Journal of Machine Learning Research 26 (2025) 1-75 Submitted 4/24; Revised 12/24; Published 2/25

Statistical Inference of Constrained Stochastic Optimization via Sketched Sequential Quadratic Programming

Sen Na SENNA@GATECH.EDU
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: Zhihua Zhang

Abstract

We consider online statistical inference of constrained stochastic nonlinear optimization problems. We apply the Stochastic Sequential Quadratic Programming (StoSQP) method to solve these problems, which can be regarded as applying second-order Newton's method to the Karush-Kuhn-Tucker (KKT) conditions. In each iteration, the StoSQP method computes the Newton direction by solving a quadratic program, and then selects a proper *adaptive* stepsize $\bar{\alpha}_t$ to update the primal-dual iterate. To reduce dominant computational cost of the method, we *inexactly* solve the quadratic program in each iteration by employing an iterative sketching solver. Notably, the approximation error of the sketching solver need not vanish as iterations proceed, meaning that the per-iteration computational cost does not blow up. For the above StoSQP method, we show that under mild assumptions, the rescaled primal-dual sequence $1/\sqrt{\bar{\alpha}_t}\cdot(\boldsymbol{x}_t-\boldsymbol{x}^\star,\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star)$ converges to a mean-zero Gaussian distribution with a nontrivial covariance matrix depending on the underlying sketching distribution. To perform inference in practice, we also analyze a plug-in covariance matrix estimator. We illustrate the asymptotic normality result of the method both on benchmark nonlinear problems in CUTEst test set and on linearly/nonlinearly constrained regression problems.

Keywords: constrained stochastic optimization, Newton sketching, online inference, uncertainty quantification, randomized numerical linear algebra

1. Introduction

We consider equality-constrained stochastic nonlinear optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f(\boldsymbol{x})=\mathbb{E}_{\mathcal{P}}[F(\boldsymbol{x};\xi)], \quad \text{s.t. } c(\boldsymbol{x})=\boldsymbol{0}, \tag{1.1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective function, $F(\cdot;\xi):\mathbb{R}^d\to\mathbb{R}$ is a realization with a random variable $\xi\sim\mathcal{P}$, and $c:\mathbb{R}^d\to\mathbb{R}^m$ provides deterministic equality constraints. Problems of this form appear widely in a variety of applications in statistics and machine learning, including constrained M-estimation (Geyer, 1991, 1994; Wets, 1999), multi-stage stochastic optimization (Dantzig and Infanger, 1993; Veliz et al., 2014), physics-informed neural networks

(Karniadakis et al., 2021; Cuomo et al., 2022), and algorithmic fairness (Zafar et al., 2019). In practice, the random variable ξ corresponds to a data sample; $F(\boldsymbol{x};\xi)$ is the loss occurred at the sample ξ when using the parameter $\boldsymbol{x}$ to fit the model; and $f(\boldsymbol{x})$ is the expected loss. Deterministic constraints are prevalent in real examples, which can encode prior model information, address identifiability issue, and/or reduce searching complexity.

In this paper, we are particularly interested in performing statistical inference on a (local) primal-dual solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ of Problem (1.1). To that end, the classical (offline) approach often generates N samples $\xi_1, \ldots, \xi_N \sim \mathcal{P}$ iid, and then solves the corresponding empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad \widehat{f}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} F(\boldsymbol{x};\xi_i), \quad \text{s.t. } c(\boldsymbol{x}) = \boldsymbol{0}. \tag{1.2}$$

Under certain regularity conditions, we can establish the asymptotic consistency and normality of the minimizer $(\widehat{\boldsymbol{x}}_N, \widehat{\boldsymbol{\lambda}}_N)$ of (1.2), also called *constrained M-estimator*, given by

$$\sqrt{N}\begin{pmatrix}\widehat{\boldsymbol{x}}_N - \boldsymbol{x}^\star \\ \widehat{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}^\star\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (G^\star)^T \\ G^\star & \boldsymbol{0}\end{pmatrix}^{-1}\begin{pmatrix}\mathrm{cov}(\nabla F(\boldsymbol{x}^\star;\xi)) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (G^\star)^T \\ G^\star & \boldsymbol{0}\end{pmatrix}^{-1}\right), \tag{1.3}$$

where $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$ is the Lagrangian function with $\boldsymbol{\lambda}\in\mathbb{R}^m$ being the dual variables associated with the constraints, $\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star$ is the Lagrangian Hessian with respect to $\boldsymbol{x}$ evaluated at $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$, and $G^\star = \nabla c(\boldsymbol{x}^\star)\in\mathbb{R}^{m\times d}$ is the constraints Jacobian. See (Shapiro et al., 2014, Chapter 5) for the result of (1.3), and Duchi and Ruan (2021) and Davis et al. (2024) for showing (1.3) attains the minimax optimality. Numerous methods can be applied to solve constrained ERM (1.2), including (exact) penalty methods, augmented Lagrangian methods, and sequential quadratic programming (SQP) methods (Nocedal and Wright, 2006).

Given the prevalence of streaming datasets in modern problems, offline methods that require dealing with a large batch set in each step are less attractive. It is desirable to design *fully online methods*, where only a *single* sample is used in each step, and to perform *online statistical inference* by leveraging those methods. Without constraints, one can apply stochastic gradient descent (SGD) and its many variates, whose statistical properties (e.g., asymptotic normality) have been comprehensively studied from different aspects (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Polyak and Juditsky, 1992; Ruppert, 1988). However, unlike solving unconstrained stochastic programs, there are limited methods proposed for constrained stochastic programs (1.1) that enable online statistical inference. We refer to Section 2.2 for a detailed literature review. One potential exception is the projection-based SGD recently studied in Duchi and Ruan (2021); Davis et al. (2024). Although the literature has shown that projected methods also exhibit asymptotic normality, there are two major concerns when applying these methods for practical statistical inference.

(a) It is unclear how to online estimate the limiting covariance matrix based on the projected primal iterates. Due to the absence of dual update, the joint primal-dual normality as in (1.3) is not (at least, immediately) achievable for projected methods. Even for the primal normality, the covariance matrix still depends on the dual solution through the Lagrangian Hessian (cf. (1.3)). However, due to intrinsic objective noise, simply using primal iterates and optimality conditions to solve for the dual solution does not yield a consistent dual estimator for the underlying plug-in covariance estimation. One possible resolution is to draw inspiration from

long-run variance estimations of stationary processes, and design batch-means covariance estimators that utilize only the projected primal iterates themselves. That said, this approach is highly nontrivial for projected methods (because of the non-stationarity), as studied in the context of vanilla SGD methods (Chen et al., 2020; Zhu et al., 2021).

(b) There are prevalent scenarios where the projection operator becomes intractable. For example, when the constraint function $c(\boldsymbol{x})$ is nonlinear and nonconvex, as in physics-informed neural networks (cf. Section 2.1), the projection onto the manifold $\{\boldsymbol{x} \in \mathbb{R}^d : c(\boldsymbol{x}) = \boldsymbol{0}\}$ is generally intractable. Additionally, if we only have local information about the constraint function $c(\boldsymbol{x})$ (e.g., function evaluation and Jacobian) at any given point $\boldsymbol{x}$, as in CUTEst benchmark nonlinear problems (cf. Section 6), the projection operator is not computable either, which requires a global characterization of the constraint set.

To perform online statistical inference of Problem (1.1) without relying on projections, we draw inspiration from a recent growing series of literature in numerical optimization, which develops various *stochastic sequential quadratic programming* (StoSQP) methods for (1.1). The SQP methods can be regarded as second-order Newton's methods applied to the Karush-Kuhn-Tucker (KKT) conditions. In particular, the StoSQP methods compute a stochastic Newton direction in each iteration by solving a quadratic program, whose objective model is estimated using the new sample. Then, the methods select a proper stepsize to achieve a sufficient reduction on the *merit function*, which balances the optimality and feasibility of the iterates. We refer to Na et al. (2022a, 2023); Berahas et al. (2021, 2023); Fang et al. (2024) and references therein for recent StoSQP designs and their promising performance on various problems. The aforementioned literature established the global convergence of StoSQP methods, where the KKT residual $\|\nabla\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\|$ converges to zero almost surely or in expectation for any initialization. However, in contrast to Duchi and Ruan (2021); Davis et al. (2024) studying projection-based methods, these literature overlooked the statistical properties and failed to quantify the uncertainty inherent in the StoSQP methods, which is yet crucial for applying these methods on online statistical inference tasks. Thus, we pose the following question:

Can we perform online inference on $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ *based on the StoSQP iterates, while further reducing the computational cost of existing second-order StoSQP methods?*

In this paper, we answer this question by complementing the global convergence guarantees and establishing the local asymptotic properties of existing StoSQP methods. Specifically, we focus on an Adaptive Inexact StoSQP scheme, referred to as `AI-StoSQP`. By *adaptive* we mean that the scheme inherits the critical merit of numerical StoSQP designs (Berahas et al., 2021; Curtis et al., 2021; Berahas et al., 2023), allowing for an adaptive stepsize $\bar{\alpha}_t$ for the Newton direction. In other words, we do not compromise the adaptivity of StoSQP to establish the local convergence guarantees. By *inexact* we mean that the scheme further reduces the computational cost of StoSQP methods by applying an iterative sketching solver to inexactly solve the Newton system in each step(Strohmer and Vershynin, 2008; Gower and Richtárik, 2015; Pilanci and Wainwright, 2016, 2017; Lacotte et al., 2020). Solving Newton systems is considered the most computationally expensive step of second-order methods; and randomized solvers offer advantages over deterministic solvers by requiring less flops and memory when equipped with proper sketching matrices (e.g., sparse sketches). Notably, we perform a constant number of sketching steps; thus, the per-iteration computational cost remains fixed even near stationarity.

For the above sketched StoSQP scheme, we quantify its uncertainty consisting of three components: random sampling, random sketching, and random stepsize. We establish the asymptotic normality of the primal-dual iterate:

$$1/\sqrt{\bar{\alpha}_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \xrightarrow{d} \mathcal{N}(0, \Xi^\star), \tag{1.4}$$

where the limiting covariance $\Xi^\star$ solves a Lyapunov equation that depends on the underlying sketching distribution used in the sketching solver. Let $\Omega^\star$ denote the limiting covariance of constrained M-estimator in (1.3). Our result suggests that if $\bar{\alpha}_t \asymp 1/t$, then we have two cases:

$$\Xi^\star = \Omega^\star \quad \text{for the exact solver,} \qquad\qquad \Xi^\star \succeq \Omega^\star \quad \text{for the sketching solver.} \tag{1.5}$$

This implies that (i) if we solve Newton systems exactly, then online StoSQP estimator (even with adaptive stepsizes) achieves the same estimation efficiency as offline M-estimator. In fact, if we focus solely on the primal variables $\boldsymbol{x}$, the marginal covariance of $\Xi^\star = \Omega^\star$ also matches the limiting covariance of online projection-based estimators established in Duchi and Ruan (2021); Davis et al. (2024), which is known to be *asymptotic minimax optimal* (see Remark 5.8). (ii) If we solve Newton systems inexactly, then the sketching solver hurts the asymptotic optimality of StoSQP as $\Xi^\star \succeq \Omega^\star$. Fortunately, the hurt is tolerable as seen from the bound (cf. Corollary 5.7)

$$\|\Xi^\star - \Omega^\star\| \lesssim \rho^\tau \quad \text{for some } \rho \in (0,1),$$

where τ is the number of iterations we run for the sketching solver at each step. In addition to asymptotic normality, we also present some by-product results of independent interest, including the local convergence rate, sample complexity, and the Berry-Esseen bound that quantitatively measures the convergence rate in (1.4). To facilitate practical inference, we also analyze a plug-in covariance estimator that can be computed in online fashion. We illustrate our results on benchmark nonlinear problems in CUTEst test set and on linearly/nonlinearly constrained regression problems.

Structure of the paper. We introduce some motivating examples of Problem (1.1) and provide a literature review in Section 2. Then, we introduce `AI-StoSQP` in Section 3 and prove the global almost sure convergence with iteration complexity in Section 4. Asymptotic normality with covariance estimation is established in Section 5. Experiments and conclusions are presented in Sections 6 and 7, respectively. We defer all the proofs to the appendices.

Notation. Throughout the paper, we use $\|\cdot\|$ to denote ℓ_2 norm for vectors and spectral norm for matrices. For scalars a, b, $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. We use $O(\cdot)$ (or $o(\cdot)$) to denote big (or small) O notation in usual almost sure sense. 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 (the identity matrix) if $j < i$. We use the bar notation, $(\bar{\cdot})$, to denote algorithmic quantities that are random (i.e., depending on realized samples), except for the iterates. We reserve the notation $G(\boldsymbol{x})$ to denote the constraints Jacobian, i.e., $G(\boldsymbol{x}) = \nabla c(\boldsymbol{x}) = (\nabla c_1(\boldsymbol{x}), \ldots, \nabla c_m(\boldsymbol{x}))^T \in \mathbb{R}^{m \times d}$.

2. Applications and Literature Review

We present two motivating examples of (1.1) in Section 2.1, and then review related literature in Section 2.2.

2.1 Motivating examples

Many statistical and machine learning problems can be cast into the form of Problem (1.1).

Example 1 (Constrained regression problems) *Let $\xi_t = (\xi_{\boldsymbol{a}_t}, \xi_{b_t})$ be the t-th sample, where $\xi_{\boldsymbol{a}_t} \in \mathbb{R}^d$ is the feature vector independently drawn from some multivariate distribution and ξ_{b_t} is the response. We consider different regression models, such as*

$$\begin{aligned} &\textit{linear models:} && \xi_{b_t} = \xi_{\boldsymbol{a}_t}^T \boldsymbol{x}^\star + \epsilon_t && \textit{with } \epsilon_t \textit{ iid noise,} \\ &\textit{logistic models:} && P\left(\xi_{b_t} \mid \xi_{\boldsymbol{a}_t}\right) = \frac{\exp\left(\xi_{b_t} \cdot \xi_{\boldsymbol{a}_t}^T \boldsymbol{x}^\star\right)}{1+\exp\left(\xi_{b_t} \cdot \xi_{\boldsymbol{a}_t}^T \boldsymbol{x}^\star\right)} && \textit{with } \xi_{b_t} \in \{-1,1\}, \end{aligned}$$

where $\boldsymbol{x}^\star \in \mathbb{R}^d$ is the true model parameter. For the above models, we define the corresponding loss functions at $\boldsymbol{x}$:

$$\begin{aligned} \textit{linear models:} \quad & F(\boldsymbol{x}; \xi_t) = \frac{1}{2}(\xi_{\boldsymbol{a}_t}^T \boldsymbol{x} - \xi_{b_t})^2, \\ \textit{logistic models:} \quad & F(\boldsymbol{x}; \xi_t) = \log\left(1+\exp(-\xi_{b_t} \cdot \xi_{\boldsymbol{a}_t}^T \boldsymbol{x})\right). \end{aligned}$$

Then, we can verify that $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{x}} \mathbb{E}[F(\boldsymbol{x}; \xi)]$. In many cases, we access prior information about the model parameters, which is encoded as constraints. For example, in portfolio selection, $\boldsymbol{x}$ represents the portfolio allocation vector that satisfies $\boldsymbol{x}^T \mathbf{1} = 1$ (total allocation is 100%). We may also fix the target percentage on each sector or each region, which is translated into the constraint $A\boldsymbol{x} = \boldsymbol{d}$. See (Fan, 2007, (4.3, 4.4)), (Fan et al., 2012, (2.1)), (Du et al., 2022, (1)) and references therein for such applications. In principle component analysis and semiparametric single/multiple index regressions, we enforce $\boldsymbol{x}$ to have a unit norm to address the identifiability issue, leading to a nonlinear constraint $\|\boldsymbol{x}\|^2 = 1$. We point to Kaufman and Pereyra (1978); Kirkegaard and Eldrup (1972); Sen (1979); Nagaraj and Fuller (1991); Dupacova and Wets (1988); Shapiro (2000); Na et al. (2019); Na and Kolar (2021) for various examples of linearly/nonlinearly constrained estimation problems. For some constrained estimation problems, projecting into the feasible set can be intractable. For instance, in factor analysis, researchers may estimate a covariance matrix Σ under so-called tetrad constraints: $\Sigma_{i_1 i_2}\Sigma_{i_3 i_4} - \Sigma_{i_1 i_4}\Sigma_{i_2 i_3} = 0$ for every set of four distinct variables $\{i_1, i_2, i_3, i_4\}$ (Bollen and Ting, 2000; Drton and Xiao, 2016; Sturma et al., 2024). For such highly nonlinear constraints, the linear-quadratic approximation performed in SQP can be a promising resolution.

Example 2 (Physics-informed machine learning) *Recent decades have seen machine learning (ML) making significant inroads into science. The major task in ML is to learn an unknown mapping $\boldsymbol{z}(\cdot) : \mathcal{A} \to \mathcal{B}$ from data that can perform well in the downstream tasks. Since $\boldsymbol{z}(\cdot)$ is infinite-dimensional, one key step in ML is to use neural networks (NNs) to parameterize $\boldsymbol{z}(\cdot)$ as $\boldsymbol{z}_{\boldsymbol{x}}(\cdot)$, and learn the optimal weight parameters $\boldsymbol{x} \in \mathbb{R}^d$ instead (called function approximation). One of the trending topics in ML now is physics-informed ML, where one requires $\boldsymbol{z}(\cdot)$ to obey some physical principles that are often characterized by partial differential equations (PDEs) (Karniadakis et al., 2021; Cuomo et al., 2022). In such applications, we can use the squared loss function, defined for the t-th sample $\xi_t = (\xi_{\boldsymbol{a}_t}, \xi_{\boldsymbol{b}_t}) \in \mathcal{A} \times \mathcal{B}$ as*

$$F(\boldsymbol{x};\xi_t)=\frac{1}{2}\left(\boldsymbol{z}_{\boldsymbol{x}}(\xi_{\boldsymbol{a}_t})-\xi_{\boldsymbol{b}_t}\right)^2.$$

Here, $\xi_{\boldsymbol{a}_t}$ is NN inputs that can be spatial and/or temporal coordinates; $\xi_{\boldsymbol{b}_t}$ is measurements that can be speed, velocity, and temperature, etc; and $\boldsymbol{z}_{\boldsymbol{x}}\in\mathcal{C}^\infty(\mathcal{A},\mathcal{B})$ is NN architecture. Let $\mathcal{F}:\mathcal{C}^\infty(\mathcal{A},\mathcal{B})\to\mathcal{C}^\infty(\mathcal{A},\mathcal{B})$ be the PDE operator, which encodes the underlying physical law (e.g., energy conservation law). We aim to find optimal weights $\boldsymbol{x}^\star$ that not only minimize the mean squared error of observed data, but also satisfy the constraints $\mathcal{F}(\boldsymbol{z}_{\boldsymbol{x}})=\mathbf{0}$. To this end, we select some leverage points $\{\xi'_{\boldsymbol{a}_i}\}_{i=1}^m$ in $\mathcal{A}$ and impose deterministic constraints:

$$\mathcal{F}(\boldsymbol{z}_{\boldsymbol{x}})(\xi'_{\boldsymbol{a}_i})=\mathbf{0},\quad\forall i=1,2,\ldots,m.$$

Here, we abuse the notation $\mathbf{0}$ to denote either a zero mapping of $\mathcal{C}^\infty(\mathcal{A},\mathcal{B})$ or a zero element of $\mathcal{B}$. For more details on this problem formulation, see (Lu et al., 2021, (2.3)), (Krishnapriyan et al., 2021, (2)), and references therein. Due to the nonlinearity nature of NNs, projection-free methods are desired, and SQP can achieve competitive performance compared to penalty methods and augmented Lagrangian methods (Cheng and Na, 2024).

2.2 Related literature and contribution

There are numerous methods for solving constrained optimization problems, such as projection-based methods, penalty methods, augmented Lagrangian methods, and sequential quadratic programming (SQP) methods (Nocedal and Wright, 2006). This paper particularly considers solving constrained stochastic optimization problems via Stochastic SQP (StoSQP) methods, which can be regarded as an application of stochastic Newton's method on constrained problems. Berahas et al. (2021) designed the very first online StoSQP scheme. At each step, the method selects a suitable penalty parameter of an ℓ_1-penalized objective; ensures the Newton direction produces a sufficient reduction on the penalized objective; and then selects an adaptive stepsize $\beta_t\le\bar{\alpha}_t\le\eta_t=\beta_t+\chi_t$ based on input sequences β_t and $\chi_t=O(\beta_t^2)$. An alternative StoSQP scheme was then reported in Na et al. (2022a), where $\bar{\alpha}_t$ is selected by performing stochastic line search on the augmented Lagrangian with batch sizes increasing as iteration proceeds. Subsequently, Curtis et al. (2021); Na et al. (2023); Berahas et al. (2023); Fang et al. (2024) proposed different variates of StoSQP to cope with inequality constraints, degenerate constraints, etc. These works all proved the global convergence of StoSQP methods — the KKT residual $\|\nabla\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)\|$ converges to zero from any initialization. However, they fall short of uncertainty quantification and online statistical inference goals.

On the other hand, a growing body of literature leverages optimization procedures to facilitate online inference, starting with Robbins and Monro (1951); Kiefer and Wolfowitz (1952) and continuing through Robbins and Siegmund (1971); Fabian (1973); Ermoliev (1983). To study the asymptotic distribution of stochastic gradient descent (SGD), Ruppert (1988) and Polyak and Juditsky (1992) averaged SGD iterates and established the optimal central limit theorem rate. Toulis and Airoldi (2017) designed an implicit SGD method and showed the asymptotics of averaged implicit SGD iterates. Li et al. (2018) designed an inference procedure for constant-stepsize SGD by averaging the iterates with recurrent burn-in periods. Mou et al. (2020) further showed the asymptotic covariance of constant-stepsize SGD with Ployak-Ruppert averaging. Liang and Su (2019) designed a moment-adjusted SGD method and

provided non-asymptotic results that characterize the statistical distribution as the batch size of each step tends to infinity. Chen et al. (2020) and Zhu et al. (2021) proposed different covariance matrix estimators constructed by grouping SGD iterates. Additionally, Chen et al. (2021) designed a distributed method for the inference of non-differentiable convex problems; Roy and Balasubramanian (2023) analyzed a batch-mean covariance estimator under a Markovian sampling setup; and Duchi and Ruan (2021) and Davis et al. (2024) applied projection-based SGD methods for the inference of inequality-constrained convex problems. The aforementioned literature all studied first-order methods with deterministic stepsizes.

The asymptotics of second-order Newton's methods for unconstrained problems have recently been investigated. Bercu et al. (2020) designed an online Newton's method for logistic regression, and Boyer and Godichon-Baggioni (2023) generalized that method to general regression problems. Compared to first-order methods that often consider averaged iterates and/or exclude the stepsize $1/t$ due to technical challenges, both works showed the normality of the *last* iterate with $1/t$ stepsize. However, those analyses are not applicable to our study for two reasons. First, they studied unconstrained regression problems with objectives in the form $F(\boldsymbol{x}^T\xi)$, resulting in objective Hessians owning rank-one updates that cannot be employed for our general problem (1.1). Second, they solved Newton systems exactly and utilized $1/t$ deterministic stepsize. In contrast, we use a randomized sketching solver to solve Newton systems inexactly to reduce the computational cost associated with higher-order methods, along with an adaptive random stepsize inspired by numerical designs in Berahas et al. (2021). Both of these components affect the uncertainty quantification and lead to a different normality result (cf. (1.5)). To our knowledge, this is the first work that performs online inference by taking into account not only the randomness of samples but also the randomness of computation (i.e., sketching and stepsize); the latter is particularly important for making second-order methods computationally promising.

We briefly review the literature on sketched Newton methods. Compared to the works below, the sketching step in StoSQP is only a subroutine for solving linear-quadratic programs; a complete method also involves merit function reduction and stepsize selection (here, stepsize refers to that of StoSQP rather than the sketching solver). This paper focuses on uncertainty quantification and statistical inference of online (sketched) Newton methods, which differs significantly from the following literature that focuses on design and convergence of sketched Newton methods. In particular, for many (regression) problems, the objective Hessian can be expressed as $\mathbf{H} = AA^T \in \mathbb{R}^{d\times d}$ with a data matrix $A \in \mathbb{R}^{d\times n}$ and $n \geq d$. Then, one can generate a sketch matrix $S\in\mathbb{R}^{n\times s}$ and compute the approximate Hessian $\widehat{\mathbf{H}}=ASS^TA^T$. Pilanci and Wainwright (2016) developed an iterative Hessian sketch algorithm for solving least-squares problems $\min_{\boldsymbol{x}} \|A\boldsymbol{x}-b\|^2$ (subject to convex constraints). The authors sketched only the data matrix A rather than both the data matrix A and vector b, and established a high-probability convergence result. Lacotte et al. (2020) later extended this study by showing the optimal stepsize and convergence rate for Haar sketches. Pilanci and Wainwright (2017) designed a sketched Newton method that approximates the Hessian using the Johnson–Lindenstrauss transform. Building on this, Agarwal et al. (2017); Derezinski and Mahoney (2019); Derezinski et al. (2020a,b, 2021); Lacotte et al. (2021) introduced various sketching methods to explore the trade-off between the computational cost of $\widehat{\mathbf{H}}$ and the convergence rate of the algorithm. In addition to the above series of literature, another type of sketched Newton method is based on Sketch-and-Project framework, where one approximates a generic Hessian inverse $\mathbf{H}^{-1}$ by $S(S^T\mathbf{H}S)^\dagger S^T$

for a sketch matrix $S \in \mathbb{R}^{d\times s}$. See Strohmer and Vershynin (2008); Gower and Richtárik (2015); Luo et al. (2016); Doikov et al. (2018); Gower et al. (2019); Dereziński and Rebrova (2024) and references therein for the convergence properties of this family of methods.

Compared to deterministic methods for solving Newton systems, such as conjugate gradient and broad preconditioned Krylov (or minimal residual) methods, randomized sketching methods may behave better in terms of improved convergence rates and range of convergence, while requiring less computation and memory (when using suitable sketches) to be scalable and parallelizable (Gower, 2016). We refer to Hong et al. (2023) for an empirical demonstration of the advantages of sketching solvers over deterministic solvers in the context of SQP methods.

3. Adaptive Inexact StoSQP Method

Let $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x})+\boldsymbol{\lambda}^T c(\boldsymbol{x})$ be the Lagrangian function of (1.1), where $\boldsymbol{\lambda}\in\mathbb{R}^m$ is the dual vector. Under certain constraint qualifications (introduced later), a necessary condition for $(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star)$ being a local solution to (1.1) is the KKT conditions: $\nabla\mathcal{L}^\star = (\nabla_{\boldsymbol{x}}\mathcal{L}^\star, \nabla_{\boldsymbol{\lambda}}\mathcal{L}^\star) = (\mathbf{0},\mathbf{0})$.

`AI-StoSQP` applies Newton's method to the equation $\nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}$, involving three steps: estimating the objective gradient and Hessian, (inexactly) solving Newton's system, and updating the primal-dual iterate. We detail each step as follows. For simplicity, we denote $c_t = c(\boldsymbol{x}_t)$ (similarly, $G_t = \nabla c(\boldsymbol{x}_t)$, $\nabla\mathcal{L}_t = \nabla\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)$, etc.).

• **Step 1: Estimate the gradient and Hessian**. We realize a sample $\xi_t\sim\mathcal{P}$ and estimate the gradient ∇f_t and Hessian $\nabla^2 f_t$ of the objective as

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

Then, we compute three quantities:

$$\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t = \bar{g}_t + G_t^T\boldsymbol{\lambda}_t, \quad \bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_t = \bar{H}_t + \sum_{i=1}^m(\boldsymbol{\lambda}_t)_i\nabla^2 c_i(\boldsymbol{x}_t), \quad B_t = \frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i + \Delta_t.$$

Here, $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$ and $\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_t$ are the estimates of the Lagrangian gradient and Hessian with respect to $\boldsymbol{x}$, respectively; and B_t is a regularized averaged Hessian used in the quadratic program (3.1). We let $\Delta_t = \Delta(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)$ be any regularization term ensuring B_t to be positive definite in the null space $\{\boldsymbol{x}\in\mathbb{R}^d : G_t\boldsymbol{x} = \mathbf{0}\}$. Note that the average $\sum_{i=0}^{t-1}\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i/t$ can be updated online. We explain the matrix B_t in the following remark.

Remark 3.1 *We note that the Lagrangian Hessian average in B_t is over samples $\{\xi_0,\dots,\xi_{t-1}\}$, meaning that the Hessian estimate $\bar{H}_t$, which depends on the new sample ξ_t, will only be used in the $(t+1)$-th iteration. Thus, B_t and Δ_t are deterministic given $(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)$. In addition, Δ_t can simply be Levenberg-Marquardt type regularization of the form $\delta_t I$ with suitably large $\delta_t > 0$ (Nocedal and Wright, 2006). The Hessian regularization is standard for nonlinear problems, together with linear independence constraint qualification (LICQ, Assumption 4.1), ensuring that the quadratic program* (3.1) *is solvable. For convex problems, we just set $\Delta_t = \mathbf{0}$, $\forall t$. See Bertsekas (1982); Nocedal and Wright (2006) for various regularization approaches. Moreover, Na et al. (2022b) showed that Hessian averaging accelerates the local rate of Newton's method on unconstrained deterministic problems.*

• **Step 2: Solve the quadratic program**. With the above estimates, we solve the quadratic program (QP):

$$\min_{\widetilde{\Delta}\boldsymbol{x}_t\in\mathbb{R}^d} \quad \frac{1}{2}\widetilde{\Delta}\boldsymbol{x}_t^T B_t\ \widetilde{\Delta}\boldsymbol{x}_t + \bar{g}_t^T\widetilde{\Delta}\boldsymbol{x}_t, \quad \text{s.t. } c_t + G_t\widetilde{\Delta}\boldsymbol{x}_t = \mathbf{0}. \tag{3.1}$$

For the above QP, the objective can be seen as a quadratic approximation of $F(\boldsymbol{x};\xi)$ at $(\boldsymbol{x}_t;\xi_t)$, and the constraint can be seen as a linear approximation of $c(\boldsymbol{x})$ at $\boldsymbol{x}_t$. It is easy to observe that solving the above QP is equivalent to solving the following Newton system

$$\underbrace{\begin{pmatrix} B_t & G_t^T \\ G_t & \mathbf{0} \end{pmatrix}}_{K_t} \underbrace{\begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_t \\ \widetilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}}_{\widetilde{\boldsymbol{z}}_t} = -\underbrace{\begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}}_{\bar{\nabla}\mathcal{L}_t}, \tag{3.2}$$

where K_t, $\bar{\nabla}\mathcal{L}_t$ are the Lagrangian Hessian and gradient, and $\widetilde{\boldsymbol{z}}_t$ is the exact Newton direction.

Instead of solving the QP (3.1) exactly, we solve it inexactly by an iterative sketching solver. This approach proves more efficient than deterministic solvers, especially when equipped with suitable sketching matrices (Strohmer and Vershynin, 2008; Gower and Richtárik, 2015; Pilanci and Wainwright, 2016, 2017; Lacotte et al., 2020). In particular, we generate a random sketching matrix $S \in \mathbb{R}^{(d+m)\times s}$, whose column dimension $s \geq 1$ can also be random, and transform the original large-scale linear system to the sketched, small-scale system as

$$K_t\boldsymbol{z}_t = -\bar{\nabla}\mathcal{L}_t \qquad \Longrightarrow \qquad S^T K_t\boldsymbol{z}_t = -S^T\bar{\nabla}\mathcal{L}_t.$$

Clearly, there are multiple solutions to the sketched system, and $\boldsymbol{z}_t = \widetilde{\boldsymbol{z}}_t$ is one of them. We prefer the solution that is closest to the current solution approximation. That is, the j-th iteration of the sketching solver has the form ($\boldsymbol{z}_{t,0} = \mathbf{0}$)

$$\boldsymbol{z}_{t,j+1} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z} - \boldsymbol{z}_{t,j}\|^2 \qquad \text{s.t.} \quad S_{t,j}^T K_t\boldsymbol{z} = -S_{t,j}^T\bar{\nabla}\mathcal{L}_t, \tag{3.3}$$

where $S_{t,j} \sim S, \forall j$ are independent and identically distributed and are also independent of ξ_t. An explicit recursion of (3.3) is given by

$$\boldsymbol{z}_{t,j+1} = \boldsymbol{z}_{t,j} - K_t S_{t,j}(S_{t,j}^T K_t^2 S_{t,j})^\dagger S_{t,j}^T(K_t\boldsymbol{z}_{t,j} + \bar{\nabla}\mathcal{L}_t), \tag{3.4}$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. One can let $s = 1$ (i.e., using sketching vectors) so that $S_{t,j}^T K_t^2 S_{t,j}$ reduces to a scalar and the pseudoinverse reduces to the reciprocal.

We perform $\tau \geq 1$ iterations of (3.4) and use

$$(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t) \coloneqq \boldsymbol{z}_{t,\tau}$$

as the approximate Newton direction. We emphasize that τ is independent of t; thus, we do not require a vanishing approximation error and blow up the computational cost as $t \to \infty$.

Remark 3.2 *A significant difference between randomized solvers and deterministic solvers is that the approximation error $\|\boldsymbol{z}_{t,j} - \widetilde{\boldsymbol{z}}_t\|$ of randomized solvers may not be monotonically decreasing as j increases. This subtlety challenges both inference and convergence analysis. In*

Algorithm 1 Adaptive Inexact StoSQP Method

Input: initial iterate $(\boldsymbol{x}_0, \boldsymbol{\lambda}_0)$, positive sequences $\{\beta_t, \eta_t\}$, an integer $\tau > 0$, $B_0 = I$;
for $t = 0, 1, 2, \ldots$ **do**
  Realize ξ_t and compute $\bar{g}_t = \nabla f(\boldsymbol{x}_t; \xi_t)$, $\bar{H}_t = \nabla^2 f(\boldsymbol{x}_t; \xi_t)$, and $\bar{\nabla}^2_{\boldsymbol{x}} \mathcal{L}_t$;
  Compute the regularized Hessian average $B_t = \frac{1}{t} \sum_{i=0}^{t-1} \bar{\nabla}^2_{\boldsymbol{x}} \mathcal{L}_i + \Delta_t$;
  Generate sketching matrices $S_{t,j} \sim S$, $\forall j$ iid and iterate (3.4) for τ times;
  Select any adaptive stepsize $\bar{\alpha}_t$ with $\beta_t \leq \bar{\alpha}_t \leq \eta_t$, and update the iterate as (3.5);
end for

classical optimization world, it is unanimously agreed that if the search direction is asymptotically close to the exact Newton direction (here $\widetilde{\boldsymbol{z}}_t$), then the algorithm will locally behave just like Newton's method with a similar convergence rate. A precise characterization is called the Dennis-Moré condition (Dennis and Moré, 1974). In our study, although Lemma 4.4 shows that the expected error $\mathbb{E}[\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\| \mid \boldsymbol{x}_t, \xi_t]$ decays exponentially in τ, $\boldsymbol{z}_{t,\tau}$ can still be far from $\widetilde{\boldsymbol{z}}_t$ for any t and large τ in the almost sure sense. In particular, (Patel et al., 2021, Theorem 4.2) proved that only a subsequence of $\{\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\|\}_{\tau=0}^{\infty}$ would decrease monotonically with the subsequence indices being also random.

