This implies $\sum_{t=1}^\infty \|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t)\| < \infty$ almost surely and, thus, $\|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t)\| \to 0$ as $t\to\infty$ almost surely. Suppose $\lim_{t\to\infty}\|\nabla\mathcal{L}_t\| \neq 0$, then $\limsup_{t\to\infty}\|\nabla\mathcal{L}_t\| = \epsilon > 0$. Then, there exist two index sequences $\{t_{1,i}\}_i, \{t_{2,i}\}_i$ with $t_{1,i+1} > t_{2,i} > t_{1,i}$, and for all $i = 1,2,\ldots$, we have

$$\|\nabla\mathcal{L}_{t_{1,i}}\| \geq \epsilon/2, \qquad \|\nabla\mathcal{L}_j\| \geq \epsilon/3 \text{ for } j = t_{1,i}+1,\ldots,t_{2,i}-1, \qquad \|\nabla\mathcal{L}_{t_{i,2}}\| < \epsilon/3. \quad \text{(B.9)}$$

Since $\sum_t\beta_t\|\nabla\mathcal{L}_t\|^2 < \infty$, we know

$$\infty > \sum_{i=1}^{\infty}\sum_{j=t_{i,1}}^{t_{i,2}-1}\beta_j\|\nabla\mathcal{L}_j\|^2 \overset{(B.9)}{\geq} \frac{\epsilon^2}{9}\sum_{i=1}^{\infty}\sum_{j=t_{i,1}}^{t_{i,2}-1}\beta_j. \quad \text{(B.10)}$$

Furthermore, by (B.8), we have

$$\mathbb{E}\left[\left\|\begin{pmatrix}\boldsymbol{x}_{t_{2,i}}-\boldsymbol{x}_{t_{1,i}}\\ \boldsymbol{\lambda}_{t_{2,i}}-\boldsymbol{\lambda}_{t_{1,i}}\end{pmatrix}\right\|\right] \overset{(B.8)}{\leq} 4\Upsilon_K(\Upsilon_u\vee\sqrt{\Upsilon_m})\sum_{j=t_{i,1}}^{t_{i,2-1}}\eta_j \overset{(9)}{=} 4\Upsilon_K(\Upsilon_u\vee\sqrt{\Upsilon_m})\left\{\sum_{j=t_{i,1}}^{t_{i,2}-1}\beta_j + \sum_{j=t_{i,1}}^{t_{i,2}-1}\chi_j\right\}.$$

Summing over $i = 1$ to ∞, and noting that $\sum_{i=1}^{\infty}\sum_{j=t_{i,1}}^{t_{i,2-1}} \beta_j < \infty$ by (B.10) and $\sum_{i=1}^{\infty}\sum_{j=t_{i,1}}^{t_{i,2-1}} \chi_j \le \sum_{j=1}^{\infty} \chi_j < \infty$, we exchange the expectation and sum by applying Fubini's theorem again. We know that the sequence $\{(\boldsymbol{x}_{t_{2,i}} - \boldsymbol{x}_{t_{1,i}}, \boldsymbol{\lambda}_{t_{2,i}} - \boldsymbol{\lambda}_{t_{1,i}})\}_i$ converges to zero as $i \to \infty$ with probability one. This contradicts with $\|\nabla\mathcal{L}_{t_{1,i}}\| \ge \epsilon/2$ and $\|\nabla\mathcal{L}_{t_{i,2}}\| < \epsilon/3$ in (B.9). We complete the proof.

B.4 Proof of Corollary 3.8

Applying Lemma 3.6 and taking full expectation, we know for some constants $h_1, h_2 > 0$ that

$$\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1}] \le \mathbb{E}[\mathcal{L}_{\mu,\nu}^{t}] - h_1\beta_t\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] + h_2(\chi_t + \eta_t^2), \quad \forall t \ge 0.$$

Rearranging the inequality and summing over $t = 0$ to $\mathcal{T}_\epsilon - 1$, we obtain

$$\begin{aligned}
h_1\sum_{t=0}^{\mathcal{T}_\epsilon-1}\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] &\le \sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{1}{\beta_t}\left((\mathbb{E}[\mathcal{L}_{\mu,\nu}^t] - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu}) - (\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1}] - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu})\right) + h_2\sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t}\\
&\le \frac{\mathbb{E}[\mathcal{L}_{\mu,\nu}^0] - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu}}{\beta_0} + \sum_{t=1}^{\mathcal{T}_\epsilon-1}\left(\frac{1}{\beta_t} - \frac{1}{\beta_{t-1}}\right)(\mathbb{E}[\mathcal{L}_{\mu,\nu}^t] - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu}) + h_2\sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t}.
\end{aligned}$$

Using the fact that $\min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} \le \mathbb{E}[\mathcal{L}_{\mu,\nu}^t] \le \max_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu}$ and denoting $\Delta\mathcal{L}_{\mu,\nu} = \max_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu}$, we further have

$$\begin{aligned}
h_1\sum_{t=0}^{\mathcal{T}_\epsilon-1}\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] &\le (\Delta\mathcal{L}_{\mu,\nu}\vee h_2)\left\{\frac{1}{\beta_0} + \sum_{t=1}^{\mathcal{T}_\epsilon-1}\left(\frac{1}{\beta_t} - \frac{1}{\beta_{t-1}}\right) + \sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t}\right\}\\
&= (\Delta\mathcal{L}_{\mu,\nu}\vee h_2)\left\{\frac{1}{\beta_{\mathcal{T}_\epsilon-1}} + \sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t}\right\} = (\Delta\mathcal{L}_{\mu,\nu}\vee h_2)\left\{\mathcal{T}_\epsilon^a + \sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t}\right\}.
\end{aligned}$$

For the last term on the right hand side, we have

$$\begin{aligned}
\sum_{t=0}^{\mathcal{T}_\epsilon-1}\frac{\chi_t+\eta_t^2}{\beta_t} &= \sum_{t=0}^{\mathcal{T}_\epsilon-1}\left\{(t+1)^{a-b} + (t+1)^a\left((t+1)^{-2a} + 2(t+1)^{-(a+b)} + (t+1)^{-2b}\right)\right\}\\
&\le \sum_{t=0}^{\mathcal{T}_\epsilon-1}(t+1)^{a-b} + 4\sum_{t=0}^{\mathcal{T}_\epsilon-1}(t+1)^{-a} = 5 + \sum_{t=1}^{\mathcal{T}_\epsilon-1}\left\{(t+1)^{a-b} + 4(t+1)^{-a}\right\}\\
&\le 5 + \int_0^{\mathcal{T}_\epsilon-1}(t+1)^{a-b} + 4(t+1)^{-a}dt \quad \text{(by the convexity of } x^p \text{ with } p<0)\\
&\le \begin{cases} 5 + \frac{\mathcal{T}_\epsilon^{1+a-b}}{1+a-b} + \frac{4\mathcal{T}_\epsilon^{1-a}}{1-a} & \text{if } 1+a>b,\\ 5 + \log(\mathcal{T}_\epsilon) + \frac{4\mathcal{T}_\epsilon^{1-a}}{1-a} & \text{if } 1+a=b,\\ 5 + \frac{1}{b-a-1} + \frac{4\mathcal{T}_\epsilon^{1-a}}{1-a} & \text{if } 1+a<b. \end{cases}
\end{aligned}$$

Combining the above two displays, dividing $\mathcal{T}_\epsilon$ on both sides, and using "$\lesssim$" to neglect constant factors (i.e., not depending on $\mathcal{T}_\epsilon$), we obtain

$$\begin{aligned}
\epsilon^2 &\leq \left(\frac{1}{\mathcal{T}_\epsilon}\sum_{t=0}^{\mathcal{T}_\epsilon-1}\mathbb{E}[\|\nabla\mathcal{L}_t\|]\right)^2 \leq \frac{1}{\mathcal{T}_\epsilon}\sum_{t=0}^{\mathcal{T}_\epsilon-1}(\mathbb{E}[\|\nabla\mathcal{L}_t\|])^2 \leq \frac{1}{\mathcal{T}_\epsilon}\sum_{t=0}^{\mathcal{T}_\epsilon-1}\mathbb{E}[\|\nabla\mathcal{L}_t\|^2] \\
&\lesssim \begin{cases} \frac{1}{\mathcal{T}_\epsilon^{1-a}} + \frac{1}{\mathcal{T}_\epsilon^{b-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} & \text{if } \ 1+a > b, \\ \frac{1}{\mathcal{T}_\epsilon^{1-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} & \text{if } \ 1+a = b, \\ \frac{1}{\mathcal{T}_\epsilon^{1-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} & \text{if } \ 1+a < b, \end{cases} \quad (\text{use } 1/\mathcal{T}_\epsilon \leq 1/\mathcal{T}_\epsilon^a \text{ and } \log(\mathcal{T}_\epsilon)/\mathcal{T}_\epsilon \lesssim 1/\mathcal{T}_\epsilon^a), \\
&\lesssim \begin{cases} \frac{1}{\mathcal{T}_\epsilon^{b-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} & \text{if } \ 1 > b, \\ \frac{1}{\mathcal{T}_\epsilon^{1-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} & \text{if } \ 1 \leq b, \end{cases} = \frac{1}{\mathcal{T}_\epsilon^{(1\wedge b)-a}} + \frac{1}{\mathcal{T}_\epsilon^{a}} \lesssim \frac{1}{\mathcal{T}_\epsilon^{a\wedge(1-a)\wedge(b-a)}}.
\end{aligned}$$

This completes the proof.

