Online Inference of Constrained Optimization: Primal-Dual Optimality and Sequential Quadratic Programming

Yihang Gao
Department of Mathematics, National University of Singapore
Michael K. Ng
Department of Mathematics, Hong Kong Baptist University
Michael W. Mahoney
ICSI and Department of Statistics, University of California, Berkeley
Sen Na
School of Industrial and Systems Engineering, Georgia Institute of Technology

Abstract

We study online statistical inference for the solutions of stochastic optimization problems with equality and inequality constraints. Such problems are prevalent in statistics and machine learning, encompassing constrained M-estimation, physics-informed models, safe reinforcement learning, and algorithmic fairness. We develop a stochastic sequential quadratic programming (SSQP) method to solve these problems, where the step direction is computed by sequentially performing a quadratic approximation of the objective and a linear approximation of the constraints. Despite having access to unbiased estimates of population gradients, a key challenge in constrained stochastic problems lies in dealing with the bias in the step direction. As such, we apply a momentum-style gradient moving-average technique within SSQP to debias the step. We show that our method achieves global almost-sure convergence and exhibits local asymptotic normality with an *optimal* primal-dual limiting covariance matrix in the sense of Hájek and Le Cam. In addition, we provide a plug-in covariance matrix estimator for practical inference. To our knowledge, the proposed SSQP method is the *first fully online method* that attains primal-dual asymptotic minimax optimality without relying on projection operators onto the constraint set, which are generally intractable for nonlinear problems. Through extensive experiments on benchmark nonlinear problems, as well as on constrained generalized linear models and portfolio allocation problems using both synthetic and real data, we demonstrate superior performance of our method, showing that the method and its asymptotic behavior not only solve constrained stochastic problems efficiently but also provide valid and practical online inference in real-world applications.

Keywords: constrained model inference; primal-dual minimax optimality; stochastic SQP; gradient momentum

1 Introduction

We consider stochastic optimization problems with equality and box inequality constraints:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f(\boldsymbol{x}) = \mathbb{E}_{\zeta\sim\mathcal{P}}\left[F(\boldsymbol{x};\zeta)\right] \quad \text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}. \tag{1}$$

Here, $\boldsymbol{\ell}, \boldsymbol{u} \in \mathbb{R}^d$ denote the lower and upper bounds, respectively, with "$\le$" representing element-wise comparison; $F(\cdot;\zeta) : \mathbb{R}^d \to \mathbb{R}$ denotes a realization of the stochastic objective f; and $\boldsymbol{c} : \mathbb{R}^d \to \mathbb{R}^m$ encodes deterministic equality constraints. Throughout the paper, we assume f, $\boldsymbol{c}$, $F(\cdot;\zeta)$ are twice continuously differentiable while potentially nonconvex and nonlinear. Problem (1) readily accommodates nonlinear inequality constraints $\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}) \le \boldsymbol{0}$ by introducing slack variables, i.e., reformulating them as $\boldsymbol{c}_{\mathcal{I}}(\boldsymbol{x}) + \boldsymbol{y} = \boldsymbol{0}$ and $\boldsymbol{y} \ge \boldsymbol{0}$.

Problem (1) is ubiquitous with many statistical and machine learning applications, where constraints can encode prior domain knowledge, ensure models' identifiability, and reduce models' intrinsic dimensionality. We introduce concrete motivating examples in Appendix A. Given the ubiquity of Problem (1), statisticians aim to estimate its (local) solution $\boldsymbol{x}^\star$ and perform statistical inference. Arguably, the most primitive estimator is constrained M-estimator, where we draw n samples $\zeta_1, \ldots, \zeta_n \overset{\text{iid}}{\sim} \mathcal{P}$ and estimate the population loss f by the empirical loss $\widehat{f}_n$:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad \widehat{f}_n(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x};\zeta_i) \quad \text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}. \tag{2}$$

In fact, it is known that the above constrained M-estimation attains *asymptotic optimality* in Hájek and Le Cam's sense (Hájek, 1972; Le Cam, 1972). Roughly speaking, under certain regularity conditions, the (local) minimizer $\widehat{\boldsymbol{x}}_n$ of Problem (2) exhibits asymptotic normality with the smallest covariance matrix, given by

$$\sqrt{n}\left(\widehat{\boldsymbol{x}}_n - \boldsymbol{x}^\star\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{L}^\dagger \mathrm{Cov}\left(\nabla F(\boldsymbol{x}^\star;\zeta)\right)\boldsymbol{L}^\dagger\right), \tag{3}$$

where $\boldsymbol{L} = \boldsymbol{P}_{\mathrm{Null}(\boldsymbol{J}^\star)}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star\boldsymbol{P}_{\mathrm{Null}(\boldsymbol{J}^\star)}$ and $\dagger$ denotes the Moore-Penrose pseudoinverse. Here, we let

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \coloneqq f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top\boldsymbol{c}(\boldsymbol{x}) + \boldsymbol{\mu}_1^\top(\boldsymbol{\ell}-\boldsymbol{x}) + \boldsymbol{\mu}_2^\top(\boldsymbol{x}-\boldsymbol{u}) \tag{4}$$

be the Lagrangian function of (1) with $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \in \mathbb{R}^{2d}$ being the dual variables

associated with the equality and inequality constraints; $\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star = \nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ be the Lagrangian Hessian at the primal-dual solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ (with respect to $\boldsymbol{x}$); and $\boldsymbol{P}_{\mathrm{Null}(\boldsymbol{J}^\star)}$ denote the projection matrix onto the null space of $\boldsymbol{J}^\star = \boldsymbol{J}(\boldsymbol{x}^\star)$, where $\boldsymbol{J}(\boldsymbol{x})$ is the Jacobian of the active constraints at $\boldsymbol{x}$ (see (9)). We refer to Duchi and Ruan, 2021; Davis et al., 2024 for the rigorous statement of the result (3); and see Theorem 2.5 for primal-dual generalization.

Although $\widehat{\boldsymbol{x}}_n$ enjoys nice properties, offline M-estimation requires dealing with a full batch of samples, leading to significant computational costs. Over the past decades, online methods that can handle online streaming data have been preferred. One of the most fundamental online methods is Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952). Recent work has significantly advanced local asymptotic analysis of SGD (and its variates) to enable online inference of $\boldsymbol{x}^\star$. See, e.g., Polyak and Juditsky, 1992; Toulis and Airoldi, 2017; Fang et al., 2018; Toulis et al., 2021; Chen et al., 2020; 2024; Zhu et al., 2021; Lee et al., 2022.

The above literature all studied unconstrained stochastic optimization, while to accommodate constraints, Duchi and Ruan, 2021 designed a Projected Riemannian Stochastic Gradient method, which to our knowledge is the *first fully online* method capable of achieving asymptotic optimality for constrained model parameters $\boldsymbol{x}^\star$. The authors also identified a gap between problems with linear and nonlinear constraints, illustrating the reason why vanilla Projected SGD,

$$\boldsymbol{x}_{k+1} = \mathrm{Proj}_\Omega\left(\boldsymbol{x}_k - \alpha_k \nabla F(\boldsymbol{x}_k; \zeta_k)\right), \qquad \Omega = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}\}, \tag{5}$$

can be sub-optimal for nonlinear constraints. Notably, the gap has recently been closed by Davis et al., 2024, showing that the averaged Projected SGD, $\bar{\boldsymbol{x}}_n = \frac{1}{n}\sum_{k=0}^{n-1} \boldsymbol{x}_k$, actually attains the same asymptotic optimality for nonlinear constraints as displayed in (3). A consistent limiting covariance estimator has also been proposed in the follow-up work Jiang et al., 2025.

Despite these advances in online constrained inference, existing methods rely on a projection operator onto the feasible set $\mathrm{Proj}_\Omega(\cdot)$, making them, as stated in Duchi and Ruan, 2021, *more intellectually intriguing than practically applicable*. In particular, projections require solving non-

linear equations that is often intractable and needs global knowledge of the feasible set Ω. However, many problems (cf. CUTEst in Section 5.1) only have access to black-box oracles that return the local constraint value and Jacobian at a given point, making the projection inapplicable. Even in the ideal case where we have global information of the feasible set, the projection at each step may only be approximately computed by solving an intensive optimization subproblem; yet the resulting approximation error is typically neglected in existing analyses, which rely heavily on exact computation of $\mathrm{Proj}_{\Omega}(\cdot)$. Furthermore, a key distinction between constrained and unconstrained problems lies in the presence of dual multipliers. The dual component is not involved in existing methods, yet it plays a crucial role in stationarity certification and active-set identification (ensured by complementary slackness theorem; see (7)). Thus, performing primal-dual joint inference is of fundamental interest. This motivation leads to the following question: *What is the asymptotic minimax optimality for the primal–dual solution* $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ *of Problem* (1)*, and how to obtain the optimal joint online estimator?*

1.1 Backbone: Sequential Quadratic Programming

In this paper, we first establish primal-dual asymptotic minimax optimality of $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$, and then design a *Stochastic Sequential Quadratic Programming* (SSQP) method to attain the lower bound. The idea is to *localize* the model along with constraints, and utilize dual information of constrained local model when updating the estimate. In particular, at the k-th step, we perform a quadratic approximation of the loss and a linear approximation of the constraints in (1):

$$\min_{\bar{\Delta}\boldsymbol{x}_k \in \mathbb{R}^d} \nabla F(\boldsymbol{x}_k;\zeta_k)^\top \bar{\Delta}\boldsymbol{x}_k + \frac{1}{2}\bar{\Delta}\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \bar{\Delta}\boldsymbol{x}_k, \quad \text{s.t.} \quad \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \bar{\Delta}\boldsymbol{x}_k = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k \le \boldsymbol{u}. \tag{6}$$

Here, we denote the constraint value and Jacobian by $\boldsymbol{c}_k = \boldsymbol{c}(\boldsymbol{x}_k)$ and $\nabla \boldsymbol{c}_k = \nabla \boldsymbol{c}(\boldsymbol{x}_k)$; and $\bar{\boldsymbol{B}}_k$, rather than estimating $\nabla^2 f_k$, estimates the Lagrangian Hessian $\nabla^2_{\boldsymbol{x}} \mathcal{L}_k$ to leverage the curvature information of nonlinear constraints. After solving (6) and obtaining its primal-dual solution $(\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\mathrm{sub}}, \bar{\boldsymbol{\mu}}_k^{\mathrm{sub}})$, we update the iterate as

$$(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}_{k+1}, \boldsymbol{\mu}_{k+1}) = (\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) + \alpha_k(\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\mathrm{sub}} - \boldsymbol{\lambda}_k, \bar{\boldsymbol{\mu}}_k^{\mathrm{sub}} - \boldsymbol{\mu}_k).$$

Although the above scheme streamlines the estimation procedure by solving a linear–quadratic program at each step, two fundamental challenges remain. First, even if the original nonlinear constraints $\boldsymbol{c}(\boldsymbol{x}) = \mathbf{0}$, $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$ admit a nonempty feasible set Ω (cf. (5)), their linearization in (6) may become infeasible. Second, the inequality constraints truncate the noise in a way such that, even if $\nabla F(\boldsymbol{x}_k; \zeta_k)$ is an unbiased estimate of the true gradient ∇f_k, the resulting step direction $\bar{\Delta}\boldsymbol{x}_k$ can still be biased (cf. Remark 3.3), thereby preventing convergence of the scheme. To address the two challenges, we respectively introduce a constraint relaxation technique and a gradient (and Hessian) moving average technique. Although momentum-based gradient method has been widely studied in optimization context since the introduction in Polyak, 1964, our method is the *first* attempt to incorporate both gradient and Hessian momentum into inequality-constrained problems and show promising asymptotic properties to facilitate inference. Furthermore, our method allows an adaptive random stepsize α_k, which is often preferred in practice.

Specifically, with all above designs, we show that the *last* SSQP iterate exhibits asymptotic normality, with an optimal primal-dual covariance matrix matching the theoretical lower bound. In particular, the marginal primal covariance of the SSQP estimator also matches that of the offline constrained M-estimator and online averaged Projected SGD estimator, as shown in (3). Moreover, we propose a plug-in limiting covariance matrix estimator to facilitate practical inference. We demonstrate the superiority of our method through extensive experiments on benchmark nonlinear problems, constrained generalized linear models, and portfolio allocation problems with both synthetic data and real `Fama-French Portfolios` and `Chicago Air Pollution` data. Note that projection-based methods may not be applicable or competitive in the experiments (e.g., Duchi and Ruan, 2021 considered a linear regression problem with $d = 2$).

1.2 Related literature and contribution

Inspired by the profound success of deterministic SQP in numerical optimization (Nocedal and Wright, 2006), recent research has extended SQP framework to stochastic settings (Na et al.,

2022; 2023; Na, 2025; Curtis et al., 2023; 2024; Berahas et al., 2021; 2023; Fang et al., 2024a; b; 2025). However, existing work has primarily focused on equality-constrained problems. Only Na et al., 2023; Curtis et al., 2024 considered inequality constraints, yet both required strong constraint qualification conditions to circumvent infeasibility issue and a gradually increasing batch size for convergence, making them inapplicable in online settings. Moreover, all above numerical work studied only convergence of SSQP but fell short on uncertainty quantification. In contrast, we not only inherit the strengths of the above numerical methods by allowing adaptive stepsizes, but also significantly refine their designs to fit in *fully online* settings. Furthermore, we bridge the two worlds of optimization and statistics, leveraging the SSQP methods to enable online constrained statistical inference, thereby addressing practical limitations of projection-based methods as stated in Duchi and Ruan, 2021; Davis et al., 2024; Jiang et al., 2025.

This paper also relates to the growing literature on online inference via second-order methods. Bercu et al., 2020; Boyer and Godichon-Baggioni, 2022; Leluc and Portier, 2023; Cénac et al., 2025 studied online Newton (or conditioned SGD) methods for (non)linear regression problems, establishing asymptotic normality for the resulting Newton estimators. Kuang et al., 2025 proposed a consistent covariance estimator for more general sketched Newton methods. Na and Mahoney, 2025 established asymptotic normality for a sketched SSQP method under equality constraints, while Du et al., 2025 leveraged a random scaling technique to construct pivotal statistics. Due to the lack of inequality constraints, none of these works address the challenges of infeasibility and biased step directions discussed in Section 1.1. In fact, to our knowledge, establishing asymptotic distribution of online methods with gradient (and Hessian) momentum is open even in unconstrained settings, while this work establishes such results for inequality-constrained problems. We clearly demonstrate how the (adaptive) stepsize and momentum weight are interrelated: gradient averaging reduces direction bias to facilitate convergence, while past gradient information is also forgotten fast enough to attain optimal asymptotic normality.

1.3 Organization and notation

Section 2 introduces the preliminaries of Problem (1), including the infeasibility issue, constraint qualification, and constraint relaxation. We also establish the primal–dual asymptotic minimax optimality of Problem (1). Section 3 introduces the proposed SSQP method utilizing gradient moving average, and establishes its global almost-sure convergence guarantee. Section 4 presents the asymptotic normality result and proposes a consistent plug-in covariance matrix estimator. Section 5 reports extensive experimental results on both synthetic and real data, followed by conclusions and future work discussions in Section 6.

Throughout the paper, we use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the spectral norm for matrices. We use boldface lowercase and capital letters ($\boldsymbol{a}$ and $\boldsymbol{A}$) to denote vectors and matrices, respectively. For a positive integer m, $[m] \coloneqq \{1, \cdots, m\}$. For an index set $\mathcal{A} \subseteq [m]$, $\mathcal{A}^- \coloneqq [m] \backslash \mathcal{A}$ and $|\mathcal{A}|$ denotes its cardinality. For a vector $\boldsymbol{a} \in \mathbb{R}^m$, $\boldsymbol{a}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ denotes the subvector containing only entries with indices in $\mathcal{A}$. Similarly, for a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$, $\boldsymbol{A}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}| \times d}$ denotes the submatrix containing only rows with indices in $\mathcal{A}$. Without ambiguity, the i-th entry of a vector $\boldsymbol{a}$ is written as either $[\boldsymbol{a}]_i$ or a_i. We use $\mathcal{O}(\cdot)$ to denote the standard big-O notation: $a_k = \mathcal{O}(b_k)$ if $|a_k / b_k| \leq C$ for some constant $C > 0$ and large enough k. We use $\odot$ to denote the Hadamard (element-wise) product. For notational brevity, we write $\boldsymbol{c}_k \coloneqq \boldsymbol{c}(\boldsymbol{x}_k)$ and $f_k \coloneqq f(\boldsymbol{x}_k)$ (similarly for $\nabla \boldsymbol{c}_k$, etc.) as abbreviations.

2 Preliminary and Asymptotic Optimality

Recall the Lagrangian function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ of Problem (1) is defined in (4). Under proper constraint qualifications, a necessary condition for $\boldsymbol{x}^\star$ being a local solution of (1) is the KKT conditions: there exist dual multipliers $\boldsymbol{\lambda}^\star \in \mathbb{R}^d$ and $\boldsymbol{\mu}^\star = (\boldsymbol{\mu}_1^\star, \boldsymbol{\mu}_2^\star) \in \mathbb{R}^{2d}$ such that

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) = \boldsymbol{0}, \qquad \boldsymbol{c}(\boldsymbol{x}^\star) = \boldsymbol{0}, \qquad \boldsymbol{\ell} \leq \boldsymbol{x}^\star \leq \boldsymbol{u}, \\ \boldsymbol{\mu}^\star \geq \boldsymbol{0}, \qquad (\boldsymbol{\mu}_1^\star)^\top (\boldsymbol{\ell} - \boldsymbol{x}^\star) = (\boldsymbol{\mu}_2^\star)^\top (\boldsymbol{x}^\star - \boldsymbol{u}) = 0. \end{aligned} \tag{7}$$

At any $\boldsymbol{x}$, we let $\nabla \boldsymbol{c}(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ be the Jacobian of $\boldsymbol{c}(\boldsymbol{x})$. We denote active sets for inequalities as

$$\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}) = \{i \in [d] : [\boldsymbol{x}]_i = [\boldsymbol{\ell}]_i\} \qquad \text{and} \qquad \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}) = \{i \in [d] : [\boldsymbol{x}]_i = [\boldsymbol{u}]_i\}. \tag{8}$$

The active constraints include equality and active inequality constraints, $(\boldsymbol{c}(\boldsymbol{x}); [\boldsymbol{\ell}-\boldsymbol{x}]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})}; [\boldsymbol{x}-\boldsymbol{u}]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})})$; and their full Jacobian is given by

$$\boldsymbol{J}(\boldsymbol{x}) := \begin{pmatrix} \nabla \boldsymbol{c}(\boldsymbol{x}) \\ -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \\ \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \end{pmatrix} \in \mathbb{R}^{(m+|\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})|+|\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})|)\times d}, \tag{9}$$

where $\boldsymbol{I} \in \mathbb{R}^{d\times d}$ is the d-dimensional identity matrix.

2.1 Infeasibility and constraint relaxation

Compared to the projection operator $\mathrm{Proj}_{\Omega}(\cdot)$ in (5), the constraint linearization in (6) requires only local information and leads to a much simpler subproblem. However, an intrinsic difficulty of linearization is the issue of *infeasibility*: even if the original constraint set $\Omega \neq \emptyset$, the linearized constraints may end up with an empty set.

Consider an illustrative example with $d = m = 1$, where we have $c(x) = x^2 - 4$ and constraints $0 \le x \le 3$. We know the constraint set $\Omega = \{x = \pm 2\} \cap [0, 3] = \{2\}$ is nonempty. However, for any $x_k \in (0, 3]$, the linearized equality constraint is $x_k^2 - 4 + 2x_k\bar{\Delta}x_k = 0 \Leftrightarrow \bar{\Delta}x_k = (4-x_k^2)/(2x_k)$, depicted by the red curve in Figure 1; while the linearized inequality constraint is $-x_k \le \bar{\Delta}x_k \le 3 - x_k$, depicted by the shaded region. From Figure 1, we see that the linearized equality and inequality constraints intersect only if $x_k \ge 3-\sqrt{5}$, suggesting an infeasibility issue for Problem (6) when $x_k \in (0, 3-\sqrt{5})$. Thus, we need constraint relaxation in this regard.

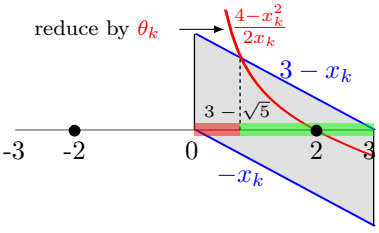


Figure 1: *Insights of our constraint relaxation.*

The constraint relaxation introduced in this paper stems from the observation that the nonlinear curve in Figure 1 can be flattened by a proper scaling factor $\theta_k \in (0, 1]$, so that the curve has more intersections with the shaded region. In particular, we consider the relaxed version of (6):

$$\theta_k \cdot \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \bar{\Delta}\boldsymbol{x}_k = \boldsymbol{0}, \qquad \boldsymbol{\ell} \le \boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k \le \boldsymbol{u}. \tag{10}$$

Naturally, we should investigate the conditions under which a relaxation parameter θ_k exists or fails to exist. We find that this is closely tied to the *extended generalized Mangasarian-Fromowitz constraint qualification* (EGMFCQ, Xu et al., 2015, Definition 2.4), which is weaker than *linear independence constraint qualification* (LICQ) that needs to be imposed for asymptotic minimax optimality analysis nevertheless; see Duchi and Ruan, 2021, Assumption B and Davis et al., 2024, Example 2.1 for references. Thus, our relaxation does not comes with stronger conditions.

Definition 2.1 (EGMFCQ v.s. LICQ). *The extended generalized Mangasarian-Fromowitz constraint qualification is said to hold at $\boldsymbol{x} \in \mathbb{R}^d$ satisfying $\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}$ if the two conditions are met:*

(a) The Jacobian $\nabla \boldsymbol{c}(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ has full row rank.

(b) There exists a vector $\boldsymbol{z} \in \mathbb{R}^d$ such that

$$\boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{z} = \boldsymbol{0}, \qquad [\boldsymbol{z}]_i > 0 \ \text{ for } \ i \in \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}), \qquad [\boldsymbol{z}]_i < 0 \ \text{ for } \ i \in \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}). \tag{11}$$

The linear independence constraint qualification is said to hold at $\boldsymbol{x} \in \mathbb{R}^d$ if the Jacobian of active constraints $\boldsymbol{J}(\boldsymbol{x})$ defined in (9) has full row rank.

Setting $\boldsymbol{p} = -(\boldsymbol{c}(\boldsymbol{x}); \boldsymbol{1}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})}; \boldsymbol{1}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})})$, we note that LICQ ensures the linear system $\boldsymbol{J}(\boldsymbol{x})\boldsymbol{z} = \boldsymbol{p}$ has a solution, and such a solution satisfies (11). Hence, LICQ implies EGMFCQ. We further establish the connection between the constraint relaxation in (10) and EGMFCQ.

Theorem 2.2. *For any point $\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}$, if EGMFCQ holds at $\boldsymbol{x}$, there exists a threshold $\bar{\theta} \in (0, 1]$ such that for any $\theta \in [0, \bar{\theta}]$,*

$$\Omega(\boldsymbol{x}; \theta) \coloneqq \{\boldsymbol{z} \in \mathbb{R}^d : \theta \cdot \boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{z} = \boldsymbol{0},\ \boldsymbol{\ell} \le \boldsymbol{x} + \boldsymbol{z} \le \boldsymbol{u}\} \neq \emptyset.$$

Conversely, suppose a sequence of points $\boldsymbol{\ell} \le \boldsymbol{x}_k \le \boldsymbol{u}$, $\forall k \ge 0$, admits a sequence of "sharp" $\theta_k \in (0, 1]$ in the sense that $\Omega(\boldsymbol{x}_k, \theta_k) \neq \emptyset$ but $\Omega(\boldsymbol{x}_k, \min\{2\theta_k, 1\}) = \emptyset$. If there exists a subsequence $\{k_l\}_l$ such that $\lim_{l\to\infty} \theta_{k_l} = 0$, then EGMFCQ fails to hold at any accumulation point of $\{\boldsymbol{x}_{k_l}\}_l$.

Theorem 2.2 suggests that the condition of EGMFCQ ensures the existence of $\theta_k > 0$ that makes the linearized constraints (10) feasible, i.e., $\Omega(\boldsymbol{x}_k; \theta_k) \neq \emptyset$. Note that EGMFCQ or even LICQ is indeed widely imposed in constrained optimization problems, which essentially charac-

terizes the quality of linear approximation to nonlinear constraints (Nocedal and Wright, 2006). Based on Theorem 2.2, we can design the algorithm with constraint relaxation in the following way: at each step, we initialize $\theta_k = 1$ and backtrack reducing θ_k until $\Omega(\boldsymbol{x}_k; \theta_k) \neq \emptyset$. Note that checking $\Omega(\boldsymbol{x}_k; \theta_k) \neq \emptyset$ can be efficiently done by solving a convex quadratic program:

$$\min_{\boldsymbol{z} \in \mathbb{R}^d} \quad \|\theta_k \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \boldsymbol{z}\|^2 \qquad \text{s.t.} \qquad \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{z} \leq \boldsymbol{u}. \tag{12}$$

We have $\Omega(\boldsymbol{x}_k; \theta_k) \neq \emptyset$ if and only if the optimal value of the above program is 0. Furthermore, if the selected sequence $\{\theta_k\}$ is not bounded away from zero, Theorem 2.2 suggests that a subsequence of $\{\boldsymbol{x}_k\}$ will converge to a point that fails EGMFCQ, and certainly fails LICQ. In that case, we should re-initiate the method to explore other local solutions, since any estimation procedures with asymptotic normality, either offline M-estimator or online projection-based estimator, target only solutions that satisfy LICQ (Duchi and Ruan, 2021; Davis et al., 2024).

2.2 Primal-dual asymptotic optimality

In this section, we explore the primal-dual asymptotic minimax optimality of Problem (1). In particular, we target a KKT triplet $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ satisfying (7) and investigate what is the best estimator one can expect in the sense of Hájek and Le Cam. To that end, we introduce some assumptions and notations beforehand.

Assumption 2.3 (Strict Complementarity)**.** *We assume* $[\boldsymbol{\mu}_1^\star]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}^\star)} > \mathbf{0}$ *and* $[\boldsymbol{\mu}_2^\star]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}^\star)} > \mathbf{0}$.

Assumption 2.4 (Second-Order Sufficient Condition)**.** *The Lagrangian Hessian with respect to* $\boldsymbol{x}$, $\nabla^2_{\boldsymbol{x}} \mathcal{L}^\star = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$, *is positive definite in the null space of active constraint Jacobian* $\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{J}(\boldsymbol{x}^\star)\boldsymbol{z} = \mathbf{0}\}$. *In other words, there exists* $\omega > 0$ *such that*

$$\boldsymbol{z}^\top \nabla^2_{\boldsymbol{x}} \mathcal{L}^\star \boldsymbol{z} \geq \omega \|\boldsymbol{z}\|^2 \qquad \textit{for all} \quad \boldsymbol{z} \in \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{J}(\boldsymbol{x}^\star)\boldsymbol{z} = \mathbf{0}\}.$$

Assumptions 2.3 and 2.4, together with LICQ at $\boldsymbol{x}^\star$, are standard in the sensitivity analysis of constrained nonlinear optimization (Shapiro, 1990; Bonnans and Shapiro, 2000), and are also imposed for minimax optimality analysis of the primal solution $\boldsymbol{x}^\star$ in the literature (Duchi and

Ruan, 2021; Davis et al., 2024). These assumptions essentially require that the KKT triplet $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ is a strict local solution of (1), admitting a local convexity structure in the sense that the Lagrangian is strongly convex in the tangent space of the constraints manifold at $\boldsymbol{x}^\star$.

For any $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \in \mathbb{R}^{2d}$, let us denote active dual components by $\boldsymbol{\mu}_{\mathcal{A}(\boldsymbol{x})} = ([\boldsymbol{\mu}_1]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})}, [\boldsymbol{\mu}_2]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})}) \in \mathbb{R}^{|\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})|+|\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})|}$ and inactive dual components by $\boldsymbol{\mu}_{\mathcal{A}(\boldsymbol{x})^-} = ([\boldsymbol{\mu}_1]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})^-}, [\boldsymbol{\mu}_2]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})^-}) \in \mathbb{R}^{2d-|\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})|-|\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})|}$. For notational simplicity, we write $\boldsymbol{\mu}^\star_{\mathcal{A}^\star} = \boldsymbol{\mu}^\star_{\mathcal{A}(\boldsymbol{x}^\star)}$ and note that $\boldsymbol{\mu}^\star_{(\mathcal{A}^\star)^-} = \boldsymbol{\mu}^\star_{\mathcal{A}(\boldsymbol{x}^\star)^-} = \mathbf{0}$ by the KKT conditions in (7). We also define the Lagrangian Hessian matrix with respect to $(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}_{\mathcal{A}(\boldsymbol{x})})$, a covariance matrix, and a sandwich matrix as

$$\boldsymbol{H}^\star = \begin{pmatrix} \nabla^2_{\boldsymbol{x}} \mathcal{L}^\star & (\boldsymbol{J}^\star)^\top \\ \boldsymbol{J}^\star & \mathbf{0} \end{pmatrix}, \quad \boldsymbol{\Sigma}^\star = \begin{pmatrix} \mathrm{Cov}(\nabla F(\boldsymbol{x}^\star; \zeta)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \boldsymbol{\Omega}^\star = (\boldsymbol{H}^\star)^{-1} \boldsymbol{\Sigma}^\star (\boldsymbol{H}^\star)^{-1}. \tag{13}$$

The non-singularity of $\boldsymbol{H}^\star$ is ensured by LICQ (Definition 2.1) and SOSC (Assumption 2.4).

The lower-bound analysis closely relates to the sensitivity analysis of Problem (1) with respect to the distribution $\mathcal{P}$. We will explore how the primal-dual solution as a function of $\mathcal{P}$, $(\boldsymbol{x}^\star(\mathcal{P}), \boldsymbol{\lambda}^\star(\mathcal{P}), \boldsymbol{\mu}^\star(\mathcal{P}))$, varies when $\mathcal{P}$ is perturbed slightly. As such, we measure the perturbation size via the ϕ-divergence, defined as

$$D_\phi(\mathcal{P}', \mathcal{P}) = \int \phi\left(\frac{d\mathcal{P}'}{d\mathcal{P}}\right) d\mathcal{P},$$

where $\phi : (0, \infty) \to \mathbb{R}$ is a C^3-smooth convex function satisfying $\phi(1) = 0$. We define an admissible neighborhood $\mathcal{B}(\mathcal{P}; \epsilon)$ of $\mathcal{P}$ to consist of all probability distributions $\mathcal{P}'$ with $D_\phi(\mathcal{P}', \mathcal{P}) \le \epsilon$, such that there exists a primal-dual solution $(\boldsymbol{x}^\star(\mathcal{P}'), \boldsymbol{\lambda}^\star(\mathcal{P}'), \boldsymbol{\mu}^\star(\mathcal{P}'))$ to the perturbed Problem (1). It is well known that the inactive constraint set $\mathcal{A}(\boldsymbol{x}^\star(\mathcal{P}))^-$ is preserved under small perturbations (cf. Shapiro, 1990 Theorem 1). That is, $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}^\star(\mathcal{P}'))^- = \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}^\star(\mathcal{P}))^-$ and $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}^\star(\mathcal{P}'))^- = \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}^\star(\mathcal{P}))^-$ for ϵ small enough; and we trivially have $\boldsymbol{\mu}^\star_{(\mathcal{A}^\star)^-}(\mathcal{P}) = [\boldsymbol{\mu}^\star(\mathcal{P})]_{\mathcal{A}(\boldsymbol{x}^\star(\mathcal{P}))^-} = \mathbf{0}$ in the neighborhood of $\mathcal{P}$. Below, we focus on the nontrivial part $\boldsymbol{w}^\star(\mathcal{P}) = (\boldsymbol{x}^\star(\mathcal{P}), \boldsymbol{\lambda}^\star(\mathcal{P}), \boldsymbol{\mu}^\star_{\mathcal{A}^\star}(\mathcal{P}))$.

The next theorem asymptotically lower bounds the performance of any estimation procedure for finding the solution to (1), by considering adversarially chosen small perturbations of the unperturbed distribution $\mathcal{P}$.

Theorem 2.5 (Local primal-dual minimax optimality). *Suppose Assumptions 2.3, 2.4, and LICQ hold at* $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$. *Let* $\ell : \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|} \to [0,\infty)$ *be any symmetric, quasiconvex, and lower semicontinuous function, and let* $\boldsymbol{w}_k(\mathcal{P}') : \zeta_{1:k} \to \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|}$ *be a sequence of estimators based on* k *samples* $\zeta_{1:k} \overset{iid}{\sim} \mathcal{P}'$. *Then, the following inequality holds:*

$$\lim_{c\to+\infty} \liminf_{k\to+\infty} \sup_{\mathcal{P}'\in\mathcal{B}(\mathcal{P};c/k)} \mathbb{E}_{\mathcal{P}_k'}\left[\ell\left(\sqrt{k}\left(\boldsymbol{w}_k(\mathcal{P}') - \boldsymbol{w}^\star(\mathcal{P}')\right)\right)\right] \geq \mathbb{E}\left[\ell\left(Z\right)\right],$$

where $\mathbb{E}_{\mathcal{P}_k'}[\cdot]$ *is taken over* k *iid samples* $\zeta_{1:k}$ *and* $Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^\star)$.

Compared with (Davis et al., 2024, Theorem 3.2) and (Duchi and Ruan, 2021, Theorem 1), our result jointly captures both the primal solution and the active dual solution (since inactive dual solution remains zero). Notably, following a similar calculation as in (Na and Mahoney, 2025, Remark 5.9), we observe that the marginal covariance of $\boldsymbol{\Omega}^\star$ with respect to the primal solution matches those in the above works (i.e., (3)). Thus, Theorem 2.5 provides a complete and sharp characterization of local asymptotic optimality for constrained estimation problems.

3 SSQP with Polyak's Momentum

In this section, we introduce the proposed online SSQP estimation procedure for Problem (1), which leverages the constraint relaxation idea discussed in Section 2.1 and employs gradient (and Hessian) moving average schemes based on Polyak's momentum to debias the step direction.

3.1 Estimation procedure

The estimation procedure involves tuning parameters $\tau \in (0,1)$, $\psi \geq 0$, $p \geq 1$, and one stepsize and two averaging weight sequences $\alpha_k, \beta_k, \gamma_k > 0$. Given the current k-th primal-dual iterate $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ with $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})$, we initialize $\theta_k = 1$ and perform

$$\theta_k \leftarrow \tau\theta_k \qquad \text{until} \qquad \text{convex problem (12) has an optimal value 0.} \tag{14}$$

The above scheme is supposed to terminate in finite time by Theorem 2.2. In particular, if θ_k falls below an infinitesimal threshold, our theory implies that $\boldsymbol{x}_k$ is approaching an undesirable

point at which even EGMFCQ (let alone LICQ) does not hold.

With the selected θ_k from (14), we then draw a single sample $\zeta_k \sim \mathcal{P}$ and estimate the objective gradient and Hessian as $\nabla F(\boldsymbol{x}_k;\zeta_k)$ and $\nabla^2 F(\boldsymbol{x}_k;\zeta_k)$, respectively. Instead of using these noisy estimates directly, we perform gradient and Hessian moving averages:

$$\bar{\boldsymbol{g}}_k = (1-\beta_k)\bar{\boldsymbol{g}}_{k-1} + \beta_k \nabla F(\boldsymbol{x}_k;\zeta_k) \qquad \text{and} \qquad \bar{\boldsymbol{Q}}_k = (1-\gamma_k)\bar{\boldsymbol{Q}}_{k-1} + \gamma_k \nabla^2 F(\boldsymbol{x}_k;\zeta_k). \tag{15}$$

The Lagrangian Hessian with respect to $\boldsymbol{x}$ depends on the constraint curvature information, which we estimate as

$$\bar{\boldsymbol{B}}_k = \bar{\boldsymbol{Q}}_k + \sum_{i=1}^{m} [\boldsymbol{\lambda}_k]_i \nabla^2 \boldsymbol{c}_i(\boldsymbol{x}_k) + \boldsymbol{\Delta}_k, \tag{16}$$

where $\boldsymbol{\Delta}_k \in \mathbb{R}^{d\times d}$ is a Hessian modification term. In fact, if only the global convergence[1] of the estimation procedure is of interest, we do not require any second-order information and can simply set $\bar{\boldsymbol{B}}_k = \boldsymbol{I}$. As such, the importance of the two averaging weight sequences, β_k and γ_k, is different: the choice of β_k is more delicate than that of γ_k, since first-order information plays a more crucial role in debiasing the step and facilitating asymptotic normality (cf. Theorem 4.6).

