Submitted to the Annals of Statistics

AN ASYMPTOTICALLY OPTIMAL METHOD FOR CONSTRAINED STOCHASTIC OPTIMIZATION*

BY SEN NA[1], YIHANG GAO[2], MICHAEL K. NG[3], AND MICHAEL W. MAHONEY[1]

[1]*ICSI and Department of Statistics, University of California, Berkeley*

[2]*Department of Mathematics, The University of Hong Kong*

[3]*Department of Mathematics, Hong Kong Baptist University*

We perform statistical inference for the solution of stochastic optimization problems with equality and box inequality constraints. The considered problems are prevalent in statistics and machine learning, encompassing constrained M-estimation, PDE-constrained problems, physics-inspired networks, and algorithmic fairness. We introduce a stochastic sequential quadratic programming method (StoSQP) to solve these problems, where we determine the search direction by 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 problems lies in dealing with the bias in the search direction. To address this challenge, we introduce a novel gradient averaging technique to debias the direction step, leading to Debiased-StoSQP. Our method achieves global almost sure convergence and exhibits local asymptotic normality with an optimal limiting covariance matrix in Hájek and Le Cam's sense. Additionally, a plug-in estimator of the covariance matrix is provided for practical inference purposes. To our knowledge, Debiased-StoSQP is the first fully online method to achieve asymptotic minimax optimality without relying on projection operators to the constraint set, which are incomputable for nonlinear problems. Through extensive experiments on benchmark nonlinear problems in the CUTEst test set, as well as on constrained generalized linear models and portfolio allocation problems, with both synthetic and real data, we demonstrate the superior performance of the method.

1. Introduction. We consider stochastic optimization problems with equality and box inequality constraints, given by the form:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^d} \quad & f(\boldsymbol{x}) = \mathbb{E}_{\zeta\sim\mathcal{P}}\left[F(\boldsymbol{x};\zeta)\right], \\ \text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}. \end{aligned} \tag{1.1}$$

Here, the vectors $\boldsymbol{\ell}$ and $\boldsymbol{u}$ denote the lower and upper bounds, respectively, with the symbol "$\le$" representing element-wise comparison; and $\zeta \sim \mathcal{P}$ is a random variable. The function $F(\cdot;\zeta):\mathbb{R}^d \to \mathbb{R}$ denotes a realization of the stochastic objective f, and $\boldsymbol{c}:\mathbb{R}^d \to \mathbb{R}^m$ encodes the deterministic equality constraints. Throughout this paper, we assume that f, $\boldsymbol{c}$, and $F(\cdot;\zeta)$ for each realization ζ are twice continuously differentiable. We aim to develop a *practical*, *fully online*, and *asymptotically optimal* method to solve Problem (1.1).

Constraints are useful tools for integrating prior models information, ensuring models' identifiability, and reducing dimensionality. We will provide concrete motivating examples in Section 1.1. Given the ubiquity of Problem (1.1), it is of particular interest to estimate its (local) solution $\boldsymbol{x}^*$ with n samples. Arguably, the most primitive estimator is the classical

*S. Na and Y. Gao contribute equally.

Keywords and phrases: Local asymptotic minimax optimality, Stochastic approximation, Sequential quadratic programming, Nonlinear nonconvex optimization.

M-estimator, where we generate samples $\zeta_1, \ldots, \zeta_n \overset{\text{iid}}{\sim} \mathcal{P}$ and solve the constrained problem by replacing the population loss f with the empirical loss $\hat{f}_n$:

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

In fact, the above constrained M-estimator is optimal in Hájek and Le Cam's sense [41, 72]. That is, the asymptotic consistency and normality of the minimizer $\hat{\boldsymbol{x}}_n$ is given by

$$\sqrt{n}\,(\hat{\boldsymbol{x}}_n - \boldsymbol{x}^*) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{L}^{\dagger}\mathrm{Cov}\left(\nabla F(\boldsymbol{x}^*;\zeta)\right)\boldsymbol{L}^{\dagger}\right), \tag{1.2}$$

where $\boldsymbol{L} = \boldsymbol{P}_J \nabla^2 \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)\boldsymbol{P}_J$, $\boldsymbol{P}_J = \boldsymbol{I} - \boldsymbol{J}^\top\left(\boldsymbol{J}\boldsymbol{J}^\top\right)^{\dagger}\boldsymbol{J}$ is the projection matrix, $\boldsymbol{J}$ is the Jacobian matrix of active constraints at $\boldsymbol{x}^*$, and $\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is the Lagrangian function at the optimal primal-dual points.

In this era of large-scale data, optimization problems such as Problem (1.1) have wide-ranging applications, including but not limited to signal processing, neural network learning, PDE-constrained optimization, and (often randomized) numerical linear algebra. By introducing auxiliary variables (also called slack variables), general equality- and inequality-constrained problems can be transformed into the form of Problem (1.1). Given this equivalency, the focus of this paper is on developing stochastic optimization algorithms to solve Problem (1.1).

Stochastic optimization algorithms for optimizing an objective $f(\boldsymbol{x})$ have a rich history and can be traced back (at least) to stochastic gradient descent (SGD), which solves Problem (1.1) in an unconstrained setting. While SGD is computationally and storage-efficient, subsequent research has developed and enhanced its global convergence and local asymptotic properties. For instance, Ruppert [62], Polyak and Juditsky[55] introduced the concept of Polyak-Ruppert averaging, achieving asymptotic normality for averaged iterates. Chen et al. [15] proposed the plug-in estimator and developed a more efficient batch-means estimator to approximate the covariance matrix and estimate the corresponding confidence intervals. Anastasiou et al. [2] developed non-asymptotic convergence rates for normal approximation of SGD with Polyak-Ruppert averaging. Leluc and Potier [42] extend the analysis to conditioned SGD, thereby encompassing a broader class of algorithms like Newton's methods and Quasi-Newton's methods.

Newton's methods are often favored over first-order methods like gradient descent, particularly for their faster convergence rates, which are made possible by incorporating (exact or approximate) Hessian information [38, 49, 79]. Beyond theoretical advantages, Newton's methods, in particular randomized versions [61, 75, 76, 78], have exceptional performance in practical applications. For example, Yao et al. [77] introduced AdaHessian, employng an adaptive Newton's methods to speed up deep neural network training; and Liu et al. [43] then used ideas very similar to AdaHessian to develop Sophia, which reduced the computational cost for training large language models. Although efforts have been made to enhance gradient descent-based algorithms by partially extracting Hessian information [1, 12, 13], the unique benefits of Newton's methods continue to make them a focal point of ongoing research.

Sequential Quadratic Programming (SQP) is recognized as a potent method for tackling constrained optimization problems, particularly when dealing with nonlinear constraints. As Nocedal and Wright emphasized in their seminal work [38], SQP stands as one of the most effective techniques for solving such problems in the deterministic setting. In contrast to deterministic SQP methods, which assume full access to the objective $f(\boldsymbol{x})$ as described in [9, 38], our work considers a stochastic objective alongside deterministic constraints, as

formulated in Problem (1.1). This paradigm introduces challenges, as the exact values of the objective function, its gradients, and Hessian matrices are generally inaccessible. While recent research has extended SQP algorithms to stochastic settings [5, 18, 19, 20, 23, 27, 47, 48, 51], these works have focused predominantly on problems with only equality constraints. A more exhaustive literature review will be provided in Section 1.3.

Asymptotic analysis serves as a critical tool for a nuanced understanding of the local behavior of iterates in stochastic algorithms. In the context of constrained optimization, we define the primal-dual solution $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$, especially the primal solution $\boldsymbol{x}^*$, as the optimal solution of the Problem (1.1) with expected objective. The dual variable $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$ corresponds to the equality constraints $\boldsymbol{c}(\boldsymbol{x}) = \mathbf{0}$, the lower-bound box constraints $\boldsymbol{\ell} - \boldsymbol{x} \leq \mathbf{0}$, and the upper-bound box constraints $\boldsymbol{x} - \boldsymbol{u} \leq \mathbf{0}$, respectively. While global convergence results offer a broad understanding of the algorithm's behavior, they often fall short in revealing detailed convergence characteristics, especially in the presence of noisy observations related to the objective $f(\boldsymbol{x})$, gradients, and Hessians. Consider $\{(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})\}$ as the sequence of primal-dual iterates generated by an algorithm for solving Problem (1.1). The statistical inference drawn from $\{(\boldsymbol{x}_k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*, \boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*)\}$ can provide more granular insights. Based on this, we can develop the local asymptotic distributions and estimate associated statistical properties, such as covariance matrices and confidence intervals. The stochastic nature of the iterates is reflected by the asymptotic distribution (especially the confidence interval), provided that the number of iterations is sufficiently large. Such statistical insights offer a quantified measure of confidence and a mechanism to manage uncertainty (and thus inference) in stochastic optimization.

Given these considerations, a natural question is:

- *Can we develop an (asymptotically) optimal algorithm, in Hájek and Le Cam's sense [41, 72], and in a manner analogous to the classical M-estimator, for the constrained stochastic optimization Problem* (1.1)*?*

In this paper, we answer this question in the affirmative. To accomplish this, we introduce a novel stochastic Sequential Quadratic Programming algorithm (Debiased-StoSQP and its refinement version Debiased-StoSQP-v2) to solve Problem (1.1), with global almost sure convergence guarantees. We also develop asymptotic normality results and practical estimators for covariance matrices of the generated iterates. The derived limiting covariance matrix matches that of the M-estimator, as in Equation (1.2), showing that our algorithm Debiased-StoSQP-v2 is asymptotically optimal.

Unlike previous analyses of SGD algorithms that focus on averaged iterates [15, 55], our statistical inference targets the last iterate, rendering our approach more aligned with practical applications. Importantly, the presence of inequality constraints in Problem (1.1) introduces a bias in the solutions of the quadratic subproblems for direction estimates, i.e., the obtained search direction is biased. This happens even when $\nabla f(\boldsymbol{x}; \zeta)$ is an unbiased estimators of $\nabla f(\boldsymbol{x})$. This makes our problem formulation more challenging, compared to problems with only equality constraints [6, 51]. To mitigate the bias, we employ moving averaging techniques for gradient estimation. Our results on the asymptotic normality of iterates establish optimality in terms of the min-max lower bound on the covariance matrix in Hájek and Le Cam's sense [41, 72].

1.1. *Motivating examples.* We now present specific examples from machine learning and statistics that can be cast into the forms of Problem (1.1). We re-emphasize that the general constrained problem can be converted into the form of Problem (1.1) by introducing auxiliary variables, where both forms share the same KKT points (where the first-order optimality condition holds).

1.1.1. *Constrained regression.* In regression models, issues like multicollinearity can lead to unreliable inference results that conflict with both intuition and empirical evidence. One way to mitigate such issues is by incorporating prior information into the model via constraints on the model parameters. For instance, we observed such complexities while working with Poisson regression models for Chicago air pollution and death rate data; further details are discussed in Section 5.4. Constraints can also be an inherent part of the problem formulation itself. For example, in portfolio allocation problems, each entry of $\boldsymbol{x}$ denotes the weight assigned to an asset. Thus, it is common to constrain the estimation within the set $\{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{1}^\top \boldsymbol{x} = 1, \boldsymbol{x} \geq \mathbf{0}\}$. In certain contexts, alternative constraints are imposed for particular purposes, including box constraints $\|\boldsymbol{x}\|_\infty \leq \boldsymbol{u}$ and affine constraints $A\boldsymbol{x} = \boldsymbol{b}$ [25, 26] (e.g., a negative weight signifies shorting the asset). In semiparametric index models, we impose $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2^2 = 1, \boldsymbol{x}_1 > 0\}$ to make models identifiable [50, 52]. In factor analysis, constraints can prevent Heywood cases (i.e., a negative estimate for the variance) [66]. In algorithmic fairness, constraints can prevent classifiers from yielding disparate outcomes based on sensitive features like gender and ethnicity [80]. For a more comprehensive review of constrained regression models, including different types of constraints, we refer the reader to [22, 24, 26, 50, 52, 65, 67].

1.1.2. *Physics-informed machine learning.* In scientific machine learning, models must adhere to domain knowledge, often described by partial differential equations (PDEs) constraints [17, 33, 40, 53]. In specialized network architectures (like neural ODEs [14], physics-informed neural networks [39, 40, 58], and physics-informed DeepONets [73]), constraints derived from PDEs are applied to the network, i.e., the neural network is motivated from the following constrained optimization:

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \mathcal{L}_{\text{data}}(\boldsymbol{x}), \\ \text{s.t.} \ & \mathcal{C}_{\text{PDE}}(\boldsymbol{x}) = \mathbf{0}, \end{aligned}$$

where $\boldsymbol{x}$ represents the neural network parameters; $\mathcal{L}_{\text{data}}(\boldsymbol{x})$ and $\mathcal{C}_{\text{PDE}}(\boldsymbol{x}) = \mathbf{0}$ are the constructed data fitting loss to be minimized and the governing PDEs, respectively. (It is known that failure to enforce these constraints can lead to serious problems/instabilities in such models [33, 40, 53].)

1.1.3. *Adversarial training.* Constraints are also frequently employed in adversarial training scenarios. For example, in the training of Wasserstein generative adversarial networks, constraints on the norm of the network parameters are often imposed to ensure model effectiveness [3]. Various types of constraints have been found to improve the adversarial robustness of the model and reduce its sensitivity to perturbations in the input data [16]. In the context of adversarial attacks, constraints are formulated to ensure that the search space of adversarial examples remains close to the original data samples [30, 68, 81].

1.1.4. *Constrained neural networks.* With advances in computational power and storage, as well as the increasing complexity of problems to solve, the number of parameters in modern neural networks can range from millions to billions or more. This scale is often much larger than the size of available training samples. In such situations, constraints play a crucial role in mitigating overfitting by limiting the flexibility of the network space. One simple yet effective way to improve a neural network's generalization capability is to apply L_2-norm (regularization) constraints on the network parameters. Well-known implementation tricks include batch normalization and layer normalization are practical tools to regularize neural network parameters to avoid gradient vanishing or explosion and accelerate convergence. Additionally, the importance of parameter constraints for neural networks is seen in generalization analysis [54]; and it is common to work with neural networks that are (implicitly) constrained to interpolate the data [35, 36].

1.2. *Our contributions.* Our main contribution is to introduce a stochastic SQP algorithm with averaged gradients for solving the constrained optimization Problem (1.1). We summarize our primary contributions, described in more detail in Sections 2-4, here.

(a) We revisit a standard SQP algorithm (namely, RelaxedSQP, Algorithm 1), which is applicable to deterministic objectives in the form of Problem (1.1), where a relaxation parameter is introduced for the feasibility of the quadratic subproblem. Significantly, we establish a connection between this relaxation scheme and constraint qualifications, providing a deeper understanding of constrained optimization problems.

(b) Building on the RelaxedSQP, we introduce Debiased-StoSQP (Algorithm 2), a stochastic counterpart of RelaxedSQP. To address bias issues and challenges due to inequality constraints, we employ averaged gradients for debiasing. We introduce separate step sizes, denoted by α_k and β_k, for iterative updates and gradient averaging, respectively; and we prove that the KKT residual of the sequence of iterates $\{\boldsymbol{x}_k\}$, along with least square estimates of dual variables, converges to zero almost surely.