B.5 Proof of Theorem 3.12

Using $\|\Delta_t\| \leq \omega_t$ in Assumption 3.11, we have

$$\begin{aligned}
\|K_t - K^\star\| &\overset{(4)}{=} \left\|\begin{pmatrix} B_t - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star & (G_t)^T - (G^\star)^T \\ G_t - G^\star & \mathbf{0} \end{pmatrix}\right\| \overset{(2)}{\leq} \left\|\begin{pmatrix} \frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star & (G_t)^T - (G^\star)^T \\ G_t - G^\star & \mathbf{0} \end{pmatrix}\right\| + w_t \\
&\leq \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\right\| + \|G_t - G^\star\| + \omega_t \\
&\leq \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| + \frac{1}{t}\sum_{i=0}^{t-1}\|\nabla_{\boldsymbol{x}}^2\mathcal{L}_i - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\| + \|G_t - G^\star\| + \omega_t \\
&\overset{(12)}{\leq} \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| + \frac{\Upsilon_L}{t}\sum_{i=0}^{t-1}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \omega_t.
\end{aligned} \tag{B.11}$$

Since $(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \to \mathbf{0}$ as $t \to \infty$, and by the fact that $a_t \to a$ implies $\frac{1}{t}\sum_{i=0}^{t-1} a_i \to a$ for any sequence a_t (known as Stolz–Cesàro theorem), it suffices to show that $(\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i)/t$ converges. In fact, noting from Assumption 3.2(15d) that $\mathbb{E}[\bar{H}_i \mid \mathcal{F}_{i-1}] = \nabla^2 f_i$ and $\mathbb{E}[\|\bar{H}_i - \nabla^2 f_i\|^2 \mid \mathcal{F}_{i-1}] \leq \Upsilon_m$, we know $(\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i)/t$ is a square integrable martingale. Thus, (Duflo, 1997, Theorem 1.3.15) suggests that for any $\upsilon > 0$,

$$\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| = o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right). \tag{B.12}$$

Combining (B.11) and (B.12), we complete the proof.

C Proofs of Section 4

C.1 Proof of Lemma 4.1

Using the scheme in Algorithm 1, we have

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star \end{pmatrix} &\overset{(8)}{=} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \varphi_t \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \varphi_t(I + C_t) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} + \varphi_t \left\{ \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} - (I + C_t) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \right\} + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(3)}{=} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} - \varphi_t(I + C_t) K_t^{-1} \begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix} + \varphi_t \left\{ \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} - (I + C_t) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \right\} + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} - \varphi_t(I + C_t) K_t^{-1} \begin{pmatrix} \nabla_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix} - \varphi_t(I + C_t) K_t^{-1} \begin{pmatrix} \bar{g}_t - \nabla f_t \\ \mathbf{0} \end{pmatrix} \\
&\quad + \varphi_t \left\{ \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} - (I + C_t) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \right\} + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(21b)}{=} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} - \varphi_t(I + C_t) K_t^{-1} \nabla\mathcal{L}_t + \varphi_t \boldsymbol{\theta}^t + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} - \varphi_t(I + C_t)(K^\star)^{-1} \nabla\mathcal{L}_t - \varphi_t(I + C_t) \left\{ K_t^{-1} - (K^\star)^{-1} \right\} \nabla\mathcal{L}_t + \varphi_t \boldsymbol{\theta}^t + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(21d)}{=} \{I - \varphi_t(I + C_t)\} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} - \varphi_t(I + C_t)(K^\star)^{-1} \boldsymbol{\psi}^t - \varphi_t(I + C_t) \left\{ K_t^{-1} - (K^\star)^{-1} \right\} \nabla\mathcal{L}_t \\
&\quad + \varphi_t \boldsymbol{\theta}^t + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&\overset{(21c)}{=} \{I - \varphi_t(I + C^\star)\} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \varphi_t(\boldsymbol{\theta}^t + \boldsymbol{\delta}^t) + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}.
\end{aligned}$$

We then apply the above result recursively and have

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star \end{pmatrix} &= \{I - \varphi_t(I + C^\star)\} \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \varphi_t(\boldsymbol{\theta}^t + \boldsymbol{\delta}^t) + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \{I - \varphi_t(I + C^\star)\} \left\{ \{I - \varphi_{t-1}(I + C^\star)\} \begin{pmatrix} \boldsymbol{x}_{t-1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^\star \end{pmatrix} + \varphi_{t-1}(\boldsymbol{\theta}^{t-1} + \boldsymbol{\delta}^{t-1}) \right. \\
&\quad \left. + (\bar{\alpha}_{t-1} - \varphi_{t-1}) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_{t-1} \\ \bar{\Delta}\boldsymbol{\lambda}_{t-1} \end{pmatrix} \right\} + \varphi_t(\boldsymbol{\theta}^t + \boldsymbol{\delta}^t) + (\bar{\alpha}_t - \varphi_t) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_t \\ \bar{\Delta}\boldsymbol{\lambda}_t \end{pmatrix} \\
&= \cdots = \mathcal{I}_{1,t} + \mathcal{I}_{2,t} + \mathcal{I}_{3,t}.
\end{aligned}$$

This shows the first part of the result. Moreover, if Assumptions 3.2, 3.3 hold, we know $\mathbb{E}[\bar{g}_i - \nabla f_i \mid \mathcal{F}_{i-1}] = \mathbf{0}$ by Assumption 3.2, and

$$\mathbb{E}\left[\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} \mid \mathcal{F}_{i-2/3}\right] \mid \mathcal{F}_{i-1}\right] = \mathbf{0} \tag{C.1}$$

by Lemma 3.4(b). Thus, $\mathbb{E}[\boldsymbol{\theta}^i \mid \mathcal{F}_{i-1}] = \mathbf{0}$ and $\boldsymbol{\theta}^i$ is a martingale difference.

C.2 Proof of Lemma 4.2

Let us denote $\text{rank}(S) = r$ in the proof. Since K_t, $K^\star$ have full rank, we know $\text{rank}(K_tS) = \text{rank}(K^\star S) = r$. Let $K^\star S = EDF^T$ be truncated singular value decomposition of $K^\star S$. We have

$$E \in \mathbb{R}^{(d+m)\times r}, \quad F \in \mathbb{R}^{q\times r}, \quad E^TE = F^TF = I, \quad D = \text{diag}(D_1, \ldots, D_r) \quad \text{with } D_1 \geq \ldots \geq D_r > 0.$$

Similarly, we let $K_tS = E'D'(F')^T$. By direct calculation, we have

$$\|K_tS(S^TK_t^2S)^\dagger S^TK_t - K^\star S(S^T(K^\star)^2S)^\dagger S^TK^\star\| = \|EE^T - E'(E')^T\|. \tag{C.2}$$

Define the principle angles θ_p between $\text{span}(E)$ and $\text{span}(E')$ to be $\theta_p = (\theta_{p,1}, \ldots, \theta_{p,r})$, so that E^TE' has singular value decomposition $E^TE' = P\cos(\theta_p)Q^T$, where $P, Q \in \mathbb{R}^{r\times r}$ are orthonormal matrices and $\cos(\theta_p) = \text{diag}(\cos(\theta_{p,1}), \ldots, \cos(\theta_{p,r}))$ (similar for $\sin(\theta_p)$). We further let $E^\perp \in \mathbb{R}^{(d+m)\times(d+m-r)}$ be the complement of E, and express E' as

$$E' = EA + E^\perp B. \tag{C.3}$$

Then, $E^TE' = A = P\cos(\theta_p)Q^T$ and $I = (E')^TE' = A^TA + B^TB$. By the above formulation,

$$\begin{aligned}
\|EE^T - E'(E')^T\| &\overset{\text{(C.3)}}{=} \left\|(E, E^\perp)\begin{pmatrix}I - AA^T & -AB^T\\ -BA^T & -BB^T\end{pmatrix}\begin{pmatrix}E^T\\ (E^\perp)^T\end{pmatrix}\right\| = \left\|\begin{pmatrix}I - AA^T & -AB^T\\ -BA^T & -BB^T\end{pmatrix}\right\| \\
&\leq \left\|\begin{pmatrix}I - AA^T & \mathbf{0}\\ \mathbf{0} & -BB^T\end{pmatrix}\right\| + \left\|\begin{pmatrix}\mathbf{0} & AB^T\\ BA^T & \mathbf{0}\end{pmatrix}\right\| \\
&\leq \max\{\|I - AA^T\|, \|BB^T\|\} + \|AB^T\| \\
&= \max\{\|I - AA^T\|, \|I - A^TA\|\} + \|AB^T\| \\
&= \|\sin(\theta_p)\|^2 + \sqrt{\|P\cos(\theta_p)\sin^2(\theta_p)\cos(\theta_p)P^T\|} \\
&= \|\sin(\theta_p)\|^2 + \|\sin(\theta_p)\cos(\theta_p)\| \leq 2\|\sin(\theta_p)\|.
\end{aligned} \tag{C.4}$$