With the above estimators, we construct an SSQP subproblem similar to (6), but using averaged gradient (and Hessian) estimators along with the constraint relaxation:

$$\min_{\bar{\Delta}\boldsymbol{x}_k \in \mathbb{R}^d} \ \bar{\boldsymbol{g}}_k^\top \bar{\Delta}\boldsymbol{x}_k + \frac{1}{2}\bar{\Delta}\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \bar{\Delta}\boldsymbol{x}_k, \ \ \text{s.t.} \ \ \theta_k \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \bar{\Delta}\boldsymbol{x}_k = \boldsymbol{0}, \ \ \boldsymbol{\ell} \le \boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k \le \boldsymbol{u}. \tag{17}$$

Since solving nonconvex inequality-constrained quadratic program without any constraint qualifications can be NP-hard (Burer and Letchford, 2009), we require the Hessian modification $\boldsymbol{\Delta}_k$ to convexify $\bar{\boldsymbol{Q}}_k + \sum_{i=1}^m [\boldsymbol{\lambda}_k]_i \nabla^2 \boldsymbol{c}_i(\boldsymbol{x}_k)$ such that $\bar{\boldsymbol{B}}_k$ is positive definite (cf. Assumption 3.5). We slightly abuse notation by letting $\bar{\Delta}\boldsymbol{x}_k$ denote the primal solution of (17), and $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_k^{\text{sub}})$ with $\bar{\boldsymbol{\mu}}_k^{\text{sub}} = (\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})$ denote any (one of) dual solutions. We then update the iterate as follows:

$$(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}_{k+1}, \boldsymbol{\mu}_{k+1}) = (\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) + \bar{\alpha}_k \cdot (\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}_k, \bar{\boldsymbol{\mu}}_k^{\text{sub}} - \boldsymbol{\mu}_k), \tag{18}$$

where $\bar{\alpha}_k > 0$ is an adaptive, potentially random stepsize that satisfies a safeguard condition:

[1] In our context of nonlinear problems, global convergence refers to convergence to a stationary point from *any* initialization, as opposed to convergence to a global solution, which is generally unachievable without specific problem structures. That said, these two notions coincide in the case of convex problems.

$\alpha_k \le \bar{\alpha}_k \le \alpha_k + \psi\alpha_k^p$ for the tuning parameters $\psi > 0$ and $p \ge 1$.

Remark 3.1 (Uniqueness of subproblem dual solution)**.** *We should emphasize that the dual solution* $(\bar{\boldsymbol{\lambda}}_k^{sub}, \bar{\boldsymbol{\mu}}_k^{sub})$ *of the strongly convex SSQP subproblem* (17) *is not necessarily unique, as we do not impose strong constraint qualifications. In fact, in global convergence analysis, when evaluating the KKT residual at the iterate* $\boldsymbol{x}_k$*, we do not rely directly on the dual iterates* $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ *or* $(\bar{\boldsymbol{\lambda}}_k^{sub}, \bar{\boldsymbol{\mu}}_k^{sub})$*. Note that the KKT condition is defined solely in terms of the primal variable (see* (7)*); therefore, we only desire the existence of dual multipliers, possibly not unique, under which the KKT residual converges to zero. In other words, even though the dual variables of individual SSQP subproblems may not be unique, the global convergence of the KKT residual with respect to the primal iterates can still be established.*

On the other hand, our local analysis in Section 4 targets a local solution $\boldsymbol{x}^\star$ *at which LICQ holds, as required by minimax optimality guarantee in Section 2. In that case, we can examine the local convergence behavior of the dual variables* $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$*, since the SSQP subproblem admits a unique dual solution* $(\bar{\boldsymbol{\lambda}}_k^{sub}, \bar{\boldsymbol{\mu}}_k^{sub})$ *under LICQ. In particular, as* $\boldsymbol{x}_k$ *approaches* $\boldsymbol{x}^\star$ *satisfying LICQ, the inactive constraints at* $\boldsymbol{x}^\star$ *for the original problem* (1) *remain inactive at* $\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k$ *for the SSQP subproblem* (17)*, implying* (17) *has a unique dual solution* $(\bar{\boldsymbol{\lambda}}_k^{sub}, \bar{\boldsymbol{\mu}}_k^{sub})$*. Therefore, our local convergence analysis on the dual variable in Section 4 (cf. Lemma 4.4) is well justified.*

Remark 3.2 (Adaptive stepsize)**.** *Compared to online projection-based methods (Duchi and Ruan, 2021; Davis et al., 2024), we allow using a random adaptive stepsize* $\bar{\alpha}_k$*, as long as* $\bar{\alpha}_k$ *that may depend on the step direction is well controlled within the interval* $[\alpha_k, \alpha_k + \psi\alpha_k^p]$*. Here, the interval length* $\psi\alpha_k^p$ *measures the degree of adaptivity. Setting* $\psi = 0$ *corresponds to deterministic stepsize. The adaptive designs have been proposed for SSQP schemes in numerical optimization literature (Curtis et al., 2024; Berahas et al., 2021; Fang et al., 2024a; Na and Mahoney, 2025), where the stepsize is chosen via either line-search or trust-region strategies. All the proposed stepsize rules in those works satisfy the safeguard condition. Motivated by promising empirical*

performance of adaptive methods, we adopt a similar design of stepsize in our estimation procedure. This additional flexibility, however, makes the statistical inference analysis more involved, as it requires controlling the effect of adaptivity on the limiting distribution.

Remark 3.3 (Bias in step direction)**.** *Compared to SSQP designs for equality-constrained problems, a crucial challenge posed by inequality constraints is the bias in the step direction estimation.*

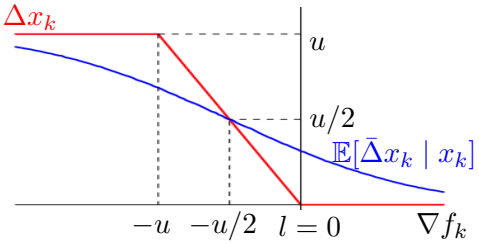


Figure 2: *Illustration of bias in step direction estimation.*

Consider the following example on $\mathbb{R}$*, where we let* $\bar{B}_k = 1$ *and suppress equality constraints in* (6)*:* $\min_{\bar{\Delta} x_k \in \mathbb{R}} 0.5\bar{\Delta} x_k^2 + \nabla F(x_k;\zeta_k)\bar{\Delta} x_k$ *subject to* $0 = l \le \bar{\Delta} x_k \le u$*. Without inequality constraints, this reduces to SGD with* $\Delta x_k = -\nabla f_k = \mathbb{E}[-\nabla F(x_k;\zeta_k) \mid x_k] = \mathbb{E}[\bar{\Delta} x_k \mid x_k]$*, as long as* $\nabla F(x_k;\zeta_k)$ *is unbiased. However, with inequality constraints, we have* $\Delta x_k = -Project(\nabla f_k, [-u, -l])$ *shown as the red line in Figure 2, differing from*

$$\mathbb{E}[\bar{\Delta} x_k \mid x_k] = -\mathbb{E}[Project(\nabla F(x_k;\zeta_k), [-u, -l]) \mid x_k],$$

shown as the blue line in Figure 2 with Gaussian noise $\nabla F(x_k;\zeta_k) \sim \mathcal{N}(\nabla f_k, \sigma^2)$*. In other words, inequality constraints truncate the gradient noise in a nonlinear manner, such that an unbiased gradient estimate does not necessarily yield an unbiased step direction estimate. To address this issue, we leverage momentum-based averaging estimators in* (15)*. We show that* $\bar{\Delta} x_k$ *with averaged gradient estimators satisfies* $\mathbb{E}[\|\bar{\Delta} x_k - \Delta x_k\|^2] \to 0$ *as* $k \to \infty$*, i.e., it is asymptotically unbiased (cf. Lemmas D.1 and D.2). One significant contribution of this paper is to further show that gradient momentum simultaneously preserves asymptotic optimality while debiasing step directions, as long as the weight sequence* β_k *is chosen properly according to the stepsize* α_k*.*

3.2 Global almost-sure convergence

In this section, we establish global almost-sure convergence for the proposed SSQP estimation procedure in Section 3.1. We show that the KKT residual of $\boldsymbol{x}_k$ converges to zero from any initialization. Our analysis generalizes existing SSQP literature on numerical optimization (Na

et al., 2023; Curtis et al., 2024) and relaxes restrictive constraint qualification conditions.

We define an ℓ_2-penalized Lyapunov function to quantify the progress toward stationarity:

$$\varphi_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho\|\boldsymbol{c}(\boldsymbol{x})\| \qquad \text{for} \quad \rho > 0. \tag{19}$$

The penalty term biases the feasibility error of equality constraints, while inequality constraints are not penalized since SSQP iterates always stay within the box constraints (cf. Lemma C.3). In contrast to unconstrained problems, where methods aim to reduce the objective function f, reducing f alone is not sufficient to justify the iteration progress toward stationarity in constrained problems, since a small objective value may be achieved by severely violating the constraints. We will show that our method decreases the Lyapunov function (19) for sufficiently large $\rho > 0$, thereby vanishing the KKT residual, which is defined at the primal-dual triplet by

$$\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \boldsymbol{c}(\boldsymbol{x}) \\ \boldsymbol{\mu}_1 \odot (\boldsymbol{\ell} - \boldsymbol{x}) \\ \boldsymbol{\mu}_2 \odot (\boldsymbol{x} - \boldsymbol{u}) \end{pmatrix}. \tag{20}$$

Here, $\odot$ denotes the Hadamard elementwise product and each entry of $\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ corresponds to one KKT condition in (7). Again, we do not measure violation of conditions $\boldsymbol{\mu} \geq 0$ and $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$ since they hold at every step. Thus, $\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{0}$ suggests we arrive at stationarity.

3.2.1 Assumptions

Now, we state the assumptions for global analysis. Assumption 3.4 is tied to our constraint relaxation technique, which has been justified by Theorem 2.2 under EGMFCQ (weaker than LICQ). Assumptions 3.5 and 3.6 are standard regularity conditions commonly imposed in the literature.

Assumption 3.4. *There exists a deterministic constant $\bar{\theta} \in (0, 1]$ such that the linearized constraints (10) are feasible for any relaxation parameter $\theta_k \in [0, \bar{\theta}]$. That is, $\Omega(\boldsymbol{x}_k; \theta_k) \neq \emptyset$ for $\theta_k \in [0, \bar{\theta}]$.*

Assumption 3.4 requires the relaxation parameter θ_k to be bounded away from zero to ensure subproblem feasibility. By Theorem 2.2, if θ_k is not bounded below, any accumulation point of the iterates $\boldsymbol{x}_k$ fails to satisfy EGMFCQ and, certainly, LICQ. In that case, it becomes intractable

to perform statistical inference for such an accumulation point even using offline M-estimation or online projected methods, since asymptotic normality result in (3) of both methods, as well as primal-dual minimax optimality in Theorem 2.5, crucially relies on LICQ.

Assumption 3.5. *We assume ∇f, $\nabla \boldsymbol{c}$ are Lipschitz continuous; that is, for all $\boldsymbol{\ell} \leq \boldsymbol{x}, \boldsymbol{y} \leq \boldsymbol{u}$,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \kappa_{\nabla f} \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \|\nabla \boldsymbol{c}(\boldsymbol{x}) - \nabla \boldsymbol{c}(\boldsymbol{y})\| \leq \kappa_{\nabla c} \|\boldsymbol{x} - \boldsymbol{y}\|$$

for some constants $\kappa_{\nabla f}, \kappa_{\nabla c} > 0$. Furthermore, the Hessian modification term $\boldsymbol{\Delta}_k \in \mathbb{R}^{d \times d}$ in (16) ensures $\bar{\boldsymbol{B}}_k$ satisfy $\kappa_1 \boldsymbol{I} \preceq \bar{\boldsymbol{B}}_k \preceq \kappa_2 \boldsymbol{I}$ for some $0 < \kappa_1 \leq \kappa_2$. In addition, we assume the dual multipliers of the SQP subproblem (17) with the exact true gradient ∇f_k, denoted by $(\boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_k^{sub})$, are bounded; that is, there exists $M_{dual} > 0$ such that $\max\{\|\boldsymbol{\lambda}_k^{sub}\|, \|\boldsymbol{\mu}_k^{sub}\|\} \leq M_{dual}$.

The Lipschitz condition is standard and naturally holds within a closed bounded feasible set (Curtis et al., 2024; Berahas et al., 2023). The condition on $\bar{\boldsymbol{B}}_k$ ensures the subproblem (17) is strongly convex and admits a unique solution; this requirement can be readily relaxed to the SOSC condition, as stated in Assumption 2.4, when a strong constraint qualification condition, e.g., LICQ, is imposed (see discussion after Assumption 4.1 in Section 4). The boundedness of dual multipliers in (deterministic) SQP subproblems is a conventional assumption in constrained optimization. We show in Theorem E.1 (Appendix E) that if the accumulation point of the iteration sequence $\boldsymbol{x}_k$ is feasible and satisfies EGMFCQ, then the multipliers indeed remain bounded. This result provides a theoretical justification for the boundedness assumption.

Assumption 3.6. *We assume, for any $k \geq 0$, $\nabla F(\boldsymbol{x}_k; \zeta_k)$ is an unbiased estimate of ∇f_k with bounded variance. In particular, $\mathbb{E}_k[\nabla F(\boldsymbol{x}_k; \zeta_k)] = \nabla f_k$ and $\mathbb{E}_k[\|\nabla F(\boldsymbol{x}_k; \zeta_k) - \nabla f_k\|^2] \leq \sigma_g^2$ for some $\sigma_g^2 > 0$. Here, $\mathbb{E}_k[\cdot]$ denotes the conditional expectation given $\boldsymbol{x}_k$.*

The unbiasedness and bounded variance of stochastic gradients are standard assumptions in both unconstrained and constrained stochastic optimization (Polyak and Juditsky, 1992; Chen et al., 2020; Curtis et al., 2024; Na and Mahoney, 2025).

3.2.2 Almost-sure convergence

The Lyapunov function φ_ρ in (19) serves as a surrogate of the original constrained problem, providing a unified measure of objective value and constraint violation. Drawing intuition from unconstrained optimization, a decrease in the Lyapunov function can be interpreted as progress toward stationarity of constrained problem. Thus, we analyze the decrease of the Lyapunov function $\varphi_\rho(\boldsymbol{x})$ along the sequence of iterates. As such, we consider a local quadratic model of $\varphi_\rho(\boldsymbol{x})$:

$$\varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \Delta\boldsymbol{x}, \boldsymbol{B}) \coloneqq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^\top \boldsymbol{B}\Delta\boldsymbol{x} + \rho \left\|\boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\Delta\boldsymbol{x}\right\|, \tag{21}$$

and its difference measures the reduction of φ_ρ at $\boldsymbol{x}$ along the step $\Delta\boldsymbol{x}$:

$$\Delta\varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \Delta\boldsymbol{x}, \boldsymbol{B}) \coloneqq \varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \boldsymbol{0}, \boldsymbol{B}) - \varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \Delta\boldsymbol{x}, \boldsymbol{B}). \tag{22}$$

The next lemma establishes a reduction argument for the local model difference $\Delta\varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \Delta\boldsymbol{x}, \boldsymbol{B})$, showing that the reduction at each step is proportional to both the step magnitude $\|\Delta\boldsymbol{x}_k\|^2$ and the constraint violation $\|\boldsymbol{c}_k\|$. Recall from Section 3.1 that $\bar{\Delta}\boldsymbol{x}_k$ and $(\bar{\boldsymbol{\lambda}}_k^{\mathrm{sub}}, \bar{\boldsymbol{\mu}}_k^{\mathrm{sub}})$ denote the primal and dual solutions of the subproblem (17), while $\Delta\boldsymbol{x}_k$ and $(\boldsymbol{\lambda}_k^{\mathrm{sub}}, \boldsymbol{\mu}_k^{\mathrm{sub}})$ denote their counterparts when replacing gradient estimate $\bar{\boldsymbol{g}}_k$ by the true gradient ∇f_k in (17).

Lemma 3.7. *Under Assumptions 3.4 and 3.5, for any $\nu \in (0,1)$, as long as $\rho \geq \bar{\rho} \coloneqq \frac{M_{dual}}{(1-\nu)\tau\theta}$,*

$$\Delta\varphi_\rho^{loc}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) \geq \frac{1}{2}\Delta\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k + \nu\rho\theta_k \left\|\boldsymbol{c}_k\right\| \quad \text{for all } \ k \geq 0.$$

Since $\varphi_\rho^{\mathrm{loc}}(\boldsymbol{x}, \Delta\boldsymbol{x}, \boldsymbol{B})$ closely approximates the Lyapunov function $\varphi_\rho(\boldsymbol{x})$, the above result implies a decrease in φ_ρ itself. Consequently, the convergence of the iterates toward a stationary point is closely linked to the vanishing of $\|\Delta\boldsymbol{x}_k\|$ and $\|\boldsymbol{c}_k\|$.

Theorem 3.8. *Under Assumptions 3.4, 3.5, and 3.6, we specify the stepsize and averaging weight sequences as $\alpha_k = \iota_1(k+1)^{-b_1}$ and $\beta_k = \iota_2(k+1)^{-b_2}$, for constants $\iota_1, \iota_2 > 0$ and exponents b_1, b_2 satisfying $b_1 \in (0.75, 1]$ and $b_2 \in (2-2b_1, 2b_1-1)$. Then,*

$$\liminf_{k\to\infty} \{\|\Delta\boldsymbol{x}_k\| + \|\boldsymbol{c}_k\|\} = 0 \qquad \text{almost surely.}$$

Finally, we establish the convergence of the KKT residual. Note that the KKT condition (7)

is defined in terms of the primal variable only; without LICQ, the associated dual multipliers of Problem (1) are not unique (see Remark 3.1). Thus, it suffices to establish the existence of (possibly nonunique) dual multipliers such that, together with $\boldsymbol{x}_k$, the KKT residual converges to zero. As such, given $\boldsymbol{x}_k$ of SSQP, we consider the multipliers defined by a least-squares problem:

$$\begin{aligned} \min_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad & \|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_k,\boldsymbol{\lambda},\boldsymbol{\mu})\|^2 + \|\boldsymbol{\mu}_1 \odot (\boldsymbol{\ell}-\boldsymbol{x}_k)\|^2 + \|\boldsymbol{\mu}_2 \odot (\boldsymbol{x}_k-\boldsymbol{u})\|^2, \\ \text{s.t.} \quad & \boldsymbol{\mu} \geq \mathbf{0}. \end{aligned} \tag{23}$$

We denote (one of) the optimal solution to (23) by $(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star)$. The next theorem shows that the SSQP iterate $\boldsymbol{x}_k$, coupled with $(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star)$, yields a vanishing KKT residual.

Theorem 3.9. *Under the same conditions as in Theorem 3.8, we have*

$$\lim_{k\to\infty} \boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star) = \mathbf{0} \qquad \textit{almost surely.}$$

We note that in global analysis, two algorithmic components, the moving averaged second-order information of $\bar{\boldsymbol{B}}_k$ in (16) and the dual update in (18), are not essential; one may simply set $\bar{\boldsymbol{B}}_k = \boldsymbol{I}$ and suppress the dual update if only the consistency of the estimation is of interest. However, these components are crucial for local inference analysis, as they ensure the minimax optimality of the SSQP method. On the other hand, our method involves two key sequences: α_k for the primal updates and β_k for gradient averaging, which exhibit *competing* effects. In particular, a larger α_k accelerates the movement of the iterates but may impede the stabilization of the averaged gradients, while a smaller β_k improves gradient accuracy but slows the adaptation of both the gradient sequence and the primal iterates. These coupled dynamics highlight the importance of carefully balancing the decay rates of α_k and β_k, as specified in Theorem 3.8, to ensure simultaneous convergence of the averaged gradients and the primal iterates.

4 Asymptotic Normality and Inference

We now set the stage to present the local asymptotic analysis of our SSQP method in Section 3.1. We establish that the joint primal-dual variables, with proper stepsize control and averaging

weight sequences $\{\alpha_k, \beta_k, \gamma_k\}$, converge in distribution to a normal variable with limiting covariance $\boldsymbol{\Omega}^\star$ defined in (13). The resulting limiting distribution attains the minimax lower bound established in Theorem 2.5, thereby demonstrating that the proposed method is asymptotically optimal in the sense of Hájek and Le Cam. Furthermore, we provide a consistent plug-in estimator for the limiting covariance matrix, which enables practical online statistical inference, including hypothesis testing and confidence interval construction.

4.1 Asymptotic normality of the last iterate

We first introduce a local assumption on the limiting point $\boldsymbol{x}^\star$ to which the sequence $\{\boldsymbol{x}_k\}$ converges. This assumption is standard in both the lower-bound analysis (cf. Theorem 2.5) and the upper-bound analyses of offline M-estimation and online projection-based methods (Duchi and Ruan, 2021; Davis et al., 2024).

Assumption 4.1. *The iteration sequence $\boldsymbol{x}_k$ converges almost surely to a strict local solution $\boldsymbol{x}^\star$, which satisfies: (1) LICQ in Definition 2.1 (implying the uniqueness of the dual solution $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$); (2) strict complementarity in Assumption 2.3; and (3) second-order sufficient condition (SOSC) in Assumption 2.4.*

With Assumption 4.1, we now elaborate on the role of the Hessian matrix $\bar{\boldsymbol{B}}_k$. Note that the global analysis does not rely on any second-order information; hence, one may simply set $\bar{\boldsymbol{B}}_k = \boldsymbol{I}$ in the SSQP subproblem (17) and all results in Section 3.2 still hold. In contrast, for the local analysis, where $\boldsymbol{x}_k$ is sufficiently close to a strict local solution $\boldsymbol{x}^\star$ that satisfies LICQ and SOSC, we expect $\bar{\boldsymbol{B}}_k$ to serve as a good approximation to the Lagrangian Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_k$, which motivates the Hessian averaging update in (15). In fact, the conditions of LICQ and SOSC at $\boldsymbol{x}^\star$ extend to a local neighborhood. Thus, as long as $\boldsymbol{x}_k$ is close to $\boldsymbol{x}^\star$ (which is the case) and $\bar{\boldsymbol{B}}_k$ is close to $\nabla^2_{\boldsymbol{x}}\mathcal{L}_k$, these conditions apply to the SSQP subproblem (17) and further imply the *local* uniqueness of its primal-dual solution. Since no constraint qualification is imposed at $\boldsymbol{x}_k$ globally, the Hessian information in the subproblem is not helpful for determining its solutions in the global

analysis. Note that Hessian approximation in the local analysis is necessary even for deterministic (quasi-)second-order methods, so called the Dennis–Moré condition; see Dennis and Moré, 1974, (Nocedal and Wright, 2006, Chapter 18), and (Liu and Yuan, 2000, Section 5) for details.

To precisely state the above argument, let us consider any $\epsilon > 0$ such that

$$0 < \epsilon \leq 0.5 \min \left\{ [\boldsymbol{u} - \boldsymbol{\ell}]_i, [\boldsymbol{x}^\star - \boldsymbol{\ell}]_j, [\boldsymbol{u} - \boldsymbol{x}^\star]_j : i \in \mathcal{A}^\star, j \in (\mathcal{A}^\star)^- \right\}, \tag{24}$$

where we recall from Section 2 that $\mathcal{A}^\star = \mathcal{A}^\star_{\boldsymbol{\ell}} \cup \mathcal{A}^\star_{\boldsymbol{u}}$ denotes the index set of active inequality constraints and $(\mathcal{A}^\star)^- = [d] \setminus \mathcal{A}^\star$ denotes its inactive complement. We define two sets at $\boldsymbol{x}_k$,

$$\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \{i : [\boldsymbol{x}_k - \boldsymbol{\ell}]_i \leq \epsilon\} \qquad \text{and} \qquad \mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \{i : [\boldsymbol{u} - \boldsymbol{x}_k]_i \leq \epsilon\}, \tag{25}$$

as our guess of true active sets $\mathcal{A}^\star_{\boldsymbol{\ell}}$ and $\mathcal{A}^\star_{\boldsymbol{u}}$, and let $\mathcal{A}_k(\epsilon) \coloneqq \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) \cup \mathcal{A}_{\boldsymbol{u},k}(\epsilon)$. We will show next that $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{\ell}}, \mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{u}}$, and both also coincide with the active set of SSQP subproblem (17) when k is large enough, indicating that our method successfully identifies the true active set. Now, since SOSC holds at $\boldsymbol{x}^\star$ (cf. Assumption 2.4), given $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ and $\mathcal{A}_{\boldsymbol{u},k}(\epsilon)$ in (25), we require the regularizer $\boldsymbol{\Delta}_k$ to convexify the estimate of $\nabla^2_{\boldsymbol{x}} \mathcal{L}_k$, i.e., $\bar{\boldsymbol{Q}}_k + \sum_{i=1}^m [\boldsymbol{\lambda}_k]_i \nabla^2 \boldsymbol{c}_i(\boldsymbol{x}_k)$ in (16), to ensure $\boldsymbol{z}^\top \bar{\boldsymbol{B}}_k \boldsymbol{z} \geq \omega \|\boldsymbol{z}\|^2, \forall \boldsymbol{z} \in \{\boldsymbol{z} : \boldsymbol{J}_k(\epsilon)\boldsymbol{z} = \boldsymbol{0}\}$, where $\boldsymbol{J}_k(\epsilon) = (\nabla \boldsymbol{c}_k; -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}; \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)})$ is active constraints Jacobian. The following lemma shows that, under this construction of $\bar{\boldsymbol{B}}_k$, the SSQP subproblem admits a *unique* solution in a neighborhood of $\boldsymbol{0}$ when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}^\star$. In particular, if we fail to solve the subproblem (17) with the above $\bar{\boldsymbol{B}}_k$, we should turn to safely regularize $\bar{\boldsymbol{B}}_k$ to be positive definite to ensure the subproblem well-defined.

Lemma 4.2. *Under Assumptions 3.6 and 4.1, we let ϵ satisfy (24) and $\beta_k = \iota_2 (k+1)^{-b_2}$ satisfy $\iota_2 > 0$, $b_2 \in (0.5, 1]$. For any run of the method, there exists a (potential random) threshold $K^\star < \infty$ such that for all $k \geq K^\star$,*

(a) The relaxation parameter satisfies $\theta_k = 1$.

(b) The estimated active sets satisfy $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{\ell}}$ and $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{u}}$.

(c) The subproblems (17) with both the averaged gradient $\bar{\boldsymbol{g}}_k$ and true gradient ∇f_k admit unique local solutions in a neighborhood of $\boldsymbol{0}$; furthermore, $\bar{\Delta}\boldsymbol{x}_k \to \boldsymbol{0}$ and $\Delta \boldsymbol{x}_k \to \boldsymbol{0}$ as $k \to \infty$.

(d) *The active sets identified by the subproblems satisfy* $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k+\bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k+\Delta\boldsymbol{x}_k) = \mathcal{A}^\star_{\boldsymbol{\ell}}$ *and* $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k+\bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k+\Delta\boldsymbol{x}_k) = \mathcal{A}^\star_{\boldsymbol{u}}$.

We additionally emphasize that, by Lemma 4.2(a) and under LICQ at $\boldsymbol{x}^\star$, the relaxation parameter θ_k eventually equals one so that the linearized constraints become locally exact. We next explore almost-sure convergence of the dual iterates $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ and the Hessian approximation $\bar{\boldsymbol{B}}_k$ of the SSQP method. For the latter purpose, we require the stochastic Hessian estimates to be unbiased and have bounded variance.

Assumption 4.3. *We assume, for any* $k \geq 0$, $\nabla^2 F(\boldsymbol{x}_k;\zeta_k)$ *is an unbiased estimate of* $\nabla^2 f_k$ *with bounded variance. In particular,* $\mathbb{E}_k[\nabla^2 F(\boldsymbol{x}_k;\zeta_k)] = \nabla^2 f_k$ *and* $\mathbb{E}_k[\|\nabla^2 F(\boldsymbol{x}_k;\zeta_k) - \nabla^2 f_k\|^2] \leq \sigma_H^2$ *for some* $\sigma_H^2 > 0$. *Here,* $\mathbb{E}_k[\cdot]$ *denotes the conditional expectation given* $\boldsymbol{x}_k$.

Lemma 4.4. *Under Assumptions 3.6 and 4.1, let* $\alpha_k = \iota_1(k+1)^{-b_1}$ *and* $\beta_k = \iota_2(k+1)^{-b_2}$ *satisfy* $\iota_1, \iota_2 > 0$, $b_1 \in (0,1]$, $b_2 \in (0.5,1]$. *Then, we have* $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \to (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ *as* $k \to \infty$ *almost surely. Furthermore, suppose Assumption 4.3 holds and let the averaging weight sequence of the Hessian in (15) satisfy* $\gamma_k = \iota_3(k+1)^{-b_3}$ *with* $\iota_3 > 0$, $b_3 \in (0.5,1]$. *Then, we also have* $\bar{\boldsymbol{B}}_k \to \nabla^2_{\boldsymbol{x}}\mathcal{L}^\star$ *as* $k \to \infty$ *almost surely.*

Before stating normality result, we strengthen the bounded variance condition on stochastic gradients in Assumption 3.6 to a bounded $(2+\delta)$ moment. For notational simplicity, we reuse the symbol σ_g^2 from Assumption 3.6 to denote the upper bound of $(2+\delta)$ moment. This moment condition is mild and standard in existing literature on both unconstrained and constrained stochastic approximation methods (Duchi and Ruan, 2021; Polyak and Juditsky, 1992; Toulis et al., 2021; Chen et al., 2020; 2024; Zhu et al., 2021; Lee et al., 2022; Na and Mahoney, 2025). Some methods may require even higher-order moments; for example, projected SGD in Davis et al., 2024 imposes a bounded fourth moment to establish normality.

Assumption 4.5. *We assume, for any* $k \geq 0$, $\nabla F(\boldsymbol{x}_k;\zeta_k)$ *is unbiased and has bounded* $(2+\delta)$ *moment for some* $\delta > 0$. *In particular,* $\mathbb{E}_k[\|\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k\|^{2+\delta}] \leq \sigma_g^2$ *for some* $\sigma_g^2 > 0$. *Fur-*

thermore, we assume $\mathbb{E}_k[\nabla F(\boldsymbol{x}_k;\zeta_k)\nabla F(\boldsymbol{x}_k;\zeta_k)^\top] \to \mathbb{E}[\nabla F(\boldsymbol{x}^\star;\zeta)\nabla F(\boldsymbol{x}^\star;\zeta)^\top]$ *as* $\boldsymbol{x}_k \to \boldsymbol{x}^\star$.

We are now ready to present the asymptotic normality result for the SSQP iterates. Following the notation in the lower-bound analysis of Theorem 2.5, we denote $\boldsymbol{w}_k = (\boldsymbol{x}_k, \boldsymbol{\lambda}_k, [\boldsymbol{\mu}_k]_{\mathcal{A}^\star})$ and $\boldsymbol{w}^\star = (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star_{\mathcal{A}^\star})$ as the nontrivial part of the primal-dual solution, noting that $\boldsymbol{\mu}^\star_{(\mathcal{A}^\star)^-} = \mathbf{0}$.

Theorem 4.6. *Consider the SSQP method in Section 3.1 and suppose Assumptions 4.1, 4.3, 4.5 hold. We specify the stepsize and averaging weight sequences* $\alpha_k = \iota_1(k+1)^{-b_1}$, $\beta_k = \iota_2(k+1)^{-b_2}$, $\gamma_k = \iota_3(k+1)^{-b_3}$, *and the stepsize adaptivity gap parameter* p *in (18) to satisfy*

$$\max\left\{0.5, \frac{2-2\delta}{2+\delta}\right\} < b_1 \le 1, \quad 0.5 < b_2 < b_1, \quad 0.5 < b_3 \le 1, \quad p > 1.5 - \frac{b_2}{2b_1},$$

and $\iota_1 > 2/3$ *if* $b_1 = 1$ *and* $\iota_3 > 0.25b_1$ *if* $b_3 = 1$. *Then, we have (recall* $\boldsymbol{\Omega}^\star$ *defined in (13))*

$$1/\sqrt{\bar{\alpha}_k}\cdot(\boldsymbol{w}_k - \boldsymbol{w}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \eta\cdot\boldsymbol{\Omega}^\star) \qquad \text{with} \quad \eta = \begin{cases} 0.5 & \text{if } b_1 < 1, \\ \iota_1/(2\iota_1 - 1) & \text{if } b_1 = 1. \end{cases}$$

For the inactive dual components, $[\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}$ *vanishes almost surely with a rate* $\|[\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}\| = o(k^{-b})$ *for any* $b > 0$ *if* $b_1 < 1$ *and* $O(k^{-\iota_1})$ *if* $b_1 = 1$.

We consider a specific setup where $\alpha_k = (k+1)^{-1}$. The following corollary shows that SSQP with this stepsize attains optimal $\sqrt{k}$-consistency with the limiting covariance $\boldsymbol{\Omega}^\star$ matching the minimax lower bound in the sense of Hájek and Le Cam, as established in Theorem 2.5.

Corollary 4.7. *Under Assumptions 4.1, 4.3, 4.5, we let* $\iota_1 = b_1 = 1$, $b_2, b_3 \in (0.5, 1)$, $\iota_2, \iota_3 > 0$, $p \ge 1.5$. *Then,* $\sqrt{k}(\boldsymbol{w}_k - \boldsymbol{w}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^\star)$.

Theorem 4.6 establishes asymptotic normality for primal and active dual variables of SSQP under general setups of parameters. To our knowledge, this result illustrates that SSQP with gradient momentum is the *first* method to attain joint primal-dual asymptotic optimality, extending prior works Duchi and Ruan, 2021; Davis et al., 2024 that explore only primal asymptotic optimality and rely on expensive projection operators. The main idea of SSQP is to perform sequential linear-quadratic approximations to original nonlinearly constrained problem.

The primary technical challenge in our analysis arises from handling the gradient momentum.

In particular, existing literature on asymptotic normality of first- and second-order stochastic approximation methods relies crucially on the conditional independence of stochastic gradients (Duchi and Ruan, 2021; Davis et al., 2024; Polyak and Juditsky, 1992; Toulis et al., 2021; Chen et al., 2020; 2024; Zhu et al., 2021; Lee et al., 2022; Leluc and Portier, 2023; Na and Mahoney, 2025). In contrast, the averaged gradients in our setting are inherently dependent across iterations. As a result, our analysis must carefully examine the interplay between the stepsize sequence α_k and the averaging weight sequence β_k. We show that achieving asymptotic normality requires the weights β_k to decay more slowly than the stepsizes α_k, i.e., $b_2 < b_1$. Although gradient momentum helps reduce stochastic noise and allows the method to mimic the behavior of its deterministic counterparts, the slowly decaying weights ensure that the method still effectively uses a gradient estimate $\bar{\boldsymbol{g}}_k$ that is close to the stochastic gradient $\nabla F(\boldsymbol{x}_k; \zeta_k)$, thereby preserving asymptotically optimal behavior of the iterates. We also note that the choice of b_1 is always feasible for any $\delta > 0$, with larger values of δ permitting larger feasible ranges. When $\delta = 1$ (i.e., $\nabla F(\boldsymbol{x}_k; \zeta_k)$ has a bounded third moment), we can take $b_1 \in (0.5, 1]$. In addition, the adaptivity gap condition is satisfied as long as $p \geq 1.5$, and we remark that numerical literature often sets $p = 2$ (Berahas et al., 2021). Finally, the conditions on ι_1, ι_3 address the corner cases when b_1 or $b_3 = 1$, and setting $\iota_1 = \iota_3 = 1$ is always sufficient to satisfy conditions.