• **Step 3: Update the iterate with an adaptive stepsize**. With the direction $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t) = \boldsymbol{z}_{t,\tau}$ from Step 2, we update the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ by an adaptive stepsize $\bar{\alpha}_t$:

$$(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\lambda}_t) + \bar{\alpha}_t \cdot (\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t). \tag{3.5}$$

In principle, the stepsize $\bar{\alpha}_t$ may rely on the random direction $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$, so it is also random. We allow using any adaptive stepsize selection schemes but require a safeguard condition on $\bar{\alpha}_t$:

$$0 < \beta_t \leq \bar{\alpha}_t \leq \eta_t \quad \text{with} \quad \eta_t \coloneqq \beta_t + \chi_t, \tag{3.6}$$

where $\{\beta_t, \eta_t\}$ are upper and lower bound sequences and χ_t is the adaptivity gap. We do not require specific stepsize selection schemes beyond the condition (3.6) to achieve our online inference goals. The schemes reported in Berahas et al. (2021, 2023); Curtis et al. (2021) all adhere to the condition (3.6). See (Berahas et al., 2021, Lemma 3.6) and (Curtis et al., 2021, (25, 28)) for details. Their numerical experiments suggest that adaptive random stepsizes offer promising empirical benefits over non-adaptive deterministic stepsizes (i.e., $\chi_t = 0$). For sake of completeness, we present a selection scheme from Berahas et al. (2021) in Appendix A.

We combine the above three steps and summarize `AI-StoSQP` in Algorithm 1. To end this section, we introduce a filtration notation for later use. We let $\mathcal{F}_t = \sigma(\{\xi_i, \{S_{i,j}\}_j, \bar{\alpha}_i\}_{i=0}^t)$, $\forall t \geq 0$ be the σ-algebra generated by the random variables $\{\xi_i, \{S_{i,j}\}_j, \bar{\alpha}_i\}_{i=0}^t$. Moreover, we let $\mathcal{F}_{t-2/3} = \sigma(\{\xi_i, \{S_{i,j}\}_j, \bar{\alpha}_i\}_{i=0}^{t-1} \cup \xi_t)$, $\mathcal{F}_{t-1/3} = \sigma(\{\xi_i, \{S_{i,j}\}_j, \bar{\alpha}_i\}_{i=0}^{t-1} \cup \xi_t \cup \{S_{t,j}\}_j)$, 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. With these notation, Algorithm 1 has a generating process as follows: given $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, we first realize ξ_t to get the estimates of the gradient $\bar{g}_t$ and Hessian $\bar{H}_t$ and derive $\mathcal{F}_{t-2/3}$; then we generate $\{S_{t,j}\}_j$ to obtain the inexact Newton direction and derive $\mathcal{F}_{t-1/3}$; then we select the stepsize $\bar{\alpha}_t$ and derive $\mathcal{F}_t$. We also let $(\Delta\boldsymbol{x}_t, \Delta\boldsymbol{\lambda}_t)$ be the exact Newton direction solved from (3.2) with $\bar{\nabla}_{\boldsymbol{x}} \mathcal{L}_t$ being replaced by $\nabla_{\boldsymbol{x}} \mathcal{L}_t$.

4. Global Almost Sure Convergence

In this section, we present the global almost sure convergence[1] for the StoSQP method. We show that the KKT residual $\|\nabla\mathcal{L}_t\|$ converges to zero from any initialization. This global convergence serves as a preliminary result of our inference analysis in Section 5.

We use an adapted augmented Lagrangian function as the Lyapunov function to show the convergence, which has two penalty terms of the form

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

The first penalty term biases the feasibility error; while, in contrast to standard augmented Lagrangian ($\nu = 0$), the extra second penalty term biases the optimality error. We will first show that the inner product between the *exact* Newton direction $(\Delta\boldsymbol{x}_t, \Delta\boldsymbol{\lambda}_t)$ and the augmented Lagrangian gradient $\nabla\mathcal{L}_{\mu,\nu}$, with proper parameters μ and ν, is sufficiently negative (Lemma 4.6). Thus, the Newton direction is a descent direction of $\mathcal{L}_{\mu,\nu}$. Then, we will show that the augmented Lagrangian $\mathcal{L}_{\mu,\nu}$ decreases at each step even with an *inexact* Newton direction (Lemma 4.7). This implies that the residual $\|\nabla\mathcal{L}_t\|$ finally vanishes to zero (Theorem 4.8).

4.1 Assumptions and preliminary results

We state the following assumptions that are standard and proposed in the optimization literature (Kushner and Clark, 1978; Bertsekas, 1982; Nocedal and Wright, 2006; Na et al., 2022a).

Assumption 4.1 *We assume the existence of a closed, bounded, convex set $\mathcal{X}\times\Lambda$ containing the iterates $\{(\boldsymbol{x}_t,\boldsymbol{\lambda}_t)\}_t$, such that f and c are twice continuously differentiable over $\mathcal{X}$. We also assume that the Hessian $\nabla^2\mathcal{L}$ is Υ_L-Lipschitz continuous over $\mathcal{X}\times\Lambda$. In other words,*

$$\|\nabla^2\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) - \nabla^2\mathcal{L}(\boldsymbol{x}',\boldsymbol{\lambda}')\| \le \Upsilon_L\|(\boldsymbol{x}-\boldsymbol{x}',\boldsymbol{\lambda}-\boldsymbol{\lambda}')\|, \quad \forall\,(\boldsymbol{x},\boldsymbol{\lambda}),(\boldsymbol{x}',\boldsymbol{\lambda}')\in\mathcal{X}\times\Lambda. \tag{4.1}$$

Furthermore, we assume that the constraints Jacobian G_t has full row rank with $G_tG_t^T \succeq \gamma_G I$ for a constant $\gamma_G>0$. Additionally, the regularization Δ_t ensures that B_t satisfies $\|B_t\|\le\Upsilon_B$ and $\boldsymbol{x}^TB_t\boldsymbol{x}\ge\gamma_{RH}\|\boldsymbol{x}\|^2$ for any $\boldsymbol{x}\in\{\boldsymbol{x}\in\mathbb{R}^d : G_t\boldsymbol{x}=\boldsymbol{0}\}$, for some constants $\gamma_{RH},\Upsilon_B>0$.

Assumption 4.1 assumes G_t has full row rank, which is referred to as the linear independence constraint qualification (LICQ). LICQ is a common constraint qualification ensuring the uniqueness of the dual solution, and is also necessary for the inference analysis (see (Duchi and Ruan, 2021, Assumption B) and (Davis et al., 2024, Example 2.1)). LICQ and conditions on B_t are critical for SQP methods; they imply QP (3.1) has a unique solution (Nocedal and Wright, 2006, Lemma 16.1). The Lipschitz continuity of the Hessian matrix is also standard for analyzing Newton's method.

The bounded iterates condition is commonly assumed in the literature on (stochastic) nonlinear nonconvex optimization, both for first-order gradient-based methods (Bolte et al., 2013; Song et al., 2014; Davis et al., 2016, 2019; Atchade et al., 2017; Asi and Duchi, 2019; Liu et al.,

[1] Global convergence in nonlinear optimization refers to the convergence to a stationary point from any initialization, in contrast to the convergence to a global solution, which is not achievable without particular problem structures (Nocedal and Wright, 2006). However, they are equivalent for convex problems as studied for projection-based methods in Duchi and Ruan (2021); Davis et al. (2024).

2023a) and for second-order SQP methods ((Bertsekas, 1982, Proposition 4.15), (Nocedal and Wright, 2006, Theorem 18.3)). This assumption ensures that all functions with their gradients and Hessians are bounded over $\mathcal{X} \times \Lambda$ as long as they are smooth. Some literature replaces the bounded iterates condition by directly imposing boundedness on the gradients and Hessians of the objective and constraints, although the main use of the condition in the proof is rather similar (Berahas et al., 2021, 2023; Curtis et al., 2021; Ramprasad et al., 2022; Liu et al., 2023b).

We provide two justifications for the boundedness condition. First, in our study, the StoSQP iterates presumably track a deterministic feasible set $\{\boldsymbol{x} \in \mathbb{R}^d : c(\boldsymbol{x}) = \boldsymbol{0}\}$, so we believe that an unbounded iteration sequence is generally rare especially when the feasible set is bounded. Second, a practical way to enforce the boundedness condition may be through adaptive truncation (Andrieu et al., 2005; Liang, 2010). Under some conditions on the Markov transition kernel of the iteration sequence, one can show that the truncation occurs only finitely many times, ensuring that the convergence and asymptotic behavior are finally not affected by the truncation.

We also impose bounded moment conditions on the stochastic estimates $\bar{g}_t$ and $\bar{H}_t$.

Assumption 4.2 *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 the following moment conditions when needed: for a constant $\Upsilon_m > 0$,*

$$\begin{aligned}
&\textit{gradient} \quad (\textit{bounded 2nd moment}): && \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^2 \mid \boldsymbol{x}_t] \le \Upsilon_m, && (4.2a)\\
&\qquad\qquad (\textit{bounded 3th moment}): && \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^3 \mid \boldsymbol{x}_t] \le \Upsilon_m, && (4.2b)\\
&\qquad\qquad (\textit{bounded 4th moment}): && \mathbb{E}[\|\bar{g}_t - \nabla f_t\|^4 \mid \boldsymbol{x}_t] \le \Upsilon_m, && (4.2c)
\end{aligned}$$

and

$$\begin{aligned}
&\textit{Hessian} \quad (\textit{bounded 2nd moment}): && \mathbb{E}[\|\bar{H}_t - \nabla^2 f_t\|^2 \mid \boldsymbol{x}_t] \le \Upsilon_m, && (4.2d)\\
&\qquad\qquad (\textit{bounded 2nd moment}): && \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla^2 f(\boldsymbol{x};\xi)\|^2] \le \Upsilon_m. && (4.2e)
\end{aligned}$$

We write $\mathbb{E}[\cdot \mid \boldsymbol{x}_t]$ to express out the conditional variable. It can also be written as $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, meaning the expectation is taken over the randomness of sample ξ_t. For conditions (4.2), we do not impose all of them at once, but impose them step by step. In this section, we only require (4.2a) to show the convergence of $\nabla \mathcal{L}_t$. In the next section, we require higher-order moments for inference. In fact, (4.2c) implies (4.2b), which implies (4.2a), and (4.2e) implies (4.2d).

Assumption 4.2 is standard for uncertainty quantification of stochastic methods. We would like to mention that (4.2e) is also required for the asymptotic analysis of averaged SGD (Chen et al., 2020), which 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 (E.11). Please refer to (Chen et al., 2020, Assumption 3.2(2) and Lemma 3.1) for further discussions. Moreover, (4.2e) is satisfied by various objectives, such as logistic and least squares losses in Example 1, as long as the feature variable $\xi_{\boldsymbol{a}}$ has bounded 4-th moment.

In terms of the sketching matrices, we need the following assumption.

Assumption 4.3 *For $t \ge 0$, we assume that the sketching matrices $S_{t,j} \overset{iid}{\sim} S$ satisfy*

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

Assumption 4.3 is required for sketching solvers to converge in expectation (Gower and Richtárik, 2015, Theorem 4.6). This assumption can be easily verified for various sketching ma-

trices. For example, for randomized Kaczmarz method where $S \in \mathbb{R}^{(d+m)\times s}$ has s columns sampled uniformly (without replacement) from the canonical bases $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d+m}\}$ (Strohmer and Vershynin, 2008), we have

$$\mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \succeq \frac{\mathbb{E}[K_t S S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]}{\sigma_{\max}(K_t^2)} = \frac{sK_t^2}{(d+m)\sigma_{\max}(K_t^2)} \succeq \frac{sI}{(d+m)\kappa(K_t^2)}, \quad (4.3)$$

where $\sigma_{\max}(K_t^2)$ denotes the largest singular value (which is the same as the largest eigenvalue in this case) of K_t^2 and $\kappa(K_t^2)$ denotes the condition number of K_t^2 (it is independent of t by Assumption 4.1). The first inequality is by the eigenvalue interlacing theorem, which leads to $\sigma_{\max}(S^T K_t^2 S) \leq \sigma_{\max}(K_t^2)$, and the second equality is by the sampling mechanism:

$$\mathbb{E}[SS^T] = \frac{1}{\binom{d+m}{s}} \sum_{\mathcal{A}\in\{\text{sets of } s \text{ indices}\}} I_{\mathcal{A}} = \frac{\binom{d+m-1}{s-1}}{\binom{d+m}{s}} I = \frac{s}{d+m} I.$$

Here, $I_{\mathcal{A}} \in \mathbb{R}^{(d+m)\times(d+m)}$ is a diagonal matrix with $[I_{\mathcal{A}}]_{i,i} = 1$ if the index $i \in \mathcal{A}$ and $[I_{\mathcal{A}}]_{i,i} = 0$ otherwise. The set $\mathcal{A}$ contains s distinct indices. We note that a recent study (Dereziński and Rebrova, 2024,(1.3)) showed a similar lower bound to (4.3) for Gaussian sketching. The authors improved the denominator from $(d+m)\sigma_{\max}^2(K_t)$ to $\|K_t\|_F^2$, though the latter still grows linearly with the problem dimension without additional spectral decay assumptions.

Assumption 4.3 directly leads to the following result.

Lemma 4.4 (Guarantees of sketching solvers) *Under Assumption 4.3, for all $t \geq 0$:*
(a): Let $\rho = 1 - \gamma_S$. We have $0 \leq \rho < 1$.
(b): $\mathbb{E}[\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t \mid \mathcal{F}_{t-2/3}] = -(I - \mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \mathcal{F}_{t-1}])^\tau \widetilde{\boldsymbol{z}}_t =: C_t \widetilde{\boldsymbol{z}}_t$, and $\|C_t\| \leq \rho^\tau$.
(c): $\mathbb{E}[\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\|^2 \mid \mathcal{F}_{t-2/3}] \leq \rho^\tau \|\widetilde{\boldsymbol{z}}_t\|^2$.

Remark 4.5 *We note that the linear convergence rate ρ of the expected error of the sketching solver depends on the lower bound $\gamma_S \in (0, 1]$ of the projection matrix, which is proportional to the sketching dimension s and inversely proportional to the problem dimension $d+m$, as shown in (4.3). The condition number $\kappa^2(K_t)$ is assumed to be uniformly bounded in our study.*

By the above relation, and given an error threshold δ, we have $\rho^\tau = (1-\gamma_S)^\tau \leq \delta \iff \tau \geq \log(1/\delta)/\log(1/\{1-\gamma_S\}) = O(1/\gamma_S) = O((d+m)/s)$. This implies that, to decay the expected error below a threshold, the number of sketching steps τ is proportional to the problem dimension $d+m$ and inversely proportional to the sketching dimension s. Certainly, a larger sketching dimension leads to a higher computational cost per step in (3.4). Using the Kaczmarz method as an example, the flops per step are $O((d+m)s^2)$ (dominated by the computation of $S^T K_t^2 S$). Thus, the total flops over τ steps are $O((d+m)^2 s)$, indicating that a smaller sketching dimension is generally preferable. That said, this analysis only reflects a worst-case scenario under the presumption that $\kappa^2(K_t)$ is uniformly bounded. The optimal choice of s is often case-by-case and depends on whether K_t exhibits a particular sparsity or eigenvalue decay structure (Dereziński and Rebrova, 2024).

4.2 Almost sure convergence

We now set the stage to show global almost sure convergence. The first result shows that the exact Newton direction $(\Delta \boldsymbol{x}_t, \Delta \boldsymbol{\lambda}_t)$ is a descent direction of $\mathcal{L}^t_{\mu,\nu} = \mathcal{L}_{\mu,\nu}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ if μ is sufficiently large and ν is sufficiently small.

Lemma 4.6 *Under Assumption 4.1, there exists a deterministic constant $\Upsilon_1 > 0$, depending only on $(\gamma_G, \gamma_{RH}, \Upsilon_B)$, such that*

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}^t_{\mu,\nu} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}^t_{\mu,\nu} \end{pmatrix}^T \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \leq -\frac{\nu}{\Upsilon_1} \left\{ \left\| \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_t \\ c_t \end{pmatrix} \right\|^2 \right\},$$

provided $\mu\nu \geq \Upsilon_1$ and $\nu \leq 1/\Upsilon_1$.

With Lemmas 4.4 and 4.6, we are able to show the following one-step recursion of $\mathcal{L}^t_{\mu,\nu}$.

Lemma 4.7 *Under Assumptions 4.1, 4.2(4.2a), 4.3, we suppose that the pair (μ, ν) satisfies the condition in Lemma 4.6 and τ satisfies $\rho^\tau \leq \nu/(\mu\Upsilon_1)$. Then, there exists a deterministic constant $\Upsilon_2 > 0$, depending on $(\gamma_G, \gamma_{RH}, \Upsilon_B, \Upsilon_L, \Upsilon_m)$, such that*

$$\mathbb{E}[\mathcal{L}^{t+1}_{\mu,\nu} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}^t_{\mu,\nu} - \frac{\nu}{2\Upsilon_1} \cdot \beta_t \|\nabla \mathcal{L}_t\|^2 + \Upsilon_2(\chi_t + \eta_t^2).$$

With Lemma 4.7, we can apply Robbins-Siegmund theorem (Robbins and Siegmund, 1971) to establish the convergence of the KKT residual $\|\nabla \mathcal{L}_t\|$.

Theorem 4.8 (Global convergence) *Consider Algorithm 1 under Assumptions 4.1, 4.2(4.2a), 4.3. Suppose we perform the sketching solver (3.4) for τ steps with $\tau \geq 4 \log \Upsilon_1 / \log\{1/(1-\gamma_S)\}$, where Υ_1 is from Lemma 4.6. Also, we let $\{\beta_t, \eta_t = \beta_t + \chi_t\}$ satisfy*

$$\sum_{t=0}^{\infty} \beta_t = \infty, \qquad \sum_{t=0}^{\infty} \beta_t^2 < \infty, \qquad \sum_{t=0}^{\infty} \chi_t < \infty. \tag{4.4}$$

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.

Theorem 4.8 indicates that all limiting points of the primal-dual iteration sequence $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ are stationary. Our global convergence guarantee aligns with the guarantee of deterministic SQP methods (Nocedal and Wright, 2006, Theorem 18.3), as well as the guarantee of recent stochastic SQP methods (Na et al., 2022a, 2023), despite the fact that Algorithm 1 possesses an additional source of randomness from the sketching solver at each step. The convergence of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is equivalent to the convergence of $\nabla \mathcal{L}_t$ if Problem (1.1) is convex. Furthermore, the results $\|\nabla \mathcal{L}_t\| \vee \|(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)\| \xrightarrow{a.s.} 0$ imply the existence of an attraction region around the local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$. Once $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ lies in the neighborhood, all subsequent iterates will stay in the neighborhood and $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \xrightarrow{a.s.} (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ (Bertsekas, 1982, Chapter 4.4).

Based on Theorem 4.8, we can immediately show the worst-case iteration complexity. Due to the online nature of the method, the iteration complexity is equivalent to the sample complexity, as we observe one sample in each iteration.

Corollary 4.9 *Consider Algorithm 1 under Assumptions 4.1, 4.2(4.2a), 4.3. Suppose τ satisfies the condition in Theorem 4.8, and let $\beta_t = (t+1)^{-a}$, $\chi_t = (t+1)^{-b}$ where $a \in (0,1)$ and $a < b$. Also, 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, $\mathcal{T}_\epsilon$ attains the minimum $O(\epsilon^{-4})$ with $a = 1/2$ and $b = 1$.

We should mention that a recent work Curtis et al. (2023) also showed an $O(\epsilon^{-4})$ iteration complexity with Newton systems being exactly solved. Our result matches theirs while allowing for the use of sketching solvers to inexactly solve Newton systems. We highlight that Corollary 4.9 is based on a *non-asymptotic* convergence rate of the averaged expected KKT residual $\frac{1}{t}\sum_{i=0}^{t-1} \mathbb{E}[\|\nabla \mathcal{L}_i\|]$. This non-asymptotic result is in contrast to our main inference analysis in Section 5, where the results hold asymptotically.

5. Statistical Inference via StoSQP

We perform online statistical inference for Problem (1.1) by leveraging StoSQP. To segue into inference analysis, we suppose in this section that the method converges to a local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ of (1.1); specifically, $G^\star = \nabla c^\star$ has full row rank and $\nabla^2_{\boldsymbol{x}} \mathcal{L}^\star$ is positive definite in the null space $\{\boldsymbol{x} \in \mathbb{R}^d : G^\star \boldsymbol{x} = \boldsymbol{0}\}$. These optimality conditions ensure that the Lagrangian Hessian $K^\star = \nabla^2 \mathcal{L}^\star$ is non-singular, as necessary for M-estimators in (1.3).

5.1 Iteration recursion

From a high-level view, our method generates a stochastic sequence

$$\begin{pmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star \end{pmatrix} = (1 - \bar{\alpha}_t) \begin{pmatrix} \boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \boldsymbol{\theta}^t_{\boldsymbol{x}} \\ \boldsymbol{\theta}^t_{\boldsymbol{\lambda}} \end{pmatrix} + \bar{\alpha}_t \begin{pmatrix} \boldsymbol{\delta}^t_{\boldsymbol{x}} \\ \boldsymbol{\delta}^t_{\boldsymbol{\lambda}} \end{pmatrix}, \tag{5.1}$$

where $(\boldsymbol{\theta}^t_{\boldsymbol{x}}, \boldsymbol{\theta}^t_{\boldsymbol{\lambda}})$ is a martingale difference with $\mathbb{E}[(\boldsymbol{\theta}^t_{\boldsymbol{x}}, \boldsymbol{\theta}^t_{\boldsymbol{\lambda}}) \mid \mathcal{F}_{t-1}] = \boldsymbol{0}$, and $(\boldsymbol{\delta}^t_{\boldsymbol{x}}, \boldsymbol{\delta}^t_{\boldsymbol{\lambda}})$ is the remaining error term. Compared to existing stochastic first- and second-order methods (Chen et al., 2020; Bercu et al., 2020; Duchi and Ruan, 2021; Davis et al., 2024; Boyer and Godichon-Baggioni, 2023), the randomness brought by the adaptivity and inexactness (AI) of the method affects all the terms in (5.1). This includes the random stepsize $\bar{\alpha}_t$, as well as the random approximation errors in $(\boldsymbol{\theta}^t_{\boldsymbol{x}}, \boldsymbol{\theta}^t_{\boldsymbol{\lambda}})$ and $(\boldsymbol{\delta}^t_{\boldsymbol{x}}, \boldsymbol{\delta}^t_{\boldsymbol{\lambda}})$ associated with the sketching solver.

We formalize the recursion (5.1) in the following lemma.

Lemma 5.1 *Let $\varphi_t = (\beta_t + \eta_t)/2$. The iteration sequence of Algorithm 1 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

$$\mathcal{I}_{1,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j (I + C^\star)\} \varphi_i \begin{pmatrix} \boldsymbol{\theta}^i_{\boldsymbol{x}} \\ \boldsymbol{\theta}^i_{\boldsymbol{\lambda}} \end{pmatrix}, \tag{5.2a}$$

$$\mathcal{I}_{2,t} = \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\} (\bar{\alpha}_i - \varphi_i) \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_i \\ \bar{\Delta}\boldsymbol{\lambda}_i \end{pmatrix}, \tag{5.2b}$$

$$\mathcal{I}_{3,t} = \prod_{i=0}^{t} \{I - \varphi_i(I + C^\star)\} \begin{pmatrix} \boldsymbol{x}_0 - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^\star \end{pmatrix} + \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j(I + C^\star)\} \varphi_i \begin{pmatrix} \boldsymbol{\delta}_{\boldsymbol{x}}^i \\ \boldsymbol{\delta}_{\boldsymbol{\lambda}}^i \end{pmatrix}, \tag{5.2c}$$

and

$$C^\star = -(I - \mathbb{E}[K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star])^\tau, \tag{5.3a}$$

$$\begin{pmatrix} \boldsymbol{\theta}_{\boldsymbol{x}}^i \\ \boldsymbol{\theta}_{\boldsymbol{\lambda}}^i \end{pmatrix} = -(I + C_i) K_i^{-1} \begin{pmatrix} \bar{g}_i - \nabla f_i \\ \mathbf{0} \end{pmatrix} + \left\{ \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_i \\ \bar{\Delta}\boldsymbol{\lambda}_i \end{pmatrix} - (I + C_i) \begin{pmatrix} \widetilde{\Delta}\boldsymbol{x}_i \\ \widetilde{\Delta}\boldsymbol{\lambda}_i \end{pmatrix} \right\}, \tag{5.3b}$$

$$\begin{pmatrix} \boldsymbol{\delta}_{\boldsymbol{x}}^i \\ \boldsymbol{\delta}_{\boldsymbol{\lambda}}^i \end{pmatrix} = -(I + C_i) \left\{ (K^\star)^{-1} \begin{pmatrix} \boldsymbol{\psi}_{\boldsymbol{x}}^i \\ \boldsymbol{\psi}_{\boldsymbol{\lambda}}^i \end{pmatrix} + \left\{ K_i^{-1} - (K^\star)^{-1} \right\} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_i \\ c_i \end{pmatrix} \right\} - (C_i - C^\star) \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star \end{pmatrix} \tag{5.3c}$$

$$\begin{pmatrix} \boldsymbol{\psi}_{\boldsymbol{x}}^i \\ \boldsymbol{\psi}_{\boldsymbol{\lambda}}^i \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_i \\ c_i \end{pmatrix} - K^\star \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star \end{pmatrix}. \tag{5.3d}$$

Under Assumptions 4.2, 4.3, $\boldsymbol{\theta}^i = (\boldsymbol{\theta}_{\boldsymbol{x}}^i, \boldsymbol{\theta}_{\boldsymbol{\lambda}}^i)$ *is a martingale difference with* $\mathbb{E}[\boldsymbol{\theta}^i \mid \mathcal{F}_{i-1}] = \mathbf{0}$.

From Lemma 5.1, we observe that the recursion consists of three terms. $\mathcal{I}_{1,t}$ is a martingale that accounts for the randomness of sampling ξ_t to estimate ∇f_t and the randomness of sketching $\{S_{t,j}\}_j$ to solve QP (3.1). $\mathcal{I}_{2,t}$ captures the randomness of the stepsize $\bar{\alpha}_t$. $\mathcal{I}_{3,t}$ contains all the remainder terms. The asymptotic analysis of each term is provided in Appendix E.4.

Next, we establish a continuity property for the projection matrix $K_t S(S^T K_t^2 S)^\dagger S^T K_t$, a critical quantity of the sketching solver appeared in C_t and $C^\star$ (cf. Lemma 4.4(b) and (5.3a)).

Lemma 5.2 *Suppose* $K_t, K^\star \in \mathbb{R}^{(d+m)\times(d+m)}$ *are non-singular. For any* $S \in \mathbb{R}^{(d+m)\times s}$,

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

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

Lemma 5.2 indicates that the difference between the projection matrices $K_t S(S^T K_t^2 S)^\dagger S^T K_t$ and $K^\star S(S^T(K^\star)^2 S)^\dagger S^T K^\star$ is proportional to the difference between the Hessian matrices K_t and $K^\star$, with a random factor scaling with the condition number of the sketching matrix S. In practice, using sketching vectors ($s = 1$) can reduce computational cost and result in a unit condition number $\|S\|\|S^\dagger\| = 1$.

Lemma 5.2 leads to the following condition to ensure the convergence of C_t, which is the expectation of the product of projection matrices.

Assumption 5.3 *We assume that* S *satisfies* $\mathbb{E}[\|S\|\|S^\dagger\|] \le \Upsilon_S$ *for a constant* $\Upsilon_S > 0$.

Corollary 5.4 *Under Assumption 5.3,* $\|C_t - C^\star\| \le 2\tau\Upsilon_S\|K_t - K^\star\|/\sigma_{\min}(K^\star)$.

5.2 Asymptotic rate and normality

We are now ready to state inference theory. Let $S_1, \ldots, S_\tau \overset{iid}{\sim} S$, and define a random matrix:

$$\widetilde{C}^\star = -\prod_{j=1}^{\tau}(I - K^\star S_j (S_j^T (K^\star)^2 S_j)^\dagger S_j^T K^\star). \tag{5.4}$$

Clearly, $\mathbb{E}[\widetilde{C}^\star] = C^\star$. Also, define the sandwich matrix that appears as the limiting covariance of M-estimators in (1.3):

$$\Omega^\star = (K^\star)^{-1}\mathrm{cov}(\bar{\nabla}\mathcal{L}^\star)(K^\star)^{-1} = \begin{pmatrix} \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (G^\star)^T \\ G^\star & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{cov}(\nabla F(\boldsymbol{x}^\star;\xi)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (G^\star)^T \\ G^\star & \mathbf{0} \end{pmatrix}^{-1} \tag{5.5}$$

To allow for general stepsize control sequences $\{\beta_t, \chi_t\}_t$, we define three quantities:

$$\beta \coloneqq \lim_{t\to\infty} t\left(1 - \frac{\beta_{t-1}}{\beta_t}\right), \qquad \widetilde{\beta} \coloneqq \lim_{t\to\infty} t\beta_t, \qquad \chi \coloneqq \lim_{t\to\infty} t\left(1 - \frac{\chi_{t-1}}{\chi_t}\right).$$

The polynomial sequences $1/t^\omega$ are specialized in Lemma 5.12. For sake of understanding, we here mention that if $\beta_t = O(1/t^\omega)$ for any $\omega > 0$, then we simply have $\beta = -\omega$.

Theorem 5.5 (Local convergence rate) *Under Assumptions 4.1, 4.2(4.2a, 4.2e), 4.3, we suppose $\{\beta_t, \chi_t\}_t$ satisfy*

$$\chi < \beta < 0, \qquad \widetilde{\beta} \in (0, \infty], \qquad 1.5(1-\rho^\tau) + \beta/\widetilde{\beta} > 0. \tag{5.6}$$

Then, for any $\upsilon > 0$ and any constant $p \in (0, 1]$ such that $(1-\rho^\tau)+p(\chi-0.5\beta)/\widetilde{\beta} > 0$, we have (if $p < 1$, the second $O(\cdot)$ in the following results can be strengthened to $o(\cdot)$)

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = o\big(\sqrt{\beta_t\{\log(1/\beta_t)\}^{1+\upsilon}}\,\big) + O(\chi_t^p/\beta_t^p) \quad a.s.$$

Furthermore, if (4.2a) is strengthened to (4.2b), then

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = O\big(\sqrt{\beta_t \log(1/\beta_t)}\,\big) + O(\chi_t^p/\beta_t^p) \quad a.s.$$

The asymptotic convergence rate consists of two terms. The first term $o(\sqrt{\beta_t\{\log(1/\beta_t)\}^{1+\upsilon}})$ comes from the strong law of large number for the martingale $\mathcal{I}_{1,t}$, which can be strengthened to $O(\sqrt{\beta_t \log(1/\beta_t)})$ if the stochastic gradient estimate $\bar{g}_t$ has a bounded moment of order higher than two (Duflo, 1997, Theorem 1.3.15). That is, the bounded 3rd moment in (4.2b) can be directly replaced by a bounded $2+\delta$ moment. The second term $O(\chi_t^p/\beta_t^p)$ comes from $\mathcal{I}_{2,t}$, which characterizes the adaptivity of the random stepsize $\bar{\alpha}_t$. This term is suppressed if we degrade the method to a non-adaptive one ($\chi_t = 0$).

We will investigate the condition (5.6) in Lemma 5.12 and demonstrate that it is week enough to allow for different setups of sequences $\{\beta_t, \chi_t\}_t$. Most importantly, the condition (5.6) covers the setup of $\beta_t = 1/t$, corresponding to $\beta = -1$ and $\widetilde{\beta} = 1$, which leads to the optimal central limit theorem rate (cf. Theorem 5.6). In fact, (5.6) reveals a relationship between the inexactness of the sketching solver (i.e., the parameter τ) and the setup of the stepsize. When $\widetilde{\beta} = \infty$ (e.g., $\beta_t = 1/t^\omega$ for $\omega < 1$), the third condition in (5.6) holds trivially.

Furthermore, we note that $(1-\rho^\tau)+p(\chi-0.5\beta)/\widetilde{\beta} > 0$ is always satisfied for small p (specifically, $p = 1$ if $\widetilde{\beta} = \infty$). Thus, our condition on the adaptivity gap χ_t is simply $\chi_t = o(\beta_t)$ (i.e., $\chi < \beta$). We have to strengthen this condition to $\chi_t = o(\beta_t^{1.5})$ to enable inference analysis in Theorem 5.6. Even that, it's worth mentioning that the methods proposed in Berahas et al. (2021, 2023); Curtis et al. (2021) all set $\chi_t = O(\beta_t^2)$ (i.e., $\chi \leq 2\beta$). Our analysis relaxes these designs to allow for a larger χ_t, resulting in a wider interval (3.6) for stepsize adaptivity.

Theorem 5.6 (Asymptotic normality) *Under Assumptions 4.1, 4.2(4.2b, 4.2e), 4.3, 5.3, we strengthen $\chi < \beta$ in the condition (5.6) to $\chi < 1.5\beta$. Then, we have*

$$\sqrt{1/\bar{\alpha}_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Xi^\star), \tag{5.7}$$

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

$$\big(\{1+\beta/(2\widetilde{\beta})\}I + C^\star\big)\Xi^\star + \Xi^\star\big(\{1+\beta/(2\widetilde{\beta})\}I + C^\star\big) = \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T]. \tag{5.8}$$

Furthermore, let us define $q = 1$ if $1-\rho^\tau+(\chi-\beta)/\widetilde{\beta} > 0$, otherwise $q = \{(1-\rho^\tau)\widetilde{\beta}+\epsilon\beta\}/(\beta-\chi) \in (0,1)$ for any $\epsilon \in (0, 1/6]$. Then, for any $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T\Xi^\star\boldsymbol{w} \neq 0$,

$$\sup_{z\in\mathbb{R}} \left| P\left(\frac{\sqrt{1/\bar{\alpha}_t} \cdot \boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}} \leq z \right) - P\left(\mathcal{N}(0,1) \leq z\right) \right| = o(\beta_t^{1/6}\log(1/\beta_t)) + O(\chi_t^q/\beta_t^{q+0.5}), \tag{5.9}$$

where $O(\cdot)$ can be strengthened to $o(\cdot)$ if $q < 1$.

We mention that, to our knowledge, (5.7) provides the first primal-dual asymptotic normality result for a constrained online estimation procedure, while existing works in Duchi and Ruan (2021); Davis et al. (2024) have only established primal asymptotic normality. Although the uncertainty quantification of primal variables $\boldsymbol{x}^\star$ has a natural meaning as they represent model parameters, the uncertainty quantification of dual variables $\boldsymbol{\lambda}^\star$ is also significant in two ways beyond the technical interest.

(a) The dual variables are widely used as optimality certificates in algorithmic designs. In particular, the normality of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ enables the construction of a confidence region for the gradient vector $\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$; it suggests that the optimality residual, $\|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\|^2$, exhibits a (generalized) chi-squared limiting distribution. See (Jones, 1983, Section 4.1) and (Shapiro et al., 2014, Section 5.6.2) for analogous constructions. By checking whether the region contains $\mathbf{0}$, we can determine whether we have sufficiently achieved the optimality condition at the given significance level and terminate the algorithm accordingly.

(b) Methods for solving inequality-constrained problems often involve equality-constrained subproblems (Na et al., 2023), and constructing confidence intervals for dual variables associated with inequality constraints is crucial for active-set identification under uncertainty. This paper serves as a first step toward that goal. Specifically, let $\boldsymbol{\lambda}^\star \in \mathbb{R}^m$ be the dual solution associated with the constraints $c(\boldsymbol{x}) \leq \mathbf{0}$. By the strict complementarity condition (Davis et al. (2024), Example 2.1), we know $\boldsymbol{\lambda}_i^\star > 0 \Leftrightarrow c_i(\boldsymbol{x}^\star) = 0$ and $\boldsymbol{\lambda}_i^\star = 0 \Leftrightarrow c_i(\boldsymbol{x}^\star) < 0$. Thus, if the confidence interval for $\boldsymbol{\lambda}_i^\star$ contains 0, it suggests that the i-th constraint is not identified as active at the given significance level, and vice versa.

The explicit form of the solution to (5.8) is given by

$$\Xi^\star = U\big(\Theta \circ U^T\mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T]\,U\big)U^T \quad \text{with} \quad [\Theta]_{k,l} = 1/(\sigma_k+\sigma_l+\beta/\widetilde{\beta}), \tag{5.10}$$

where $I+C^\star = U\Sigma U^T$ with $\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_{d+m})$ is the eigenvalue decomposition of $I+C^\star$, and $\circ$ denotes the matrix Hadamard product.