On the other hand, by Wedin's $\sin(\Theta)$ theorem (Wedin, 1972, (3.1)), we know that

$$\|\sin(\theta_p)\| \leq \frac{\|(K^\star - K_t)S\|}{D_r}. \tag{C.5}$$

Moreover, we let F_r be the r-th column of F and note that

$$D_r^2 = F_r^TS^T(K^\star)^2SF_r \geq (\sigma_{\min}(K^\star))^2F_r^TS^TSF_r.$$

Since $\text{kernel}(K^\star S) = \text{kernel}(S)$ and $F_r \in \text{kernel}^\perp(K^\star S)$, we know $F_r \in \text{kernel}^\perp(S) = \text{span}(S^T)$. Thus, $F_r^T S^T S F_r \geq \lambda_{\min}^+(S^T S)$, where $\lambda_{\min}^+(S^T S) = (\sigma_{\min}^+(S))^2$ is the least positive eigenvalue of $S^T S$. Therefore, we have

$$D_r \geq \sigma_{\min}(K^\star)\sigma_{\min}^+(S). \tag{C.6}$$

Combining all above derivations, we have

$$\begin{aligned}\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star\| &\overset{\text{(C.2)}}{=} \|EE^T - E'(E')^T\| \overset{\text{(C.4)}}{\leq} 2\|\sin(\theta_p)\| \\ &\overset{\text{(C.5)}}{\leq} \frac{2\|K_t - K^\star\| \cdot \|S\|}{D_r} \overset{\text{(C.6)}}{\leq} \frac{2\|K_t - K^\star\|}{\sigma_{\min}(K^\star)} \cdot \frac{\|S\|}{\sigma_{\min}^+(S)}.\end{aligned}$$

This completes the proof.

C.3 Proof of Corollary 4.4

Noting that $\|K_t S(S^T K_t^2 S)^\dagger S^T K_t\| \leq 1$, we apply dominated convergence theorem (Durrett, 2019, Theorem 1.6.7) and Lemma 4.2, and almost surely have

$$\lim_{t\to\infty} \mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] = \mathbb{E}[K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star],$$

where the expectation is taken over randomness of S only. This shows $C_t \to C^\star$. Further, denoting $A_t = I - \mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]$ and $A^\star = I - \mathbb{E}[K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star]$, we have

$$\begin{aligned}\|C_t - C^\star\| &= \|A_t^\tau - (A^\star)^\tau\| \leq \left\|A_t^{\tau-1}(A_t - A^\star)\right\| + \left\|(A_t^{\tau-1} - (A^\star)^{\tau-1})A^\star\right\| \\ &\leq \|A_t - A^\star\| + \left\|A_t^{\tau-1} - (A^\star)^{\tau-1}\right\| \quad (\|A_t\| \vee \|A^\star\| \leq 1) \\ &\leq \tau\|A_t - A^\star\| \leq \tau\mathbb{E}\left[\left\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star\right\| \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t\right] \\ &\leq \frac{2\tau\|K_t - K^\star\|}{\sigma_{\min}(K^\star)}\mathbb{E}\left[\frac{\|S\|}{\sigma_{\min}^+(S)}\right] \leq \frac{2\tau\Upsilon_S}{\sigma_{\min}(K^\star)}\|K_t - K^\star\| \quad \text{(by Assumption 4.3)}.\end{aligned}$$

This completes the proof.

C.4 Proof of Lemma 4.5

We note that

$$\mathcal{I}_{1,t} \overset{\text{(20a)}}{=} \sum_{i=0}^t \prod_{j=i+1}^t \{I - \varphi_j(I + C^\star)\}\varphi_i \boldsymbol{\theta}^i \overset{\text{(22)}}{=} U\sum_{i=0}^t \prod_{j=i+1}^t \{I - \varphi_j\Sigma\}\varphi_i U^T \boldsymbol{\theta}^i.$$

Since $\mathbb{E}[\boldsymbol{\theta}^i \mid \mathcal{F}_{i-1}] = \boldsymbol{0}$, $\mathcal{I}_{1,t}$ is a martingale. We hence aim to apply the strong law of large number (Duflo, 1997, Theorem 1.3.15), the central limit theorem (Duflo, 1997, Corollary 2.1.10), and the Berry-Esseen bound (Fan, 2019, Theorem 2.1) for martingales to show each result. We first characterize the conditional covariance of $\mathcal{I}_{1,t}$ defined as (cf. (Duflo, 1997, Proposition 1.3.7))

$$\langle \mathcal{I}_1 \rangle_t \coloneqq U\sum_{i=0}^t \prod_{j=i+1}^t \{I - \varphi_j\Sigma\}\varphi_i^2 U^T \mathbb{E}[\boldsymbol{\theta}^i(\boldsymbol{\theta}^i)^T \mid \mathcal{F}_{i-1}]U\Big(\prod_{j=i+1}^t \{I - \varphi_j\Sigma\}\Big)^T U^T. \tag{C.7}$$

For the term $\mathbb{E}[\boldsymbol{\theta}^i(\boldsymbol{\theta}^i)^T \mid \mathcal{F}_{i-1}]$, we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}^i(\boldsymbol{\theta}^i)^T \mid \mathcal{F}_{i-1}] &\overset{(21b)}{=} \mathbb{E}\Bigg[\left\{(I+C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i\\ \boldsymbol{0}\end{pmatrix} - \left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix}\right\}\right\}\\
&\quad \left\{(I+C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i\\ \boldsymbol{0}\end{pmatrix} - \left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix}\right\}\right\}^T \mid \mathcal{F}_{i-1}\Bigg]\\
&\overset{(C.1)}{=} (I+C_i)K_i^{-1}\mathbb{E}\left[\begin{pmatrix}\bar{g}_i - \nabla f_i\\ \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\bar{g}_i - \nabla f_i\\ \boldsymbol{0}\end{pmatrix}^T \mid \mathcal{F}_{i-1}\right]K_i^{-1}(I+C_i)\\
&\quad + \mathbb{E}\left[\left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix}\right\}\left\{\begin{pmatrix}\bar{\Delta}\boldsymbol{x}_i\\ \bar{\Delta}\boldsymbol{\lambda}_i\end{pmatrix} - (I+C_i)\begin{pmatrix}\widetilde{\Delta}\boldsymbol{x}_i\\ \widetilde{\Delta}\boldsymbol{\lambda}_i\end{pmatrix}\right\}^T \mid \mathcal{F}_{i-1}\right] =: \mathcal{J}_{1,i} + \mathcal{J}_{2,i}. \quad \text{(C.8)}
\end{aligned}$$

For the term $\mathcal{J}_{1,i}$, we apply Assumption 3.2 and have

$$\mathbb{E}[(\bar{g}_i - \nabla f_i)(\bar{g}_i - \nabla f_i)^T \mid \mathcal{F}_{i-1}] = \mathbb{E}[\bar{g}_i\bar{g}_i^T \mid \mathcal{F}_{i-1}] - \nabla f_i \nabla^T f_i.$$

We also have

$$\begin{aligned}
&\left\|\mathbb{E}[\bar{g}_i\bar{g}_i^T - \nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi) \mid \mathcal{F}_{i-1}]\right\|\\
&\leq 2\mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\| \cdot \|\bar{g}_i\| \mid \mathcal{F}_{i-1}\right] + \mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\|^2 \mid \mathcal{F}_{i-1}\right]\\
&\leq 2\sqrt{\mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\|^2 \mid \mathcal{F}_{i-1}\right]}\sqrt{\mathbb{E}\left[\|\bar{g}_i\|^2 \mid \mathcal{F}_{i-1}\right]} + \mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\|^2 \mid \mathcal{F}_{i-1}\right],
\end{aligned}$$

and

$$\mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\|^2 \mid \mathcal{F}_{i-1}\right] \leq \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{X}}\|\nabla^2 f(\boldsymbol{x};\xi)\|^2] \cdot \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2 \overset{(15e)}{\leq} \Upsilon_m\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2,$$

and

$$\mathbb{E}[\|\bar{g}_i\|^2 \mid \mathcal{F}_{i-1}] = \|\nabla f_i\|^2 + \mathbb{E}[\|\bar{g}_i - \nabla f_i\|^2 \mid \mathcal{F}_{i-1}] \overset{(13),(15a)}{\leq} \Upsilon_u^2 + \Upsilon_m \leq 2(\Upsilon_u^2 \vee \Upsilon_m).$$