4.2 Practical inference: plug-in covariance estimation

To perform online statistical inference based on Theorem 4.6, we must also estimate the limiting covariance matrix $\boldsymbol{\Omega}^\star$. Existing literature considers either plug-in estimators (Chen et al., 2020; Na and Mahoney, 2025) or batch-means estimators (Zhu et al., 2021; Kuang et al., 2025). In this paper, we provide a simple plug-in estimator since the dominant computational cost of SSQP lies in solving the subproblem instead of covariance estimation. We also note that the plug-in estimator does not require any additional gradient evaluations. Define

$$\boldsymbol{H}_k = \begin{pmatrix} \bar{\boldsymbol{B}}_k & \boldsymbol{J}_k^\top \\ \boldsymbol{J}_k & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} \mathrm{Cov}(\{\nabla F(\boldsymbol{x}_i; \zeta_i)\}_{i=0}^k) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{\Omega}_k = \boldsymbol{H}_k^{-1} \boldsymbol{\Sigma}_k \boldsymbol{H}_k^{-1},$$

where $\bar{\boldsymbol{B}}_k$ is from (16); $\boldsymbol{J}_k = (\nabla \boldsymbol{c}_k; -\boldsymbol{I}_{\mathcal{A}_\ell(\boldsymbol{x}_k+\bar{\Delta}\boldsymbol{x}_k)}; \boldsymbol{I}_{\mathcal{A}_u(\boldsymbol{x}_k+\bar{\Delta}\boldsymbol{x}_k)})$ is the Jacobian of the active constraints with the active set identified by the SSQP subproblem (cf. Lemma 4.2); and the sample covariance is defined as

$$\mathrm{Cov}(\{\nabla F(\boldsymbol{x}_i;\zeta_i)\}_{i=0}^k) = \frac{1}{k+1}\sum_{i=0}^k \nabla F(\boldsymbol{x}_i;\zeta_i)\nabla F(\boldsymbol{x}_i;\zeta_i)^\top - \left(\frac{1}{k+1}\sum_{i=0}^k \nabla F(\boldsymbol{x}_i;\zeta_i)\right)\left(\frac{1}{k+1}\sum_{i=0}^k \nabla F(\boldsymbol{x}_i;\zeta_i)\right)^\top.$$

Note that $\boldsymbol{\Omega}_k$ can be updated recursively during the iterations, as it relies on the same stochastic gradient evaluations as the averaged gradient $\bar{\boldsymbol{g}}_k$, namely $\{\nabla F(\boldsymbol{x}_i;\zeta_i)\}_{i=0}^k$. Consequently, the plug-in estimator $\boldsymbol{\Omega}_k$ incurs negligible additional computational cost.

The following theorem establishes the almost-sure convergence of $\boldsymbol{\Omega}_k$. We strengthen the bounded $(2+\delta)$ moment condition of gradient estimates to bounded 4-th moment, which is also standard for other plug-in covariance estimators (Chen et al., 2020; Na and Mahoney, 2025).

Theorem 4.8. *Under the conditions of Theorem 4.6 and further assuming $\delta \geq 2$ in Assumption 4.5, we have $\boldsymbol{\Omega}_k \to \boldsymbol{\Omega}^\star$ as $k \to \infty$ almost surely.*

With the above theorem, we are now able to construct confidence intervals or regions for $\boldsymbol{w}^\star = (\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star_{\mathcal{A}^\star})$. For example, fixing a desired coverage probability $1-q$ with $q \in (0,1)$, the $100(1-q)\%$ confidence region of $\boldsymbol{w}^\star$ is given by

$$P(\boldsymbol{w}^\star \in \mathcal{C}_{k,q}) \to 1-q \quad \text{as} \quad k \to \infty, \tag{26}$$

where $\mathcal{C}_{k,q} = \{\boldsymbol{w} : (\boldsymbol{w}-\boldsymbol{w}_k)^\top \boldsymbol{\Omega}_k^\dagger (\boldsymbol{w}-\boldsymbol{w}_k)/\bar{\alpha}_k \leq \chi^2_{d,1-q}\}$. Here, $\chi^2_{d,1-q}$ is the $(1-q)$-quantile of χ^2_d distribution and $\boldsymbol{\Omega}_k^\dagger$ is the pseudo-inverse of $\boldsymbol{\Omega}_k$. The result (26) follows from the distribution of quadratic forms and the fact that $\mathrm{rank}(\boldsymbol{\Omega}^\star) = d$ (Christensen, 2020, Corollary 1.3.6a).

5 Experimental Studies

In this section, we perform comprehensive experiments to demonstrate the effectiveness of the SSQP estimation procedure in Section 3.1, validate asymptotic normality results, and illustrate

its applicability to practical statistical inference tasks, such as performing hypothesis testing and constructing confidence intervals for constrained model parameters. We evaluate the empirical performance of our method on a diverse set of tasks, including benchmark nonlinear constrained optimization problems from the CUTEst library (Gould et al., 2014; Fowkes et al., 2022), constrained regression problems involving linear, logistic, and Poisson models, as well as portfolio allocation problems. We study both synthetic data and real `Fama-French Portfolios` and `Chicago Air Pollution` data.

We use the same set of hyper-parameters for all experiments. Specifically, we set $\tau = 0.5$, $\psi = 1$, and $p = 2$. The stepsize control sequence is set to $\alpha_k = (k+1)^{-0.751}$, while the weight sequences for the gradient and Hessian are set to $\beta_k = (k+1)^{-0.501}$ and $\gamma_k = (k+1)^{-1}$, respectively, which satisfy the conditions in Theorems 3.9 and 4.6. Our implementation code is provided in the public repository: https://github.com/yihang-gao/SSQP.

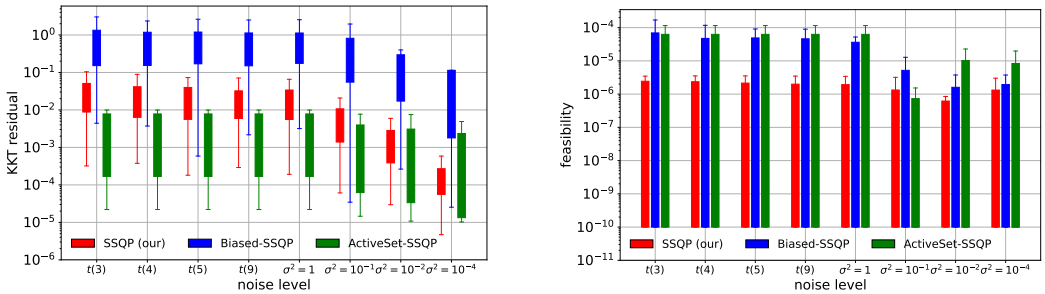
5.1 CUTEst benchmark problems

The CUTEst library collects numerous constrained problems that are widely used for benchmarking optimization methods. We apply our method to a subset of CUTEst problems and introduce stochastic perturbations to gradients and Hessians in order to simulate noisy environments. In particular, at each step, we let $\nabla F(\boldsymbol{x}_k;\zeta_k) = \nabla f_k + \boldsymbol{E}_{k,\nabla f}$ and $\nabla^2 F(\boldsymbol{x}_k;\zeta_k) = \nabla^2 f_k + \boldsymbol{E}_{k,\nabla^2 f}$, where the deterministic quantities ∇f_k and $\nabla^2 f_k$ are provided by the CUTEst package, and $\boldsymbol{E}_{k,\nabla f}$ and $\boldsymbol{E}_{k,\nabla^2 f}$ denote the noise vector and matrix, respectively. We study two types of noise: light-tailed Gaussian noise, where each entry $\boldsymbol{E}_{k,\nabla f}$, $\boldsymbol{E}_{k,\nabla^2 f}$ follows $\mathcal{N}(0,\sigma^2)$, and heavy-tailed t-distribution noise, where each entry $\boldsymbol{E}_{k,\nabla f}$, $\boldsymbol{E}_{k,\nabla^2 f}$ follows t_{df}. We vary $\sigma^2 \in \{1, 10^{-1}, 10^{-2}, 10^{-4}\}$ and $df \in \{3, 4, 5, 9\}$. Note that the noise of t_9 closely approximates that of $\mathcal{N}(0,1)$.

We compare our SSQP method with two state-of-the-art baselines: ActiveSet-SSQP (Na et al., 2023) and Biased-SSQP (Curtis et al., 2024). The ActiveSet-SSQP method adaptively increases the batch size to reduce the noise in the step direction, and is hence not an online method

and is more sample-intensive. In contrast, the Biased-SSQP method is an online method but lacks bias reduction through gradient momentum.

• **Global convergence.** For each problem instance, we run $K = 10^5$ iterations; and for each method, we evaluate it using the KKT residual $\|\boldsymbol{R}_K\|$ defined in (20) and the feasibility error $\|\boldsymbol{c}_K\|$. The performance comparison is shown in Figure 3. From the figure, we observe that Biased-SSQP exhibits noticeable deviation from the optimal solution, primarily due to the bias in its step direction induced by the inequality constraints. ActiveSet-SSQP, on the other hand, performs robustly even under high noise levels but heavily depends on increasing batch size to ensure a sufficiently accurate step direction. In contrast, our proposed method requires only a single sample per iteration to estimate both the gradient and Hessian, while effectively reducing stochastic noise and debiasing the step through gradient averaging. We see that when the noise level is moderate (Gaussian noise with $\sigma^2 \in \{10^{-1}, 10^{-2}, 10^{-4}\}$), our method achieves performance comparable to, or even better than, ActiveSet-SSQP.



(a) KKT residuals (b) Feasibility errors

Figure 3: *Boxplots of KKT residuals and feasibility errors for CUTEst problems. For each noise setting, three boxplots are shown, corresponding to the SSQP (ours), Biased-SSQP, and ActiveSet-SSQP methods.*

• **Local normality.** We next examine local asymptotic normality of the SSQP iterates. For each problem, we estimate the averaged model parameter $\mathbf{1}^\top \boldsymbol{x}^\star/d$ and construct its nominal 95% confidence interval. The performance of the constructed intervals is evaluated in terms of their empirical coverage rate (CovRate) and average length (AvgLen), each averaged over 200 independent runs. Results of 5 HS-type CUTEst problems are reported in Table 1. We find that the confidence intervals produced by SSQP achieve empirical coverage probabilities that closely

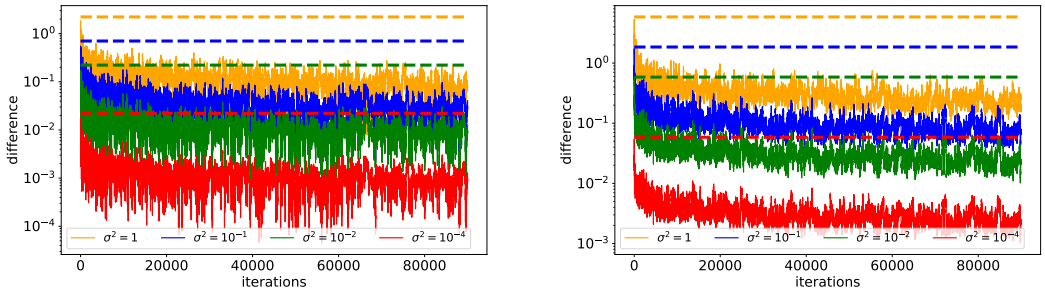
match the nominal 95% level, thereby providing strong empirical support for the theoretical asymptotic normality guarantee established in Theorem 4.6. Furthermore, the average length of the confidence intervals increases as the noise level grows, a behavior also consistent with our theoretical expectations, since the asymptotic covariance matrix $\mathbf{\Omega}^\star$ is proportional to $\mathrm{Cov}(\nabla F(\boldsymbol{x}^\star;\zeta))$ as defined in (13).

Table 1: *The coverage rate (CovRate) and length of confidence intervals (AvgLen) for 5 HS-type CUTEst constrained problems. The standard deviation of the interval length is also reported.*

Problem	**Noise Level**	**Gaussian**		**Noise Level**	**Student t**	
	(σ^2)	**CovRate**(%)	**AvgLen**	(*df*)	**CovRate**(%)	**AvgLen**
HS41	1	97.0	2.50E-2 (7.73E-4)	3	86.0	3.77E-2 (1.90E-3)
	10^{-1}	97.5	7.59E-3 (7.03E-5)	4	93.0	3.06E-2 (1.28E-3)
	10^{-2}	97.0	2.40E-3 (8.69E-6)	5	94.0	2.79E-2 (9.25E-4)
	10^{-4}	97.5	2.40E-4 (5.95E-7)	9	97.0	2.45E-2 (7.10E-4)
HS65	1	94.5	1.87E-3 (6.82E-6)	3	96.5	3.18E-3 (1.66E-4)
	10^{-1}	94.5	5.92E-4 (1.59E-6)	4	95.0	2.59E-3 (1.72E-5)
	10^{-2}	95.0	1.87E-4 (4.96E-7)	5	95.0	2.37E-3 (1.11E-5)
	10^{-4}	94.5	1.87E-5 (4.97E-8)	9	94.5	2.08E-3 (8.36E-6)
HS68	1	97.0	2.31E-1 (4.85E-2)	3	95.5	3.00E-1 (1.32E-1)
	10^{-1}	98.0	5.09E-2 (2.33E-3)	4	94.5	2.08E-1 (6.09E-2)
	10^{-2}	98.5	1.58E-2 (2.23E-4)	5	95.0	1.81E-1 (5.07E-2)
	10^{-4}	95.5	1.58E-3 (4.56E-6)	9	94.5	1.48E-1 (3.52E-2)
HS71	1	97.0	1.95E-3 (1.44E-5)	3	94.0	3.34E-3 (1.23E-4)
	10^{-1}	96.5	6.17E-4 (1.93E-6)	4	96.0	2.74E-3 (6.79E-3)
	10^{-2}	96.5	1.95E-4 (5.20E-7)	5	96.5	2.49E-3 (2.51E-5)
	10^{-4}	98.5	1.95E-5 (5.08E-8)	9	95.0	2.19E-3 (2.12E-5)
HS81	1	94.5	3.49E-2 (3.17E-3)	3	91.0	5.04E-2 (7.56E-3)
	10^{-1}	97.0	1.13E-2 (4.77E-5)	4	94.0	4.21E-2 (3.51E-3)
	10^{-2}	98.0	3.58E-3 (9.63E-6)	5	94.5	3.88E-2 (2.42E-3)
	10^{-4}	98.0	3.59E-4 (9.22E-7)	9	95.0	3.43E-2 (2.10E-3)

• **Effectiveness of gradient averaging.** Empirically, our online method consistently outperforms the other online Biased-SSQP method. This improvement is primarily attributed to the momentum-style gradient moving-average scheme: as iterations proceed, the averaged gradient progressively approximates the true population gradient, enabling the method to emulate the behavior of deterministic SQP while maintaining full stochasticity. In contrast, Biased-SSQP updates its iterates using noisy gradients, leading to inherent biases in step estimates that accumulate over time and result in deviations from the optimal solution. As shown in Figure 4, we visualize the difference between the averaged gradients and the true gradients across iterations. The results align with our theory that averaged gradients (solid lines) progressively approach

the true gradients; while non-averaged gradients (dashed lines) exhibit constant variance.



(a) HS32 (b) FCCU

Figure 4: *Differences between the averaged gradients and the true gradients on HS32 and FCCU problems. Solid lines: trajectories of gradient difference between the averaged gradients and the exact gradients during iterations, i.e.,* $\|\bar{\boldsymbol{g}}_k - \nabla f_k\|$. *Dashed lines: expected error without averaging, i.e.,* $\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k\| \mid \boldsymbol{x}_k\right]$.

5.2 Constrained regression problems

We further implement our method on constrained regression problems, including both linear and logistic regression models. In this study, each sample is denoted by the covariate-response pair $\zeta_k = (\zeta_k^{\boldsymbol{a}}, \zeta_k^b)$, and the response is generated based on different regression models. In particular, for linear regression, we let $\zeta_k^b = \zeta_k^{\boldsymbol{a}\top}\boldsymbol{x}^\star + \varepsilon_k$ with $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0,1)$ and set $F(\boldsymbol{x};\zeta_k) = \frac{1}{2}(\zeta_k^b - \zeta_k^{\boldsymbol{a}\top}\boldsymbol{x})^2$. For logistic regression, we let $P(\zeta_k^b \mid \zeta_k^{\boldsymbol{a}}) = 1/\{1+\exp(-\zeta_k^b \cdot \zeta_k^{\boldsymbol{a}\top}\boldsymbol{x}^\star)\}$ with $\zeta_k^b \in \{-1, 1\}$ and set $F(\boldsymbol{x};\zeta_k) = \log(1+\exp(-\zeta_k^b \cdot \zeta_k^{\boldsymbol{a}\top}\boldsymbol{x}))$. For both models, we set the true model parameters as $\boldsymbol{x}^\star = \left(\frac{3}{2d}, \ldots, \frac{3}{2d}, \frac{1}{2d}, \ldots, \frac{1}{2d}\right)$, where the first half of the components are $\frac{3}{2d}$ and the second half are $\frac{1}{2d}$. We then impose the probability simplex constraints $\Omega := \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{1}^\top\boldsymbol{x} = 1,\ \boldsymbol{x} \geq \mathbf{0}\}$ in the estimation procedure. We generate covariate vector $\zeta_k^{\boldsymbol{a}}$ from $\zeta_k^{\boldsymbol{a}} \sim \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{a}}, \boldsymbol{\Sigma}^{\boldsymbol{a}})$, where $\boldsymbol{\mu}^{\boldsymbol{a}} = (1, \ldots, 1, -1, \ldots, -1)$. Following unconstrained regression settings (Chen et al., 2020; Zhu et al., 2021), we consider three structures for $\boldsymbol{\Sigma}^{\boldsymbol{a}}$: (i) Identity: $\boldsymbol{\Sigma}^{\boldsymbol{a}} = \boldsymbol{I}$; (ii) Toeplitz: $[\boldsymbol{\Sigma}^{\boldsymbol{a}}]_{ij} = r^{|i-j|}$ with $r \in \{0.4, 0.5, 0.6\}$; and (iii) Equi-correlation: $[\boldsymbol{\Sigma}^{\boldsymbol{a}}]_{ij} = r$ for $i \neq j$ and $[\boldsymbol{\Sigma}^{\boldsymbol{a}}]_{ii} = 1$ with $r \in \{0.1, 0.2, 0.3\}$. For each setting of $\boldsymbol{\Sigma}^{\boldsymbol{a}}$, we vary the problem dimension $d \in \{5, 10, 20, 30\}$ and construct 95% confidence interval for $\boldsymbol{\mu}^{\boldsymbol{a}\top}\boldsymbol{x}^\star$, i.e., the parameters difference of two groups.

We measure the performance by repeating each experiment 200 times with varying random seeds and reporting the empirical coverage rate (CovRate) and average length (AvgLen) of the

constructed confidence intervals. The results are summarized in Tables 2 and 3. From the tables, we observe that the empirical coverage probabilities are close to the nominal 95% level across all settings, thereby providing strong empirical support for our local asymptotic normality guarantees and demonstrating the practical effectiveness of our method. Furthermore, the average length of the confidence intervals remains on the order of 10^{-2}, consistent with the findings reported by Chen et al., 2020; Zhu et al., 2021; Na and Mahoney, 2025. The low standard deviation of the interval lengths relative to their means further indicates the robustness of the proposed inference procedure across different covariance structures.

Table 2: *The coverage rate (CovRate) and length of confidence intervals (AvgLen) for constrained linear regression problems. The standard deviation of the interval length is also reported.*

Cov Matrix	**Dim** d	**CovRate**(%)	**AvgLen**	**Dim** d	**CovRate**(%)	**AvgLen**
Identity	5	93.5	3.73E-2 (1.74E-4)	20	92.5	4.00E-2 (1.33E-4)
	10	96.5	3.91E-2 (1.47E-4)	30	92.5	4.03E-2 (1.53E-4)
Toeplitz ($r = 0.4$)	5	94.0	3.71E-2 (1.68E-4)	20	96.0	3.93E-2 (1.38E-4)
	10	94.5	3.82E-2 (1.62E-4)	30	93.0	3.98E-2 (1.52E-4)
Toeplitz ($r = 0.5$)	5	94.0	3.74E-2 (1.67E-4)	20	96.0	3.91E-2 (1.38E-4)
	10	95.5	3.82E-2 (1.60E-4)	30	93.0	3.95E-2 (1.61E-4)
Toeplitz ($r = 0.6$)	5	94.5	3.78E-2 (1.70E-4)	20	96.5	3.90E-2 (1.36E-4)
	10	94.5	3.83E-2 (1.68E-4)	30	93.5	3.94E-2 (1.60E-4)
EquiCorr ($r = 0.1$)	5	93.5	3.76E-2 (1.58E-4)	20	94.0	4.01E-2 (1.35E-4)
	10	93.0	3.92E-2 (1.40E-4)	30	92.5	4.05E-2 (1.56E-4)
EquiCorr ($r = 0.2$)	5	92.5	3.79E-2 (1.59E-4)	20	93.5	4.02E-2 (1.26E-4)
	10	95.0	3.94E-2 (1.50E-4)	30	96.0	4.05E-2 (1.44E-4)
EquiCorr ($r = 0.3$)	5	92.5	3.83E-2 (1.65E-4)	20	93.0	4.03E-2 (1.31E-4)
	10	95.0	3.96E-2 (1.46E-4)	30	93.5	4.05E-2 (1.49E-4)

5.3 Portfolio allocation: Fama-French dataset

We implement our method on portfolio allocation problems, using 30 portfolios selected from the Fama–French Portfolios dataset, subject to the gross-exposure constraint (Fan, 2007; Fan et al., 2012; Du et al., 2022): $\Omega := \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{1}^\top \boldsymbol{x} = 1,\ \|\boldsymbol{x}\|_1 \leq c\}$, where we set $c = 3$ and $\boldsymbol{x}$ denotes the portfolio weight vector; a negative weight signifies shorting an asset. Let $\boldsymbol{\mu}^{\boldsymbol{a}}, \boldsymbol{\Sigma}^{\boldsymbol{a}}$ denote the mean vector and covariance matrix of the asset returns $\zeta^{\boldsymbol{a}} \in \mathbb{R}^{30}$. We consider four widely studied portfolio models: (i) Global Minimum Variance **(GMV)**: $\min_{\boldsymbol{x}\in\Omega} \boldsymbol{x}^\top \boldsymbol{\Sigma}^{\boldsymbol{a}} \boldsymbol{x}$; (ii) Mean-Variance **(MV)**: $\min_{\boldsymbol{x}\in\Omega} -\boldsymbol{x}^\top \boldsymbol{\mu}^{\boldsymbol{a}} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{\boldsymbol{a}} \boldsymbol{x}$; (iii) Exponential Utility **(EXP)**: $\min_{\boldsymbol{x}\in\Omega} \mathbb{E}[\exp(-\eta_1 \boldsymbol{x}^\top \zeta^{\boldsymbol{a}})]$, where

Table 3: *The coverage rate (CovRate) and length of confidence intervals (AvgLen) for constrained logistic regression problems. The standard deviation of the interval length is also reported.*

Cov Matrix	Dim d	CovRate(%)	AvgLen	Dim d	CovRate(%)	AvgLen
Identity	5	96.5	4.46E-2 (7.97E-5)	20	94.5	5.87E-2 (7.13E-5)
	10	94.5	5.87E-2 (7.13E-5)	30	93.0	7.34E-2 (7.90E-5)
Toeplitz ($r = 0.4$)	5	94.5	4.46E-2 (9.06E-5)	20	92.5	6.86E-2 (1.01E-4)
	10	95.5	5.83E-2 (8.59E-5)	30	93.5	7.30E-2 (1.13E-4)
Toeplitz ($r = 0.5$)	5	95.0	4.46E-2 (8.91E-5)	20	94.0	6.84E-2 (1.08E-4)
	10	94.5	5.83E-2 (8.77E-5)	30	93.0	7.28E-2 (1.24E-4)
Toeplitz ($r = 0.6$)	5	94.5	4.47E-2 (9.63E-5)	20	92.5	6.82E-2 (1.19E-4)
	10	94.0	5.83E-2 (8.77E-5)	30	94.5	7.26E-2 (1.32E-4)
EquiCorr ($r = 0.1$)	5	95.0	4.47E-2 (9.22E-5)	20	93.0	6.69E-2 (9.40E-5)
	10	94.0	5.89E-2 (7.81E-5)	30	93.5	7.40E-2 (9.27E-5)
EquiCorr ($r = 0.2$)	5	96.0	4.47E-2 (8.86E-4)	20	95.0	7.00E-2 (1.05E-4)
	10	95.0	5.92E-2 (7.32E-5)	30	92.5	7.46E-2 (1.02E-4)
EquiCorr ($r = 0.3$)	5	95.0	4.48E-2 (8.59E-5)	20	93.5	7.05E-2 (1.09E-4)
	10	96.0	5.95E-2 (7.94E-4)	30	94.5	7.52E-2 (1.09E-4)

$\eta_1 > 0$ is the risk-aversion parameter and set to $\eta_1 = 0.1$; and (iv) Logarithmic Utility **(LOG)**: $\min_{\boldsymbol{x}\in\Omega} -\mathbb{E}[\log(\boldsymbol{x}^\top \zeta^{\boldsymbol{a}} + \eta_2)]$, where $\eta_2 > 0$ is a regularization parameter ensuring the validity of the logarithm and set to $\eta_2 = 15$.

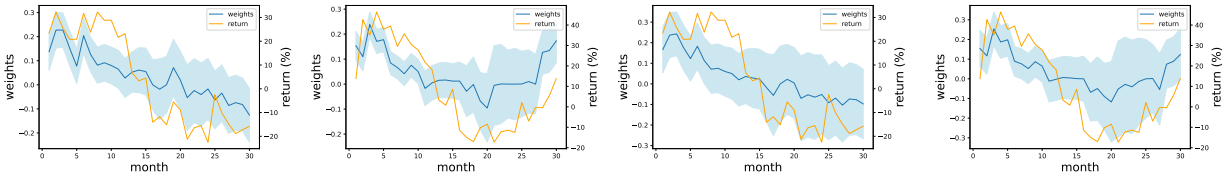
For each month, we use historical daily data from the preceding year as training samples and apply both the SSQP method (dealing with one sample at a time) and the M-estimation method (using full samples to estimate $\boldsymbol{\mu}^{\boldsymbol{a}}, \boldsymbol{\Sigma}^{\boldsymbol{a}}$) to solve the resulting constrained stochastic problem and obtain a portfolio weight vector. This weight vector is then held fixed throughout the following month for evaluation. Specifically, the out-of-sample performance of the estimated portfolio weights is evaluated using four standard metrics computed over 30 months during the period 2021–2023: accumulative return, maximum drawdown, Sharpe ratio, and Sortino ratio. The accumulative return quantifies the overall gain or loss of a portfolio strategy, while the other three metrics balance return with risk aspect: the maximum drawdown measures the largest observed decline from a peak to a trough; the Sharpe ratio relates the portfolio's mean return to its total risk measured by the standard deviation of returns; and the Sortino ratio refines this measure by considering only downside volatility, i.e., the standard deviation of negative returns. The results are summarized in Table 4.

Table 4: *The comparison of SSQP and M-estimation on Fama-French Portfolios Dataset from 2021-2023. For each metric, the best-performing model is shown in bold.*

Model	**Return** (%)	**Max Drawdown**	**Sharpe Ratio**	**Sortino Ratio**
Equal Weight	15.10	**0.22**	0.73	1.15
GMV (ours)	34.94	0.27	2.81	4.28
GMV (M-est)	33.43	0.27	2.71	4.14
MV (ours)	42.21	0.28	**3.36**	**5.09**
MV (M-est)	40.31	0.28	**3.29**	**5.02**
EXP (ours)	52.50	0.32	2.60	3.98
EXP (M-est)	51.85	0.31	2.55	3.86
LOG (ours)	**54.86**	0.33	2.45	3.59
LOG (M-est)	**55.08**	0.32	2.46	3.57

From Table 4, we see that SSQP achieves comparable performance to offline M-estimation across all metrics. In particular, the performance differences across all four models and four metrics are within 0.15, with the largest observed in the global minimum variance model under the Sortino ratio metric, where SSQP outperforms M-estimation by 0.14. Among different portfolio models, we observe that the portfolio strategy based on the logarithmic utility model achieves the highest accumulative return, which is consistent with the empirical findings of Du et al., 2022. The mean-variance model yields higher Sharpe and Sortino ratios, while the simple equal-weight strategy yields the smallest maximum drawdown.

Figure 5 further visualizes the estimated weights of two representative stocks under two portfolio models, exponential utility and logarithmic utility. From the figure, we observe a strong correlation between the temporal adjustments of the weights and the dynamics of the corresponding stock returns. In particular, as the return decreases, the weight assigned to the stock also decreases. Moreover, an abrupt drop in the stock return is promptly followed by a widening blue band, indicating an increase in the estimated standard deviation of the associated weight. This behavior aligns with economic intuition and demonstrates the interpretability and validity of the estimated weights under constrained model estimation.



(a) Stock 1, EXP (b) Stock 2, EXP (c) Stock 1, LOG (d) Stock 2, LOG

Figure 5: *Trajectories of portfolio weights and corresponding stock returns. We show two stocks under two portfolio models. The blue lines depict the predicted weights for the stock, while the shaded blue bands indicate the estimated standard deviations of these weights, computed based on the derived asymptotic normality results. The yellow lines depict the returns of the same stock.*

5.4 Poisson regression: Chicago air pollution data

In this section, we compare unconstrained and constrained methods for generalized linear regression problems. We use daily air pollution and death rate data for Chicago (Wood, 2017) to analyze the relationship between death rates and air pollution levels. Following Toulis and Airoldi, 2017, we fit a Poisson regression model that regresses the death counts (ζ^b) on six variables: Intercept, Time, PM_{10}, $PM_{2.5}$, SO_2, and O_3 ($\zeta^{\boldsymbol{a}}$). In particular, we assume $\zeta^b|\zeta^{\boldsymbol{a}} \sim \text{Poisson}(\lambda(\zeta^{\boldsymbol{a}}))$ with $\log \lambda(\zeta^{\boldsymbol{a}}) = \zeta^{\boldsymbol{a}\top}\boldsymbol{x}^\star$. In this study, we may want to incorporate the prior domain knowledge that higher pollutant concentrations are expected to increase mortality, and thus impose non-negativity constraints on the coefficients of the pollution covariates. Accordingly, we compare the following two methods:

$$\min_{\boldsymbol{x}} \ \mathbb{E}_\zeta[\zeta^b\, \zeta^{\boldsymbol{a}\top}\boldsymbol{x} - \exp(\zeta^{\boldsymbol{a}\top}\boldsymbol{x})], \qquad \min_{\boldsymbol{x}} \ \mathbb{E}_\zeta[\zeta^b\, \zeta^{\boldsymbol{a}\top}\boldsymbol{x} - \exp(\zeta^{\boldsymbol{a}\top}\boldsymbol{x})], \quad \text{s.t.} \ \ \boldsymbol{x}_{3:6} \geq \boldsymbol{0}. \tag{27}$$

The above unconstrained Poisson model is fitted using the `statsmodels` package in Python (Seabold and Perktold, 2010), while the constrained model is fitted using the proposed SSQP method. We summarize the estimated coefficients along with their 95% confidence intervals and p-values in Table 5 (left). From this table, we observe that the unconstrained method estimates the coefficient of O_3 to be negative and significant, which contradicts prior domain knowledge and can be difficult to interpret. In contrast, the constrained method correctly identifies the coefficient of O_3 as active. Among other variables, the two methods produce comparable coefficient

estimates. For example, the coefficients estimated by the unconstrained method lie within the 95% confidence intervals constructed by the constrained method, and vice versa. Furthermore, both methods identify the Intercept, Time, and SO_2 as significant, while PM_{10} as insignificant. The only difference occurs with $PM_{2.5}$, which is insignificant under the unconstrained method but significant under the constrained method.

Table 5: *Summary of two Poisson regression models* (27) *(left) and* (28) *(right) applied to the Chicago air pollution data.*

Var	Method	Coeff (10^{-2})	95% CI (10^{-2})	p-Value
Intercept	Uncons	4.6968	[4.690, 4.704]	<0.001
	Cons	4.6974	[4.692, 4.703]	<0.001
Time	Uncons	0.95	[0.17, 1.74]	0.008
	Cons	1.13	[0.64, 1.63]	<0.001
PM_{10}	Uncons	0.42	[-0.57, 1.42]	0.396
	Cons	0.13	[-0.56,0.79]	0.371
$PM_{2.5}$	Uncons	0.72	[-0.08, 1.52]	0.103
	Cons	0.65	[0.02, 1.28]	0.023
SO_2	Uncons	1.38	[0.58, 2.20]	0.001
	Cons	2.08	[1.43, 2.73]	<0.001
O_3	Uncons	-2.97	[-3.70, -2.24]	<0.001
	Cons	0.00	active	

Var	Method	Coeff (10^{-2})	95% CI (10^{-2})	p-Value
Intercept	Uncons	4.6972	[4.690, 4.704]	<0.001
	Cons	4.6973	[4.692, 4.703]	<0.001
Time	Uncons	1.21	[0.53, 1.89]	0.001
	Cons	1.13	[0.64, 1.63]	<0.001
PM_{10}	Uncons	-0.86	[-1.82, 0.10]	0.062
	Cons	0.11	[-0.52, 0.74]	0.362
$PM_{2.5}$	Uncons	1.37	[0.51, 2.23]	0.001
	Cons	0.65	[0.01, 1.28]	0.022
SO_2	Uncons	2.06	[1.28, 2.84]	<0.001
	Cons	2.08	[1.42, 2.73]	<0.001

When we encounter a negative coefficient of O_3 after fitting the model, we may next perform model modification by simply removing this variable for sake of interpretability. As such, we next consider a sub-model comparison where we remove O_3 variable:

$$\min_{\boldsymbol{x}} \ \mathbb{E}_\zeta[\zeta^b \zeta^{\boldsymbol{a}^\top \boldsymbol{x}} - \exp(\zeta^{\boldsymbol{a}^\top \boldsymbol{x}})], \qquad \min_{\boldsymbol{x}} \ \mathbb{E}_\zeta[\zeta^b \zeta^{\boldsymbol{a}^\top \boldsymbol{x}} - \exp(\zeta^{\boldsymbol{a}^\top \boldsymbol{x}})], \quad \text{s.t.} \quad \boldsymbol{x}_{3:5} \geq \mathbf{0}. \tag{28}$$

The results are summarized in Table 5 (right). From this table, we observe that the two methods output largely similar results. In particular, both methods identify the Intercept, Time, and SO_2 as significant, and PM_{10} as insignificant, consistent with the conclusions drawn from the full model. Notably, both methods also identify $PM_{2.5}$ as significant, which aligns with the constrained method fitted on the full model. To summarize, the unconstrained method may yield inconsistency between fitting the full and sub models, while our constrained method preserves consistency when active variables on the boundary are removed. This illustrates the benefits of applying our constrained methods when incorporating prior domain knowledge into model fitting.

6 Conclusion and Future Work

In this paper, we studied online statistical inference for the solutions of stochastic optimization problems with equality and inequality constraints. We developed a stochastic sequential quadratic programming method that incorporates a moving-average gradient scheme to correct the bias in the stochastic step direction induced by inequality constraints. We established global almost-sure convergence and proved that the proposed method achieves local asymptotic normality with a minimax-optimal primal–dual limiting covariance matrix in the sense of Hájek and Le Cam. Furthermore, we proposed a plug-in covariance matrix estimator for practical inference. Extensive experiments on benchmark nonlinear problems, as well as on constrained generalized linear models and portfolio allocation tasks using both synthetic and real data, demonstrated the superior empirical performance of our method and confirmed its effectiveness of online constrained inference in practice.

As for future directions, it would be significant to provide a non-asymptotic analysis that quantifies how quickly the stochastic iterates approach the limiting distribution. For example, Anastasiou et al., 2019 derived a non-asymptotic convergence rate of averaged SGD to a normal distribution by applying a non-asymptotic martingale central limit theorem. Establishing comparable bounds for SSQP methods on constrained problems would further highlight the effectiveness of this type of methods. Furthermore, recent work has shown promising advances in applying SGD methods to high-dimensional settings (Li et al., 2025). Developing a high-dimensional theory for constrained model estimation would require incorporating regularization techniques such as sparsity-inducing penalties, manifold constraints, or low-rank constraints to enable valid analysis. Integrating these techniques with the proposed SSQP method can open new pathways toward scalable, structure-aware online inference for high-dimensional constrained problems.