From Theorem 5.6, we see that the limiting covariance $\Xi^\star$ depends on the sandwich matrix $\Omega^\star$ in (5.5), which is the same as the optimal one for M-estimators in (1.3), but it also depends on the underlying sketching distribution. The sketching matrices affect both the left- and right-hand sides of the Lyapunov equation (5.8). If we degrade the randomized sketching solver to an exact QP solver, then $\tau = \infty$, $C^\star = \widetilde{C}^\star = \mathbf{0}$, and

$$\Xi^\star = \frac{\Omega^\star}{2+\beta/\widetilde{\beta}}. \tag{5.11}$$

We present the following corollary to further connect Theorem 5.6 with the asymptotic minimax optimality of constrained estimation problems established in (Duchi and Ruan, 2021, Theorem 1) and (Davis et al., 2024, Theorem 3.2).

Corollary 5.7 *Let $\beta_t = 1/t$, $\chi_t = o(\beta_t^{1.5})$, and τ such that $\rho^\tau < 1/3$. Then,*
(a): Exact QP solver: $\Xi^\star = \Omega^\star$,
(b): Sketching solver: $\Xi^\star \succeq \Omega^\star$ but $\|\Xi^\star - \Omega^\star\| \le 3\rho^\tau\|\Omega^\star\|$.

By Corollary 5.7, the limiting covariance of StoSQP obtained by suppressing the sketching solver matches the asymptotic minimax optimum $\Omega^\star$ that is achieved by offline M-estimators and online projection-based estimators (Duchi and Ruan, 2021; Davis et al., 2024). To the best of our knowledge, our estimator based on StoSQP is the first online estimator that does not rely on projection operators; we instead replace projection by employing a series of linear-quadratic approximation of nonlinear problems. On the other hand, when employing the sketching solver to inexactly solve QPs, the limiting covariance of our method exceeds the optimum $\Omega^\star$, indicating that the sketching solver (to address the computation concern of second-order methods) indeed compromises optimality. Fortunately, this compromise is marginal since the distance between $\Xi^\star$ and $\Omega^\star$ decays exponentially fast with the number of iterations τ performed for the sketching solver.

The Berry-Esseen bound in (5.9) consists of two terms. The first term is due to the random sample and random sketching, while the second term is due to the random stepsize. Our choice of q always guarantees that $\chi_t^q/\beta_t^{q+0.5} = o(1)$. We note that $q = 1$ when $\widetilde{\beta} = \infty$ (e.g., $\beta_t = 1/t^\omega$ for $\omega < 1$).

Remark 5.8 *Since existing studies in Duchi and Ruan (2021); Davis et al. (2024) heavily used projection notation and only established the normality of the primal estimator, we further elucidate in this remark the connection between our joint covariance $\Omega^\star$ (i.e., StoSQP with exact QP solver) and the covariance in (Davis et al., 2024, Corollary 5.2) and (Duchi and Ruan, 2021, Theorem 5).*

Recall that $G^\star = \nabla c(\boldsymbol{x}^\star) \in \mathbb{R}^{m\times d}$ is the constraints Jacobian. Let $Z^\star \in \mathbb{R}^{d\times(d-m)}$ be a matrix whose columns are orthonormal and form the bases of $ker(G^\star)$. Then, using the relation

$G^{\star T}(G^\star G^{\star T})^{-1}G^\star + Z^\star Z^{\star T} = I_d$, *we can verify that*

$$\begin{pmatrix} \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & G^{\star T} \\ G^\star & \mathbf{0}_m \end{pmatrix}^{-1} = \begin{pmatrix} A_1 & A_2^T \\ A_2 & A_3 \end{pmatrix} \tag{5.12}$$

where

$$A_1 = Z^\star(Z^{\star T}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star Z^\star)^{-1}Z^{\star T}, \quad A_2 = (G^\star G^{\star T})^{-1}G^\star(I - \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star A_1),$$
$$A_3 = (G^\star G^{\star T})^{-1}G^\star(\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star A_1 \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star - \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star)(G^\star G^{\star T})^{-1}G^\star.$$

Note that all above matrix inverses are well-defined under the conditions of (Davis et al., 2024, Example 2.1 and Section 5.1) (or, equivalently, under our conditions). Plugging the above display into (5.5)*, we see that the marginal covariance of* $\boldsymbol{x}$ *is*

$$\Omega^\star_{\boldsymbol{x},\boldsymbol{x}} = A_1\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_1. \tag{5.13}$$

Furthermore, we note that $A_1 = (Z^\star Z^{\star T}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star Z^\star Z^{\star T})^\dagger$ *by verifying the definition of the Moore-Penrose pseudoinverse. Since* $Z^\star Z^{\star T} = Proj_{ker(G^\star)}$ *and* $ker(G^\star)$ *is the tangent space of the manifold* $c(\boldsymbol{x})$ *at* $\boldsymbol{x}^\star$*, we see* (5.13) *matches the result in (Davis et al., 2024, Corollary 5.2).*

Remark 5.9 *We note that* $\Omega^\star$ *is clearly singular. We investigate in this remark the support subspace of the limiting distribution. Let us write out the expression of* $\Omega^\star$ *using the notation in Remark 5.8. We have*

$$\Omega^\star = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} cov(\nabla F(\boldsymbol{x}^\star;\xi)) \begin{pmatrix} A_1 & A_2^T \end{pmatrix} = \begin{pmatrix} A_1\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_1 & A_1\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_2^T \\ A_2\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_1 & A_2\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_2^T \end{pmatrix}.$$

Note that $rank(A_1) = rank(Z^\star(Z^{\star T}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star Z^\star)^{-1/2}) = d-m$*, which implies* $rank(A_2) = m$ *(since the first block-column matrix* $[A_1; A_2]$ *in* (5.12) *has rank* d*). In the following presentation, we suppose* $rank(cov(\nabla F(\boldsymbol{x}^\star;\xi))) = d$*.*

• ***Primal covariance.*** *Since*

$$rank(\Omega^\star_{\boldsymbol{x},\boldsymbol{x}}) = rank(A_1\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_1^T) = rank(A_1\, cov(\nabla F(\boldsymbol{x}^\star;\xi))^{1/2}) = rank(A_1) = d-m,$$

the support subspace of $\Omega^\star_{\boldsymbol{x},\boldsymbol{x}}$ *has* $d-m$ *dimensions. Specifically, we decompose* $\mathbb{R}^d$ *into* $\mathbb{R}^d = \ker(G^\star) \oplus span(G^{\star T})$*, where the tangent space* $\ker(G^\star)$ *is a* $(d-m)$*-dimensional subspace of* $\mathbb{R}^d$ *and the normal space* $span(G^{\star T})$ *is an* m*-dimensional subspace of* $\mathbb{R}^d$*. Then, by the definition of* A_1*, we know that* $\Omega^\star_{\boldsymbol{x},\boldsymbol{x}}$ *has support in the tangent space* $\ker(G^\star)$ *and vanishes in the normal space* $span(G^{\star T})$*.*

• ***Dual covariance.*** *Since*

$$rank(\Omega^\star_{\boldsymbol{\lambda},\boldsymbol{\lambda}}) = rank(A_2\, cov(\nabla F(\boldsymbol{x}^\star;\xi))A_2^T) = rank(A_2\, cov(\nabla F(\boldsymbol{x}^\star;\xi))^{1/2}) = rank(A_2) = m,$$

we know $\Omega^\star_{\boldsymbol{\lambda},\boldsymbol{\lambda}}$ *is non-degenerate along all directions in* $\mathbb{R}^m$ *and has full support over* $\mathbb{R}^m$*.*

• ***Joint primal-dual covariance.*** *Since* $rank(\Omega^\star) = d$*, we know the support subspace of* $\Omega^\star$ *has* d *dimensions. Let us decompose* $\mathbb{R}^{d+m}$ *into* $\mathbb{R}^{d+m} = (ker(G^\star) \otimes \mathbb{R}^m) \oplus (span(G^{\star T}) \otimes \mathbf{0}_m)$*.*

More clearly, ker$(G^\star)\otimes\mathbb{R}^m$ is a d-dimensional subspace of $\mathbb{R}^{d+m}$ and span$(G^{\star T})\otimes\mathbf{0}_m$ is an m-dimensional subspace of $\mathbb{R}^{d+m}$; there bases can be expressed as:

$$\begin{pmatrix} Z^\star & \mathbf{0}_{d\times m} \\ \mathbf{0}_{m\times(d-m)} & I_m \end{pmatrix} \in \mathbb{R}^{(d+m)\times d} \qquad \oplus \qquad \begin{pmatrix} G^{\star T} \\ \mathbf{0}_m \end{pmatrix} \in \mathbb{R}^{(d+m)\times m}.$$

Then, the joint covariance $\Omega^\star$ has support in the subspace ker$(G^\star)\otimes\mathbb{R}^m$ and vanishes in the subspace span$(G^{\star T})\otimes\mathbf{0}_m$.

5.3 An estimator of the covariance matrix

We analyze a plug-in covariance matrix estimator. The sketch-related quantities in (5.8), $C^\star = \mathbb{E}[\widetilde{C}^\star]$ and $P^\star = \mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T]$, can be estimated by computing (note $\widetilde{C}^\star$ is defined in (5.4) and Ω_t is defined in (5.14))

$$\widetilde{C}_t \coloneqq -\prod_{j=0}^{\tau-1}\left(I - K_t S_{t,j}(S_{t,j}^T K_t^2 S_{t,j})^\dagger S_{t,j}^T K_t\right),$$

$$\widehat{C}_t \coloneqq \frac{1}{t}\sum_{i=0}^{t-1}\widetilde{C}_i, \qquad \text{and} \qquad \widehat{P}_t \coloneqq \frac{1}{t}\sum_{i=0}^{t-1}(I+\widetilde{C}_i)\Omega_i(I+\widetilde{C}_i)^T.$$

Since solving Lyapunov equation requires additional effort and $\widehat{P}_t$ requires matrix inverse even if we do not perform inference at the current step, in what follows, we are motivated by Corollary 5.7 and simply neglect the sketch-related quantities in (5.8) by estimating (5.11) instead. We demonstrate that such negligence only leads to an $O(\rho^\tau)$ error term, which is generally small even for a moderate τ. Specifically, our estimator of $\Xi^\star$ is defined as:

$$\Omega_t = K_t^{-1}\begin{pmatrix} sample\_cov(\{\bar{g}_i\}_{i=0}^{t-1}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} K_t^{-1} \quad \text{and} \quad \Xi_t = \frac{\Omega_t}{2+\beta/\widetilde{\beta}}, \tag{5.14}$$

where $sample\_cov(\{\bar{g}_i\}_{i=0}^{t-1}) = \frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)^T$ is the sample covariance.

Theorem 5.10 *Consider* (5.14) *under the conditions of Theorem 5.6 with (4.2c). For any $\upsilon > 0$,*

$$\|\Xi_t - \Xi^\star\| = O\big(\sqrt{\beta_t\log(1/\beta_t)}\,\big) + o\big(\sqrt{(\log t)^{1+\upsilon}/t}\big) + O(\rho^\tau) \quad a.s.$$

Furthermore, for any $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T\Xi_t\boldsymbol{w} \neq 0$,

$$\sup_{z\in\mathbb{R}}\left|P\left(\frac{\sqrt{1/\bar{\alpha}_t}\cdot\boldsymbol{w}^T(\boldsymbol{x}_t-\boldsymbol{x}^\star,\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi_t\boldsymbol{w}}}\le z\right) - P\left(\mathcal{N}(0,1)\le z\right)\right| = o(\beta_t^{1/6}\log(1/\beta_t)) + O(\chi_t^q/\beta_t^{q+0.5}) + O(\rho^\tau),$$

where $q \in (0,1]$ is defined in Theorem 5.6.

The second term $o\big(\sqrt{(\log t)^{1+\upsilon}/t}\big)$ can be absorbed into the first term $O\big(\sqrt{\beta_t\log(1/\beta_t)}\,\big)$ if $\beta > -1$, e.g., $\beta_t = 1/t^\omega$ for $\omega \in (0,1)$. We require condition (4.2c) for the estimation of the limiting covariance, which ensures the convergence of the sample covariance of $\{\bar{g}_i\}_{i=0}^{t-1}$. The bounded 4-th moment of $\bar{g}_t$ is also standard in the literature, as required for analyzing different covariance estimators for SGD methods (Chen et al., 2020; Zhu et al., 2021) .

Remark 5.11 (Discussion on control sequences $\{\beta_t, \chi_t = \eta_t - \beta_t\}$) *We note that the normality in* (5.7) *utilizes the adaptive stepsize $\bar{\alpha}_t$, which nevertheless is controlled by the sequences $\{\beta_t, \chi_t\}$. We discuss their conditions in this remark.*

The global convergence requires (4.4) (Theorem 4.8); the local convergence requires (5.6) (Theorem 5.5); and the inference additionally requires $\chi < 1.5\beta$ (Theorem 5.6). In fact, by Raabe's test, (4.4) also relates to the quantities β and χ: (4.4) holds if $-1 \leq \beta < -0.5$ and $\chi < -1$. We now specialize β_t and χ_t to be polynomial in t, and demonstrate that the conditions can be easily satisfied.

Lemma 5.12 *Suppose $\beta_t = c_1/t^{c_2}$ and $\chi_t = \beta_t^{c_3}$. Then,*

(a): (4.4) holds ***if*** *$c_1 > 0$, $c_2 \in (0.5, 1]$, $c_3 > \frac{1}{c_2} \Longrightarrow$ global convergence.*

(b): (4.4) and (5.6) hold ***if*** *additionally $c_1 > \frac{1}{1.5(1-\rho^\tau)}$ when $c_2 = 1 \Longrightarrow$ local convergence.*

(c): (4.4) and (5.6) hold with $\chi < 1.5\beta$ ***if*** *additionally $c_3 > 1.5 \vee \frac{1}{c_2} \Longrightarrow$ asymptotic inference.*

The proof of the above lemma is immediate by noting that $\beta = -c_2$, $\chi = -c_2 c_3$, and $\widetilde{\beta} = c_1$ if $c_2 = 1$ and ∞ if $c_2 < 1$. Thus, we omit it.

6. Numerical Experiments

We provide experimental results in this section. We apply `AI-StoSQP` to both benchmark constrained nonlinear optimization problems in CUTEst set (Gould et al., 2014) and to linearly/ nonlinearly constrained regression problems. For regression problems, we explore both squared loss and logistic loss. For conciseness, some of results are deferred to Appendix F.

6.1 Benchmark constrained problems

The CUTEst test set collects a number of nonlinear optimization problems with and without constraints. We implement eight equality-constrained problems: `MARATOS`, `ORTHREGB`, `HS7`, `HS48`, `HS78`, `BT9`, `GENHS28`, `HS39`. The solution of each problem is solved by IPOPT solver (Wächter and Biegler, 2006) with the initialization specified by the CUTEst package. Note that the benchmark problems may not have unique solutions; however, we observed that by initializing at the same point, our StoSQP method consistently converges to the same solution as IPOPT, which is also a widely used (deterministic) SQP-based solver.

For our method, we perform 10^5 iterations and, at each step, we perform $\tau = 40$ randomized Kaczmarz steps to approximately solve QPs. Given the iterate $\boldsymbol{x}_t$, we generate $\bar{g}_t \sim \mathcal{N}(\nabla f_t, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$, where $\mathbf{1} \in \mathbb{R}^d$ is an all-one vector. We also generate the (i,j) and (j,i) entries of $\bar{H}_t$ from $\mathcal{N}((\nabla^2 f_t)_{i,j}, \sigma^2)$. We vary $\sigma^2 \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$ and let $\beta_t = 1/t^{0.501}$ (power slightly larger than 0.5) and $\chi_t = \beta_t^2$. We randomly choose $\bar{\alpha}_t \sim \text{Uniform}([\beta_t, \eta_t])$ with $\eta_t = \beta_t + \chi_t$. For each problem, we aim to perform inference for each individual variable $\{\boldsymbol{x}_i^\star\}_{i=1}^d$ by setting the nominal coverage probability to 95%. Note from Remark 5.9 that $\Omega^\star_{\boldsymbol{x},\boldsymbol{x}}$ may vanish along the direction in the normal space $\text{span}(G^{\star T})$. Thus, we consider inferring only those $\boldsymbol{x}_i^\star$ such that the canonical basis $\boldsymbol{e}_i \notin \text{span}(G^{\star T})$. The confidence intervals are constructed by estimating the limiting covariance using Theorem 5.10. The performance of the method is measured by the mean absolute error (MAE) $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$, the average coverage rate (Avg Cov) of the

confidence intervals among individuals $\boldsymbol{x}_i^\star$, the average length (Avg Len) of the confidence intervals, and the computational flops per iteration. We repeat the experiments 200 times when computing the coverage rate.

The results are summarized in Table 1. From Table 1, we have the following observations.

(a) In terms of MAE, our method achieves reasonably small MAE values across all problems. As the noise level σ^2 for the objective gradient and Hessian estimates gradually increases, the MAE also increases. This aligns with our intuition, as noisier estimates of the objective quantities will require more samples, i.e., longer iterations, to ensure the accuracy of the estimate $\boldsymbol{x}_t$.

(b) In terms of coverage rate, we observe for the majority of cases that our constructed confidence intervals cover the true solution with probability of at least 95%, and the coverage rate is robust to the sampling variance σ^2. There are scenarios, such as `HS78` and `ORTHREGB` ($\sigma^2 = 10^{-4}$), where our confidence intervals may exhibit over- or under-coverage. This phenomenon can be attributed to two factors: (i) our limiting covariance estimate has an $O(\rho^\tau)$ bias due to sketching techniques, which may either inflate or deflate the estimated variance for each individual variable, and (ii) we run only a limited number of steps (i.e., the standardized sequence may have not yet reached a stationary stage), given the problem's scale and the challenges inherent in the online inference task. Nevertheless, our observation suggests that neglecting the sketching randomness in the estimation of the limiting covariance does not obviously deteriorate the coverage rate. However, using (sparse) sketching vectors to solve QPs as in (3.4) is computationally more efficient than exact second-order methods.

(c) In terms of confidence intervals' length, we see that the average length gradually increases as σ^2 increases. When σ^2 increases from 10^{-4} to 1, the length increases from 10^{-4} to 10^{-2}. This outcome is expected, as the asymptotic covariance $\Xi^\star$ depends on $\mathrm{cov}(\nabla f(\boldsymbol{x}^\star;\xi))$ in (5.8).

(d) In terms of computational flops per iteration, it is uniform over different σ^2 and independent runs. This quantity basically reflects the problem scale.

We testify the almost sure convergence rate of our method (Theorem 5.5) in Appendix F.1.

6.2 Constrained regression problems

We implement our method on constrained regression problems, considering both linear regression and logistic regression (see Example 1). We also allow for either linear constraints $A\boldsymbol{x} = \boldsymbol{d}$ or nonlinear constraints $\|\boldsymbol{x}\|^2 = b$. Therefore, there are four cases in total. We compare our online inference method with offline constrained M-estimation. For fair comparisons, we solve M-estimation problems, which are constrained finite-sum problems, using ℓ_1-penalized SQP methods with backtracking line search to select proper stepsizes (Nocedal and Wright, 2006). While offline M-estimators enjoy asymptotic minimax optimal performance with the least covariance (as introduced in (1.3)), the proposed online StoSQP method is promising due to its lower per-iteration computational complexity. (Our method also achieves optimal performance under appropriate setups; cf. Corollary 5.7.)

To be specific, for each case (regression model + constraint type), we vary the parameter dimension $d \in \{5, 20, 40, 60\}$, and the true solution $\boldsymbol{x}^\star$ is linearly spaced between 0 and 1. For each d, our method randomly samples a covariate $\xi_{\boldsymbol{a}} \sim \mathcal{N}(\boldsymbol{0}, 5I + \Sigma_a)$ at each step, with three different choices of Σ_a (also considered in Chen et al. (2020)). (i) Identity: $\Sigma_a = I$. (ii) Toeplitz: $[\Sigma_a]_{i,j} = r^{|i-j|}$. (iii) Equi-correlation: $[\Sigma_a]_{i,j} = r$ for $i \neq j$ and $[\Sigma_a]_{i,i} = 1$. For linear constraints, we let $A \in \mathbb{R}^{m\times d}$ with $m = \lceil\sqrt{d}\rceil$ and entries being independently generated from stan-

Prob	σ^2	MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
MARATOS	10^{-4}	0.04(0.03)	96.00	0.11(<0.01)	137.00
	10^{-2}	0.41(0.31)	95.00	1.10(0.01)	
	10^{-1}	1.42(1.07)	93.50	3.48(0.09)	
	1	4.44(3.29)	95.50	10.96(0.74)	
ORTHREGB	10^{-4}	1.76(1.17)	88.90	27.64(4.80)	3867.40
	10^{-2}	4.44(3.51)	96.42	58.02(12.64)	
	10^{-1}	9.90(17.99)	95.84	40.80(9.06)	
	1	38.19(44.51)	96.72	24.36(32.92)	
HS7	10^{-4}	0.02(0.01)	93.00	0.03(<0.01)	137.00
	10^{-2}	0.14(0.12)	92.50	0.35(<0.01)	
	10^{-1}	0.45(0.34)	95.00	1.10(0.01)	
	1	1.28(0.98)	97.00	3.48(0.09)	
HS48	10^{-4}	0.03(0.01)	94.50	0.02(0.01)	379.00
	10^{-2}	0.25(0.11)	95.70	0.24(0.11)	
	10^{-1}	0.85(0.38)	94.00	0.76(0.35)	
	1	2.51(1.25)	97.50	2.41(1.12)	
HS78	10^{-4}	0.01(<0.01)	97.80	0.02(<0.01)	434.01
	10^{-2}	0.15(0.07)	99.10	0.17(0.04)	
	10^{-1}	0.51(0.27)	96.80	0.55(0.12)	
	1	1.41(0.69)	99.10	1.74(0.37)	
BT9	10^{-4}	0.06(0.03)	93.00	0.05(<0.01)	308.01
	10^{-2}	0.55(0.29)	96.25	0.55(0.01)	
	10^{-1}	1.91(0.92)	95.50	1.76(0.07)	
	1	24.54(1.07)	94.50	6.58(5.57)	
GENHS28	10^{-4}	0.03(0.01)	97.75	0.04(0.04)	1244.07
	10^{-2}	0.26(0.09)	98.33	0.42(0.37)	
	10^{-1}	0.76(0.27)	98.17	1.34(1.18)	
	1	2.44(0.94)	97.83	4.22(3.73)	
HS39	10^{-4}	0.06(0.03)	93.00	0.05(<0.01)	308.01
	10^{-2}	0.55(0.29)	96.25	0.55(0.01)	
	10^{-1}	1.91(0.92)	95.50	1.76(0.07)	
	1	24.54(1.07)	94.50	6.58(5.57)	

Table 1: *Results of eight benchmark CUTEst problems. We measure performance using the mean absolute error (MAE), the average coverage rate (Ave Cov) and the average length (Ave Len) of the confidence intervals, and the computational flops per iteration. Standard errors are reported in the bracket for MAE and Ave Len. We do not report standard errors for Ave Cov as they are meaningless for a 0-1 indicator vector; the standard error is given by* $\sqrt{r(1-r)}$ *with* r *being the coverage rate. The flops/iter is uniform over different noise level* σ^2 *and different runs (i.e., the standard error is 0).*

dard normal distribution. For logistic models, we regularize the loss by a quadratic penalty with unit parameter. Following Section 6.1, we perform inference for each individual variable $\{\boldsymbol{x}_i^\star\}_{i=1}^d$

by setting the nominal coverage probability to 95%. We also follow Section 6.1 to implement our method, including the number of iterations, the setup of the stepsize, and the number of sketching steps. In contrast to online methods, the offline M-estimation method generates all 10^5 samples at once and uses those fixed samples to compute the estimator. The covariance matrix is also estimated by a plug-in estimator. We report the results only for $r = 0.5$ for Toeplitz and $r = 0.2$ for Equi-correlation. The comprehensive comparisons between inexact and exact methods with varying d, r and τ are reported in Section F.2.

We summarize the comparison results in Tables 2 and 3, including the mean absolute error (MAE) $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$, the average coverage rate (Avg Cov) of the confidence intervals among individuals $\boldsymbol{x}_i^\star$, the average length (Avg Len) of the confidence intervals, and the computational flops per iteration. From Tables 2 and 3, we summarize the following observations.

(a) In terms of MAE and Ave Len, offline constrained M-estimators achieve results that are *an order of magnitude smaller* than those of online StoSQP. This gap aligns with our analysis from two perspectives. First, M-estimators exhibit $\sqrt{n}$-consistency and optimal asymptotic covariance, while StoSQP estimators exhibit only $\sqrt{1/\bar{\alpha}_n}$-consistency, where $\bar{\alpha}_n$ denotes the stepsize (in this case, $\sqrt{n^{0.501}}$). Clearly, as long as $\bar{\alpha}_n \not\asymp 1/n$, the covariance of StoSQP estimators is not comparable to that of M-estimators when the estimators are scaled by the same scalar. Second, the randomization of the sketching solver within StoSQP introduces additional uncertainty to the estimators, further enlarging the asymptotic covariance (cf. Corollary 5.7), although this enlargement is controlled by the precision of the sketching solver and decreases exponentially with the number of sketching steps.

(b) In terms of Ave Cov, the proposed online StoSQP method achieves promising coverage rates that are very close to 95% for both linear and logistic models as well as for both linear and nonlinear constraints, matching the performance of offline M-estimators. The only potential exception occurs when $d = 5$, where StoSQP may exhibit undercoverage (around 90%) for linear and logistic models. Upon closer examination of these scenarios, we find that the condition number of the Lagrangian Hessian of these problems exceeds $d^3 = 125$, indicating that these problems are ill-conditioned and difficult to solve, particularly when using a sketching solver that seems overkill (see (4.3)). Furthermore, SGD-based estimators are also observed to exhibit undercoverage for various problems due to significant challenges of online inference tasks (Zhu et al., 2021). To our knowledge, StoSQP is the first method capable of conducting online statistical inference for constrained model parameters. While projection-based estimators (Duchi and Ruan, 2021; Davis et al., 2024) may demonstrate similar asymptotic normality, estimating their limiting covariance remains unclear.

(c) In terms of computational flops per iteration, offline M-estimation involves processing the full batch of samples, resulting in significant computational and memory costs. In contrast, our online method processes a single sample, requiring significantly fewer computations. Additionally, the inexact sketching solver for solving Newton systems further reduces the dominant computational cost of the proposed second-order method. Overall, the reduced computational flops per iteration are a major advantage of our online StoSQP method over offline methods.

Further discussions of exact and inexact StoSQP methods are provided in Section F.2.

Obj	Cons	d	Design Cov	MAE (10^{-2})	Ave Cov	Ave Len (10^{-2})	Flops/iter
Lin	Lin	5	Identity	0.20(0.08) 3.09(1.19)	95.40 89.00	0.19(0.03) 2.43(0.39)	$> 1.6 \times 10^7$ 380.00
			Toeplitz ($r = 0.5$)	0.20(0.08) 2.84(1.08)	94.60 90.70	0.19(0.03) 2.39(0.41)	$> 1.7 \times 10^7$ 380.00
			Equi-corr ($r = 0.2$)	0.21(0.09) 3.03(1.04)	94.50 90.70	0.19(0.03) 2.41(0.40)	$> 1.6 \times 10^7$ 380.00
		20	Identity	0.51(0.09) 6.82(1.18)	94.98 93.67	0.23(0.02) 2.87(0.22)	$> 7.1 \times 10^7$ 2340.11
			Toeplitz ($r = 0.5$)	0.52(0.10) 6.83(1.07)	94.35 93.42	0.23(0.02) 2.89(0.22)	$> 6.5 \times 10^7$ 2340.11
			Equi-corr ($r = 0.2$)	0.52(0.09) 6.70(1.10)	94.85 94.40	0.23(0.02) 2.89(0.23)	$> 6.3 \times 10^7$ 2340.11
		40	Identity	0.75(0.09) 10.02(1.15)	94.99 94.32	0.23(0.01) 3.01(0.14)	$> 2.3 \times 10^8$ 7160.90
			Toeplitz ($r = 0.5$)	0.75(0.09) 9.84(1.49)	94.97 94.60	0.23(0.01) 3.03(0.16)	$> 2.0 \times 10^8$ 7160.90
			Equi-corr ($r = 0.2$)	0.75(0.09) 9.84(1.16)	95.35 94.69	0.24(0.01) 3.03(0.15)	$> 1.9 \times 10^8$ 7160.90
		60	Identity	0.95(0.09) 12.34(1.18)	94.47 94.91	0.24(0.01) 3.12(0.12)	$> 4.4 \times 10^8$ 14382.86
			Toeplitz ($r = 0.5$)	0.94(0.09) 12.29(1.18)	94.81 95.21	0.24(0.01) 3.14(0.12)	$> 3.5 \times 10^8$ 14382.86
			Equi-corr ($r = 0.2$)	0.94(0.10) 12.07(1.22)	95.07 95.66	0.24(0.01) 3.14(0.13)	$> 2.9 \times 10^8$ 14382.86
	Non	5	Identity	0.25(0.10) 3.25(1.10)	94.90 93.40	0.22(0.03) 2.82(0.37)	$> 9.3 \times 10^6$ 330.00
			Toeplitz ($r = 0.5$)	0.26(0.10) 3.16(1.01)	93.90 95.00	0.23(0.03) 2.88(0.37)	$> 9.9 \times 10^6$ 330.00
			Equi-corr ($r = 0.2$)	0.25(0.08) 3.21(1.07)	94.80 93.60	0.23(0.03) 2.85(0.37)	$> 9.7 \times 10^6$ 330.00
		20	Identity	0.56(0.09) 7.08(1.11)	94.65 94.55	0.25(0.01) 3.14(0.09)	$> 3.4 \times 10^7$ 2100.07
			Toeplitz ($r = 0.5$)	0.56(0.10) 7.15(1.11)	94.83 95.35	0.25(0.01) 3.18(0.08)	$> 3.7 \times 10^7$ 2100.07
			Equi-corr ($r = 0.2$)	0.56(0.09) 7.11(1.14)	95.23 94.77	0.25(0.01) 3.18(0.08)	$> 3.6 \times 10^7$ 2100.07
		40	Identity	0.79(0.09) 10.32(1.15)	95.23 94.94	0.25(0.01) 3.23(0.06)	$> 6.7 \times 10^7$ 6560.62
			Toeplitz ($r = 0.5$)	0.81(0.09) 10.34(1.20)	95.20 95.45	0.25(0.01) 3.27(0.06)	$> 8.0 \times 10^7$ 6560.62
			Equi-corr ($r = 0.2$)	0.81(0.09) 10.53(1.17)	94.94 94.99	0.25(0.01) 3.28(0.06)	$> 8.0 \times 10^7$ 6560.62
		60	Identity	0.99(0.09) 12.95(1.28)	94.89 95.04	0.25(0.01) 3.30(0.06)	$> 1.0 \times 10^8$ 13422.14
			Toeplitz ($r = 0.5$)	1.00(0.09) 12.80(1.29)	94.82 95.46	0.25(0.01) 3.34(0.05)	$> 1.1 \times 10^8$ 13422.14
			Equi-corr ($r = 0.2$)	0.99(0.09) 12.88(1.20)	95.28 95.63	0.25(0.01) 3.35(0.05)	$> 1.2 \times 10^8$ 13422.14

Table 2: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems (linear models). For each cell, the top row shows the result of the M-estimator, while the bottom row shows the result of StoSQP.*

Obj	Cons	d	Design Cov	MAE (10^{-2})	Ave Cov	Ave Len (10^{-2})	Flops/iter
Logit	Lin	5	Identity	0.13(0.06) 2.14(0.86)	95.00 86.40	0.13(0.03) 1.56(0.42)	$> 1.8 \times 10^7$ 380.00
			Toeplitz ($r = 0.5$)	0.13(0.06) 1.97(0.77)	95.60 89.20	0.12(0.03) 1.53(0.42)	$> 1.2 \times 10^7$ 380.00
			Equi-corr ($r = 0.2$)	0.14(0.06) 2.03(0.81)	93.40 88.00	0.12(0.03) 1.55(0.41)	$> 1.3 \times 10^7$ 380.00
		20	Identity	0.31(0.05) 4.28(0.79)	95.20 92.25	0.14(0.01) 1.73(0.15)	$> 1.0 \times 10^8$ 2340.11
			Toeplitz ($r = 0.5$)	0.30(0.05) 4.01(0.83)	94.60 92.28	0.13(0.01) 1.65(0.15)	$> 9.4 \times 10^7$ 2340.11
			Equi-corr ($r = 0.2$)	0.29(0.05) 3.86(0.73)	95.27 93.20	0.13(0.01) 1.61(0.14)	$> 9.8 \times 10^7$ 2340.11
		40	Identity	0.40(0.05) 5.13(0.64)	95.05 94.94	0.13(0.01) 1.59(0.09)	$> 3.6 \times 10^8$ 7160.90
			Toeplitz ($r = 0.5$)	0.39(0.05) 4.89(0.68)	94.65 94.84	0.12(0.01) 1.51(0.09)	$> 3.5 \times 10^8$ 7160.90
			Equi-corr ($r = 0.2$)	0.35(0.04) 4.28(0.61)	95.07 95.81	0.11(0.01) 1.38(0.09)	$> 3.2 \times 10^8$ 7160.90
		60	Identity	0.48(0.05) 5.85(0.66)	94.74 95.20	0.12(0.01) 1.51(0.08)	$> 6.2 \times 10^8$ 14382.86
			Toeplitz ($r = 0.5$)	0.45(0.04) 5.40(0.61)	94.89 95.66	0.11(0.01) 1.42(0.08)	$> 6.2 \times 10^8$ 14382.86
			Equi-corr ($r = 0.2$)	0.39(0.03) 4.61(0.52)	95.00 96.19	0.10(0.01) 1.24(0.07)	$> 5.9 \times 10^8$ 14382.86
	Non	5	Identity	0.17(0.07) 2.74(0.91)	94.70 89.30	0.17(0.02) 2.08(0.27)	$> 5.2 \times 10^6$ 330.00
			Toeplitz ($r = 0.5$)	0.18(0.07) 2.45(0.85)	94.30 92.00	0.16(0.02) 2.02(0.28)	$> 5.0 \times 10^6$ 330.00
			Equi-corr ($r = 0.2$)	0.18(0.07) 2.47(0.93)	94.50 91.10	0.16(0.02) 2.05(0.28)	$> 5.4 \times 10^6$ 330.00
		20	Identity	0.34(0.06) 4.64(0.85)	94.98 93.20	0.15(0.01) 1.92(0.07)	$> 1.8 \times 10^7$ 2100.07
			Toeplitz ($r = 0.5$)	0.33(0.05) 4.38(0.73)	94.93 93.52	0.15(0.01) 1.83(0.07)	$> 1.8 \times 10^7$ 2100.07
			Equi-corr ($r = 0.2$)	0.32(0.05) 4.31(0.72)	94.50 92.93	0.14(0.01) 1.78(0.07)	$> 1.7 \times 10^7$ 2100.07
		40	Identity	0.45(0.05) 5.82(0.71)	94.60 93.62	0.14(0.01) 1.72(0.06)	$> 3.6 \times 10^7$ 6560.62
			Toeplitz ($r = 0.5$)	0.43(0.05) 5.47(0.77)	94.54 93.88	0.13(0.01) 1.64(0.06)	$> 3.7 \times 10^7$ 6560.62
			Equi-corr ($r = 0.2$)	0.39(0.04) 4.92(0.73)	94.89 94.35	0.12(0.01) 1.50(0.06)	$> 3.4 \times 10^7$ 6560.62
		60	Identity	0.51(0.05) 6.48(0.71)	95.09 94.39	0.13(0.01) 1.60(0.06)	$> 5.6 \times 10^7$ 13422.14
			Toeplitz ($r = 0.5$)	0.47(0.04) 6.16(0.65)	95.36 93.99	0.12(0.01) 1.51(0.05)	$> 5.8 \times 10^7$ 13422.14
			Equi-corr ($r = 0.2$)	0.42(0.04) 5.30(0.66)	95.13 94.34	0.11(0.01) 1.32(0.06)	$> 5.6 \times 10^7$ 13422.14

Table 3: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems (logistic models). For each cell, the top row shows the result of the M-estimator, while the bottom row shows the result of StoSQP.*

7. Conclusion and Future Work

We performed statistical inference of nonlinearly constrained stochastic optimization problems using a fully online second-order method called Stochastic Sequential Quadratic Programming (StoSQP). In each iteration, the scheme selects a proper adaptive stepsize and inexactly solves the Newton system (a quadratic program) by a randomized sketching solver. Consequently, the considered method is more adaptive and computationally efficient than existing exact second-order methods. For this method, we established an almost sure convergence rate and iteration complexity, and proved the asymptotic normality property for the last iterate. We observed that although the limiting covariance is worse than the minimax optimum achieved by constrained M-estimators and online projection-based estimators, the gap decays exponentially fast in terms of the number of iterations employed for the sketching solver (e.g., the covariance matches the optimum if using exact QP solvers). Additionally, we analyzed a plug-in covariance matrix estimator. Our analysis precisely quantified the uncertainty of the stochastic process generated by StoSQP methods, which encompasses the randomness of sampling as well as the computation (sketching and stepsize). The randomness of computation is particularly important for second-order methods to be efficient in practice. With our results, one can apply `AI-StoSQP` to perform online inference for constrained estimation problems.