(c) We perform statistical inference on the iterates generated by Debiased-StoSQP-v2 algorithm (Algorithm 3), which is a refinement of Algorithm 2. To do so, an averaging scheme is introduced for the Hessian, along with an additional update for dual variables. Under mild conditions, we show almost sure convergence of the dual variables $\{(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})\}$ to their optimal values $\{(\boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)\}$. We establish the asymptotic normality for the iterates,

$$1/\sqrt{\alpha_k^{\min}}(\boldsymbol{x}_k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*, \boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Theta\boldsymbol{\Omega}^*),$$

where $\Theta\boldsymbol{\Omega}^*$ is the Fisher information matrix of the algorithm (more specifically, the covariance matrix revealing the uncertainty of iterates). We achieve asymptotically optimal normality in terms of the covariance matrix, according to the min-max lower bound by Duchi and Ruan [23], in Hájek and Le Cam's sense [41, 72]. We also provide a practical estimator for the unknown covariance matrix $\boldsymbol{\Omega}^*$ and show that $\boldsymbol{\Omega}_k$ is convergent to $\boldsymbol{\Omega}^*$ almost surely, where $\boldsymbol{\Omega}_k$ is the estimation of $\boldsymbol{\Omega}^*$ (details can be found in Section 4). It is a surprising and novel result that the algorithm with averaged gradients can indeed achieve the asymptotic normality.

There are additional details upon which we would like to elaborate. The concept of using a relaxation technique in the SQP subproblem was initially proposed by Powell [56]. He provided an intuitive explanation that the constrained problem is difficult/challenging if the relaxation technique is invalid. In Section 2, we extend this intuition by providing a more rigorous treatment of the relaxation technique, examining it from the perspective of constraint qualification. Unlike the approach in [20], which relies on increasing sample sizes to reduce bias (a computationally expensive and often impractical process), we use moving averaging techniques. These techniques allow for fully stochastic and online settings, and they achieve convergence with just a single sample of $f(\boldsymbol{x}; \zeta)$, for gradient and Hessian estimation.

A key novelty in our approach is that we introduce two different step sizes with different decay rates for updating the iterates $\boldsymbol{x}_k$ and the gradient, respectively. This can be regarded as a "competition" between the iterates and the gradients. Specifically, for the global almost sure convergence and local asymptotic normality to be achieved, it is essential that the gradients converge faster than the iterates. This ensures that the algorithm is driven by the most current and relevant gradients, contributing to its effective performance. Our work develops the almost sure "$\lim$" results that the KKT residual of generated iterates converges to zero almost surely, with the help of the averaging gradient technique and the least squares estimation on dual variables. This provides a more complete analysis than existing work [20],

which only proved the almost sure "$\liminf$" convergence. We also employ statistical inference techniques to gain insights into the locally asymptotic behavior of Debiased-StoSQP-v2 algorithm. Using recent advancements in martingale difference arrays, our asymptotic analysis is developed to deal with the algorithm with averaging gradients where the gradients are highly correlated. By setting $\alpha_k^{\min} = 1/(k+1)$, we can achieve the asymptotically optimal normality

$$\sqrt{k}(\boldsymbol{x}_k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*, \boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^*),$$

in terms of the covariance matrix, as shown in [23]. Our estimator $\boldsymbol{\Omega}_k$ for $\boldsymbol{\Omega}^*$ is more like the plug-in estimator in [15] and we further show that $\boldsymbol{\Omega}_k$ is convergent to $\boldsymbol{\Omega}^*$ almost surely.

Note that, unlike SGD, SQP is a constrained Newton's method, where we calculate the noisy Hessian matrix $\nabla^2 f(\boldsymbol{x}; \zeta)$ in each iteration. Therefore, the plug-in estimator for the covariance matrix does not significantly increase the computational complexity. Previous works studying the asymptotic behavior of algorithms typically rely on the independence of gradients, while we consider averaged gradients that are highly correlated. The technique we apply involves the use of two distinct step sizes with different rates of decay for iterate and gradient updates. This enables us to balance the convergence behavior of both the iterates and the gradients, thereby achieving asymptotic normality even under conditions of gradient averaging. This represents a significant difference from existing methods, and it brings new perspectives into the study of the asymptotic behavior of algorithms. More details can be found in Section 4.

1.3. *Related works.* Constrained optimization problems in stochastic settings have gained increasing attention in recent years. Berahas et al. [6] initiated the study of stochastic SQP algorithms, with a focus on equality-constrained problems. They incorporated an ℓ_1-penalized merit function and adaptive selection mechanisms for both penalty parameters and step sizes to ensure the sufficient decrease of Newton step on the merit function, proving "$\liminf$" convergence for the expectation of the KKT residual. Na and Mahoney [51] extended this line of work by developing an algorithm with inexact subproblem solutions, and they showed the almost sure convergence of the KKT residual based on the sufficient decrease of the exact augmented Lagrangian merit function. An alternative method is the stochastic line search SQP proposed by Na et al. [47], where they adaptively select batch size depending on the decrease of the exact augmented Lagrangian merit function. Their method is more adaptive and powerful than fully stochastic algorithms, due to the growing batch size. However, the stochastic line search method is usually more computationally expensive, and some safeguarded techniques are required in practice, as the batch size cannot grow arbitrarily. Curtis et al. [19] aimed to reduce computational overhead by allowing inexact solutions for the quadratic subproblems, subject to specific termination tests. This approach effectively reduces computational effort, especially in high-dimensional scenarios. Similarly, Na and Mahoney [51] considered the sketch-and-project method in stochastic SQP, a randomized iterative solver introduced in [32], to approximately solve the Newton system in each iteration and reduce the total computation. Berahas et al. [7] also explored variance reduction techniques in gradient approximations, adding robustness to the algorithms at the expense of requiring exact gradient estimations in the outer loops. However, their method still requires exact estimations of gradients, which may be intractable in some applications. Most of the aforementioned works focus on equality-constrained problems, leaving inequality constraints as an area open for further research.

Equality- and inequality-constrained problems pose more formidable challenges than their equality-constrained analogs, particularly due to complications like SQP subproblem infeasibility and solution bias. Recent advancements, such as the method of Na et al. [47], use exact

augmented Lagrangian merit functions, concentrating on identifying each iteration's active set of constraints. However, this approach may impose stringent requirements concerning the linear independence of Jacobians for the active constraints. Alternatively, Curtis et al. [20] introduce an innovative two-stage algorithm. The first stage is designed to improve feasibility by solving a box-constrained, strongly convex quadratic problem. The second stage then zeroes in on optimizing the objective function using quadratic expansion. A comparable two-stage algorithm has been introduced by Qiu and Kungurtsev [57]. The primary distinction between the two methods lies in their handling of stochasticity and step-size selection. Specifically, Curtis et al. mandate increasing the sample size for gradient estimations to ensure convergence and employ adaptive step sizes. In contrast, Qiu and Kungurtsev's approach [57] necessitates a lower bound for the batch size to control gradient uncertainty and uses stochastic line search techniques. Duchi and Ruan [23] have formulated a Riemannian stochastic gradient algorithm that employs dual averaging to address inequality-constrained problems. To guarantee the feasibility of the solution iterates, their method incorporates manifold projections, a technique that tends to be computationally demanding.

There exists an extensive body of work focusing on the statistical properties of SGD and its various adaptations [15, 55]. We begin by reviewing some seminal contributions to the area of SGD. For instance, Toulis and Airoldi [70] introduced the concept of implicit SGD, which achieves asymptotic normality accompanied by an optimal covariance matrix. Mou et al. [45] further contributed by investigating the asymptotic behavior of SGD when fixed step sizes and Polyak-Ruppert averaging are employed in solving linear systems. Duchi and Ruan [23] extended this line of research by developing projected Riemannian SGD and offering statistical inferences for inequality-constrained convex problems. However, the local statistical behavior of stochastic Newton's methods remains relatively unexplored. More recently, Boyer and Godichon-Baggioni [10] turned their focus to the asymptotic normality of an advanced stochastic Quasi-Newton method tailored for regression issues. Na and Mahoney [51] provided a particularly interesting insight by showing that the iterates generated by stochastic SQP in equality-constrained problems tend towards an asymptotic Gaussian distribution with a nearly optimal covariance matrix. The basis for this near-optimality is the min-max lower bounds on the covariance matrices, as proven by Duchi and Ruan [23]. Despite these strides, a significant gap persists in literature concerning local statistical analyses for stochastic algorithms applied to both equality and inequality-constrained problems. Our research bridges this critical lacuna.

1.4. *Structure of the paper.* Our paper is organized as follows. In Section 2, we revisit the concept of constraint relaxation and establish a link with constraint qualifications. Section 3 is devoted to the introduction of the proposed relaxed stochastic SQP method (Debiased-StoSQP, Algorithm 2), where we also derive its global almost-sure convergence properties. In Section 4, we delve into the algorithm's (Debiased-StoSQP-v2, Algorithm 3) asymptotic behavior, establishing both asymptotic normality and the convergence rate of the iterates generated by our method. We also introduce a practical estimator designed for statistical inference. Section 5 presents experimental results, focusing on applications to CUTEst benchmark problems and regression analyses. Throughout the paper, we provide sketches of proofs following the theorems to aid in comprehension, but we defer all detailed proofs to the Appendices.

1.5. *Notations.* Throughout the paper, we use $\|\cdot\|_2$ to denote the 2-norm (Euclidean norm) for vectors and the corresponding spectral norm for matrices. The ∞-norm of a vector, representing the maximal absolute value among its elements, is symbolized as $\|\cdot\|_\infty$. We use boldface capital and lowercase letters (e.g., $\boldsymbol{A}$ and $\boldsymbol{a}$) to denote matrices and vectors respectively. Given a positive integer m, the symbol $[m]$ represents the set $\{1, 2, \cdots, m\}$. If we

know without any ambiguity that $\mathcal{I} \subseteq [m]$, then we define $\mathcal{I}^- := [m] \setminus \mathcal{I}$, which implies that $\mathcal{I} \cup \mathcal{I}^- = [m]$ and $\mathcal{I} \cap \mathcal{I}^- = \emptyset$. Let $\mathcal{I} \subseteq [m]$ be the set of indices and $\boldsymbol{A}$ be an $m \times m$ matrix. Then the notation $\boldsymbol{A}_{\mathcal{I}}$ indicates the submatrix composed by columns of $\boldsymbol{A}$ with corresponding columns indices in $\mathcal{I}$, i.e., $\boldsymbol{A}_{\mathcal{I}} = [\boldsymbol{a}_{i_1}, \boldsymbol{a}_{i_2}, \cdots, \boldsymbol{a}_{i_{|\mathcal{I}|}}]$, where $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m]$, $|\mathcal{I}|$ denotes the number of elements in the set $\mathcal{I}$, and $\mathcal{I} = \{i_1, i_2, \cdots, i_{|\mathcal{I}|}\}$. For an m-dimensional vector $\boldsymbol{a}$ and a set of indices $\mathcal{I}$, we denote $[\boldsymbol{a}]_{\mathcal{I}}$ as the subvector of the vector $\boldsymbol{a}$ with $[\boldsymbol{a}]_{\mathcal{I}} = (a_{i_1}, a_{i_2}, \cdots, a_{i_{|\mathcal{I}|}})^\top$, where $\boldsymbol{a} = (a_1, a_2, \cdots, a_m)^\top$ and $\mathcal{I} = \{i_1, i_2, \cdots, i_{|\mathcal{I}|}\}$. Individual elements of a vector $\boldsymbol{a}$ are expressed as either $(\boldsymbol{a})_i$ or a_i, depending on the context. Unimportant constants are subsumed within the big O notation, $\mathcal{O}(\cdot)$, implying that $f = \mathcal{O}(g)$ if $f \leq C \cdot g$ for some constant $C > 0$. We use $\mathcal{F}_k$ to denote the σ-algebra defined by event $\{\zeta_i\}_{i=0}^k$. The conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{k-1}]$ on ζ_k is abbreviated as $\mathbb{E}_k[\cdot]$. We use $\odot$ to denote the element-wise multiplication between two vectors.

2. Constraints Relaxation and Deterministic SQP Algorithm. In this section, we start by considering the deterministic constrained problem, defined as follows:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^d} \quad & f(\boldsymbol{x}), \\ \text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}, \qquad \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}, \end{aligned} \tag{2.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the objective whose derivatives and Hessian are fully accessible under the deterministic setting, and $\boldsymbol{c} : \mathbb{R}^d \to \mathbb{R}^r$ is the equality constraint. Throughout the paper, we assume that the constraints $\boldsymbol{c}$ are second-order continuously differentiable. The vectors $\boldsymbol{\ell}$ and $\boldsymbol{u}$ define the lower and upper bounds, respectively. Here, we require $-\infty < \boldsymbol{\ell} < \boldsymbol{u} < \infty$ and that the feasible region $\Omega := \{\boldsymbol{x} : \boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}, \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$ is non-empty. At the current iterate $\boldsymbol{x}_k$, the classical SQP algorithm obtains the search direction by solving the following subproblem:

$$\begin{aligned} \min_{\boldsymbol{p} \in \mathbb{R}^d} \quad & \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\ \text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}, \quad \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p} \leq \boldsymbol{u}. \end{aligned} \tag{2.2}$$

This is a quadratic expansion of the objective with a linearization of the constraints.

The solution $\boldsymbol{p}_k$ of Problem (2.2) then serves as the direction for updating the variable $\boldsymbol{x}$ in Problem (2.1), i.e., $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$, where $\alpha_k > 0$ is the step size. Interestingly, even though the feasible region Ω of the original Problem (1.1) is non-empty, the Problem (2.2) may yield an infeasible region that

$$\Omega_k := \{\boldsymbol{p} : \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p} \leq \boldsymbol{u}\} = \emptyset.$$

We demonstrate this through the following example.

EXAMPLE 1. *Consider the following constrained optimization problem*

$$\begin{aligned} \min_{(x,y) \in \mathbb{R}^2} \quad & x + 2x^2 + 3y^2, \\ \textit{s.t.} \quad & c(\boldsymbol{x}) := x^2 + y^2 - 9 = 0, \quad \boldsymbol{\ell} := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \leq \begin{pmatrix} x \\ y \end{pmatrix} \leq \begin{pmatrix} 3 \\ 2 \end{pmatrix} := \boldsymbol{u}. \end{aligned} \tag{2.3}$$

Here, the feasible region $\Omega = \{(x, y) : x^2 + y^2 - 9 = 0, 0 \leq x \leq 3, 0 \leq y \leq 2\}$ *is non-empty. If* $(x_k, y_k) = (2, 1)$*, then the region* $\Omega_k = \{(\Delta x, \Delta y) : -4 + 4\Delta x + 2\Delta y = 0, 0 \leq 2 + \Delta x \leq 3, 0 \leq 1 + \Delta y \leq 2\}$ *is non-empty; but the region* Ω_k *at* $(x_k, y_k) = (1, 1)$ *(i.e.,* $\Omega_k = \{(\Delta x, \Delta y) : -7 + 2\Delta x + 2\Delta y = 0, 0 \leq 1 + \Delta x \leq 3, 0 \leq 1 + \Delta y \leq 2\}$*) is empty.*

Fortunately, constraint relaxation provides an effective approach to circumvent this issue of infeasibility in the SQP subproblem. Specifically, we can relax the constraints by introducing a factor $\theta_k \in (0, 1]$, resulting in a relaxed feasible region, defined as

$$\widetilde{\Omega}_k := \{\boldsymbol{p} : \theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p} \leq \boldsymbol{u}\}.$$

For certain $\theta_k \in (0, 1]$, this relaxed feasible region $\widetilde{\Omega}_k$ can be non-empty, even when Ω_k is empty. To illustrate, recall Example 1, where the SQP subproblem yielded an empty region Ω_k at $(x_k, y_k) = (1, 1)$. Introducing the relaxation factor $\theta_k = \frac{1}{2}$ makes the relaxed feasible region $\widetilde{\Omega}_k$ non-empty. This constraint relaxation strategy was originally proposed by Powell [56], whose insight was that the absence of a suitable relaxation parameter signifies that the nonlinear constraints cannot be locally improved in a first-order sense (e.g., linearization).