References

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*. Vol. 1, p. 175–194.

Le Cam, L. (1972). Limits of experiments. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press Berkeley-Los Angeles, p. 245–261.

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

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

Robbins, H. and S. Monro (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3): p. 400–407.

Kiefer, J. and J. Wolfowitz (1952). Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3): p. 462–466.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4): p. 838–855.

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

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

Toulis, P., T. Horel, and E. M. Airoldi (2021). The Proximal Robbins–Monro Method. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1): p. 188–212.

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

Chen, X., Z. Lai, H. Li, and Y. Zhang (2024). Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods. *Journal of the American Statistical Association*, 119(548): p. 2972–2982.

Zhu, W., X. Chen, and W. B. Wu (2021). Online Covariance Matrix Estimation in Stochastic Gradient Descent. *Journal of the American Statistical Association*, 118(541): p. 393–404.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2022). Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): p. 7381–7389.

Jiang, L. et al. (2025). Online covariance estimation in nonsmooth stochastic approximation. *The 38th Annual Conference on Learning Theory (COLT)*.

Polyak, B. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5): p. 1–17.

Nocedal, J. and S. J. Wright (2006). Numerical optimization. Springer New York.

Na, S., M. Anitescu, and M. Kolar (2022). An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199(1–2): p. 721–791.

— (2023). Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*, 202(1–2): p. 279–353.

Na, S. (2025). Derivative-Free Sequential Quadratic Programming for Equality-Constrained Stochastic Optimization. *arXiv preprint arXiv:2510.22458*.

Curtis, F. E., M. J. O'Neill, and D. P. Robinson (2023). Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*.

Curtis, F. E., D. P. Robinson, and B. Zhou (2024). Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints. *SIAM Journal on Optimization*, 34(4): p. 3592–3622.

Berahas, A. S., F. E. Curtis, D. Robinson, and B. Zhou (2021). Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization. *SIAM Journal on Optimization*, 31(2): p. 1352–1379.

Berahas, A. S., F. E. Curtis, M. J. O'Neill, and D. P. Robinson (2023). A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear-Equality-Constrained Optimization with Rank-Deficient Jacobians. *Mathematics of Operations Research*.

Fang, Y., S. Na, M. W. Mahoney, and M. Kolar (2024a). Fully Stochastic Trust-Region Sequential Quadratic Programming for Equality-Constrained Optimization Problems. *SIAM Journal on Optimization*, 34(2): p. 2007–2037.

Fang, Y., S. Na, M. W. Mahoney, and M. Kolar (2024b). Trust-Region Sequential Quadratic Programming for Stochastic Optimization with Random Models. *arXiv preprint arXiv:2409.15734*.

Fang, Y., J. Lavaei, and S. Na (2025). High Probability Complexity Bounds of Trust-Region Stochastic Sequential Quadratic Programming with Heavy-Tailed Noise. *arXiv preprint arXiv:2503.19091*.

Bercu, B., A. Godichon, and B. Portier (2020). An Efficient Stochastic Newton Algorithm for Parameter Estimation in Logistic Regressions. *SIAM Journal on Control and Optimization*, 58(1): p. 348–367.

Boyer, C. and A. Godichon-Baggioni (2022). On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions. *Computational Optimization and Applications*, 84(3): p. 921–972.

Leluc, R. and F. Portier (2023). Asymptotic Analysis of Conditioned Stochastic Gradient Descent. *Transactions on Machine Learning Research*.

Cénac, P., A. Godichon-Baggioni, and B. Portier (2025). An efficient averaged stochastic Gauss-Newton algorithm for estimating parameters of nonlinear regressions models. *Bernoulli*, 31(1).

Kuang, W., M. Anitescu, and S. Na (2025). Online covariance matrix estimation in sketched newton methods. *arXiv preprint arXiv:2502.07114*.

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

Du, X., W. Zhu, W. B. Wu, and S. Na (2025). Online Statistical Inference of Constrained Stochastic Optimization via Random Scaling. *arXiv preprint arXiv:2505.18327*.

Xu, M., J. J. Ye, and L. Zhang (2015). Smoothing SQP Methods for Solving Degenerate Nonsmooth Constrained Optimization Problems with Applications to Bilevel Programs. *SIAM Journal on Optimization*, 25(3): p. 1388–1410.

Shapiro, A. (1990). On differential stability in stochastic programming. *Mathematical Programming*, 47(1–3): p. 107–116.

Bonnans, J. F. and A. Shapiro (2000). Perturbation Analysis of Optimization Problems. Springer New York.

Burer, S. and A. N. Letchford (2009). On Nonconvex Quadratic Programming with Box Constraints. *SIAM Journal on Optimization*, 20(2): p. 1073–1089.

Dennis, J. E. and J. J. Moré (1974). A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28(126): p. 549–560.

Liu, X.-w. and Y.-x. Yuan (2000). A Robust Algorithm for Optimization with General Equality and Inequality Constraints. *SIAM Journal on Scientific Computing*, 22(2): p. 517–534.

Christensen, R. (2020). Plane Answers to Complex Questions: The Theory of Linear Models. Springer International Publishing.

Gould, N. I. M., D. Orban, and P. L. Toint (2014). CUTEst: a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3): p. 545–557.

Fowkes, J., L. Roberts, and Á. Bűrmen (2022). PyCUTEst: an open source Python package of optimization test problems. *Journal of Open Source Software*, 7(78): p. 4377.

Fan, J. (2007). Variable screening in high-dimensional feature space. *Proceedings of the 4th International Congress of Chinese Mathematicians*. Vol. 2. Citeseer, p. 735–747.

Fan, J., J. Zhang, and K. Yu (2012). Vast Portfolio Selection With Gross-Exposure Constraints. *Journal of the American Statistical Association*, 107(498): p. 592–606.

Du, J.-H., Y. Guo, and X. Wang (2022). High-Dimensional Portfolio Selection with Cardinality Constraints. *Journal of the American Statistical Association*, 118(542): p. 779–791.

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Seabold, S. and J. Perktold (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*. Vol. 57. (61).

Anastasiou, A., K. Balasubramanian, and M. A. Erdogdu (2019). Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. *Conference on Learning Theory (COLT)*. PMLR, p. 115–137.

Li, J., Z. Lou, J. Schmidt-Hieber, and W. B. Wu (2025). Statistical Guarantees for High-Dimensional Stochastic Gradient Descent. *Conference on Neural Information Processing Systems (NeurIPS)*.

Na, S., Z. Yang, Z. Wang, and M. Kolar (2019). High-dimensional Varying Index Coefficient Models via Stein's Identity. *Journal of Machine Learning Research*, 20: p. 152–1.

Na, S. and M. Kolar (2021). High-dimensional index volatility models via Stein's identity. *Bernoulli*, 27(2).

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1): p. 133–144.

Zafar, M. B., I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): p. 2737–2778.

Sen, P. K. (1979). Asymptotic Properties of Maximum Likelihood Estimators Based on Conditional Specification. *The Annals of Statistics*, 7(5).

Dupacova, J. and R. Wets (1988). Asymptotic Behavior of Statistical Estimators and of Optimal Solutions of Stochastic Optimization Problems. *The Annals of Statistics*, 16(4).

Nagaraj, N. K. and W. A. Fuller (1991). Estimation of the Parameters of Linear Time Series Models Subject to Nonlinear Restrictions. *The Annals of Statistics*, 19(3).

Shapiro, A. (2000). On the asymptotics of constrained local M-estimators. *The Annals of Statistics*, 28(3).

Cuomo, S. et al. (2022). Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. *Journal of Scientific Computing*, 92(3).

Chen, R. T., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

Wang, S., H. Wang, and P. Perdikaris (2021). Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances*, 7(40).

Raissi, M., P. Perdikaris, and G. Karniadakis (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: p. 686–707.

Krishnapriyan, A. et al. (2021). Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

Négiar, G., M. W. Mahoney, and A. Krishnapriyan (2023). Learning differentiable solvers for systems with hard constraints. *International Conference on Learning Representations*.

Robinson, S. M. (1976). Stability Theory for Systems of Inequalities, Part II: Differentiable Nonlinear Systems. *SIAM Journal on Numerical Analysis*, 13(4): p. 497–513.

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

Daniel, J. W. (1973). Stability of the solution of definite quadratic programs. *Mathematical Programming*, 5(1): p. 41–53.

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

Duflo, M. (1997). Random iterative models. Ed. by S. S. Wilson. Applications of mathematics 34. Aus dem Franz. übers. Berlin: Springer. ISBN: 3540571000.

Dontchev, A. L. and R. T. Rockafellar (2009). Implicit Functions and Solution Mappings: A View from Variational Analysis. Springer New York. ISBN: 9780387878218.

Hall, P. and C. C. Heyde (2014). Martingale limit theory and its application. Academic press.

Appendix: Online Inference of Constrained Optimization: Primal-Dual Optimality and Sequential Quadratic Programming

A Motivating Examples

The question of when constraints are needed in statistical machine learning problems may be divided into two scenarios. The first scenario refers to problems that are otherwise ill-posed and impossible to solve. For example, in principal component analysis, we estimate the leading eigenvector of the population covariance matrix $\boldsymbol{\Sigma}$. However, maximizing $\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}$ is not well-defined without the constraint $\|\boldsymbol{x}\|^2 = 1$. The second scenario refers to problems where constraints incorporate prior model information and enable data to be explained by simpler hypotheses. We introduce several examples that can be cast into the form of Problem (1).

A.1 Constrained regression

Given a set of feature-response data $\{(\zeta_1^{\boldsymbol{a}}, \zeta_1^b), \ldots, (\zeta_n^{\boldsymbol{a}}, \zeta_n^b)\}$, we consider fitting a smooth monotone function $z : \zeta^{\boldsymbol{a}} \to \zeta^b$. Monotonic regression has many applications, such as estimating cumulative distribution functions, and monotonic relationships between two variables are prevalent in various studies. For example, demand is typically a non-increasing function of price, and the rate of chemical processes is an increasing function of temperature. See our study on environmental data in Section 5.4. A simple approach to fit a monotone function is to select some basis functions $\{z_1(\cdot), \ldots, z_d(\cdot)\}$ and define

$$z(\zeta^{\boldsymbol{a}}; \boldsymbol{x}) = \sum_{i=1}^{d} x_i z_i(\zeta^{\boldsymbol{a}}),$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$ are regression coefficients. Then, we minimize the mean-squared error and impose the monotonicity, e.g., non-increasing, via constraints on the coefficients $\boldsymbol{x}$:

$$\min_{\boldsymbol{x}} \ \frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{x}; \zeta_i) \coloneqq \frac{1}{n}\sum_{i=1}^{n} (\zeta_i^b - z(\zeta_i^{\boldsymbol{a}}; \boldsymbol{x}))^2 \quad \text{s.t.} \quad z'(\omega_j; \boldsymbol{x}) \coloneqq \sum_{i=1}^{d} x_i z_i'(\omega_j) \le 0, \quad j \in [m],$$

where $\omega_1, \ldots, \omega_m$ are a fine grid of points over the range of $\zeta^{\boldsymbol{a}}$.

Constraints are also imposed in other statistical problems. In portfolio optimization (cf.

Section 5.3), $\boldsymbol{x}$ refers to the asset weights, and it is common to constrain the estimation within the probability simplex $\{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{1}^\top \boldsymbol{x} = 1, \boldsymbol{x} \geq \mathbf{0}\}$ (Du et al., 2022). Alternatively, one may control the risk exposures to a known threshold by imposing box constraints $\|\boldsymbol{x}\|_\infty \leq c$, or control allocations among sectors by imposing affine constraints $A\boldsymbol{x} = \boldsymbol{b}$, where a negative weight signifies shorting the asset (Fan, 2007; Fan et al., 2012). In semiparametric index models, we impose $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|^2 = 1, x_1 > 0\}$ to ensure models' identifiability (Na et al., 2019; Na and Kolar, 2021). In factor analysis, constraints can prevent Heywood cases that cause negative estimates for the variance (Shapiro, 1985). In algorithmic fairness, constraints are applied to prevent classifiers from yielding disparate outcomes based on sensitive features such as gender and ethnicity (Zafar et al., 2019). For more examples of constrained regression models, we refer to Sen, 1979; Dupacova and Wets, 1988; Nagaraj and Fuller, 1991; Shapiro, 2000.

A.2 Physics-informed machine learning

A recent trending topic in machine learning is its applications on scientific problems, where models must adhere to domain knowledge (such as physical laws) that is often described by partial differential equation (PDE) constraints (Cuomo et al., 2022). Consider a smooth neural network model $z(\cdot; \boldsymbol{x}) : \zeta^{\boldsymbol{a}} \to \zeta^b$, where $\zeta^{\boldsymbol{a}} = (s, t)$ is the spatial-temporal pair, ζ^b is the measurement of some quantity, and $\boldsymbol{x}$ denotes the network parameters (e.g., weight matrices). If $z(\cdot; \boldsymbol{x})$ models the transport of the quantity, then it should satisfy the transport equation (γ is given):

$$\partial_t z(\zeta^{\boldsymbol{a}}; \boldsymbol{x}) + \gamma \cdot \partial_s z(\zeta^{\boldsymbol{a}}; \boldsymbol{x}) = \mathbf{0}. \tag{A.1}$$

Applying PDE constraints to the model $z(\cdot; \boldsymbol{x})$ leads to many important network architectures, including Neural ODEs (Chen et al., 2018), DeepONets (Wang et al., 2021), and physics-informed neural networks (Raissi et al., 2019). In these problems, the desired model is solved from the following constrained optimization:

$$\min_{\boldsymbol{x}} \ \mathbb{E}_{\zeta \sim \mathcal{P}}[F(\boldsymbol{x}; \zeta)] \quad \text{s.t.} \quad \mathcal{C}_{\text{PDE}}(z(\omega_j; \boldsymbol{x})) = \mathbf{0}, \ \ j \in [m],$$

where $F(\boldsymbol{x};\zeta)$ is the data fitting loss (e.g., mean-squared error in Section A.1), $(\omega_1,\ldots,\omega_m)$ are sampled points in the space of $\zeta^{\boldsymbol{a}}$ including both interior and boundary points, and $\mathcal{C}_{\text{PDE}}(\cdot)$ is a PDE operator with respect to $\zeta^{\boldsymbol{a}}$ that encodes domain knowledge (e.g., transport equation as in (A.1)). Numerous research has shown that failure to enforce these constraints can lead to serious interpretability and instability issues in neural network models (Krishnapriyan et al., 2021; Négiar et al., 2023).

B Preparation Definitions and Lemmas

Lemma B.1 (Theorem 3 in Robinson, 1976). *If $\boldsymbol{\ell} \le \bar{\boldsymbol{x}} \le \boldsymbol{u}$ satisfies EGMFCQ, then there exists a neighborhood $\mathcal{B}(\bar{\boldsymbol{x}};\bar{r}) := \{\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u} : \|\boldsymbol{x}-\bar{\boldsymbol{x}}\| \le \bar{r}\}$ for some sufficiently small radius $\bar{r} > 0$, such that all points in the neighborhood also satisfy EGMFCQ.*

Definition B.2 (Regular mapping sequence, Vaart, 1998). *Let $\mathcal{W} \subseteq \mathbb{R}^r$ be a neighborhood of the origin. A sequence of mappings $\Gamma_k : \mathcal{W} \to \mathbb{R}^d$ is said to be regular with derivative $\boldsymbol{D} \in \mathbb{R}^{d\times r}$ if*

$$\lim_{k\to\infty} \sqrt{k}\,(\Gamma_k(\boldsymbol{w}) - \Gamma_k(\boldsymbol{0})) = \boldsymbol{D}\boldsymbol{w}, \quad \text{for all } \boldsymbol{w} \in \mathcal{W}.$$

Lemma B.3 (Hájek and Le Cam's local minimax theorem, Vaart, 1998). *Let $\{\mathcal{T}_k, \mathcal{F}_k, \mathcal{P}_{k,\boldsymbol{v}}\}_{\boldsymbol{v}\in\mathcal{W}}$ be a locally asymptotically normal family with precision $\boldsymbol{Q} \succeq \boldsymbol{0}$. Let $\Gamma_k : \mathcal{W} \to \mathbb{R}^d$ be a regular mapping sequence with derivative $\boldsymbol{D}$, and let $\ell : \mathbb{R}^d \to [0,\infty)$ be any symmetric, quasiconvex, and lower semicontinuous function. Then, for any sequence of estimators $T_k : \mathcal{T}_k \to \mathbb{R}^d$, the following inequality holds:*

$$\sup_{\mathcal{W}_0\subseteq\mathcal{W},|\mathcal{W}_0|<\infty} \liminf_{k\to\infty} \max_{\boldsymbol{v}\in\mathcal{W}_0} \mathbb{E}_{\mathcal{P}_{k,\boldsymbol{v}}}[\ell(\sqrt{k}(T_k - \Gamma_k(\boldsymbol{v})))] \ge \mathbb{E}\left[\ell(Z)\right],$$

where $Z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}\boldsymbol{Q}^{-1}\boldsymbol{D}^\top)$ when $\boldsymbol{Q}$ is invertible. If $\boldsymbol{Q}$ is singular, then the above inequality holds with $Z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{Q}+\gamma\boldsymbol{I})^{-1}\boldsymbol{D}^\top)$ for any $\gamma > 0$. Here, $\mathbb{E}_{\mathcal{P}_{k,\boldsymbol{v}}}[\cdot]$ is taken over the randomness of samples $\mathcal{T}_k$.

Lemma B.4 (Lemma B.3 in Na and Mahoney, 2025). *Let $\alpha_k = \iota_1(k+1)^{-b_1}$ and $\beta_k = \iota_2(k+1)^{-b_2}$ for some $\iota_1,\iota_2 > 0$ and $b_1, b_2 > 0$. Then the following statements hold:*

(a) Define $\chi = 0$ *if* $0 < b_2 < 1$ *and* $\chi = -b_1/\iota_2$ *if* $b_2 = 1$. *If* $\sum_{t=1}^{l} a_t + \chi > 0$, *then*

$$\lim_{k\to\infty} \frac{1}{\alpha_k} \sum_{i=0}^{k} \prod_{j=i+1}^{k} \prod_{t=1}^{l} (1 - a_t\beta_j)\, \beta_i\alpha_i = \frac{1}{\sum_{t=1}^{l} a_t + \chi}.$$

Moreover, for any $b \in \mathbb{R}$ *and any sequence* e_k *satisfying* $\lim_{k\to\infty} e_k \to 0$, *we have*

$$\lim_{k\to\infty} \left\{ \frac{1}{\alpha_k} \sum_{i=0}^{k} \prod_{j=i+1}^{k} \prod_{t=1}^{l} (1 - a_t\beta_j)\, \beta_i\alpha_i e_i + b \prod_{j=0}^{k} \prod_{t=1}^{l} (1 - a_t\beta_j) \right\} = 0.$$

(b) Suppose $b_2 = 1$ *and let* $\chi = -b_1/\iota_2$. *If* $\sum_{t=1}^{l} a_t + p\chi > 0$ *for some* $p \in (0, 1)$, *then*

$$\lim_{k\to\infty} \frac{1}{\alpha_k^p} \sum_{i=0}^{k} \prod_{j=i+1}^{k} \prod_{t=1}^{l} (1 - a_t\beta_j)\, \beta_i\alpha_i = 0.$$

(c) If $0 < b_2 < b_1 \leq 1$, *then*

$$\lim_{k\to\infty} \frac{1}{\alpha_k} \sum_{i=0}^{k} \prod_{j=i+1}^{k} (1 - \alpha_j)\,(1 - \beta_j)\,\alpha_i\beta_i = 1.$$

Lemma B.5 (Stability of Quadratic Programs, Theorem 2.1 in Daniel, 1973)**.** *Consider two constrained strongly convex quadratic problems*

$$\boldsymbol{y}^\star \in \min_{\boldsymbol{y}\in\Omega} \boldsymbol{g}^\top\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{Q}\boldsymbol{y} \qquad \textit{and} \qquad \boldsymbol{y}^{\star\star} \in \min_{\boldsymbol{y}\in\Omega} \boldsymbol{g}'^\top\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{Q}'\boldsymbol{y},$$

where the feasible region Ω *is convex. Suppose* $\boldsymbol{Q}, \boldsymbol{Q}' \succeq \nu\boldsymbol{I}$ *and* $\max\{\|\boldsymbol{y}^\star\|, \|\boldsymbol{y}^{\star\star}\|\} \leq M_y$ *for some* $\nu, M_y > 0$. *Then,*

$$\|\boldsymbol{y}^\star - \boldsymbol{y}^{\star\star}\| \leq \upsilon^{-1}(1 + M_y) \max\{\|\boldsymbol{g} - \boldsymbol{g}'\|, \|\boldsymbol{Q} - \boldsymbol{Q}'\|\}.$$

C Proofs of Main Results

C.1 Proof of Theorem 2.2

Suppose EGMFCQ holds at $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$, we denote $\mathcal{A}(\boldsymbol{x}) = \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}) \cup \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})$, $\mathcal{A}(\boldsymbol{x})^- = [d] \setminus \mathcal{A}(\boldsymbol{x})$, and let $\boldsymbol{z}$ satisfy (11). If $\boldsymbol{z} = \boldsymbol{0}$, then we define $\bar{\theta} = 1$. Otherwise, we define

$$\bar{\theta} = \frac{\epsilon}{\|\boldsymbol{z}\|} \qquad \text{where} \qquad \epsilon = \min\left\{ [\boldsymbol{u} - \boldsymbol{\ell}]_i, [\boldsymbol{x} - \boldsymbol{\ell}]_j, [\boldsymbol{u} - \boldsymbol{x}]_j : i \in \mathcal{A}(\boldsymbol{x}), j \in \mathcal{A}(\boldsymbol{x})^- \right\}. \quad \text{(C.1)}$$

Note that $\bar{\theta} > 0$ since $\epsilon > 0$, although we do not restrict $\bar{\theta} \leq 1$. We claim $\Omega(\boldsymbol{x}; \bar{\theta}) \neq \emptyset$. In fact, $\bar{\theta}\boldsymbol{z} \in \Omega(\boldsymbol{x}; \bar{\theta})$ since $\bar{\theta}\boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})(\bar{\theta}\boldsymbol{z}) \overset{(11)}{=} \boldsymbol{0}$ and

$$\begin{aligned}
&\boldsymbol{\ell}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \overset{(8)}{=} \boldsymbol{x}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \overset{(11)}{\leq} [\boldsymbol{x} + \bar{\theta}\boldsymbol{z}]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \overset{(8)}{=} [\boldsymbol{\ell} + \bar{\theta}\boldsymbol{z}]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \leq [\boldsymbol{\ell} + \epsilon\boldsymbol{1}]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})} \overset{(C.1)}{\leq} \boldsymbol{u}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x})},\\
&\boldsymbol{\ell}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \overset{(C.1)}{\leq} [\boldsymbol{u} - \epsilon\boldsymbol{1}]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \leq [\boldsymbol{u} + \bar{\theta}\boldsymbol{z}]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \overset{(8)}{=} [\boldsymbol{x} + \bar{\theta}\boldsymbol{z}]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \overset{(11)}{\leq} \boldsymbol{x}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})} \overset{(8)}{=} \boldsymbol{u}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x})},\\
&\boldsymbol{\ell}_{\mathcal{A}(\boldsymbol{x})^-} \overset{(C.1)}{\leq} [\boldsymbol{x} - \epsilon\boldsymbol{1}]_{\mathcal{A}(\boldsymbol{x})^-} \leq [\boldsymbol{x} + \bar{\theta}\boldsymbol{z}]_{\mathcal{A}(\boldsymbol{x})^-} \leq [\boldsymbol{x} + \epsilon\boldsymbol{1}]_{\mathcal{A}(\boldsymbol{x})^-} \overset{(C.1)}{\leq} \boldsymbol{u}_{\mathcal{A}(\boldsymbol{x})^-},
\end{aligned}$$

where $\boldsymbol{1} = (1, \ldots, 1) \in \mathbb{R}^d$ denotes the all-one vector and the above display also uses $-\boldsymbol{1} \leq \boldsymbol{z}/\|\boldsymbol{z}\| \leq \boldsymbol{1}$. From the above display, we know $\boldsymbol{\ell} \leq \boldsymbol{x} + \bar{\theta}\boldsymbol{z} \leq \boldsymbol{u}$; thus, $\bar{\theta}\boldsymbol{z} \in \Omega(\boldsymbol{x}; \bar{\theta})$. Now, we show that $\Omega(\boldsymbol{x}; \theta) \neq \emptyset$ for any $\theta \in [0, \bar{\theta}]$. In fact, for any $\boldsymbol{v} \in \Omega(\boldsymbol{x}; \bar{\theta})$, we consider $\theta/\bar{\theta} \cdot \boldsymbol{v}$. First, we have $\theta\boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})(\theta/\bar{\theta}\boldsymbol{v}) = \boldsymbol{0}$ since $\bar{\theta}\boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{v} = \boldsymbol{0}$. Second, we bound $\boldsymbol{x} + \theta/\bar{\theta}\boldsymbol{v}$ by considering two cases based on the sign of $\boldsymbol{v}$. We apply $\boldsymbol{\ell} \leq \boldsymbol{x} + \boldsymbol{v} \leq \boldsymbol{u}$ and have

$$\boldsymbol{\ell}_{\{i:\boldsymbol{v}_i \geq 0\}} \leq \boldsymbol{x}_{\{i:\boldsymbol{v}_i \geq 0\}} \leq [\boldsymbol{x} + \theta/\bar{\theta}\boldsymbol{v}]_{\{i:\boldsymbol{v}_i \geq 0\}} \leq [\boldsymbol{x} + \boldsymbol{v}]_{\{i:\boldsymbol{v}_i \geq 0\}} \leq \boldsymbol{u}_{\{i:\boldsymbol{v}_i \geq 0\}},$$

$$\boldsymbol{\ell}_{\{i:\boldsymbol{v}_i < 0\}} \leq [\boldsymbol{x} + \boldsymbol{v}]_{\{i:\boldsymbol{v}_i < 0\}} \leq [\boldsymbol{x} + \theta/\bar{\theta}\boldsymbol{v}]_{\{i:\boldsymbol{v}_i < 0\}} \leq \boldsymbol{x}_{\{i:\boldsymbol{v}_i < 0\}} \leq \boldsymbol{u}_{\{i:\boldsymbol{v}_i < 0\}}.$$

Thus, $\boldsymbol{\ell} \leq \boldsymbol{x} + \theta/\bar{\theta}\boldsymbol{v} \leq \boldsymbol{u}$. This completes the proof of the first part of the theorem.

For the second part of the theorem, without loss of generality, we suppose $\lim_{k\to\infty} \theta_k = 0$ and $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^\star$. Suppose $\boldsymbol{x}^\star$ satisfies EGMFCQ, then there exists $\boldsymbol{z}^\star$ such that (11) holds. On the other hand, by Lemma B.1 we know $\boldsymbol{x}_k$ satisfies EGMFCQ for k large enough, implying that $\nabla \boldsymbol{c}_k$ has full row rank. Noting that

$$\boldsymbol{z}^\star = -(\nabla \boldsymbol{c}^\star)^\top (\nabla \boldsymbol{c}^\star (\nabla \boldsymbol{c}^\star)^\top)^{-1} \boldsymbol{c}^\star + \left(\boldsymbol{I} - (\nabla \boldsymbol{c}^\star)^\top (\nabla \boldsymbol{c}^\star (\nabla \boldsymbol{c}^\star)^\top)^{-1} \nabla \boldsymbol{c}^\star\right) \boldsymbol{z}^\star,$$

we define

$$\boldsymbol{z}_k = -\nabla \boldsymbol{c}_k^\top (\nabla \boldsymbol{c}_k \nabla \boldsymbol{c}_k^\top)^{-1} \boldsymbol{c}_k + \left(\boldsymbol{I} - \nabla \boldsymbol{c}_k^\top (\nabla \boldsymbol{c}_k \nabla \boldsymbol{c}_k^\top)^{-1} \nabla \boldsymbol{c}_k\right) \boldsymbol{z}^\star$$

and have $\boldsymbol{c}_k + \nabla \boldsymbol{c}_k \boldsymbol{z}_k = \boldsymbol{0}$ and $\boldsymbol{z}_k \to \boldsymbol{z}^\star$ as $k \to \infty$. Since $\mathcal{A}(\boldsymbol{x}^\star)^- \subseteq \mathcal{A}(\boldsymbol{x}_k)^-$, we know $\mathcal{A}(\boldsymbol{x}_k) \subseteq \mathcal{A}(\boldsymbol{x}^\star)$. Since each entry of $\boldsymbol{z}^\star_{\mathcal{A}(\boldsymbol{x}^\star)}$ is nonzero by (11), we know each entry of $[\boldsymbol{z}_k]_{\mathcal{A}(\boldsymbol{x}_k)}$ is also nonzero and has the same sign as $\boldsymbol{z}^\star_{\mathcal{A}(\boldsymbol{x}_k)}$. This shows that the constructed $\boldsymbol{z}_k$ satisfies (11) for $\boldsymbol{x}_k$. Let $l_k \coloneqq \inf\{\|\boldsymbol{z}\| : \boldsymbol{z} \text{ satisfies (11) at } \boldsymbol{x}_k\}$. For any $\boldsymbol{z}$ satisfies (11) at $\boldsymbol{x}_k$ with k large

enough, we can follow the analysis of the firs part, define

$$\epsilon_k = \min\{0.5[\boldsymbol{u}-\boldsymbol{\ell}]_i, [\boldsymbol{x}_k-\boldsymbol{\ell}]_j, [\boldsymbol{u}-\boldsymbol{x}_k]_j : i \in \mathcal{A}(\boldsymbol{x}^\star), j \in \mathcal{A}(\boldsymbol{x}^\star)^-\},$$

and show $\Omega(\boldsymbol{x}_k; \epsilon_k/\|\boldsymbol{z}\|) \neq \emptyset$. Since $\epsilon_k \geq \epsilon$ for some $\epsilon > 0$ independent of k, we know $\theta_k \to 0$ implies $\limsup_{k\to\infty} l_k = \infty$. This leads to a contradiction since $\infty = \limsup_{k\to\infty} l_k \leq \|\boldsymbol{z}^\star\| =: l^\star < \infty$. Thus, EGMFCQ does not hold at $\boldsymbol{x}^\star$. This completes the proof of the second part of the theorem.

C.2 Proof of Theorem 2.5

First, by Assumption 2.3, we know $\boldsymbol{\mu}^\star_{\mathcal{A}^\star} > \boldsymbol{0}$. For any KKT triplet $(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ that is close to $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$, the corresponding active and inactive sets coincide with those at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$. In particular, for any $i \in \mathcal{A}^\star = \mathcal{A}^\star_{\boldsymbol{\ell}} \cup \mathcal{A}^\star_{\boldsymbol{u}}$, we have $\boldsymbol{\mu}_{\mathcal{A}^\star} > \boldsymbol{0}$ since $\boldsymbol{\mu}^\star_{\mathcal{A}^\star} > \boldsymbol{0}$, which implies $\mathcal{A}^\star \subseteq \mathcal{A}(\boldsymbol{x})$ due to the KKT conditions (7) at $(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$. On the other hand, since $\boldsymbol{\ell}_{(\mathcal{A}^\star)^-} < \boldsymbol{x}^\star{}_{(\mathcal{A}^\star)^-} < \boldsymbol{u}_{(\mathcal{A}^\star)^-}$, the above inequalities also hold for $\boldsymbol{x}_{(\mathcal{A}^\star)^-}$, leading to $(\mathcal{A}^\star)^- \subseteq \mathcal{A}(\boldsymbol{x})^-$ and hence $\mathcal{A}(\boldsymbol{x}) \subseteq \mathcal{A}^\star$. This shows $\mathcal{A}(\boldsymbol{x}) = \mathcal{A}^\star$.

With this observation, we define the Lagrangian gradient with respect to $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}_{\mathcal{A}^\star}) \in \mathbb{R}^{d+m+|\mathcal{A}^\star_{\boldsymbol{\ell}}|+|\mathcal{A}^\star_{\boldsymbol{u}}|}$ as

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = \begin{pmatrix} \nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{\ell}}} [\boldsymbol{\mu}_1]_{\mathcal{A}^\star_{\boldsymbol{\ell}}} + \boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{u}}} [\boldsymbol{\mu}_2]_{\mathcal{A}^\star_{\boldsymbol{u}}} \\ \boldsymbol{c}(\boldsymbol{x}) \\ [\boldsymbol{\ell} - \boldsymbol{x}]_{\mathcal{A}^\star_{\boldsymbol{\ell}}} \\ [\boldsymbol{x} - \boldsymbol{u}]_{\mathcal{A}^\star_{\boldsymbol{u}}} \end{pmatrix} \in \mathbb{R}^{d+m+|\mathcal{A}^\star_{\boldsymbol{\ell}}|+|\mathcal{A}^\star_{\boldsymbol{u}}|}, \tag{C.2}$$

and define the mapping $\sigma(\boldsymbol{\delta}) = \{\boldsymbol{w} : \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = \boldsymbol{\delta}\}$. This mapping characterizes the set of primal-dual points satisfying the perturbed optimality conditions up to a residual vector $\boldsymbol{\delta}$. In fact, $\sigma(\boldsymbol{\delta})$ is a single-valued mapping in a neighborhood of $(\boldsymbol{\delta}, \boldsymbol{w}) = (\boldsymbol{0}, \boldsymbol{w}^\star)$ with $\nabla \sigma(\boldsymbol{\delta})|_{\boldsymbol{\delta}=\boldsymbol{0}} = (\boldsymbol{H}^\star)^{-1}$. To see this, let $\widehat{\boldsymbol{w}}, \bar{\boldsymbol{w}} \in \sigma(\boldsymbol{\delta})$ for sufficiently small $\boldsymbol{\delta}$. By SOSC and LICQ in Assumption 2.4 and Definition 2.1, we know $\nabla^2_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^\star) = \boldsymbol{H}^\star$ is nonsingular. Thus, there exists a neighborhood of $\boldsymbol{w}^\star$ within which $\nabla^2_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ remains nonsingular. We then have

$$\boldsymbol{0} = \nabla_{\boldsymbol{w}} \mathcal{L}(\widehat{\boldsymbol{w}}) - \nabla_{\boldsymbol{w}} \mathcal{L}(\bar{\boldsymbol{w}}) = \left(\int_0^1 \nabla^2_{\boldsymbol{w}} \mathcal{L}\left(\bar{\boldsymbol{w}} + t(\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}})\right) dt \right) \cdot (\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}}). \tag{C.3}$$

Since both $\widehat{\boldsymbol{w}}$ and $\bar{\boldsymbol{w}}$ lie sufficiently close to $\boldsymbol{w}^\star$, the integral matrix is nonsingular, which implies

$\widehat{\boldsymbol{w}} - \bar{\boldsymbol{w}} = \mathbf{0}$. Hence, $\sigma(\cdot)$ is locally single-valued. Furthermore, by (C.3) we have

$$\boldsymbol{\delta} = \nabla_{\boldsymbol{w}} \mathcal{L}(\widehat{\boldsymbol{w}}) - \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^\star) = \left(\int_0^1 \nabla_{\boldsymbol{w}}^2 \mathcal{L} \left(\boldsymbol{w}^\star + t(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star) \right) dt \right) \cdot (\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star),$$

leading to

$$\sigma(\boldsymbol{\delta}) = \widehat{\boldsymbol{w}} = \boldsymbol{w}^\star + \left(\int_0^1 \nabla_{\boldsymbol{w}}^2 \mathcal{L} \left(\boldsymbol{w}^\star + t(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star) \right) dt \right)^{-1} \boldsymbol{\delta}.$$

Differentiating $\sigma(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ at $\boldsymbol{\delta} = \mathbf{0}$ yields

$$\nabla \sigma(\boldsymbol{\delta})|_{\boldsymbol{\delta}=\mathbf{0}} = \lim_{\boldsymbol{\delta} \to \mathbf{0}} \left(\int_0^1 \nabla_{\boldsymbol{w}}^2 \mathcal{L}(\boldsymbol{w}^\star + t(\widehat{\boldsymbol{w}} - \boldsymbol{w}^\star)) dt \right)^{-1} = \left(\nabla_{\boldsymbol{w}}^2 \mathcal{L}(\boldsymbol{w}^\star) \right)^{-1} = (\boldsymbol{H}^\star)^{-1}.$$

Let us use $\mathcal{T}$ to denote the corresponding sample space of the probability measure $\mathcal{P}$. We define the function class

$$\mathcal{G} = \left\{ \boldsymbol{g} : \mathcal{T} \to \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|} : \mathbb{E}_{\zeta \sim \mathcal{P}} \left[\boldsymbol{g}(\zeta) \right] = \mathbf{0}, \mathbb{E}_{\zeta \sim \mathcal{P}} \left\| \boldsymbol{g}(\zeta) \right\|^2 < \infty \right\}.$$

Fix an arbitrary function $\boldsymbol{g} \in \mathcal{G}$, and let $h : \mathbb{R} \to [-1, 1]$ be any C^3-smooth function whose first three derivatives are globally bounded and that satisfies $h(t) = t$ for $t \in [-1/2, 1/2]$. For each $\boldsymbol{v} \in \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|}$, we define a perturbed distribution whose density with respect to $\mathcal{P}$ is

$$d\mathcal{P}_{\boldsymbol{v}}(\zeta) = \frac{1 + h(\boldsymbol{v}^\top \boldsymbol{g}(\zeta))}{C(\boldsymbol{v})} d\mathcal{P}(\zeta) \qquad \text{where} \qquad C(\boldsymbol{v}) = 1 + \int h(\boldsymbol{v}^\top \boldsymbol{g}(\zeta)) d\mathcal{P}(\zeta).$$

Each vector $\boldsymbol{v}$ specifies a perturbed optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \quad f_{\boldsymbol{v}}(\boldsymbol{x}) = \mathbb{E}_{\zeta \sim \mathcal{P}_{\boldsymbol{v}}} \left[F(\boldsymbol{x}; \zeta) \right] \qquad \text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) = \mathbf{0}, \quad \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}. \tag{C.4}$$

We further define the associated perturbed optimality mapping $\nabla_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})$ analogously to (C.2), except that $\nabla f(\boldsymbol{x})$ is replaced by $\nabla f_{\boldsymbol{v}}(\boldsymbol{x})$. We introduce the following two lemmas regarding the mapping $\nabla_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})$, which are proved in Appendices D.1 and D.2, respectively.