As for future directions, it is of interest to provide a non-asymptotic analysis for StoSQP methods. Such a result would complement our analysis by bounding the distance between the distribution of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ and the normal distribution for any given t. Furthermore, recent literature on online inference has explored different test statistics, whose asymptotic distributions rely on the application of the Functional Central Limit Theorem (Lee et al., 2022; Luo et al., 2022; Li et al., 2023; Roy and Balasubramanian, 2023; Chen et al., 2024). Establishing a functional CLT for second-order methods and studying the limiting distribution of random scaling estimators is also an interesting research direction. Finally, incorporating nonlinear inequality constraints into the problems and developing second-order methods without projections also deserves further study in future work.

Acknowledgments

SN would like to acknowledge Xinchen Du, Yuefeng Han, and Wanrong Zhu for the helpful discussions of the work. MWM would like to acknowledge the NSF, ONR, and a J. P. Morgan Chase Faculty Research Award for providing partial support of this work. This work was also supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program, under Contract Number DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1): 4148–4187, 2017.

Christophe Andrieu, Eric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, January 2005. ISSN 1095-7138. doi: 10.1137/s0363012902417267.

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, January 2019. ISSN 1095-7189. doi: 10.1137/18m1230323.

Yves F Atchade, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

Albert S. Berahas, Frank E. Curtis, Daniel Robinson, and Baoyu Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, jan 2021. doi: 10.1137/20m1354556.

Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient jacobians. *Mathematics of Operations Research*, October 2023. ISSN 1526-5471. doi: 10.1287/moor.2021.0154.

Bernard Bercu, Antoine Godichon, and Bruno Portier. An efficient stochastic Newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367, jan 2020. doi: 10.1137/19m1261717.

Dimitri Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Elsevier, Belmont, Mass, 1982. ISBN 1886529043. doi: 10.1016/c2013-0-10366-2.

Kenneth A. Bollen and Kwok-fai Ting. A tetrad test for causal indicators. *Psychological Methods*, 5(1):3–22, 2000. ISSN 1082-989X. doi: 10.1037/1082-989x.5.1.3.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2): 459–494, July 2013. ISSN 1436-4646. doi: 10.1007/s10107-013-0701-9.

Claire Boyer and Antoine Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972, dec 2023. doi: 10.1007/s10589-022-00442-3.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, feb 2020. doi: 10.1214/18-aos1801.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17, apr 2021. doi: 10.1080/01621459.2021.1891925.

Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for stochastic optimization via kiefer-wolfowitz methods. *Journal of the American Statistical Association*, pages 1–24, January 2024. ISSN 1537-274X. doi: 10.1080/01621459.2023.2296703.

Xiaoran Cheng and Sen Na. Physics-informed neural networks with trust-region sequential quadratic programming. *arXiv preprint arXiv:2409.10777*, 2024.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3): 88, jul 2022. doi: 10.1007/s10915-022-01939-z.

Frank E. Curtis and Daniel P. Robinson. Exploiting negative curvature in deterministic and stochastic optimization. *Mathematical Programming*, 176(1-2):69–94, oct 2018. doi: 10.1007/s10107-018-1335-8.

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

Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, jun 2023. doi: 10.1007/s10107-023-01981-1.

George B. Dantzig and Gerd Infanger. Multi-stage stochastic linear programs for portfolio optimization. *Annals of Operations Research*, 45(1):59–76, dec 1993. doi: 10.1007/bf02282041.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, mar 1970. doi: 10.1137/0707001.

Damek Davis, Brent Edmunds, and Madeleine Udell. The sound of apalm clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm. *Advances in Neural Information Processing Systems*, 29, 2016.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1): 119–154, January 2019. ISSN 1615-3383. doi: 10.1007/s10208-018-09409-5.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *To appear in Annals of Statistics*, 2024.

J. E. Dennis and Jorge J. Moré. A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of Computation*, 28(126):549–560, 1974. doi: 10.1090/s0025-5718-1974-0343581-1.

Michal Derezinski and Michael W Mahoney. Distributed estimation of the inverse hessian by determinantal averaging. *Advances in Neural Information Processing Systems*, 32, 2019.

Michal Derezinski, Burak Bartan, Mert Pilanci, and Michael W Mahoney. Debiasing distributed second order optimization with surrogate sketching and scaled regularization. *Advances in Neural Information Processing Systems*, 33:6684–6695, 2020a.

Michal Derezinski, Feynman T Liang, Zhenyu Liao, and Michael W Mahoney. Precise expressions for random projections: Low-rank approximation and randomized newton. *Advances in Neural Information Processing Systems*, 33:18272–18283, 2020b.

Michal Derezinski, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. Newton-less: Sparsification without trade-offs for the sketched newton update. *Advances in Neural Information Processing Systems*, 34:2835–2847, 2021.

Michał Dereziński and Elizaveta Rebrova. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science*, 6(1):127–153, February 2024. ISSN 2577-0187. doi: 10.1137/23m1545537.

Nikita Doikov, Peter Richtárik, et al. Randomized block cubic newton method. In *International Conference on Machine Learning*, pages 1290–1298. PMLR, 2018.

Mathias Drton and Han Xiao. Wald tests of singular hypotheses. *Bernoulli*, 22(1), February 2016. ISSN 1350-7265. doi: 10.3150/14-bej620.

Jin-Hong Du, Yifeng Guo, and Xueqin Wang. High-dimensional portfolio selection with cardinality constraints. *Journal of the American Statistical Association*, 118(542):779–791, nov 2022. doi: 10.1080/01621459.2022.2133718.

John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1), feb 2021. doi: 10.1214/19-aos1831.

Marie Duflo. *Random iterative models*, volume 34. Springer, Berlin New York, 1997. ISBN 9783540571001. URL `https://dl.acm.org/doi/10.5555/548484`.

Jitka Dupacova and Roger Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics*, 16(4):1517–1549, dec 1988. doi: 10.1214/aos/1176351052.

Rick Durrett. *Probability*, volume 49. Cambridge University Press, apr 2019. doi: 10.1017/9781108591034.

Yuri Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9(1-2):1–36, jan 1983. doi: 10.1080/17442508308833246.

Vaclav Fabian. Asymptotically efficient stochastic approximation; the RM case. *The Annals of Statistics*, 1(3):486–495, may 1973. doi: 10.1214/aos/1176342414.

Jianqing Fan. Variable screening in high-dimensional feature space. In *Proceedings of the 4th international congress of chinese mathematicians*, volume 2, pages 735–747, 2007. URL `https://fan.princeton.edu/sites/g/files/toruqf5476/files/documents/Screening1.pdf`.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, jun 2012. doi: 10.1080/01621459.2012.682825.

Xiequan Fan. Exact rates of convergence in some martingale central limit theorems. *Journal of Mathematical Analysis and Applications*, 469(2):1028–1044, jan 2019. doi: 10.1016/j.jmaa.2018.09.049.

Yuchen Fang, Sen Na, Michael W Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization*, 2024. URL `https://arxiv.org/abs/2211.15943`.

Charles J. Geyer. Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, 86(415):717–724, sep 1991. doi: 10.1080/01621459.1991.10475100.

Charles J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 22(4):1993–2010, dec 1994. doi: 10.1214/aos/1176325768.

Nicholas I. M. Gould, Dominique Orban, and Philippe L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3):545–557, aug 2014. ISSN 0926-6003. doi: 10.1007/s10589-014-9687-3.

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. Rsn: randomized subspace newton. *Advances in Neural Information Processing Systems*, 32, 2019.

Robert M Gower. Sketch and project: Randomized iterative methods for linear systems and inverting matrices. *arXiv preprint arXiv:1612.06013*, 2016.

Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, jan 2015. doi: 10.1137/15m1025487.

Ilgee Hong, Sen Na, Michael W Mahoney, and Mladen Kolar. Constrained optimization via exact augmented lagrangian and randomized iterative sketching. In *International Conference on Machine Learning*, pages 13174–13198. PMLR, 2023.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, December 1985. ISBN 9780511810817. doi: 10.1017/cbo9780511810817.

DA Jones. Statistical analysis of empirical models fitted by optimization. *Biometrika*, 70(1):67–88, 1983. ISSN 1464-3510. doi: 10.1093/biomet/70.1.67.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, may 2021. doi: 10.1038/s42254-021-00314-5.

Linda Kaufman and Victor Pereyra. A method for separable nonlinear least squares problems with separable nonlinear equality constraints. *SIAM Journal on Numerical Analysis*, 15(1):12–20, feb 1978. ISSN 0036-1429. doi: 10.1137/0715002.

Hassan K. Khalil. *Khalil*, volume [Hauptbd.]. Prentice Hall, Upper Saddle River, NJ, 3. ed. edition, 2002. ISBN 0130673897.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, sep 1952. doi: 10.1214/aoms/1177729392.

Peter Kirkegaard and Morten Eldrup. POSITRONFIT: A versatile program for analysing positron lifetime spectra. *Computer Physics Communications*, 3(3):240–255, apr 1972. doi: 10.1016/0010-4655(72)90070-7.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html`.

Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, volume 26. Springer New York, 1978. doi: 10.1007/978-1-4684-9352-8.

Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci. Optimal iterative sketching methods with the subsampled randomized hadamard transform. *Advances in Neural Information Processing Systems*, 33:9725–9735, 2020.

Jonathan Lacotte, Yifei Wang, and Mert Pilanci. Adaptive newton sketch: Linear-time optimization with quadratic convergence and effective hessian dimensionality. In *International Conference on Machine Learning*, pages 5926–5936. PMLR, 2021.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7381–7389, June 2022. ISSN 2159-5399. doi: 10.1609/aaai.v36i7.20701.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using SGD. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Association for the Advancement of Artificial Intelligence (AAAI), apr 2018. doi: 10.1609/aaai.v32i1.11686.

Xiang Li, Jiadong Liang, and Zhihua Zhang. Online statistical inference for nonlinear stochastic approximation with markovian data. *arXiv preprint arXiv:2302.07690*, 2023.

Faming Liang. Trajectory averaging for stochastic approximation mcmc algorithms. *The Annals of Statistics*, 38(5), October 2010. ISSN 0090-5364. doi: 10.1214/10-aos807.

Tengyuan Liang and Weijie J. Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):431–456, feb 2019. doi: 10.1111/rssb.12313.

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

Weidong Liu, Jiyuan Tu, Yichen Zhang, and Xi Chen. Online estimation and inference for robust policy evaluation in reinforcement learning. *arXiv preprint arXiv:2310.02581*, 2023b.

Lu Lu, Raphaël Pestourie, Wenjie Yao, Zhicheng Wang, Francesc Verdugo, and Steven G. Johnson. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6):B1105–B1132, jan 2021. doi: 10.1137/21m1397908.

Haipeng Luo, Alekh Agarwal, Nicolo Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems*, pages 902–910, 2016.

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

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020. URL `https://proceedings.mlr.press/v125/mou20a.html`.

Sen Na and Mladen Kolar. High-dimensional index volatility models via stein's identity. *Bernoulli*, 27(2), may 2021. doi: 10.3150/20-bej1238.

Sen Na, Zhuoran Yang, Zhaoran Wang, and Mladen Kolar. High-dimensional varying index coefficient models via stein's identity. *J. Mach. Learn. Res.*, 20:152–1, 2019. URL `https://jmlr.csail.mit.edu/papers/v20/18-705.html`.

Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, jun 2022a. doi: 10.1007/s10107-022-01846-z.

Sen Na, Michał Dereziński, and Michael W. Mahoney. Hessian averaging in stochastic Newton methods achieves superlinear convergence. *Mathematical Programming*, dec 2022b. doi: 10.1007/s10107-022-01913-5.

Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*, mar 2023. doi: 10.1007/s10107-023-01935-7.

Neerchal K. Nagaraj and Wayne A. Fuller. Estimation of the parameters of linear time series models subject to nonlinear restrictions. *The Annals of Statistics*, 19(3):1143–1154, sep 1991. ISSN 0090-5364. doi: 10.1214/aos/1176348242.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2nd edition, 2006. ISBN 978-0387-30303-1; 0-387-30303-0. doi: 10.1007/978-0-387-40065-5.

Vivak Patel, Mohammad Jahangoshahi, and Daniel A. Maldonado. An implicit representation and iterative solution of randomly sketched linear systems. *SIAM Journal on Matrix Analysis and Applications*, 42(2):800–831, January 2021. ISSN 1095-7162. doi: 10.1137/19m1259481.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17 (53):1–38, 2016.

Mert Pilanci and Martin J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1): 205–245, jan 2017. doi: 10.1137/15m1021106.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, jul 1992. doi: 10.1137/0330046.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, July 2022. ISSN 1537-274X. doi: 10.1080/01621459.2022.2096620.

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. doi: 10.1016/b978-0-12-604550-5.50015-8.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586.

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

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, 1988. URL `https://ecommons.cornell.edu/bitstream/handle/1813/8664/TR000781.pdf`.

Pranab Kumar Sen. Asymptotic properties of maximum likelihood estimators based on conditional specification. *The Annals of Statistics*, 7(5):1019–1033, sep 1979. ISSN 00905364. doi: 10.1214/aos/1176344785.

Alexander Shapiro. On the asymptotics of constrained local m-estimators. *The Annals of Statistics*, 28(3):948–960, may 2000. ISSN 0090-5364. doi: 10.1214/aos/1015952006.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, jan 2014. doi: 10.1137/1.9781611973433.

Qifan Song, Mingqi Wu, and Faming Liang. Weak convergence rates of population versus single-chain stochastic approximation mcmc algorithms. *Advances in Applied Probability*, 46(4):1059–1083, December 2014. ISSN 1475-6064. doi: 10.1239/aap/1418396243.

Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, apr 2008. doi: 10.1007/s00041-008-9030-4.

Nils Sturma, Mathias Drton, and Dennis Leung. Testing many constraints in possibly irregular models using incomplete u-statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, March 2024. ISSN 1467-9868. doi: 10.1093/jrsssb/qkae022.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, aug 2017. doi: 10.1214/16-aos1506.

Fernando Badilla Veliz, Jean-Paul Watson, Andres Weintraub, Roger J.-B. Wets, and David L. Woodruff. Stochastic optimization models in forest planning: a progressive hedging solution approach. *Annals of Operations Research*, 232:259–274, may 2014. doi: 10.1007/s10479-014-1608-4.

Andreas Wächter and Lorenz 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. ISSN 0025-5610. doi: 10.1007/s10107-004-0559-y.

Jia-Gang Wang. The asymptotic behavior of locally square integrable martingales. *The Annals of Probability*, 23(2):552–585, apr 1995. doi: 10.1214/aop/1176988279.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1):99–111, mar 1972. doi: 10.1007/bf01932678.

Roger J-B Wets. Statistical estimation from an optimization viewpoint. *Annals of Operations Research*, 85(0):79–101, 1999. doi: 10.1023/a:1018934214007.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019. URL `https://jmlr.org/papers/v20/18-262.html`.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–30, may 2021. doi: 10.1080/01621459.2021.1933498.

Appendix A. An Example of Stepsize Selection Scheme

We consider selecting a stepsize to decrease the penalized objective

$$\phi_\nu(\boldsymbol{x};\xi) \coloneqq \nu F(\boldsymbol{x};\xi) + \|c(\boldsymbol{x})\|.$$

We note that decreasing the objective $F(\boldsymbol{x};\xi)$ only is not reasonable for constrained problems, since we may violate constraints arbitrarily. The local linear approximation of $\phi_\nu(\boldsymbol{x};\xi)$ at $(\boldsymbol{x}_t;\xi_t)$ along the direction $\bar{\Delta}\boldsymbol{x}_t$ is

$$\phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t) \coloneqq \nu(F(\boldsymbol{x}_t;\xi_t) + \bar{g}_t^T\bar{\Delta}\boldsymbol{x}_t) + \|c_t + G_t\bar{\Delta}\boldsymbol{x}_t\|.$$

We can further define the local model reduction, a negative quantity for sufficiently small $\nu > 0$ and approximation error, as

$$\begin{aligned}\Delta\phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t) &\coloneqq \phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t) - \phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\boldsymbol{0})\\ &= \nu\bar{g}_t^T\bar{\Delta}\boldsymbol{x}_t + \|c_t + G_t\bar{\Delta}\boldsymbol{x}_t\| - \|c_t\|. \end{aligned} \tag{A.1}$$

For a given scalar $\kappa_t \in (0,1)$, we select $\bar{\alpha}_t$ such that $\phi_\nu(\boldsymbol{x}_t + \bar{\alpha}_t\bar{\Delta}\boldsymbol{x}_t;\xi_t)$ decreases $\phi_\nu(\boldsymbol{x}_t;\xi_t)$ by at least a factor of $\kappa_t\bar{\alpha}_t$ of the local model reduction $\Delta\phi_\nu^{\mathrm{loc}}$ (so called the Armijo condition). Specifically, we require

$$\phi_\nu(\boldsymbol{x}_t + \bar{\alpha}_t\bar{\Delta}\boldsymbol{x}_t;\xi_t) \le \phi_\nu(\boldsymbol{x}_t;\xi_t) + \kappa_t\bar{\alpha}_t\cdot\Delta\phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t). \tag{A.2}$$

To satisfy (A.2), we suppose $\nabla F(\boldsymbol{x};\xi_t)$ and $G(\boldsymbol{x})$ are local Lipschitz continuous around $\boldsymbol{x}_t$, and suppose $\bar{\alpha}_t \le 1$. Then, there exists a constant $\Upsilon_{\nu,t} > 0$, such that

$$\begin{aligned}&\phi_\nu(\boldsymbol{x}_t + \bar{\alpha}_t\bar{\Delta}\boldsymbol{x}_t;\xi_t) = \nu F(\boldsymbol{x}_t + \bar{\alpha}_t\bar{\Delta}\boldsymbol{x};\xi_t) + \|c(\boldsymbol{x}_t + \bar{\alpha}_t\bar{\Delta}\boldsymbol{x})\|\\ &\le \phi_\nu(\boldsymbol{x}_t;\xi_t) + \nu\bar{\alpha}_t\bar{g}_t^T\bar{\Delta}\boldsymbol{x}_t + \|c_t + \bar{\alpha}_tG_t\bar{\Delta}\boldsymbol{x}_t\| - \|c_t\| + \Upsilon_{\nu,t}\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \ \text{(since } \nabla F, G \text{ are Lip)}\\ &\le \phi_\nu(\boldsymbol{x}_t;\xi_t) + \nu\bar{\alpha}_t\bar{g}_t^T\bar{\Delta}\boldsymbol{x}_t + \bar{\alpha}_t\|c_t + G_t\bar{\Delta}\boldsymbol{x}_t\| - \bar{\alpha}_t\|c_t\| + \Upsilon_{\nu,t}\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2 \ \text{(since } \bar{\alpha}_t \le 1)\\ &\overset{(A.1)}{=} \phi_\nu(\boldsymbol{x}_t;\xi_t) + \bar{\alpha}_t\cdot\Delta\phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t) + \Upsilon_{\nu,t}\bar{\alpha}_t^2\|\bar{\Delta}\boldsymbol{x}_t\|^2.\end{aligned}$$

Therefore, (A.2) is satisfied as long as

$$\bar{\alpha}_t \le \frac{(\kappa_t - 1)\cdot\Delta\phi_\nu^{\mathrm{loc}}(\boldsymbol{x}_t;\xi_t,\bar{\Delta}\boldsymbol{x}_t)}{\Upsilon_{\nu,t}\|\bar{\Delta}\boldsymbol{x}_t\|^2} \wedge 1 =: \bar{\alpha}_{t,thres}. \tag{A.3}$$

The Lipschitz constant $\Upsilon_{\nu,t}$ can be estimated around $\boldsymbol{x}_t$ (Curtis and Robinson, 2018) or simply prespecified as a large constant. The condition (A.3) leads us to propose $\bar{\alpha}_t \coloneqq \mathrm{Proj}_{[\beta_t,\eta_t]}(\bar{\alpha}_{t,thres})$. See Berahas et al. (2021, 2023); Curtis et al. (2021) for detailed random projections and Hong et al. (2023) for adaptive selection of parameters of line search functions (e.g. ν).

Appendix B. Preparation Lemmas

Lemma B.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 $p \ge 0$, we have $\lim_{i\to\infty} i\left(1 - \varphi_{i-1}^p/\varphi_i^p\right) = p\cdot\varphi$.*

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

Lemma B.3 *Let $\{\phi_i\}_i$, $\{\varphi_i\}_i$, $\{\sigma_i\}_i$ be three positive sequences. Suppose*

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

for a constant ϕ and a (possibly infinite) constant $\widetilde{\varphi} \in (0,\infty]$. For any $l \geq 1$, if we further have $\sum_{k=1}^{l} \sigma_k + p\phi/\widetilde{\varphi} > 0$ for some constant $p \in (0,1]$, then the following results hold as $t \to \infty$

(a): When $p = 1$,

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

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

(b): When $p \in (0,1)$,

$$\begin{aligned} &\frac{1}{\phi_t^p}\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i \longrightarrow 0, \\ &\frac{1}{\phi_t^p}\left\{\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i^p a_i + b\cdot\prod_{j=0}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\right\} \longrightarrow 0, \end{aligned} \tag{B.3}$$

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

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

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

$$\sup_{z\in\mathbb{R}} |P\left(A_t \leq z\right) - \Phi(z)| \leq a_t, \quad |B_t| \leq b_t, \quad |C_t| \leq c_t \quad \textit{almost surely} \tag{B.4}$$

where $a_t, b_t \geq 0$ and $0 \leq c_t < 1$. Then, we have

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

Appendix C. Proofs of Preparation Lemmas

C.1 Proof of Lemma B.1

By the condition, we know $\varphi_{i-1}/\varphi_i = 1 - \varphi/i + o\left(1/i\right)$. Thus, we have

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

This completes the proof.

C.2 Proof of Lemma B.2

By Lemma B.1, we know for any positive constant p, $\lim_{i\to\infty} i\left(1-\varphi_{i-1}^p/\varphi_i^p\right) = p\varphi$. 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.

C.3 Proof of Lemma B.3

For any scalar A, we have

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

For the last term, we have

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

Combining the above two displays, we obtain

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

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

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

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, we know

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

With these two facts, we have

$$\varphi_i\phi_i^{1-p} - A\big\{1 - \frac{\phi_{i-1}^p}{\phi_i^p}\prod_{k=1}^{l}(1-\varphi_i\sigma_k)\big\} = \varphi_i\phi_i^{1-p} - A\big\{1 - (1 - \frac{p\phi}{\widetilde{\varphi}}\cdot\varphi_i + o(\varphi_i))(1 - \varphi_i\sum_{k=1}^{l}\sigma_k + o(\varphi_i))\big\}$$

$$= \varphi_i\phi_i^{1-p} - A\left(\frac{p\phi}{\widetilde{\varphi}} + \sum_{k=1}^{l}\sigma_k\right)\varphi_i + o(\varphi_i). \tag{C.2}$$

Thus, we let $A = 1/(\sum_{k=1}^{l}\sigma_k + \phi/\widetilde{\varphi})$ if $p = 1$ and $A = 0$ if $p \in (0,1)$. Noting that $\phi_i^{1-p} \to 0$, (C.1) leads to

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

$$= \frac{1}{\phi_t^p}\prod_{j=0}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\Big\{\sum_{i=1}^{t}\prod_{j=0}^{i}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)^{-1}\phi_i^p\cdot o(\varphi_i)$$

$$+ \phi_0^p\prod_{k=1}^{l}(1-\varphi_0\sigma_k)^{-1}\left(\varphi_0\phi_0^{1-p} - A\right)\Big\}.$$

Comparing the above display with (B.2) and (B.3), we note that the first results in (B.2) and (B.3) are implied by the second results. Thus, it suffices to prove the second results. We define

$$\Psi_t = \frac{1}{\phi_t^p}\left\{\sum_{i=0}^{t}\prod_{j=i+1}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i^p a_i + b\cdot\prod_{j=0}^{t}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\right\}, \tag{C.3}$$

then

$$\Psi_t = \frac{1}{\phi_t^p}\left\{\varphi_t\phi_t^p a_t + \prod_{k=1}^{l}(1-\varphi_t\sigma_k)\left(\sum_{i=0}^{t-1}\prod_{j=i+1}^{t-1}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\varphi_i\phi_i^p a_i + b\cdot\prod_{j=0}^{t-1}\prod_{k=1}^{l}(1-\varphi_j\sigma_k)\right)\right\}$$

$$\overset{\text{(C.3)}}{=} \frac{\phi_{t-1}^p}{\phi_t^p}\prod_{k=1}^{l}(1-\varphi_t\sigma_k)\Psi_{t-1} + \varphi_t a_t.$$

By (C.2), we know that

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

Since $\sum_{k=1}^{l}\sigma_k + p\phi/\widetilde{\varphi} > 0$, we immediately conclude 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 this 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

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

Since $n\varphi_i \to \widetilde{\varphi} \in (0,\infty]$, we know $\sum_t \varphi_t \to \infty$. Thus, for the above $\epsilon > 0$, there exists $t_2 \ge t_1$ 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.

C.4 Proof of Lemma B.4

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

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

where the last inequality uses $a\exp(-a^2/2) \le 1$ for all a. This completes the proof.

C.5 Proof of Lemma B.5

We only prove the result for $z > 0$. The result of $z \le 0$ can be shown in the same way. We know from (B.4) that $\frac{A_t-b_t}{\sqrt{1+c_t}} \le \frac{A_t+B_t}{\sqrt{1+C_t}} \le \frac{A_t+b_t}{\sqrt{1-c_t}}$, almost surely. Therefore, we have

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

On the other hand, we have

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

Combining the above two displays completes the proof.

Appendix D. Proofs of Section 4

D.1 Proof of Lemma 4.4

We note that $\gamma_S \le \|\mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]\| \le \mathbb{E}[\|K_t S(S^T K_t^2 S)^\dagger S^T K_t\| \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \le 1$, where the second inequality is by Jensen's inequality; the third inequality is by the fact that $K_t S(S^T K_t^2 S)^\dagger S^T K_t$ is a projection matrix. This shows (a). Let us define for $j = 1, \ldots, \tau$, $C_{t,j} = I - K_t S_{t,j}(S_{t,j}^T K_t^2 S_{t,j})^\dagger S_{t,j}^T K_t$. Then, we obtain from (3.4) that

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

Thus, (b) follows from (D.1) and the independence among $\{S_{t,j}\}_j$. Moreover, (c) is proved by (Gower and Richtárik, 2015, Theorem 4.6).

D.2 Proof of Lemma 4.6

By Assumption 4.1, there exists a constant $\Upsilon_u \ge 1$ such that

$$\|\nabla^2 \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\| \vee \|\nabla \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})\| \le \Upsilon_u, \quad \forall (\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{X} \times \Lambda. \tag{D.2}$$

By direct calculation, we have (the evaluation point is suppressed for simplicity)

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\mu,\nu} \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mu,\nu} \end{pmatrix} = \begin{pmatrix} I + \nu \nabla_{\boldsymbol{x}}^2 \mathcal{L} & \mu G^T \\ \nu G & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L} \\ c \end{pmatrix}. \tag{D.3}$$

Using (D.3) and the definition of $(\Delta \boldsymbol{x}_t, \Delta \boldsymbol{\lambda}_t)$, we have

$$\begin{aligned}
\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} &\overset{\text{(D.3)}}{=} \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} I + \nu \nabla_{\boldsymbol{x}}^2 \mathcal{L}_t & \mu G_t^T \\ \nu G_t & I \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_t \\ c_t \end{pmatrix} \\
&\overset{(3.2)}{=} -\begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} I + \nu \nabla_{\boldsymbol{x}}^2 \mathcal{L}_t & \mu G_t^T \\ \nu G_t & I \end{pmatrix} \begin{pmatrix} B_t & G_t^T \\ G_t & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \\
&= -\begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}^T \begin{pmatrix} B_t + \nu \nabla_{\boldsymbol{x}}^2 \mathcal{L}_t B_t + \mu G_t^T G_t & G_t^T + \nu \nabla_{\boldsymbol{x}}^2 \mathcal{L}_t G_t^T \\ G_t + \nu G_t B_t & \nu G_t G_t^T \end{pmatrix} \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}.
\end{aligned}$$

Furthermore, using (D.2) and Assumption 4.1, we obtain

$$\begin{aligned}
&\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \\
&\le -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu \Upsilon_B \Upsilon_u \|\Delta \boldsymbol{x}_t\|^2 - \mu \|G_t \Delta \boldsymbol{x}_t\|^2 - 2\Delta \boldsymbol{\lambda}_t^T G_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_u + \Upsilon_B)\|\Delta \boldsymbol{x}_t\| \|G_t^T \Delta \boldsymbol{\lambda}_t\| - \nu \|G_t^T \Delta \boldsymbol{\lambda}_t\|^2 \\
&\overset{(3.2)}{\le} -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu \Upsilon_B \Upsilon_u \|\Delta \boldsymbol{x}_t\|^2 - \mu \|c_t\|^2 + 2c_t^T \Delta \boldsymbol{\lambda}_t + \frac{\nu(\Upsilon_u + \Upsilon_B)^2}{2} \|\Delta \boldsymbol{x}_t\|^2 - \frac{\nu}{2} \|G_t^T \Delta \boldsymbol{\lambda}_t\|^2 \\
&\le -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 - \mu \|c_t\|^2 + \frac{8}{\nu \gamma_G} \|c_t\|^2 + \frac{\nu \gamma_G}{8} \|\Delta \boldsymbol{\lambda}_t\|^2 \\
&\quad - \frac{\nu \gamma_G}{4} \|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{4} \|G_t^T \Delta \boldsymbol{\lambda}_t\|^2 \quad \text{(Young's inequality and Assumption 4.1)} \\
&\overset{(3.2)}{=} -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu \gamma_G}\right) \|c_t\|^2 - \frac{\nu \gamma_G}{8} \|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{4} \|B_t \Delta \boldsymbol{x}_t + \nabla_{\boldsymbol{x}} \mathcal{L}_t\|^2
\end{aligned}$$

$$\leq -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + \nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu\gamma_G}\right)\|c_t\|^2 - \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{8}\|\nabla_{\boldsymbol{x}} \mathcal{L}_t\|^2 + \frac{\nu \Upsilon_B^2}{4}\|\Delta \boldsymbol{x}_t\|^2$$
$$\leq -\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 - \left(\mu - \frac{8}{\nu\gamma_G}\right)\|c_t\|^2 - \frac{\nu\gamma_G}{8}\|\Delta \boldsymbol{\lambda}_t\|^2 - \frac{\nu}{8}\|\nabla_{\boldsymbol{x}} \mathcal{L}_t\|^2, \quad \text{(D.4)}$$

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 (D.4), we decompose the step $\Delta \boldsymbol{x}_t$ as

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

Then, the first two terms of (D.4) can be simplified as

$$\begin{aligned}
&-\Delta \boldsymbol{x}_t^T B_t \Delta \boldsymbol{x}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 \\
&\quad = -\Delta \boldsymbol{u}_t^T B_t \Delta \boldsymbol{u}_t - 2\Delta \boldsymbol{u}_t^T B_t \Delta \boldsymbol{v}_t - \Delta \boldsymbol{v}_t^T B_t \Delta \boldsymbol{v}_t + 2\nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 \\
&\quad \leq \Upsilon_B \|\Delta \boldsymbol{u}_t\|^2 + 2\Upsilon_B \|\Delta \boldsymbol{u}_t\| \|\Delta \boldsymbol{v}_t\| - \gamma_{RH} \|\Delta \boldsymbol{v}_t\|^2 + 2\nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 \quad \text{(Assumption 4.1)} \\
&\quad \leq \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}}\right)\|\Delta \boldsymbol{u}_t\|^2 - \frac{\gamma_{RH}}{2}\|\Delta \boldsymbol{v}_t\|^2 + 2\nu(\Upsilon_B + \Upsilon_u)^2 \|\Delta \boldsymbol{x}_t\|^2 \quad \text{(Young's inequality)} \\
&\quad = \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}} + \frac{\gamma_{RH}}{2}\right)\|\Delta \boldsymbol{u}_t\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2\right)\|\Delta \boldsymbol{x}_t\|^2 \\
&\quad \leq \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}} + \frac{\gamma_{RH}}{2}\right)\frac{1}{\gamma_G}\|c_t\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2\right)\|\Delta \boldsymbol{x}_t\|^2,
\end{aligned}$$

where the last inequality uses the fact that $\|c_t\|^2 = \|G_t \Delta \boldsymbol{x}_t\|^2 = \|G_t \Delta \boldsymbol{u}_t\|^2 \geq \gamma_G \|\Delta \boldsymbol{u}_t\|^2$. Here, the inequality is due to Assumption 4.1. Combining the above display with (D.4), we have

$$\begin{aligned}
\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_{\mu,\nu}^t \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\mu,\nu}^t \end{pmatrix}^T \begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix} \leq &-\frac{\nu\gamma_G}{8}\left\|\begin{pmatrix} \Delta \boldsymbol{x}_t \\ \Delta \boldsymbol{\lambda}_t \end{pmatrix}\right\|^2 - \frac{\nu}{8}\left\|\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}_t \\ c_t \end{pmatrix}\right\|^2 - \left(\frac{\gamma_{RH}}{2} - 2\nu(\Upsilon_B + \Upsilon_u)^2 - \frac{\nu\gamma_G}{8}\right)\|\Delta \boldsymbol{x}_t\|^2 \\
&- \left\{\mu - \frac{8}{\nu\gamma_G} - \left(\Upsilon_B + \frac{2\Upsilon_B^2}{\gamma_{RH}} + \frac{\gamma_{RH}}{2}\right)\frac{1}{\gamma_G} - \frac{\nu}{8}\right\}\|c_t\|^2.
\end{aligned}$$

Thus, choosing Υ_1 large enough (depending only on $\gamma_G, \gamma_{RH}, \Upsilon_B$), we complete the proof.