Naturally, this leads us to investigate the conditions under which a relaxation parameter exists or fails to exist. We found that this is intimately tied to the extended generalized Mangasarian-Fromowitz constraint qualification (EGMFCQ, as defined in Definition 2.4 of [74]). Constraint qualifications serve as conditions that assess the compatibility between nonlinear constraints and their linear approximations. When these qualifications are not met, the linear approximations are inadequate to capture the local geometric properties of the nonlinear constraints. We demonstrate that if $\boldsymbol{x}_k$ moves away from points where EGMFCQ is violated, then $\widetilde{\Omega}_k$ is feasible for some $\theta_k \in (0, 1]$. The relationship between constraint relaxation and constraint qualifications is elaborated further in Appendix A.

DEFINITION 1 (EGMFCQ, Definition 2.4 in [74]). *The extended generalized Mangasarian-Fromowitz constraint qualification (EGMFCQ) is said to be satisfied at a point $\bar{\boldsymbol{x}} \in \mathbb{R}^d$, with respect to the equality constraints $\boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{0}$ and the box constraints $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$, if the following conditions are met:*

- *there is a vector $\boldsymbol{z} \in \mathbb{R}^d$ such that*

$$\begin{aligned} &\boldsymbol{c}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \boldsymbol{z} = \boldsymbol{0}, \\ &(\boldsymbol{z})_i > 0, \ \textit{if } (\bar{\boldsymbol{x}})_i = (\boldsymbol{\ell})_i, \\ &(\boldsymbol{z})_i < 0, \ \textit{if } (\bar{\boldsymbol{x}})_i = (\boldsymbol{u})_i; \end{aligned} \tag{2.4}$$

- *columns of $\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})$ are linearly independent.*

Remark. It is not difficult to verify that EGMFCQ is weaker than linear independence constraint qualification (LICQ, Definition 2 in Appendix A.1) [38, Definition 12.4]. Note that the point $\bar{\boldsymbol{x}}$ does not necessarily satisfy the equality constraints $\boldsymbol{c}(\boldsymbol{x}) = 0$, but it is required to lie within the box constraints $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$.

LEMMA 1. *For Problem (1.1) and the current iterate $\boldsymbol{x}_k$, if EGMFCQ is satisfied at $\boldsymbol{x}_k$, then the relaxed feasible region $\widetilde{\Omega}_k$ is nonempty for some $\theta_k \in (0, 1]$. Moreover, let θ_k be selected within the interval $(0, 1]$ such that $\widetilde{\Omega}_k$ is nonempty with this θ_k but becomes empty when the relaxation parameter θ_k is replaced by $\min\{1.1\theta_k, 1\}$. This selection of θ_k can always be achieved. If $\liminf_{k\to\infty} \theta_k = 0$, then there exists an accumulation point $\boldsymbol{x}^*$ of the sequence $\{\boldsymbol{x}_k\}$ where EGMFCQ fails to hold at $\boldsymbol{x}^*$.*

Remark. Here, θ_k is selected such that it approximates the maximal relaxation parameter to make the relaxed region $\widetilde{\Omega}_k$ feasible. As indicated by Lemma 1, if the relaxation parameter is not uniformly lower-bounded, a subsequence of $\{\boldsymbol{x}_k\}$ will converge to a non-EGMFCQ point. In light of this, we assume that the maximal relaxation parameter remains

lower-bounded throughout the iterative process, as in Assumption 1 (below). At the beginning of each iteration, we first examine the relaxation parameter to ensure that it exceeds a predefined threshold. Failing this, we deduce that the current point is approaching a non-EGMFCQ point, implying that the nonlinear constraints may not be effectively approximated by linearization.

ASSUMPTION 1. *For the iterates $\{\boldsymbol{x}_k\}$ generated by the algorithm, there exists $\tilde{\theta} \in (0,1]$ such that the relaxed feasible region $\widetilde{\Omega}_k$ with $\theta_k \leq \tilde{\theta}$ is always nonempty.*

Remark. We would like to discuss more the search direction derived from Problem (2.2). It is possible to contemplate a more cost-effective algorithm, where the search direction is first obtained from the equality-constrained QP subproblem, followed by applying projections to the box-constrained regions. However, the following example illustrates that this approach can potentially lead to a stagnation at a specific point, despite being under MFCQ/LICQ conditions.

EXAMPLE 2. *Consider the following QP problem*

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathbb{R}^3} \quad & \boldsymbol{g}^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}\boldsymbol{p}, \\ \textit{s.t.} \quad & 1 + \boldsymbol{J}^\top \boldsymbol{p} = 0, \ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \leq \boldsymbol{x} + \boldsymbol{p} \leq \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \end{aligned}$$

where $\boldsymbol{g} = (-1,0)^\top$, $\boldsymbol{B} = \boldsymbol{I}_2$, $\boldsymbol{J} = (1,1)^\top$ and $\boldsymbol{x} = (1,0)^\top$. If we first solve the equality-constrained problem, the search direction is $\boldsymbol{p} = (0,-1)^\top$, then the projection $P(\boldsymbol{x} + \alpha\boldsymbol{p}) = \boldsymbol{x}$ gets stuck for any $\alpha > 0$.

Algorithm 1 RelaxedSQP

Input: $\boldsymbol{\ell} \leq \boldsymbol{x}_0 \leq \boldsymbol{u}$, $\tau, \tilde{\tau} \in (0,1)$, $\sigma \in (0,1)$, $\rho_{-1} > 0$, $\epsilon > 0$, $\beta \in (0,1)$.

for $k = 0, 1, 2, \cdots$ **do**
  $\theta_k = 1$;
  while $\widetilde{\Omega}_k$ with θ_k is empty **do**
    $\theta_k = \theta_k \cdot \tilde{\tau}$;
  end while
  Compute an positive definite Hessian matrix $\boldsymbol{B}_k$ and the gradient $\nabla f(\boldsymbol{x}_k)$.
  Solve the relaxed SQP Subproblem (2.7) with θ_k, where the solution is denoted as $\boldsymbol{p}_k$;
  Let

$$\rho_k^{\text{trial}} = \begin{cases} 0, & \text{if } -\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k \geq 0, \\ \dfrac{\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k}{(1-\sigma)\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2}, & \text{otherwise;} \end{cases} \tag{2.5}$$

  and

$$\rho_k = \begin{cases} \rho_{k-1}, & \text{if } \rho_k^{\text{trial}} \leq \rho_{k-1}, \\ (1+\epsilon)\rho_k^{\text{trial}}, & \text{otherwise;} \end{cases} \tag{2.6}$$

  $\alpha_k = 1$;
  while $\phi(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \rho_k) > \phi(\boldsymbol{x}_k, \rho_k) - \beta\alpha_k \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k; \rho_k)$ **do**
    $\alpha_k = \alpha_k \cdot \tau$;
  end while
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$
end for

Now, we are ready to deal with the standard SQP Subproblem (2.2), which is probably infeasible as discussed, by including a relaxation parameter $\theta_k \in (0, 1]$ in each iteration. More specifically, throughout the paper, instead of solving Problem (2.2) for the search direction in each iteration, we alternatively focus on the following relaxed SQP subproblem:

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathbb{R}^n} \quad & \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\ \text{s.t.} \quad & \theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}, \\ & \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}, \end{aligned} \tag{2.7}$$

for some $\theta_k \in (0, 1]$. We assume that Assumption 1 consistently holds for iterates generated by Algorithms 1-3.

We next offer a concise overview of the line-search technique incorporated into SQP for constrained optimization. To do so, we adopt the ℓ_2 regularized merit function, defined as

$$\phi(\boldsymbol{x}; \rho) := f(\boldsymbol{x}) + \rho \|\boldsymbol{c}(\boldsymbol{x})\|_2, \tag{2.8}$$

to perform the backtracking line search. We define the expanded merit function at $\boldsymbol{x}_k$ with step $\boldsymbol{p}_k$ as

$$q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \rho_k) := f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \rho_k \|\boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p}_k\|_2, \tag{2.9}$$

which combines the second-order approximation of the objective with the first-order linearization of constraints, and the corresponding improvement

$$\begin{aligned} &\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \rho_k) \\ &\quad := q(\boldsymbol{x}_k, \boldsymbol{0}, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \rho_k) - q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \rho_k) \\ &\quad = -\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \rho_k \left(\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 - \left\| \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p}_k \right\|_2 \right) \\ &\quad = -\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \rho_k \theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2, \end{aligned} \tag{2.10}$$

where the last equality comes from the equality constraints of the relaxed SQP Subproblem, Problem (2.7). To make sufficient improvement, we let $\rho_k > 0$ to be large enough such that $\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k; \rho_k) \ge \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \sigma \rho_k \theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$ for some $\sigma \in (0, 1)$, i.e., we introduce the following strategy

$$\rho_k^{\text{trial}} = \begin{cases} 0, & \text{if } -\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k \ge 0, \\ \frac{\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k}{(1-\sigma)\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2}, & \text{otherwise;} \end{cases}$$

and

$$\rho_k = \begin{cases} \rho_{k-1}, & \text{if } \rho_k^{\text{trial}} \le \rho_{k-1}, \\ (1+\epsilon)\rho_k^{\text{trial}}, & \text{otherwise;} \end{cases}$$

for some $\epsilon > 0$. Here, the strategy guarantees that the sequence $\{\rho_k\}$ is monotonically increasing and sufficient improvement is secured. We summarize this algorithm in Algorithm 1.

In addition to the feasibility assumption (Assumption 1) on the constraints, the smoothness and boundedness assumptions on the objective and constraints, as stated in Assumption 2, are standard for convergence analysis.

ASSUMPTION 2. *The objective function f and the constraints $\boldsymbol{c}$ are second-order continuously differentiable. Then for all $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$, there exist $M_{\nabla f}, M_{\boldsymbol{\ell},\boldsymbol{u}}, \kappa_{\nabla f}, \kappa_{\nabla c} > 0$ such that*

$$\|\boldsymbol{u} - \boldsymbol{\ell}\|_2 = M_{\boldsymbol{\ell},\boldsymbol{u}}, \ \ \|\nabla f(\boldsymbol{x})\|_2 \leq M_{\nabla f},$$

and for all $\boldsymbol{\ell} \leq \boldsymbol{x}, \boldsymbol{y} \leq \boldsymbol{u}$ one has

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \kappa_{\nabla f} \|\boldsymbol{x} - \boldsymbol{y}\|_2, \|\nabla \boldsymbol{c}(\boldsymbol{x}) - \nabla \boldsymbol{c}(\boldsymbol{y})\|_2 \leq \kappa_{\nabla c} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

The approximate Hessian matrix $\boldsymbol{B}_k$ is positive definite, i.e., $\kappa_1 \mathbf{I} \preceq \boldsymbol{B}_k \preceq \kappa_2 \mathbf{I}$ for some $0 < \kappa_1 \leq \kappa_2$. The Lagrangian multipliers $\{(\boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_{1,k}^{sub}, \boldsymbol{\mu}_{2,k}^{sub})\}$ for the Problem (2.7) are bounded, i.e., there exist $M_{Lag} > 0$ such that

$$\max\{\|\boldsymbol{\lambda}_k^{sub}\|_2, \|\boldsymbol{\mu}_{1,k}^{sub}\|_2, \|\boldsymbol{\mu}_{2,k}^{sub}\|_2\} \leq M_{Lag}.$$

Here, the boundedness assumption for the Lagrangian multipliers guarantees that the penalty parameter ρ_k is upper bounded, as illustrated by existing literature [8, 11]. Potential concerns regarding the boundedness of Lagrange multipliers are addressed by invoking the EGMFCQ condition. This reveals that the Lagrange multipliers for the SQP subproblems are indeed bounded under EGMFCQ condition. A detailed exposition of this can be found in Appendix A.2, with references [11, 29].

Before delving into the properties of RelaxedSQP (Algorithm 1), it is imperative to articulate the Karush-Kuhn-Tucker (KKT) optimality conditions, as well as the associated KKT residual specific to Problem (1.1). The KKT condition and the corresponding KKT residual for Problem (1.1) at $\boldsymbol{x}$ is formalized as

$$\begin{aligned} &\nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\lambda} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 = \mathbf{0}, \\ &\boldsymbol{c}(\boldsymbol{x}) = \mathbf{0}, \quad \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}, \\ &\boldsymbol{\mu}_1^\top (\boldsymbol{x} - \boldsymbol{\ell}) = 0, \boldsymbol{\mu}_2^\top (\boldsymbol{x} - \boldsymbol{u}) = 0, \\ &\boldsymbol{\mu}_1 \geq \mathbf{0}, \quad \boldsymbol{\mu}_2 \geq \mathbf{0}, \end{aligned} \tag{2.11}$$

and $$\boldsymbol{R}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \begin{pmatrix} \nabla f(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})\boldsymbol{\lambda} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 \\ \boldsymbol{c}(\boldsymbol{x}) \\ \boldsymbol{\mu}_1 \odot (\boldsymbol{x} - \boldsymbol{\ell}) \\ \boldsymbol{\mu}_2 \odot (\boldsymbol{x} - \boldsymbol{u}) \end{pmatrix},$$

for some dual variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \in \mathbb{R}^r \times \mathbb{R}^n_+ \times \mathbb{R}^n_+$. Notably, we exclude the inequality constraints $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$ in the residual definition, as they are intrinsically satisfied by the sequences generated via the proposed RelaxedSQP algorithm. Consequently, if the sequence $\{\boldsymbol{x}_k\}$ with the accompanying Lagrangian multipliers $\{(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})\}$ satisfy $\boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k}) \to \mathbf{0}$, then any accumulation point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$ of $\{(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)\}$ satisfies the KKT condition in Equation (2.11), rendering $\boldsymbol{x}^*$ as a KKT (first-order) optimal point.

THEOREM 1. *Under Assumptions 1 and 2, there exist sufficiently large $\widetilde{K} \in \mathbb{Z}_+$ and $\tilde{\rho} > 0$, such that $\rho_k = \tilde{\rho}$ for all $k \geq \widetilde{K}$ and*

$$\lim_{k\to\infty} \|\boldsymbol{p}_k\|_2 = 0 \text{ and } \lim_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 = 0. \tag{2.12}$$

Furthermore, if we let $(\boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_{1,k}^{sub}, \boldsymbol{\mu}_{2,k}^{sub})$ be the Lagrangian multipliers of Problem (2.7) at $\boldsymbol{x}_k$, then

$$\lim_{k\to\infty} \left\|\boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_{1,k}^{sub}, \boldsymbol{\mu}_{2,k}^{sub})\right\|_2 = 0. \tag{2.13}$$

Sketch of the proof. We begin by establishing that the penalty parameter ρ_k stabilizes after a sufficient number of iterations. Specifically, we show that there exists a sufficiently

large integer $\widetilde{K} \in \mathbb{Z}_+$ and a constant $\tilde{\rho} > 0$, such that $\rho_k = \tilde{\rho}$ for all $k \geq \widetilde{K}$. Capitalizing on the construction of the penalty parameter, we achieve a sufficient improvement in the merit function, formalized as $\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k; \rho_k) \geq \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \sigma \rho_k \theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$. The lower-boundedness of the merit function, in conjunction with this sufficient improvement, ensures global convergence. This rationale is analogous to the convergence analysis in gradient descent and Newton's methods in unconstrained problems. While unconstrained optimization techniques ensure convergence by achieving a sufficient reduction in the objective function, constrained optimization employs a merit function that amalgamates both the objective function and constraint violations to achieve a similar result. Complete details are provided in Appendix A.3.