Lemma C.1. *The mapping $\nabla_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})$ is continuously differentiable in a neighborhood of $(\boldsymbol{w}, \boldsymbol{v}) = (\boldsymbol{w}^\star, \mathbf{0})$, with derivatives given by*

$$\nabla_{\boldsymbol{w}}^2 \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}^\star)|_{\boldsymbol{v}=\mathbf{0}} = \nabla_{\boldsymbol{w}}^2 \mathcal{L}(\boldsymbol{w}^\star) = \boldsymbol{H}^\star, \quad \nabla_{\boldsymbol{v}\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})|_{\boldsymbol{v}=\mathbf{0}} = \mathbb{E}_{\zeta \sim \mathcal{P}} \left[\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; \zeta) \boldsymbol{g}(\zeta)^\top \right], \tag{C.5}$$

where $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; \zeta)$ has the same form as $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ in (C.2) but replaces $\nabla f(\boldsymbol{x})$ by $\nabla F(\boldsymbol{x}; \zeta)$.

Lemma C.2. *The solution mapping $S(\boldsymbol{v}) = \{\boldsymbol{w} : \nabla_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}) = \mathbf{0}\}$ admits a single-valued lo-*

calization $s(\boldsymbol{v})$ *in a neighborhood of* $(\boldsymbol{w}, \boldsymbol{v}) = (\boldsymbol{w}^\star, \boldsymbol{0})$. *Moreover,* $s(\cdot)$ *is differentiable at* $\boldsymbol{v} = \boldsymbol{0}$ *with Jacobian* $\nabla s(\boldsymbol{0}) = -\left(\boldsymbol{H}^\star\right)^{-1} \mathbb{E}_{\zeta\sim\mathcal{P}}[\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star;\zeta)\boldsymbol{g}(\zeta)^\top]$.

With the above two lemmas, we then consider the mapping sequence $\Gamma_k(\boldsymbol{v}) = s(\boldsymbol{v}/\sqrt{k})$, where $s(\cdot)$ is the single-valued localization of the solution map from Lemma C.2. In fact, $\Gamma_k(\boldsymbol{v})$ is regular with derivative $\nabla s(\boldsymbol{0})$ (cf. Definition B.2). To see this, we just note that by the differentiability of $s(\boldsymbol{v})$ at $\boldsymbol{v} = \boldsymbol{0}$, we have for all small enough $\boldsymbol{v}$,

$$\lim_{k\to\infty} \sqrt{k}\left(\Gamma_k(\boldsymbol{v}) - \Gamma_k(\boldsymbol{0})\right) = \lim_{k\to\infty} \sqrt{k}(s(\boldsymbol{v}/\sqrt{k}) - s(\boldsymbol{0})) = \nabla s(\boldsymbol{0})\boldsymbol{v}.$$

Now, we are ready to combine all above pieces to finalize the proof of Theorem 2.5. For each $\boldsymbol{v} \in \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|}$, we define the product probability space $(\mathcal{T}_k, \mathcal{F}_k, \mathcal{P}_{k,\boldsymbol{v}}) := (\mathcal{T}, \mathcal{F}, \mathcal{P}_{\boldsymbol{v}/\sqrt{k}})^{\otimes k}$, that is, the k-fold product of $(\mathcal{T}, \mathcal{F}, \mathcal{P}_{\boldsymbol{v}/\sqrt{k}})$. It has been shown in (Duchi and Ruan, 2021, Lemma 8.3) that the sequence $(\mathcal{T}_k, \mathcal{F}_k, \mathcal{P}_{k,\boldsymbol{v}})$ is locally asymptotically normal with the precision matrix $\boldsymbol{Q} = \mathbb{E}_{\zeta\sim\mathcal{P}}\left[\boldsymbol{g}(\zeta)\boldsymbol{g}(\zeta)^\top\right]$. Let us specify

$$\boldsymbol{g}(\zeta) = \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star;\zeta) - \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star) = \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star;\zeta),$$

then we have $\boldsymbol{\Sigma}^\star = \mathbb{E}_{\zeta\sim\mathcal{P}}\left[\boldsymbol{g}(\zeta)\boldsymbol{g}(\zeta)^\top\right]$ (see (13)). Let ℓ be any symmetric, quasiconvex, and lower semicontinuous function, and let $\boldsymbol{w}_k$ denote any sequence of estimators, and $\boldsymbol{w}_{\boldsymbol{v}}^\star$ denote the primal and (active) dual solution of the perturbed problem (C.4) for small enough $\boldsymbol{v}$, i.e., $\boldsymbol{w}_{\boldsymbol{v}}^\star = s(\boldsymbol{v})$. Note that the active set is consistent with the unperturbed active set $\mathcal{A}^\star$. For any finite subset $\mathcal{W}_0 \subseteq \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|}$, we define $c = \max_{\boldsymbol{v}\in\mathcal{W}_0} \|\boldsymbol{v}\|$. Since $\boldsymbol{w}_{\boldsymbol{v}}^\star = s(\boldsymbol{v})$ and $\Gamma_k(\boldsymbol{v}) = s(\boldsymbol{v}/\sqrt{k})$, we have

$$\liminf_{k\to\infty} \sup_{\|\boldsymbol{v}\|\le c/\sqrt{k}} \mathbb{E}_{\mathcal{P}_{\boldsymbol{v}}^{\otimes k}}[\ell(\sqrt{k}(\boldsymbol{w}_k - \boldsymbol{w}_{\boldsymbol{v}}^\star))] \ge \liminf_{k\to\infty} \max_{\boldsymbol{v}\in\mathcal{W}_0} \mathbb{E}_{\mathcal{P}_{k,\boldsymbol{v}}}[\ell(\sqrt{k}(\boldsymbol{w}_k - \Gamma_k(\boldsymbol{v})))].$$

Taking the supremum over all finite $\mathcal{W}_0 \subseteq \mathbb{R}^{d+m+|\mathcal{A}_\ell^\star|+|\mathcal{A}_u^\star|}$ (equivalently, letting $c \to \infty$), and applying the Hájek and Le Cam's local minimax theorem in Theorem B.3, we obtain

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\|\boldsymbol{v}\|\le c/\sqrt{k}} \mathbb{E}_{\mathcal{P}_{\boldsymbol{v}}^{\otimes k}}[\ell(\sqrt{k}(\boldsymbol{w}_k - \boldsymbol{w}_{\boldsymbol{v}}^\star))] \ge \mathbb{E}\left[\ell(Z_\gamma)\right],$$

where $Z_\gamma \sim \mathcal{N}(\boldsymbol{0}, \left(\boldsymbol{H}^\star\right)^{-1}\boldsymbol{\Sigma}^\star(\boldsymbol{\Sigma}^\star + \gamma\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^\star\left(\boldsymbol{H}^\star\right)^{-1})$ for any $\gamma > 0$. Since $\lim_{\gamma\to 0}\boldsymbol{\Sigma}^\star(\boldsymbol{\Sigma}^\star +$

$\gamma \boldsymbol{I})^{-1}\boldsymbol{\Sigma}^\star = \boldsymbol{\Sigma}^\star$, the right-hand side converges to $\mathbb{E}\left[\ell(Z)\right]$ with $Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^\star)$. For the left-hand side, it has been shown in Duchi and Ruan, 2021 that $\mathcal{P}_{\boldsymbol{v}}$ with $\|\boldsymbol{v}\| \le c/\sqrt{k}$ implies $\mathcal{P}_{\boldsymbol{v}} \in \mathcal{B}(\mathcal{P}; c'/k)$ for some $c' > 0$. This establishes the claimed lower bound.

C.3 Proof of Lemma 3.7

We first state a preliminary lemma. The proof is omitted, as it follows directly from the observation that $\boldsymbol{x} + \alpha\Delta\boldsymbol{x} = (1-\alpha)\boldsymbol{x} + \alpha(\boldsymbol{x} + \Delta\boldsymbol{x})$ and that the box constraint set is convex.

Lemma C.3. *Suppose $\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}$ and $\boldsymbol{\ell} \le \boldsymbol{x} + \Delta\boldsymbol{x} \le \boldsymbol{u}$. Then, for any $\alpha \in [0,1]$, $\boldsymbol{\ell} \le \boldsymbol{x} + \alpha\Delta\boldsymbol{x} \le \boldsymbol{u}$. In particular, if $\boldsymbol{\ell} \le \boldsymbol{x}_0 \le \boldsymbol{u}$, then the entire SSQP sequence $\{\boldsymbol{x}_k\}$ satisfies $\boldsymbol{\ell} \le \boldsymbol{x}_k \le \boldsymbol{u}$.*

The KKT conditions for the SQP subproblem at $\boldsymbol{x}_k$ with true gradient ∇f_k show that there exist some dual multipliers $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})$ satisfying

$$\begin{aligned}
&\nabla f_k + \bar{\boldsymbol{B}}_k\Delta\boldsymbol{x}_k + \nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}} + \boldsymbol{\mu}_{2,k}^{\text{sub}} = \mathbf{0},\\
&\theta_k \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \Delta\boldsymbol{x}_k = \mathbf{0}, \qquad\qquad \boldsymbol{\ell} \le \boldsymbol{x}_k + \Delta\boldsymbol{x}_k \le \boldsymbol{u},\\
&\boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{\ell} - \boldsymbol{x}_k - \Delta\boldsymbol{x}_k) = 0, \quad \boldsymbol{\mu}_{2,k}^{\text{sub}\top}(\boldsymbol{x}_k + \Delta\boldsymbol{x}_k - \boldsymbol{u}) = 0,\\
&\boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}} \ge \mathbf{0}.
\end{aligned} \tag{C.6}$$

Multiplying both sides of the first equality by $\Delta\boldsymbol{x}_k$, we obtain

$$\begin{aligned}
\nabla f_k^\top \Delta\boldsymbol{x}_k + \Delta\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k &= -\Delta\boldsymbol{x}_k^\top \nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda}_k^{\text{sub}} + \Delta\boldsymbol{x}_k^\top \boldsymbol{\mu}_{1,k}^{\text{sub}} - \Delta\boldsymbol{x}_k^\top \boldsymbol{\mu}_{2,k}^{\text{sub}}\\
&\overset{(C.6)}{=} \theta_k \boldsymbol{\lambda}_k^{\text{sub}\top} \boldsymbol{c}_k - \boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{x}_k - \boldsymbol{\ell}) + \boldsymbol{\mu}_{2,k}^{\text{sub}\top}(\boldsymbol{x}_k - \boldsymbol{u})\\
&\le \theta_k \boldsymbol{\lambda}_k^{\text{sub}\top} \boldsymbol{c}_k \le \|\boldsymbol{\lambda}_k^{\text{sub}}\| \|\boldsymbol{c}_k\| \le M_{\text{dual}} \|\boldsymbol{c}_k\|,
\end{aligned} \tag{C.7}$$

where the third inequality comes from $\boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}} \ge \mathbf{0}$ and $\boldsymbol{\ell} \le \boldsymbol{x}_k \le \boldsymbol{u}$ (by Lemma C.3). By the definitions (21) and (22), we have

$$\Delta\varphi_\rho^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) = -\nabla f_k^\top \Delta\boldsymbol{x}_k - \frac{1}{2}\Delta\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k + \rho\theta_k \left\|\boldsymbol{c}_k\right\|. \tag{C.8}$$

Thus, to satisfy $\Delta\varphi_\rho^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) \ge \frac{1}{2}\Delta\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k + \nu\rho\theta_k\|\boldsymbol{c}_k\|$, it suffices to satisfy

$$\nabla f_k^\top \Delta\boldsymbol{x}_k + \Delta\boldsymbol{x}_k^\top \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k \le (1-\nu)\rho\theta_k\|\boldsymbol{c}_k\|.$$

Combining with (C.7), we know the above display holds if

$$M_{\text{dual}}\|\boldsymbol{c}_k\| \le (1-\nu)\rho\theta_k \|\boldsymbol{c}_k\|,$$

which is further implied by the stated condition $\rho \ge \bar{\rho} := \frac{M_{\text{dual}}}{(1-\nu)\tau\bar{\theta}}$ since $\theta_k \ge \tau\bar{\theta}$ by Assumption 3.4. This completes the proof.

C.4 Proof of Theorem 3.8

To simplify notation, we slightly abuse the symbol $\mathbb{E}_k[\cdot]$ throughout the proof to denote the conditional expectation given the randomness of $\boldsymbol{x}_0$ through $\boldsymbol{x}_k$. Since $\bar{\alpha}_k \le \alpha_k + \psi\alpha_k^p$ and $\theta_k \le 1$, we may, without loss of generality, assume that $\bar{\alpha}_k\theta_k \le 1$ (otherwise, we can simply restrict our analysis for $k \ge k_0$ for some deterministic k_0). For any $\rho > 0$, it follows from Assumptions 3.4, 3.5, and 3.6 that

$$\begin{aligned}
&\varphi_\rho(\boldsymbol{x}_{k+1}) - \varphi_\rho(\boldsymbol{x}_k) \\
&= f(\boldsymbol{x}_k + \bar{\alpha}_k\bar{\Delta}\boldsymbol{x}_k) - f(\boldsymbol{x}_k) + \rho(\|\boldsymbol{c}(\boldsymbol{x}_k + \bar{\alpha}_k\bar{\Delta}\boldsymbol{x}_k)\| - \|\boldsymbol{c}_k\|) \\
&\le \bar{\alpha}_k\nabla f_k^\top\bar{\Delta}\boldsymbol{x}_k + \frac{\kappa_{\nabla f}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2 + \rho\left(\|\boldsymbol{c}_k + \bar{\alpha}_k\nabla\boldsymbol{c}_k\bar{\Delta}\boldsymbol{x}_k\| - \|\boldsymbol{c}_k\| + \frac{\kappa_{\nabla c}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2\right) \\
&= \bar{\alpha}_k\nabla f_k^\top\bar{\Delta}\boldsymbol{x}_k - \bar{\alpha}_k\rho\theta_k\|\boldsymbol{c}_k\| + \frac{\kappa_{\nabla f} + \rho\kappa_{\nabla c}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2 \\
&= \bar{\alpha}_k\nabla f_k^\top\Delta\boldsymbol{x}_k + \bar{\alpha}_k\nabla f_k^\top\left(\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\right) - \bar{\alpha}_k\rho\theta_k\|\boldsymbol{c}_k\| + \frac{\kappa_{\nabla f} + \rho\kappa_{\nabla c}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2 \\
&\overset{(C.8)}{=} -\bar{\alpha}_k\Delta\varphi_\rho^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) - \frac{\bar{\alpha}_k}{2}\Delta\boldsymbol{x}_k^\top\bar{\boldsymbol{B}}_k\Delta\boldsymbol{x}_k + \bar{\alpha}_k\nabla f_k^\top\left(\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\right) + \frac{\kappa_{\nabla f} + \rho\kappa_{\nabla c}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2 \\
&\le -\bar{\alpha}_k\Delta\varphi_\rho^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) + \bar{\alpha}_k\nabla f_k^\top\left(\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\right) + \frac{\kappa_{\nabla f} + \rho\kappa_{\nabla c}}{2}\bar{\alpha}_k^2\|\bar{\Delta}\boldsymbol{x}_k\|^2 \\
&\le -\bar{\alpha}_k\Delta\varphi_\rho^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) + \bar{\alpha}_k M_{\nabla f}\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\| + \frac{\kappa_{\nabla f} + \bar{\rho}\kappa_{\nabla c}}{2}\bar{\alpha}_k^2 M_{\boldsymbol{\ell},\boldsymbol{u}}^2,
\end{aligned}$$

where the last inequality holds for some constants $M_{\nabla f}, M_{\boldsymbol{\ell},\boldsymbol{u}}^2 > 0$ since box constraints form a compact set. Setting $\rho = \bar{\rho}$, we apply Lemma 3.7, Assumptions 3.4 and 3.5, and obtain

$$\Delta\varphi_{\bar{\rho}}^{\text{loc}}(\boldsymbol{x}_k, \Delta\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k) \ge \frac{1}{2}\Delta\boldsymbol{x}_k^\top\bar{\boldsymbol{B}}_k\Delta\boldsymbol{x}_k + \nu\bar{\rho}\tau\bar{\theta}\|\boldsymbol{c}_k\| \ge \frac{\kappa_1}{2}\|\Delta\boldsymbol{x}_k\|^2 + \nu\bar{\rho}\tau\bar{\theta}\|\boldsymbol{c}_k\|.$$

Combining the above two displays and taking conditional expectation over $\boldsymbol{x}_0$ to $\boldsymbol{x}_k$, we obtain

$$\mathbb{E}_k\left[\varphi_{\bar{\rho}}(\boldsymbol{x}_{k+1}) - \min_{\boldsymbol{\ell}\le\boldsymbol{x}\le\boldsymbol{u}}\varphi_{\bar{\rho}}(\boldsymbol{x})\right] \le \left\{\varphi_{\bar{\rho}}(\boldsymbol{x}_k) - \min_{\boldsymbol{\ell}\le\boldsymbol{x}\le\boldsymbol{u}}\varphi_{\bar{\rho}}(\boldsymbol{x})\right\} - \alpha_k\mathbb{E}_k\left[\frac{\kappa_1}{2}\|\Delta\boldsymbol{x}_k\|^2 + \nu\bar{\rho}\tau\bar{\theta}\|\boldsymbol{c}_k\|\right]$$

$$+ M_{\nabla f}\left(\alpha_k + \psi\alpha_k^p\right)\mathbb{E}_k[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] + \frac{\kappa_{\nabla f} + \bar{\rho}\kappa_{\nabla c}}{2} M_{\boldsymbol{\ell},\boldsymbol{u}}^2\left(\alpha_k + \psi\alpha_k^p\right)^2, \quad \forall k \geq 0. \tag{C.9}$$

Now, we introduce the following lemma, which is proved in Appendix D.3.

Lemma C.4. *Under the conditions of Theorem 3.8, we have* $\sum_{k=0}^{\infty}\alpha_k\mathbb{E}[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] < \infty$.

With the above lemma, we directly have $\mathbb{E}[\sum_{k=0}^{\infty}\alpha_k\mathbb{E}_k[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|]] = \sum_{k=0}^{\infty}\alpha_k\mathbb{E}[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] < \infty$, which implies that $\sum_{k=0}^{\infty}\left(\alpha_k + \psi\alpha_k^p\right)\mathbb{E}_k[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] < \infty$ almost surely. From (C.9), together with the condition $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, we apply the Robbins-Siegmund theorem (Robbins and Siegmund, 1971) and obtain $\sum_{k=0}^{\infty}\alpha_k(\|\Delta\boldsymbol{x}_k\|^2 + \|\boldsymbol{c}_k\|) < \infty$. This completes the proof by noting that $\sum_{k=0}^{\infty}\alpha_k = \infty$.

C.5 Proof of Theorem 3.9

Let us denote the objective function of Problem (23) by

$$G(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}_k) = \|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{\lambda}, \boldsymbol{\mu})\|^2 + \|\boldsymbol{\mu}_1 \odot (\boldsymbol{\ell} - \boldsymbol{x}_k)\|^2 + \|\boldsymbol{\mu}_2 \odot (\boldsymbol{x}_k - \boldsymbol{u})\|^2.$$

By the definition of the KKT residual $\boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star)$ in (20), it suffices to show $G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k) \to 0$ and $\|\boldsymbol{c}_k\| \to 0$ as $k \to \infty$ almost surely.

• **Convergence of $G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k)$.** Recall from Assumption 3.5 that $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})$ denotes the dual multipliers of the SQP subproblem (17) with the true gradient ∇f_k. It follows from the KKT conditions in (C.6) that

$$\begin{aligned} G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k) \leq G(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}}; \boldsymbol{x}_k) &\leq \|\boldsymbol{B}_k\Delta\boldsymbol{x}_k\|^2 + \left\|\boldsymbol{\mu}_{1,k}^{\text{sub}} \odot \Delta\boldsymbol{x}_k\right\|^2 + \left\|\boldsymbol{\mu}_{2,k}^{\text{sub}} \odot \Delta\boldsymbol{x}_k\right\|^2 \\ &\leq \left(\kappa_2^2 + 2M_{\text{dual}}^2\right)\|\Delta\boldsymbol{x}_k\|^2, \end{aligned} \tag{C.10}$$

which implies from Theorem 3.8 that $\liminf_{k\to\infty} G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k) = 0$. Suppose $\limsup_{k\to\infty} G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k) > 0$, we can find a sufficiently small number $\varepsilon > 0$ and two infinite sequences $\{m_i, n_i\}$ with $m_i < n_i$, such that

$$G(\boldsymbol{\lambda}_{m_i}^\star, \boldsymbol{\mu}_{m_i}^\star; \boldsymbol{x}_{m_i}) > 2\varepsilon, \quad \|\Delta\boldsymbol{x}_k\| \geq \sqrt{\frac{\varepsilon}{\kappa_2^2 + 2M_{\text{dual}}^2}} \text{ for } k \in [m_i, n_i), \quad \|\Delta\boldsymbol{x}_{n_i}\| \leq \sqrt{\frac{\varepsilon}{\kappa_2^2 + 2M_{\text{dual}}^2}}.$$

We can always find such sequences since $G(\boldsymbol{\lambda}_{m_i}^\star, \boldsymbol{\mu}_{m_i}^\star; \boldsymbol{x}_{m_i}) > 2\varepsilon$ implies $\|\Delta\boldsymbol{x}_{m_i}\| \geq \sqrt{\frac{2\varepsilon}{\kappa_2^2 + 2M_{\text{dual}}^2}}$;

and due to $\liminf_{k\to\infty} \|\Delta \boldsymbol{x}_k\| = 0$, there must exist $n_i > m_i$ such that $\|\Delta \boldsymbol{x}_{n_i}\| \le \sqrt{\frac{\varepsilon}{\kappa_2^2 + 2M_{\text{dual}}^2}}$. Let

$$\widetilde{G}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}) = G(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}) + \frac{\varepsilon}{6M_{\text{dual}}^2} \|(\boldsymbol{\lambda}, \boldsymbol{\mu})\|^2$$

and denote $\boldsymbol{w}_k = (\widetilde{\boldsymbol{\lambda}}_k, \widetilde{\boldsymbol{\mu}}_k) \in \arg\min_{\boldsymbol{\lambda}, \boldsymbol{\mu} \ge \boldsymbol{0}} \widetilde{G}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}_k)$. By the above definition of $\widetilde{G}$ and the construction of the sequences $\{m_i, n_i\}$, we have

$$\begin{aligned} &\widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{m_i}, \widetilde{\boldsymbol{\mu}}_{m_i}; \boldsymbol{x}_{m_i}) \ge G(\widetilde{\boldsymbol{\lambda}}_{m_i}, \widetilde{\boldsymbol{\mu}}_{m_i}; \boldsymbol{x}_{m_i}) \ge G(\boldsymbol{\lambda}^\star_{m_i}, \boldsymbol{\mu}^\star_{m_i}; \boldsymbol{x}_{m_i}) \ge 2\varepsilon, \\ &\widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{n_i}, \widetilde{\boldsymbol{\mu}}_{n_i}; \boldsymbol{x}_{n_i}) \le \widetilde{G}(\boldsymbol{\lambda}^{\text{sub}}_{n_i}, \boldsymbol{\mu}^{\text{sub}}_{n_i}; \boldsymbol{x}_{n_i}) \overset{\text{(C.10)}}{\le} (\kappa_2^2 + 2M_{\text{dual}}^2) \|\Delta \boldsymbol{x}_{n_i}\|^2 + \frac{\varepsilon}{3} \le \frac{4}{3}\varepsilon. \end{aligned} \tag{C.11}$$

Next, we analyze the consecutive difference of $\widetilde{G}$. In particular, let us write $\widetilde{G}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}_k)$ into a quadratic form of $\boldsymbol{w} = (\boldsymbol{\lambda}, \boldsymbol{\mu}) = (\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$:

$$\widetilde{G}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{x}_k) = \|\nabla f_k\|^2 + \boldsymbol{q}_k^\top \boldsymbol{w} + \frac{1}{2} \boldsymbol{w}^\top \nabla^2 \widetilde{G}_k \boldsymbol{w},$$

where

$$\boldsymbol{q}_k = \begin{pmatrix} 2\nabla \boldsymbol{c}_k \nabla f_k \\ -2\nabla f_k \\ 2\nabla f_k \end{pmatrix}$$

and

$$\nabla^2 \widetilde{G}_k = \begin{pmatrix} 2\nabla \boldsymbol{c}_k \nabla \boldsymbol{c}_k^\top & -2\nabla \boldsymbol{c}_k & 2\nabla \boldsymbol{c}_k \\ -2\nabla \boldsymbol{c}_k^\top & 2\boldsymbol{I} + 2\text{diag}^2(\boldsymbol{\ell} - \boldsymbol{x}_k) & -2\boldsymbol{I} \\ 2\nabla \boldsymbol{c}_k^\top & -2\boldsymbol{I} & 2\boldsymbol{I} + 2\text{diag}^2(\boldsymbol{x}_k - \boldsymbol{u}) \end{pmatrix} + \frac{\varepsilon}{3M_{\text{dual}}^2} \boldsymbol{I}.$$

Then, we have for any $k \ge 0$,

$$\begin{aligned} &|\widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{k+1}, \widetilde{\boldsymbol{\mu}}_{k+1}; \boldsymbol{x}_{k+1}) - \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_k, \widetilde{\boldsymbol{\mu}}_k; \boldsymbol{x}_k)| \\ &\le \left| \boldsymbol{q}_{k+1}^\top \boldsymbol{w}_{k+1} + \frac{1}{2} \boldsymbol{w}_{k+1}^\top \nabla^2 \widetilde{G}_{k+1} \boldsymbol{w}_{k+1} - \boldsymbol{q}_k^\top \boldsymbol{w}_k - \frac{1}{2} \boldsymbol{w}_k^\top \nabla^2 \widetilde{G}_k \boldsymbol{w}_k \right| + \left| \|\nabla f_{k+1}\|^2 - \|\nabla f_k\|^2 \right| \\ &\le \left| \boldsymbol{q}_{k+1}^\top \boldsymbol{w}_{k+1} + \frac{1}{2} \boldsymbol{w}_{k+1}^\top \nabla^2 \widetilde{G}_{k+1} \boldsymbol{w}_{k+1} - \boldsymbol{q}_k^\top \boldsymbol{w}_{k+1} - \frac{1}{2} \boldsymbol{w}_{k+1}^\top \nabla^2 \widetilde{G}_k \boldsymbol{w}_{k+1} \right| \\ &\quad + \left| \boldsymbol{q}_k^\top \boldsymbol{w}_{k+1} + \frac{1}{2} \boldsymbol{w}_{k+1}^\top \nabla^2 \widetilde{G}_k \boldsymbol{w}_{k+1} - \boldsymbol{q}_k^\top \boldsymbol{w}_k - \frac{1}{2} \boldsymbol{w}_k^\top \nabla^2 \widetilde{G}_k \boldsymbol{w}_k \right| + \left| \|\nabla f_{k+1}\|^2 - \|\nabla f_k\|^2 \right| \\ &\le \|\boldsymbol{w}_{k+1}\| \|\boldsymbol{q}_{k+1} - \boldsymbol{q}_k\| + \frac{1}{2} \|\boldsymbol{w}_{k+1}\|^2 \|\nabla^2 \widetilde{G}_{k+1} - \nabla^2 \widetilde{G}_k\| + \|\boldsymbol{q}_k\| \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\| \\ &\quad + \frac{1}{2} (\|\boldsymbol{w}_{k+1}\| + \|\boldsymbol{w}_k\|) \|\nabla^2 \widetilde{G}_k\| \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\| + \|\nabla f_{k+1} - \nabla f_k\| \left(\|\nabla f_{k+1}\| + \|\nabla f_k\| \right) \end{aligned} \tag{C.12}$$

We provide the bound for each of term in the above display. We note that for any $k \ge 0$,

$$\frac{\varepsilon}{6M_{\text{dual}}^2} \|\boldsymbol{w}_k\|^2 \le \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_k, \widetilde{\boldsymbol{\mu}}_k; \boldsymbol{x}_k)$$

$$\leq G(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}}; \boldsymbol{x}_k) + \frac{\varepsilon}{6M_{\text{dual}}^2}\|(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})\|^2 \overset{\text{(C.10)}}{\leq} (\kappa_2^2 + 2M_{\text{dual}}^2)\|\Delta \boldsymbol{x}_k\|^2 + \frac{\varepsilon}{3}.$$

Thus, we have

$$\|\boldsymbol{w}_k\| \leq \sqrt{\frac{6M_{\text{dual}}^2\,(\kappa_2^2 + 2M_{\text{dual}}^2)\,M_{\boldsymbol{\ell},\boldsymbol{u}}^2}{\varepsilon}} + \sqrt{2}M_{\text{dual}}. \tag{C.13}$$

Furthermore, by the smoothness of the objective $f(\boldsymbol{x})$ and the constraints $\boldsymbol{c}(\boldsymbol{x})$, as well as the boundedness of $\nabla \boldsymbol{c}, \nabla f, \bar{\Delta}\boldsymbol{x}$, we know there exists a constant $M_{\boldsymbol{q},\nabla^2\widetilde{G}} > 0$ such that for $k \geq 0$,

$$\max\{\|\boldsymbol{q}_{k+1} - \boldsymbol{q}_k\|, \|\nabla^2\widetilde{G}_{k+1} - \nabla^2\widetilde{G}_k\|\} \leq M_{\boldsymbol{q},\nabla^2\widetilde{G}} \cdot \alpha_k. \tag{C.14}$$

Finally, we note that for any $k \geq 0$ and any vector $\boldsymbol{w} = (\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \in \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^d$,

$$\frac{1}{2}\boldsymbol{w}^\top \left(\nabla^2\widetilde{G}_k - \frac{\varepsilon}{3M_{\text{dual}}^2}\boldsymbol{I}\right)\boldsymbol{w} = \|\nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|^2 + \|\,(\boldsymbol{x}_k - \boldsymbol{\ell}) \odot \boldsymbol{\mu}_1\|^2 + \|\,(\boldsymbol{x}_k - \boldsymbol{u}) \odot \boldsymbol{\mu}_2\|^2 \geq 0.$$

We combine the above display with (C.13) and (C.14), apply Lemma B.5, and have $\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\| = \mathcal{O}\left(\frac{\alpha_k}{\varepsilon^{3/2}}\right)$, where we have omitted universal deterministic constants that are independent of ε. Combining this result with (C.13) and (C.14), and plugging into (C.12), we obtain

$$|\widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{k+1}, \widetilde{\boldsymbol{\mu}}_{k+1}; \boldsymbol{x}_{k+1}) - \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_k, \widetilde{\boldsymbol{\mu}}_k; \boldsymbol{x}_k)| \leq M_G \frac{\alpha_k}{\varepsilon^2}, \qquad \forall k \geq 0,$$

for some universal constant $M_G > 0$ that is independent of α_k and ε. Combining the above display with (C.11), we obtain

$$\begin{aligned}\frac{2}{3}\varepsilon &\leq \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{m_i}, \widetilde{\boldsymbol{\mu}}_{m_i}; \boldsymbol{x}_{m_i}) - \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{n_i}, \widetilde{\boldsymbol{\mu}}_{n_i}; \boldsymbol{x}_{n_i}) \\ &\leq \sum_{k=m_i}^{n_i-1} \left|\widetilde{G}(\widetilde{\boldsymbol{\lambda}}_k, \widetilde{\boldsymbol{\mu}}_k; \boldsymbol{x}_k) - \widetilde{G}(\widetilde{\boldsymbol{\lambda}}_{k+1}, \widetilde{\boldsymbol{\mu}}_{k+1}; \boldsymbol{x}_{k+1})\right| \leq \frac{M_G}{\varepsilon^2}\sum_{k=m_i}^{n_i-1}\alpha_k.\end{aligned}$$

Summing up both sides from $i = 1$ to ∞, we have

$$\infty = \sum_{i=1}^{\infty}\frac{2\varepsilon^3}{3M_G} \leq \sum_{i=1}^{\infty}\sum_{k=m_i}^{n_i-1}\alpha_k.$$

On the other hand, since $\|\Delta \boldsymbol{x}_k\| \geq \sqrt{\frac{\varepsilon}{\kappa_2^2 + 2M_{\text{dual}}^2}}$ for $m_i \leq k \leq n_i - 1$, we know that

$$\sum_{i=1}^{\infty}\sum_{k=m_i}^{n_i-1}\alpha_k \leq \frac{\kappa_2^2 + 2M_{\text{dual}}^2}{\varepsilon}\sum_{i=1}^{\infty}\sum_{k=m_i}^{n_i-1}\alpha_k\|\Delta \boldsymbol{x}_k\|^2 \leq \frac{\kappa_2^2 + 2M_{\text{dual}}^2}{\varepsilon}\sum_{k=0}^{\infty}\alpha_k\|\Delta \boldsymbol{x}_k\|^2 < \infty,$$

where the last inequality is due to the proof of Theorem 3.8. This derives the contradiction and we complete the proof of $\lim_{k\to\infty} G(\boldsymbol{\lambda}_k^\star, \boldsymbol{\mu}_k^\star; \boldsymbol{x}_k) = 0$.