D.3 Proof of Lemma 4.7

Let us denote $\boldsymbol{z}_t = (\Delta \boldsymbol{x}_t, \Delta \boldsymbol{\lambda}_t)$ and recall that $\boldsymbol{z}_{t,\tau} = (\bar{\Delta} \boldsymbol{x}_t, \bar{\Delta} \boldsymbol{\lambda}_t)$ and $\widetilde{\boldsymbol{z}}_t = (\widetilde{\Delta} \boldsymbol{x}_t, \widetilde{\Delta} \boldsymbol{\lambda}_t)$. By Assumption 4.1 and the expression (D.3), it is straightforward to see that $\nabla \mathcal{L}_{\mu,\nu}$ is Lipschitz continuous with a constant $\Upsilon_{\mathcal{AL}} > 0$ depending on (μ, ν, Υ_L). Thus, using (3.5) we have

$$\begin{aligned}
\mathcal{L}_{\mu,\nu}^{t+1} &\leq \mathcal{L}_{\mu,\nu}^t + \bar{\alpha}_t (\nabla \mathcal{L}_{\mu,\nu}^t)^T \boldsymbol{z}_{t,\tau} + \frac{\Upsilon_{\mathcal{AL}} \bar{\alpha}_t^2}{2}\|\boldsymbol{z}_{t,\tau}\|^2 \\
&= \mathcal{L}_{\mu,\nu}^t + \bar{\alpha}_t (\nabla \mathcal{L}_{\mu,\nu}^t)^T (I + C_t)\boldsymbol{z}_t + \bar{\alpha}_t (\nabla \mathcal{L}_{\mu,\nu}^t)^T \{\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\} + \frac{\Upsilon_{\mathcal{AL}} \bar{\alpha}_t^2}{2}\|\boldsymbol{z}_{t,\tau}\|^2, \quad \text{(D.5)}
\end{aligned}$$

where C_t is from Lemma 4.4(b). By Lemmas 4.6 and 4.4, the second term can be bounded as

$$(\nabla \mathcal{L}_{\mu,\nu}^t)^T (I + C_t)\boldsymbol{z}_t \leq -\frac{\nu}{\Upsilon_1}\left(\|\boldsymbol{z}_t\|^2 + \|\nabla \mathcal{L}_t\|^2\right) + \|C_t\| \left\|\nabla \mathcal{L}_{\mu,\nu}^t\right\| \|\boldsymbol{z}_t\|$$

$$\begin{aligned}
&\overset{\text{(D.3),(D.2)}}{\leq} -\frac{\nu}{\Upsilon_1}\left(\|\boldsymbol{z}_t\|^2 + \|\nabla\mathcal{L}_t\|^2\right) + \rho^\tau\left(1 + (2\nu+\mu)\Upsilon_u\right)\|\nabla\mathcal{L}_t\|\,\|\boldsymbol{z}_t\| \\
&\leq -\frac{\nu}{\Upsilon_1}\left(\|\boldsymbol{z}_t\|^2 + \|\nabla\mathcal{L}_t\|^2\right) + 2\rho^\tau\mu\Upsilon_u\|\nabla\mathcal{L}_t\|\,\|\boldsymbol{z}_t\| \\
&\leq -\left(\frac{\nu}{\Upsilon_1} - \rho^\tau\mu\Upsilon_u\right)\left(\|\boldsymbol{z}_t\|^2 + \|\nabla\mathcal{L}_t\|^2\right),
\end{aligned}$$

where the third inequality uses the facts that $\Upsilon_u \geq 1$ and $1 + 2\nu \leq \mu$ (as long as $\Upsilon_1 \geq 2$). Thus, we can re-define Υ_1 as $\Upsilon_1 \leftarrow 2\Upsilon_1\Upsilon_u$. If $\rho^\tau \leq \nu/(\mu\Upsilon_1)$, then we have

$$(\nabla\mathcal{L}^t_{\mu,\nu})^T(I + C_t)\boldsymbol{z}_t \leq -\frac{\nu}{2\Upsilon_1}\left(\|\boldsymbol{z}_t\|^2 + \|\nabla\mathcal{L}_t\|^2\right). \tag{D.6}$$

Now, we deal with the last two terms of (D.5). By Lemma 4.4(b), we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{z}_{t,\tau} \mid \mathcal{F}_{t-2/3}\right] \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[(I + C_t)\widetilde{\boldsymbol{z}}_t \mid \mathcal{F}_{t-1}\right] \\
&\overset{(3.2)}{=} -(I + C_t)K_t^{-1}\mathbb{E}\left[\bar{\nabla}\mathcal{L}_t \mid \mathcal{F}_{t-1}\right] = -(I + C_t)K_t^{-1}\nabla\mathcal{L}_t \quad \text{(Assumption 4.2)} \\
&\overset{(3.2)}{=} (I + C_t)\boldsymbol{z}_t.
\end{aligned} \tag{D.7}$$

By Lemma 4.4(b, c), we also have

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
&\leq 3\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau} - \widetilde{\boldsymbol{z}}_t\|^2 \mid \mathcal{F}_{t-1}\right] + 3\mathbb{E}\left[\|\widetilde{\boldsymbol{z}}_t - \boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] + 3\|C_t\|^2\|\boldsymbol{z}_t\|^2 \\
&\leq 3\rho^\tau\mathbb{E}\left[\|\widetilde{\boldsymbol{z}}_t\|^2 \mid \mathcal{F}_{t-1}\right] + 3\mathbb{E}\left[\|\widetilde{\boldsymbol{z}}_t - \boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] + 3\rho^{2\tau}\|\boldsymbol{z}_t\|^2 \\
&= 3(\rho^\tau + \rho^{2\tau})\|\boldsymbol{z}_t\|^2 + 3(1 + \rho^\tau)\mathbb{E}\left[\|\widetilde{\boldsymbol{z}}_t - \boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] \quad \text{(bias-variance decomposition)}.
\end{aligned}$$

By Assumption 4.1 and (Na et al., 2022a, Lemma 1), there exists a constant $\Upsilon_K \geq 1$ depending on $(\gamma_G, \gamma_{RH}, \Upsilon_B)$ such that $\|K_t^{-1}\| \leq \Upsilon_K$. Thus, we apply (3.2) and (D.2), and obtain

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] &\leq 3(\rho^\tau + \rho^{2\tau})\Upsilon_K^2\Upsilon_u^2 + 3(1 + \rho^\tau)\Upsilon_K^2\mathbb{E}[\|\bar{g}_t - \nabla f_t\|^2 \mid \mathcal{F}_{t-1}] \\
&\leq 3(1 + \rho^\tau)\Upsilon_K^2(\rho^\tau\Upsilon_u^2 + \Upsilon_m) \quad \text{(Assumption 4.2(4.2a))}.
\end{aligned} \tag{D.8}$$

Thus, using (D.7) and (D.8), we have

$$\begin{aligned}
&\mathbb{E}\left[\bar{\alpha}_t(\nabla\mathcal{L}^t_{\mu,\nu})^T\left\{\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\right\} \mid \mathcal{F}_{t-1}\right] \\
&\quad\overset{\text{(D.7)}}{=} \mathbb{E}\left[\{\bar{\alpha}_t - (\beta_t + \eta_t)/2\}\cdot(\nabla\mathcal{L}^t_{\mu,\nu})^T\left\{\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\right\} \mid \mathcal{F}_{t-1}\right] \\
&\quad\overset{(3.6)}{\leq} \frac{\eta_t - \beta_t}{2}\mathbb{E}\left[\left\|\nabla\mathcal{L}^t_{\mu,\nu}\right\|\|\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\| \mid \mathcal{F}_{t-1}\right] \\
&\quad\overset{\text{(D.2)}}{\leq} \frac{\eta_t - \beta_t}{2}(1 + (2\nu+\mu)\Upsilon_u)\Upsilon_u\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\| \mid \mathcal{F}_{t-1}\right] \\
&\quad\leq (\eta_t - \beta_t)\mu\Upsilon_u^2\sqrt{\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau} - (I + C_t)\boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right]} \quad (1 \leq \Upsilon_u \text{ and } 1 + 2\nu \leq \mu)
\end{aligned}$$

$$\overset{\text{(D.8)}}{\leq} 2\mu\Upsilon_K\Upsilon_u^2(1+\rho^\tau)(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t) \leq 4\mu\Upsilon_K\Upsilon_u^2(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t), \quad \text{(D.9)}$$

and

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{z}_{t,\tau}\|^2 \mid \mathcal{F}_{t-1}] &\overset{\text{(D.7)}}{=} \|(I+C_t)\boldsymbol{z}_t\|^2 + \mathbb{E}\left[\|\boldsymbol{z}_{t,\tau}-(I+C_t)\boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\ &\overset{\text{(3.2),(D.2)}}{\leq} (1+\rho^\tau)^2\Upsilon_K^2\Upsilon_u^2 + \mathbb{E}\left[\|\boldsymbol{z}_{t,\tau}-(I+C_t)\boldsymbol{z}_t\|^2 \mid \mathcal{F}_{t-1}\right] \quad \text{(also use Lemma 4.4(b))} \\ &\overset{\text{(D.8)}}{\leq} (1+\rho^\tau)^2\Upsilon_K^2\Upsilon_u^2 + 3(1+\rho^\tau)\Upsilon_K^2(\rho^\tau\Upsilon_u^2+\Upsilon_m) \leq 16\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m). \quad \text{(D.10)}\end{aligned}$$

Combining (D.10) with (D.9) and (D.6), plugging into (D.5), and using (3.6), we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}_{\mu,\nu}^t - \frac{\nu\beta_t}{2\Upsilon_1}\left(\|\boldsymbol{z}_t\|^2 + \|\nabla\mathcal{L}_t\|^2\right) \\ + 4\mu\Upsilon_K\Upsilon_u^2(\sqrt{\Upsilon_m}\vee\Upsilon_u)(\eta_t-\beta_t) + 8\Upsilon_{\mathcal{AL}}\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m)\eta_t^2.\end{aligned}$$

Choosing Υ_2 large enough that depends on $(\mu,\nu,\gamma_G,\gamma_{RH},\Upsilon_B,\Upsilon_m,\Upsilon_L)$, and noting that (μ,ν) are determined by $(\gamma_G,\gamma_{RH},\Upsilon_B)$, we complete the proof.

D.4 Proof of Theorem 4.8

Note that the condition of τ in the statement implies that we can select (μ,ν) to satisfy the condition in Lemma 4.6 and have $\rho^\tau \leq \nu/(\mu\Upsilon_1)$ with $\rho = 1-\gamma_S$. Thus, Lemma 4.7 leads to

$$\mathbb{E}[\mathcal{L}_{\mu,\nu}^{t+1} - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} \mid \mathcal{F}_{t-1}] \leq \mathcal{L}_{\mu,\nu}^t - \min_{\mathcal{X}\times\Lambda}\mathcal{L}_{\mu,\nu} - \frac{\nu\beta_t}{2\Upsilon_1}\|\nabla\mathcal{L}_t\|^2 + \Upsilon_2(\chi_t+\eta_t^2).$$

By Robbins-Siegmund theorem (see Robbins and Siegmund (1971) or (Duflo, 1997, Theorem 1.3.12)), we conclude that $\sum_t \beta_t\|\nabla\mathcal{L}_t\|^2 < \infty$. Since $\sum_t\beta_t = \infty$ from (4.4), we know that $\liminf_{t\to\infty}\|\nabla\mathcal{L}_t\| = 0$. Furthermore, we note that

$$\begin{aligned}\mathbb{E}\left[\|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t,\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t)\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t,\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t)\|^2 \mid \mathcal{F}_{t-1}\right]\right] \\ &\overset{\text{(3.5)}}{\leq} \eta_t^2\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{z}_{t,\tau}\|^2 \mid \mathcal{F}_{t-1}\right]\right] \overset{\text{(D.10)}}{\leq} 16\Upsilon_K^2(\Upsilon_u^2\vee\Upsilon_m)\cdot\eta_t^2. \quad \text{(D.11)}\end{aligned}$$

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

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

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 for any run of the algorithm $\lim_{t\to\infty}\|\nabla\mathcal{L}_t\| \neq 0$, then we have $\limsup_{t\to\infty}\|\nabla\mathcal{L}_t\| = \epsilon > 0$. Then, there exist two index sequences $\{t_{1,i}\}_i, \{t_{2,i}\}_i$ with $t_{1,i+1} > t_{2,i} > t_{1,i}$ such that, for all $i = 1,2,\ldots,$

$$\|\nabla\mathcal{L}_{t_{1,i}}\| \geq \epsilon/2, \qquad \|\nabla\mathcal{L}_j\| \geq \epsilon/3 \text{ for } j = t_{1,i}+1,\ldots,t_{2,i}-1, \qquad \|\nabla\mathcal{L}_{t_{2,i}}\| < \epsilon/3. \quad \text{(D.12)}$$

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

$$\infty > \sum_{i=1}^{\infty} \sum_{j=t_{1,i}}^{t_{2,i}-1} \beta_j \|\nabla \mathcal{L}_j\|^2 \overset{(D.12)}{\geq} \frac{\epsilon^2}{9} \sum_{i=1}^{\infty} \sum_{j=t_{1,i}}^{t_{2,i}-1} \beta_j. \tag{D.13}$$

Furthermore, by (D.11), we have

$$\begin{aligned}\mathbb{E}\left[\left\|(\boldsymbol{x}_{t_{2,i}} - \boldsymbol{x}_{t_{1,i}}, \boldsymbol{\lambda}_{t_{2,i}} - \boldsymbol{\lambda}_{t_{1,i}})\right\|\right] &\overset{(D.11)}{\leq} 4\Upsilon_K(\Upsilon_u \vee \sqrt{\Upsilon_m}) \sum_{j=t_{1,i}}^{t_{2,i}-1} \eta_j \\ &\overset{(3.6)}{=} 4\Upsilon_K(\Upsilon_u \vee \sqrt{\Upsilon_m}) \left\{ \sum_{j=t_{1,i}}^{t_{2,i}-1} \beta_j + \sum_{j=t_{1,i}}^{t_{2,i}-1} \chi_j \right\}.\end{aligned}$$

Summing over $i = 1$ to ∞, and noting that $\sum_i \sum_{j=t_{1,i}}^{t_{2,i}-1} \beta_j < \infty$ by (D.13) and $\sum_i \sum_{j=t_{1,i}}^{t_{2,i}-1} \chi_j \leq \sum_{j=1}^{\infty} \chi_j < \infty$, we exchange the expectation and summation 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 $\|\nabla \mathcal{L}_{t_{1,i}}\| \geq \epsilon/2$ and $\|\nabla \mathcal{L}_{t_{2,i}}\| < \epsilon/3$ in (D.12). We complete the proof.

D.5 Proof of Corollary 4.9

Applying Lemma 4.7 and taking full expectation, we know for some constants $h_1, h_2 > 0$,

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

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

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

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

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

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

$$\sum_{t=0}^{\mathcal{T}_\epsilon - 1} \frac{\chi_t + \eta_t^2}{\beta_t} = \sum_{t=0}^{\mathcal{T}_\epsilon - 1} \left\{ (t+1)^{a-b} + (t+1)^a \left((t+1)^{-2a} + 2(t+1)^{-(a+b)} + (t+1)^{-2b} \right) \right\}$$

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

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

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

This completes the proof.

Appendix E. Proofs of Section 5

E.1 Proof of Lemma 5.1

For notational brevity, we let $\boldsymbol{\omega}_t=(\boldsymbol{x}_t-\boldsymbol{x}^\star,\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^\star)$. By the scheme of Algorithm 1, we have

$$
\begin{aligned}
\boldsymbol{\omega}_{t+1}&\overset{(3.5)}{=}\boldsymbol{\omega}_t+\bar{\alpha}_t\boldsymbol{z}_{t,\tau}=\boldsymbol{\omega}_t+\varphi_t\boldsymbol{z}_{t,\tau}+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&=\boldsymbol{\omega}_t+\varphi_t(I+C_t)\widetilde{\boldsymbol{z}}_t+\varphi_t\left\{\boldsymbol{z}_{t,\tau}-(I+C_t)\widetilde{\boldsymbol{z}}_t\right\}+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&\overset{(3.2)}{=}\boldsymbol{\omega}_t-\varphi_t(I+C_t)K_t^{-1}\bar{\nabla}\mathcal{L}_t+\varphi_t\left\{\boldsymbol{z}_{t,\tau}-(I+C_t)\widetilde{\boldsymbol{z}}_t\right\}+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&=\boldsymbol{\omega}_t-\varphi_t(I+C_t)K_t^{-1}\nabla\mathcal{L}_t-\varphi_t(I+C_t)K_t^{-1}(\bar{\nabla}\mathcal{L}_t-\nabla\mathcal{L}_t)+\varphi_t\left\{\boldsymbol{z}_{t,\tau}-(I+C_t)\widetilde{\boldsymbol{z}}_t\right\}\\
&\quad+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&\overset{(5.3b)}{=}\boldsymbol{\omega}_t-\varphi_t(I+C_t)K_t^{-1}\nabla\mathcal{L}_t+\varphi_t\boldsymbol{\theta}^t+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&=\boldsymbol{\omega}_t-\varphi_t(I+C_t)(K^\star)^{-1}\nabla\mathcal{L}_t-\varphi_t(I+C_t)\left\{K_t^{-1}-(K^\star)^{-1}\right\}\nabla\mathcal{L}_t+\varphi_t\boldsymbol{\theta}^t+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&\overset{(5.3d)}{=}\left\{I-\varphi_t(I+C_t)\right\}\boldsymbol{\omega}_t-\varphi_t(I+C_t)(K^\star)^{-1}\boldsymbol{\psi}^t-\varphi_t(I+C_t)\left\{K_t^{-1}-(K^\star)^{-1}\right\}\nabla\mathcal{L}_t\\
&\quad+\varphi_t\boldsymbol{\theta}^t+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}\\
&\overset{(5.3c)}{=}\left\{I-\varphi_t(I+C^\star)\right\}\boldsymbol{\omega}_t+\varphi_t(\boldsymbol{\theta}^t+\boldsymbol{\delta}^t)+(\bar{\alpha}_t-\varphi_t)\,\boldsymbol{z}_{t,\tau}.
\end{aligned}
$$

We apply the above equation recursively and show the result. Moreover, under Assumptions 4.2 and 4.3, we know $\mathbb{E}[\bar{g}_i - \nabla f_i \mid \mathcal{F}_{i-1}] = \mathbf{0}$ and, by (D.7), $\mathbb{E}\left[\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i \mid \mathcal{F}_{i-1}\right] = \mathbf{0}$. Thus, $\mathbb{E}[\boldsymbol{\theta}^i \mid \mathcal{F}_{i-1}] = \mathbf{0}$ and $\boldsymbol{\theta}^i$ is a martingale difference.

E.2 Proof of Lemma 5.2

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

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

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

$$\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star\| = \|EE^T - E'(E')^T\|. \tag{E.1}$$

Define the principle angles θ_p between $\operatorname{span}(E)$ and $\operatorname{span}(E')$ to be $\theta_p = (\theta_{p,1}, \ldots, \theta_{p,r})$, so that $E^T E'$ has the singular value decomposition $E^T E' = P\cos(\theta_p)Q^T$, where $P, Q \in \mathbb{R}^{r\times r}$ are orthonormal matrices and $\cos(\theta_p) = \operatorname{diag}(\cos(\theta_{p,1}), \ldots, \cos(\theta_{p,r}))$ (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

$$E' = EA + E^\perp B. \tag{E.2}$$

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

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

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

$$\|\sin(\theta_p)\| \leq \frac{\|(K^\star - K_t)S\|}{D_r}. \tag{E.4}$$

We let F_r be the r-th column of F and have $D_r^2 = F_r^T S^T (K^\star)^2 S F_r \geq (\sigma_{\min}(K^\star))^2 F_r^T S^T S F_r$. Since $\operatorname{kernel}(K^\star S) = \operatorname{kernel}(S)$ and $F_r \in \operatorname{kernel}^\perp(K^\star S)$, we know $F_r \in \operatorname{kernel}^\perp(S) = \operatorname{span}(S^T)$. 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

$$D_r \geq \sigma_{\min}(K^\star)\sigma_{\min}^+(S). \tag{E.5}$$

Combining all above derivations, we obtain

$$\begin{aligned}
\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star\| &\overset{\text{(E.1)}}{=} \|EE^T - E'(E')^T\| \overset{\text{(E.3)}}{\leq} 2\|\sin(\theta_p)\| \\
&\overset{\text{(E.4)}}{\leq} \frac{2\|K_t - K^\star\| \cdot \|S\|}{D_r} \overset{\text{(E.5)}}{\leq} \frac{2\|K_t - K^\star\|}{\sigma_{\min}(K^\star)} \cdot \frac{\|S\|}{\sigma_{\min}^+(S)}.
\end{aligned}$$

This completes the proof.

E.3 Proof of Corollary 5.4

Denote $A_t = I - \mathbb{E}[K_t S(S^T K_t^2 S)^\dagger S^T K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t]$ and $A^\star = I - \mathbb{E}[K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star]$. We have

$$\begin{aligned}
\|C_t - C^\star\| = \|A_t^\tau - (A^\star)^\tau\| &\le \left\|A_t^{\tau-1}(A_t - A^\star)\right\| + \left\|(A_t^{\tau-1} - (A^\star)^{\tau-1})A^\star\right\| \\
&\le \|A_t - A^\star\| + \left\|A_t^{\tau-1} - (A^\star)^{\tau-1}\right\| \quad (\|A_t\| \vee \|A^\star\| \le 1) \\
&\le \tau \|A_t - A^\star\| \le \tau \mathbb{E}\left[\left\|K_t S(S^T K_t^2 S)^\dagger S^T K_t - K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star\right\| \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t\right] \\
&\le \frac{2\tau \|K_t - K^\star\|}{\sigma_{\min}(K^\star)} \mathbb{E}\left[\|S\| \|S^\dagger\|\right] \le \frac{2\tau \Upsilon_S}{\sigma_{\min}(K^\star)} \|K_t - K^\star\| \quad \text{(by Assumption 5.3).}
\end{aligned}$$

This completes the proof.

E.4 Proof of Theorem 5.5

We present some lemmas that bound $\mathcal{I}_{1,t}$, $\mathcal{I}_{2,t}$, and $\mathcal{I}_{3,t}$ in (5.2a), (5.2b), and (5.2c), respectively. The proofs of these lemmas are presented in Appendix E.4.1 – E.4.4.

Lemma E.1 *Under Assumptions 4.1, 4.2(4.2a, 4.2e), 4.3, and $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, suppose*

$$\lim_{t\to\infty} t\,(1 - \varphi_{t-1}/\varphi_t) = \varphi < 0, \qquad \lim_{t\to\infty} t\varphi_t = \widetilde{\varphi} \in (0, \infty], \qquad 1.5(1-\rho^\tau) + \varphi/\widetilde{\varphi} > 0. \tag{E.6}$$

Then, for any $\upsilon > 0$,

$$\mathcal{I}_{1,t} = o\big(\sqrt{\varphi_t \{\log(1/\varphi_t)\}^{1+\upsilon}}\,\big) \quad \text{a.s.} \tag{E.7}$$

Furthermore, if (4.2a) is strengthened to (4.2b), then we have

(a): (asymptotic rate) $\mathcal{I}_{1,t} = O(\sqrt{\varphi_t \log(1/\varphi_t)})$ a.s.

(b): (asymptotic normality) $\sqrt{1/\varphi_t} \cdot \mathcal{I}_{1,t} \xrightarrow{d} \mathcal{N}(0, \Xi^\star)$ where $\Xi^\star$ is from (5.10).

(c): (Berry-Esseen bound) For any vector $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ such that $\boldsymbol{w}^T \Xi^\star \boldsymbol{w} \neq 0$,

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

Lemma E.2 *Under the conditions of Lemma E.1 with (4.2a) and assume for the adaptivity gap χ_t that for some $p, q \in (0, 1]$,*

$$\lim_{t\to\infty} t\,(1 - \chi_{t-1}/\chi_t) = \chi < \varphi, \;\; (1-\rho^\tau) + p(\chi - 0.5\varphi)/\widetilde{\varphi} > 0, \;\; (1-\rho^\tau) + q(\chi - \varphi)/\widetilde{\varphi} > 0. \tag{E.8}$$

Then, for any $\nu > 0$ (if $q < 1$, the second $O(\cdot)$ can be strengthened to $o(\cdot)$)

$$\mathcal{I}_{2,t} = o\big(\chi_t^p / \varphi_t^{0.5p} \sqrt{\{\log(1/\chi_t)\}^{1+\nu}}\,\big) + O(\chi_t^q/\varphi_t^q) \quad \text{a.s.}$$

Furthermore, if (4.2a) is strengthened to (4.2b), then we have (if $p < 1$, the first $O(\cdot)$ can also be strengthened to $o(\cdot)$)

$$\mathcal{I}_{2,t} = O\big(\chi_t^p / \varphi_t^{0.5p} \sqrt{\log(1/\chi_t)}\,\big) + O(\chi_t^q/\varphi_t^q) \quad \text{a.s.}$$

Lemma E.3 *Under the conditions of Lemma E.2 with (4.2a), we have for any $\upsilon > 0$,*

$$\mathcal{I}_{3,t} = o\big(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\upsilon}}\,\big) + o\big(\chi_t^p/\varphi_t^{0.5p}\sqrt{\{\log(1/\chi_t)\}^{1+\nu}}\big) + o(\chi_t^q/\varphi_t^q) = o(\mathcal{I}_{1,t}+\mathcal{I}_{2,t}) \quad a.s.$$

If (4.2a) is strengthened to (4.2b), the above result holds with $\upsilon = 0$.

We apply the above lemmas. We first check the conditions (E.6) and (E.8). Since $\varphi_t = (\beta_t+\eta_t)/2 = \beta_t+\chi_t/2$ and $\chi_t = o(\beta_t)$ (as implied by $\chi < \beta$), we know $\beta_t \le \varphi_t \le \beta_t + o(\beta_t)$ and $\lim_{t\to\infty} t\varphi_t = \lim_{t\to\infty} t\beta_t = \widetilde{\beta}$. Furthermore, we have

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

The above derivations show that $\varphi = \beta$ and $\widetilde{\varphi} = \widetilde{\beta}$. Thus, (5.6) implies (E.6) holds. Moreover, for any constant $p \in (0,1]$ such that $(1-\rho^\tau)+p(\chi-0.5\beta)/\widetilde{\beta} = (1-\rho^\tau)+p(\chi-0.5\varphi)/\widetilde{\varphi} > 0$, we simply let $q = p$ and have $(1-\rho^\tau)+q(\chi-\varphi)/\widetilde{\varphi} > 0$. Thus, (E.8) holds with $q = p$. We note that for any $\nu \ge 0$,

$$\begin{aligned}
&\lim_{t\to\infty} t\left(1-\frac{\varphi_{t-1}^{0.5p}\{\log(1/\chi_{t-1})\}^{0.5(1+\nu)}}{\varphi_t^{0.5p}\{\log(1/\chi_t)\}^{0.5(1+\nu)}}\right) \overset{\text{Lem. B.1}}{=} 0.5p\varphi + 0.5(1+\nu)\lim_{t\to\infty} t\left(1-\frac{\log(1/\chi_{t-1})}{\log(1/\chi_t)}\right)\\
&= 0.5p\varphi + 0.5(1+\nu)\lim_{t\to\infty} t\left(\frac{\log(\chi_{t-1}/\chi_t)}{\log(1/\chi_t)}\right) = 0.5p\varphi + 0.5(1+\nu)\lim_{t\to\infty} t\left(\frac{\frac{\chi_{t-1}-\chi_t}{\chi_t}+O\left(\frac{(\chi_{t-1}-\chi_t)^2}{\chi_t^2}\right)}{\log(1/\chi_t)}\right)\\
&= 0.5p\varphi - 0.5(1+\nu)\chi\lim_{t\to\infty} 1/\log(1/\chi_t) = 0.5p\varphi < 0 \quad (\text{by (E.6) and } \chi_t \to 0).
\end{aligned}$$

Thus, by Lemma B.2, we know $\chi_t^p/\varphi_t^{0.5p}\sqrt{\{\log(1/\chi_t)\}^{1+\nu}} = o(\chi_t^p/\varphi_t^p)$. The convergence rate of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ comes from Lemmas 5.1, E.1, E.2, E.3, the above fact, and the fact that $\beta_t \le \varphi_t \le 2\beta_t$. We complete the proof.

E.4.1 Proof of Lemma E.1

We need a preparation lemma. Recall that we suppose $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ is a local solution of (1.1) with $G^\star$ being full row rank and $\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star$ being positive definite in the null space $\{\boldsymbol{x} \in \mathbb{R}^d : G^\star\boldsymbol{x} = \boldsymbol{0}\}$.

Lemma E.4 *Under Assumptions 4.1, 4.2(4.2d) and $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, we have $\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i$ $\to \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star$ as $t \to \infty$. Further, with a small γ_{RH} and a large Υ_B, $\Delta_t = \boldsymbol{0}$ for all large enough t.*

Let $I + C^\star = U\Sigma U^T$ with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{d+m})$ be the eigenvalue decomposition. Then,

$$\mathcal{I}_{1,t} \overset{(5.2a)}{=} \sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j(I+C^\star)\}\varphi_i\boldsymbol{\theta}^i = U\sum_{i=0}^{t}\prod_{j=i+1}^{t}\{I-\varphi_j\Sigma\}\varphi_i U^T\boldsymbol{\theta}^i.$$

Since $\mathbb{E}[\boldsymbol{\theta}^i|\mathcal{F}_{i-1}] = \mathbf{0}$, we 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 inequality (Fan, 2019, Theorem 2.1) to show each result in the lemma. We compute the conditional covariance of $\mathcal{I}_{1,t}$, which is defined as (Duflo, 1997, Proposition 1.3.7)

$$\langle \mathcal{I}_1 \rangle_t \coloneqq U \sum_{i=0}^{t} \prod_{j=i+1}^{t} \{I - \varphi_j \Sigma\} \varphi_i^2 U^T \mathbb{E}[\boldsymbol{\theta}^i (\boldsymbol{\theta}^i)^T \mid \mathcal{F}_{i-1}] U \big(\prod_{j=i+1}^{t} \{I - \varphi_j \Sigma\} \big)^T U^T. \tag{E.9}$$

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

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta}^i(\boldsymbol{\theta}^i)^T \mid \mathcal{F}_{i-1}] &\overset{(5.3b)}{=} \mathbb{E}\big[\big\{(I + C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i) - \{\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i\}\big\} \\
&\qquad \big\{(I + C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i) - \{\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i\}\big\}^T \mid \mathcal{F}_{i-1}\big] \\
&\overset{(D.7)}{=} (I + C_i)K_i^{-1}\mathbb{E}\left[(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^T \mid \mathcal{F}_{i-1}\right] K_i^{-1}(I + C_i) \\
&\qquad + \mathbb{E}[\{\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i\}\{\boldsymbol{z}_{i,\tau} - (I + C_i)\widetilde{\boldsymbol{z}}_i\}^T \mid \mathcal{F}_{i-1}] =: \mathcal{J}_{1,i} + \mathcal{J}_{2,i}.
\end{aligned} \tag{E.10}$$

For the term $\mathcal{J}_{1,i}$, we apply Assumption 4.2 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 note that

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

and

$$\mathbb{E}\left[\|\bar{g}_i - \nabla f(\boldsymbol{x}^\star;\xi)\|^2 \mid \mathcal{F}_{i-1}\right] \le \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla^2 f(\boldsymbol{x};\xi)\|^2] \cdot \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2 \overset{(4.2e)}{\le} \Upsilon_m \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2.$$

By Assumptions 4.1, 4.2(4.2a), we suppose $\|\nabla f_i\| \le \Upsilon_u$ (we abuse Υ_u from (D.2)) and obtain

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

Combining the above three displays, we have

$$\begin{aligned}
&\left\|\mathbb{E}[\bar{g}_i\bar{g}_i^T - \nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi) \mid \mathcal{F}_{i-1}]\right\| \\
&\qquad \le 2\sqrt{2\Upsilon_m}(\Upsilon_u \vee \sqrt{\Upsilon_m})(\|\boldsymbol{x}_i - \boldsymbol{x}^\star\| + \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2) \to 0.
\end{aligned} \tag{E.11}$$

This implies that

$$\lim_{i\to\infty} \mathbb{E}[(\bar{g}_i - \nabla f_i)(\bar{g}_i - \nabla f_i)^T \mid \mathcal{F}_{i-1}] = \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] - \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star). \tag{E.12}$$

Furthermore, by Lemma E.4 we know $K_i \to K^\star$ as $i \to \infty$. Since $\|K_i S(S^T K_t^2 S)^\dagger S^T K_i\| \le 1$, we apply dominated convergence theorem (Durrett, 2019, Theorem 1.6.7) and Lemma 5.2, and have $\lim_{i\to\infty} \mathbb{E}[K_i S(S^T K_i^2 S)^\dagger S^T K_i \mid \boldsymbol{x}_i, \boldsymbol{\lambda}_i] = \mathbb{E}[K^\star S(S^T (K^\star)^2 S)^\dagger S^T K^\star]$. Here, the expectation is taken over randomness of S. Thus, $C_i \to C^\star$. By the definition (5.5), we obtain

$$\mathcal{J}_{1,i} = (I + C^\star)\Omega^\star(I + C^\star) + O(\mathcal{K}_{1,i}) \tag{E.13}$$

with $\mathcal{K}_{1,i} \to 0$ as $i \to \infty$ almost surely. For the term $\mathcal{J}_{2,i}$, we apply (D.1) and define $\widetilde{C}_i \coloneqq -\prod_{j=0}^{\tau-1} C_{i,j}$. Then, we have

$$\begin{aligned}
\mathcal{J}_{2,i} &= \mathbb{E}[(\widetilde{C}_i - C_i)\widetilde{\boldsymbol{z}}_i\widetilde{\boldsymbol{z}}_i^T(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \overset{(3.2)}{=} \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\bar{\nabla}\mathcal{L}_i\bar{\nabla}^T\mathcal{L}_iK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&= \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&\quad + \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i\nabla^T\mathcal{L}_iK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&\quad + \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)\nabla^T\mathcal{L}_iK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&\quad + \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}]. \qquad \text{(E.14)}
\end{aligned}$$

For the last two terms, we apply the tower property of conditional expectation by first conditioning on the randomness of $\{S_{i,j}\}_j$ to take expectation over the randomness of ξ_i, and then taking expectation over the randomness of $\{S_{i,j}\}_j$. In particular, we have (similar for the second last term in (E.14))

$$\begin{aligned}
&\mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&= \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i\mathbb{E}[\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i \mid \mathcal{F}_{i-1} \cup \sigma(\{S_{i,j}\}_j)]^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&= \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\nabla\mathcal{L}_i\mathbb{E}[\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i \mid \mathcal{F}_{i-1}]^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] = \mathbf{0}.
\end{aligned}$$

For the second term in (E.14), it converges to zero almost surely as $i \to \infty$ since $\|\widetilde{C}_i\| \vee \|C_i\| \leq 1$, $\|K_i^{-1}\| \leq \Upsilon_K$, and $\nabla\mathcal{L}_i \to 0$. For the first term in (E.14), we have

$$\begin{aligned}
&\mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^TK_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&= \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\mathbb{E}\left[(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^T \mid \mathcal{F}_{i-1} \cup \sigma(\{S_{i,j}\}_j)\right]K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&= \mathbb{E}[(\widetilde{C}_i - C_i)K_i^{-1}\mathbb{E}\left[(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)(\bar{\nabla}\mathcal{L}_i - \nabla\mathcal{L}_i)^T \mid \mathcal{F}_{i-1}\right]K_i^{-1}(\widetilde{C}_i^T - C_i^T) \mid \mathcal{F}_{i-1}] \\
&\overset{(5.4)}{\longrightarrow} \mathbb{E}[(\widetilde{C}^\star - C^\star)\Omega^\star((\widetilde{C}^\star)^T - C^\star)] = \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] - C^\star\Omega^\star C^\star.
\end{aligned}$$

Again, the convergence here is due to the dominated convergence theorem, (E.12), and $K_i \to K^\star$; 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 (E.14), we have

$$\mathcal{J}_{2,i} = \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] - C^\star\Omega^\star C^\star + O(\mathcal{K}_{2,i}) \qquad \text{(E.15)}$$

with $\mathcal{K}_{2,i} \to 0$ as $i \to \infty$ almost surely. Combining (E.15), (E.13), and (E.10), 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 (E.9), let us denote $\Gamma \coloneqq U^T\mathbb{E}[(I+\widetilde{C}^\star)\Omega^\star(I+\widetilde{C}^\star)^T]U$. For any $k, l \in \{1, \ldots, d+m\}$, the (k,l) entry of the matrix $U^T\langle\mathcal{I}_1\rangle_tU$ can be written as

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

where $r_{i,kl} \to 0$ as $i \to \infty$ almost surely. By Lemma 4.4(b) and the fact that $C_i \to C^\star$ as $i \to \infty$, we know $\|C^\star\| \leq \rho^\tau$. Since $C^\star \preceq \mathbf{0}$, we have $0 < 1 - \rho^\tau \leq \sigma_i \leq 1$ for $i = 1, \ldots, d+m$,

which implies $\sigma_k+\sigma_l \geq 2(1-\rho^\tau)$. Using the condition (E.6), Lemmas B.2 and B.3, we obtain $[U^T\langle \mathcal{I}_1\rangle_t U]_{k,l}/\varphi_t \to \Gamma_{kl}/(\sigma_k+\sigma_l+\varphi/\widetilde{\varphi})$ as $t\to\infty$ almost surely. Thus, by (5.10), we have

$$\langle \mathcal{I}_1\rangle_t/\varphi_t \xrightarrow{a.s.} U(\Theta\circ\Gamma)U^T = \Xi^\star. \tag{E.16}$$

Then, (Duflo, 1997, Theorem 1.3.15) indicates (E.7) holds. This shows the first part of the results. For the second part of the results, we assume the condition (4.2b) and have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\theta}^i\|^3 \mid \mathcal{F}_{i-1}] &\overset{(5.3b)}{\leq} 4\left(\mathbb{E}[\|(I+C_i)K_i^{-1}(\bar{\nabla}\mathcal{L}_i-\nabla\mathcal{L}_i)\|^3 \mid \mathcal{F}_{i-1}] + \mathbb{E}[\|\boldsymbol{z}_{i,\tau}-(I+C_i)\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right)\\
&\overset{(D.1)}{\leq} 4\left(8\Upsilon_K^3\mathbb{E}[\|\bar{g}_i-\nabla f_i\|^3 \mid \mathcal{F}_{i-1}] + \mathbb{E}[\|(\widetilde{C}_i-C_i)\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\|C_i\|\leq 1, \|K_i^{-1}\|\leq\Upsilon_K)\\
&\overset{(4.2b)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\mathbb{E}[\|\widetilde{\boldsymbol{z}}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\|\widetilde{C}_i\|\vee\|C_i\|\leq 1)\\
&\overset{(3.2)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\mathbb{E}[\|\bar{\nabla}\mathcal{L}_i\|^3 \mid \mathcal{F}_{i-1}]\right) \quad (\|K_i^{-1}\|\leq\Upsilon_K)\\
&\overset{(3.2)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\left\{4\|\nabla\mathcal{L}_i\|^3 + 4\mathbb{E}[\|\bar{g}_i-\nabla f_i\|^3 \mid \mathcal{F}_{i-1}]\right\}\right)\\
&\overset{(4.2b)}{\leq} 4\left(8\Upsilon_K^3\Upsilon_m + 8\Upsilon_K^3\left\{4\Upsilon_u^3 + 4\Upsilon_m\right\}\right) \quad \text{(also use (D.2))}.
\end{aligned} \tag{E.17}$$

Thus, $\boldsymbol{\theta}^i$ has bounded third moment; and (Wang, 1995, pp. 554) together with (E.16) give the result (a). For (b), we verify the Lindeberg's condition. For any $\epsilon>0$, we have

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

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

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

By (E.17) and $\mathbb{E}[|[U^T\boldsymbol{\theta}^i]_k|^3|\mathcal{F}_{i-1}] \leq \mathbb{E}[\|\boldsymbol{\theta}^i\|^3|\mathcal{F}_{i-1}]$, we only 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 $1-\varphi_j\sigma_k\geq 0$ for all $j\geq 1$ and show

$$\sum_{i=0}^t\prod_{j=i+1}^t(1-\varphi_j\sigma_k)^3\varphi_i^3 = o(\varphi_t^{3/2}). \tag{E.18}$$

Otherwise, since $\varphi<0$ from (E.6), Lemma B.2 shows that $\varphi_i\to 0$. Thus, there exists $\widetilde{t}$ such that $1-\varphi_j\sigma_k\geq 0$, $\forall j\geq\widetilde{t}$. Then,

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

$$= \prod_{j=\widetilde{t}}^{t}(1-\varphi_j\sigma_k)^3 \sum_{i=0}^{\widetilde{t}-2}\prod_{j=i+1}^{\widetilde{t}-1}\left|1-\varphi_j\sigma_k\right|^3\varphi_i^3 + \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3\varphi_i^3$$
$$= \sum_{i=\widetilde{t}-1}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^3(\varphi_i')^3, \tag{E.19}$$

where

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

Note that (E.19) has the same form as (E.18), and φ_i' differs from φ_i only at $i = \widetilde{t}-1$. Thus, (E.19) and (E.18) have the same limit. For (E.18), we apply Lemma B.1 and observe that

$$\lim_{i\to\infty} i\left(1-\varphi_{i-1}^2/\varphi_i^2\right) \overset{\text{(E.6)}}{=} 2\varphi \qquad \text{and} \qquad 3\sigma_k + 2\varphi/\widetilde{\varphi} \overset{\text{(E.6)}}{>} 0.$$

Thus, Lemma B.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 (E.18) and further verifies the 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 the Lindeberg's condition above, and obtain the result immediately. This completes the proof.