The RelaxedSQP specializes to the conventional SQP if the unit relaxation parameter (i.e., $\theta_k = 1$) is accepted, under the condition that $\boldsymbol{x}_k$ is close enough to a feasible and EGMFCQ point. Remarkably, the RelaxedSQP algorithm achieves superlinear local convergence, akin to the classical SQP method, under mild conditions. We do not elaborate on the local superlinear convergence of classical SQP algorithms, which are well-studied and can be found in [9, 38, 44, 63, 71]. Our principal focus here is on the intriguing behavior associated with the acceptance of unit relaxation parameters, which is given in the following lemma. The detailed proof is included in Appendix A.4.

LEMMA 2. *Suppose that Assumptions 1 and 2 hold, and* $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^*$, *where* $\boldsymbol{c}(\boldsymbol{x}^*) = \boldsymbol{0}$ *and EGMFCQ condition holds at* $\boldsymbol{x}^*$. *Then the unit relaxation parameter* $\theta_k = 1$ *will be accepted when* k *is sufficiently large.*

3. The Stochastic SQP Algorithm. In this section, we developed a stochastic optimization algorithm with almost sure convergence guarantees (Debiased-StoSQP, Algorithm 2) for solving Problem (1.1). To accomplish this, we first modify the deterministic SQP algorithm (presented in Algorithm 1) into a fully stochastic algorithm (see Algorithm 2), where the averaging gradient is used to reduce the biasedness introduced by the uncertainty in the SQP subproblem. Initially, we will establish the global almost sure "$\liminf$" convergence for the iterates. Subsequently, we extend these results to achieve the almost sure "$\lim$" convergence by incorporating the least squares estimates of dual variables. Before analyzing the convergence performances of Debiased-StoSQP (Algorithm 2), we describe the algorithm the following steps:

- **Step 1:** Selection of relaxation parameter. The relaxation parameter is initialized to be 1 for k-th iterate $\boldsymbol{x}_k$, i.e., $\theta_k = 1$. The feasibility of the region $\widetilde{\Omega}_k$ with θ_k is then assessed. The relaxation parameter θ_k is adjusted iteratively by scaling it down by a factor $\tilde{\tau} \in (0, 1)$, i.e., $\theta_k \leftarrow \theta_k \cdot \tilde{\tau}$, until $\widetilde{\Omega}_k$ is confirmed to be feasible. Under Assumption 1, a suitable relaxation parameter θ_k can be found after at most $\lceil \log_{\tilde{\tau}} \tilde{\theta} \rceil$ steps. In practice, we include $\tilde{\theta} \in (0, 1]$ as a tolerance in the algorithm for θ_k. There are various ways to verify the feasibility of $\widetilde{\Omega}_k$ with θ_k. A direct and practical way is to solve the following convex quadratic problem:

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \left\| \theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} \right\|_2^2, \\ \text{s.t.} \quad & \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p} \leq \boldsymbol{u}, \end{aligned}$$

  and verify whether the minimal objective is zero. Projected gradient descent and projected Newton's methods are popular and efficient solvers for the box-constrained quadratic problem. If the algorithm terminates with $\theta_k < \tilde{\theta}$, where the small $\tilde{\theta} \in (0, 1]$ is the tolerance included in Algorithm 2, then the iterate $\boldsymbol{x}_k$ approaches an undesirable point, where EGMFCQ does not hold.

- **Step 2:** Derivative and Hessian estimation. Since the exact derivative and Hessian are inaccessible at $\boldsymbol{x}_k$, we obtain an estimated derivative $\boldsymbol{g}_k := \nabla f(\boldsymbol{x}_k; \zeta_k)$ and approximated Hessian $\boldsymbol{B}_k$. Note that $\boldsymbol{B}_k$ here is an approximation to the Hessian of the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) := f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{c}(\boldsymbol{x}) + \boldsymbol{\mu}_1^\top (\boldsymbol{\ell} - \boldsymbol{x}) + \boldsymbol{\mu}_2^\top (\boldsymbol{x} - \boldsymbol{u})$ at the primal variable $\boldsymbol{x}_k$ and the estimated dual variable $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})$. A practical and cheap way to calculate $\boldsymbol{B}_k$ follows

$$\boldsymbol{B}_k = \nabla^2 f(\boldsymbol{x}_k; \zeta_k) + \sum_{j=1}^{r} (\boldsymbol{\lambda}_{k-1}^{\text{sub}})_j \nabla^2 c_j(\boldsymbol{x}_k) + \boldsymbol{\Delta}_k, \tag{3.1}$$

  where $\boldsymbol{\lambda}_{k-1}^{\text{sub}}$ is the dual variable of the SQP subproblem at $(k-1)$-th iteration and $\boldsymbol{\Delta}_k$ is a regularizer to make $\boldsymbol{B}_k$ positive definite. To achieve the first-order optimality convergence, $\boldsymbol{B}_k$ is not necessarily an accurate approximation to $\nabla^2 \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})$. In fact, $\boldsymbol{B}_k$ is required to be positive definite to achieve a sufficient decrease direction, i.e., $\kappa_1 \mathbf{I} \preceq \boldsymbol{B}_k \preceq \kappa_2 \mathbf{I}$ for some $0 < \kappa_1 \leq \kappa_2$. For example, the approximate Hessian $\boldsymbol{B}_k$ is set to an identity matrix in [20]. To achieve the local "optimal" convergence, we expect $\boldsymbol{B}_k$ to be an accurate approximation to the Hessian of Lagrangian, and the averaging technique is employed to reduce the noise of the stochasticity. For more details, see Lemma 4, where we show the almost sure convergence of $\boldsymbol{B}_k$ to the exact Hessian of Lagrangian $\boldsymbol{B}^*$ (defined later). In this part for the first-order global convergence, only the positive-definiteness of $\boldsymbol{B}_k$ is enforced. Different from [20], where the estimated derivative $\boldsymbol{g}_k$ is directly used in the SQP subproblem, we apply the averaging technique for reducing the noise, i.e., $\bar{\boldsymbol{g}}_k = \bar{\boldsymbol{g}}_{k-1} + \beta_k(\boldsymbol{g}_k - \bar{\boldsymbol{g}}_{k-1})$. The averaging of derivatives is essential to inequality-constrained problems, in both theoretical convergence and experimental performance. It is not difficult to verify that the averaged gradients converge to the exact gradient, at least in expectation, i.e., $\lim_{k\to\infty} \mathbb{E}\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] = 0$, under some mild conditions. However, the simple estimated gradient is not close to the exact gradient, i.e., $\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] = \mathcal{O}(1)$. Without the averaging of derivatives, [20] achieves global convergence by reducing the noise level manually, i.e., increasing the sample size during the iterations. However, our algorithms (Debiased-StoSQP and its variant Debiased-StoSQP-v2) are still fully stochastic, i.e., the derivative estimate is only required to have bounded variance, and the noise level is reduced by the imposed averaging.
- **Step 3:** Obtaining the direction from SQP subproblem. Equipped with the estimated derivative $\bar{\boldsymbol{g}}_k$, the approximate Hessian $\boldsymbol{B}_k$ and the relaxation parameter θ_k, we acquire the search direction $\bar{\boldsymbol{p}}_k$ as a solution of the following SQP subproblem

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathbb{R}^n} \quad & \bar{\boldsymbol{g}}_k^\top \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\ \text{s.t.} \quad & \theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \mathbf{0}, \quad \boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p} \leq \boldsymbol{u}. \end{aligned} \tag{3.2}$$

  Here, the direction $\bar{\boldsymbol{p}}_k$ is "descent" for the merit function, owing to the convergence of $\bar{\boldsymbol{g}}_k$ to $\nabla f(\boldsymbol{x}_k)$ and the positive-definiteness of $\boldsymbol{B}_k$.
- **Step 4:** Adaptive step size selection. We first require that the pre-defined step size $\{\gamma_k\}$ decays (asymptotically) polynomially, i.e., $\gamma_k = \iota_0 (k+1)^{-b_1}$ for some $\iota_0 > 0$ and $b_1 \in (0, 1]$. The strategy is similar to the adaptive strategy in the deterministic algorithm. We alternatively select ρ_k such that $\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k)$ enjoys the sufficient decrease, i.e., $\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k) \geq \frac{1}{2} \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \sigma \rho_k \theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$ for some $\sigma \in (0, 1)$. The adaptivity parameter $\xi_k \leq \xi_k^{\text{trial}} := \frac{\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k)}{\|\bar{\boldsymbol{p}}_k\|_2^2}$ measures the quality of the direction $\bar{\boldsymbol{p}}_k$ in reducing the merit function. If ξ_k is large, which implies that $\bar{\boldsymbol{p}}_k$ is probably a promising direction, then a more aggressive step size $\alpha_k \propto \xi_k \gamma_k$ is preferred, and vise versa. We

select the step size $\alpha_k \in \left[\alpha_k^{\min}, \alpha_k^{\max}\right] := \left[\frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}}, \frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2\right]$, where $\kappa_{\nabla f}$ and $\kappa_{\nabla c}$ are Lipschitz constant for ∇f and $\nabla \boldsymbol{c}$, respectively. We may efficiently estimate the Lipschitz constants by finite differences, an idea quite similar to the Armijo condition.

- **Step 5:** Updating the variable. The primal variable is updated as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \bar{\boldsymbol{p}}_k$. Notably, the algorithm does not necessitate the explicit use of dual variables for updating the primal variable, thus omitting an update scheme for the dual variables in this section. However, the "optimal" local convergence requires an accurate approximation of $\boldsymbol{B}_k$ to the Hessian of the Lagrangian function. This, in turn, demands a satisfactory estimation of dual variables. An update scheme for these dual variables, which is crucial for examining the local convergence properties of the iterates, is provided in Equation (4.1) in Section 4.

Algorithm 2 Debiased-StoSQP

Input: $\boldsymbol{\ell} \le \boldsymbol{x}_0 \le \boldsymbol{u}$, $\tau, \tilde{\tau} \in (0,1)$, $\sigma \in (0,1)$, $\rho_{-1} > 0$, $\epsilon_\rho, \epsilon_\xi, \beta \in (0,1)$, $\mu \in (0,1)$, $\varrho > 0$, $\{\beta_k\}_{k=0}^{\infty}$, $\{\gamma_k\}_{k=0}^{\infty}$.

for $k = 0, 1, 2, \cdots$ **do**
  (Step 1.) $\theta_k = 1$;
  while $\widetilde{\Omega}_k$ with θ_k is empty **do**
    $\theta_k = \theta_k \cdot \tilde{\tau}$;
  end while
  (Step 2.) Compute a positive definite approximate Hessian matrix $\boldsymbol{B}_k$ and the estimated gradient $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k; \zeta_k)$;
  Let

$$\bar{\boldsymbol{g}}_k = \bar{\boldsymbol{g}}_{k-1} + \beta_k(\boldsymbol{g}_k - \bar{\boldsymbol{g}}_{k-1});$$

  (Step 3.) Solve the relaxed SQP Subproblem (2.7) with θ_k, $\boldsymbol{B}_k$ and $\bar{\boldsymbol{g}}_k$, where the solution is denoted as $\bar{\boldsymbol{p}}_k$;
  (Step 4.) Let

(3.3)
$$\rho_k^{\text{trial}} = \begin{cases} 0, & \text{if } -\bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k - \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k \ge 0, \\ \frac{\bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k + \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k}{(1-\sigma)\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2}, & \text{otherwise;} \end{cases} \quad \text{and} \quad \rho_k = \begin{cases} \rho_{k-1}, & \text{if } \rho_k^{\text{trial}} \le \rho_{k-1}, \\ (1+\epsilon_\rho)\rho_k^{\text{trial}}, & \text{otherwise;} \end{cases}$$

  Let

(3.4)
$$\xi_k^{\text{trial}} = \frac{\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k)}{\|\bar{\boldsymbol{p}}_k\|_2^2}, \text{ and } \quad \xi_k = \begin{cases} \xi_{k-1}, & \text{if } \xi_{k-1} \le \xi_k^{\text{trial}}, \\ \min\{(1-\epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\}, & \text{otherwise;} \end{cases}$$

  Select $\alpha_k \in \left[\alpha_k^{\min}, \alpha_k^{\max}\right] := \left[\frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}}, \frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2\right]$;
  $\alpha_k = \min\{\alpha_k, 1/\theta_k\}$;
  (Step 5.) $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \bar{\boldsymbol{p}}_k$.
end for

We make the following two key assumptions regarding the gradient estimate: that it is unbiased; and that it has bounded variance. Similar to the deterministic algorithm, we assume that the penalty parameter becomes stable after a sufficient number of iterations, in line with existing literature [6, 57]. See Assumption 3 (below). In addition, in Assumption 4 (below), we impose an additional condition stipulating that this stabilized penalty parameter must be sufficiently large, when compared to the corresponding parameter in deterministic algorithms. A similar assumption also appears in [6] for the convergence of the fully stochastic algorithms.

ASSUMPTION 3. *Suppose that $\boldsymbol{g}_k := \nabla f(\boldsymbol{x}_k; \zeta_k)$ is a unbiased estimate of $\nabla f(\boldsymbol{x}_k)$, i.e., $\mathbb{E}_k[\boldsymbol{g}_k] = \mathbb{E}_\zeta[\nabla f(\boldsymbol{x}_k; \zeta)|\mathcal{F}_{k-1}]$, and there exists a positive number $\sigma_g > 0$ such that*

$\mathbb{E}_k \|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|_2^2 \leq \sigma_g^2$. *We further assume that the penalty parameter becomes stable after* $\bar{K}$ *iterations, i.e.,* $\rho_k = \bar{\rho}$, $\forall k \geq \bar{K}$.

ASSUMPTION 4. *Suppose that the stable penalty parameter* $\bar{\rho}$ *for the stochastic algorithm is sufficiently large such that* $\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k; \bar{\rho}) \geq \frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \sigma\bar{\rho}\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$ *holds for some* $\sigma \in (0,1)$ *and for all* $k \geq \bar{K}$.

Under these two additional assumptions, we have the following theorem. This theorem is an intermediate result that the KKT residual of the iterates obtained by the proposed Debiased-StoSQP algorithm achieves the "$\liminf$" convergence. The detailed proofs are available in Appendix B.