• **Convergence of $\|\boldsymbol{c}_k\|$.** We follow the same proof strategy. By Theorem 3.8 we know that $\liminf_{k\to\infty} \|\boldsymbol{c}_k\| = 0$. Suppose $\limsup_{k\to\infty} \|\boldsymbol{c}_k\| > 0$, then we can find a sufficiently small number $\varepsilon > 0$ and two infinite sequences $\{m_i, n_i\}$ with $m_i < n_i$, such that

$$\|\boldsymbol{c}_{m_i}\| > 2\varepsilon, \qquad \|\boldsymbol{c}_k\| \geq \varepsilon \ \text{ for } \ k \in [m_i, n_i), \qquad \|\boldsymbol{c}_{n_i}\| \leq \varepsilon.$$

It follows from the construction of the sequences that

$$\varepsilon \leq \|\boldsymbol{c}_{m_i}\| - \|\boldsymbol{c}_{n_i}\| = \sum_{k=m_i}^{n_i-1} (\|\boldsymbol{c}_k\| - \|\boldsymbol{c}_{k+1}\|) \leq \sum_{k=m_i}^{n_i-1} \|\boldsymbol{c}_k - \boldsymbol{c}_{k+1}\| \leq M_c \sum_{k=m_i}^{n_i-1} \alpha_k$$

for some universal constant $M_c > 0$ that is independent of α_k and ε. Multiplying both sides by ε and noting the fact that $\|\boldsymbol{c}_k\| \geq \varepsilon$ for $m_i \leq k < n_i$, we have

$$\varepsilon^2 \leq M_c \sum_{k=m_i}^{n_i-1} \alpha_k \|\boldsymbol{c}_k\|,$$

which implies that $\infty \leq \sum_{i=1}^{\infty} \sum_{k=m_i}^{n_i-1} \alpha_k \|\boldsymbol{c}_k\| \leq \sum_{k=0}^{\infty} \alpha_k \|\boldsymbol{c}_k\| < \infty$. This leads to the contradiction and we complete the proof.

C.6 Proof of Lemma 4.2

• **Proof of (a).** By Assumption 4.1, we know LICQ holds at $\boldsymbol{x}^\star$. Thus, when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}^\star$, the rows of $\widetilde{\boldsymbol{J}}_k = (\nabla \boldsymbol{c}_k; -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell}}^\star}; \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u}}^\star})$ are linearly independent. To show $\theta_k = 1$ for k large enough, it suffices to show that there exists $\boldsymbol{z}_k \in \mathbb{R}^d$ such that $\boldsymbol{c}_k + \nabla \boldsymbol{c}_k \boldsymbol{z}_k = \boldsymbol{0}$ and $\boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{z}_k \leq \boldsymbol{u}$. To this end, let us define the linear system at $\boldsymbol{x}_k$ as

$$\begin{pmatrix} \nabla \boldsymbol{c}_k \\ -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell}}^\star} \\ \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u}}^\star} \end{pmatrix} \boldsymbol{z}_k = - \begin{pmatrix} \boldsymbol{c}_k \\ [\boldsymbol{\ell} - \boldsymbol{x}_k]_{\mathcal{A}_{\boldsymbol{\ell}}^\star} \\ [\boldsymbol{x}_k - \boldsymbol{u}]_{\mathcal{A}_{\boldsymbol{u}}^\star} \end{pmatrix}.$$

By the full row-rankness of $\widetilde{\boldsymbol{J}}_k$, the above linear system has a solution given by

$$\boldsymbol{z}_k = -\widetilde{\boldsymbol{J}}_k^\top (\widetilde{\boldsymbol{J}}_k \widetilde{\boldsymbol{J}}_k^\top)^{-1} \begin{pmatrix} \boldsymbol{c}_k \\ [\boldsymbol{\ell} - \boldsymbol{x}_k]_{\mathcal{A}_{\boldsymbol{\ell}}^\star} \\ [\boldsymbol{x}_k - \boldsymbol{u}]_{\mathcal{A}_{\boldsymbol{u}}^\star} \end{pmatrix}.$$

Clearly, the above $\boldsymbol{z}_k$ satisfies $\boldsymbol{c}_k + \nabla \boldsymbol{c}_k \boldsymbol{z}_k = \boldsymbol{0}$ and $[\boldsymbol{l}]_i \leq [\boldsymbol{x}_k + \boldsymbol{z}_k]_i \leq [\boldsymbol{u}]_i$ for all $i \in \mathcal{A}^\star = \mathcal{A}_{\boldsymbol{\ell}}^\star \cup \mathcal{A}_{\boldsymbol{u}}^\star$. Furthermore, since $\boldsymbol{z}_k \to \boldsymbol{0}$ as $\boldsymbol{x}_k \to \boldsymbol{x}^\star$, we also have $[\boldsymbol{l}]_i \leq [\boldsymbol{x}_k + \boldsymbol{z}_k]_i \leq [\boldsymbol{u}]_i$ for all $i \in (\mathcal{A}^\star)^-$. This implies that $\boldsymbol{z}_k \in \Omega(\boldsymbol{x}_k; \theta_k)$ with $\theta_k = 1$ and we complete the proof.

• **Proof of (b).** Let us consider $\boldsymbol{x}_k$ such that $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_{\max} = \max_{i\in[d]} |[\boldsymbol{x}_k - \boldsymbol{x}^\star]_i| \leq 0.5\epsilon$. For any

$i \in \mathcal{A}_{\boldsymbol{\ell}}^\star$, we have $[\boldsymbol{x}_k - \boldsymbol{\ell}]_i \le \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_{\max} \le 0.5\epsilon$, implying that $\mathcal{A}_{\boldsymbol{\ell}}^\star \subseteq \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ by the definition of (25). On the other hand, for any $i \in \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$, we have $[\boldsymbol{x}^\star - \boldsymbol{\ell}]_i \le |[\boldsymbol{x}_k - \boldsymbol{x}^\star]_i| + [\boldsymbol{x}_k - \boldsymbol{\ell}]_i \le \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_{\max} + [\boldsymbol{x}_k - \boldsymbol{\ell}]_i \le 1.5\epsilon < [\boldsymbol{x}^\star - \boldsymbol{\ell}]_j$, for any $j \in (\mathcal{A}^\star)^-$, where the last inequality is by the definition (24). Thus, we know $i \in \mathcal{A}^\star$. By the same derivations, we also have $[\boldsymbol{x}^\star - \boldsymbol{\ell}]_i < [\boldsymbol{u} - \boldsymbol{\ell}]_i$, which implies $i \in \mathcal{A}_{\boldsymbol{\ell}}^\star$. Therefore, we have $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}_{\boldsymbol{\ell}}^\star$ as long as $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_{\max} \le 0.5\epsilon$. Similar arguments hold for showing that $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}_{\boldsymbol{u}}^\star$, and we complete the proof.

• **Proof of (c) and (d).** We prove **(c)** and **(d)** jointly. We need the following supporting lemma, which is proved in Appendix D.6.

Lemma C.5. *Under Assumptions 3.6 and 4.1, let $\beta_k = \iota_2 (k+1)^{-b_2}$ satisfy $\iota_2 > 0, b_2 \in (0.5, 1]$. We have $\bar{\boldsymbol{g}}_k \to \nabla f^\star$ as $k \to \infty$ almost surely.*

We will first show that the subproblem (17) admits a solution $\bar{\Delta}\boldsymbol{x}_k$ near the origin for k large enough, with $\bar{\Delta}\boldsymbol{x}_k \to 0$ as $k \to \infty$ and satisfying $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell}}^\star$ and $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{u}}^\star$. We then show that the subproblem solution is also unique in a neighborhood of the origin. Let us consider the following linear system:

$$\begin{pmatrix} \bar{\boldsymbol{B}}_k & \nabla \boldsymbol{c}_k^\top & -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}^\top & \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)}^\top \\ \nabla \boldsymbol{c}_k & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_k \\ \bar{\boldsymbol{\lambda}}_k^{\mathrm{sub}} \\ [\bar{\boldsymbol{\mu}}_{1,k}^{\mathrm{sub}}]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} \\ [\bar{\boldsymbol{\mu}}_{2,k}^{\mathrm{sub}}]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} \end{pmatrix} = - \begin{pmatrix} \bar{\boldsymbol{g}}_k \\ \boldsymbol{c}_k \\ [\boldsymbol{\ell} - \boldsymbol{x}_k]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} \\ [\boldsymbol{x}_k - \boldsymbol{u}]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} \end{pmatrix}. \tag{C.15}$$

By the construction of $\bar{\boldsymbol{B}}_k$, the local LICQ in Assumption 4.1, and the properties $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}_{\boldsymbol{\ell}}^\star$ and $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}_{\boldsymbol{u}}^\star$ proved in **(b)**, we know the above system has a (unique) solution. Let us rewrite the system as

$$\begin{aligned} &\begin{pmatrix} \bar{\boldsymbol{B}}_k & \nabla \boldsymbol{c}_k^\top & -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}^\top & \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)}^\top \\ \nabla \boldsymbol{c}_k & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \bar{\Delta}\boldsymbol{x}_k \\ \bar{\boldsymbol{\lambda}}_k^{\mathrm{sub}} - \boldsymbol{\lambda}^\star \\ [\bar{\boldsymbol{\mu}}_{1,k}^{\mathrm{sub}} - \boldsymbol{\mu}_1^\star]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} \\ [\bar{\boldsymbol{\mu}}_{2,k}^{\mathrm{sub}} - \boldsymbol{\mu}_2^\star]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} \end{pmatrix} \\ &\qquad = - \begin{pmatrix} \bar{\boldsymbol{g}}_k + \nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda}^\star - \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}^\top [\boldsymbol{\mu}_1^\star]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} + \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)}^\top [\boldsymbol{\mu}_2^\star]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} \\ \boldsymbol{c}_k \\ [\boldsymbol{\ell} - \boldsymbol{x}_k]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)} \\ [\boldsymbol{x}_k - \boldsymbol{u}]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)} \end{pmatrix}. \end{aligned} \tag{C.16}$$

By the convergence of $\boldsymbol{x}_k$, $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$, $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$, and Lemma C.5, we know the right hand side of (C.16)

converges to 0 because it is the KKT conditions at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$, implying that $(\bar{\Delta}\boldsymbol{x}_k; \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}^\star; [\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} - \boldsymbol{\mu}_1^\star]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}; [\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} - \boldsymbol{\mu}_2^\star]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)}) \to 0$. Now, let us show that $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}_{\boldsymbol{\ell}}^\star$ and $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}_{\boldsymbol{u}}^\star$ when $\|\bar{\Delta}\boldsymbol{x}_k\|_{\max} \le \epsilon$. By the system (C.16), we have $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) \subseteq \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k)$. On the other hand, for any $i \in \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k)$, we have $|[\boldsymbol{x}_k]_i - [\boldsymbol{\ell}]_i| \le |[\bar{\Delta}\boldsymbol{x}_k]_i| \le \epsilon$, suggesting that $i \in \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ and hence $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) \subseteq \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$. This leads to $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k)$. Similarly, we can show $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k)$. Next, we complement $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k(\epsilon)} = ([\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}, [\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}}]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)})$ by defining $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k^-(\epsilon)} = \mathbf{0}$, and aim to show that $(\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_k^{\text{sub}})$ is the local solution of the SSQP subproblem (17). By (C.15) and the complement $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k^-(\epsilon)} = \mathbf{0}$, we have

$$\begin{aligned} &\bar{\boldsymbol{g}}_k + \bar{\boldsymbol{B}}_k \bar{\Delta}\boldsymbol{x}_k + \nabla \boldsymbol{c}_k^\top \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} + \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} = \mathbf{0}, \\ &\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}\top} \left(\boldsymbol{\ell} - \boldsymbol{x}_k - \bar{\Delta}\boldsymbol{x}_k\right) = \mathbf{0}, \\ &\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}\top} \left(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k - \boldsymbol{u}\right) = \mathbf{0}. \end{aligned} \tag{C.17}$$

Furthermore, since $[\boldsymbol{\mu}^\star]_{\mathcal{A}_k(\epsilon)} > 0$ by strict complementarity condition and $[\bar{\boldsymbol{\mu}}_k^{\text{sub}} - \boldsymbol{\mu}^\star]_{\mathcal{A}_k(\epsilon)} \to 0$ as $k \to \infty$, we have $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k(\epsilon)} > 0$ for k large enough. This verifies the KKT conditions of the subproblem (17). The conditions of LICQ, SOSC, and strict complementarity of the subproblem are trivial due to the facts that $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ and $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$.

Finally, we show the uniqueness of the solution $\bar{\Delta}\boldsymbol{x}_k$ in the neighborhood of $\mathbf{0}$. Suppose there exists another solution $\bar{\Delta}\boldsymbol{x}_k'$ with $\|\bar{\Delta}\boldsymbol{x}_k'\|_{\max} \le \epsilon$. We know for any $i \in \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')$, $[\boldsymbol{x}_k - \boldsymbol{\ell}]_i \le \|\bar{\Delta}\boldsymbol{x}_k'\|_{\max} \le \epsilon$, suggesting that $i \in \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ and hence $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k') \subseteq \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$. Similarly, we can show $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k') \subseteq \mathcal{A}_{\boldsymbol{u},k}(\epsilon)$. This suggests that the LICQ of the subproblem holds at $\bar{\Delta}\boldsymbol{x}_k'$ when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}^\star$, as $\boldsymbol{J}_k = (\nabla \boldsymbol{c}_k; -\boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}; \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)})$ has full row-rank when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}^\star$. Then, the KKT condition of the subproblem at $\bar{\Delta}\boldsymbol{x}_k'$ implies the existence of unique dual multipliers $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}\prime}, \bar{\boldsymbol{\mu}}_k^{\text{sub}\prime})$ such that

$$\bar{\boldsymbol{g}}_k + \bar{\boldsymbol{B}}_k \bar{\Delta}\boldsymbol{x}_k' + \nabla \boldsymbol{c}_k^\top \bar{\boldsymbol{\lambda}}_k^{\text{sub}\prime} - \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')}^\top [\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}\prime}]_{\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')} + \boldsymbol{I}_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')}^\top [\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}\prime}]_{\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')} = \mathbf{0}.$$

Comparing the above display with (C.17), we obtain

$$\boldsymbol{J}_k^\top \left[\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \bar{\boldsymbol{\lambda}}_k^{\text{sub}\prime}; [\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} - \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}\prime}]_{\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)}; [\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} - \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}\prime}]_{\mathcal{A}_{\boldsymbol{u},k}(\epsilon)}\right] = \bar{\boldsymbol{B}}_k(\bar{\Delta}\boldsymbol{x}_k' - \bar{\Delta}\boldsymbol{x}_k).$$

By the full row-rankness of $\boldsymbol{J}_k$ as implied by LICQ at $\boldsymbol{x}^\star$, we know $\|[\bar{\boldsymbol{\mu}}_k^{\text{sub}} - \bar{\boldsymbol{\mu}}_k^{\text{sub}\prime}]_{\mathcal{A}_k(\epsilon)}\| \le \texttt{const} \cdot (\|\bar{\Delta}\boldsymbol{x}_k'\| + \|\bar{\Delta}\boldsymbol{x}_k\|)$ for some $\texttt{const} > 0$. Thus, using the fact that $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k(\epsilon)} > 0$, we can choose the neighborhood of the origin small enough such that if $\bar{\Delta}\boldsymbol{x}_k', \bar{\Delta}\boldsymbol{x}_k$ are in the neighborhood, we have $[\bar{\boldsymbol{\mu}}_k^{\text{sub}\prime}]_{\mathcal{A}_k(\epsilon)} > 0$ as well, further implying that $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) \subseteq \mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')$ and $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) \subseteq \mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k')$ due to the complementarity conditions of the subproblem. This leads to $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k') = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon)$ and $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k') = \mathcal{A}_{\boldsymbol{u},k}(\epsilon)$. However, by the KKT condition of the SSQP subproblem, the solution $\bar{\Delta}\boldsymbol{x}_k'$ must also follow the linear system (C.15), which admits a unique solution. Thus, the solution of SSQP subproblem is unique near $\mathbf{0}$. The argument about $\Delta\boldsymbol{x}_k$ follows the same analysis by noting that $\nabla f_k \to \nabla f^\star$ almost surely. This completes the proof of **(c)** and **(d)**.

C.7 Proof of Lemma 4.4

We first prove the convergence of $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$. By the proof of Lemma 4.2(c) in Appendix C.6, we know $\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}^\star \to \infty$ and $[\bar{\boldsymbol{\mu}}_k^{\text{sub}} - \boldsymbol{\mu}^\star]_{\mathcal{A}_k(\epsilon)} \to \mathbf{0}$ as $k \to \infty$ almost surely. Let ϵ satisfy (24) and we have $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star$ for k large enough by Lemma 4.2(b). Since $[\bar{\boldsymbol{\mu}}_k^{\text{sub}}]_{\mathcal{A}_k^-(\epsilon)} = \mathbf{0} = \boldsymbol{\mu}^\star_{(\mathcal{A}^\star)^-}$, we have $\bar{\boldsymbol{\mu}}_k^{\text{sub}} - \boldsymbol{\mu}^\star \to \mathbf{0}$ as $k \to \infty$ almost surely. With this result, we now prove $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \to (\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$. We take $\boldsymbol{\lambda}_k$ as an example. By the update (18), we have

$$\begin{aligned}\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^\star = \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star + \bar{\alpha}_k\left(\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}_k\right) &= (1-\bar{\alpha}_k)(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star) + \bar{\alpha}_k(\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}^\star) \\ &= \prod_{i=0}^{k}(1-\bar{\alpha}_i)(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^\star) + \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\bar{\alpha}_j)\bar{\alpha}_i(\bar{\boldsymbol{\lambda}}_i^{\text{sub}} - \boldsymbol{\lambda}^\star).\end{aligned}$$

To proceed, we present the following lemma, which is proved in Appendix D.7.

Lemma C.6. *Suppose $\beta_k \ge 0$ satisfies $\beta_k \to 0$ and $\sum_{k=0}^\infty \beta_k = \infty$, then for any scalar a and sequence $e_k \to 0$, we have $\sum_{i=0}^k \prod_{j=i+1}^k (1-\beta_j)\beta_i e_i + a\prod_{j=0}^k(1-\beta_j) \to 0$ as $k \to \infty$.*

By Lemma C.6 and the fact that $\sum_{k=0}^\infty \bar{\alpha}_k = \infty$, we know $\boldsymbol{\lambda}_k \to \boldsymbol{\lambda}^\star$ almost surely. Following

the same derivations, we obtain $\boldsymbol{\mu}_k \to \boldsymbol{\mu}^\star$. Now, we consider the convergence of $\bar{\boldsymbol{B}}_k$. By the construction of $\boldsymbol{\Delta}_k$, the facts that $\mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{\ell}}$ and $\mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{u}}$, and the conditions of LICQ and SOSC at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$, it suffices to show $\bar{\boldsymbol{Q}}_k \to \nabla^2 f^\star$ almost surely. This is an analogy of Lemma C.5 and we summarize it to the following lemma. See Appendix D.8 for the proof.

Lemma C.7. *Under Assumptions 4.1 and 4.3, let $\gamma_k = \iota_3(k+1)^{-b_3}$ satisfy $\iota_3 > 0, b_3 \in (0.5, 1]$. We have $\bar{\boldsymbol{Q}}_k \to \nabla^2 f^\star$ as $k \to \infty$ almost surely.*

The above lemma directly implies $\bar{\boldsymbol{B}}_k \to \nabla^2_{\boldsymbol{x}} \mathcal{L}^\star$ and we complete the proof.

C.8 Proof of Theorem 4.6

We introduce some notation first. We let

$$\boldsymbol{H}_k = \begin{pmatrix} \bar{\boldsymbol{B}}_k & \widetilde{\boldsymbol{J}}_k^\top \\ \widetilde{\boldsymbol{J}}_k & \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{B}}_k & \nabla \boldsymbol{c}_k^\top & -\boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{\ell}}} & \boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{u}}} \\ \nabla \boldsymbol{c}_k & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{I}_{\mathcal{A}^\star_{\boldsymbol{\ell}}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_{\mathcal{A}^\star_{\boldsymbol{u}}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \quad \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k = \begin{pmatrix} \bar{\boldsymbol{g}}_k + \nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda}_k - \boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{\ell}}} [\boldsymbol{\mu}_{1,k}]_{\mathcal{A}^\star_{\boldsymbol{\ell}}} + \boldsymbol{I}^\top_{\mathcal{A}^\star_{\boldsymbol{u}}} [\boldsymbol{\mu}_{2,k}]_{\mathcal{A}^\star_{\boldsymbol{u}}} \\ \boldsymbol{c}_k \\ [\boldsymbol{\ell} - \boldsymbol{x}_k]_{\mathcal{A}^\star_{\boldsymbol{\ell}}} \\ [\boldsymbol{x}_k - \boldsymbol{u}]_{\mathcal{A}^\star_{\boldsymbol{u}}} \end{pmatrix},$$

and $\nabla_{\boldsymbol{w}} \mathcal{L}_k$ is defined in the same way as $\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k$ except that $\bar{\boldsymbol{g}}_k$ is replaced by the true gradient ∇f_k. We further define the step direction

$$\bar{\Delta} \boldsymbol{w}_k = \begin{pmatrix} \bar{\Delta} \boldsymbol{x}_k \\ \bar{\Delta} \boldsymbol{\lambda}_k \\ [\bar{\Delta} \boldsymbol{\mu}_{1,k}]_{\mathcal{A}^\star_{\boldsymbol{\ell}}} \\ [\bar{\Delta} \boldsymbol{\mu}_{2,k}]_{\mathcal{A}^\star_{\boldsymbol{u}}} \end{pmatrix} = -\boldsymbol{H}_k^{-1} \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k \qquad \text{and} \qquad [\bar{\Delta} \boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-} = -[\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}. \tag{C.18}$$

By the KKT conditions of the SSQP subproblem (C.15), the facts from Lemma 4.2 that $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta} \boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{\ell}}$ and $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta} \boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{u}}$, and using the relation of $(\bar{\boldsymbol{\lambda}}^{\text{sub}}_k, \bar{\boldsymbol{\mu}}^{\text{sub}}_k) = (\boldsymbol{\lambda}_k + \bar{\Delta} \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k + \bar{\Delta} \boldsymbol{\mu}_k)$, we know the update of the iterates when k is large enough can be written by utilizing (C.18) as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \bar{\alpha}_k \bar{\Delta} \boldsymbol{w}_k \qquad \text{and} \qquad [\boldsymbol{\mu}_{k+1}]_{(\mathcal{A}^\star)^-} = (1 - \bar{\alpha}_k) \cdot [\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}. \tag{C.19}$$

Let $\Upsilon > 0$ be a constant and we propose the following conditions:

(a) $\max\{\beta_k, \bar{\alpha}_k\} \le 1$; $\|\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^\star)^{-1}\| \le \Upsilon \|\boldsymbol{H}_k - \boldsymbol{H}^\star\|$; $\|\boldsymbol{H}_k^{-1}\| \le \Upsilon$.

(b) $\max\{\|\nabla f_k - \nabla f^\star\|, \|\nabla^2 f_k - \nabla^2 f^\star\|\} \le \Upsilon \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|$; $\|\nabla f_k - \nabla f_{k-1}\| \le \Upsilon \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|$.

(c) $\max\{\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|, \|\nabla f_k - \nabla f^\star\|\} \le \frac{1}{\Upsilon}$; $\max\{\|\nabla_{\boldsymbol{w}} \mathcal{L}_k\|, \|\nabla^2_{\boldsymbol{w}} \mathcal{L}_k - \nabla^2_{\boldsymbol{w}} \mathcal{L}^\star\|\} \le \Upsilon \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|$.

(d) $\|\nabla_{\boldsymbol{w}}\mathcal{L}_k - \boldsymbol{H}^\star(\boldsymbol{w}_k - \boldsymbol{w}^\star)\| \le \Upsilon\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2$; $\|\boldsymbol{w}_k - \boldsymbol{w}^\star - \boldsymbol{H}_k^{-1}\nabla_{\boldsymbol{w}}\mathcal{L}_k\| \le \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|/\Upsilon$.

(e) $\mathcal{A}_{\boldsymbol{\ell}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{\ell},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{\ell}}$; $\mathcal{A}_{\boldsymbol{u}}(\boldsymbol{x}_k + \bar{\Delta}\boldsymbol{x}_k) = \mathcal{A}_{\boldsymbol{u},k}(\epsilon) = \mathcal{A}^\star_{\boldsymbol{u}}$.

(f) The subproblem solution $(\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_k^{\text{sub}}) = (\bar{\Delta}\boldsymbol{x}_k, \boldsymbol{\lambda}_k + \bar{\Delta}\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k + \bar{\Delta}\boldsymbol{\mu}_k)$ is from (C.18).

For any $k_0 \ge 0$, we also define the stopping time

$$\tau_{k_0} := \inf_j\{j \ge k_0 : \text{any of the above conditions does not hold at } j\text{-th iteration}\}. \quad \text{(C.20)}$$

By Lemmas 4.2, 4.4, Assumption 4.1, and choosing Υ to be large enough, we know for any run of the method, there exists a (potentially random) $\widetilde{k}_0 < \infty$ such that $\tau_{k_0} = \infty$, $\forall k_0 \ge \widetilde{k}_0$.

With the above preparation definitions, we first focus on the inactive dual components. For any run of the method, we know from (C.19) that there exists a (potentially random) $\widetilde{k}_0 < \infty$ such that for all $k \ge \widetilde{k}_0$,

$$\begin{aligned}\|[\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}\| &= \prod_{j=\widetilde{k}_0}^{k-1}(1-\bar{\alpha}_j)\cdot\|[\boldsymbol{\mu}_{\widetilde{k}_0}]_{(\mathcal{A}^\star)^-}\| \le \prod_{j=0}^{\widetilde{k}_0-1}(1-\alpha_j)^{-1}\|[\boldsymbol{\mu}_{\widetilde{k}_0}]_{(\mathcal{A}^\star)^-}\|\cdot\prod_{j=0}^{k-1}(1-\alpha_j)\\ &\le \prod_{j=0}^{\widetilde{k}_0-1}(1-\alpha_j)^{-1}\|[\boldsymbol{\mu}_{\widetilde{k}_0}]_{(\mathcal{A}^\star)^-}\|\exp\left(-\sum_{j=0}^{k-1}\alpha_j\right).\end{aligned}$$

Note that

$$\sum_{i=0}^{k-1}\alpha_k \ge \iota_1\int_1^{k+1}x^{-b_1}dx = \begin{cases}\frac{\iota_1}{1-b_1}\left((k+1)^{1-b_1}-1\right) & \text{if } b_1 < 1,\\ \iota_1\log(k+1) & \text{if } b_1 = 1.\end{cases} \quad \text{(C.21)}$$

Combining the above two displays, we know

$$\|[\boldsymbol{\mu}_k]_{(\mathcal{A}^\star)^-}\| = \begin{cases}o(k^{-b}) \text{ for any } b>0 & \text{if } b_1 < 1,\\ \mathcal{O}(k^{-\iota_1}) & \text{if } b_1 = 1.\end{cases}$$

We now focus on the primal and active dual components. For notational simplicity, we denote $\bar{\Delta}\boldsymbol{w}_k = (\bar{\Delta}\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}_k, [\bar{\boldsymbol{\mu}}_k^{\text{sub}} - \boldsymbol{\mu}_k]_{\mathcal{A}^\star})$ for any $k \ge 0$, while noting from (C.18) that we only have $\bar{\Delta}\boldsymbol{w}_k = -\boldsymbol{H}_k^{-1}\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k$ for $k_0 \le k < \tau_{k_0}$ with any given $k_0 \ge 0$. For any $k \ge 0$, we have

$$\begin{aligned}&\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\\ &= \boldsymbol{w}_k - \boldsymbol{w}^\star + \bar{\alpha}_k\bar{\Delta}\boldsymbol{w}_k = \boldsymbol{w}_k - \boldsymbol{w}^\star + \alpha_k\bar{\Delta}\boldsymbol{w}_k + (\bar{\alpha}_k - \alpha_k)\bar{\Delta}\boldsymbol{w}_k\\ &= \boldsymbol{w}_k - \boldsymbol{w}^\star - \alpha_k(\boldsymbol{H}^\star)^{-1}\nabla_{\boldsymbol{w}}\mathcal{L}_k + \alpha_k(\boldsymbol{H}^\star)^{-1}(\nabla_{\boldsymbol{w}}\mathcal{L}_k - \bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k) + \alpha_k((\boldsymbol{H}^\star)^{-1} - \boldsymbol{H}_k^\dagger)\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k\end{aligned}$$

$$
\begin{aligned}
&+ \alpha_k(\boldsymbol{H}_k^\dagger \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k + \bar{\Delta}\boldsymbol{w}_k) + (\bar{\alpha}_k - \alpha_k)\bar{\Delta}\boldsymbol{w}_k \\
&= \boldsymbol{w}_k - \boldsymbol{w}^\star - \alpha_k \boldsymbol{H}_k^\dagger \nabla_{\boldsymbol{w}} \mathcal{L}_k + \alpha_k (\boldsymbol{H}^\star)^{-1}(\nabla_{\boldsymbol{w}} \mathcal{L}_k - \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k) + \alpha_k((\boldsymbol{H}^\star)^{-1} - \boldsymbol{H}_k^\dagger)(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k) \\
&\quad + \alpha_k(\boldsymbol{H}_k^\dagger \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k + \bar{\Delta}\boldsymbol{w}_k) + (\bar{\alpha}_k - \alpha_k)\bar{\Delta}\boldsymbol{w}_k \\
&= (1-\alpha_k)(\boldsymbol{w}_k - \boldsymbol{w}^\star) - \alpha_k \boldsymbol{H}_k^\dagger (\nabla_{\boldsymbol{w}} \mathcal{L}_k - \boldsymbol{H}^\star(\boldsymbol{w}_k - \boldsymbol{w}^\star)) - \alpha_k(\boldsymbol{H}_k^\dagger \boldsymbol{H}^\star - \boldsymbol{I})(\boldsymbol{w}_k - \boldsymbol{w}^\star) \\
&\quad - \alpha_k(\boldsymbol{H}_k^\dagger - (\boldsymbol{H}^\star)^{-1})(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k) + \alpha_k(\boldsymbol{H}_k^\dagger \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k + \bar{\Delta}\boldsymbol{w}_k) + (\bar{\alpha}_k - \alpha_k)\bar{\Delta}\boldsymbol{w}_k \\
&\quad - \alpha_k(\boldsymbol{H}^\star)^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k).
\end{aligned}
$$

Applying the above recursion from 0 to k, we have

$$
\begin{aligned}
&\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star - \prod_{i=0}^{k}(1-\alpha_i)(\boldsymbol{w}_0 - \boldsymbol{w}^\star) \\
&= -\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \Big\{ \boldsymbol{H}_i^\dagger(\nabla_{\boldsymbol{w}} \mathcal{L}_i - \boldsymbol{H}^\star(\boldsymbol{w}_i - \boldsymbol{w}^\star)) + (\boldsymbol{H}_i^\dagger \boldsymbol{H}^\star - \boldsymbol{I})(\boldsymbol{w}_i - \boldsymbol{w}^\star) \Big\} \\
&\quad - \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \Big\{ (\boldsymbol{H}_i^\dagger - (\boldsymbol{H}^\star)^{-1})(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_i - \nabla_{\boldsymbol{w}} \mathcal{L}_i) - (\boldsymbol{H}_i^\dagger \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_i + \bar{\Delta}\boldsymbol{w}_i) - \frac{\bar{\alpha}_i - \alpha_i}{\alpha_i}\bar{\Delta}\boldsymbol{w}_i \Big\} \\
&\quad - \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i(\boldsymbol{H}^\star)^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_i - \nabla_{\boldsymbol{w}} \mathcal{L}_i) =: -\mathcal{J}_{1,k} - \mathcal{J}_{2,k} - \mathcal{J}_{3,k}. \qquad \text{(C.22)}
\end{aligned}
$$

First, for the term $\prod_{i=0}^{k}(1-\alpha_i)(\boldsymbol{w}_0 - \boldsymbol{w}^\star)$ on the left, we note from (C.21) that $\prod_{i=0}^{k}(1-\alpha_i)(\boldsymbol{w}_0 - \boldsymbol{w}^\star) = o(\sqrt{\alpha_k})$ as long as $\iota_1 > 0.5$ when $b_1 = 1$. Next, we aim to show $\mathcal{J}_{1,k} = o_p(\sqrt{\alpha_k})$, $\mathcal{J}_{2,k} = o_p(\sqrt{\alpha_k})$, and asymptotic normality of $\mathcal{J}_{3,k}$. Throughout the proof, we use $o_p(\cdot)$ and $\mathcal{O}_p(\cdot)$ to denote the standard small- and big-O notation in the probability sense. We state the following preparation lemma.

Lemma C.8 (Lemma C.3 in Na, 2025). *Consider a sequence of random variables $\{X_k\}_{k=0}^{\infty}$ and a sequence of events $\{\mathcal{A}_k\}_{k=0}^{\infty}$. Let $\tau_{k_0} = \inf\{k \geq k_0 : \mathcal{A}_k \text{ happens}\}$ be the first index k after k_0 such that $\mathcal{A}_k$ happens. Suppose that for each realization of the sequence, there exists a (potentially random) $\widetilde{k}_0 < \infty$ such that $\tau_{\widetilde{k}_0} = \infty$. Also, for the sequence $\alpha_k = \iota_1(k+1)^{-b_1}$ with $b_1 \in (0, 1]$ and $\iota_1 > 0.5\mathbf{1}_{b_1=1}$, suppose there exists a deterministic $\bar{k}_0 > 0$ such that for any fixed $k_0 \geq \bar{k}_0$,*

$X_k \mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k})$. *Then, we have*

$$\sum_{i=0}^{k} \prod_{j=i+1}^{k} (1-\alpha_j)\alpha_i X_i = o_p(\sqrt{\alpha_k}).$$

By the above lemma and (C.22), we know $\mathcal{J}_{1,k} = o_p(\sqrt{\alpha_k})$ and $\mathcal{J}_{2,k} = o_p(\sqrt{\alpha_k})$ as long as

$$\|\boldsymbol{H}_k^\dagger(\nabla_{\boldsymbol{w}}\mathcal{L}_k - \boldsymbol{H}^\star(\boldsymbol{w}_k - \boldsymbol{w}^\star))\|\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k}), \tag{C.23a}$$

$$\|(\boldsymbol{H}_k^\dagger \boldsymbol{H}^\star - \boldsymbol{I})(\boldsymbol{w}_k - \boldsymbol{w}^\star)\|\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k}), \tag{C.23b}$$

$$\|(\boldsymbol{H}_k^\dagger - (\boldsymbol{H}^\star)^{-1})(\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k - \nabla_{\boldsymbol{w}}\mathcal{L}_k)\|\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k}), \tag{C.23c}$$

$$\|(\boldsymbol{H}_k^\dagger \bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k + \bar{\Delta}\boldsymbol{w}_k)\|\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k}), \tag{C.23d}$$

$$\frac{\bar{\alpha}_k - \alpha_k}{\alpha_k}\|\bar{\Delta}\boldsymbol{w}_k\|\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k}). \tag{C.23e}$$

For (C.23a), we apply the definition of τ_{k_0} in (C.20) and have

$$\|\boldsymbol{H}_k^\dagger(\nabla_{\boldsymbol{w}}\mathcal{L}_k - \boldsymbol{H}^\star(\boldsymbol{w}_k - \boldsymbol{w}^\star))\|\mathbf{1}_{\tau_{k_0}>k} \le \Upsilon^2\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2\mathbf{1}_{\tau_{k_0}>k}.$$

For (C.23b), we have

$$\begin{aligned}\|(\boldsymbol{H}_k^\dagger \boldsymbol{H}^\star - \boldsymbol{I})(\boldsymbol{w}_k - \boldsymbol{w}^\star)\|\mathbf{1}_{\tau_{k_0}>k} &\overset{(C.20)}{\le} \Upsilon\|\boldsymbol{H}_k - \boldsymbol{H}^\star\|\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|\mathbf{1}_{\tau_{k_0}>k} \\ &\le 0.5\Upsilon\|\boldsymbol{H}_k - \boldsymbol{H}^\star\|^2\mathbf{1}_{\tau_{k_0}>k} + 0.5\Upsilon\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2\mathbf{1}_{\tau_{k_0}>k}.\end{aligned}$$

For (C.23c), we have

$$\begin{aligned}\|(\boldsymbol{H}_k^\dagger - (\boldsymbol{H}^\star)^{-1})(\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k - \nabla_{\boldsymbol{w}}\mathcal{L}_k)\|\mathbf{1}_{\tau_{k_0}>k} &\overset{(C.20)}{\le} \Upsilon\|\boldsymbol{H}_k - \boldsymbol{H}^\star\|\|\bar{\boldsymbol{g}}_k - \nabla f_k\|\mathbf{1}_{\tau_{k_0}>k} \\ &\le 0.5\Upsilon\|\boldsymbol{H}_k - \boldsymbol{H}^\star\|^2\mathbf{1}_{\tau_{k_0}>k} + 0.5\Upsilon\|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2\mathbf{1}_{\tau_{k_0}>k}.\end{aligned}$$

For (C.23d), it is trivial due to (C.18). For (C.23e), we have

$$\begin{aligned}\frac{\bar{\alpha}_k - \alpha_k}{\alpha_k}\|\bar{\Delta}\boldsymbol{w}_k\|\mathbf{1}_{\tau_{k_0}>k} \le \psi\alpha_k^{p-1}\|\bar{\Delta}\boldsymbol{w}_k\|\mathbf{1}_{\tau_{k_0}>k} &\overset{(C.18)}{\le} \Upsilon\psi\alpha_k^{p-1}\|\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k\|\mathbf{1}_{\tau_{k_0}>k} \\ &\overset{(C.20)}{\le} \Upsilon\psi\alpha_k^{p-1}\|\bar{\boldsymbol{g}}_k - \nabla f_k\|\mathbf{1}_{\tau_{k_0}>k} + \Upsilon^2\psi\alpha_k^{p-1}\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|\mathbf{1}_{\tau_{k_0}>k}.\end{aligned}$$

Combining the above four displays, we know (C.23) is implied by showing that

$$\begin{aligned}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 + \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 + \|\boldsymbol{H}_k - \boldsymbol{H}^\star\|^2\right)\mathbf{1}_{\tau_{k_0}>k} &= o_p(\sqrt{\alpha_k}), \\ \alpha_k^{p-1}\left(\|\boldsymbol{w}_k - \boldsymbol{w}^\star\| + \|\bar{\boldsymbol{g}}_k - \nabla f_k\|\right)\mathbf{1}_{\tau_{k_0}>k} &= o_p(\sqrt{\alpha_k}).\end{aligned} \tag{C.24}$$

We present the following two lemmas to corroborate the above result (C.24). See Appendices

D.9 and D.11 for the proofs, respectively.