Thus, we combine the above three displays, and have

$$\left\|\mathbb{E}[\bar{g}_i\bar{g}_i^T - \nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi) \mid \mathcal{F}_{i-1}]\right\| \leq 2\sqrt{2}(\Upsilon_u^2 \vee \Upsilon_m)(\|\boldsymbol{x}_i - \boldsymbol{x}^\star\| + \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2) \rightarrow 0. \quad \text{(C.9)}$$

Thus, we obtain

$$\lim_{i\rightarrow\infty}\mathbb{E}[(\bar{g}_i - \nabla f_i)(\bar{g}_i - \nabla f_i)^T \mid \mathcal{F}_{i-1}] = \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star). \quad \text{(C.10)}$$

Further, by Theorem 3.12, we know $K_i \rightarrow K^\star$ as $i \rightarrow \infty$; and Corollary 4.4 shows $C_i \rightarrow C^\star$. Thus, we use the definition (25) and have

$$\mathcal{J}_{1,i} = (I + C^\star)\Omega^\star(I + C^\star) + O(\mathcal{K}_{1,i}) \quad \text{(C.11)}$$

with $\mathcal{K}_{1,i} \rightarrow 0$ as $i \rightarrow \infty$ almost surely.

For the term $\mathcal{J}_{2,i}$, we have (recall from Section 2 that $\boldsymbol{z}_{i,\tau} = (\bar{\Delta}\boldsymbol{x}_i, \bar{\Delta}\boldsymbol{\lambda}_i)$ and $\widetilde{\boldsymbol{z}}_i = (\widetilde{\Delta}\boldsymbol{x}_i, \widetilde{\Delta}\boldsymbol{\lambda}_i)$)

$$
\begin{aligned}
\mathcal{J}_{2,i} &= \mathbb{E}\left[(\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i)(\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i)^T \mid \mathcal{F}_{i-1}\right] \stackrel{(7)}{=} \mathbb{E}[(\widetilde{C}_i - C_i)\widetilde{\boldsymbol{z}}_i\widetilde{\boldsymbol{z}}_i^T(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&\stackrel{(3)}{=} \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}\begin{pmatrix}\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&\quad + \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i\nabla^T\mathcal{L}_iK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&\quad + \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&\quad + \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right].
\end{aligned} \tag{C.12}
$$

For the last two terms, we apply the tower property of conditional expectation by first conditioning on the randomness of ζ_i to take expectation over the randomness of ξ_i, and then taking expectation over the randomness of ζ_i. This is allowed even ζ_i is generated after ξ_i in the algorithm, because ξ_i and ζ_i are independent. In particular, we have (similar for the second last term in (C.12))

$$
\begin{aligned}
&\mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}\begin{pmatrix}\mathbb{E}[\bar{g}_i - \nabla f_i \mid \mathcal{F}_{i-1} \cup \zeta_i] \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}_i \\ c_i\end{pmatrix}\begin{pmatrix}\mathbb{E}[\bar{g}_i - \nabla f_i \mid \mathcal{F}_{i-1}] \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] = \boldsymbol{0}.
\end{aligned}
$$

For the second term in (C.12), it converges to zero almost surely as i goes to infinity since $\|\widetilde{C}_i\| \vee \|C_i\| \leq 1$, $\|K_i^{-1}\| \leq \Upsilon_K$, and $\nabla\mathcal{L}_i \rightarrow 0$. For the first term in (C.12), we have

$$
\begin{aligned}
&\mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \boldsymbol{0}\end{pmatrix}^T K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\mathbb{E}\left[(\bar{g}_i - \nabla f_i)(\bar{g}_i - \nabla f_i)^T \mid \mathcal{F}_{i-1} \cup \zeta_i\right] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[(\widetilde{C}_i - C_i)K_i^{-1}\begin{pmatrix}\mathbb{E}\left[(\bar{g}_i - \nabla f_i)(\bar{g}_i - \nabla f_i)^T \mid \mathcal{F}_{i-1}\right] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}\right] \\
&\xrightarrow{(24)} \mathbb{E}[(\widetilde{C}^\star - C^\star)\Omega^\star((\widetilde{C}^\star)^T - C^\star)] = \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] - C^\star\Omega^\star C^\star.
\end{aligned}
$$

Again, the convergence here is due to the dominated convergence theorem, (C.10), and $K_i \rightarrow K^\star$; and the expectation is taken over the randomness of τ sketch matrices $S_1, \ldots, S_\tau$ only. Thus, combining the above two displays with (C.12), we have

$$
\mathcal{J}_{2,i} = \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] - C^\star\Omega^\star C^\star + O(\mathcal{K}_{2,i}) \tag{C.13}
$$

with $\mathcal{K}_{2,i} \to 0$ as $i \to \infty$ almost surely.

Combining (C.13), (C.11), and (C.8), we obtain

$$\mathbb{E}[\boldsymbol{\theta}^i(\boldsymbol{\theta}^i) \mid \mathcal{F}_{i-1}] = \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T] + O(\mathcal{K}_{1,i} + \mathcal{K}_{2,i}).$$

By the definition of $\langle \mathcal{I}_1 \rangle_t$ in (C.7), let us denote

$$\Gamma = U^T \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]U,$$

and for any $k, l \in \{1, \ldots, d+m\}$, the (k, l) entry of the matrix $U^T\langle \mathcal{I}_1 \rangle_t U$ can be written as

$$[U^T\langle \mathcal{I}_1 \rangle_t U]_{k,l} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} (1 - \varphi_j\sigma_k)(1 - \varphi_j\sigma_l)\varphi_i^2(\Gamma_{kl} + r_{i,kl}),$$

where $r_{i,kl} \to 0$ as $i \to \infty$ almost surely. Noting from (23) that $\sigma_k + \sigma_l \geq 2(1 - \rho^\tau)$, and using the condition (26), Lemmas A.2, A.3 lead to $[U^T\langle \mathcal{I}_1 \rangle_t U]_{k,l}/\varphi_t \to \Gamma_{kl}/(\sigma_k + \sigma_l + \varphi/\widetilde{\varphi})$ as $t \to \infty$ almost surely. Thus, we have

$$\frac{1}{\varphi_t} \cdot \langle \mathcal{I}_1 \rangle_t \xrightarrow{a.s.} U(\Theta \circ \Gamma)U^T \overset{(28)}{=} \Xi^\star. \tag{C.14}$$

Thus, (Duflo, 1997, Theorem 1.3.15) shows that (27) holds. This shows the first part of the results.

For the second part of the results, we assume a higher order moment condition (15b). We have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\theta}^i\|^3 \mid \mathcal{F}_{i-1}] \overset{(21b)}{\leq} 4\left(\mathbb{E}\left[\left\|(I + C_i)K_i^{-1}\begin{pmatrix}\bar{g}_i - \nabla f_i \\ \mathbf{0}\end{pmatrix}\right\|^3 \mid \mathcal{F}_{i-1}\right] + \mathbb{E}[\|\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \\
&\overset{(7)}{\leq} 4\left(8\Upsilon_K^3\mathbb{E}[\|\bar{g}_i - \nabla f_i\|^3 \mid \mathcal{F}_{i-1}] + \mathbb{E}[\|(\widetilde{C}_i - C_i)\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\text{also use } \|C_i\| \leq 1, \|K_i^{-1}\| \leq \Upsilon_K) \\
&\overset{(15b)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\mathbb{E}[\|\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\text{also use } \|\widetilde{C}_i\| \vee \|C_i\| \leq 1) \\
&\overset{(3)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\mathbb{E}[\|\bar{\nabla}\mathcal{L}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\text{also use } \|K_i^{-1}\| \leq \Upsilon_K) \\
&\overset{(3)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\left\{4\|\nabla\mathcal{L}_i\|^3 + 4\mathbb{E}[\|\bar{g}_i - \nabla f_i\|^3 \mid \mathcal{F}_{i-1}]\right\}\right) \\
&\overset{(15b)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\left\{4\Upsilon_u^3 + 4\Upsilon_m\right\}\right) \quad (\text{also use } (13)).
\end{aligned} \tag{C.15}$$