E.4.2 Proof of Lemma E.2

We have

$$\mathcal{I}_{2,t} \overset{\text{(5.2b)}}{=} \sum_{i=0}^{t}\prod_{j=i+1}^{t}\left\{I-\varphi_j(I+C^\star)\right\}(\bar{\alpha}_i-\varphi_i)\boldsymbol{z}_{i,\tau} = U\sum_{i=0}^{t}\prod_{j=i+1}^{t}\left\{I-\varphi_j\Sigma\right\}(\bar{\alpha}_i-\varphi_i)U^T\boldsymbol{z}_{i,\tau}.$$

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 (E.18) and (E.19), we suppose for any $j \geq 0$ that $1-\varphi_j\sigma_k \geq 0$. Then,

$$\left|[U^T\mathcal{I}_{2,t}]_k\right| \leq \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}|1-\varphi_j\sigma_k|\chi_i\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| = \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|$$
$$= \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| \mid \mathcal{F}_{i-1}\right] + \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\Big\{\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|$$
$$- \mathbb{E}\left[\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right| \mid \mathcal{F}_{i-1}\right]\Big\} =: \mathcal{J}_{3,t,k} + \mathcal{J}_{4,t,k}. \tag{E.20}$$

We analyze $\mathcal{J}_{3,t,k}$ and $\mathcal{J}_{4,t,k}$ separately as follows. We first show $\left|[U^T\boldsymbol{z}_{i,\tau}]_k\right|$ has bounded variance. We have

$$\mathbb{E}\big[\big\{\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| - \mathbb{E}\big[\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| \mid \mathcal{F}_{i-1}\big]\big\}^2 \mid \mathcal{F}_{i-1}\big]$$
$$\leq \mathbb{E}\big[\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big|^2 \mid \mathcal{F}_{i-1}\big] \leq \mathbb{E}\big[\|\boldsymbol{z}_{i,\tau}\|^2 \mid \mathcal{F}_{i-1}\big] \overset{\text{(D.10)}}{\leq} 16\Upsilon_K^2(\Upsilon_u^2 \vee \Upsilon_m). \quad \text{(E.21)}$$

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

$$\langle \mathcal{J}_{4,k}\rangle_t \coloneqq \frac{1}{4}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^2\chi_i^2\mathbb{E}\left[\big\{\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| - \mathbb{E}\big[\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| \mid \mathcal{F}_{i-1}\big]\big\}^2 \mid \mathcal{F}_{i-1}\right]$$
$$\overset{\text{(E.21)}}{\leq} 4\Upsilon_K^2(\Upsilon_u^2 \vee \Upsilon_m)\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)^2\chi_i^2.$$

Using (E.6) and (E.8), we know

$$\lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^2/\varphi_{i-1}}{\chi_i^2/\varphi_i}\right) = \lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^2}{\chi_i^2} + \frac{\chi_{i-1}^2}{\chi_i^2}\left(1 - \frac{\varphi_i}{\varphi_{i-1}}\right)\right)$$
$$= \lim_{i\to\infty} i\left\{\left(1-\frac{\chi_{i-1}}{\chi_i}\right)\left(1+\frac{\chi_{i-1}}{\chi_i}\right) - \frac{\chi_{i-1}^2}{\chi_i^2}\frac{\varphi_i}{\varphi_{i-1}}\left(1-\frac{\varphi_{i-1}}{\varphi_i}\right)\right\} = 2\chi - \varphi. \quad \text{(E.22)}$$

Further, (E.8) implies $2\sigma_k + p(2\chi-\varphi)/\widetilde{\varphi} > 0$ for some constant $p \in (0,1]$. Thus, Lemma B.3 leads to $\langle \mathcal{J}_{4,k}\rangle_t = O(\chi_t^{2p}/\varphi_t^p)$ (when $p \in (0,1)$, $O(\cdot)$ can be strengthened to $o(\cdot)$); and the strong law of large number (Duflo, 1997, Theorem 1.3.15) suggests that for any $\nu > 0$,

$$\mathcal{J}_{4,t,k} = o\left(\sqrt{\chi_t^{2p}/\varphi_t^p \cdot \{\log(\varphi_t^p/\chi_t^{2p})\}^{1+\nu}}\right) = o\left(\sqrt{\chi_t^{2p}/\varphi_t^p \cdot \{\log(1/\chi_t)\}^{1+\nu}}\right).$$

If (4.2a) is strengthened to (4.2b), then we follow (E.21), (D.10), and (E.17), and can show $\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| - \mathbb{E}\big[\big|[U^T\boldsymbol{z}_{i,\tau}]_k\big| \mid \mathcal{F}_{i-1}\big]$ has bounded third moment. Thus, (Wang, 1995, pp. 554) suggests that $\mathcal{J}_{4,t,k} = O(\chi_t^p/\varphi_t^{0.5p}\sqrt{\log(1/\chi_t)})$. When $p \in (0,1)$, $O(\cdot)$ can be strengthened to $o(\cdot)$ due to $\langle \mathcal{J}_{4,k}\rangle_t = o(\chi_t^{2p}/\varphi_t^p)$. For the term $\mathcal{J}_{3,t,k}$, we have

$$\mathcal{J}_{3,t,k} \leq \frac{1}{2}\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i\sqrt{\mathbb{E}[|[U^T\boldsymbol{z}_{i,\tau}]_k|^2 \mid \mathcal{F}_{i-1}]}$$
$$\overset{\text{(E.21)}}{\leq} 2\Upsilon_K(\Upsilon_u \vee \sqrt{\Upsilon_m})\sum_{i=0}^{t}\prod_{j=i+1}^{t}(1-\varphi_j\sigma_k)\chi_i.$$

Using (E.6), (E.8), and the facts that $\lim_{i\to\infty} i(1-\frac{\chi_{i-1}/\varphi_{i-1}}{\chi_i/\varphi_i}) = \chi-\varphi$ and $\sigma_k + q(\chi-\varphi)/\widetilde{\varphi} > 0$ (as implied by (E.8)), we apply Lemma B.3 and obtain $\mathcal{J}_{3,t,k} = O(\chi_t^q/\varphi_t^q)$. When $q \in (0,1)$, $O(\cdot)$ can be strengthened to $o(\cdot)$. Combining with (E.20) and the bound of $\mathcal{J}_{4,t,k}$, we complete the proof.

E.4.3 Proof of Lemma E.3

Based on the definition of $\mathcal{I}_{3,t}$ in (5.2c), we have the recursion

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

By Assumption 4.1 and the fact that $\|C_t\| \le 1$, we have

$$\begin{aligned}\left\|\boldsymbol{\delta}^t\right\| &\overset{(5.3c)}{\le} 2\left(\|(K^\star)^{-1}\|\|\boldsymbol{\psi}^t\| + \|K_t^{-1} - (K^\star)^{-1}\| \cdot \|\nabla \mathcal{L}_t\|\right) + \|C_t - C^\star\| \cdot \left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| \\ &\le 2\Upsilon_K \Upsilon_L \left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|^2 + \left(2\Upsilon_K^2 \Upsilon_u \|K_t - K^\star\| + \|C_t - C^\star\|\right)\left\|\begin{pmatrix}\boldsymbol{x}_t - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|. \quad \text{(E.24)}\end{aligned}$$

Since $K_t \to K^\star$ (cf. Lemma E.4) and $C_t \to C^\star$, we know

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

Using $\|C^\star\| \le \rho^\tau$, we know for any $a \in (0,1)$, there exists an integer t_1 such that for any $t \ge t_1$,

$$\begin{aligned}\|\mathcal{I}_{3,t+1}\| &\le \{1 - \varphi_{t+1}(1-\rho^\tau)\}\|\mathcal{I}_{3,t}\| + \varphi_{t+1} \cdot o\left(\|(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^\star)\|\right) \\ &\le \{1 - \varphi_{t+1}(1-\rho^\tau) + o(\varphi_{t+1})\}\|\mathcal{I}_{3,t}\| + \varphi_{t+1} \cdot o(\|\mathcal{I}_{1,t}\| + \|\mathcal{I}_{2,t}\|) \quad \text{(by Lemma 5.1)} \\ &\le \{1 - a(1-\rho^\tau)\varphi_{t+1}\}\|\mathcal{I}_{3,t}\| + \varphi_{t+1} \cdot o(\|\mathcal{I}_{1,t}\| + \|\mathcal{I}_{2,t}\|).\end{aligned}$$

We apply the above inequality recursively and obtain

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

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

$$\begin{aligned}\lim_{i\to\infty} i\left(1 - \frac{\sqrt{\varphi_{i-1}\{\log(1/\varphi_{i-1})\}^{1+\upsilon}}}{\sqrt{\varphi_i\{\log(1/\varphi_i)\}^{1+\upsilon}}}\right) &\overset{(E.6)}{=} \lim_{i\to\infty} i\left(1 - \frac{\sqrt{\varphi_{i-1}}}{\sqrt{\varphi_i}}\right) + \lim_{i\to\infty} i\left(1 - \frac{\{\log(1/\varphi_{i-1})\}^{\frac{1+\upsilon}{2}}}{\{\log(1/\varphi_i)\}^{\frac{1+\upsilon}{2}}}\right) \\ &\overset{(E.6)}{=} \frac{\varphi}{2} + \lim_{i\to\infty} i\left(1 - \frac{\{\log(1/\varphi_{i-1})\}^{\frac{1+\upsilon}{2}}}{\{\log(1/\varphi_i)\}^{\frac{1+\upsilon}{2}}}\right) \quad \text{(Lemma B.1)}.\end{aligned}$$

Furthermore, we have

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

where the last equality is due to $\varphi_i \to 0$, as implied by Lemma B.2. Combining the above two displays with Lemma B.1, we have

$$\lim_{i\to\infty} i\left(1 - \frac{\sqrt{\varphi_{i-1}\{\log(1/\varphi_{i-1})\}^{1+\upsilon}}}{\sqrt{\varphi_i\{\log(1/\varphi_i)\}^{1+\upsilon}}}\right) = \frac{\varphi}{2} \qquad \text{for any} \quad \nu \ge 0. \quad \text{(E.27)}$$

Moreover, we have for any $p, q \in (0, 1]$ and $\nu \geq 0$,

$$\begin{aligned} \lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^p/\varphi_{i-1}^{0.5p}\sqrt{\{\log(1/\chi_{i-1})\}^{1+\nu}}}{\chi_i^p/\varphi_i^{0.5p}\sqrt{\{\log(1/\chi_i)\}^{1+\nu}}}\right) &\overset{\text{(E.27)}}{\overset{\text{(E.22)}}{=}} p(\chi - 0.5\varphi), \\ \lim_{i\to\infty} i\left(1 - \frac{\chi_{i-1}^q/\varphi_{i-1}^q}{\chi_i^q/\varphi_i^q}\right) &\overset{\text{(E.22)}}{=} q(\chi - \varphi). \end{aligned} \tag{E.28}$$

For the constants p, q in (E.8), we let a be any scalar such that

$$0 < \frac{-\varphi/\widetilde{\varphi}}{2(1-\rho^\tau)} \vee \frac{-p(\chi - 0.5\varphi)/\widetilde{\varphi}}{1-\rho^\tau} \vee \frac{-q(\chi-\varphi)/\widetilde{\varphi}}{1-\rho^\tau} < a < 1,$$

which is guaranteed to exist due to (E.6) and (E.8). Then, we obtain

$$a(1-\rho^\tau) + \frac{\varphi}{2\widetilde{\varphi}} > 0 \quad a(1-\rho^\tau) + \frac{p(\chi-0.5\varphi)}{\widetilde{\varphi}} > 0 \quad \text{and} \quad a(1-\rho^\tau) + \frac{q(\chi-\varphi)}{\widetilde{\varphi}} > 0.$$

Thus, combining (E.26), (E.27), and (E.28) with Lemma B.3, we obtain the results under either (4.2a) or (4.2b). This completes the proof.

E.4.4 Proof of Lemma E.4

We note that

$$\begin{aligned} \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\right\| &\leq \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| + \frac{1}{t}\sum_{i=0}^{t-1}\left\|\nabla_{\boldsymbol{x}}^2\mathcal{L}_i - \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\right\| \\ &\overset{(4.1)}{\leq} \left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| + \frac{\Upsilon_L}{t}\sum_{i=0}^{t-1}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|. \end{aligned} \tag{E.29}$$

Since $(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \to \mathbf{0}$, by the fact that $a_t \to a$ implies $\frac{1}{t}\sum_{i=0}^{t-1} a_i \to a$ (also known as Stolz–Cesàro theorem), it suffices to show $(\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i)/t$ converges to zero. In fact, by Assumption 4.2(4.2d) 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 notice $(\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i)/t$ is a square integrable martingale. Thus, (Duflo, 1997, Theorem 1.3.15) suggests that for any $\upsilon > 0$,

$$\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{H}_i - \nabla^2 f_i\right\| = o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right). \tag{E.30}$$

Combining (E.29) and (E.30), we obtain $\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i \to \nabla_{\boldsymbol{x}}^2\mathcal{L}^\star$ as $t \to \infty$. For the second result, we suppose $\|\nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\| \leq \Upsilon_B^\star$ and $\boldsymbol{x}^T\nabla_{\boldsymbol{x}}^2\mathcal{L}^\star\boldsymbol{x} \geq \gamma_{RH}^\star\|\boldsymbol{x}\|^2$ in the space $\{\boldsymbol{x} \in \mathbb{R}^d : G^\star\boldsymbol{x} = \mathbf{0}\}$. Whenever $\gamma_{RH} < \gamma_{RH}^\star$ and $\Upsilon_B > \Upsilon_B^\star$, we know $\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i\| \leq \Upsilon_B$ for large enough t. In addition, we let $Z_t, Z^\star \in \mathbb{R}^{d\times(d-m)}$ be the matrices whose columns are orthonormal and span the spaces of $\ker(G_t)$, $\ker(G^\star)$, respectively. Since $G_t \to G^\star$, Davis-Kahan $\sin(\theta)$ theorem suggests that $Z_tZ_t^T \to Z^\star(Z^\star)^T$, implying $\inf_Q\|Z_t - Z^\star Q\| \to 0$ with Q chosen over all $(d-m)\times(d-m)$ orthogonal matrices (Davis and Kahan, 1970). Thus, we have

$$\lambda_{\min}(Z_t^T(\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i/t)Z_t) = \lambda_{\min}(QZ_t^T(\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i/t)Z_tQ^T) \to \lambda_{\min}((Z^\star)^T\nabla_{\boldsymbol{x}}^2\mathcal{L}^\star Z^\star),$$

which implies $\lambda_{\min}(Z_t^T(\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i/t)Z_t) \geq \gamma_{RH}$ for large enough t. This completes the proof.

E.5 Proof of Theorem 5.6

We first improve the rate of $\mathcal{I}_{3,t}$. Lemma E.5 differs from Lemma E.3 in the bound of $\boldsymbol{\delta}^t$. We propose a more precise bound on $\boldsymbol{\delta}^t$ compared to (E.25). The new bound relies on the convergence rate of the Hessian K_t in Lemma E.6 and Assumption 5.3 for applying Corollary 5.4.

Lemma E.5 *Under the conditions of Theorem 5.6, for any $\upsilon > 0$,*

$$\mathcal{I}_{3,t} = o\big(\{\varphi_t \log(1/\varphi_t)\}^{2/3}\{\log(1/\varphi_t)\}^{\nu}\big) \quad a.s.$$

We note that

$$|\sqrt{1/\bar{\alpha}_t} - \sqrt{1/\varphi_t}| = |\sqrt{\bar{\alpha}_t} - \sqrt{\varphi_t}|/\sqrt{\bar{\alpha}_t\varphi_t} = |\bar{\alpha}_t - \varphi_t|/(\sqrt{\bar{\alpha}_t\varphi_t}(\sqrt{\bar{\alpha}_t} + \sqrt{\varphi_t})) \leq \chi_t/(4\beta_t^{1.5}).$$

Since $\chi < 1.5\beta$, we know $\chi_t = o(\beta_t^{1.5})$. By the almost sure convergence of $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, we only need to show the normality of $1/\sqrt{\varphi_t}(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)$. Let us choose $p \in (0, 1]$ such that

$$\begin{aligned} p\chi - 0.5p\varphi - 2\varphi/3 < 0 &\Longleftrightarrow p > \frac{2\varphi/3}{\chi - 0.5\varphi}, \\ (1-\rho^\tau) + p(\chi - 0.5\varphi)/\widetilde{\varphi} > 0 &\Longleftrightarrow p < \frac{(1-\rho^\tau)\widetilde{\varphi}}{0.5\varphi - \chi}, \end{aligned} \tag{E.31}$$

which is guaranteed to exist due to the fact that

$$0 < \frac{2\varphi/3}{\chi - 0.5\varphi} < 1 \wedge \frac{(1-\rho^\tau)\widetilde{\varphi}}{0.5\varphi - \chi}.$$

We also choose q as stated in the theorem, which guarantees that (by the proof of Theorem 5.5, $\varphi = \beta$ and $\widetilde{\varphi} = \widetilde{\beta}$)

$$q\chi - q\varphi - 0.5\varphi < 0 \Leftrightarrow q > \frac{0.5\varphi}{\chi - \varphi} \quad \text{and} \quad (1-\rho^\tau) + q(\chi - \varphi)/\widetilde{\varphi} > 0 \Leftrightarrow q < \frac{(1-\rho^\tau)\widetilde{\varphi}}{\varphi - \chi}. \tag{E.32}$$

With the above choices of p and q, we know from (E.28), Lemmas E.2 and E.5 that for any $\nu > 0$

$$\mathcal{I}_{2,t} + \mathcal{I}_{3,t} = O(\chi_t^q/\varphi_t^q) + o\big(\{\varphi_t \log(1/\varphi_t)\}^{2/3}\{\log(1/\varphi_t)\}^{\nu}\big).$$

The first $O(\cdot)$ can be strengthened to $o(\cdot)$ when $q < 1$. Noting that $1/\sqrt{\varphi_t}(\mathcal{I}_{2,t} + \mathcal{I}_{3,t}) = o(1)$ a.s., the Slutsky's theorem together with Lemma E.1 leads to the asymptotic normality. Furthermore, Lemma B.5 with

$$A_t = \frac{\sqrt{1/\varphi_t}\cdot \boldsymbol{w}^T\mathcal{I}_{1,t}}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}}, \quad B_t = \frac{\sqrt{1/\varphi_t}\cdot \boldsymbol{w}^T(\mathcal{I}_{2,t} + \mathcal{I}_{3,t})}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}} + \frac{(\sqrt{1/\bar{\alpha}_t} - \sqrt{1/\varphi_t})\cdot \boldsymbol{w}^T(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^T\Xi^\star\boldsymbol{w}}}$$

and $C_t = 0$ leads to the Berry-Esseen bound. This completes the proof.

E.5.1 Proof of Lemma E.5

We need the following lemma to establish the convergence rate of K_t. The conditions are the same as those for showing the convergence rate of $(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)$, which are weaker than Theorem 5.6. The proof is provided in Appendix E.5.2.

Lemma E.6 *Under Assumptions 4.1, 4.2(4.2a, 4.2e), 4.3 and suppose $\{\beta_t, \chi_t\}_t$ satisfy (5.6). Then, for any $\nu > 0$ and any constants $p, q \in (0, 1]$ such that $(1-\rho^\tau)+p(\chi-0.5\beta)/\widetilde{\beta} > 0$ and $(1-\rho^\tau)+q(\chi-\beta)/\widetilde{\beta} > 0$, we have*

$$\|K_t - K^\star\| = o\big(\sqrt{\beta_t\{\log(1/\beta_t)\}^{1+\upsilon}}\,\big) + o\big(\chi_t^p/\beta_t^{0.5p}\sqrt{\{\log(1/\chi_t)\}^{1+\nu}}\,\big) + O(\chi_t^q/\beta_t^q) \ \ a.s.$$

Furthermore, if (4.2a) is strengthened to (4.2b), then

$$\|K_t - K^\star\| = O\big(\sqrt{\beta_t \log(1/\beta_t)}\,\big) + O\big(\chi_t^p/\beta_t^{0.5p}\sqrt{\log(1/\chi_t)}\,\big) + O(\chi_t^q/\beta_t^q) + o\big(\sqrt{(\log t)^{1+\upsilon}/t}\,\big) \ a.s.$$

If $p < 1$ (and/or $q < 1$), the second (and/or third) $O(\cdot)$ in the above results can be strengthened to $o(\cdot)$.

Applying Lemma E.6 with p, q chosen to satisfy (E.31) and (E.32), we know for any $\nu > 0$

$$\|K_t - K^\star\| = O\big(\sqrt{\varphi_t \log(1/\varphi_t)}\,\big) + o\big(\sqrt{(\log t)^{1+\upsilon}/t}\,\big).$$

Combining the above result with (E.24), Lemmas 5.1, E.1, E.2, E.3, and Corollary 5.4, we have

$$\|\boldsymbol{\delta}^t\| = O(\varphi_t \log(1/\varphi_t)) + o\big(\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \sqrt{(\log t)^{1+\upsilon}/t}\,\big). \tag{E.33}$$

We plug the above bound into the recursion (E.23). We note that

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

In the following proof, we consider φ such that $0.5\varphi - 0.5 \geq \varphi$. Otherwise, the second term in (E.33) is absorbed into the first term. Applying Lemma B.3 and noting that

$$1.5(1-\rho^\tau) + (0.5\varphi - 0.5)/\widetilde{\varphi} \geq 1.5(1-\rho^\tau) + \varphi/\widetilde{\varphi} > 0,$$

we know

$$\begin{aligned}
\mathcal{I}_{3,t} &= o\big(\{\varphi_t \log(1/\varphi_t)\}^{2/3}\,\big) + o\big(\{\sqrt{\varphi_t \log(1/\varphi_t)} \cdot \sqrt{(\log t)^{1+\upsilon}/t}\}^{2/3}\big) \\
&= o\big(\{\varphi_t \log(1/\varphi_t)\}^{2/3}\{\log(1/\varphi_t)\}^{\nu}\big),
\end{aligned}$$

where the second equality is due to $\sqrt{(\log t)^{1+\upsilon}/t} = O\big(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\nu}}\big)$ as implied by the fact that $t\varphi_t \to \widetilde{\varphi} \in (0, \infty]$. This completes the proof.

E.5.2 Proof of Lemma E.6

We note from the proof of Theorem 5.5 that $\varphi = \beta$ and $\widetilde{\varphi} = \widetilde{\beta}$. By Lemmas 5.1, E.1, E.2, E.3, for any $\nu > 0$, if (4.2a) holds, then we have

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = o\big(\sqrt{\varphi_t \{\log(1/\varphi_t)\}^{1+\nu}}\big) + o\big(\chi_t^p/\varphi_t^{0.5p}\sqrt{\{\log(1/\chi_t)\}^{1+\nu}}\,\big) + O(\chi_t^q/\varphi_t^q). \quad \text{(E.34)}$$

Here, $p, q \in (0, 1]$ are any constants such that $(1 - \rho^\tau) + p(\chi - 0.5\varphi)/\widetilde{\varphi} > 0$ and $(1 - \rho^\tau) + q(\chi - \varphi)/\widetilde{\varphi} > 0$. Furthermore, if $q < 1$ the $O(\cdot)$ in the third term can be strengthened to $o(\cdot)$. In the following proof, we only consider p, q such that

$$p(\chi - 0.5\varphi) \geq 0.5\varphi \qquad \text{and} \qquad q(\chi - \varphi) \geq 0.5\varphi. \quad \text{(E.35)}$$

Otherwise, by (E.27), (E.28), Lemmas B.1 and B.2, we know the second (and/or third) term in (E.34) can be absorbed into the first term. We now apply (4.1), combine (E.29) and (E.30), and have for any $\upsilon > 0$ and large enough t that

$$\begin{aligned}
\|K_t - K^\star\| &\leq \frac{\Upsilon_L}{t}\sum_{i=0}^{t-1}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right)\\
&= \frac{\Upsilon_L}{t}\left\|\begin{pmatrix}\boldsymbol{x}_0 - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \frac{1}{j}\right)\frac{1}{i}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right)\\
&= \Upsilon_L\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \frac{1}{j}\right)\frac{1}{i}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\| + \Upsilon_L\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + o\left(\sqrt{\frac{(\log t)^{1+\upsilon}}{t}}\right). \quad \text{(E.36)}
\end{aligned}$$

We claim that $\varphi = \beta > -2$. Otherwise, $\varphi + 1.5 \leq -0.5 < 0$. We apply Lemma B.1 and have

$$\lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}(t-1)^{1.5}}{\varphi_t t^{1.5}}\right) = \lim_{t\to\infty} t\left(1 - \frac{\varphi_{t-1}}{\varphi_t} + \frac{\varphi_{t-1}}{\varphi_t}\left(1 - \frac{(t-1)^{1.5}}{t^{1.5}}\right)\right) \overset{\text{(E.6)}}{=} \varphi + 1.5 < 0.$$

Then, Lemma B.2 suggests $\varphi_t t^{1.5} \to 0$, which cannot hold under (5.6). Thus, $\varphi > -2$. Using (E.27), (E.28), (E.35), and Lemma B.3, and noting that $1 + p(\chi - 0.5\varphi) \geq 1 + 0.5\varphi > 0$ and $1 + q(\chi - \varphi) \geq 1 + 0.5\varphi > 0$, we obtain

$$\begin{aligned}
&\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \frac{1}{j}\right)\frac{1}{i}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|\\
&\qquad = o\big(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\nu}}\big) + o\big(\chi_t^p/\varphi_t^{0.5p}\sqrt{\{\log(1/\chi_t)\}^{1+\nu}}\,\big) + O(\chi_t^q/\varphi_t^q). \quad \text{(E.37)}
\end{aligned}$$

Following the same derivation, we know that if (4.2b) holds, then

$$\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\| = O\big(\sqrt{\varphi_t \log(1/\varphi_t)}\big) + O\big(\chi_t^p/\varphi_t^{0.5p}\sqrt{\log(1/\chi_t)}\,\big) + O(\chi_t^q/\varphi_t^q)$$

and

$$\begin{aligned}
&\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \frac{1}{j}\right)\frac{1}{i}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^\star\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star\end{pmatrix}\right\|\\
&\qquad = O\big(\sqrt{\varphi_t \log(1/\varphi_t)}\big) + O\big(\chi_t^p/\varphi_t^{0.5p}\sqrt{\log(1/\chi_t)}\,\big) + O(\chi_t^q/\varphi_t^q). \quad \text{(E.38)}
\end{aligned}$$

Combining (E.36), (E.37), (E.38) together, and noting that $\beta_t \leq \varphi_t \leq 2\beta_t$ and $o(\sqrt{(\log t)^{1+\upsilon}/t})$ can be absorbed into $o\big(\sqrt{\varphi_t\{\log(1/\varphi_t)\}^{1+\nu}}\big)$ under (4.2a) (as implied by the fact that $t\varphi_t \to \widetilde{\varphi} \in (0, \infty]$), we complete the proof.

E.6 Proof of Corollary 5.7

The result in (a) is immediate by plugging $\beta = -1$ and $\widetilde{\beta} = 1$ into (5.11). For (b), we plug $\beta = -1$ and $\widetilde{\beta} = 1$ into (5.8), and know that $\Xi^\star$ solves the equation

$$(0.5I + C^\star)\Xi^\star + \Xi^\star(0.5I + C^\star) = \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T].$$

Thus, we have

$$\begin{aligned}&(0.5I + C^\star)(\Xi^\star - \Omega^\star) + (\Xi^\star - \Omega^\star)(0.5I + C^\star)\\ &\qquad = \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T] - \Omega^\star - C^\star\Omega^\star - \Omega^\star C^\star = \mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T] \succeq \mathbf{0}.\end{aligned}$$

Since $\|C^\star\| \le \rho^\tau < 0.5$, the basic Lyapunov theorem (cf. (Khalil, 2002, Theorem 4.6)) suggests that $\Xi^\star \succeq \Omega^\star$. Furthermore, with the notation in (5.10), we know

$$\Xi^\star - \Omega^\star = U\big(\Theta \circ U^T\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\, U\big)U^T \quad \text{with} \quad [\Theta]_{k,l} = 1/(\sigma_k + \sigma_l - 1). \tag{E.39}$$

The matrix Θ is positive semidefinite since for any vector $\boldsymbol{\omega}$,

$$\begin{aligned}\boldsymbol{\omega}^T\Theta\boldsymbol{\omega} &= \sum_{k,l=1}^{d+m} \frac{\boldsymbol{\omega}_k\boldsymbol{\omega}_l}{\sigma_k + \sigma_l - 1} = \sum_{k,l=1}^{d+m} \boldsymbol{\omega}_k\boldsymbol{\omega}_l \int_0^\infty \exp(-s(\sigma_k + \sigma_l - 1))ds \quad (\text{since } \sigma_k + \sigma_l - 1 > 0)\\ &= \int_0^\infty \left(\sum_{k=1}^{d+m} \boldsymbol{\omega}_k \exp(-s(\sigma_k - 0.5))\right)^2 ds \ge 0.\end{aligned}$$

By (Horn and Johnson, 1985, 7.5.P24), we have

$$\begin{aligned}\|\Xi^\star - \Omega^\star\| &\overset{(E.39)}{=} \|\Theta \circ U^T\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\, U\|\\ &\le \max_k [\Theta]_{k,k}\|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\| \le \frac{1}{1 - 2\rho^\tau}\|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\| \le 3\|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\|,\end{aligned}$$

where the second last inequality is due to $\sigma_k \ge 1 - \rho^\tau$ and the last inequality is due to $\rho^\tau < 1/3$. Furthermore, we have

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

$$\preceq \rho\|\Omega^\star\| \cdot \mathbb{E}\left[\left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}\left\{\prod_{j=2}^{\tau}(I - K^\star S_j(S_j^T(K^\star)S)^\dagger S_j^T K^\star)\right\}^T\right]$$
$$\preceq \rho^\tau\|\Omega^\star\| \cdot I, \tag{E.40}$$

where the second last inequality is from Assumption 4.3 and $K_t \to K^\star$; and the last inequality applies the same reason for sketch matrices $S_{2:\tau}$. Combining the above two displays completes the proof.