Lemma C.9. *Under Assumption 3.6 and suppose* $\alpha_k = \iota_1(k+1)^{-b_1}$, $\beta_k = \iota_2(k+1)^{-b_2}$ *satisfy* $\iota_1, \iota_2 > 0$, $0 < b_2 < b_1 \leq 1$, *and* $\iota_1 > 0.5$ *if* $b_1 = 1$. *There exists a deterministic integer* $\bar{k}_0 > 0$ *such that for any* $k_0 \geq \bar{k}_0$, *the following result holds for some deterministic constant* $\Upsilon(k_0) > 0$*:*

$$\mathbb{E}\big[(\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 + \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2)\mathbf{1}_{\tau_{k_0} > k}\big] \leq \Upsilon(k_0)\beta_k, \qquad \forall k \geq k_0.$$

Lemma C.10. *Under Assumptions 3.6, 4.3 and suppose* $\alpha_k = \iota_1(k+1)^{-b_1}$, $\beta_k = \iota_2(k+1)^{-b_2}$, $\gamma_k = \iota_3(k+1)^{-b_3}$ *satisfy* $\iota_1, \iota_2, \iota_3 > 0$, $b_1 \in (0, 1]$, $b_2 \in (0.5b_1, b_1)$, $b_3 \in (0.5b_1, 1]$, *and* $\iota_1 > 0.5$ *if* $b_1 = 1$ *and* $\iota_3 > 0.25b_1$ *if* $b_3 = 1$. *There exists a deterministic integer* $\bar{k}_0 > 0$ *such that for any* $k_0 \geq \bar{k}_0$, *the following result holds for some deterministic constant* $\Upsilon(k_0) > 0$*:*

$$\mathbb{E}[\|\boldsymbol{H}_k - \boldsymbol{H}^\star\|^2 \mathbf{1}_{\tau_{k_0} > k}]/(\Upsilon(k_0)\sqrt{\alpha_k}) \to 0 \quad \text{as} \quad k \to \infty.$$

With Lemmas C.9 and C.10, and applying the conditions that $b_2 > 0.5 \geq 0.5b_1$ and $p > 1.5 - 0.5b_2/b_1$, we know (C.24) holds, leading to $\mathcal{J}_{1,k} = o_p(\sqrt{\alpha_k})$ and $\mathcal{J}_{2,k} = o_p(\sqrt{\alpha_k})$. Finally, we present the following lemma demonstrating the asymptotic normality property of $\mathcal{J}_{3,k}$. See Appendix D.12 for the proof.

Lemma C.11. *Under Assumptions 4.1, 4.5 and suppose* $\alpha_k = \iota_1(k+1)^{-b_1}$, $\beta_k = \iota_2(k+1)^{-b_2}$ *satisfy* $\iota_1, \iota_2 > 0$, $b_1 \in ((2-2\delta)/(2+\delta), 1]$, $b_2 \in (0, b_1)$, *and* $\iota_1 > 2/3$ *if* $b_1 = 1$. *Then, we have* $1/\sqrt{\alpha_k} \cdot \mathcal{J}_{3,k} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \eta \cdot \boldsymbol{\Omega}^\star)$, *where* η *and* $\boldsymbol{\Omega}^\star$ *are defined in Theorem 4.6.*

Noting that $\alpha_k/\bar{\alpha}_k \to 1$ almost surely since $p > 1.5 - 0.5b_2/b_1$ and $b_1 > b_2 > 0$ imply $p > 1$, we combine Lemma C.11 with (C.22) and complete the proof.

C.9 Proof of Theorem 4.8

By Assumption 4.1 and Lemma 4.4, it suffices to show the convergence of $\mathrm{Cov}(\{\nabla F(\boldsymbol{x}_i; \zeta_i)\}_{i=0}^k)$. In particular, we have

$$\big\|\mathrm{Cov}(\{\nabla F(\boldsymbol{x}_i; \zeta_i)\}_{i=0}^k) - \mathrm{Cov}(\nabla F(\boldsymbol{x}^\star; \zeta))\big\|$$

$$\leq \left\| \frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) \nabla F(\boldsymbol{x}_i;\zeta_i)^\top - \mathbb{E}\left[\nabla F(\boldsymbol{x}^\star;\zeta) \nabla F(\boldsymbol{x}^\star;\zeta)^\top\right] \right\|$$
$$+ \left\| \left(\frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) \right) \left(\frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) \right)^\top - \nabla f^\star \nabla f^{\star\top} \right\|. \quad \text{(C.25)}$$

We analyze the two terms on the right hand side separately. For the first term in (C.25), we have

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) \nabla F(\boldsymbol{x}_i;\zeta_i)^\top - \mathbb{E}\left[\nabla F(\boldsymbol{x}^\star;\zeta) \nabla F(\boldsymbol{x}^\star;\zeta)^\top\right] \right\|$$
$$\leq \left\| \frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) \nabla F(\boldsymbol{x}_i;\zeta_i)^\top - \mathbb{E}_i\left[\nabla F(\boldsymbol{x}_i;\zeta_i) \nabla F(\boldsymbol{x}_i;\zeta_i)^\top\right] \right\|$$
$$+ \left\| \frac{1}{k+1} \sum_{i=0}^{k} \mathbb{E}_i\left[\nabla F(\boldsymbol{x}_i;\zeta_i) \nabla F(\boldsymbol{x}_i;\zeta_i)^\top\right] - \mathbb{E}\left[\nabla F(\boldsymbol{x}^\star;\zeta) \nabla F(\boldsymbol{x}^\star;\zeta)^\top\right] \right\|.$$

The first term converges to zero almost surely by bounded 4-th moment condition of $\nabla F(\boldsymbol{x}_k;\zeta_k)$ and the strong law of large numbers for square integrable martingale (Duflo, 1997, Theorem 1.3.15); the second term converges to zero almost surely by Assumptions 4.1, 4.5, and the fact that $x_k \to 0 \Rightarrow \frac{1}{k+1}\sum_{i=0}^{k} x_i \to 0$. Similarly, for the second term in (C.25), we just note that

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} \nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f^\star \right\| \leq \left\| \frac{1}{k+1} \sum_{i=0}^{k} (\nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f_i) \right\| + \left\| \frac{1}{k+1} \sum_{i=0}^{k} (\nabla f_i - \nabla f^\star) \right\|,$$

and it converges to zero almost surely due to the same reasons as above. This completes the proof.

D Proofs of Supporting Lemmas

D.1 Proof of Lemma C.1

We adapt the proof from (Davis et al., 2024, Lemma B.1). Consider the normalizing constant $C(\boldsymbol{v}) = 1 + \int h(\boldsymbol{v}^\top \boldsymbol{g}(\zeta)) d\mathcal{P}(\zeta)$. Since h is C^3-smooth with bounded derivatives and $\boldsymbol{g} \in \mathcal{G}$ satisfies $\mathbb{E}\|\boldsymbol{g}(\zeta)\|^2 < \infty$, the dominated convergence theorem ensures $C(\boldsymbol{v})$ is twice continuously differentiable with $\nabla C(\boldsymbol{v}) = \int h'(\boldsymbol{v}^\top \boldsymbol{g}(\zeta)) \boldsymbol{g}(\zeta) d\mathcal{P}(\zeta)$, $\nabla^2 C(\boldsymbol{v}) = \int h''(\boldsymbol{v}^\top \boldsymbol{g}(\zeta)) \boldsymbol{g}(\zeta) \boldsymbol{g}(\zeta)^\top d\mathcal{P}(\zeta)$. Similarly, $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ is C^1-smooth with $\nabla^2_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = \mathbb{E}_{\zeta\sim\mathcal{P}}[\nabla^2_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w};\zeta)]$. Define $\widetilde{\nabla}_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}) = \int h(\boldsymbol{v}^\top \boldsymbol{g}(\zeta)) \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w};\zeta) d\mathcal{P}(\zeta)$. By the dominated convergence theorem, $\widetilde{\nabla}_{\boldsymbol{w}} \mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})$ is contin-

uously differentiable in both $\boldsymbol{v}$ and $\boldsymbol{w}$, and its partial derivatives are given by

$$\begin{aligned}\nabla_{\boldsymbol{v}}(\widetilde{\nabla}_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})) &= \int h'(\boldsymbol{v}^\top \boldsymbol{g}(\zeta))\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w};\zeta)\boldsymbol{g}(\zeta)^\top d\mathcal{P}(\zeta),\\ \nabla_{\boldsymbol{w}}(\widetilde{\nabla}_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})) &= \int h(\boldsymbol{v}^\top \boldsymbol{g}(\zeta))\nabla^2_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w};\zeta)d\mathcal{P}(\zeta),\end{aligned} \tag{D.1}$$

both of which are continuous in $(\boldsymbol{w},\boldsymbol{v})$. Thus, in a small neighborhood of $(\boldsymbol{w},\boldsymbol{v})=(\boldsymbol{w}^\star,\boldsymbol{0})$, we apply $h(t)=t$ and $\mathbb{E}_{\zeta\sim\mathcal{P}}[\boldsymbol{g}(\zeta)]=\boldsymbol{0}$, and know $\nabla_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})=\widetilde{\nabla}_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})/C(\boldsymbol{v})$ is C^1-smooth. Finally, the first identity in (C.5) directly follows from the fact that $\nabla_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})|_{\boldsymbol{v}=\boldsymbol{0}}=\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w})$ for any $\boldsymbol{w}$. For the second identity, we note that $\nabla C(\boldsymbol{0})=\boldsymbol{0}$ and $\nabla^2 C(\boldsymbol{0})=\boldsymbol{0}$, implying $C(\boldsymbol{v})=1+o(\|\boldsymbol{v}\|^2)$. Thus, $\nabla_{\boldsymbol{v}\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w})|_{\boldsymbol{v}=\boldsymbol{0}}=\nabla_{\boldsymbol{v}}(\widetilde{\nabla}_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}))|_{\boldsymbol{v}=\boldsymbol{0}}\overset{(D.1)}{=}\mathbb{E}_{\zeta\sim\mathcal{P}}[\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w};\zeta)\boldsymbol{g}(\zeta)^\top]$. This completes the proof.

D.2 Proof of Lemma C.2

We adapt the proof from (Davis et al., 2024, Lemma B.2). Define the linearized mapping $\Psi(\boldsymbol{w})=\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star)+\nabla^2_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star)(\boldsymbol{w}-\boldsymbol{w}^\star)=\boldsymbol{H}^\star(\boldsymbol{w}-\boldsymbol{w}^\star)$, which is invertible and is C^1-smooth. By Lemma C.1, we can equivalently express $\Psi(\boldsymbol{w})=\nabla_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}^\star)|_{\boldsymbol{v}=\boldsymbol{0}}+\nabla^2_{\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}^\star)|_{\boldsymbol{v}=\boldsymbol{0}}(\boldsymbol{w}-\boldsymbol{w}^\star)$. By the implicit function theorem (Dontchev and Rockafellar, 2009, Theorem 2D.6), the mapping $S(\boldsymbol{v})$ admits a single-valued localization $s(\boldsymbol{v})$ around $\boldsymbol{w}^\star$ and $\boldsymbol{v}=\boldsymbol{0}$ with $\nabla s(\boldsymbol{0})=-\nabla\sigma(\boldsymbol{0})\cdot\nabla_{\boldsymbol{v}\boldsymbol{w}}\mathcal{L}_{\boldsymbol{v}}(\boldsymbol{w}^\star)|_{\boldsymbol{v}=\boldsymbol{0}}=-(\boldsymbol{H}^\star)^{-1}\cdot\mathbb{E}_{\zeta\sim\mathcal{P}}\left[\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}^\star;\zeta)\boldsymbol{g}(\zeta)^\top\right]$. This completes the proof.

D.3 Proof of Lemma C.4

We present two preparation lemmas that are proved in Appendices D.4 and D.5, respectively.

Lemma D.1. *Let $\alpha_k=\iota_1(k+1)^{-b_1}$, $\beta_k=\iota_2(k+1)^{-b_2}$ for $\iota_1,\iota_2,b_1,b_2>0$ with $b_2<\min\{b_1,1\}$. Under Assumption 3.6 and the Lipschitz continuity of ∇f (cf. Assumption 3.5), we have*

$$\mathbb{E}[\|\bar{\boldsymbol{g}}_k-\nabla f_k\|^2]=\mathcal{O}\left(\beta_k+\alpha_k^2/\beta_k^2\right).$$

Lemma D.2. *Suppose $\boldsymbol{B}_k\succeq\kappa_1\boldsymbol{I}$, then*

$$\left\|\bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k\right\|\leq\kappa_1^{-1}\left\|\bar{\boldsymbol{g}}_k-\nabla f_k\right\|,$$

where $\bar{\Delta}\boldsymbol{x}_k$ is the solution of (17) and $\Delta\boldsymbol{x}_k$ is the solution with $\bar{\boldsymbol{g}}_k$ replaced by ∇f_k.

Now, we apply the results in Lemmas D.1 and D.2 and obtain

$$\mathbb{E}[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] \le \kappa_1^{-1}\mathbb{E}[\|\bar{\boldsymbol{g}}_k - \nabla f_k\|] \le \kappa_1^{-1}\left\{\mathbb{E}[\|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2]\right\}^{1/2} = \mathcal{O}\left(\sqrt{\beta_k} + \alpha_k/\beta_k\right).$$

By the setup of α_k, β_k, it follows that $\sum_{k=0}^{\infty}\alpha_k\mathbb{E}[\|\bar{\Delta}\boldsymbol{x}_k - \Delta\boldsymbol{x}_k\|] < \infty$ if $b_1 + 0.5b_2 > 1$ and $2b_1 - b_2 > 1$. Equivalently, this requires $b_1 \in (0.75, 1]$ and $b_2 \in (2 - 2b_1, 2b_1 - 1)$. We complete the proof.

D.4 Proof of Lemma D.1

From the update rule of $\bar{\boldsymbol{g}}_k$ in (15), we have

$$\begin{aligned}
\bar{\boldsymbol{g}}_k - \nabla f_k &= \beta_k(\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k) + (1-\beta_k)\left(\bar{\boldsymbol{g}}_{k-1} - \nabla f_{k-1}\right) + (1-\beta_k)\left(\nabla f_{k-1} - \nabla f_k\right) \\
&= \beta_k(\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k) + (1-\beta_k)\left\{\beta_{k-1}(\nabla F(\boldsymbol{x}_{k-1};\zeta_{k-1}) - \nabla f_{k-1})\right. \\
&\quad \left. +(1-\beta_{k-1})\left(\bar{\boldsymbol{g}}_{k-2} - \nabla f_{k-2}\right) + (1-\beta_{k-1})\left(\nabla f_{k-2} - \nabla f_{k-1}\right)\right\} + (1-\beta_k)\left(\nabla f_{k-1} - \nabla f_k\right) \\
&= \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)\beta_i\left(\nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f_i\right) + \sum_{i=0}^{k}\left(\prod_{j=i}^{k}(1-\beta_j)\right)\left(\nabla f_{i-1} - \nabla f_i\right) \\
&:= \mathcal{W}_{1,k} + \mathcal{W}_{2,k}, \qquad \text{(D.2)}
\end{aligned}$$

where we denote $\nabla f_{-1} = \bar{\boldsymbol{g}}_{-1}$ in the second last equality. By Lemma B.4 and noting that $\alpha_k \le \bar{\alpha}_k \le \alpha_k + \psi\alpha_k^p$, we have

$$\|\mathcal{W}_{2,k}\| \le \prod_{j=0}^{k}(1-\beta_j)(\|\bar{\boldsymbol{g}}_{-1}\| + \|\nabla f_0\|) + \sum_{i=1}^{k}\left(\prod_{j=i}^{k}(1-\beta_j)\right)\bar{\alpha}_{i-1}M_{\boldsymbol{\ell},\boldsymbol{u}} = \mathcal{O}(\alpha_k/\beta_k).$$

Moreover, by Assumption 3.6 and Lemma B.4, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{W}_{1,k}\|^2] &= \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)^2\beta_i^2\mathbb{E}[\|\nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f_i\|^2] \\
&\le \sigma_g^2\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)^2\beta_i^2 = \mathcal{O}\left(\beta_k\right).
\end{aligned}$$

Combining the above three displays together, we complete the proof.

D.5 Proof of Lemma D.2

The subproblem (17) at $\boldsymbol{x}_k$ with the averaged gradient $\bar{\boldsymbol{g}}_k$ can be written as

$$\min_{\bar{\Delta}\boldsymbol{x}\in\Omega(\boldsymbol{x}_k;\theta_k)} \frac{1}{2}\left\|\bar{\Delta}\boldsymbol{x}+\boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k\right\|_{\boldsymbol{B}_k}^2,$$

where $\|\boldsymbol{x}\|_{\boldsymbol{B}_k}^2=\boldsymbol{x}^\top\boldsymbol{B}_k\boldsymbol{x}$. By the optimality condition using variational inequalities, we have

$$\left\langle \Delta\boldsymbol{x}_k-\bar{\Delta}\boldsymbol{x}_k, -\boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k-\bar{\Delta}\boldsymbol{x}_k\right\rangle_{\boldsymbol{B}_k}\le 0.$$

Similarly, since $\Delta\boldsymbol{x}_k$ is the solution with true gradient ∇f_k, it satisfies

$$\left\langle \bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k, -\boldsymbol{B}_k^{-1}\nabla f_k-\Delta\boldsymbol{x}_k\right\rangle_{\boldsymbol{B}_k}\le 0.$$

Adding the two inequalities yields

$$\begin{aligned}
0 &\ge \left\langle \Delta\boldsymbol{x}_k-\bar{\Delta}\boldsymbol{x}_k, -\boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k-\bar{\Delta}\boldsymbol{x}_k+\boldsymbol{B}_k^{-1}\nabla f_k+\Delta\boldsymbol{x}_k\right\rangle_{\boldsymbol{B}_k}\\
&= \left\|\Delta\boldsymbol{x}_k-\bar{\Delta}\boldsymbol{x}_k\right\|_{\boldsymbol{B}_k}^2+\left\langle \bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k, \bar{\boldsymbol{g}}_k-\nabla f_k\right\rangle\\
&\ge \|\bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k\|_{\boldsymbol{B}_k}^2-\|\bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k\|\cdot\|\bar{\boldsymbol{g}}_k-\nabla f_k\|.
\end{aligned}$$

Since $\boldsymbol{B}_k\succeq\kappa_1\boldsymbol{I}$, we know $\|\bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k\|_{\boldsymbol{B}_k}^2\ge\kappa_1\|\bar{\Delta}\boldsymbol{x}_k-\Delta\boldsymbol{x}_k\|^2$. This completes the proof.

D.6 Proof of Lemma C.5

We follow the proof of Lemma D.1 in Appendix D.4 and slightly abuse the notation of $\mathcal{W}_{1,k}$ and $\mathcal{W}_{2,k}$ there. In particular, we revisit the definition of $\bar{\boldsymbol{g}}_k$ and have

$$\begin{aligned}
\bar{\boldsymbol{g}}_k-\nabla f^\star &= (1-\beta_k)(\bar{\boldsymbol{g}}_{k-1}-\nabla f^\star)+\beta_k(\nabla F(\boldsymbol{x}_k;\zeta_k)-\nabla f_k)+\beta_k(\nabla f_k-\nabla f^\star)\\
&= \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)\beta_i(\nabla F(\boldsymbol{x}_i;\zeta_i)-\nabla f_i)\\
&\quad+\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_j)\right)\beta_i(\nabla f_i-\nabla f^\star)+\prod_{j=0}^{k}(1-\beta_j)(\bar{\boldsymbol{g}}_{-1}-\nabla f^\star)\\
&:= \mathcal{W}_{1,k}+\mathcal{W}_{2,k}.
\end{aligned}$$

Applying Lemma C.6 and the fact that $\nabla f_i \to \nabla f^\star$, we immediately have $\mathcal{W}_{2,k} \to \mathbf{0}$ as $k \to \infty$.

For the term $\mathcal{W}_{1,k}$, we rewrite it as

$$\mathcal{W}_{1,k} = (1-\beta_k)\mathcal{W}_{1,k-1} + \beta_k \left(\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k\right).$$

Thus, for large enough k, we have from Assumption 3.6 that

$$\begin{aligned}\mathbb{E}_k[\|\mathcal{W}_{1,k}\|^2] &= (1-\beta_k)^2\|\mathcal{W}_{1,k-1}\|^2 + \beta_k^2\mathbb{E}_k[\|\nabla F(\boldsymbol{x}_k;\zeta_k) - \nabla f_k\|^2] \\ &\le (1-\beta_k)\|\mathcal{W}_{1,k-1}\|^2 + \beta_k^2\sigma_g^2.\end{aligned}$$

Since $\sum_{k=0}^\infty \beta_k = \infty$ and $\sum_{k=0}^\infty \beta_k^2 < \infty$, by Robbins-Siegmund theorem (Robbins and Siegmund, 1971), we know $\|\mathcal{W}_{1,k}\|^2$ is convergent and $\sum_{k=0}^\infty \beta_k\|\mathcal{W}_{1,k-1}\|^2 < \infty$ almost surely. This further implies $\|\mathcal{W}_{1,k}\| \to 0$ almost surely and we complete the proof.

D.7 Proof of Lemma C.6

For any $\epsilon > 0$, there exists $i' > 0$ such that $|e_i| \le \epsilon$ and $\beta_i \in [0,1]$ for any $i \ge i'$. Thus, for $k \ge i'$,

$$\left|a\prod_{j=0}^k(1-\beta_j)\right| \le |a|\prod_{j=0}^{i'-1}|1-\beta_j| \cdot \prod_{j=i'}^k(1-\beta_j) \le |a|\prod_{j=0}^{i'-1}|1-\beta_j| \cdot \exp\left(-\sum_{j=i'}^k \beta_j\right),$$

and

$$\begin{aligned}\left|\sum_{i=0}^k\prod_{j=i+1}^k(1-\beta_j)\beta_i e_i\right| &\le \sum_{i=0}^{i'-1}\prod_{j=i+1}^k|1-\beta_j|\beta_i|e_i| + \sum_{i=i'}^k\prod_{j=i+1}^k(1-\beta_j)\beta_i|e_i| \\ &\le \prod_{j=i'}^k(1-\beta_j)\cdot\sum_{i=0}^{i'-1}\prod_{j=i+1}^{i'-1}|1-\beta_j|\beta_i|e_i| + \epsilon\sum_{i=i'}^k\prod_{j=i+1}^k(1-\beta_j)\beta_i \\ &= \prod_{j=i'}^k(1-\beta_j)\cdot\sum_{i=0}^{i'-1}\prod_{j=i+1}^{i'-1}|1-\beta_j|\beta_i|e_i| + \epsilon\left\{1-\prod_{j=i'}^k(1-\beta_j)\right\} \\ &\le \exp\left(-\sum_{j=i'}^k\beta_j\right)\cdot\sum_{i=0}^{i'-1}\prod_{j=i+1}^{i'-1}|1-\beta_j|\beta_i|e_i| + \epsilon.\end{aligned}$$

Since $\sum_{i=0}^\infty \beta_i = \infty$, we can find $k' \ge i'$ large enough such that $\exp(-\sum_{j=i'}^k \beta_j)\cdot\{|a|\prod_{j=0}^{i'-1}|1-\beta_j| + \sum_{i=0}^{i'-1}\prod_{j=i+1}^{i'-1}|1-\beta_j|\beta_i|e_i|\} \le \epsilon$ for any $k \ge k'$. Then, we obtain for $k \ge k'$ that

$$\left|a\prod_{j=0}^k(1-\beta_j)\right| + \left|\sum_{i=0}^k\prod_{j=i+1}^k(1-\beta_j)\beta_i e_i\right| \le 2\epsilon.$$

This shows $a\prod_{j=0}^k(1-\beta_j) + \sum_{i=0}^k\prod_{j=i+1}^k(1-\beta_j)\beta_i e_i \to 0$ as $k \to \infty$. We complete the proof.

D.8 Proof of Lemma C.7

We follow the proof of Lemma C.5 in Appendix D.6. By the update rule (15), we have

$$\begin{aligned}\bar{\boldsymbol{Q}}_k - \nabla^2 f^\star &= (1-\gamma_k)(\bar{\boldsymbol{Q}}_{k-1} - \nabla^2 f^\star) + \gamma_k(\nabla^2 F(\boldsymbol{x}_k;\zeta_k) - \nabla^2 f_k) + \gamma_k(\nabla^2 f_k - \nabla^2 f^\star) \\ &= \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\gamma_j)\right)\gamma_i\left(\nabla^2 F(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f_i\right) \\ &\quad + \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\gamma_j)\right)\gamma_i\left(\nabla^2 f_i - \nabla^2 f^\star\right) + \prod_{j=0}^{k}(1-\gamma_j)(\bar{\boldsymbol{Q}}_{-1} - \nabla^2 f^\star) \\ &:= \mathcal{V}_{1,k} + \mathcal{V}_{2,k}.\end{aligned}$$

Applying Lemma C.6 and the fact that $\nabla^2 f_i \to \nabla^2 f^\star$, we immediately have $\mathcal{V}_{2,k} \to \mathbf{0}$ as $k \to \infty$. For the term $\mathcal{V}_{1,k}$, we rewrite it as

$$\mathcal{V}_{1,k} = (1-\gamma_k)\mathcal{V}_{1,k-1} + \gamma_k(\nabla^2 F(\boldsymbol{x}_k;\zeta_k) - \nabla^2 f_k).$$

Thus, for large enough k, we have from Assumption 4.3 that

$$\begin{aligned}\mathbb{E}_k[\|\mathcal{V}_{1,k}\|_F^2] &= (1-\gamma_k)^2\|\mathcal{V}_{1,k-1}\|_F^2 + \gamma_k^2\mathbb{E}_k[\|\nabla^2 F(\boldsymbol{x}_k;\zeta_k) - \nabla^2 f_k\|_F^2] \\ &\le (1-\gamma_k)\|\mathcal{V}_{1,k-1}\|_F^2 + \gamma_k^2 d\sigma_H^2.\end{aligned}$$

Since $\sum_{k=0}^{\infty}\gamma_k = \infty$ and $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$, by Robbins-Siegmund theorem (Robbins and Siegmund, 1971), we know $\|\mathcal{V}_{1,k}\|_F^2$ is convergent and $\sum_{k=0}^{\infty}\gamma_k\|\mathcal{V}_{1,k-1}\|_F^2 < \infty$ almost surely. This further implies $\|\mathcal{V}_{1,k}\|_F \to 0$ almost surely and we complete the proof.

D.9 Proof of Lemma C.9

We introduce a lemma that shows the recursive relation between $\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2$ and $\|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2$. See Appendix D.10 for the proof.

Lemma D.3. *Under Assumption 3.6 and suppose $\alpha_k = \iota_1(k+1)^{-b_1}$, $\beta_k = \iota_2(k+1)^{-b_2}$ satisfy $\iota_1, \iota_2, b_1 > 0$, $b_2 \in (0,1)$. There exists a deterministic integer $\bar{k}_0 > 0$ such that for any $k_0 \ge \bar{k}_0$, the following result holds for some deterministic constant $\Upsilon_1(k_0) > 0$: for any $k \ge k_0$,*

$$\mathbb{E}[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0} > k+1}]$$

$$\leq \{1-2(1-2/\Upsilon)\alpha_k\}\,\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{w}^\star\|^2\mathbf{1}_{\tau_{k_0}>k}]+(2+\psi)\Upsilon^3\alpha_k\mathbb{E}[\|\bar{\boldsymbol{g}}_k-\nabla f_k\|^2\mathbf{1}_{\tau_{k_0}>k}],$$

$$\mathbb{E}[\|\bar{\boldsymbol{g}}_{k+1}-\nabla f_{k+1}\|^2\mathbf{1}_{\tau_{k_0}>k+1}]$$
$$\leq \Upsilon_1(k_0)\left\{\beta_k+\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\left\{\mathbb{E}[(\|\bar{\boldsymbol{g}}_i-\nabla f_i\|^2+\|\boldsymbol{w}_i-\boldsymbol{w}^\star\|^2)\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2\right\},$$

where $\Upsilon\geq 2$ *is defined in the stopping time* τ_{k_0} *in* (C.20).

With the above recursion, we claim that for any $q\geq 0$, there exists a deterministic integer $\bar{k}_0>0$ such that for any $k_0\geq\bar{k}_0$, we have for some constant $\Upsilon_2(k_0)>0$ that

$$\mathbb{E}\big[(\|\boldsymbol{w}_k-\boldsymbol{w}^\star\|^2+\|\bar{\boldsymbol{g}}_k-\nabla f_k\|^2)\mathbf{1}_{\tau_{k_0}>k}\big]\leq\Upsilon_2(k_0)\left(\beta_k+(\alpha_k/\beta_k)^{2q}\right),\qquad\forall k\geq k_0.\quad\text{(D.3)}$$

We prove the above statement by induction. When $q=0$, we apply the definition of τ_{k_0} in (C.20) and have $\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{w}^\star\|^2\mathbf{1}_{\tau_{k_0}>k}]\leq 1/\Upsilon$ for any $k\geq k_0$. Furthermore, by the recursive form of $\bar{\boldsymbol{g}}_k$ in (15) and Assumption 3.6, for any $k\geq k_0$,

$$\begin{aligned}\mathbb{E}[\|\bar{\boldsymbol{g}}_k\|^2\mathbf{1}_{\tau_{k_0}>k}]&\leq(1-\beta_k)\mathbb{E}[\|\bar{\boldsymbol{g}}_{k-1}\|^2\mathbf{1}_{\tau_{k_0}>k-1}]+\beta_k\mathbb{E}[\|\nabla F(\boldsymbol{x}_k;\zeta_k)\|^2\mathbf{1}_{\tau_{k_0}>k}]\\&\leq(1-\beta_k)\mathbb{E}[\|\bar{\boldsymbol{g}}_{k-1}\|^2\mathbf{1}_{\tau_{k_0}>k-1}]+3\beta_k(\sigma_g^2+\|\nabla f_k-\nabla f^\star\|^2\mathbf{1}_{\tau_{k_0}>k}+\|\nabla f^\star\|^2)\\&\overset{\text{(C.20)}}{\leq}(1-\beta_k)\mathbb{E}[\|\bar{\boldsymbol{g}}_{k-1}\|^2\mathbf{1}_{\tau_{k_0}>k-1}]+3\beta_k\left(\sigma_g^2+1/\Upsilon+\|\nabla f^\star\|^2\right).\end{aligned}$$

The above display, together with the fact that $\mathbb{E}[\|\bar{\boldsymbol{g}}_{k_0}\|^2]\leq\Upsilon_3(k_0)$ for some constant $\Upsilon_3(k_0)$ due to Assumption 3.6, implies that $\mathbb{E}[\|\bar{\boldsymbol{g}}_k-\nabla f_k\|^2\mathbf{1}_{\tau_{k_0}>k}]\leq\Upsilon_4(k_0)$, $\forall k\geq k_0$ for some constant $\Upsilon_4(k_0)$. This verifies (D.3) for $q=0$. Suppose that (D.3) holds for $q\geq 0$, then we establish the results for $q+1$. In particular, we apply Lemmas D.3 and B.4 and have ($\lesssim$ hides deterministic constants that may depend on k_0)

$$\begin{aligned}&\mathbb{E}[\|\bar{\boldsymbol{g}}_{k+1}-\nabla f_{k+1}\|^2\mathbf{1}_{\tau_{k_0}>k+1}]\\&\lesssim\beta_k+\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\left\{\mathbb{E}[(\|\bar{\boldsymbol{g}}_i-\nabla f_i\|^2+\|\boldsymbol{w}_i-\boldsymbol{w}^\star\|^2)\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2\\&\lesssim\beta_k+\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\left\{\beta_i+\left(\frac{\alpha_i}{\beta_i}\right)^{2q}\right\}^{1/2}\right)^2\lesssim\beta_k+\left(\frac{\alpha_k}{\beta_k}\right)^{2(q+1)}.\end{aligned}$$

Moreover, we still apply Lemma D.3 and have for any $b > 0$ and $k \geq k_0$,

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0} > k+1}] \\
&\quad \leq \{1 - 2(1-2/\Upsilon)\alpha_k\}\, \mathbb{E}[\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0} > k}] + (2+\psi)\Upsilon^3 \alpha_k \mathbb{E}[\|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0} > k}] \\
&\quad \leq (2+\psi)\Upsilon^3 \sum_{i=k_0}^{k} \prod_{j=i+1}^{k} \{1 - 2(1-2/\Upsilon)\alpha_j\}\, \alpha_i \mathbb{E}[\|\bar{\boldsymbol{g}}_i - \nabla f_i\|^2 \mathbf{1}_{\tau_{k_0} > i}] \\
&\qquad + \prod_{j=k_0}^{k} \{1 - 2(1-2/\Upsilon)\alpha_j\}\, \mathbb{E}[\|\boldsymbol{w}_{k_0} - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0} > k_0}] \\
&\quad \overset{(C.21)}{\lesssim} \sum_{i=k_0}^{k} \prod_{j=i+1}^{k} \{1 - 2(1-2/\Upsilon)\alpha_j\}\, \alpha_i \mathbb{E}[\|\bar{\boldsymbol{g}}_i - \nabla f_i\|^2 \mathbf{1}_{\tau_{k_0} > i}] + \left(k^{-b}\mathbf{1}_{b_1<1} + k^{-2(1-2/\Upsilon)\iota_1}\mathbf{1}_{b_1=1}\right) \\
&\quad \lesssim \sum_{i=0}^{k} \prod_{j=i+1}^{k} \{1 - 2(1-2/\Upsilon)\alpha_j\}\, \alpha_i \left\{\beta_i + \left(\frac{\alpha_i}{\beta_i}\right)^{2(q+1)}\right\} + \left(k^{-b}\mathbf{1}_{b_1<1} + k^{-2(1-2/\Upsilon)\iota_1}\mathbf{1}_{b_1=1}\right) \\
&\quad \lesssim \beta_k + (\alpha_k/\beta_k)^{2(q+1)},
\end{aligned}
$$

where the last inequality is because when $b_1 = 1$, we have $2(1-2/\Upsilon)\iota_1 > 1 > b_2$ by choosing Υ is large enough. This completes the induction step of (D.3) and further completes the proof.