THEOREM 2. *Under Assumptions 1, 2 and 3, if* $\alpha_k^{min} = \iota_1(k+1)^{-b_1}$ *and* $\beta_k = \iota_2(k+1)^{-b_2}$ *for some* $\iota_1, \iota_2 > 0$ *and some* b_1, b_2 *satisfying* $b_1 \in (\frac{3}{4}, 1]$ *and* $b_2 \in (2-2b_1, 2b_1-1)$, *then*

$$\liminf_{k\to\infty} \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k; \bar{\rho}) = 0, \text{ almost surely}.$$

If we further assume that Assumption 4 holds, then

$$\liminf_{k\to\infty} \left[\|\boldsymbol{p}_k\|_2^2 + \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2\right] = 0, \text{ almost surely}.$$

Furthermore, let $(\boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_{1,k}^{sub}, \boldsymbol{\mu}_{2,k}^{sub})$ *be the Lagrangian multipliers of Problem* (2.7) *at* $\boldsymbol{x}_k$ *with full gradient* $\nabla f(\boldsymbol{x}_k)$, *then*

$$\liminf_{k\to\infty} \left\|\boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k^{sub}, \boldsymbol{\mu}_{1,k}^{sub}, \boldsymbol{\mu}_{2,k}^{sub})\right\|_2 = 0, \text{ almost surely}. \tag{3.5}$$

On top of this result, we also aim to enhance this "$\liminf$" convergence to "$\lim$" convergence by employing the least squares estimates of dual variables, rather than Lagrangian multipliers obtained from subproblems. Given an iterate $\boldsymbol{x}_k$, the Lagrangian multipliers can be determined from the following least-square optimization problem

$$\min_{\boldsymbol{\lambda},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2} F(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}) = \left\|\nabla f(\boldsymbol{x}) + \nabla c(\boldsymbol{x})^\top \boldsymbol{\lambda} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right\|_2^2 + \|\boldsymbol{\mu}_1 \odot (\boldsymbol{x} - \boldsymbol{\ell})\|_2^2 + \|\boldsymbol{\mu}_2 \odot (\boldsymbol{x} - \boldsymbol{u})\|_2^2, \tag{3.6}$$

s.t. $\boldsymbol{\mu}_1 \geq \boldsymbol{0}, \boldsymbol{\mu}_2 \geq \boldsymbol{0}$.

The estimated optimal Lagrangian multipliers $(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*)$ corresponding to $\boldsymbol{x}_k$ serve as one of the feasible solutions of Problem (3.6) evaluated at $\boldsymbol{x}_k$. The detailed proofs can be found in Appendix B.

THEOREM 3. *Under Assumptions 1, 2, 3 and 4, if* $\alpha_k^{min} = \iota_1(k+1)^{-b_1}$ *and* $\beta_k = \iota_2(k+1)^{-b_2}$ *for some* $\iota_1, \iota_2 > 0$ *and some* b_1, b_2 *satisfying* $b_1 \in (\frac{3}{4}, 1]$ *and* $b_2 \in (2-2b_1, 2b_1-1)$, *then we have*

$$\lim_{k\to\infty} \boldsymbol{R}(\boldsymbol{x}_k, \boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*) = \boldsymbol{0}, \text{ almost surely}. \tag{3.7}$$

Practical step size selection. In line 10 of Algorithm 2, the step size α_k is chosen from the interval that $\alpha_k \in \left[\frac{\xi_k\gamma_k}{\kappa_{\nabla f} + \rho_k\kappa_{\nabla c}}, \frac{\xi_k\gamma_k}{\kappa_{\nabla f} + \rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2\right]$. By the definition of the adaptivity parameter $\xi_k \leq \xi_k^{\text{trial}} = \frac{\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \nabla\bar{f}(\boldsymbol{x}_k), \boldsymbol{B}_k, \rho_k)}{\|\bar{\boldsymbol{p}}_k\|_2^2}$ and the step size $\alpha_k \geq \frac{\xi_k\gamma_k}{\kappa_{\nabla f} + \rho_k\kappa_{\nabla c}}$, there is a potential

risk of underestimation, i.e., $\xi_k \ll \frac{\Delta q(\boldsymbol{x}_k,\bar{\boldsymbol{p}}_k,\nabla \bar{f}(\boldsymbol{x}_k),\boldsymbol{B}_k,\rho_k)}{\|\bar{\boldsymbol{p}}_k\|_2^2}$. This occurs especially due to the non-increasing strategy outlined in Equation (3.4) for the construction of the sequence $\{\xi_k\}$. To address this, we introduce additional flexibility in the upper bound of the step size selection by incorporating $\varrho\gamma_k^2$. Defining a more aggressive trial step size $\alpha_k^{\text{trial}} = \frac{\xi_k^{\text{trial}}\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}}$, we project the trail step size into the predefined interval $\left[\frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}}, \frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2\right]$, resulting in

$$\alpha_k = \begin{cases} \alpha_k^{\text{trial}}, & \text{if } \alpha_k^{\text{trial}} \le \frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2, \\ \frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}} + \varrho\gamma_k^2, & \text{otherwise.} \end{cases} \tag{3.8}$$

4. Asymptotic Normality and Convergence Rate. In this section, we further refine Debiased-StoSQP (Algorithm 2) to Debiased-StoSQP-v2 (Algorithm 3) and show the asymptotic normality property, where Debiased-StoSQP-v2 (Algorithm 3) is asymptotically optimal in Hájek and Le Cam's sense, as shown in Equation (1.2). Compared with Debiased-StoSQP (Algorithm 2), in Debiased-StoSQP-v2 (Algorithm 3), we provide detailed update scheme for approximate Hessian matrix $\boldsymbol{B}_k$ (as shown in **Step 2′**) and dual variables (as shown in **Step 6**). Note that Debiased-StoSQP-v2 is a special version of Debiased-StoSQP, therefore, they share all properties developed in Section 3.

Recall that in Algorithm 2, the approximate Hessian matrix can be any bounded and positive definite matrices. Here, to show the optimal local convergence of $\{\boldsymbol{x}_k\}$ to a local minimizer $\boldsymbol{x}^*$, the convergence of the approximate Hessian $\boldsymbol{B}_k$ to the exact Hessian matrix $\nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^r (\boldsymbol{\lambda}^*)_i \nabla^2 \boldsymbol{c}(\boldsymbol{x}^*)$ is essential. Besides the update scheme for the primal variable $\boldsymbol{x}_k$ in Algorithm 2, we must include extra updates for dual variables which are only useful for the calculating the approximate Hessian matrix $\boldsymbol{B}_k$. More specifically, for the current primal-dual variables $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})$, the approximate Hessian matrix $\boldsymbol{B}_k$ is estimated by

$$\boldsymbol{B}_k = \frac{1}{k}\sum_{i=1}^{k}\left(\nabla^2 f(\boldsymbol{x}_i;\zeta_i) + \sum_{j=1}^{r}(\boldsymbol{\lambda}_i)_j \nabla^2 c_j(\boldsymbol{x}_i)\right) + \boldsymbol{\Delta}_k,$$

where $\boldsymbol{\Delta}_k$ is a regularization matrix that guarantees the positive-definiteness of $\boldsymbol{B}_k$. We also include the averaging for the approximate Hessian to reduce the stochasticity and achieve the almost sure convergence (see Lemma 4). We describe it as **Step 2′** in Algorithm 3, since it is a specific case of Step 2 in Algorithm 2. Let $\left(\bar{\boldsymbol{p}}_k, \boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}}\right)$ be the primal-dual solution of the SQP subproblem

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathbb{R}^n} \quad & \bar{\boldsymbol{g}}_k^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\ \text{s.t.} \quad & \theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}, \end{aligned}$$

where $\bar{\boldsymbol{g}}_k$ is the averaged gradient, as in Algorithm 2. Then the dual variables $\boldsymbol{\lambda}_{k+1}, \boldsymbol{\mu}_{1,k+1}, \boldsymbol{\mu}_{2,k+1}$ are obtained by

$$\begin{pmatrix} \boldsymbol{\lambda}_{k+1} \\ \boldsymbol{\mu}_{1,k+1} \\ \boldsymbol{\mu}_{2,k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\lambda}_k \\ \boldsymbol{\mu}_{1,k} \\ \boldsymbol{\mu}_{2,k} \end{pmatrix} + \alpha_k \begin{pmatrix} \boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\lambda}_k \\ \boldsymbol{\mu}_{1,k}^{\text{sub}} - \boldsymbol{\mu}_{1,k} \\ \boldsymbol{\mu}_{2,k}^{\text{sub}} - \boldsymbol{\mu}_{2,k} \end{pmatrix}, \tag{4.1}$$

as in **Step 6** in Algorithm 3. Note that the regularization matrix $\boldsymbol{\Delta}_k$ guarantees that the Algorithm 3 is a specific case of Algorithm 2. Consequently, the almost sure convergence results

established in Section 3 for Algorithm 2 are also applicable to Algorithm 3. The following assumption requires that the LICQ condition, strictly complementary slackness condition, and strongly convexity conditions, namely second-order sufficient conditions (SOSC), are satisfied at the local minimizer. This kind of local condition is commonly considered crucial for analyzing local convergence behaviors, both in unconstrained and constrained optimization problems [38].

Algorithm 3 Debiased-StoSQP-v2

Input: $\boldsymbol{\ell} \le \boldsymbol{x}_0 \le \boldsymbol{u}$, $\tau, \tilde{\tau} \in (0,1)$, $\sigma \in (0,1)$, $\rho_{-1} > 0$, $\epsilon_\rho, \epsilon_\xi, \beta \in (0,1)$, $\mu \in (0,1)$, $\varrho > 0$, $\{\beta_k\}_{k=0}^{\infty}$, $\{\gamma_k\}_{k=0}^{\infty}$.

for $k = 0, 1, 2, \cdots$ **do**
  (Step 1.) $\theta_k = 1$;
  while $\widetilde{\Omega}_k$ with θ_k is empty **do**
    $\theta_k = \theta_k \cdot \tilde{\tau}$;
  end while
  (Step 2′). Compute the estimated gradient $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k; \zeta_k)$ and let

$$\bar{\boldsymbol{g}}_k = \bar{\boldsymbol{g}}_{k-1} + \beta_k(\boldsymbol{g}_k - \bar{\boldsymbol{g}}_{k-1});$$

  Compute the positive definite approximate Hessian matrix $\boldsymbol{B}_k$, i.e.,

$$\boldsymbol{B}_k = \frac{1}{k}\sum_{i=1}^{k}\left(\nabla^2 f(\boldsymbol{x}_i;\zeta_i) + \sum_{j=1}^{r}(\boldsymbol{\lambda}_i)_j \nabla^2 c_j(\boldsymbol{x}_i)\right) + \boldsymbol{\Delta}_k,$$

  where $\boldsymbol{\Delta}_k$ is a regularization matrix that guarantees the positive-definiteness of $\boldsymbol{B}_k$.
  (Step 3.) Solve the relaxed SQP Subproblem (2.7) with θ_k, $\boldsymbol{B}_k$ and $\bar{\boldsymbol{g}}_k$, where the primal-dual solution is denoted as $\left(\bar{\boldsymbol{p}}_k, \boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}}\right)$;
  (Step 4.) Let

$$\rho_k^{\text{trial}} = \begin{cases} 0, & \text{if } -\bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k - \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k \ge 0, \\ \dfrac{\bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k + \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k}{(1-\sigma)\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2}, & \text{otherwise;} \end{cases} \quad \text{and} \quad \rho_k = \begin{cases} \rho_{k-1}, & \text{if } \rho_k^{\text{trial}} \le \rho_{k-1}, \\ (1+\epsilon_\rho)\rho_k^{\text{trial}}, & \text{otherwise;} \end{cases}$$

  Let

$$\xi_k^{\text{trial}} = \frac{\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k)}{\|\bar{\boldsymbol{p}}_k\|_2^2}, \text{ and } \quad \xi_k = \begin{cases} \xi_{k-1}, & \text{if } \xi_{k-1} \le \xi_k^{\text{trial}}, \\ \min\{(1-\epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\}, & \text{otherwise;} \end{cases}$$

  Select $\alpha_k \in \left[\alpha_k^{\min}, \alpha_k^{\max}\right] := \left[\frac{\xi_k \gamma_k}{\kappa_{\nabla f} + \rho_k \kappa_{\nabla c}}, \frac{\xi_k \gamma_k}{\kappa_{\nabla f} + \rho_k \kappa_{\nabla c}} + \varrho \gamma_k^2\right]$;
  $\alpha_k = \min\{\alpha_k, 1/\theta_k\}$;
  (Step 5.) $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \bar{\boldsymbol{p}}_k$.
  (Step 6.) $\begin{pmatrix} \boldsymbol{\lambda}_{k+1} \\ \boldsymbol{\mu}_{1,k+1} \\ \boldsymbol{\mu}_{2,k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\lambda}_k \\ \boldsymbol{\mu}_{1,k} \\ \boldsymbol{\mu}_{2,k} \end{pmatrix} + \alpha_k \begin{pmatrix} \Delta\boldsymbol{\lambda}_k \\ \Delta\boldsymbol{\mu}_{1,k} \\ \Delta\boldsymbol{\mu}_{2,k} \end{pmatrix}$, where $\begin{pmatrix} \Delta\boldsymbol{\lambda}_k \\ \Delta\boldsymbol{\mu}_{1,k} \\ \Delta\boldsymbol{\mu}_{2,k} \end{pmatrix} := \begin{pmatrix} \boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\lambda}_k \\ \boldsymbol{\mu}_{1,k}^{\text{sub}} - \boldsymbol{\mu}_{1,k} \\ \boldsymbol{\mu}_{2,k}^{\text{sub}} - \boldsymbol{\mu}_{2,k} \end{pmatrix}$.
end for

ASSUMPTION 5. *We assume that the generated sequence $\{\boldsymbol{x}_k\}$ is convergent almost surely to a strict local solution $\boldsymbol{x}^*$, where (i) LICQ holds for active constraints at $\boldsymbol{x}^*$; (ii) strictly complementary slackness condition holds, i.e., $(\boldsymbol{\mu}_1^*)_i > 0$ if $(\boldsymbol{x})_i = (\boldsymbol{\ell})_i$, and $(\boldsymbol{\mu}_2^*)_i > 0$ if $(\boldsymbol{x})_i = (\boldsymbol{u})_i$; (iii) $\nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^{r}(\boldsymbol{\lambda}^*)_i \nabla^2 c_i(\boldsymbol{x}^*)$ is positive definite.*

LEMMA 3. *Under Assumptions 2 and 5, the followings hold:*

1. $\boldsymbol{p}_k \to \boldsymbol{0}$ *almost surely, where $\boldsymbol{p}_k$ is the solution of the relaxed SQP subproblem at $\boldsymbol{x}_k$ with exact gradient $\nabla f(\boldsymbol{x}_k)$ and the approximate Hessian matrix $\boldsymbol{B}_k$;*

2. $\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k) \to \mathbf{0}$, *almost surely;*
3. *there exist sufficiently sufficiently large* K^*, *such that* $\mathcal{I}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{I}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{I}(\boldsymbol{x}^*)$ *and* $\mathcal{J}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{J}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{J}(\boldsymbol{x}^*)$, *for* $k \geq K^*$;
4. $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k}) \to (\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$ *almost surely.*

We put the detailed proof in Appendix C.1. In Lemma 3, we show that: (1) the exact search direction $\boldsymbol{p}_k$ converges almost surely to zero, implying that the iterate $\boldsymbol{x}_k$ approximately satisfies KKT conditions; (2) the averaged gradients are arbitrarily close to the exact gradients after a sufficient number of iterations, due to the updating schemes for the averaged gradients and iterates; (3) the active and inactive sets of constraints can be correctly identified; and (4) the primal-dual iterates converge almost surely to the optimal solution. By the correct identification of active and inactive sets in Lemma 3, the KKT condition and the strong convexity of the SQP subproblem further imply that $\boldsymbol{p}_k$ and $\bar{\boldsymbol{p}}_k$ are the solution of the following equality-constrained problems, respectively:

(4.2)

$$\begin{aligned}
\boldsymbol{p}_k = \underset{\boldsymbol{p}\in\mathbb{R}^d}{\arg\min} \quad & \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\
\text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \mathbf{0}, \\
& (\boldsymbol{x}_k + \boldsymbol{p})_i = (\boldsymbol{\ell})_i, \text{ for } i \in \mathcal{I}(\boldsymbol{x}^*), \\
& (\boldsymbol{x}_k + \boldsymbol{p})_i = (\boldsymbol{u})_i, \text{ for } i \in \mathcal{J}(\boldsymbol{x}^*),
\end{aligned}
\qquad \text{and} \qquad
\begin{aligned}
\bar{\boldsymbol{p}}_k = \underset{\boldsymbol{p}\in\mathbb{R}^d}{\arg\min} \quad & \bar{\boldsymbol{g}}_k^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p}, \\
\text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \mathbf{0}, \\
& (\boldsymbol{x}_k + \boldsymbol{p})_i = (\boldsymbol{\ell})_i, \text{ for } i \in \mathcal{I}(\boldsymbol{x}^*), \\
& (\boldsymbol{x}_k + \boldsymbol{p})_i = (\boldsymbol{u})_i, \text{ for } i \in \mathcal{J}(\boldsymbol{x}^*).
\end{aligned}$$

Without the loss of generality, we assume that the relaxation parameter is unit according to Lemma 2. The LICQ condition at $\boldsymbol{x}^*$ also implies the LICQ at $\boldsymbol{x}_k$ when $\boldsymbol{x}_k$ is sufficiently close to $\boldsymbol{x}^*$. Then the KKT system of Problem (4.2) shows that

(4.3)

$$\begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ \left[\Delta\boldsymbol{\mu}_{1,k}\right]_{\mathcal{I}(\boldsymbol{x}^*)} \\ \left[\Delta\boldsymbol{\mu}_{2,k}\right]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_k & \nabla \boldsymbol{c}(\boldsymbol{x}_k) & [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)} & [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)} \\ \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} -\bar{\boldsymbol{g}}_k - \boldsymbol{\lambda}_k \nabla \boldsymbol{c}(\boldsymbol{x}_k) + \boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{2,k} \\ -\boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{x}_k - \boldsymbol{\ell}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{u} - \boldsymbol{x}_k]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix},$$

$$[\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}^-(\boldsymbol{x}^*)} = -[\boldsymbol{\mu}_{1,k}]_{\mathcal{I}^-(\boldsymbol{x}^*)},$$

and

$$[\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}^-(\boldsymbol{x}^*)} = -[\boldsymbol{\mu}_{2,k}]_{\mathcal{J}^-(\boldsymbol{x}^*)},$$

under the almost sure convergence of primal-dual iterates and conditions in Assumption 5. For the parameters and the step size, we only consider the case where the penalty parameter ρ_k and adaptivity parameter ξ_k become stable, and we let $\alpha_k^{\min} = \iota_1(k+1)^{-b_1}$ and $\alpha_k^{\max} = \iota_1(k+1)^{-b_1} + \iota_0(k+1)^{-2b_1}$, where $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$. We denote the Jacobian matrix of the (estimated) KKT system at $\boldsymbol{x}_k$ and $\boldsymbol{x}^*$ as

$$\boldsymbol{H}_k = \begin{pmatrix} \boldsymbol{B}_k & \nabla \boldsymbol{c}(\boldsymbol{x}_k) & [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)} & [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)} \\ \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and

$$\boldsymbol{H}^* = \begin{pmatrix} \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^r (\boldsymbol{\lambda}^*)_i \nabla^2 c_i(\boldsymbol{x}^*) & \nabla \boldsymbol{c}(\boldsymbol{x}^*) & [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)} & [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)} \\ \nabla \boldsymbol{c}(\boldsymbol{x}^*)^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

respectively. Let the core covariance (also the Fisher information matrix of the algorithm) at $\boldsymbol{x}^*$ be defined as

(4.4)
$$\boldsymbol{\Omega}^* = \boldsymbol{H}^{*-1} \begin{pmatrix} \mathbb{E}\left[\nabla f(\boldsymbol{x}^*;\zeta)\nabla f(\boldsymbol{x}^*;\zeta)^\top\right] - \nabla f(\boldsymbol{x}^*)\nabla f(\boldsymbol{x}^*)^\top & \mathbf{0}\,\mathbf{0}\,\mathbf{0} \\ \mathbf{0} & \mathbf{0}\,\mathbf{0}\,\mathbf{0} \\ \mathbf{0} & \mathbf{0}\,\mathbf{0}\,\mathbf{0} \\ \mathbf{0} & \mathbf{0}\,\mathbf{0}\,\mathbf{0} \end{pmatrix} \boldsymbol{H}^{*-1} := \boldsymbol{H}^{*-1}\boldsymbol{\Sigma}\boldsymbol{H}^{*-1}.$$

LEMMA 4. *Under Assumptions 2 and 5, we have $\boldsymbol{B}_k \to \boldsymbol{B}^*$ and $\boldsymbol{H}_k \to \boldsymbol{H}^*$ almost surely, where $\boldsymbol{B}^* := \nabla^2 f(\boldsymbol{x}^*) + \sum_{i=1}^r (\boldsymbol{\lambda}^*)_i \nabla^2 c_i(\boldsymbol{x}^*)$.*

The proof for Lemma 4 can be found in Appendix C.2. According to the almost sure convergence in Assumption 5, and Lemmas 3 and 4, we deduce that there exists a sufficiently large integer K^*, such that the active set of box inequalities remains constant. Specifically, $\mathcal{I}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{I}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{I}(\boldsymbol{x}^*)$ and $\mathcal{J}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{J}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{J}(\boldsymbol{x}^*)$, for $k \geq K^*$. Therefore, we can equivalently consider the equality-constrained SQP subproblem as given by Problem (4.2) for $k \geq K^*$.

ASSUMPTION 6. *Assume that the random gradient has finite third-moment, i.e., the conditioned expectation*

$$\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|_2^3 \,|\mathcal{F}_{k-1}\right] = \mathbb{E}_\zeta\left[\|\nabla f(\boldsymbol{x}_k;\zeta) - \nabla f(\boldsymbol{x}_k)\|_2^3 \,|\mathcal{F}_{k-1}\right] \leq M_m,$$

for some $M_m > 0$ and all $\boldsymbol{x}_k$ in the feasible region $\{\boldsymbol{x} : \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$. The Lipschitzness of stochastic gradient holds, i.e., $\mathbb{E}\left[\|\nabla f(\boldsymbol{x};\zeta) - \nabla f(\boldsymbol{y};\zeta)\|_2^2\right] \leq \kappa_{\nabla f}^2 \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ for some $\kappa_{\nabla f} > 0$.

THEOREM 4. *Under Assumptions 2, 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, and $b_2 > \frac{1}{2}b_1$, then*

$$\frac{1}{\sqrt{\alpha_k^{min}}} \begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Theta\boldsymbol{\Omega}^*\right), \tag{4.5}$$

and

$$\left\| \begin{pmatrix} [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}^-(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}^-(\boldsymbol{x}^*)} \end{pmatrix} \right\|_2 = \begin{cases} o\left(\alpha_k^{min}\right), & \text{if } b_1 < 1, \\ \mathcal{O}\left(\alpha_k^{min}\right), & \text{if } b_1 = 1. \end{cases}, \tag{4.6}$$

where

$$\Theta := \begin{cases} 1/2, & \text{if } b_1 < 1, \\ 1/\left(2 - \frac{1}{\iota_1}\right), & \text{if } b_1 = 1. \end{cases}$$

Sketch of proof: We start by decomposing the primal-dual variable

$$(\boldsymbol{x}_{k+1} - \boldsymbol{x}^*, \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^*, [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)})$$

into three terms $\mathcal{Q}_{1,k}$, $\mathcal{Q}_{2,k}$ and $\mathcal{Q}_{3,k}$. Using the central limit theorem for martingale difference array, we establish that $\frac{1}{\sqrt{\alpha_k^{\min}}}\mathcal{Q}_{1,k} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Theta\boldsymbol{\Omega}^*\right)$. Under the given conditions, we

can further show the remaining two terms satisfying $\mathbb{E}[\mathcal{Q}_{2,k}] = o\left(\sqrt{\alpha_k^{\min}}\right)$ and $\mathbb{E}[\mathcal{Q}_{3,k}] = o\left(\sqrt{\alpha_k^{\min}}\right)$. Then, the result is obtained by Slutsky's theorem. Detailed proof can be found in Appendix C.3.

It is a surprising and novel result that the algorithm with averaged gradients can achieve asymptotic normality. Previous works [15, 42, 51, 70] studying the asymptotic normality of algorithms mostly rely on the independence of gradients. However, the averaged gradients in our algorithm are highly correlated. The key idea here is the introduction of two distinct step sizes, with different decay rates for iterates and gradient updates. This can be regarded as a "competition" between the iterates and the gradients. Specifically, for asymptotic normality to be achieved, it is essential that the gradients converge faster than the iterates. This ensures that the algorithm is driven by the most current and relevant gradients, contributing to its effective performance. To the best of our knowledge, it is the first work establishing the asymptotic normality for the algorithm with averaged gradients.

COROLLARY 1. *Under Assumptions 2, 5 and 6, and let* $\iota_1 = 1$, $b_1 = 1$, $\iota_2 > 0$ *and* $b_2 \in \left(\frac{1}{2}, 1\right)$, *then*

$$\sqrt{k}\begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^*). \tag{4.7}$$

The asymptotic normality for M-estimators in Equation (1.2) establishes a lower bound for stochastic algorithms in solving Problem (1.1), in Hájek and Le Cam's sense [41, 72]. Our result in Corollary 1 complements this by demonstrating that the optimal local convergence behavior of the primal variable is achieved, as characterized by the covariance matrix $\boldsymbol{\Omega}^*$. Here, considering only the primal variable $\boldsymbol{x}_k$, our results in Equation (4.7) match Equation (1.2) by applying knowledge of block matrix inverse [69, Corollary 2.3].

A practical estimator of the covariance matrix. Here, we provide a practical plug-in estimator for the unknown covariance matrix $\boldsymbol{\Omega}^*$. We then show that the estimator is convergent to the exact one, and therefore, it can be adopted in analyzing the local behavior of algorithms and conducting statistical inference. Let

$$\boldsymbol{\Omega}_k = \boldsymbol{H}_k^{-1} \boldsymbol{\Sigma}_k \boldsymbol{H}_k^{-1}, \tag{4.8}$$

where

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} \frac{1}{k+1}\sum_{i=0}^k \boldsymbol{g}_i \boldsymbol{g}_i^\top - \left(\frac{1}{k+1}\sum_{i=0}^k \boldsymbol{g}_i\right)\left(\frac{1}{k+1}\sum_{i=0}^k \boldsymbol{g}_i\right)^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Observe that $\boldsymbol{\Omega}_k$ can be cheaply calculated in each iteration, without additional sampling/estimation of gradients, since it shares the gradient estimate $\boldsymbol{g}_i$ with averaged gradients $\bar{\boldsymbol{g}}_k$. The estimator $\boldsymbol{\Omega}_k$ for the covariance matrix $\boldsymbol{\Omega}^*$ can be used as a surrogate to the exact matrix for analyzing the local behavior of algorithms and conducting statistical inference. The following theorem establishes the almost sure convergence of the practical plug-in estimator to the exact covariance matrix. The proof can be found in Appendix C.4.

THEOREM 5. *Under Assumptions 2, 5 and 6, and suppose that* $\iota_1 > b_2$ *if* $b_1 = 1$, *and* $b_2 > \frac{1}{2}b_1$, *then* $\boldsymbol{\Sigma}_k \to \boldsymbol{\Sigma}^*$ *and* $\boldsymbol{\Omega}_k \to \boldsymbol{\Omega}^*$, *almost surely.*

5. Experiments. In this section, we describe comprehensive experiments we have conducted to demonstrate the effectiveness of the Debiased-StoSQP-v2 (Algorithm 3). Specifically, we applied it to a variety of problems, including benchmark optimization problems from CUTEst library [28, 31] as well as constrained regression problems. For regression problems, we consider the linear, logistic, and Poisson models. Under the classical regression setup, we define $\zeta_k = (\zeta_{b_k}, \zeta_{\boldsymbol{a}_k})$, where ζ_{b_k} represents the k-th response and $\zeta_{\boldsymbol{a}_k}$ the corresponding observation (attributes). In linear regression, the response is generated according to:

$$\zeta_{b_k} = \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}^* + \varepsilon_k,$$

where $\boldsymbol{x}^*$ is the true parameter and $\{\varepsilon_k\}_k$ are i.i.d. noise terms. For logistic regression models, we consider binary responses $\zeta_{b_k} \in \{-1, 1\}$ generated via:

$$\mathbb{P}\left(\zeta_{b_k} | \zeta_{\boldsymbol{a}_k}\right) = \frac{1}{1 + \exp\left(-\zeta_{b_k} \cdot \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}^*\right)}.$$

In the case of Poisson regression, the response follows a conditional Poisson distribution depending on the observation, i.e.,

$$\zeta_{b_k} \sim \text{Pois}\left(\lambda(\zeta_{\boldsymbol{a}_k})\right), \text{ where } \log(\lambda(\zeta_{\boldsymbol{a}_k})) = \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}^*.$$

For each of these models, we can define an objective function corresponding to the model parameter $\boldsymbol{x}$:

$$\begin{aligned}
\text{linear models:} \quad & f(\boldsymbol{x}; \zeta_k) = \frac{1}{2}\left(\zeta_{b_k} - \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}\right), \\
\text{logistic models:} \quad & f(\boldsymbol{x}; \zeta_k) = \log\left(1 + \exp\left(-\zeta_{b_k} \cdot \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}\right)\right), \\
\text{Poisson models:} \quad & f(\boldsymbol{x}; \zeta_k) = \zeta_{b_k} \cdot \zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x} - \exp\left(\zeta_{\boldsymbol{a}_k}^\top \boldsymbol{x}\right).
\end{aligned}$$

It is straightforward to verify that $\boldsymbol{x}^*$ is the optimal solution of the stochastic objective $f(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{x}; \zeta_k)\right]$. Constraints on the model parameters $\boldsymbol{x}$ may be incorporated based on prior knowledge or specific problem requirements. We also explore portfolio optimization problems featuring exponential and logarithmic utility functions as the objective. In terms of the hyperparameters, we fix them for all experiments. The step sizes are set as $\gamma_k = 1/(k+1)^{0.751}$ and $\beta_k = 1/(k+1)^{0.5}$, which satisfy conditions in Theorems 3 and 4. Quadratic subproblems are solved by the ProxQP solver [4].[1] Implementation details are available as our supplementary code, which can be accessed at https://github.com/yihang-gao/Debiased-StoSQP/tree/main/code.

5.1. *CUTEst benchmark problems.* The CUTEst library collects various types of constrained and unconstrained optimization problems for evaluating the performances of optimization algorithms. We select a subset of the constrained optimization problems (e.g., HS problems) from the library, and we artificially add noise to gradient and Hessian as follows:

- Gaussian noise: Let $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k, \zeta_k)$ be perturbed such that $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k, \zeta_k) \sim \mathcal{N}(\nabla f(\boldsymbol{x}_k), \epsilon(\boldsymbol{I} + \boldsymbol{e}\boldsymbol{e}^\top))$. Similarly, the Hessian is given by $\nabla^2 f(\boldsymbol{x}_k, \zeta_k) \sim \nabla^2 f(\boldsymbol{x}_k) + \boldsymbol{E}$, where $\boldsymbol{E}_{i,j} \sim \mathcal{N}(0, \epsilon)$. The noise level ϵ is chosen from $\{1, 10^{-1}, 10^{-2}, 10^{-4}\}$.