Thus, $\boldsymbol{\theta}^i$ has bounded third moment; and (Wang, 1995, pp. 554) together with (C.14) give the result (a). For (b), we verify Lindeberg's condition. For any $\epsilon > 0$, we have

$$\begin{aligned}
&\frac{1}{\varphi_t}\sum_{i=0}^{t}\mathbb{E}[\|\prod_{j=i+1}^{t}\{I - \varphi_j(I + C^\star)\}\varphi_i\boldsymbol{\theta}^i\|^2 \cdot \mathbf{1}_{\|\prod_{j=i+1}^{t}\{I - \varphi_j(I+C^\star)\}\varphi_i\boldsymbol{\theta}^i\| \geq \epsilon\sqrt{\varphi_t}} \mid \mathcal{F}_{i-1}] \\
&\leq \frac{1}{\epsilon\varphi_t^{3/2}}\sum_{i=0}^{t}\mathbb{E}[\|\prod_{j=i+1}^{t}\{I - \varphi_j(I + C^\star)\}\varphi_i\boldsymbol{\theta}^i\|^3 \mid \mathcal{F}_{i-1}] \\
&\overset{(22)}{=} \frac{1}{\epsilon\varphi_t^{3/2}}\sum_{i=0}^{t}\mathbb{E}[\|\prod_{j=i+1}^{t}\{I - \varphi_j\Sigma\}\varphi_i U^T\boldsymbol{\theta}^i\|^3 \mid \mathcal{F}_{i-1}].
\end{aligned}$$

To show the right hand side converges to zero, it suffices to show that each entry of the vector on the right hand side converges to zero. In particular, it suffices to show that for any $1 \le k \le d+m$,

$$\frac{1}{\epsilon\varphi_t^{3/2}}\sum_{i=0}^{t}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|^3\,\varphi_i^3\mathbb{E}[|[U^T\boldsymbol{\theta}^i]_k|^3\mid\mathcal{F}_{i-1}]\longrightarrow 0 \qquad \text{as} \qquad t\to\infty.$$

By (C.15) and the fact $\mathbb{E}[|[U^T\boldsymbol{\theta}^i]_k|^3\mid\mathcal{F}_{i-1}]\le\mathbb{E}[\|\boldsymbol{\theta}^i\|^3\mid\mathcal{F}_{i-1}]$, we only need to show $\sum_{i=0}^{t}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|^3\varphi_i^3 = o(\varphi_t^{3/2})$. Without loss of generality, we suppose $1-\varphi_j\sigma_k\ge 0$ for all $j\ge 1$, and show

$$\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3\varphi_i^3 = o(\varphi_t^{3/2}). \tag{C.16}$$

Otherwise, since $\varphi<0$ from (26), Lemma A.2 shows that $\varphi_i\to 0$. Thus, there exists $\widetilde{t}$ such that $1-\varphi_j\sigma_k\ge 0$, $\forall j\ge\widetilde{t}$. Therefore,

$$\begin{aligned}
\sum_{i=0}^{t}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|^3\,\varphi_i^3 &= \sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|^3\,\varphi_i^3+\sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3\varphi_i^3\\
&= \prod_{j=\widetilde{t}}^{t}(1-\varphi_j\sigma_k)^3\sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{\widetilde{t}-1}|1-\varphi_j\sigma_k|^3\,\varphi_i^3+\sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3\varphi_i^3\\
&= \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3(\varphi_i')^3,
\end{aligned}\tag{C.17}$$

where

$$\varphi_{\widetilde{t}-1}' = \left(\sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{\widetilde{t}-1}|1-\varphi_j\sigma_k|^3\,\varphi_i^3+\varphi_{\widetilde{t}-1}^3\right)^{1/3}, \qquad \text{and} \qquad \varphi_i'=\varphi_i, \quad \forall i\ge\widetilde{t}.$$

Note that (C.17) has the same form as (C.16), and φ_i' differs from φ_i only at $i=\widetilde{t}-1$. Thus, (C.17) and (C.16) have the same limit. For (C.16), we apply Lemma A.1, and note that

$$\lim_{i\to\infty} i\left(1-\frac{\varphi_{i-1}^2}{\varphi_i^2}\right)\overset{(26)}{=}2\varphi \qquad \text{and} \qquad 3\sigma_k+2\varphi/\widetilde{\varphi}\overset{(26)}{>}0.$$

Thus, Lemma A.3 suggests that

$$\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3\varphi_i^3 = O(\varphi_t^2).$$

This verifies (C.16) and further verifies Lindeberg's condition. Thus, the central limit theorem of martingale in (Duflo, 1997, Corollary 2.1.10) leads to (b). For (c), we apply (Fan, 2019, Theorem 2.1) with $\epsilon=\sqrt{\varphi_t}$, $\delta=0$, $\rho=1$ (in their notation), as proved for verifying Lindeberg's condition above, and obtain the normalized result immediately. This completes the proof.

C.5 Proof of Lemma 4.6

We note that

$$\mathcal{I}_{2,t} \overset{(20b)}{=} \sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j(I+C^\star)\}(\bar{\alpha}_i-\varphi_i)\boldsymbol{z}_{i,\tau} \overset{(22)}{=} U\sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j\Sigma\}(\bar{\alpha}_i-\varphi_i)U^T\boldsymbol{z}_{i,\tau}.$$

Thus, for any $1\leq k\leq d+m$, we have

$$[U^T\mathcal{I}_{2,t}]_k = \sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)(\bar{\alpha}_i-\varphi_i)[U^T\boldsymbol{z}_{i,\tau}]_k.$$

For the same reason as (C.16) and (C.17), we suppose for any $j\geq 0$ that $1-\varphi_j\sigma_k\geq 0$. Then, we know

$$\begin{aligned}
\left|[U^T\mathcal{I}_{2,t}]_k\right| &\leq \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|\chi_i\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| = \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| \\
&= \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|\mid\mathcal{F}_{i-1}\right] \\
&\quad + \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\left\{\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| - \mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|\mid\mathcal{F}_{i-1}\right]\right\} =: \mathcal{J}_{3,t,k}+\mathcal{J}_{4,t,k}. \quad \text{(C.18)}
\end{aligned}$$

Intuitively, $\mathcal{J}_{3,t,k}$ dominates $\mathcal{J}_{4,t,k}$ since the latter measures the error to the mean. We precisely show such result in the following. We first show that $\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|$ has bounded variance. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\{\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| - \mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|\mid\mathcal{F}_{i-1}\right]\right\}^2\mid\mathcal{F}_{i-1}\right] \\
&\qquad \leq \mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|^2\mid\mathcal{F}_{i-1}\right] \leq \mathbb{E}\left[\|\boldsymbol{z}_{i,\tau}\|^2\mid\mathcal{F}_{i-1}\right] \overset{(B.7)}{\leq} 16\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m). \quad \text{(C.19)}
\end{aligned}$$

Thus, $\mathcal{J}_{4,t,k}$ is a square integrable martingale. Its variance is bounded by

$$\begin{aligned}
\langle\mathcal{J}_{4,k}\rangle_t &\coloneqq \frac{1}{4}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^2\chi_i^2\mathbb{E}\left[\left\{\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| - \mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|\mid\mathcal{F}_{i-1}\right]\right\}^2\mid\mathcal{F}_{i-1}\right] \\
&\overset{(C.19)}{\leq} 4\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m)\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^2\chi_i^2.
\end{aligned}$$

Using (26) and (29), we know

$$\begin{aligned}
\lim_{i\to\infty} i\left(1-\frac{\chi_{i-1}^2/\varphi_{i-1}}{\chi_i^2/\varphi_i}\right) &= \lim_{i\to\infty} i\left(1-\frac{\chi_{i-1}^2}{\chi_i^2}+\frac{\chi_{i-1}^2}{\chi_i^2}\left(1-\frac{\varphi_i}{\varphi_{i-1}}\right)\right) \\
&= \lim_{i\to\infty} i\left\{\left(1-\frac{\chi_{i-1}}{\chi_i}\right)\left(1+\frac{\chi_{i-1}}{\chi_i}\right)-\frac{\chi_{i-1}^2}{\chi_i^2}\frac{\varphi_i}{\varphi_{i-1}}\left(1-\frac{\varphi_{i-1}}{\varphi_i}\right)\right\} = 2\chi-\varphi. \quad \text{(C.20)}
\end{aligned}$$

Further, (29) and (23) imply $2\sigma_k + (2\chi - \varphi)/\widetilde{\varphi} > 0$. Thus, Lemma A.3 leads to $\langle \mathcal{J}_{4,k} \rangle_t = O(\chi_t^2/\varphi_t)$; and the strong law of large number (Duflo, 1997, Theorem 1.3.15) suggests that

$$\mathcal{J}_{4,t,k} = o\left(\sqrt{\chi_t^2/\varphi_t \cdot \{\log(\varphi_t/\chi_t^2)\}^{1+\upsilon'}}\right) = o\left(\sqrt{\chi_t^2/\varphi_t \cdot \{\log(1/\chi_t)\}^{1+\upsilon'}}\right) \overset{(29)}{=} o(\chi_t/\varphi_t). \quad \text{(C.21)}$$

For the term $\mathcal{J}_{3,t,k}$, we have

$$\mathcal{J}_{3,t,k} \le \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\sqrt{\mathbb{E}[|[U^T\boldsymbol{z}_{i,\tau}]_k|^2 \mid \mathcal{F}_{i-1}]} \overset{\text{(C.19)}}{\le} 2\Upsilon_K(\Upsilon_u \vee \sqrt{\Upsilon_m})\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i.$$

Using (26) and (29), and the fact that

$$\lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}/\varphi_{i-1}}{\chi_i/\varphi_i}\right) \overset{\text{(C.20)}}{=} \chi - \varphi,$$

and $\sigma_k + (\chi - \varphi)/\widetilde{\varphi} > 0$ (as implied by (29)), we apply Lemma A.3 and obtain

$$\mathcal{J}_{3,t,k} = O(\chi_t/\varphi_t). \quad \text{(C.22)}$$

Plugging (C.22), (C.21) into (C.18), we complete the proof.