E.7 Proof of Theorem 5.10

We have

$$\|\Xi_t - \Xi^\star\| \le \|\Xi^\star - \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]/(2 + \beta/\widetilde{\beta})\|$$
$$+ \|\mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T] - \Omega^\star\|/(2 + \beta/\widetilde{\beta}) + \|\Omega^\star - \Omega_t\|/(2 + \beta/\widetilde{\beta}). \tag{E.41}$$

For the first term in (E.41), we have (Horn and Johnson, 1985, 7.7.P27)

$$\begin{aligned}
&\|\Xi^\star - \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]/(2 + \beta/\widetilde{\beta})\| \\
&\overset{(5.10)}{=} \|(\Theta - \mathbf{1}\mathbf{1}^T/(2 + \beta/\widetilde{\beta})) \circ U^T\mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]U\| \\
&\le \|\Theta - \mathbf{1}\mathbf{1}^T/(2 + \beta/\widetilde{\beta})\| \cdot \|\mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]\| \quad (\|A \circ B\| \le \|A\| \cdot \|B\|) \\
&\le 4\|\Theta - \mathbf{1}\mathbf{1}^T/(2 + \beta/\widetilde{\beta})\| \cdot \|\Omega^\star\| \quad (\|\widetilde{C}^\star\| \le 1),
\end{aligned}$$

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

$$|\Theta_{k,l} - 1/(2 + \beta/\widetilde{\beta})| = |1/(\sigma_k + \sigma_l + \beta/\widetilde{\beta}) - 1/(2 + \beta/\widetilde{\beta})| = \frac{|2 - \sigma_k - \sigma_l|}{(\sigma_k + \sigma_l + \beta/\widetilde{\beta})(2 + \beta/\widetilde{\beta})}$$
$$\le \frac{2\rho^\tau}{(2 - 2\rho^\tau + \beta/\widetilde{\beta})(2 + \beta/\widetilde{\beta})} \overset{(5.6)}{\le} \frac{2\rho^\tau}{(2 - 2(1 + \beta/(1.5\widetilde{\beta})) + \beta/\widetilde{\beta})(2 + \beta/\widetilde{\beta})} = \frac{6\rho^\tau}{-\beta/\widetilde{\beta}(2 + \beta/\widetilde{\beta})}.$$

Therefore, the above two displays lead to

$$\|\Xi^\star - \mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T]/(2 + \beta/\widetilde{\beta})\| = O(\rho^\tau). \tag{E.42}$$

For the second term in (E.41), we have

$$\|\mathbb{E}[(I + \widetilde{C}^\star)\Omega^\star(I + \widetilde{C}^\star)^T] - \Omega^\star\| \le \|C^\star\Omega^\star\| + \|\Omega^\star C^\star\| + \|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\| \quad (\mathbb{E}[\widetilde{C}^\star] = C^\star)$$
$$\le 2\|\Omega^\star\|\rho^\tau + \|\mathbb{E}[\widetilde{C}^\star\Omega^\star(\widetilde{C}^\star)^T]\| \overset{(E.40)}{=} O(\rho^\tau). \tag{E.43}$$

For the third term in (E.41), we have

$$\|\Omega_t - \Omega^\star\| \overset{(5.5)}{=} O(\|K_t - K^\star\|)$$
$$+ O\left(\left\|\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\bar{g}_i^T - \left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)\left(\frac{1}{t}\sum_{i=0}^{t-1}\bar{g}_i\right)^T - \mathbb{E}[\nabla f(\boldsymbol{x}^\star;\xi)\nabla^T f(\boldsymbol{x}^\star;\xi)] + \nabla f(\boldsymbol{x}^\star)\nabla^T f(\boldsymbol{x}^\star)\right\|\right).$$

Furthermore, we have

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

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

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

By (4.2c), 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 for any $\nu > 0$

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

By (E.11), (E.38), and the choices of p, q in (E.31) and (E.32), the second term on the right hand side can be bounded by

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

Combining the above five displays with Lemma E.6, we have

$$\|\Omega_t - \Omega^\star\| = O\left(\sqrt{\varphi_t \log(1/\varphi_t)}\right) + o\big(\sqrt{(\log t)^{1+\upsilon}/t}\big). \tag{E.44}$$

Combining (E.41), (E.42), (E.43), and (E.44), we complete the first part of the proof. For the Berry-Esseen inequality, we simply 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, we apply Theorem 5.6 and Lemma B.5, and complete the proof.

Appendix F. Additional Experimental Results

In this section, we provide more implementation details and show additional results. We follow the introduction in Section 6, and implement the method on eight problems in CUTEst test set and on linearly/nonlinearly constrained regression problems. For both implementation, we run

10^5 iterations, set $\beta_t = 1/t^{0.501}$, $\chi_t = \beta_t^2$, and $\bar{\alpha}_t \sim \text{Uniform}([\beta_t, \eta_t])$ with $\eta_t = \beta_t + \chi_t$. Regarding the Hessian regularization Δ_t, we let ($\lambda_{\min}(\cdot)$ denotes the least eigenvalue)

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

Here, $Z_t \in \mathbb{R}^{d\times(d-m)}$ has orthonormal columns that span the space $\{\boldsymbol{x} \in \mathbb{R}^d : G_t \boldsymbol{x} = \boldsymbol{0}\}$, which is obtained from the QR decomposition.

F.1 CUTEst problems

In this section, we testify the convergence rate in Theorem 5.5. In particular, we randomly pick one run across 200 runs, and show the convergence plots of the KKT residual $\|\nabla \mathcal{L}_t\|$, the mean absolute error $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|$, and the Hessian error $\|K_t - K^\star\|$. By Theorem 5.5, Lemma E.6, and the Lipschitz continuity of the Hessian, the theoretical convergence rate for these three quantities is $O(\sqrt{\beta_t \log(1/\beta_t)})$.

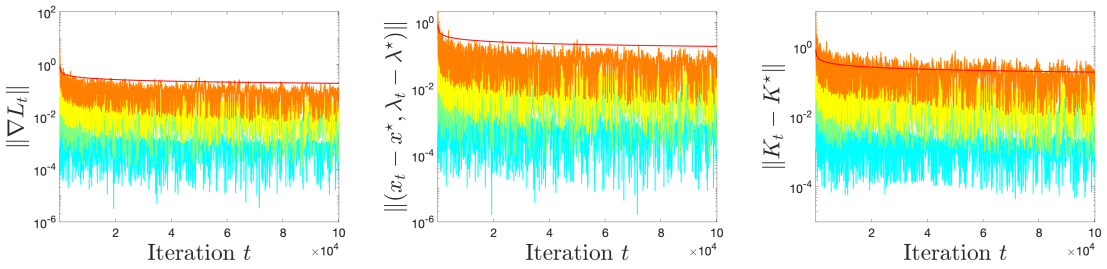
The convergence plots are shown in Figures 1 and 2. We use six problems for illustration. From the figures, we observe that the method converges faster for small sampling variance σ^2 and converges slower for large σ^2. Specifically, our theoretical convergence rate precisely characterizes asymptotic behavior of the method, and σ^2 only affects the rate as a constant factor.

F.2 Constrained regression problems

We follow the experiments in Section 6.2, and provide comprehensive comparisons between inexact and exact second-order methods on linearly/nonlinearly constrained regression problems. The coefficient matrix A in linear constraints is sampled from standard normal; the objective of logistic loss is regularized by a quadratic penalty term with a unit parameter. We vary the parameters d, r and τ. In particular, we let $d \in \{5, 20, 40, 60\}$, $r \in \{0.4, 0.5, 0.6\}$ for Toeplitz Σ_a and $r \in \{0.1, 0.2, 0.3\}$ for Equi-correlation Σ_a, and $\tau \in \{\infty, 20, 40, 60\}$. We mention that $\tau = \infty$ corresponds to the exact method. For each setup, we perform 200 independent runs.

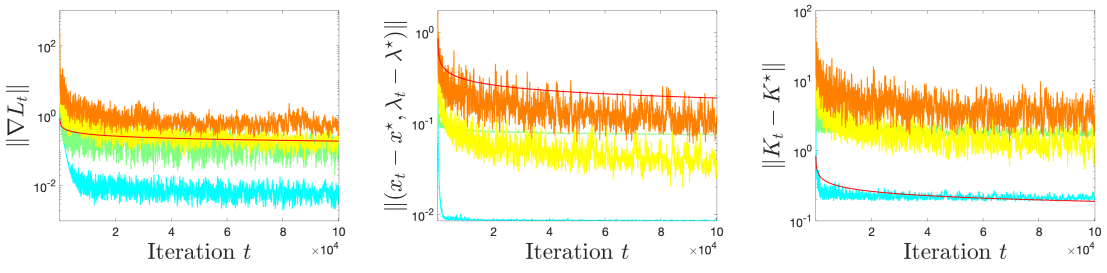
The extensive comparison results of offline M-estimation and StoSQP with different τ are reported in Tables 4-11. Specifically, Tables 4 and 5 summarize the results of linear model + linear constraints; Tables 6 and 7 summarize the results of linear model + nonlinear constraints; Tables 8 and 9 summarize the results of logistic model + linear constraints; while Tables 10 and 11 summarize the results of logistic model + nonlinear constraints. For all four cases, we have the following observations.

For MAE, we observe that M-estimation achieves results that are an order of magnitude smaller than those of the StoSQP methods. Among the different setups of τ of StoSQP, we find that exact StoSQP ($\tau = \infty$) generally yields smaller MAE compared to inexact StoSQP. Furthermore, a larger τ (i.e., more sketching steps) in StoSQP tends to result in smaller MAE, although the differences are less evident than those observed between StoSQP and M-estimation methods. This trend is robust across different setups of the design covariance Σ_a. For instance, for $d = 40$ in Table 5, StoSQP with $\tau = 20$ achieves an MAE of approximately 0.2–0.25, while StoSQP with $\tau = 40, 60, \infty$ achieves an MAE of less than 0.1. This observation suggests that, given problem parameters such as the condition number of the Lagrangian Hessian and problem dimension, $\tau = 20$ may be insufficient for solving Newton systems in this scenario. Specifically, solving exact Newton systems requires $O(46^3) = O(97,336)$ flops while a sketching solver



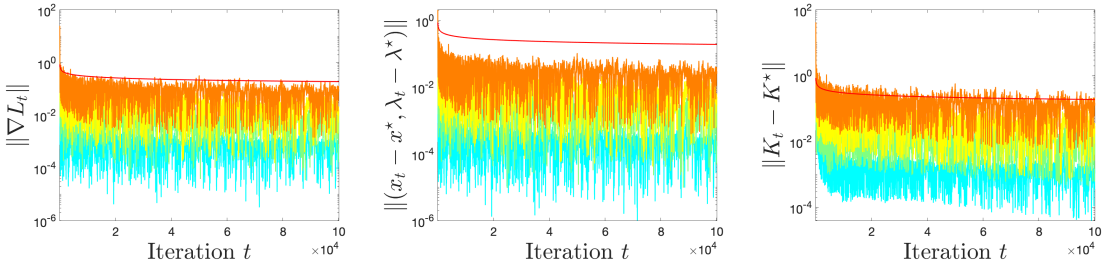
(a) KKT residual (b) Iteration error (c) Hessian error

Problem `MARATOS`



(d) KKT residual (e) Iteration error (f) Hessian error

Problem `ORTHREGB`

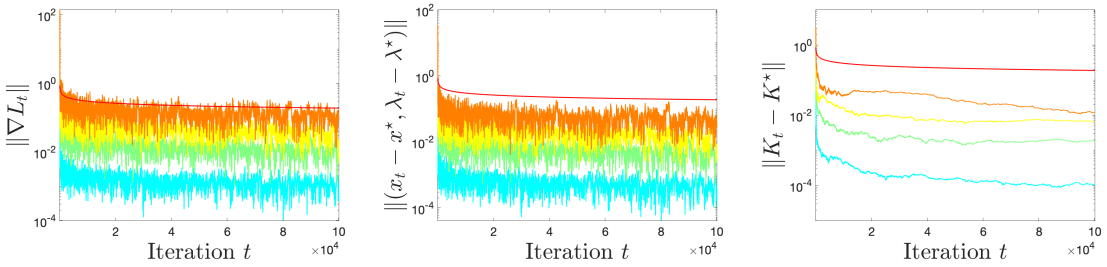


(g) KKT residual (h) Iteration error (i) Hessian error

Problem `HS7`

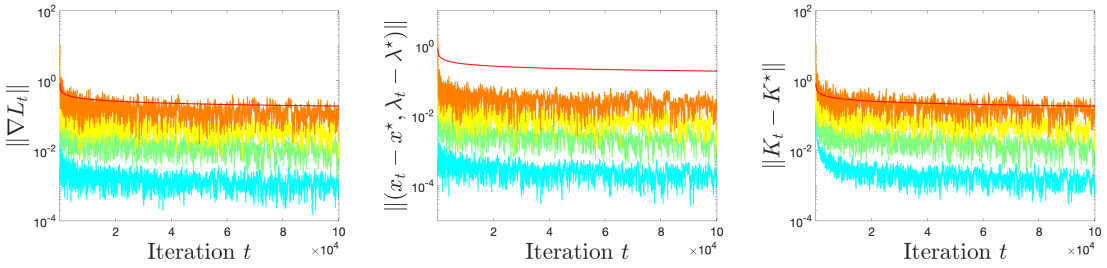
$\sigma^2 = 10^{-4}$ $\sigma^2 = 10^{-2}$ $\sigma^2 = 10^{-1}$ $\sigma^2 = 1$ Theory

Figure 1: *Convergence plots of CUTEst problems. Each row corresponds to one problem and has three figures in the* log *scale. From the left to the right, they correspond to* $\|\nabla \mathcal{L}_t\|$ *v.s.* t, $\|(\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)\|$ *v.s.* t, *and* $\|K_t - K^\star\|$ *v.s.* t. *Each figure has five lines; four lines correspond to four setups of* σ^2, *and the red line corresponds to* $\sqrt{\beta_t \log(1/\beta_t)}$ *v.s.* t, *which is the theoretical asymptotic rate.*



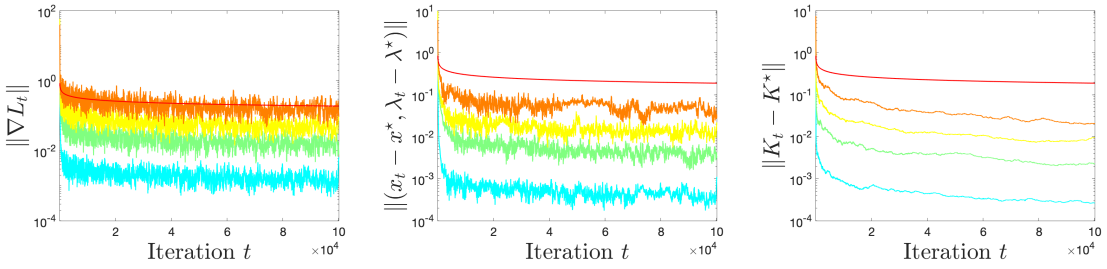
(a) KKT residual (b) Iteration error (c) Hessian error

Problem `HS48`



(d) KKT residual (e) Iteration error (f) Hessian error

Problem `HS78`



(g) KKT residual (h) Iteration error (i) Hessian error

Problem `GENHS28`

$\sigma^2 = 10^{-4}$ $\sigma^2 = 10^{-2}$ $\sigma^2 = 10^{-1}$ $\sigma^2 = 1$ Theory

Figure 2: *Convergence plots of CUTEst problems. See Figure 1 for the interpretation.*

with $\tau = 20$ requires only $O(20 \times 46) = O(920)$ flops. This substantial reduction in computational cost can result in insufficient precision for the Newton direction at each step, leading to a larger MAE when the sample size is fixed.

For coverage rate, we observe that StoSQP with different τ generally achieves a valid coverage rate that is very close to the nominal rate 95%, matching the performance of offline M-estimation. There are two potential exceptions. The first exception occurs when $d = 5$, where StoSQP exhibits undercoverage with rates ranging from 87%-92% (cf. Tables 4, 8, and 10). This undercoverage occurs because the condition numbers of the Lagrangian Hessian in these problems are significantly larger than in other scenarios, despite the small problem dimension of 5. Consequently, the iteration sequence of StoSQP may not have reached stationarity given the limited sample size. As noted in Zhu et al. (2021), even SGD-based estimators require 3×10^5 to 4×10^5 samples to alleviate undercoverage issue in challenging online inference tasks for unconstrained problems, while we only have 10^5 samples for constrained problems. Furthermore, applying a sketching solver to such small-scale problems seems an overkill. We recommend using a simple linear system solver for small-scale problems to reduce the additional uncertainty introduced by the sketching solver. The second exception occurs when $d = 40$, where StoSQP with $\tau = 20$ exhibits undercoverage with rates ranging from 82%-86%. In contrast, StoSQP with $\tau = 40, 60, \infty$ achieves valid coverage in this case (cf. Table 5). As explained for MAE, this undercoverage results from insufficient sketching steps, which not only make the StoSQP iteration sequence noisy but also make the bias of covariance matrix estimation non-negligible. For this scenario, slightly increasing the number of sketching steps (e.g., setting $\tau = 40$) can resolve the undercoverage issue while still preserving computational efficiency compared to exact methods.

For average length of confidence intervals, we observe that M-estimation produces intervals that are an order of magnitude shorter than those of the StoSQP methods. Among the StoSQP methods, the inexact settings ($\tau < \infty$) yield average lengths very similar to the exact setting ($\tau = \infty$), indicating that the inexact methods do not lead to overly conservative intervals. For both Toeplitz and Equi-correlation Σ_a, the length of the confidence intervals remains largely unchanged across different setups of r. Moreover, for both linear and logistic models, nonlinear constraints tend to result in wider confidence intervals for both offline and online methods, as shown by comparisons of Tables 4, 5 with Tables 6, 7 and Tables 8, 9 with Tables 10, 11.

For computational flops per iteration, we observe that online StoSQP methods are significantly more efficient than the offline method. A sketching solver can further reduce the computational costs of StoSQP. Choosing the sketching step τ involves a trade-off between computational and statistical efficiency. As shown in Table 5, when $d = 40$, $\tau = 20$ requires fewer flops but achieves larger MAE and lower coverage rates, whereas $\tau = 60$ requires more flops (though still fewer than those of M-estimation and $\tau = \infty$) but achieves lower MAE and better coverage rates. In this case, $\tau = 40$ strikes a better balance between the two aspects. We should also mention that for small-scale problems ($d = 5$), using a sketching solver with a large τ is counterproductive, as it may cause the flops to exceed those of the exact solver.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
5	Identity		M-estimation		0.20(0.08)	95.40	0.19(0.03)	16189377.50
			StoSQP	$\tau = \infty$	2.64(1.10)	94.00	2.43(0.38)	443.00
				$\tau = 20$	3.19(1.03)	88.80	2.43(0.38)	240.00
				$\tau = 40$	3.09(1.19)	89.00	2.43(0.39)	380.00
				$\tau = 60$	2.93(0.99)	91.80	2.43(0.39)	520.00
	Toeplitz	$r = 0.4$	M-estimation		0.20(0.08)	95.50	0.19(0.03)	17085960.18
			StoSQP	$\tau = \infty$	2.61(1.05)	95.00	2.40(0.40)	443.00
				$\tau = 20$	2.92(1.12)	90.70	2.40(0.40)	240.00
				$\tau = 40$	3.01(1.07)	90.30	2.40(0.40)	380.00
				$\tau = 60$	2.93(1.05)	90.80	2.40(0.40)	520.00
		$r = 0.5$	M-estimation		0.20(0.08)	94.60	0.19(0.03)	17117046.95
			StoSQP	$\tau = \infty$	2.58(1.02)	94.50	2.39(0.40)	443.00
				$\tau = 20$	2.93(1.05)	91.40	2.40(0.40)	240.00
				$\tau = 40$	2.84(1.08)	90.70	2.39(0.41)	380.00
				$\tau = 60$	3.00(1.06)	90.60	2.40(0.41)	520.00
		$r = 0.6$	M-estimation		0.20(0.08)	95.60	0.19(0.03)	16699894.12
			StoSQP	$\tau = \infty$	2.61(1.07)	94.40	2.39(0.41)	443.00
				$\tau = 20$	2.89(0.95)	92.20	2.39(0.41)	240.00
				$\tau = 40$	3.09(1.11)	89.10	2.39(0.41)	380.00
				$\tau = 60$	2.85(1.07)	92.20	2.39(0.41)	520.00
	Equi-correlation	$r = 0.1$	M-estimation		0.22(0.08)	94.10	0.19(0.03)	16591372.59
			StoSQP	$\tau = \infty$	2.68(1.14)	94.20	2.42(0.39)	443.00
				$\tau = 20$	3.09(0.94)	89.20	2.42(0.39)	240.00
				$\tau = 40$	2.94(1.06)	90.70	2.42(0.39)	380.00
				$\tau = 60$	2.96(1.08)	91.10	2.42(0.39)	520.00
		$r = 0.2$	M-estimation		0.21(0.09)	94.50	0.19(0.03)	16564921.51
			StoSQP	$\tau = \infty$	2.65(1.12)	93.30	2.41(0.40)	443.00
				$\tau = 20$	3.09(1.07)	90.40	2.41(0.40)	240.00
				$\tau = 40$	3.03(1.04)	90.70	2.41(0.40)	380.00
				$\tau = 60$	2.94(1.05)	90.80	2.41(0.40)	520.00
		$r = 0.3$	M-estimation		0.20(0.08)	96.90	0.19(0.03)	17816134.75
			StoSQP	$\tau = \infty$	2.57(1.14)	93.20	2.40(0.41)	443.00
				$\tau = 20$	3.05(1.04)	90.60	2.41(0.41)	240.00
				$\tau = 40$	3.13(1.14)	88.70	2.40(0.41)	380.00
				$\tau = 60$	2.98(1.01)	90.90	2.41(0.41)	520.00
20	Identity		M-estimation		0.51(0.09)	94.98	0.23(0.02)	71149851.21
			StoSQP	$\tau = \infty$	6.49(1.17)	94.60	2.87(0.22)	15203.99
				$\tau = 20$	6.85(1.09)	93.50	2.87(0.22)	1860.12
				$\tau = 40$	6.82(1.18)	93.67	2.87(0.22)	2340.11
				$\tau = 60$	6.81(1.14)	93.35	2.87(0.22)	2820.11
	Toeplitz	$r = 0.4$	M-estimation		0.51(0.09)	94.77	0.23(0.02)	72063611.92
			StoSQP	$\tau = \infty$	6.43(1.15)	95.22	2.88(0.22)	15203.99
				$\tau = 20$	6.79(1.09)	93.95	2.88(0.22)	1860.12
				$\tau = 40$	6.87(1.13)	93.80	2.89(0.22)	2340.11
				$\tau = 60$	6.83(1.20)	94.00	2.88(0.22)	2820.11
		$r = 0.5$	M-estimation		0.52(0.10)	94.35	0.23(0.02)	67237313.19
			StoSQP	$\tau = \infty$	6.44(1.12)	95.12	2.89(0.22)	15203.99
				$\tau = 20$	6.75(1.16)	93.80	2.89(0.22)	1860.12
				$\tau = 40$	6.83(1.07)	93.42	2.89(0.22)	2340.11
				$\tau = 60$	6.86(1.04)	94.03	2.89(0.22)	2820.11
		$r = 0.6$	M-estimation		0.51(0.09)	95.07	0.23(0.02)	69161097.88
			StoSQP	$\tau = \infty$	6.57(1.28)	94.47	2.90(0.22)	15203.99
				$\tau = 20$	6.82(1.10)	93.83	2.90(0.22)	1860.12
				$\tau = 40$	6.79(1.15)	93.78	2.90(0.22)	2340.11
				$\tau = 60$	6.76(1.13)	94.30	2.91(0.22)	2820.11
	Equi-correlation	$r = 0.1$	M-estimation		0.51(0.09)	95.25	0.23(0.02)	71077050.43
			StoSQP	$\tau = \infty$	6.46(1.15)	95.07	2.88(0.22)	15203.99
				$\tau = 20$	6.86(1.06)	93.03	2.88(0.23)	1860.12
				$\tau = 40$	6.82(1.13)	93.93	2.88(0.22)	2340.11
				$\tau = 60$	6.71(1.07)	94.07	2.88(0.22)	2820.11
		$r = 0.2$	M-estimation		0.52(0.09)	94.85	0.23(0.02)	63103131.95
			StoSQP	$\tau = \infty$	6.44(1.20)	94.95	2.89(0.23)	15203.99
				$\tau = 20$	6.78(0.98)	94.25	2.89(0.23)	1860.12
				$\tau = 40$	6.70(1.10)	94.40	2.89(0.23)	2340.11
				$\tau = 60$	6.76(1.07)	94.17	2.89(0.23)	2820.11
		$r = 0.3$	M-estimation		0.51(0.08)	95.05	0.23(0.02)	71184473.01
			StoSQP	$\tau = \infty$	6.78(1.12)	94.50	2.90(0.23)	15203.99
				$\tau = 20$	6.67(1.04)	94.72	2.91(0.23)	1860.12
				$\tau = 40$	6.70(1.10)	94.20	2.91(0.23)	2340.11
				$\tau = 60$	6.73(1.16)	94.05	2.91(0.23)	2820.11

Table 4: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(linear model + linear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
40	Identity		M-estimation		0.75 (0.09)	94.99	0.23 (0.01)	230134815.43
			StoSQP	$\tau = \infty$	9.53 (1.24)	95.30	3.00 (0.14)	102655.95
				$\tau = 20$	25.15 (65.81)	82.88	4.02 (5.33)	6240.91
				$\tau = 40$	10.02 (1.15)	94.32	3.01 (0.14)	7160.90
				$\tau = 60$	9.92 (1.24)	94.21	3.01 (0.14)	8080.89
	Toeplitz	$r = 0.4$	M-estimation		0.75 (0.09)	94.97	0.23 (0.01)	231495041.21
			StoSQP	$\tau = \infty$	9.76 (1.16)	94.74	3.01 (0.15)	102655.95
				$\tau = 20$	22.55 (57.70)	83.07	3.83 (5.02)	6240.91
				$\tau = 40$	9.90 (1.23)	94.36	3.02 (0.15)	7160.90
				$\tau = 60$	9.83 (1.16)	94.47	3.02 (0.15)	8080.89
		$r = 0.5$	M-estimation		0.75 (0.09)	94.97	0.23 (0.01)	204073927.80
			StoSQP	$\tau = \infty$	9.61 (1.22)	95.16	3.02 (0.15)	102655.95
				$\tau = 20$	17.85 (23.63)	84.56	3.46 (1.70)	6240.91
				$\tau = 40$	9.84 (1.49)	94.60	3.03 (0.16)	7160.90
				$\tau = 60$	9.82 (1.10)	94.79	3.03 (0.15)	8080.89
		$r = 0.6$	M-estimation		0.75 (0.09)	94.73	0.24 (0.01)	208581042.73
			StoSQP	$\tau = \infty$	9.61 (1.14)	95.61	3.03 (0.15)	102655.95
				$\tau = 20$	25.70 (63.84)	83.76	4.11 (5.20)	6240.91
				$\tau = 40$	9.60 (1.26)	95.27	3.03 (0.15)	7160.90
				$\tau = 60$	9.78 (1.20)	94.58	3.03 (0.15)	8080.89
	Equi-correlation	$r = 0.1$	M-estimation		0.74 (0.09)	95.25	0.23 (0.01)	173820750.11
			StoSQP	$\tau = \infty$	9.69 (1.25)	95.03	3.01 (0.14)	102655.95
				$\tau = 20$	19.61 (25.43)	84.30	3.58 (1.78)	6240.91
				$\tau = 40$	10.04 (1.19)	93.90	3.02 (0.14)	7160.90
				$\tau = 60$	9.80 (1.20)	94.55	3.02 (0.14)	8080.89
		$r = 0.2$	M-estimation		0.75 (0.09)	95.35	0.24 (0.01)	192694961.83
			StoSQP	$\tau = \infty$	9.81 (1.24)	94.89	3.03 (0.15)	102655.95
				$\tau = 20$	18.61 (24.12)	84.55	3.53 (1.76)	6240.91
				$\tau = 40$	9.84 (1.16)	94.69	3.03 (0.15)	7160.90
				$\tau = 60$	9.82 (1.16)	94.69	3.03 (0.15)	8080.89
		$r = 0.3$	M-estimation		0.76 (0.10)	95.15	0.24 (0.01)	215092766.42
			StoSQP	$\tau = \infty$	9.83 (1.20)	95.16	3.05 (0.15)	102655.95
				$\tau = 20$	14.45 (10.11)	86.52	3.25 (0.54)	6240.91
				$\tau = 40$	9.64 (1.21)	95.28	3.05 (0.15)	7160.90
				$\tau = 60$	9.83 (1.18)	95.00	3.05 (0.15)	8080.89
60	Identity		M-estimation		0.95(0.09)	94.47	0.24(0.01)	441552192.70
			StoSQP	$\tau = \infty$	12.21(1.28)	94.97	3.11(0.12)	312462.88
				$\tau = 20$	12.56(1.67)	94.57	3.12(0.14)	13042.88
				$\tau = 40$	12.34(1.18)	94.91	3.12(0.12)	14382.86
				$\tau = 60$	12.39(1.15)	94.74	3.12(0.12)	15722.85
	Toeplitz	$r = 0.4$	M-estimation		0.93(0.09)	95.27	0.24(0.01)	354867730.00
			StoSQP	$\tau = \infty$	12.41(1.23)	94.91	3.13(0.12)	312462.88
				$\tau = 20$	12.32(1.61)	94.95	3.13(0.13)	13042.88
				$\tau = 40$	12.25(1.18)	95.13	3.13(0.12)	14382.86
				$\tau = 60$	12.38(1.24)	94.76	3.13(0.12)	15722.85
		$r = 0.5$	M-estimation		0.94(0.09)	94.81	0.24(0.01)	353760329.36
			StoSQP	$\tau = \infty$	12.31(1.17)	95.17	3.13(0.12)	312462.88
				$\tau = 20$	12.28(1.13)	95.11	3.14(0.12)	13042.88
				$\tau = 40$	12.29(1.18)	95.21	3.14(0.12)	14382.86
				$\tau = 60$	12.32(1.25)	95.16	3.14(0.12)	15722.85
		$r = 0.6$	M-estimation		0.95(0.09)	94.91	0.24(0.01)	353373321.65
			StoSQP	$\tau = \infty$	12.39(1.19)	95.07	3.15(0.13)	312462.88
				$\tau = 20$	12.19(1.15)	95.36	3.15(0.13)	13042.88
				$\tau = 40$	12.21(1.25)	95.38	3.15(0.12)	14382.86
				$\tau = 60$	12.27(1.31)	95.22	3.15(0.12)	15722.85
	Equi-correlation	$r = 0.1$	M-estimation		0.94(0.09)	94.86	0.24(0.01)	299536428.06
			StoSQP	$\tau = \infty$	12.29(1.21)	95.01	3.13(0.12)	312462.88
				$\tau = 20$	12.37(1.20)	94.84	3.13(0.12)	13042.88
				$\tau = 40$	12.31(1.23)	95.04	3.13(0.12)	14382.86
				$\tau = 60$	12.33(1.12)	94.73	3.13(0.12)	15722.85
		$r = 0.2$	M-estimation		0.94(0.10)	95.07	0.24(0.01)	296068514.17
			StoSQP	$\tau = \infty$	12.45(1.34)	94.77	3.15(0.12)	312462.88
				$\tau = 20$	12.14(1.21)	95.22	3.14(0.13)	13042.88
				$\tau = 40$	12.07(1.22)	95.66	3.14(0.13)	14382.86
				$\tau = 60$	12.13(1.15)	95.37	3.15(0.13)	15722.85
		$r = 0.3$	M-estimation		0.95(0.09)	95.08	0.24(0.01)	319702105.39
			StoSQP	$\tau = \infty$	12.48(1.17)	94.94	3.17(0.13)	312462.88
				$\tau = 20$	11.83(1.23)	96.01	3.16(0.13)	13042.88
				$\tau = 40$	12.02(1.18)	95.87	3.16(0.13)	14382.86
				$\tau = 60$	11.99(1.24)	96.00	3.16(0.13)	15722.85

Table 5: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(linear model + linear constraints)****.*

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
5	Identity		M-estimation		0.25(0.10)	94.90	0.22(0.03)	9383817.60
			StoSQP	$\tau = \infty$	3.09(1.13)	94.90	2.82(0.37)	306.00
				$\tau = 20$	3.33(1.13)	93.50	2.82(0.37)	210.00
				$\tau = 40$	3.25(1.10)	93.40	2.82(0.37)	330.00
				$\tau = 60$	3.11(1.05)	93.90	2.82(0.38)	450.00
	Toeplitz	$r = 0.4$	M-estimation		0.24(0.10)	94.90	0.23(0.03)	10077576.37
			StoSQP	$\tau = \infty$	3.11(1.09)	94.90	2.87(0.37)	306.00
				$\tau = 20$	3.41(1.07)	92.60	2.87(0.37)	210.00
				$\tau = 40$	3.35(1.15)	92.60	2.87(0.37)	330.00
				$\tau = 60$	3.34(1.15)	92.10	2.87(0.37)	450.00
		$r = 0.5$	M-estimation		0.26(0.10)	93.90	0.23(0.03)	9986598.34
			StoSQP	$\tau = \infty$	3.18(1.24)	93.70	2.89(0.37)	306.00
				$\tau = 20$	3.21(1.21)	93.60	2.89(0.37)	210.00
				$\tau = 40$	3.16(1.01)	95.00	2.88(0.37)	330.00
				$\tau = 60$	3.29(1.13)	93.00	2.89(0.37)	450.00
		$r = 0.6$	M-estimation		0.24(0.09)	94.70	0.23(0.03)	10003553.94
			StoSQP	$\tau = \infty$	3.23(1.14)	94.30	2.90(0.37)	306.00
				$\tau = 20$	3.30(1.03)	94.10	2.91(0.36)	210.00
				$\tau = 40$	3.36(1.16)	93.20	2.90(0.37)	330.00
				$\tau = 60$	3.23(1.11)	94.20	2.90(0.36)	450.00
	Equi-correlation	$r = 0.1$	M-estimation		0.25(0.09)	95.50	0.23(0.03)	9287856.46
			StoSQP	$\tau = \infty$	3.18(1.15)	93.10	2.83(0.37)	306.00
				$\tau = 20$	3.06(1.15)	93.70	2.83(0.37)	210.00
				$\tau = 40$	3.17(1.10)	93.00	2.84(0.37)	330.00
				$\tau = 60$	3.26(1.22)	93.60	2.83(0.37)	450.00
		$r = 0.2$	M-estimation		0.25(0.08)	94.80	0.23(0.03)	9739704.68
			StoSQP	$\tau = \infty$	3.07(1.19)	94.30	2.85(0.37)	306.00
				$\tau = 20$	3.36(1.00)	92.20	2.85(0.37)	210.00
				$\tau = 40$	3.21(1.07)	93.60	2.85(0.37)	330.00
				$\tau = 60$	3.11(1.18)	95.20	2.85(0.37)	450.00
		$r = 0.3$	M-estimation		0.25(0.09)	95.00	0.23(0.03)	9545761.59
			StoSQP	$\tau = \infty$	3.18(1.05)	95.50	2.87(0.37)	306.00
				$\tau = 20$	3.28(1.15)	93.10	2.87(0.36)	210.00
				$\tau = 40$	3.19(1.07)	94.00	2.87(0.36)	330.00
				$\tau = 60$	3.16(1.12)	95.10	2.87(0.37)	450.00
20	Identity		M-estimation		0.56(0.09)	94.65	0.25(0.01)	34394136.60
			StoSQP	$\tau = \infty$	7.07(1.23)	95.10	3.14(0.09)	10520.99
				$\tau = 20$	7.07(1.13)	94.90	3.14(0.09)	1680.08
				$\tau = 40$	7.08(1.11)	94.55	3.14(0.09)	2100.07
				$\tau = 60$	7.09(1.16)	95.00	3.14(0.09)	2520.07
	Toeplitz	$r = 0.4$	M-estimation		0.55(0.09)	95.35	0.25(0.01)	36385429.73
			StoSQP	$\tau = \infty$	7.13(1.21)	94.55	3.17(0.08)	10520.99
				$\tau = 20$	7.04(1.05)	95.10	3.17(0.08)	1680.08
				$\tau = 40$	7.21(1.05)	94.67	3.17(0.08)	2100.07
				$\tau = 60$	7.26(1.13)	94.45	3.17(0.08)	2520.07
		$r = 0.5$	M-estimation		0.56(0.10)	94.83	0.25(0.01)	37880561.11
			StoSQP	$\tau = \infty$	7.05(1.09)	95.43	3.18(0.08)	10520.99
				$\tau = 20$	7.19(1.16)	95.17	3.18(0.08)	1680.08
				$\tau = 40$	7.15(1.11)	95.35	3.18(0.08)	2100.07
				$\tau = 60$	7.15(1.08)	94.92	3.18(0.08)	2520.07
		$r = 0.6$	M-estimation		0.57(0.10)	94.65	0.25(0.01)	38951806.15
			StoSQP	$\tau = \infty$	7.15(1.08)	95.13	3.20(0.08)	10520.99
				$\tau = 20$	7.19(1.16)	95.35	3.20(0.08)	1680.08
				$\tau = 40$	7.21(1.16)	95.07	3.20(0.08)	2100.07
				$\tau = 60$	7.33(1.15)	94.75	3.20(0.08)	2520.07
	Equi-correlation	$r = 0.1$	M-estimation		0.57(0.09)	94.47	0.25(0.01)	36242778.83
			StoSQP	$\tau = \infty$	7.01(1.17)	95.62	3.16(0.09)	10520.99
				$\tau = 20$	7.24(1.18)	95.07	3.16(0.09)	1680.08
				$\tau = 40$	7.27(1.16)	95.17	3.16(0.09)	2100.07
				$\tau = 60$	7.13(1.10)	94.97	3.16(0.09)	2520.07
		$r = 0.2$	M-estimation		0.56(0.09)	95.23	0.25(0.01)	36865114.61
			StoSQP	$\tau = \infty$	7.16(1.24)	94.60	3.18(0.09)	10520.99
				$\tau = 20$	7.30(1.08)	94.73	3.18(0.08)	1680.08
				$\tau = 40$	7.11(1.14)	94.77	3.18(0.08)	2100.07
				$\tau = 60$	7.12(1.17)	95.10	3.18(0.08)	2520.07
		$r = 0.3$	M-estimation		0.56(0.10)	95.27	0.25(0.01)	36995422.08
			StoSQP	$\tau = \infty$	7.33(1.11)	94.50	3.20(0.08)	10520.99
				$\tau = 20$	7.09(1.17)	95.12	3.20(0.08)	1680.08
				$\tau = 40$	7.14(1.14)	95.48	3.20(0.08)	2100.07
				$\tau = 60$	7.18(1.15)	95.10	3.20(0.08)	2520.07