[1] In our implementation, we import the 'qpsolvers' package from https://qpsolvers.github.io/qpsolvers/.

- Student's t-distribution noise: Let $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k, \zeta_k)$ be perturbed by noise following a Student's t-distribution, i.e., $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k, \zeta) \sim \nabla f(\boldsymbol{x}_k) + \boldsymbol{s}$, where each entry $\boldsymbol{s}_i \sim \mathrm{t}(m)$. Similarly, the Hessian is perturbed as $\nabla^2 f(\boldsymbol{x}_k, \zeta) \sim \nabla^2 f(\boldsymbol{x}_k) + \boldsymbol{E}$, where each element $\boldsymbol{E}_{i,j} \sim \mathrm{t}(m)$. Here, $\mathrm{t}(m)$ denotes the t-distributional noise, m denotes the degrees of freedom, and m is selected from the set $\{3, 4, 5\}$.

We first conduct a comparison between the proposed Debiased-StoSQP-v2 (Algorithm 3), ActiveSet-SQP [48] and StoSQP [20], where we evaluate each method by the KKT residual in Equation (2.11) and feasibility error. For each algorithm and problem, we run 10^5 iterations. Our empirical results indicate that the Debiased-StoSQP-v2 algorithm consistently outperforms the StoSQP (without debiasing techniques), both of which adopt fully stochastic gradients and Hessian. This superior performance can be attributed to our algorithm's use of gradient averaging. After a sufficient number of iterations, the averaged gradient approaches the exact gradient, thereby approximating the behavior of deterministic algorithms.

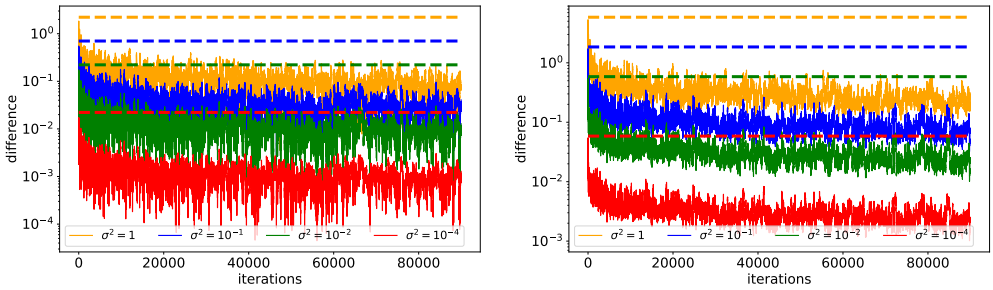
In particular, as shown in Figure 1, we plot the difference between the averaged gradients and the exact gradients during iterations, and the results validate our expectations and intuitions. The averaged gradients (the solid lines) move closer to the exact gradients, compared with estimated gradients without averaging (the dashed lines). In contrast, StoSQP (without debiasing techniques) lacks this beneficial property, and it suffers from oscillations brought by the stochastic gradients. The ActiveSet-SQP, which uses a stochastic line search method, necessitates an increasing sample size, and it employs a safeguard technique to ensure the accuracy of the line search. Consequently, it requires a sufficiently large sample size to make the line search practically effective. In contrast, Debiased-StoSQP-v2 requires only a single sample to estimate both the gradient and the Hessian in each iteration. Therefore, it is unsurprising that ActiveSet-SQP performs better with higher noise levels. However, when the noise level is relatively low, Debiased-StoSQP-v2 can effectively mitigate the noise through averaging gradient, and it can achieve similar and even better performances than ActiveSet-SQP. The visualized results are shown in Figure 2.

We next test the local asymptotic normality behavior of the generated iterates. For each problem, we aim to estimate $\frac{1}{d}\mathbf{1}^\top \boldsymbol{x}^*$ and set the nominal coverage probability to 95%. Here, the confidence interval is constructed by

$$\left[\frac{1}{d}\mathbf{1}^\top \boldsymbol{x}_k - \frac{1.96}{d}\sqrt{\alpha_k^{\min}}\sqrt{\Theta \boldsymbol{e}_{[1:d]}^\top \boldsymbol{\Omega}_t \boldsymbol{e}_{[1:d]}}, \frac{1}{d}\mathbf{1}^\top \boldsymbol{x}_k + \frac{1.96}{d}\sqrt{\alpha_k^{\min}}\sqrt{\Theta \boldsymbol{e}_{[1:d]}^\top \boldsymbol{\Omega}_t \boldsymbol{e}_{[1:d]}}\right],$$

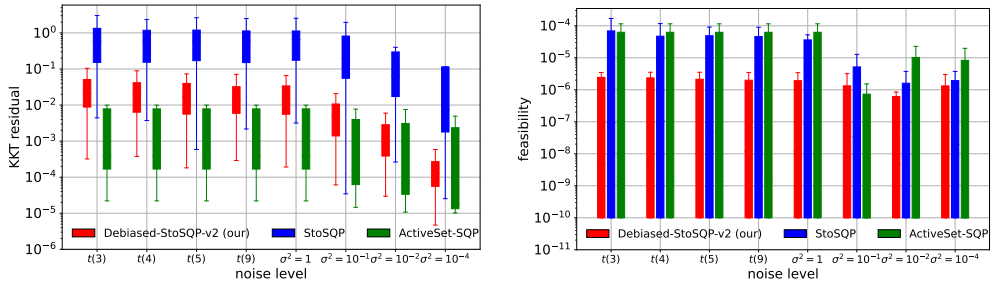
using the estimators and limiting normality results in Theorem 5. The performance of the method in terms of asymptotic normality is measured by the coverage rate (CovRate) of the confidence intervals and their average length (AvgLen) over 200 runs. The aggregated results are summarized in Table 1. We observe that the constructed (95%) confidence intervals by Debiased-StoSQP-v2 cover the true solution in probability closely aligned to 95%, thereby empirically validating our theoretical derivations on asymptotic normality. From the table, we note that the length of the confidence intervals tends to expand as the noise level increases, a behavior which is in line with our expectations, as the covariance matrix $\boldsymbol{\Omega}^*$ is dependent on the $\mathrm{Cov}\left(\nabla f(\boldsymbol{x}^*; \zeta)\right)$.

5.2. *Constrained regression problems.* Here, we implement Debiased-StoSQP-v2 algorithm (Algorithm 3) on constrained regression problems, including both the linear and the logistic regression. The response ζ_{b_k} is generated based on observations $\zeta_{\boldsymbol{a}_k} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, where the mean vector is set as $\boldsymbol{\mu}_a = (1, \cdots, 1, -1, \cdots, -1)$. We explore three different choices of the covariance matrix, as in [15]: (i) Identity matrix, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_d$; (ii) Toeplitz matrix, i.e., $(\boldsymbol{\Sigma}_a)_{i,j} = r^{|i-j|}$ for some $r > 0$; and (iii) Equicorrelation matrix, i.e., $(\boldsymbol{\Sigma}_a)_{i,j} = r$ for all $i \neq j$ and $(\boldsymbol{\Sigma}_a)_{i,i} = 1$, for some $r > 0$. The true parameter vector of



(a) HS32 (b) FCCU

Fig 1: Difference between the averaged gradients and the exact gradients on HS32 and FCCU problems. Solid lines: trajectories of gradient difference between the averaged gradients and the exact gradients during iterations, i.e., $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2$. *Dashed lines: expected error without averaging, i.e.,* $\mathbb{E}_k\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|_2\right]$.



(a) KKT residuals (b) Feasibility errors

Fig 2: KKT residuals and feasibility errors of Debiased-StoSQP-v2, StoSQP, and ActiveSet-SQP on CUTEst problems.

both two regression models is configured as $\boldsymbol{x}^* = \left(\frac{3}{2d}, \cdots, \frac{3}{2d}, \frac{1}{2d}, \cdots, \frac{1}{2d}\right)^\top$. We consider the non-negativity constraints, denoted by $\Omega := \left\{\boldsymbol{x} : \mathbf{1}^\top \boldsymbol{x} = 1, \boldsymbol{x} \geq \mathbf{0}\right\}$. In the linear regression problem, the noise ε_k is sampled from $\varepsilon_k \sim \mathcal{N}(0, 1)$. We aim to estimate $\hat{\boldsymbol{e}}^\top \boldsymbol{x}^*$, where $\hat{\boldsymbol{e}} = (1, \cdots, 1, -1, \cdots, -1)^\top$, by constructing 95% confidence intervals.

We report the results in Tables 2 and 3, highlighting different settings for the Toeplitz matrix and Equicorrelation matrix with $r = 0.4, 0.5, 0.6$ and $r = 0.1, 0.2, 0.3$, respectively. In each experiment, we run 200 times with varying random seeds to calculate the coverage rate (CovRate) and the average length (AvgLen) of the confidence interval. Our results affirm that the constructed 95% confidence intervals closely achieve a 95% coverage rate, thus empirically validating our theoretical conclusions on asymptotic normality. Moreover, we also observe that the average length of confidence intervals are in order of 10^{-2}, matching the experimental results reported by Chen et al. [15] and Na et al. [51]. The low standard deviation of these intervals' length relative to their average length suggests robustness across different random seeds.

5.3. *Portfolio optimization problems.* Here, we investigate portfolio optimization problems using 30 portfolios selected from the Fama-French 100 Portfolios DataSet, subject to the well-known gross-exposure constraint [26]: $\Omega_c := \left\{\boldsymbol{x} : \mathbf{1}^\top \boldsymbol{x} = 1, \|\boldsymbol{x}\|_1 \leq c\right\}$, where we set $c = 3$, and where $\boldsymbol{x}$ denotes the weights for corresponding stocks. We consider four prevalent portfolio optimization models:

TABLE 1
The coverage rate (CovRate) and length of confidence intervals (AvgLen) for some CUTEst (constrained) problems. The standard deviation of the interval length is also reported.

Problem	Noise Level	Gaussian		Freedom	Student t	
		CovRate(%)	AvgLen		CovRate(%)	AvgLen
HS41	1E+0	97.0	2.50E-2 (7.73E-4)	3	86.0	3.77E-2 (1.90E-3)
	1E-1	97.5	7.59E-3 (7.03E-5)	4	93.0	3.06E-2 (1.28E-3)
	1E-2	97.0	2.40E-3 (8.69E-6)	5	94.0	2.79E-2 (9.25E-4)
	1E-4	97.5	2.40E-4 (5.95E-7)	9	97.0	2.45E-2 (7.10E-4)
HS65	1E+0	94.5	1.87E-3 (6.82E-6)	3	96.5	3.18E-3 (1.66E-4)
	1E-1	94.5	5.92E-4 (1.59E-6)	4	95.0	2.59E-3 (1.72E-5)
	1E-2	95.0	1.87E-4 (4.96E-7)	5	95.0	2.37E-3 (1.11E-5)
	1E-4	94.5	1.87E-5 (4.97E-8)	9	94.5	2.08E-3 (8.36E-6)
HS68	1E+0	97.0	2.31E-1 (4.85E-2)	3	95.5	3.00E-1 (1.32E-1)
	1E-1	98.0	5.09E-2 (2.33E-3)	4	94.5	2.08E-1 (6.09E-2)
	1E-2	98.5	1.58E-2 (2.23E-4)	5	95.0	1.81E-1 (5.07E-2)
	1E-4	95.5	1.58E-3 (4.56E-6)	9	94.5	1.48E-1 (3.52E-2)
HS71	1E+0	97.0	1.95E-3 (1.44E-5)	3	94.0	3.34E-3 (1.23E-4)
	1E-1	96.5	6.17E-4 (1.93E-6)	4	96.0	2.74E-3 (6.79E-3)
	1E-2	96.5	1.95E-4 (5.20E-7)	5	96.5	2.49E-3 (2.51E-5)
	1E-4	98.5	1.95E-5 (5.08E-8)	9	95.0	2.19E-3 (2.12E-5)
HS81	1E+0	94.5	3.49E-2 (3.17E-3)	3	91.0	5.04E-2 (7.56E-3)
	1E-1	97.0	1.13E-2 (4.77E-5)	4	94.0	4.21E-2 (3.51E-3)
	1E-2	98.0	3.58E-3 (9.63E-6)	5	94.5	3.88E-2 (2.42E-3)
	1E-4	98.0	3.59E-4 (9.22E-7)	9	95.0	3.43E-2 (2.10E-3)

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	Dimension	CovRate(%)	AvgLen	Dimension	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 (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 (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 (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 (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 (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 (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)

- Global minimum variance (GMV)

$$\min_{\boldsymbol{x}\in\Omega_c} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x},$$

  where $\boldsymbol{\Sigma}$ is the covariance matrix of target stocks.
- Mean-variance (MV)

$$\min_{\boldsymbol{x}\in\Omega_c} -\boldsymbol{x}^\top \boldsymbol{\mu} + \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x},$$

  where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance matrix of target stocks.
- Exponential utility (EXP)

$$\min_{\boldsymbol{x}\in\Omega_c} \mathbb{E}\left[\exp\left(-\eta\left(\boldsymbol{x}^\top \zeta_{\boldsymbol{a}}\right)\right)\right],$$

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	Dimension	CovRate(%)	AvgLen	Dimension	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 (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 (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 (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 (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 (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 (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)

where $\zeta_{\boldsymbol{a}}$ is the observed price changes and $\eta > 0$ is a scaling parameter set to be $\eta = 0.1$.

- Logarithmic utility (LOG)

$$\min_{\boldsymbol{x} \in \Omega_c} -\mathbb{E}\left[\log\left(\boldsymbol{x}^\top \zeta_{\boldsymbol{a}} + \eta\right)\right],$$

where $\zeta_{\boldsymbol{a}}$ is the observed price changes and $\eta > 0$ serves as the regularization parameter to ensure the feasibility of the logarithm, where we set $\eta = 15$.