C.6 Proof of Lemma 4.7

Based on the definition of $\mathcal{I}_{3,t}$ in (20c), we have the recursion

$$\mathcal{I}_{3,t+1} = \{I - \varphi_{t+1}(I + C^\star)\}\,\mathcal{I}_{3,t} + \varphi_{t+1}\boldsymbol{\delta}^{t+1}. \quad \text{(C.23)}$$

By Assumption 3.1, we have

$$\begin{aligned}\|\boldsymbol{\delta}^t\| &\overset{(21c)}{\le} 2\left(\|(K^\star)^{-1}\|\|\boldsymbol{\psi}^t\| + \|K_t^{-1} - (K^\star)^{-1}\|\cdot\|\nabla\mathcal{L}_t\|\right) + \|C_t - C^\star\|\cdot\left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| \quad (\|C_t\|\le 1)\\ &\le 2\Upsilon_K\Upsilon_L\left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|^2 + 2\Upsilon_K^2\Upsilon_u\|K_t - K^\star\|\cdot\left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \|C_t - C^\star\|\cdot\left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| \quad \text{(C.24)}\end{aligned}$$

Using the fact that $K_t \to K^\star$ (cf. Theorem 3.12) and $C_t \to C^\star$ (cf. Corollary 4.4), we know

$$\boldsymbol{\delta}^t = o(\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|). \quad \text{(C.25)}$$

Furthermore, we use $\|C^\star\| \le \rho^\tau$ and know that for any $a \in (0,1)$, there exists a threshold t_1 such that for any $t \ge t_1$,

$$\begin{aligned}\|\mathcal{I}_{3,t+1}\| &\le \{1 - \varphi_{t+1}(1-\rho^\tau)\}\,\|\mathcal{I}_{3,t}\| + \varphi_{t+1}\cdot o\left(\left\|\begin{pmatrix}\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|\right)\\ &\le \{1 - \varphi_{t+1}(1-\rho^\tau) + o(\varphi_{t+1})\}\,\|\mathcal{I}_{3,t}\| + \varphi_{t+1}\cdot o(\|\mathcal{I}_{1,t}\| + \|\mathcal{I}_{2,t}\|) \quad \text{(Lemma 4.1)}\\ &\le \{1 - a(1-\rho^\tau)\varphi_{t+1}\}\,\|\mathcal{I}_{3,t}\| + \varphi_{t+1}\cdot o(\|\mathcal{I}_{1,t}\| + \|\mathcal{I}_{2,t}\|).\end{aligned}$$

We apply the above inequality recursively, and obtain

$$\|\mathcal{I}_{3,t+1}\| \leq \prod_{j=t_1+1}^{t+1} \{1 - a(1-\rho^\tau)\varphi_j\} \|\mathcal{I}_{3,t_1}\| + \sum_{i=t_1+1}^{t+1} \prod_{j=i+1}^{t+1} \{1 - a(1-\rho^\tau)\varphi_j\} \varphi_i o(\|\mathcal{I}_{1,i-1}\| + \|\mathcal{I}_{2,i-1}\|). \quad \text{(C.26)}$$

We apply Lemmas 4.5 and 4.6 for bounding $\|\mathcal{I}_{1,i-1}\|$ and $\|\mathcal{I}_{2,i-1}\|$. In particular, we note that for any $\upsilon \geq 0$,

$$\begin{aligned}
&\lim_{i\to\infty} i \left(1 - \frac{\sqrt{\varphi_{i-2} \{\log(1/\varphi_{i-2})\}^{1+\upsilon}}}{\sqrt{\varphi_{i-1} \{\log(1/\varphi_{i-1})\}^{1+\upsilon}}}\right) = \lim_{i\to\infty} i \left(1 - \frac{\sqrt{\varphi_{i-2}}}{\sqrt{\varphi_{i-1}}} + \frac{\sqrt{\varphi_{i-2}}}{\sqrt{\varphi_{i-1}}} \left(1 - \frac{\{\log(1/\varphi_{i-2})\}^{\frac{1+\upsilon}{2}}}{\{\log(1/\varphi_{i-1})\}^{\frac{1+\upsilon}{2}}}\right)\right) \\
&\overset{(26)}{=} \lim_{i\to\infty} i \left(1 - \frac{\sqrt{\varphi_{i-2}}}{\sqrt{\varphi_{i-1}}}\right) + \lim_{i\to\infty} i \left(1 - \frac{\{\log(1/\varphi_{i-2})\}^{\frac{1+\upsilon}{2}}}{\{\log(1/\varphi_{i-1})\}^{\frac{1+\upsilon}{2}}}\right) \\
&\overset{(26)}{=} \frac{\varphi}{2} + \lim_{i\to\infty} i \left(1 - \frac{\{\log(1/\varphi_{i-2})\}^{\frac{1+\upsilon}{2}}}{\{\log(1/\varphi_{i-1})\}^{\frac{1+\upsilon}{2}}}\right). \quad \text{(Lemma A.1)}
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\lim_{i\to\infty} i \left(1 - \frac{\log(1/\varphi_{i-2})}{\log(1/\varphi_{i-1})}\right) &= \lim_{i\to\infty} \frac{i \log(\varphi_{i-2}/\varphi_{i-1})}{\log(1/\varphi_{i-1})} = \lim_{i\to\infty} \frac{i \log\left(1 + (\varphi_{i-2} - \varphi_{i-1})/\varphi_{i-1}\right)}{\log(1/\varphi_{i-1})} \\
&= \lim_{i\to\infty} \frac{i \left\{\frac{\varphi_{i-2}-\varphi_{i-1}}{\varphi_{i-1}} + O\left(\frac{(\varphi_{i-2}-\varphi_{i-1})^2}{\varphi_{i-1}^2}\right)\right\}}{\log(1/\varphi_{i-1})} = \lim_{i\to\infty} \frac{-\varphi}{\log(1/\varphi_{i-1})} = 0,
\end{aligned}$$

where the last equality uses the fact that $\varphi_i \to 0$, as implied by Lemma A.2. Combining the above two displays with Lemma A.1, we have

$$\lim_{i\to\infty} i \left(1 - \frac{\sqrt{\varphi_{i-2} \{\log(1/\varphi_{i-2})\}^{1+\upsilon}}}{\sqrt{\varphi_{i-1} \{\log(1/\varphi_{i-1})\}^{1+\upsilon}}}\right) = \frac{\varphi}{2}. \quad \text{(C.27)}$$

Moreover, we have

$$\lim_{i\to\infty} i \left(1 - \frac{\chi_{i-2}/\varphi_{i-2}}{\chi_{i-1}/\varphi_{i-1}}\right) \overset{\text{(C.20)}}{=} \chi - \varphi. \quad \text{(C.28)}$$

Letting a be any scalar such that

$$0 < \frac{-\varphi/\widetilde{\varphi}}{2(1-\rho^\tau)} \vee \frac{-(\chi - \varphi)/\widetilde{\varphi}}{1-\rho^\tau} < a < 1,$$

which is guaranteed to exist due to (26) and (29), we can see that

$$a(1-\rho^\tau) + \frac{\varphi}{2\widetilde{\varphi}} > 0 \qquad \text{and} \qquad a(1-\rho^\tau) + \frac{\chi - \varphi}{\widetilde{\varphi}} > 0.$$

Thus, combining (C.26), (C.27), and (C.28) with Lemma A.3, we obtain the results for both (15a) and (15b). This completes the proof.