D.10 Proof of Lemma D.3

By the local update (C.19), we know for any $k \geq k_0$,

$$
\begin{aligned}
&\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|^2 = \|\boldsymbol{w}_k - \boldsymbol{w}^\star + \bar{\alpha}_k \bar{\Delta}\boldsymbol{w}_k\|^2 = \|\boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k\|^2 \\
&= \|\boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k - \bar{\alpha}_k \boldsymbol{H}_k^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k)\|^2 \\
&= \|\boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\|^2 + \bar{\alpha}_k^2 \|\boldsymbol{H}_k^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k)\|^2 \\
&\quad - 2\bar{\alpha}_k \langle \boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k, \boldsymbol{H}_k^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k)\rangle \\
&\leq \left(1 + \frac{1.5\bar{\alpha}_k}{\Upsilon}\right) \|\boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\|^2 + \left(\bar{\alpha}_k^2 + \frac{\Upsilon \bar{\alpha}_k}{1.5}\right) \|\boldsymbol{H}_k^{-1}(\bar{\nabla}_{\boldsymbol{w}} \mathcal{L}_k - \nabla_{\boldsymbol{w}} \mathcal{L}_k)\|^2.
\end{aligned}
$$

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

$$
\begin{aligned}
&\|\boldsymbol{w}_k - \boldsymbol{w}^\star - \bar{\alpha}_k \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\|^2 \mathbf{1}_{\tau_{k_0} > k} \\
&= \left(\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 - 2\bar{\alpha}_k \langle \boldsymbol{w}_k - \boldsymbol{w}^\star, \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\rangle + \bar{\alpha}_k^2 \|\boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\|^2\right) \mathbf{1}_{\tau_{k_0} > k} \\
&= \left((1 - 2\bar{\alpha}_k)\|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 + 2\bar{\alpha}_k \langle \boldsymbol{w}_k - \boldsymbol{w}^\star, \boldsymbol{w}_k - \boldsymbol{w}^\star - \boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\rangle + \bar{\alpha}_k^2 \|\boldsymbol{H}_k^{-1} \nabla_{\boldsymbol{w}} \mathcal{L}_k\|^2\right) \mathbf{1}_{\tau_{k_0} > k}
\end{aligned}
$$

$$\overset{(C.20)}{\leq} \left(1 - 2\bar{\alpha}_k + \frac{2\bar{\alpha}_k}{\Upsilon} + \Upsilon^4\bar{\alpha}_k^2\right) \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k}.$$

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

$$\|\boldsymbol{H}_k^{-1}(\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_k - \nabla_{\boldsymbol{w}}\mathcal{L}_k)\|^2 \mathbf{1}_{\tau_{k_0}>k} \overset{(C.20)}{\leq} \Upsilon^2 \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0}>k}.$$

Combining the above three displays and noting that $\mathbf{1}_{\tau_{k_0}>k+1} \leq \mathbf{1}_{\tau_{k_0}>k}$, we obtain

$$\begin{aligned}
\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k+1} &\leq \left(1 + \frac{1.5\bar{\alpha}_k}{\Upsilon}\right)\left(1 - 2\bar{\alpha}_k + \frac{2\bar{\alpha}_k}{\Upsilon} + \Upsilon^4\bar{\alpha}_k^2\right) \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k} \\
&\quad + \Upsilon^2\left(\bar{\alpha}_k^2 + \frac{\Upsilon\bar{\alpha}_k}{1.5}\right) \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0}>k} \\
&\leq \left(1 - 2\bar{\alpha}_k + \frac{3.5\bar{\alpha}_k}{\Upsilon} + \Upsilon^4\bar{\alpha}_k^2 + \frac{3\bar{\alpha}_k^2}{\Upsilon^2} + 1.5\Upsilon^3\bar{\alpha}_k^3\right) \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k} \\
&\quad + \Upsilon^2\left(\bar{\alpha}_k^2 + \frac{\Upsilon\bar{\alpha}_k}{1.5}\right) \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0}>k}.
\end{aligned}$$

To simplify the above display, we let $\bar{k}_0$ be large enough (with a deterministic threshold) such that

$$\Upsilon^4\bar{\alpha}_k + \frac{3\bar{\alpha}_k}{\Upsilon^2} + 1.5\Upsilon^3\bar{\alpha}_k^2 \leq \frac{0.5}{\Upsilon} \quad \text{and} \quad \left(\bar{\alpha}_k + \frac{\Upsilon}{1.5}\right)\bar{\alpha}_k \leq (2+\psi)\Upsilon\alpha_k, \quad \forall k \geq \bar{k}_0. \tag{D.4}$$

This is achievable since $b_1 > 0$ and $\operatorname{limsup}_{k\to\infty} \bar{\alpha}_k/\alpha_k \leq 1 + \psi$. Then, we obtain

$$\begin{aligned}
&\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k+1} \\
&\quad \leq \{1 - 2(1 - 2/\Upsilon)\bar{\alpha}_k\} \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k} + (2+\psi)\Upsilon^3\alpha_k \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0}>k} \\
&\quad \leq \{1 - 2(1 - 2/\Upsilon)\alpha_k\} \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|^2 \mathbf{1}_{\tau_{k_0}>k} + (2+\psi)\Upsilon^3\alpha_k \|\bar{\boldsymbol{g}}_k - \nabla f_k\|^2 \mathbf{1}_{\tau_{k_0}>k},
\end{aligned}$$

where the last inequality uses $\Upsilon \geq 2$. This shows the first argument of the lemma. For the second argument of the lemma, we recall (D.2) from the proof in Appendix D.4 and have

$$\begin{aligned}
\bar{\boldsymbol{g}}_{k+1} - \nabla f_{k+1} &= \sum_{i=0}^{k+1} \prod_{j=i+1}^{k+1} (1-\beta_j)\beta_i \left(\nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f_i\right) + \sum_{i=0}^{k+1}\prod_{j=i}^{k+1}(1-\beta_j)\left(\nabla f_{i-1} - \nabla f_i\right) \\
&\coloneqq \mathcal{W}_{1,k+1} + \mathcal{W}_{2,k+1}.
\end{aligned} \tag{D.5}$$

By Assumption 3.6 and Lemma B.4, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{W}_{1,k+1}\|^2 \mathbf{1}_{\tau_{k_0}>k+1}] &\leq \mathbb{E}[\|\mathcal{W}_{1,k+1}\|^2] \\
&= \sum_{i=0}^{k+1}\prod_{j=i+1}^{k+1}(1-\beta_j)^2\beta_i^2 \mathbb{E}[\|\nabla F(\boldsymbol{x}_i;\zeta_i) - \nabla f_i\|^2] = \mathcal{O}\left(\beta_k\right).
\end{aligned} \tag{D.6}$$

For the term $\mathcal{W}_{2,k+1}$, we use $\beta_k \le 1$ for $k \ge k_0$ and separate the sum in $\mathcal{W}_{2,k+1}$ into two terms:

$$
\begin{aligned}
&\|\mathcal{W}_{2,k+1}\|\mathbf{1}_{\tau_{k_0}>k+1} \\
&\le \sum_{i=0}^{k_0}\prod_{j=i}^{k+1}|1-\beta_j|\|\nabla f_{i-1}-\nabla f_i\| + \sum_{i=k_0+1}^{k+1}\prod_{j=i}^{k+1}(1-\beta_j)\|\nabla f_{i-1}-\nabla f_i\|\mathbf{1}_{\tau_{k_0}>k+1} \\
&\overset{(C.20)}{\le} \prod_{j=0}^{k+1}|1-\beta_j|\sum_{i=0}^{k_0}\prod_{j=0}^{i-1}|1-\beta_j|^{-1}\|\nabla f_{i-1}-\nabla f_i\| + \Upsilon\sum_{i=k_0+1}^{k+1}\prod_{j=i}^{k+1}(1-\beta_j)\bar{\alpha}_{i-1}\|\bar{\Delta}\boldsymbol{x}_{i-1}\|\mathbf{1}_{\tau_{k_0}>k+1} \\
&\le \Upsilon_1(k_0)\prod_{j=0}^{k+1}|1-\beta_j| + \Upsilon\sum_{i=k_0}^{k}\prod_{j=i+1}^{k+1}(1-\beta_j)\bar{\alpha}_i\|\bar{\Delta}\boldsymbol{x}_i\|\mathbf{1}_{\tau_{k_0}>i} \\
&\overset{\substack{(C.20)\\(D.4)}}{\le} \Upsilon_1(k_0)\prod_{j=0}^{k+1}|1-\beta_j| + (3+1.5\psi)\Upsilon\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\|\bar{\Delta}\boldsymbol{x}_i\|\mathbf{1}_{\tau_{k_0}>i}, \qquad \text{(D.7)}
\end{aligned}
$$

where the second last inequality holds for some deterministic constant $\Upsilon_1(k_0) > 0$ due to the boundedness of ∇f_i. For the first term on the right hand side, we apply (C.21) and know that $\prod_{j=0}^{k+1}|1-\beta_j| = o(\beta_k)$. For the second term on the right hand side, we have for $k_0 \le i < \tau_{k_0}$,

$$
\begin{aligned}
\|\bar{\Delta}\boldsymbol{x}_i\|^2 = \|\boldsymbol{H}_i^{-1}\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_i\|^2 &\overset{(C.20)}{\le} 2\Upsilon^2(\|\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_i - \nabla_{\boldsymbol{w}}\mathcal{L}_i\|^2 + \|\nabla_{\boldsymbol{w}}\mathcal{L}_i\|^2) \\
&\overset{(C.20)}{\le} 2\Upsilon^4(\|\bar{\boldsymbol{g}}_i - \nabla f_i\|^2 + \|\boldsymbol{w}_i - \boldsymbol{w}^\star\|^2). \qquad \text{(D.8)}
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\|\bar{\Delta}\boldsymbol{x}_i\|\mathbf{1}_{\tau_{k_0}>i}\right)^2\right] \le \left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\left\{\mathbb{E}[\|\bar{\Delta}\boldsymbol{x}_i\|^2\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2 \\
&\overset{(D.8)}{\le} 2\Upsilon^4\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\beta_j)\alpha_i\left\{\mathbb{E}[(\|\bar{\boldsymbol{g}}_i - \nabla f_i\|^2 + \|\boldsymbol{w}_i - \boldsymbol{w}^\star\|^2)\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2. \qquad \text{(D.9)}
\end{aligned}
$$

Finally, combining (D.5), (D.6), (D.7), (D.9), we complete the proof.

D.11 Proof of Lemma C.10

We follow the proof in Appendix D.8 and have

$$
\begin{aligned}
\bar{\boldsymbol{Q}}_k - \nabla^2 f^\star &= \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\left(\nabla^2 F(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f_i\right) \\
&\quad + \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\left(\nabla^2 f_i - \nabla^2 f^\star\right) + \prod_{j=0}^{k}(1-\gamma_j)(\bar{\boldsymbol{Q}}_{-1} - \nabla^2 f^\star) := \mathcal{V}_{1,k} + \mathcal{V}_{2,k}.
\end{aligned}
$$

We analyze each term separately. For the term $\mathcal{V}_{1,k}$, we apply Assumption 4.3 and have

$$\mathbb{E}[\|\mathcal{V}_{1,k}\|_F^2] = \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)^2\gamma_i^2\mathbb{E}[\|\nabla^2 F(\boldsymbol{x}_i;\zeta_i)-\nabla^2 f_i\|_F^2] = \begin{cases}\mathcal{O}(\gamma_k) & \text{if } b_3<1\\ o(\sqrt{\alpha_k}) & \text{if } b_3=1\end{cases} = o(\sqrt{\alpha_k}),$$

where the second last equality uses Lemma B.4 and the condition that $\iota_3 > 0.25b_1$ if $b_3 = 1$; and the last equality uses the condition $b_3 > 0.5b_1$. For the term $\mathcal{V}_{2,k}$, we apply (C.21) and have

$$\prod_{j=0}^{k}|1-\gamma_j|^2 = \begin{cases}o(\gamma_k) & \text{if } b_3<1\\ \mathcal{O}\left(\frac{1}{k^{2\iota_3}}\right) & \text{if } b_3=1\end{cases} = o(\sqrt{\alpha_k}). \tag{D.10}$$

Next, we analyze the first term in $\mathcal{V}_{2,k}$. We have for some constants $\Upsilon_1(k_0), \Upsilon_2(k_0) > 0$ that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i(\nabla^2 f_i-\nabla^2 f^\star)\right\|^2\mathbf{1}_{\tau_{k_0}>k}\right]\\
&\le 2\mathbb{E}\left[\left\|\sum_{i=0}^{k_0}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i(\nabla^2 f_i-\nabla^2 f^\star)\right\|^2+\left\|\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i(\nabla^2 f_i-\nabla^2 f^\star)\right\|^2\mathbf{1}_{\tau_{k_0}>k}\right]\\
&\overset{\text{(D.10)}}{\le} o(\Upsilon_1(k_0)\sqrt{\alpha_k})+2\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\left\{\mathbb{E}[\|\nabla f_i-\nabla f^\star\|^2\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2\\
&\overset{\text{(C.20)}}{\le} o(\Upsilon_1(k_0)\sqrt{\alpha_k})+2\Upsilon^2\left(\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\left\{\mathbb{E}[\|\boldsymbol{x}_i-\boldsymbol{x}^\star\|^2\mathbf{1}_{\tau_{k_0}>i}]\right\}^{1/2}\right)^2 = o(\Upsilon_2(k_0)\sqrt{\alpha_k}),
\end{aligned}$$

where the last equality is due to Lemmas B.4, C.9, and the facts that

(a) $b_3 \in (0,1)$: We apply $b_2 > 0.5b_1$ and have

$$\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\sqrt{\beta_i} = O(\sqrt{\beta_k}) = o(\alpha_k^{0.25}).$$

(b) $b_3 = 1$: We apply $b_2 > 0.5b_1$ and $\iota_3 > 0.25b_1 \Leftrightarrow 1-\frac{0.5b_1}{b_2}\cdot\frac{0.5b_2}{\iota_3} > 0$, and have

$$\frac{1}{\alpha_k^{0.25}}\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\sqrt{\beta_i} = \frac{(\sqrt{\beta_k})^{0.5b_1/b_2}}{\alpha_k^{0.25}}\frac{1}{(\sqrt{\beta_k})^{0.5b_1/b_2}}\sum_{i=k_0}^{k}\prod_{j=i+1}^{k}(1-\gamma_j)\gamma_i\sqrt{\beta_i} = o(1).$$

Thus, we have shown that $\mathbb{E}[\|\bar{\boldsymbol{Q}}_k-\nabla^2 f^\star\|^2\mathbf{1}_{\tau_{k_0}>k}] = o(\Upsilon_3(k_0)\sqrt{\alpha_k})$ for some constant $\Upsilon_3(k_0) > 0$. Finally, by noting from (C.20) that

$$\begin{aligned}
\|\boldsymbol{H}_k-\boldsymbol{H}^\star\|\mathbf{1}_{\tau_{k_0}>k} &\le \left(\|\bar{\boldsymbol{Q}}_k-\nabla^2 f^\star\|+\|\nabla^2_{\boldsymbol{w}}\mathcal{L}_k-\nabla^2_{\boldsymbol{w}}\mathcal{L}^\star\|+\|\nabla^2 f_k-\nabla^2 f^\star\|\right)\mathbf{1}_{\tau_{k_0}>k}\\
&\le \left(\|\bar{\boldsymbol{Q}}_k-\nabla^2 f^\star\|+2\Upsilon\|\boldsymbol{w}_k-\boldsymbol{w}^\star\|\right)\mathbf{1}_{\tau_{k_0}>k},
\end{aligned}$$

and applying Lemma C.9, we complete the proof.

D.12 Proof of Lemma C.11

Recalling the definition of $\mathcal{J}_{3,k}$ from (C.22), we have

$$\begin{aligned}
\mathcal{J}_{3,k} &= \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i(\boldsymbol{H}^\star)^{-1}(\bar{\nabla}_{\boldsymbol{w}}\mathcal{L}_i - \nabla_{\boldsymbol{w}}\mathcal{L}_i)\\
&\overset{(13)}{=} \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (\boldsymbol{J}^\star)^\top\\ \boldsymbol{J}^\star & \boldsymbol{0}\end{pmatrix}^{-1}\begin{pmatrix}\bar{\boldsymbol{g}}_i - \nabla f_i\\ \boldsymbol{0}\end{pmatrix}\\
&\overset{(D.2)}{=} \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (\boldsymbol{J}^\star)^\top\\ \boldsymbol{J}^\star & \boldsymbol{0}\end{pmatrix}^{-1}\left\{\begin{pmatrix}\mathcal{W}_{1,i}\\ \boldsymbol{0}\end{pmatrix}+\begin{pmatrix}\mathcal{W}_{2,i}\\ \boldsymbol{0}\end{pmatrix}\right\} = \mathcal{J}^{(1)}_{3,k}+\mathcal{J}^{(2)}_{3,k}.
\end{aligned}$$

We first analyze $\mathcal{J}^{(2)}_{3,k}$ and aim to show that $\mathcal{J}^{(2)}_{3,k} = o_p(\sqrt{\alpha_k})$. By Lemma C.8, it suffices to show $\mathcal{W}_{2,k}\mathbf{1}_{\tau_{k_0}>k} = o_p(\sqrt{\alpha_k})$. From the definition of $\mathcal{W}_{2,k}$ in (D.2) and Lemma C.8, this result is further implied by

$$\|\nabla f_k - \nabla f_{k+1}\|\mathbf{1}_{\tau_{k_0}>k+1} \overset{(C.20)}{\leq} \Upsilon\bar{\alpha}_k\|\bar{\Delta}\boldsymbol{x}_k\|\mathbf{1}_{\tau_{k_0}>k} \leq o_p(\beta_k\sqrt{\alpha_k}).$$

By (D.8), Lemma C.9, and the fact that $\alpha_k\sqrt{\beta_k} = o(\beta_k\sqrt{\alpha_k})$, we obtain the above desired result.

Next, we analyze $\mathcal{J}^{(1)}_{3,k}$. By the definition of $\mathcal{W}_{1,k}$ in (D.2), we have

$$\begin{aligned}
\mathcal{J}^{(1)}_{3,k} &= \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (\boldsymbol{J}^\star)^\top\\ \boldsymbol{J}^\star & \boldsymbol{0}\end{pmatrix}^{-1}\begin{pmatrix}\mathcal{W}_{1,i}\\ \boldsymbol{0}\end{pmatrix}\\
&= \sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i\sum_{h=0}^{i}\prod_{l=h+1}^{i}(1-\beta_l)\beta_h \underbrace{\begin{pmatrix}\nabla^2_{\boldsymbol{x}}\mathcal{L}^\star & (\boldsymbol{J}^\star)^\top\\ \boldsymbol{J}^\star & \boldsymbol{0}\end{pmatrix}^{-1}\begin{pmatrix}\nabla F(\boldsymbol{x}_h;\zeta_h) - \nabla f_h\\ \boldsymbol{0}\end{pmatrix}}_{\boldsymbol{\phi}^\star_h}\\
&= \sum_{h=0}^{k}\underbrace{\sum_{i=h}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\,\alpha_i\prod_{l=h+1}^{i}(1-\beta_l)\beta_h}_{a_{h,k}}\boldsymbol{\phi}^\star_h = \sum_{h=0}^{k}a_{h,k}\boldsymbol{\phi}^\star_h.
\end{aligned}$$

We first claim that $\mathbb{E}_h[\boldsymbol{\phi}^\star_h\boldsymbol{\phi}^{\star\top}_h] \to \boldsymbol{\Omega}^\star$ almost surely as $h\to\infty$. In fact, by Assumptions 4.1, 4.5, and the smoothness of $\nabla f(\boldsymbol{x})$, we have

$$\begin{aligned}
&\left\|\mathbb{E}_h[\nabla F(\boldsymbol{x}_h,\zeta_h)\nabla F(\boldsymbol{x}_h,\zeta_h)^\top - \nabla f_h\nabla f_h^\top] - \mathbb{E}[\nabla F(\boldsymbol{x}^\star;\zeta)\nabla F(\boldsymbol{x}^\star;\zeta)^\top - \nabla f^\star\nabla f^{\star\top}]\right\|\\
&\leq \left\|\mathbb{E}_h[\nabla F(\boldsymbol{x}_h;\zeta_h)\nabla F(\boldsymbol{x}_h;\zeta_h)^\top] - \mathbb{E}[\nabla F(\boldsymbol{x}^\star;\zeta)\nabla F(\boldsymbol{x}^\star;\zeta)^\top]\right\|\\
&\quad + (\|\nabla f_h\| + \|\nabla f^\star\|)\,\|\nabla f_h - \nabla f^\star\| \to 0 \quad \text{as} \quad h\to\infty.
\end{aligned}$$

With this result, we then analyze the conditional variance process. We have

$$
\begin{aligned}
&\frac{1}{\alpha_k}\sum_{h=0}^{k} a_{h,k}^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&= \frac{1}{\alpha_k}\sum_{h=0}^{k}\sum_{i=h}^{k}\sum_{i'=h}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{l=h+1}^{i}(1-\beta_l)\beta_h \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \prod_{l'=h+1}^{i'}(1-\beta_{l'})\beta_h \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&= \frac{1}{\alpha_k}\sum_{i=0}^{k}\sum_{i'=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \sum_{h=0}^{\min\{i,i'\}}\prod_{l=h+1}^{i}(1-\beta_l)\prod_{l'=h+1}^{i'}(1-\beta_{l'})\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&= \frac{2}{\alpha_k}\sum_{i=0}^{k}\sum_{i'=0}^{i}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \sum_{h=0}^{i'}\prod_{l=h+1}^{i}(1-\beta_l)\prod_{l'=h+1}^{i'}(1-\beta_{l'})\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&\quad - \frac{1}{\alpha_k}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^2\alpha_i^2 \sum_{h=0}^{i}\prod_{l=h+1}^{i}(1-\beta_l)^2\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&= \frac{2}{\alpha_k}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^2\alpha_i \sum_{i'=0}^{i}\prod_{j'=i'+1}^{i}(1-\alpha_{j'})(1-\beta_{j'})\alpha_{i'} \sum_{h=0}^{i'}\prod_{l'=h+1}^{i'}(1-\beta_{l'})^2\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] \\
&\quad - \frac{1}{\alpha_k}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^2\alpha_i^2 \sum_{h=0}^{i}\prod_{l=h+1}^{i}(1-\beta_l)^2\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}].
\end{aligned}
$$

We apply Lemma B.4 and note that

$$
\begin{aligned}
&\lim_{i\to\infty}\frac{1}{\beta_i}\sum_{h=0}^{i}\prod_{l=h+1}^{i}(1-\beta_l)^2\beta_h^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] = 0.5\boldsymbol{\Omega}^\star, \\
&\lim_{i\to\infty}\frac{1}{\alpha_i}\sum_{i'=0}^{i}\prod_{j'=i'+1}^{i}(1-\alpha_{j'})(1-\beta_{j'})\alpha_{i'}\beta_{i'} = 1, \\
&\lim_{k\to\infty}\frac{1}{\alpha_k}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^2\alpha_i^2 = \eta := \begin{cases} 0.5, & \text{if} \quad b_1 \in (0,1, \\ \frac{\iota_1}{2\iota_1-1}, & \text{if} \quad b_1 = 1, \end{cases} \\
&\lim_{k\to\infty}\frac{1}{\alpha_k}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^2\alpha_i^2\beta_i = 0.
\end{aligned}
$$

Combining the above two displays, we obtain almost surely,

$$
\lim_{k\to\infty}\frac{1}{\alpha_k}\sum_{h=0}^{k} a_{h,k}^2 \mathbb{E}_h[\boldsymbol{\phi}_h^\star \boldsymbol{\phi}_h^{\star\top}] = \eta\cdot\boldsymbol{\Omega}^\star. \tag{D.11}
$$

We next verify the Lindeberg condition. It is equivalent to showing that for any $\epsilon > 0$,

$$
\lim_{k\to\infty}\frac{1}{\alpha_k}\sum_{h=0}^{k} a_{h,k}^2 \mathbb{E}_h\left[\|\boldsymbol{\phi}_h^\star\|^2 \cdot \mathbf{1}_{\|a_{h,k}\boldsymbol{\phi}_h^\star\| \geq \epsilon\sqrt{\alpha_k}}\right] \leq \lim_{k\to\infty}\frac{1}{\epsilon^\delta \alpha_k^{1+0.5\delta}}\sum_{h=0}^{k} a_{h,k}^{2+\delta}\mathbb{E}_h[\|\boldsymbol{\phi}_h^\star\|^{2+\delta}] = 0.
$$

By Assumption 4.5, it suffices to show

$$\frac{1}{\alpha_k^{1+0.5\delta}}\sum_{h=0}^{k} a_{h,k}^{2+\delta} \;\le\; \frac{k+1}{\alpha_k^{1+0.5\delta}}\left(\frac{1}{k+1}\sum_{h=0}^{k} a_{h,k}^{3}\right)^{\frac{2+\delta}{3}} \;=\; \left(\frac{\sum_{h=0}^{k} a_{h,k}^{3}}{\alpha_k^{\frac{3+1.5\delta}{2+\delta}+\frac{1-\delta}{b_1(2+\delta)}}}\right)^{\frac{2+\delta}{3}} \;\to\; 0. \quad \text{(D.12)}$$

In particular, we have

$$\begin{aligned}
\sum_{h=0}^{k} a_{h,k}^{3} &= \sum_{h=0}^{k}\sum_{i=h}^{k}\sum_{i'=h}^{k}\sum_{i''=h}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{l=h+1}^{i}(1-\beta_l)\beta_h \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \prod_{l'=h+1}^{i'}(1-\beta_{l'})\beta_h\cdot \\
&\qquad \prod_{j''=i''+1}^{k}(1-\alpha_{j''})\alpha_{i''} \prod_{l''=h+1}^{i''}(1-\beta_{l''})\beta_h \\
&= \sum_{i=0}^{k}\sum_{i'=0}^{k}\sum_{i''=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \prod_{j''=i''+1}^{k}(1-\alpha_{j''})\alpha_{i''}\cdot \\
&\qquad \sum_{h=0}^{\min\{i,i',i''\}} \prod_{l=h+1}^{i}(1-\beta_l) \prod_{l'=h+1}^{i'}(1-\beta_{l'}) \prod_{l''=h+1}^{i''}(1-\beta_{l''})\beta_h^3 \\
&\le 6\sum_{i=0}^{k}\sum_{i'=0}^{i}\sum_{i''=0}^{i'}\prod_{j=i+1}^{k}(1-\alpha_j)\alpha_i \prod_{j'=i'+1}^{k}(1-\alpha_{j'})\alpha_{i'} \prod_{j''=i''+1}^{k}(1-\alpha_{j''})\alpha_{i''}\cdot \\
&\qquad \sum_{h=0}^{i''} \prod_{l=h+1}^{i}(1-\beta_l) \prod_{l'=h+1}^{i'}(1-\beta_{l'}) \prod_{l''=h+1}^{i''}(1-\beta_{l''})\beta_h^3 \qquad (i\ge i'\ge i'') \\
&= 6\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^3\alpha_i \sum_{i'=0}^{i}\prod_{j'=i'+1}^{i}(1-\alpha_{j'})^2(1-\beta_{j'})\alpha_{i'}\cdot \\
&\qquad \sum_{i''=0}^{i'}\prod_{j''=i''+1}^{i'}(1-\alpha_{j''})(1-\beta_{j''})^2\alpha_{i''} \sum_{h=0}^{i''}\prod_{l''=h+1}^{i''}(1-\beta_{l''})^3\beta_h^3.
\end{aligned}$$

We apply Lemma B.4 and note that

$$\begin{aligned}
&\lim_{i''\to\infty}\frac{1}{\beta_{i''}^2}\sum_{h=0}^{i''}\prod_{l''=h+1}^{i''}(1-\beta_{l''})^3\beta_h^3 = \frac{1}{3},\\
&\lim_{i'\to\infty}\frac{1}{\alpha_{i'}\beta_{i'}}\sum_{i''=0}^{i'}\prod_{j''=i''+1}^{i'}(1-\alpha_{j''})(1-\beta_{j''})^2\alpha_{i''}\beta_{i''}^2 = \frac{1}{2},\\
&\lim_{i\to\infty}\frac{1}{\alpha_i^2}\sum_{i'=0}^{i}\prod_{j'=i'+1}^{i}(1-\alpha_{j'})^2(1-\beta_{j'})\alpha_{i'}^2\beta_{i'} = 1,\\
&\lim_{k\to\infty}\frac{1}{\alpha_k^{\frac{3+1.5\delta}{2+\delta}+\frac{1-\delta}{b_1(2+\delta)}}}\sum_{i=0}^{k}\prod_{j=i+1}^{k}(1-\alpha_j)^3\alpha_i^3 = 0,
\end{aligned}$$

where the last equality holds due to the following facts:

(a) $b_1 \in (0,1)$: We apply $b_1 > (1-\delta)/(1+0.5\delta)$ and have

$$\frac{1}{\alpha_k^{\frac{3+1.5\delta}{2+\delta}+\frac{1-\delta}{b_1(2+\delta)}}} \sum_{i=0}^{k} \prod_{j=i+1}^{k} (1-\alpha_j)^3 \alpha_i^3 = \mathcal{O}\left(\frac{\alpha_k^2}{\alpha_k^{\frac{3+1.5\delta}{2+\delta}+\frac{1-\delta}{b_1(2+\delta)}}} \right) \to 0.$$

(b) $b_1 = 1$: We apply $\iota_1 > 2/3$ and have

$$\frac{1}{\alpha_k^{\frac{4+0.5\delta}{2+\delta}}} \sum_{i=0}^{k} \prod_{j=i+1}^{k} (1-\alpha_j)^3 \alpha_i^3 = \mathcal{O}\left(\frac{\alpha_k^2}{\alpha_k^{\frac{4+0.5\delta}{2+\delta}}} \right) \to 0.$$

Thus, we have verified the Lindeberg condition. By the central limit theorem of martingale arrays (Hall and Heyde, 2014, Corollary 3.1), the results (D.11) and (D.12) lead to $1/\sqrt{\alpha_k} \cdot \mathcal{J}_{3,k}^{(1)} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \eta \cdot \boldsymbol{\Omega}^\star)$. This completes the proof.

E Boundedness of Dual Multipliers

Theorem E.1 shows that if the iteration sequence $\boldsymbol{x}_k$ generated by our method converges to a feasible point $\bar{\boldsymbol{x}}$ satisfying EGMFCQ (Definition 2.1), then the corresponding Lagrangian dual multipliers $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})$ associated with the subproblem (17) with the exact true gradient ∇f_k are bounded. This result validates and provides theoretical justification for the boundedness assumption on the dual variables in Assumption 3.5.

Theorem E.1. *Suppose $\bar{\boldsymbol{x}}$ is feasible for Problem (1) and satisfies EGMFCQ. If $\lim_{k\to\infty} \boldsymbol{x}_k = \bar{\boldsymbol{x}}$, then the corresponding dual multipliers $(\boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_k^{sub})$ of the SQP subproblem (17) are bounded.*

Proof. We prove it by contradiction. Suppose that there exists a sequence $\{(\boldsymbol{x}_k, \bar{\boldsymbol{B}}_k, \boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})\}$ such that $\boldsymbol{x}_k \to \bar{\boldsymbol{x}}$, $\left\|(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})\right\| \to \infty$ and $\kappa_1 \mathbf{I} \preceq \bar{\boldsymbol{B}}_k \preceq \kappa_2 \mathbf{I}$. Recall that $\Delta\boldsymbol{x}_k$ and $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})$ denote the primal and dual solutions of the SQP subproblem (17) at $\boldsymbol{x}_k$ with the true gradient ∇f_k, satisfying the following KKT conditions:

$$\begin{aligned}
&\nabla f_k + \bar{\boldsymbol{B}}_k \Delta\boldsymbol{x}_k + \nabla \boldsymbol{c}_k^\top \boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}} + \boldsymbol{\mu}_{2,k}^{\text{sub}} = \mathbf{0}, \\
&\theta_k \boldsymbol{c}_k + \nabla \boldsymbol{c}_k \Delta\boldsymbol{x}_k = \mathbf{0}, \qquad \boldsymbol{\ell} \le \boldsymbol{x}_k + \Delta\boldsymbol{x}_k \le \boldsymbol{u}, \qquad \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}} \ge \mathbf{0}, \\
&\boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{\ell} - \boldsymbol{x}_k - \Delta\boldsymbol{x}_k) = 0, \quad \boldsymbol{\mu}_{2,k}^{\text{sub}\top}(\boldsymbol{x}_k + \Delta\boldsymbol{x}_k - \boldsymbol{u}) = 0.
\end{aligned} \tag{E.1}$$

Note that the sequences $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})/\|(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_k^{\text{sub}})\|$ and $\Delta\boldsymbol{x}_k$ are bounded. Without loss of gen-

erality, we assume $(\boldsymbol{\lambda}_k^{\mathrm{sub}}, \boldsymbol{\mu}_k^{\mathrm{sub}})/\|(\boldsymbol{\lambda}_k^{\mathrm{sub}}, \boldsymbol{\mu}_k^{\mathrm{sub}})\| \to (\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})$ and $\Delta \boldsymbol{x}_k \to \Delta \bar{\boldsymbol{x}}$. Dividing the two sides of (E.1) by $\|(\boldsymbol{\lambda}_k^{\mathrm{sub}}, \boldsymbol{\mu}_k^{\mathrm{sub}})\|$, taking the limit $k \to \infty$, and applying the feasibility of $\bar{\boldsymbol{x}}$, we have

$$\begin{aligned} &\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\mu}}_1 + \bar{\boldsymbol{\mu}}_2 = \boldsymbol{0}, \qquad \nabla \boldsymbol{c}(\bar{\boldsymbol{x}}) \Delta \bar{\boldsymbol{x}} = \boldsymbol{0}, \\ &\bar{\boldsymbol{\mu}}_1^\top (\boldsymbol{\ell} - \bar{\boldsymbol{x}}) = \bar{\boldsymbol{\mu}}_1^\top \Delta \bar{\boldsymbol{x}}, \qquad \bar{\boldsymbol{\mu}}_2^\top (\bar{\boldsymbol{x}} - \boldsymbol{u}) = -\bar{\boldsymbol{\mu}}_2^\top \Delta \bar{\boldsymbol{x}}. \end{aligned} \tag{E.2}$$

The above four equalities imply

$$\bar{\boldsymbol{\mu}}_1^\top (\boldsymbol{\ell} - \bar{\boldsymbol{x}}) + \bar{\boldsymbol{\mu}}_2^\top (\bar{\boldsymbol{x}} - \boldsymbol{u}) = 0. \tag{E.3}$$

Since $\bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2 \geq \boldsymbol{0}$, we deduce from (E.3) that $[\bar{\boldsymbol{\mu}}_1]_i > 0$ only if $[\bar{\boldsymbol{x}}]_i = [\boldsymbol{\ell}]_i$ and $[\bar{\boldsymbol{\mu}}_2]_i > 0$ only if $[\bar{\boldsymbol{x}}]_i = [\boldsymbol{u}]_i$. The EGMFCQ condition at $\bar{\boldsymbol{x}}$ (cf. Definition 2.1) implies that there exists $\boldsymbol{z} \in \mathbb{R}^d$ such that $\boldsymbol{c}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}}) \boldsymbol{z} = \boldsymbol{0}$, $[\boldsymbol{z}]_i > 0$ if $[\bar{\boldsymbol{x}}]_i = [\boldsymbol{\ell}]_i$, and $[\boldsymbol{z}]_i < 0$ if $[\bar{\boldsymbol{x}}]_i = [\boldsymbol{u}]_i$. Then, we have $-\boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_1 + \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_2 < 0$ if $\bar{\boldsymbol{x}}$ is on the boundary of the box constraints. Multiplying $-\boldsymbol{z}$ on both sides of the first equality in (E.2), we have

$$0 = -\boldsymbol{z}^\top (\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\mu}}_1 + \bar{\boldsymbol{\mu}}_2) = \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{\lambda}} + \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_1 - \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_2 = \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_1 - \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_2,$$

which contradicts $-\boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_1 + \boldsymbol{z}^\top \bar{\boldsymbol{\mu}}_2 < 0$. On the other hand, if $\bar{\boldsymbol{x}}$ is in the interior of the box constraints, then $\bar{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{\mu}}_2 = \boldsymbol{0}$. Together with the first equality of (E.2), the linear independence of the rows of $\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})$ shows $\bar{\boldsymbol{\lambda}} = \boldsymbol{0}$, which contradicts to the fact that $\|(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}})\| = 1$. This completes the proof. ■