Table 6: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(linear model + nonlinear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
40	Identity		M-estimation		0.79(0.09)	95.23	0.25(0.01)	67887974.71
			StoSQP	$\tau = \infty$	10.44(1.14)	94.99	3.24(0.06)	73840.95
				$\tau = 20$	10.63(1.20)	94.44	3.23(0.06)	5740.63
				$\tau = 40$	10.32(1.15)	94.94	3.23(0.06)	6560.62
				$\tau = 60$	10.34(1.21)	95.06	3.24(0.06)	7380.62
	Toeplitz	$r = 0.4$	M-estimation		0.80(0.09)	95.39	0.25(0.01)	72554106.22
			StoSQP	$\tau = \infty$	10.40(1.11)	94.97	3.25(0.06)	73840.95
				$\tau = 20$	10.33(1.16)	95.11	3.26(0.06)	5740.63
				$\tau = 40$	10.26(1.11)	95.51	3.26(0.06)	6560.62
				$\tau = 60$	10.35(1.25)	95.00	3.26(0.06)	7380.62
		$r = 0.5$	M-estimation		0.81(0.09)	95.20	0.25(0.01)	80163038.17
			StoSQP	$\tau = \infty$	10.45(1.20)	95.09	3.27(0.06)	73840.95
				$\tau = 20$	10.16(1.21)	95.63	3.27(0.06)	5740.63
				$\tau = 40$	10.34(1.20)	95.45	3.27(0.06)	6560.62
				$\tau = 60$	10.40(1.20)	95.51	3.27(0.06)	7380.62
		$r = 0.6$	M-estimation		0.81(0.10)	94.99	0.25(0.01)	81932666.57
			StoSQP	$\tau = \infty$	10.42(1.17)	95.61	3.29(0.06)	73840.95
				$\tau = 20$	10.31(1.18)	95.51	3.29(0.06)	5740.63
				$\tau = 40$	10.30(1.20)	95.35	3.28(0.06)	6560.62
				$\tau = 60$	10.41(1.21)	95.06	3.28(0.06)	7380.62
	Equi-correlation	$r = 0.1$	M-estimation		0.81(0.09)	94.86	0.25(0.01)	74034192.12
			StoSQP	$\tau = \infty$	10.45(1.17)	94.86	3.25(0.06)	73840.95
				$\tau = 20$	10.52(1.11)	95.12	3.26(0.06)	5740.63
				$\tau = 40$	10.39(1.28)	95.02	3.26(0.06)	6560.62
				$\tau = 60$	10.42(1.14)	95.17	3.26(0.06)	7380.62
		$r = 0.2$	M-estimation		0.81(0.09)	94.94	0.25(0.01)	80720232.66
			StoSQP	$\tau = \infty$	10.62(1.19)	94.96	3.28(0.06)	73840.95
				$\tau = 20$	10.36(1.20)	95.55	3.28(0.06)	5740.63
				$\tau = 40$	10.53(1.17)	94.99	3.28(0.06)	6560.62
				$\tau = 60$	10.38(1.17)	95.26	3.28(0.06)	7380.62
		$r = 0.3$	M-estimation		0.82(0.10)	94.88	0.26(0.01)	82412488.17
			StoSQP	$\tau = \infty$	10.68(1.26)	94.94	3.30(0.06)	73840.95
				$\tau = 20$	10.23(1.08)	95.90	3.30(0.06)	5740.63
				$\tau = 40$	10.28(1.10)	95.83	3.30(0.06)	6560.62
				$\tau = 60$	10.49(1.26)	94.97	3.30(0.06)	7380.62
60	Identity		M-estimation		0.99(0.09)	94.89	0.25(0.01)	103580157.11
			StoSQP	$\tau = \infty$	13.05(1.11)	94.82	3.31(0.06)	237960.89
				$\tau = 20$	13.08(1.19)	94.69	3.31(0.06)	12202.15
				$\tau = 40$	12.95(1.28)	95.04	3.30(0.06)	13422.14
				$\tau = 60$	13.21(1.29)	94.66	3.31(0.06)	14642.12
	Toeplitz	$r = 0.4$	M-estimation		0.99(0.09)	95.18	0.25(0.01)	109848502.87
			StoSQP	$\tau = \infty$	13.07(1.26)	95.22	3.33(0.05)	237960.89
				$\tau = 20$	12.92(1.22)	95.14	3.32(0.06)	12202.15
				$\tau = 40$	13.11(1.20)	94.87	3.33(0.06)	13422.14
				$\tau = 60$	12.99(1.23)	95.28	3.32(0.05)	14642.12
		$r = 0.5$	M-estimation		1.00(0.09)	94.82	0.25(0.01)	119048067.81
			StoSQP	$\tau = \infty$	13.12(1.32)	94.97	3.33(0.05)	237960.89
				$\tau = 20$	12.98(1.15)	95.59	3.34(0.06)	12202.15
				$\tau = 40$	12.80(1.29)	95.46	3.34(0.05)	13422.14
				$\tau = 60$	12.90(1.20)	95.69	3.34(0.05)	14642.12
		$r = 0.6$	M-estimation		1.00(0.09)	94.91	0.25(0.01)	121269402.35
			StoSQP	$\tau = \infty$	13.22(1.19)	94.76	3.36(0.05)	237960.89
				$\tau = 20$	12.83(1.18)	95.68	3.35(0.06)	12202.15
				$\tau = 40$	13.00(1.21)	95.22	3.35(0.05)	13422.14
				$\tau = 60$	12.82(1.24)	95.70	3.35(0.05)	14642.12
	Equi-correlation	$r = 0.1$	M-estimation		0.99(0.09)	94.90	0.25(0.01)	114985905.90
			StoSQP	$\tau = \infty$	13.05(1.23)	95.31	3.33(0.06)	237960.89
				$\tau = 20$	12.94(1.15)	95.44	3.33(0.05)	12202.15
				$\tau = 40$	13.03(1.18)	95.16	3.33(0.06)	13422.14
				$\tau = 60$	12.91(1.21)	95.22	3.33(0.06)	14642.12
		$r = 0.2$	M-estimation		0.99(0.09)	95.28	0.25(0.01)	123893833.21
			StoSQP	$\tau = \infty$	13.10(1.34)	95.21	3.35(0.05)	237960.89
				$\tau = 20$	12.97(1.17)	95.33	3.35(0.06)	12202.15
				$\tau = 40$	12.88(1.20)	95.63	3.35(0.05)	13422.14
				$\tau = 60$	13.08(1.17)	94.89	3.35(0.06)	14642.12
		$r = 0.3$	M-estimation		1.00(0.10)	95.15	0.26(0.01)	130752008.95
			StoSQP	$\tau = \infty$	13.42(1.22)	94.51	3.38(0.05)	237960.89
				$\tau = 20$	12.46(1.29)	96.46	3.36(0.05)	12202.15
				$\tau = 40$	12.83(1.21)	95.77	3.37(0.05)	13422.14
				$\tau = 60$	12.62(1.13)	96.19	3.37(0.06)	14642.12

Table 7: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(linear model + nonlinear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
5	Identity		M-estimation		0.13(0.06)	95.00	0.13(0.03)	18631964.09
			StoSQP	$\tau = \infty$	1.89(0.85)	91.70	1.57(0.42)	443.00
				$\tau = 20$	2.16(0.86)	86.10	1.56(0.42)	240.00
				$\tau = 40$	2.14(0.86)	86.40	1.56(0.42)	380.00
				$\tau = 60$	2.03(0.84)	89.50	1.57(0.42)	520.00
	Toeplitz	$r = 0.4$	M-estimation		0.14(0.06)	93.50	0.12(0.03)	12591949.55
			StoSQP	$\tau = \infty$	1.85(0.83)	91.70	1.54(0.42)	443.00
				$\tau = 20$	2.12(0.78)	87.10	1.54(0.42)	240.00
				$\tau = 40$	2.07(0.83)	88.90	1.54(0.42)	380.00
				$\tau = 60$	1.99(0.77)	90.10	1.54(0.42)	520.00
		$r = 0.5$	M-estimation		0.13(0.06)	95.60	0.12(0.03)	12485859.36
			StoSQP	$\tau = \infty$	1.87(0.83)	91.80	1.53(0.42)	443.00
				$\tau = 20$	2.12(0.79)	85.60	1.53(0.42)	240.00
				$\tau = 40$	1.97(0.77)	89.20	1.53(0.42)	380.00
				$\tau = 60$	1.97(0.82)	89.10	1.54(0.42)	520.00
		$r = 0.6$	M-estimation		0.13(0.06)	95.40	0.12(0.03)	12788454.60
			StoSQP	$\tau = \infty$	1.82(0.85)	92.70	1.52(0.42)	443.00
				$\tau = 20$	2.06(0.87)	86.60	1.52(0.42)	240.00
				$\tau = 40$	2.08(0.84)	87.00	1.53(0.42)	380.00
				$\tau = 60$	1.92(0.78)	91.00	1.52(0.42)	520.00
	Equi-correlation	$r = 0.1$	M-estimation		0.13(0.06)	96.20	0.12(0.03)	15479121.99
			StoSQP	$\tau = \infty$	1.96(0.88)	90.70	1.56(0.42)	443.00
				$\tau = 20$	2.11(0.79)	86.70	1.56(0.42)	240.00
				$\tau = 40$	2.10(0.77)	88.90	1.56(0.42)	380.00
				$\tau = 60$	2.04(0.79)	89.50	1.55(0.42)	520.00
		$r = 0.2$	M-estimation		0.14(0.06)	93.40	0.12(0.03)	13563734.48
			StoSQP	$\tau = \infty$	1.87(0.82)	91.80	1.55(0.42)	443.00
				$\tau = 20$	2.01(0.81)	88.50	1.55(0.42)	240.00
				$\tau = 40$	2.03(0.81)	88.00	1.55(0.41)	380.00
				$\tau = 60$	2.05(0.92)	89.20	1.55(0.42)	520.00
		$r = 0.3$	M-estimation		0.13(0.06)	95.30	0.12(0.03)	13559119.41
			StoSQP	$\tau = \infty$	1.90(0.85)	91.90	1.54(0.42)	443.00
				$\tau = 20$	2.13(0.83)	87.80	1.54(0.41)	240.00
				$\tau = 40$	2.05(0.88)	88.50	1.54(0.42)	380.00
				$\tau = 60$	1.93(0.78)	91.20	1.54(0.41)	520.00
20	Identity		M-estimation		0.31(0.05)	95.20	0.14(0.01)	102490458.22
			StoSQP	$\tau = \infty$	4.25(0.82)	92.62	1.73(0.15)	15203.99
				$\tau = 20$	4.21(0.79)	92.68	1.73(0.15)	1860.12
				$\tau = 40$	4.28(0.79)	92.25	1.73(0.15)	2340.11
				$\tau = 60$	4.23(0.79)	92.72	1.73(0.15)	2820.11
	Toeplitz	$r = 0.4$	M-estimation		0.30(0.05)	94.97	0.13(0.01)	104414364.13
			StoSQP	$\tau = \infty$	4.11(0.74)	92.77	1.67(0.15)	15203.99
				$\tau = 20$	3.99(0.75)	93.75	1.67(0.15)	1860.12
				$\tau = 40$	4.04(0.73)	93.15	1.67(0.14)	2340.11
				$\tau = 60$	4.09(0.71)	93.23	1.67(0.14)	2820.11
		$r = 0.5$	M-estimation		0.30(0.05)	94.60	0.13(0.01)	94453742.76
			StoSQP	$\tau = \infty$	3.99(0.75)	92.90	1.65(0.15)	15203.99
				$\tau = 20$	4.06(0.71)	92.75	1.65(0.14)	1860.12
				$\tau = 40$	4.01(0.83)	92.28	1.65(0.15)	2340.11
				$\tau = 60$	3.96(0.72)	93.33	1.65(0.15)	2820.11
		$r = 0.6$	M-estimation		0.29(0.06)	94.50	0.13(0.01)	98941404.11
			StoSQP	$\tau = \infty$	3.90(0.77)	93.05	1.62(0.14)	15203.99
				$\tau = 20$	3.84(0.74)	93.77	1.62(0.14)	1860.12
				$\tau = 40$	3.92(0.77)	92.65	1.62(0.14)	2340.11
				$\tau = 60$	3.93(0.76)	92.95	1.61(0.14)	2820.11
	Equi-correlation	$r = 0.1$	M-estimation		0.30(0.06)	95.30	0.13(0.01)	103331497.32
			StoSQP	$\tau = \infty$	4.05(0.83)	93.07	1.67(0.14)	15203.99
				$\tau = 20$	3.97(0.77)	93.28	1.66(0.14)	1860.12
				$\tau = 40$	4.14(0.71)	92.48	1.67(0.14)	2340.11
				$\tau = 60$	4.08(0.74)	92.90	1.67(0.14)	2820.11
		$r = 0.2$	M-estimation		0.29(0.05)	95.27	0.13(0.01)	98571633.30
			StoSQP	$\tau = \infty$	3.92(0.78)	92.92	1.61(0.14)	15203.99
				$\tau = 20$	3.80(0.64)	93.77	1.61(0.14)	1860.12
				$\tau = 40$	3.86(0.73)	93.20	1.61(0.14)	2340.11
				$\tau = 60$	3.91(0.75)	93.25	1.61(0.14)	2820.11
		$r = 0.3$	M-estimation		0.28(0.05)	95.48	0.12(0.01)	82384974.00
			StoSQP	$\tau = \infty$	3.66(0.69)	93.67	1.56(0.13)	15203.99
				$\tau = 20$	3.82(0.61)	92.77	1.55(0.13)	1860.12
				$\tau = 40$	3.75(0.69)	93.20	1.56(0.13)	2340.11
				$\tau = 60$	3.69(0.71)	93.67	1.56(0.13)	2820.11

Table 8: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(logistic model + linear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
40	Identity		M-estimation		0.40(0.05)	95.05	0.13(0.01)	368280379.99
			StoSQP	$\tau = \infty$	5.39(0.83)	93.46	1.59(0.09)	102655.95
				$\tau = 20$	5.11(0.68)	94.85	1.59(0.09)	6240.91
				$\tau = 40$	5.13(0.64)	94.94	1.59(0.09)	7160.90
				$\tau = 60$	5.21(0.72)	94.45	1.59(0.09)	8080.89
	Toeplitz	$r = 0.4$	M-estimation		0.39(0.05)	95.15	0.12(0.01)	354955718.18
			StoSQP	$\tau = \infty$	5.13(0.83)	93.66	1.54(0.09)	102655.95
				$\tau = 20$	4.97(0.67)	94.71	1.54(0.09)	6240.91
				$\tau = 40$	4.98(0.60)	94.78	1.53(0.09)	7160.90
				$\tau = 60$	5.00(0.74)	94.60	1.53(0.09)	8080.89
		$r = 0.5$	M-estimation		0.39(0.05)	94.65	0.12(0.01)	356017169.13
			StoSQP	$\tau = \infty$	5.05(0.74)	94.08	1.51(0.09)	102655.95
				$\tau = 20$	4.85(0.63)	95.17	1.51(0.09)	6240.91
				$\tau = 40$	4.89(0.68)	94.84	1.51(0.09)	7160.90
				$\tau = 60$	4.88(0.70)	94.76	1.51(0.09)	8080.89
		$r = 0.6$	M-estimation		0.38(0.05)	94.79	0.12(0.01)	330139055.54
			StoSQP	$\tau = \infty$	4.93(0.74)	93.71	1.47(0.09)	102655.95
				$\tau = 20$	4.66(0.60)	95.40	1.47(0.09)	6240.91
				$\tau = 40$	4.83(0.65)	94.69	1.48(0.09)	7160.90
				$\tau = 60$	4.79(0.65)	94.20	1.47(0.09)	8080.89
	Equi-correlation	$r = 0.1$	M-estimation		0.37(0.04)	95.37	0.12(0.01)	340887002.09
			StoSQP	$\tau = \infty$	4.97(0.80)	93.54	1.47(0.09)	102655.95
				$\tau = 20$	4.60(0.63)	95.75	1.47(0.09)	6240.91
				$\tau = 40$	4.75(0.56)	95.06	1.47(0.09)	7160.90
				$\tau = 60$	4.83(0.73)	94.04	1.47(0.09)	8080.89
		$r = 0.2$	M-estimation		0.35(0.04)	95.07	0.11(0.01)	323813851.17
			StoSQP	$\tau = \infty$	4.60(0.74)	93.80	1.38(0.09)	102655.95
				$\tau = 20$	4.40(0.66)	95.10	1.38(0.09)	6240.91
				$\tau = 40$	4.28(0.61)	95.81	1.38(0.09)	7160.90
				$\tau = 60$	4.50(0.70)	94.65	1.38(0.09)	8080.89
		$r = 0.3$	M-estimation		0.34(0.04)	95.15	0.11(0.01)	247729814.80
			StoSQP	$\tau = \infty$	4.47(0.77)	92.90	1.31(0.09)	102655.95
				$\tau = 20$	4.11(0.52)	95.42	1.31(0.09)	6240.91
				$\tau = 40$	4.14(0.59)	95.25	1.31(0.09)	7160.90
				$\tau = 60$	4.17(0.64)	95.02	1.31(0.09)	8080.89
60	Identity		M-estimation		0.48(0.05)	94.74	0.12(0.01)	621851092.85
			StoSQP	$\tau = \infty$	6.15(0.79)	94.14	1.51(0.08)	312462.88
				$\tau = 20$	5.64(0.56)	95.92	1.51(0.08)	13042.88
				$\tau = 40$	5.85(0.66)	95.20	1.51(0.08)	14382.86
				$\tau = 60$	5.79(0.65)	95.71	1.51(0.08)	15722.85
	Toeplitz	$r = 0.4$	M-estimation		0.46(0.05)	94.85	0.12(0.01)	629594361.50
			StoSQP	$\tau = \infty$	6.04(0.82)	93.42	1.45(0.08)	312462.88
				$\tau = 20$	5.43(0.61)	95.80	1.45(0.08)	13042.88
				$\tau = 40$	5.55(0.64)	95.71	1.45(0.08)	14382.86
				$\tau = 60$	5.60(0.68)	95.26	1.45(0.08)	15722.85
		$r = 0.5$	M-estimation		0.45(0.04)	94.89	0.11(0.01)	625412635.60
			StoSQP	$\tau = \infty$	5.83(0.82)	93.77	1.42(0.08)	312462.88
				$\tau = 20$	5.34(0.58)	96.02	1.43(0.08)	13042.88
				$\tau = 40$	5.40(0.61)	95.66	1.42(0.08)	14382.86
				$\tau = 60$	5.54(0.69)	95.15	1.42(0.08)	15722.85
		$r = 0.6$	M-estimation		0.44(0.04)	95.26	0.11(0.01)	622986528.38
			StoSQP	$\tau = \infty$	5.75(0.82)	93.65	1.39(0.07)	312462.88
				$\tau = 20$	5.19(0.54)	95.94	1.39(0.08)	13042.88
				$\tau = 40$	5.24(0.60)	95.81	1.39(0.07)	14382.86
				$\tau = 60$	5.32(0.66)	95.47	1.39(0.08)	15722.85
	Equi-correlation	$r = 0.1$	M-estimation		0.42(0.04)	95.28	0.11(0.01)	584884084.37
			StoSQP	$\tau = \infty$	5.54(0.80)	93.85	1.35(0.07)	312462.88
				$\tau = 20$	5.04(0.57)	95.97	1.34(0.07)	13042.88
				$\tau = 40$	5.07(0.62)	95.64	1.34(0.07)	14382.86
				$\tau = 60$	5.14(0.68)	95.54	1.34(0.07)	15722.85
		$r = 0.2$	M-estimation		0.39(0.03)	95.00	0.10(0.01)	583320213.75
			StoSQP	$\tau = \infty$	5.04(0.79)	93.96	1.24(0.07)	312462.88
				$\tau = 20$	4.52(0.47)	96.30	1.24(0.07)	13042.88
				$\tau = 40$	4.61(0.52)	96.19	1.24(0.07)	14382.86
				$\tau = 60$	4.70(0.68)	95.47	1.24(0.07)	15722.85
		$r = 0.3$	M-estimation		0.37(0.04)	94.97	0.09(0.01)	603554390.71
			StoSQP	$\tau = \infty$	4.66(0.78)	94.34	1.16(0.07)	312462.88
				$\tau = 20$	4.26(0.46)	96.17	1.16(0.07)	13042.88
				$\tau = 40$	4.36(0.58)	95.83	1.16(0.07)	14382.86
				$\tau = 60$	4.30(0.58)	96.00	1.16(0.07)	15722.85

Table 9: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(logistic model + linear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
5	Identity		M-estimation		0.17(0.07)	94.70	0.17(0.02)	5284663.88
			StoSQP	$\tau = \infty$	2.68(0.83)	91.40	2.08(0.28)	306.00
				$\tau = 20$	2.71(0.93)	90.00	2.08(0.27)	210.00
				$\tau = 40$	2.74(0.91)	89.30	2.08(0.27)	330.00
				$\tau = 60$	2.63(0.97)	90.50	2.08(0.27)	450.00
	Toeplitz	$r = 0.4$	M-estimation		0.18(0.06)	95.10	0.16(0.02)	5071684.79
			StoSQP	$\tau = \infty$	2.48(0.89)	91.80	2.03(0.28)	306.00
				$\tau = 20$	2.51(0.89)	91.10	2.03(0.28)	210.00
				$\tau = 40$	2.68(0.88)	89.00	2.03(0.28)	330.00
				$\tau = 60$	2.42(0.88)	92.10	2.04(0.28)	450.00
		$r = 0.5$	M-estimation		0.18(0.07)	94.30	0.16(0.02)	5073768.73
			StoSQP	$\tau = \infty$	2.45(0.87)	91.50	2.02(0.28)	306.00
				$\tau = 20$	2.48(0.95)	91.20	2.02(0.28)	210.00
				$\tau = 40$	2.45(0.85)	92.00	2.02(0.28)	330.00
				$\tau = 60$	2.44(0.89)	92.40	2.02(0.28)	450.00
		$r = 0.6$	M-estimation		0.17(0.06)	95.50	0.16(0.02)	5097272.67
			StoSQP	$\tau = \infty$	2.42(0.96)	92.70	2.01(0.28)	306.00
				$\tau = 20$	2.56(0.92)	90.80	2.01(0.28)	210.00
				$\tau = 40$	2.45(0.87)	92.00	2.01(0.28)	330.00
				$\tau = 60$	2.48(0.84)	91.20	2.01(0.28)	450.00
	Equi-correlation	$r = 0.1$	M-estimation		0.17(0.06)	96.00	0.16(0.02)	5345830.96
			StoSQP	$\tau = \infty$	2.40(0.91)	92.10	2.06(0.28)	306.00
				$\tau = 20$	2.56(0.94)	91.90	2.06(0.27)	210.00
				$\tau = 40$	2.58(0.94)	90.80	2.06(0.28)	330.00
				$\tau = 60$	2.62(0.98)	91.20	2.06(0.28)	450.00
		$r = 0.2$	M-estimation		0.18(0.07)	94.50	0.16(0.02)	5466386.61
			StoSQP	$\tau = \infty$	2.50(0.94)	91.40	2.05(0.28)	306.00
				$\tau = 20$	2.58(0.97)	90.70	2.05(0.28)	210.00
				$\tau = 40$	2.47(0.93)	91.10	2.05(0.28)	330.00
				$\tau = 60$	2.57(0.90)	90.40	2.05(0.28)	450.00
		$r = 0.3$	M-estimation		0.18(0.06)	94.70	0.16(0.02)	5556551.22
			StoSQP	$\tau = \infty$	2.54(0.94)	91.10	2.03(0.28)	306.00
				$\tau = 20$	2.57(0.93)	90.10	2.03(0.28)	210.00
				$\tau = 40$	2.42(0.89)	93.00	2.04(0.27)	330.00
				$\tau = 60$	2.51(0.94)	90.80	2.03(0.28)	450.00
20	Identity		M-estimation		0.34(0.06)	94.98	0.15(0.01)	18562224.33
			StoSQP	$\tau = \infty$	4.55(0.80)	93.80	1.92(0.07)	10520.99
				$\tau = 20$	4.63(0.78)	93.13	1.92(0.07)	1680.08
				$\tau = 40$	4.64(0.85)	93.20	1.92(0.07)	2100.07
				$\tau = 60$	4.64(0.74)	93.40	1.92(0.07)	2520.07
	Toeplitz	$r = 0.4$	M-estimation		0.34(0.05)	94.60	0.15(0.01)	18582079.94
			StoSQP	$\tau = \infty$	4.51(0.84)	92.72	1.86(0.07)	10520.99
				$\tau = 20$	4.29(0.77)	94.32	1.85(0.07)	1680.08
				$\tau = 40$	4.53(0.81)	93.00	1.85(0.07)	2100.07
				$\tau = 60$	4.58(0.79)	92.68	1.86(0.07)	2520.07
		$r = 0.5$	M-estimation		0.33(0.05)	94.93	0.15(0.01)	18629501.62
			StoSQP	$\tau = \infty$	4.49(0.82)	92.60	1.83(0.07)	10520.99
				$\tau = 20$	4.31(0.83)	93.50	1.83(0.07)	1680.08
				$\tau = 40$	4.38(0.73)	93.52	1.83(0.07)	2100.07
				$\tau = 60$	4.33(0.79)	93.82	1.83(0.07)	2520.07
		$r = 0.6$	M-estimation		0.32(0.05)	94.97	0.14(0.01)	18907211.70
			StoSQP	$\tau = \infty$	4.29(0.79)	93.70	1.79(0.07)	10520.99
				$\tau = 20$	4.21(0.73)	94.17	1.79(0.07)	1680.08
				$\tau = 40$	4.24(0.70)	93.87	1.80(0.07)	2100.07
				$\tau = 60$	4.31(0.84)	93.38	1.80(0.07)	2520.07
	Equi-correlation	$r = 0.1$	M-estimation		0.33(0.06)	94.98	0.15(0.01)	18557849.29
			StoSQP	$\tau = \infty$	4.49(0.82)	92.65	1.85(0.07)	10520.99
				$\tau = 20$	4.40(0.77)	93.75	1.85(0.07)	1680.08
				$\tau = 40$	4.46(0.74)	93.00	1.85(0.07)	2100.07
				$\tau = 60$	4.41(0.86)	93.52	1.85(0.07)	2520.07
		$r = 0.2$	M-estimation		0.32(0.05)	94.50	0.14(0.01)	17095346.92
			StoSQP	$\tau = \infty$	4.24(0.85)	93.23	1.78(0.07)	10520.99
				$\tau = 20$	4.17(0.70)	93.13	1.79(0.07)	1680.08
				$\tau = 40$	4.31(0.72)	92.93	1.78(0.07)	2100.07
				$\tau = 60$	4.26(0.78)	93.50	1.79(0.07)	2520.07
		$r = 0.3$	M-estimation		0.31(0.05)	95.00	0.14(0.01)	16972264.61
			StoSQP	$\tau = \infty$	4.10(0.77)	93.93	1.73(0.07)	10520.99
				$\tau = 20$	4.08(0.69)	93.82	1.73(0.07)	1680.08
				$\tau = 40$	4.17(0.74)	93.08	1.73(0.07)	2100.07
				$\tau = 60$	4.21(0.72)	92.95	1.73(0.07)	2520.07

Table 10: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(logistic model + nonlinear constraints)***.

d	Design Cov		Method		MAE (10^{-2})	Ave Cov (%)	Ave Len (10^{-2})	Flops/iter
40	Identity		M-estimation		0.45(0.05)	94.60	0.14(0.01)	36407191.34
			StoSQP	$\tau = \infty$	5.92(0.78)	92.85	1.73(0.05)	73840.95
				$\tau = 20$	5.82(0.74)	94.05	1.73(0.06)	5740.63
				$\tau = 40$	5.82(0.71)	93.62	1.72(0.06)	6560.62
				$\tau = 60$	5.86(0.79)	93.62	1.73(0.05)	7380.62
	Toeplitz	$r = 0.4$	M-estimation		0.42(0.05)	95.30	0.13(0.01)	36941246.80
			StoSQP	$\tau = \infty$	5.65(0.74)	93.55	1.66(0.06)	73840.95
				$\tau = 20$	5.47(0.72)	94.61	1.67(0.06)	5740.63
				$\tau = 40$	5.63(0.78)	93.70	1.67(0.06)	6560.62
				$\tau = 60$	5.60(0.76)	93.60	1.66(0.06)	7380.62
		$r = 0.5$	M-estimation		0.43(0.05)	94.54	0.13(0.01)	37104530.58
			StoSQP	$\tau = \infty$	5.48(0.81)	93.99	1.63(0.06)	73840.95
				$\tau = 20$	5.39(0.69)	94.29	1.63(0.06)	5740.63
				$\tau = 40$	5.47(0.77)	93.88	1.64(0.06)	6560.62
				$\tau = 60$	5.41(0.62)	94.04	1.63(0.06)	7380.62
		$r = 0.6$	M-estimation		0.41(0.04)	95.11	0.13(0.01)	36313253.38
			StoSQP	$\tau = \infty$	5.36(0.75)	93.67	1.60(0.06)	73840.95
				$\tau = 20$	5.32(0.66)	94.12	1.60(0.06)	5740.63
				$\tau = 40$	5.30(0.73)	94.27	1.60(0.06)	6560.62
				$\tau = 60$	5.44(0.80)	93.26	1.60(0.06)	7380.62
	Equi-correlation	$r = 0.1$	M-estimation		0.41(0.05)	94.83	0.13(0.01)	34995599.67
			StoSQP	$\tau = \infty$	5.34(0.82)	93.44	1.59(0.06)	73840.95
				$\tau = 20$	5.28(0.74)	94.13	1.59(0.06)	5740.63
				$\tau = 40$	5.25(0.71)	94.38	1.60(0.06)	6560.62
				$\tau = 60$	5.37(0.82)	93.56	1.60(0.06)	7380.62
		$r = 0.2$	M-estimation		0.39(0.04)	94.89	0.12(0.01)	34765391.77
			StoSQP	$\tau = \infty$	5.00(0.75)	93.80	1.50(0.06)	73840.95
				$\tau = 20$	4.92(0.65)	94.24	1.50(0.06)	5740.63
				$\tau = 40$	4.92(0.73)	94.35	1.50(0.06)	6560.62
				$\tau = 60$	4.88(0.70)	94.61	1.50(0.06)	7380.62
		$r = 0.3$	M-estimation		0.36(0.04)	95.39	0.11(0.01)	34923085.62
			StoSQP	$\tau = \infty$	4.75(0.69)	93.89	1.42(0.06)	73840.95
				$\tau = 20$	4.63(0.61)	94.44	1.42(0.06)	5740.63
				$\tau = 40$	4.68(0.68)	94.21	1.42(0.06)	6560.62
				$\tau = 60$	4.67(0.75)	94.36	1.42(0.06)	7380.62
60	Identity		M-estimation		0.51(0.05)	95.09	0.13(0.01)	56525866.64
			StoSQP	$\tau = \infty$	6.58(0.84)	94.02	1.61(0.05)	237960.89
				$\tau = 20$	6.57(0.71)	93.93	1.61(0.05)	12202.15
				$\tau = 40$	6.48(0.71)	94.39	1.60(0.06)	13422.14
				$\tau = 60$	6.50(0.76)	94.47	1.61(0.05)	14642.12
	Toeplitz	$r = 0.4$	M-estimation		0.49(0.05)	94.77	0.12(0.01)	56939372.59
			StoSQP	$\tau = \infty$	6.37(0.93)	93.73	1.54(0.06)	237960.89
				$\tau = 20$	6.26(0.72)	94.27	1.55(0.06)	12202.15
				$\tau = 40$	6.18(0.73)	94.53	1.54(0.06)	13422.14
				$\tau = 60$	6.29(0.79)	93.88	1.54(0.06)	14642.12
		$r = 0.5$	M-estimation		0.47(0.04)	95.36	0.12(0.01)	58241815.59
			StoSQP	$\tau = \infty$	6.18(0.96)	93.92	1.52(0.06)	237960.89
				$\tau = 20$	6.14(0.65)	94.18	1.52(0.06)	12202.15
				$\tau = 40$	6.16(0.65)	93.99	1.51(0.05)	13422.14
				$\tau = 60$	6.06(0.67)	94.73	1.52(0.06)	14642.12
		$r = 0.6$	M-estimation		0.46(0.04)	95.05	0.12(0.01)	58496455.48
			StoSQP	$\tau = \infty$	6.03(0.88)	93.86	1.48(0.06)	237960.89
				$\tau = 20$	6.03(0.65)	94.03	1.48(0.06)	12202.15
				$\tau = 40$	5.96(0.70)	94.53	1.48(0.06)	13422.14
				$\tau = 60$	6.01(0.78)	93.98	1.48(0.06)	14642.12
	Equi-correlation	$r = 0.1$	M-estimation		0.45(0.04)	94.85	0.11(0.01)	57063361.06
			StoSQP	$\tau = \infty$	5.85(0.94)	93.83	1.43(0.06)	237960.89
				$\tau = 20$	5.74(0.63)	94.32	1.43(0.06)	12202.15
				$\tau = 40$	5.82(0.70)	93.99	1.43(0.06)	13422.14
				$\tau = 60$	5.69(0.79)	94.47	1.43(0.05)	14642.12
		$r = 0.2$	M-estimation		0.42(0.04)	95.13	0.11(0.01)	56860272.87
			StoSQP	$\tau = \infty$	5.30(0.84)	94.18	1.32(0.06)	237960.89
				$\tau = 20$	5.26(0.62)	94.66	1.32(0.06)	12202.15
				$\tau = 40$	5.30(0.66)	94.34	1.32(0.06)	13422.14
				$\tau = 60$	5.34(0.75)	93.77	1.32(0.06)	14642.12
		$r = 0.3$	M-estimation		0.39(0.04)	94.83	0.10(0.01)	56511408.20
			StoSQP	$\tau = \infty$	4.97(0.86)	93.77	1.23(0.06)	237960.89
				$\tau = 20$	4.87(0.53)	94.65	1.23(0.06)	12202.15
				$\tau = 40$	5.00(0.76)	93.87	1.24(0.06)	13422.14
				$\tau = 60$	4.90(0.74)	94.49	1.23(0.06)	14642.12

Table 11: *Comparison results of online StoSQP and offline M-estimation for constrained regression problems* ***(logistic model + nonlinear constraints)***.