C.7 Proof of Lemma 4.9

Combining (B.11) and (B.12), for any $\upsilon > 0$, we have

$$\begin{aligned}
\|K_t - K^\star\| &\leq \frac{\Upsilon_L}{t}\sum_{i=0}^{t-1}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t \\
&= \frac{\Upsilon_L}{t}\left\|\begin{pmatrix}\boldsymbol{x}_0 - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\frac{1}{j}\right)\frac{1}{i}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t \\
&= \Upsilon_L\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\frac{1}{j}\right)\frac{1}{i}\left\{a_i + O\left(\frac{\chi_i}{\varphi_i}\right)\right\} + \Upsilon_L\left(a_t + O\left(\frac{\chi_t}{\varphi_t}\right)\right) + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t, \qquad \text{(C.29)}
\end{aligned}$$

where, by Theorem 4.8, $a_i = o(\sqrt{\varphi_i}\{\log(1/\varphi_i)\}^{(1+\upsilon)/2})$ under (15a) and $a_i = O(\sqrt{\varphi_i\log(1/\varphi_i)})$ under (15b). We claim that $\varphi > -2$. Otherwise, $\varphi + 1.5 \leq -0.5 < 0$. We apply Lemma A.1 and have

$$\lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}(t-1)^{1.5}}{\varphi_t t^{1.5}}\right) = \lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}}{\varphi_t} + \frac{\varphi_{t-1}}{\varphi_t}\left(1 - \frac{(t-1)^{1.5}}{t^{1.5}}\right)\right) \overset{(26)}{=} \varphi + 1.5 < 0.$$

Then, Lemma A.2 suggests that $\varphi_t t^{1.5} \to 0$, which cannot hold under (26). Thus, $\varphi > -2$. Using (C.27) and Lemma A.3, and noting that $1 + \varphi/2 > 0$, we obtain

$$\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\frac{1}{j}\right)\frac{a_i}{i} = O(a_t). \qquad \text{(C.30)}$$

Furthermore, we deal with the term that involves χ_i/φ_i in (C.29). Without loss of generality, we suppose $\chi \geq 3\varphi/2$. Otherwise, we know

$$\lim_{t\to\infty} t\left(1 - \frac{\chi_{t-1}/\varphi_{t-1}^{3/2}}{\chi_t/\varphi_t^{3/2}}\right) \overset{(C.20)}{=} \chi - \frac{3\varphi}{2} < 0.$$

This implies that $\chi_t/\varphi_t = o(\sqrt{\varphi_t}) = o(a_t)$ and, thus, all terms $O(\chi_i/\varphi_i)$ in (C.29) are negligible and the argument of the lemma holds immediately. Supposing $\chi \geq 3\varphi/2$ and noting that

$$\lim_{t\to\infty} t\left(1 - \frac{\chi_{t-1}/\varphi_{t-1}}{\chi_t/\varphi_t}\right) \overset{(C.20)}{=} \chi - \varphi \qquad \text{and} \qquad 1 + \chi - \varphi \geq 1 + \frac{\varphi}{2} > 0,$$

we obtain from Lemma A.3 that

$$\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\frac{1}{j}\right)\frac{1}{i}\cdot O\left(\frac{\chi_i}{\varphi_i}\right) = O\left(\frac{\chi_t}{\varphi_t}\right). \qquad \text{(C.31)}$$

Combining (C.29), (C.30), (C.31) together, and noting that $o(\sqrt{(\log t)^{1+\upsilon}/t}) = o(a_t)$ under (15a) (as implied by the fact that $t\varphi_t \to \widetilde{\varphi} \in (0, \infty]$), we complete the proof.

C.8 Proof of Lemma 4.10

Combining (C.24), Theorem 4.8, Lemma 4.9, and Corollary 4.4, we have for any $\upsilon > 0$,

$$\begin{aligned}\|\boldsymbol{\delta}^t\| &= O\left(\varphi_t \log(1/\varphi_t)\right) + O\left(\frac{\chi_t^2}{\varphi_t^2}\right) + o\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + o\left(\frac{\chi_t}{\varphi_t}\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) \\ &\quad + O\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \omega_t\right) + O\left(\frac{\chi_t}{\varphi_t} \cdot \omega_t\right) \\ &\overset{(30)}{=} O\left(\varphi_t \log(1/\varphi_t)\right) + o\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + O\left(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \omega_t\right).\end{aligned}$$

We plug the above bound into the recursion (C.23) and apply Lemma A.3. In particular, we note that

$$\begin{aligned}&\lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}\log(1/\varphi_{t-1})}{\varphi_t \log(1/\varphi_t)}\right) \overset{(C.27)}{=} \varphi, \\ &\lim_{t\to\infty} t\left(1 - \frac{\sqrt{\varphi_{t-1}\log(1/\varphi_{t-1})}\sqrt{(\log(t-1))^{1+\upsilon}/(t-1)}}{\sqrt{\varphi_t \log(1/\varphi_t)}\sqrt{(\log t)^{1+\upsilon}/t}}\right) = \frac{\varphi}{2} - \frac{1}{2}, \\ &\lim_{t\to\infty} t\left(1 - \frac{\sqrt{\varphi_{t-1}\log(1/\varphi_{t-1})}\omega_{t-1}}{\sqrt{\varphi_t \log(1/\varphi_t)}\omega_t}\right) = \frac{\varphi}{2} + \omega.\end{aligned}$$

Thus, Lemma A.3 suggests that $\mathcal{I}_{3,t}$ has the same order as $\boldsymbol{\delta}^t$ provided

$$1 - \rho^\tau + \frac{\varphi}{\widetilde{\varphi}} > 0, \quad 1 - \rho^\tau + \frac{\varphi - 1}{2\widetilde{\varphi}} > 0, \quad 1 - \rho^\tau + \frac{\varphi/2 + \omega}{\widetilde{\varphi}} > 0.$$

The above conditions are implied by (26) and (30). If $\omega_t = 0$ for all sufficiently large t, the results hold trivially. This completes the proof.

C.9 Proof of Theorem 4.11

By Lemmas 4.6 and 4.10, we know that

$$\begin{aligned}&\sqrt{1/\varphi_t} \cdot \mathcal{I}_{2,t} = O(\chi_t/\varphi_t^{3/2}), \\ &\sqrt{1/\varphi_t} \cdot \mathcal{I}_{3,t} = O(\sqrt{\varphi_t}\log(1/\varphi_t)) + o\left(\frac{\sqrt{\log(1/\varphi_t)(\log t)^{1+\upsilon}}}{\sqrt{t}}\right) + O\left(\sqrt{\log(1/\varphi_t)} \cdot \omega_t\right).\end{aligned}$$

Using Lemma A.1 and the fact that

$$\begin{aligned}&\lim_{t\to\infty} t\left(1 - \frac{\sqrt{\varphi_{t-1}}\log(1/\varphi_{t-1})}{\sqrt{\varphi_t}\log(1/\varphi_t)}\right) = \frac{\varphi}{2} < 0, \\ &\lim_{t\to\infty} t\left(1 - \frac{\sqrt{\log(1/\varphi_{t-1})(\log(t-1))^{1+\upsilon}}/\sqrt{t-1}}{\sqrt{\log(1/\varphi_t)(\log t)^{1+\upsilon}}/\sqrt{t}}\right) = -\frac{1}{2} < 0, \\ &\lim_{t\to\infty} t\left(1 - \frac{\sqrt{\log(1/\varphi_{t-1})} \cdot \omega_{t-1}}{\sqrt{\log(1/\varphi_t)} \cdot \omega_t}\right) = \omega < 0,\end{aligned}$$

we know $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{2,t} = o(1)$ and $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{3,t} = o(1)$ almost surely. Thus, the Slutsky's theorem together with Lemma 4.5 lead to the asymptotic normality. Furthermore, Lemma A.5 with $C_t = 0$, $B_t = \mathcal{I}_{2,t} + \mathcal{I}_{3,t}$ leads to the Berry-Esseen bound. This completes the proof.

C.10 Proof of Lemma 4.12

We note that

$$\|\Xi^\star - \Xi_t\| \leq \left\| \Xi^\star - \frac{1}{2+\varphi/\widetilde{\varphi}} \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] \right\| \\ + \frac{1}{2+\varphi/\widetilde{\varphi}} \left\| \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] - \Omega^\star \right\| + \frac{1}{2+\varphi/\widetilde{\varphi}} \|\Omega^\star - \Omega_t\|. \quad \text{(C.32)}$$

For the first term in (C.32), we have

$$\begin{aligned} &\left\| \Xi^\star - \frac{1}{2+\varphi/\widetilde{\varphi}} \cdot \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] \right\| \\ &\overset{(28)}{=} \left\| \left(\Theta - \frac{1}{2+\varphi/\widetilde{\varphi}} \mathbf{1}\mathbf{1}^T \right) \circ U^T \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] U \right\| \\ &\leq \left\| \Theta - \frac{1}{2+\varphi/\widetilde{\varphi}} \mathbf{1}\mathbf{1}^T \right\| \cdot \|\mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T]\| \quad (\|A \circ B\| \leq \|A\| \cdot \|B\|) \\ &\leq 4 \left\| \Theta - \frac{1}{2+\varphi/\widetilde{\varphi}} \mathbf{1}\mathbf{1}^T \right\| \|\Omega^\star\|, \quad (\|\widetilde{C}^\star\| \leq 1) \end{aligned}$$

and for any $1 \leq k, l \leq d+m$,

$$\begin{aligned} \left| \Theta_{k,l} - \frac{1}{2+\varphi/\widetilde{\varphi}} \right| &= \left| \frac{1}{\sigma_k + \sigma_l + \varphi/\widetilde{\varphi}} - \frac{1}{2+\varphi/\widetilde{\varphi}} \right| \\ &= \frac{|2 - \sigma_k - \sigma_l|}{(\sigma_k + \sigma_l + \varphi/\widetilde{\varphi})(2+\varphi/\widetilde{\varphi})} \overset{(23)}{\leq} \frac{2\rho^\tau}{(2 - 2\rho^\tau + \varphi/\widetilde{\varphi})(2+\varphi/\widetilde{\varphi})} \\ &\overset{(26)}{\leq} \frac{2\rho^\tau}{(2 - 2(1+\varphi/\widetilde{\varphi}) + \varphi/\widetilde{\varphi})(2+\varphi/\widetilde{\varphi})} = \frac{2\rho^\tau}{-\varphi/\widetilde{\varphi}(2+\varphi/\widetilde{\varphi})}. \end{aligned}$$

Therefore, the above two displays lead to

$$\left\| \Xi^\star - \frac{1}{2+\varphi/\widetilde{\varphi}} \cdot \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] \right\| = O(\rho^\tau). \quad \text{(C.33)}$$

For the second term in (C.32), we have

$$\begin{aligned} \left\| \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T] - \Omega^\star \right\| &\leq \|C^\star\Omega^\star\| + \|\Omega^\star C^\star\| + \|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\| \quad (\text{since } \mathbb{E}[\widetilde{C}^\star] = C^\star) \\ &\leq O(\rho^\tau) + \|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\|. \quad (\|C^\star\| \leq \rho^\tau) \end{aligned}$$

Furthermore, since $\Omega^\star \preceq \Upsilon_K^2\Upsilon_m \cdot I$, we obtain

$$\begin{aligned}
\mathbf{0} &\preceq \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] \preceq \Upsilon_K^2\Upsilon_m\mathbb{E}[\widetilde{C}^\star(\widetilde{C}^\star)^T] \\
&= \Upsilon_K^2\Upsilon_m\mathbb{E}\left[\left\{\prod_{j=1}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}\left\{\prod_{j=1}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}^T\right] \\
&= \Upsilon_K^2\Upsilon_m\mathbb{E}\Bigg[\left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S_j)^\dagger S_j^T K^\star)\right\}\mathbb{E}[(I - K^\star S_1(S_1^T(K^\star)S_1)^\dagger S_j^T K^\star) \mid S_{2:\tau}] \\
&\qquad \left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S_j)^\dagger S_j^T K^\star)\right\}^T\Bigg] \\
&\preceq \Upsilon_K^2\Upsilon_m\rho\mathbb{E}\left[\left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}\left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}^T\right] \\
&\preceq \Upsilon_K^2\Upsilon_m\rho^\tau \cdot I,
\end{aligned}$$

where the second inequality from the end is from Assumption 3.3 and Corollary 4.4; and the last inequality applies the same derivation for sketch matrices $S_{2:\tau}$. Combining the above two displays,

$$\left\|\mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T] - \Omega^\star\right\| \leq O(\rho^\tau). \tag{C.34}$$

For the third term in (C.32), we have

$$\begin{aligned}
&\|\Omega_t - \Omega^\star\| \overset{(25)}{=} O(\|K_t - K^\star\|) \\
&+ O\left(\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)^T - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star)\right\|\right).
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)^T - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star)\right\| \\
&\leq \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)]\right\| + \left\|\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)^T - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star)\right\|.
\end{aligned}$$

We take the first term as an example, while the second term has the same guarantee following the same derivations. We note that

$$\begin{aligned}
\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)]\right\| &\leq \left\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{g}_i\bar{g}_i^T) - \mathbb{E}[\bar{g}_i\bar{g}_i^T \mid \mathcal{F}_{i-1}]\right\| \\
&+ \left\|\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}[\bar{g}_i\bar{g}_i^T \mid \mathcal{F}_{i-1}] - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)]\right\|.
\end{aligned}$$

By (15c), we know the first term on the right hand side is a square integrable martingale. The strong law of large number (Duflo, 1997, Theorem 1.3.15) suggests that

$$\left\|\frac{1}{t}\sum_{i=0}^{t-1}(\bar{g}_i\bar{g}_i^T) - \mathbb{E}[\bar{g}_i\bar{g}_i^T \mid \mathcal{F}_{i-1}]\right\| = o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right).$$

By (C.9), (C.29), (C.30), (C.31), the second term on the right hand side can be bounded by

$$\left\|\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}[\bar{g}_i\bar{g}_i^T \mid \mathcal{F}_{i-1}] - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)]\right\| = O\left(\sqrt{\varphi_t\log(1/\varphi_t)}\right) + O\left(\frac{\chi_t}{\varphi_t}\right).$$

Combining the above five displays with Lemma 4.9, we have

$$\|\Omega_t - \Omega^\star\| = O\left(\sqrt{\varphi_t\log(1/\varphi_t)}\right) + O\left(\frac{\chi_t}{\varphi_t}\right) + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right) + \omega_t. \tag{C.35}$$

Combining (C.32), (C.33), (C.34), and (C.35), we complete the proof.

C.11 Proof of Corollary 4.13

We let τ large enough such that $\boldsymbol{w}^T\Xi^\star\boldsymbol{w} \neq \boldsymbol{0}$. We note that

$$\frac{\boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi_t\boldsymbol{w}}} = \frac{\boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}\cdot\sqrt{1 + \frac{\boldsymbol{w}^T\Xi_t\boldsymbol{w} - \boldsymbol{w}^T\Xi^\star\boldsymbol{w}}{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}}}.$$

Thus, combining Lemmas 4.12, A.5 with Theorem 4.11, we complete the proof.

C.12 Proof of Theorem 4.14

By the Raabe's test, we know that (32) implies that (18) holds. Thus, the convergence of $\|\nabla\mathcal{L}_t\|$ comes from Theorem 3.7. The convergence of K_t comes from Theorem 3.12. Furthermore, we note that (33) implies (26) and (29). To show this, we first note from (32) that $\chi - \beta < 0$. Thus, $\chi_t = o(\beta_t)$ (cf. Lemma A.2). This implies $\beta_t \leq \varphi_t \leq \beta_t + o(\beta_t)$, and further $\lim_{t\to\infty} t\varphi_t = \lim_{t\to\infty} t\beta_t = \widetilde{\beta}$. Moreover, we have

$$\lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}}{\varphi_t}\right) = \lim_{t\to\infty} t\left(1 - \frac{2\beta_{t-1}+\chi_{t-1}}{2\beta_t+\chi_t}\right) = \lim_{t\to\infty} t\left(1 - \frac{\beta_{t-1}}{\beta_t} + \frac{\beta_{t-1}}{\beta_t}\left\{1 - \frac{2+\chi_{t-1}/\beta_{t-1}}{2+\chi_t/\beta_t}\right\}\right)$$
$$= \beta + \lim_{t\to\infty} t\left(1 - \frac{2+\chi_{t-1}/\beta_{t-1}}{2+\chi_t/\beta_t}\right) = \beta + \frac{1}{2}\lim_{t\to\infty} t\left(\frac{\chi_t}{\beta_t} - \frac{\chi_{t-1}}{\beta_{t-1}}\right) \quad (\text{since } \chi_t = o(\beta_t))$$
$$= \beta + \frac{1}{2}\lim_{t\to\infty}\frac{\chi_t}{\beta_t}\cdot t\left(1 - \frac{\chi_{t-1}}{\chi_t}\cdot\frac{\beta_t}{\beta_{t-1}}\right) = \beta + \frac{\chi-\beta}{2}\lim_{t\to\infty}\frac{\chi_t}{\beta_t} = \beta.$$

Thus, (33) implies (26) and (29); and the asymptotic convergence rates of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ and K_t come from Theorem 4.8 and Lemma 4.9. Finally, it is easy to see (32), (33), and (34) imply (30). In fact, it suffices to show $1 - \rho^\tau + (\beta-1)/(2\widetilde{\beta}) > 0$. When $\widetilde{\beta} = \infty$, it holds naturally. When $\widetilde{\beta} \in (0, \infty)$, then we know $\beta = -1$ (otherwise, $\beta \in (-1, -0.5)$ implies $t\beta_t \to \infty$). Thus, $1 - \rho^\tau + (\beta-1)/(2\widetilde{\beta}) = 1 - \rho^\tau + \beta/\widetilde{\beta} > 0$, as implied by (33). Thus, the asymptotic normality and Berry-Esseen bound come from Theorem 4.11 and Corollary 4.13.