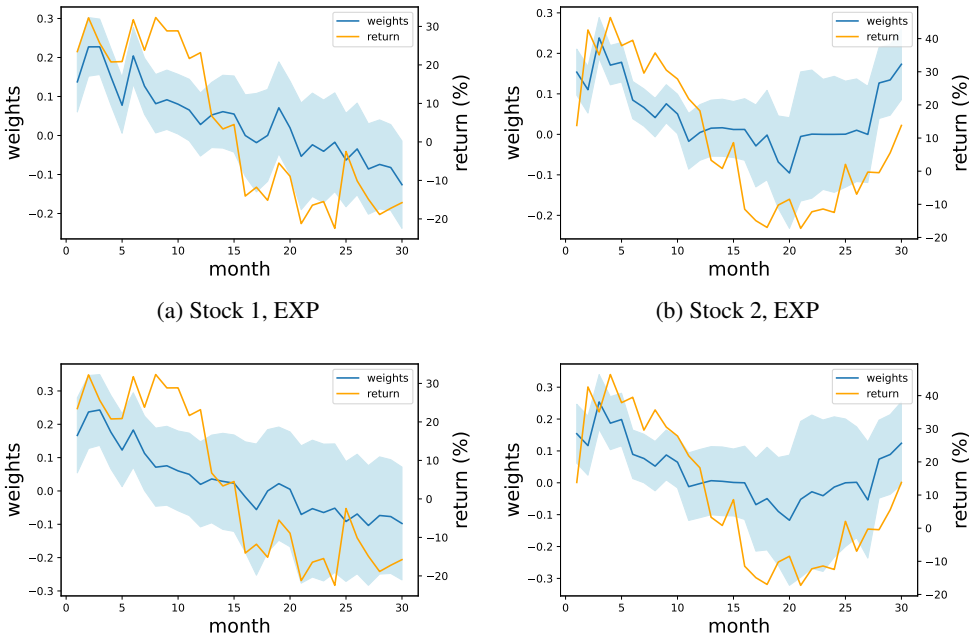
TABLE 4
Fama-French 100 Portfolios DataSet, 2021-2023

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

The exact mean, covariance, and expectations are inaccessible in practice. Instead, we estimate the stochastic gradient and Hessian of the expected objective using available observations, and we apply Debiased-StoSQP-v2 to solve the problems. For the portfolio strategy $\boldsymbol{x}$, we use historical data from the past year as training samples. We assess the performance of our portfolio strategies using four key metrics, calculated over the data from years 2021-2023: the accumulative return, maximum drawdown, sharp ratio, and Sortino ratio. The accumulative return captures the overall gain or loss of the portfolio strategy. The other three are related to the risk of the strategy: the maximum drawdown measures the maximum observed loss from a peak to a trough; the sharp ratio compares the portfolio's return to its risk, taking into account the standard deviation of the portfolio returns; and the Sortino ratio is a variation of the sharp ratio, considering the standard deviation of negative portfolio returns. The results are summarized in Table 4.

Interestingly, we observe that the model of logarithmic utility achieves the best accumulative return, consistent with the results reported by [22]. In terms of risk control, however, the mean-variance model is more favorable. We also perform a comparison between Debiased-StoSQP-v2 and the deterministic approach denoted as "det". We found that the performance metrics across the two methods are quite similar, suggesting that Debiased-StoSQP-v2 approaches deterministic algorithms because of the use of the averaging gradient and Hessian.

In Figure 3, we visualize the weights of two stocks as an example, evaluated by the exponential and the logarithmic utility models. The blue line traces the trajectory of the weight corresponding to the stock over time. This is accompanied by a blue band, which represents the estimated standard deviation of the weight, as evaluated by the developed asymptotic normality. The yellow line is the accumulative return of the stock. We observe a significant correlation between the weight adjustments and the stock's return trajectory. Notably, abrupt changes in the stock's return are promptly followed by widened blue bands, indicating a surge in the estimated variance of the weight. This behavior matches well with intuition and underscores the hypothesis that the variance of the weight may serve as an indicator of the stock's inherent risk.



(a) Stock 1, EXP

(b) Stock 2, EXP

(c) Stock 1, LOG

(d) Stock 2, LOG

Fig 3: Weights and returns. Blue lines: the predicted weight for a specific stock in different months. Blue bands: the estimated standard deviation of the weight, evaluated by the derived asymptotic normality results. Yellow lines: the accumulative return of the stock. The predicted stock weights are highly correlated with the stock's price. Abrupt stock price (return) changes shown by the yellow lines are followed by widened blue bands.

5.4. *Poisson regression: Chicago air pollution and death rate data.* Here, we study the relationship between different attributes related to air pollution (e.g., PM10, PM25, O3, SO2) and time, with the death rate, by using Poisson regression. Let $\zeta_{\boldsymbol{a}} \in \mathbb{R}^d$ represent the vector of air pollution and time attributes, and let $\zeta_b \in \mathbb{N}$ denote the death rate. We model the conditional distribution of death ζ_b given $\zeta_{\boldsymbol{a}}$ as a Poisson distribution: $\zeta_b|\zeta_{\boldsymbol{a}} \sim \text{Pois}(\lambda(\zeta_{\boldsymbol{a}}))$, where

$\log \lambda(\zeta_{\boldsymbol{a}}) = \zeta_{\boldsymbol{a}}^\top \boldsymbol{x}^*$ and $\boldsymbol{x}^*$ is the true, but unknown, parameter vector for the Poisson linear model. The unconstrained Poisson regression problem is formulated as

$$\min_{\boldsymbol{x}} \quad \mathbb{E}\left[f(\boldsymbol{x};\zeta)\right] := \mathbb{E}_{(\zeta_b,\zeta_{\boldsymbol{a}})}\left[\zeta_b \cdot \zeta_{\boldsymbol{a}}^\top \boldsymbol{x} - \exp\left(\zeta_{\boldsymbol{a}}^\top \boldsymbol{x}\right)\right]. \tag{5.1}$$

However, based on prior knowledge that air pollution attributes are likely to contribute to an increase in the death rate, we impose non-negativity constraints on the corresponding weights $\boldsymbol{x}$. The constrained Poisson regression model is

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \mathbb{E}\left[f(\boldsymbol{x};\zeta)\right] := \mathbb{E}_{(\zeta_b,\zeta_{\boldsymbol{a}})}\left[\zeta_b \cdot \zeta_{\boldsymbol{a}}^\top \boldsymbol{x} - \exp\left(\zeta_{\boldsymbol{a}}^\top \boldsymbol{x}\right)\right], \\ \text{s.t.} \quad & \boldsymbol{x}_{\mathcal{B}} \geq \boldsymbol{0}, \end{aligned} \tag{5.2}$$

where $\mathcal{B}$ is the set of indices of weights corresponding to pollution attributes.

TABLE 5
Summary of Poisson regression (Model 1) on Chicago air pollution and death rate data

Variables	Model	Coefficient (10^{-2})	95 % CI (10^{-2})	p-Value
PM10	**Model 1**	0.42	[-0.57, 1.42]	0.396
	Ours	0.13	[-0.56,0.79]	0.371
PM25	**Model 1**	0.72	[-0.08, 1.52]	0.103
	Ours	0.65	[0.02, 1.28]	0.023
O3	**Model 1**	-2.97	[-3.70, -2.24]	0.000
	Ours	0.00	active	
SO2	**Model 1**	1.38	[0.58, 2.20]	0.001
	Ours	2.08	[1.43, 2.73]	0.000
Time	**Model 1**	0.95	[0.17, 1.74]	0.008
	Ours	1.13	[0.64, 1.63]	0.000
Interc	**Model 1**	4.6968	[4.690, 4.704]	0.000
	Ours	4.6974	[4.692, 4.703]	0.000

TABLE 6
Summary of Poisson regression (Model 2) on Chicago air pollution and death rate data

Variables	Model	Coefficient (10^{-2})	95 % CI (10^{-2})	p-Value
PM10	**Model 2**	-0.86	[-1.82, 0.10]	0.062
	Ours	0.11	[-0.52, 0.74]	0.362
PM25	**Model 2**	1.37	[0.51, 2.23]	0.001
	Ours	0.65	[0.01, 1.28]	0.022
SO2	**Model 2**	2.06	[1.28, 2.84]	0.000
	Ours	2.08	[1.42, 2.73]	0.000
Time	**Model 2**	1.21	[0.53, 1.89]	0.001
	Ours	1.13	[0.64, 1.63]	0.000
Interc	**Model 1**	4.6972	[4.690, 4.704]	0.000
	Ours	4.6973	[4.692, 4.703]	0.000

We first consider the unconstrained Poisson regression model in Equation (5.1), including all five attributes, denoted as Model 1. The estimated model coefficients and their confidence

intervals and p-values are provided by the 'statsmodels' package in Python [64]. Surprisingly, we observe that the coefficient for O3 is significantly negative, which contradicts our prior understanding of the likely impact of air pollution on death rates. To address this issue, we turn to the constrained Poisson model in Equation (5.2), solved by Debiased-StoSQP-v2 (Algorithm 3). We also estimate confidence intervals and p-values using the derived asymptotic normality. We list the results in Table 5. Remarkably, under the non-negativity constraints, the weight of O3 is recognized to be active with the constraints, i.e., it is equal to zero. The estimated model coefficients from the constrained Poisson model are more consistent with our prior beliefs.

Next, we consider a reduced model that excludes the O3 attribute, denoted as Model 2. We find that in this model, the weight for PM10 becomes negative, once again contradicting our prior beliefs. Similarly, employing the constrained Poisson model and solving it via Debiased-StoSQP-v2 (Algorithm 3), the weight for PM10 becomes positive, although the significance level revealed by the p-value is not particularly strong. The results are reported in Table 6. These findings highlight the importance of incorporating domain-specific constraints in statistical models. They also emphasize the effectiveness of our approach in addressing such constrained optimization problems.

6. Conclusion. In this work, we proposed Debiased-StoSQP and its refinement Debiased-StoSQP-v2, fully stochastic Newton's methods for solving constrained optimization problems. We include the averaging technique for both the gradient and Hessian, reducing the impact of stochastic noise and improving the algorithm's performance, compared to existing fully stochastic algorithms. We then established the almost sure global convergence in terms of the first-order optimality (KKT) conditions. Furthermore, under certain mild conditions, we developed the asymptotic normality for Debiased-StoSQP-v2 (Algorithm 3). This is a particularly surprising and novel result since the gradients in Debiased-StoSQP-v2 are highly correlated, in contrast with previous work that primarily relies on the independence of gradients. We also provided a practical plug-in estimator for the covariance matrix. With our results, we are capable of applying Debiased-StoSQP-v2 to perform online inference for constrained optimization problems, as we have demonstrated with our empirical results.

While our algorithm Debiased-StoSQP-v2 has demonstrated promising results, there is still potential for further investigation and improvement. Specifically, the current implementation and analysis require the exact solution of quadratic subproblems, which could be computationally expensive. A possible extension of this work would be to explore the use of inexact solutions for the quadratic subproblems, possibly implemented within the RandBLAS/RandLAPACK library [46]. A recent work by Na and Mahoney [51] employed sketching techniques to inexactly solve linear systems for equality-constrained subproblems. The asymptotic normality behavior still holds for the StoSQP algorithm with sketching. It remains an open question whether the global almost sure convergence and the local asymptotic normality properties of Debiased-StoSQP-v2 are preserved when fast and inexact solvers are adopted. Investigating these areas would help in developing a more efficient and versatile algorithm, with broader applicability in constrained optimization scenarios.

Acknowledgements. MWM would like to acknowledge the IARPA, NSF, and ONR for providing partial support of this work.

REFERENCES

[1] ALLEN-ZHU, Z. (2018). How to make the gradients small stochastically: Even faster convex and nonconvex SGD. *Advances in Neural Information Processing Systems* **31**.

[2] ANASTASIOU, A., BALASUBRAMANIAN, K. and ERDOGDU, M. A. (2019). Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory* 115–137. PMLR.

[3] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **70** 214–223. PMLR.

[4] BAMBADE, A., EL-KAZDADI, S., TAYLOR, A. and CARPENTIER, J. (2022). Prox-QP: Yet another quadratic programming solver for robotics and beyond. In *RSS 2022-Robotics: Science and Systems*.

[5] BERAHAS, A. S., CURTIS, F. E., O'NEILL, M. J. and ROBINSON, D. P. (2021). A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv preprint arXiv:2106.13015*. https://doi.org/arXiv:2106.13015

[6] BERAHAS, A. S., CURTIS, F. E., ROBINSON, D. and ZHOU, B. (2021). Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization. *SIAM Journal on Optimization* **31** 1352–1379. https://doi.org/10.1137/20m1354556

[7] BERAHAS, A. S., SHI, J., YI, Z. and ZHOU, B. (2023). Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *Computational Optimization and Applications* **86** 79–116. https://doi.org/10.1007/s10589-023-00483-2

[8] BERTSEKAS, D. P. (1997). Nonlinear Programming. *Journal of the Operational Research Society* **48** 334–334. https://doi.org/10.1057/palgrave.jors.2600425

[9] BOGGS, P. T. and TOLLE, J. W. (1995). Sequential Quadratic Programming. *Acta Numerica* **4** 1–51. https://doi.org/10.1017/s0962492900002518

[10] BOYER, C. and GODICHON-BAGGIONI, A. (2022). On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions. *Computational Optimization and Applications* **84** 921–972. https://doi.org/10.1007/s10589-022-00442-3

[11] BURKE, J. V. and HAN, S.-P. (1989). A robust sequential quadratic programming method. *Mathematical Programming* **43** 277–303. https://doi.org/10.1007/bf01582294

[12] CARMON, Y., DUCHI, J. C., HINDER, O. and SIDFORD, A. (2017). "Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions. In *International Conference on Machine Learning* 654–663. PMLR.

[13] CARMON, Y., DUCHI, J. C., HINDER, O. and SIDFORD, A. (2018). Accelerated Methods for NonConvex Optimization. *SIAM Journal on Optimization* **28** 1751–1772. https://doi.org/10.1137/17m1114296

[14] CHEN, R. T., RUBANOVA, Y., BETTENCOURT, J. and DUVENAUD, D. K. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems* **31**.

[15] CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics* **48** 251–273. https://doi.org/10.1214/18-aos1801

[16] CISSE, M., BOJANOWSKI, P., GRAVE, E., DAUPHIN, Y. and USUNIER, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning* 854–863. PMLR.

[17] CUOMO, S., COLA, V. S. D., GIAMPAOLO, F., ROZZA, G., RAISSI, M. and PICCIALLI, F. (2022). Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. *Journal of Scientific Computing* **92** 88. https://doi.org/10.1007/s10915-022-01939-z

[18] CURTIS, F. E., O'NEILL, M. J. and ROBINSON, D. P. (2023). Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*. https://doi.org/10.1007/s10107-023-01981-1

[19] CURTIS, F. E., ROBINSON, D. P. and ZHOU, B. (2021). Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*. https://doi.org/arXiv:2107.03512

[20] CURTIS, F. E., ROBINSON, D. P. and ZHOU, B. (2023). Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints. *arXiv preprint arXiv:2302.14790*. https://doi.org/arXiv:2302.14790

[21] DANIEL, J. W. (1973). Stability of the solution of definite quadratic programs. *Mathematical Programming* **5** 41–53. https://doi.org/10.1007/bf01580110

[22] DU, J.-H., GUO, Y. and WANG, X. (2022). High-Dimensional Portfolio Selection with Cardinality Constraints. *Journal of the American Statistical Association* **118** 779–791. https://doi.org/10.1080/01621459.2022.2133718

[23] DUCHI, J. C. and RUAN, F. (2021). Asymptotic optimality in stochastic optimization. *The Annals of Statistics* **49**. https://doi.org/10.1214/19-aos1831

[24] DUPACOVA, J. and WETS, R. (1988). Asymptotic Behavior of Statistical Estimators and of Optimal Solutions of Stochastic Optimization Problems. *The Annals of Statistics* **16**. https://doi.org/10.1214/aos/1176351052

[25] FAN, J. (2007). Variable screening in high-dimensional feature space. In *Proceedings of the 4th International Congress of Chinese Mathematicians* **2** 735–747. Citeseer.
[26] FAN, J., ZHANG, J. and YU, K. (2012). Vast Portfolio Selection With Gross-Exposure Constraints. *Journal of the American Statistical Association* **107** 592–606. https://doi.org/10.1080/01621459.2012.682825
[27] FANG, Y., NA, S., MAHONEY, M. W. and KOLAR, M. (2022). Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *arXiv preprint arXiv:2211.15943*. https://doi.org/arXiv:2211.15943
[28] FOWKES, J., ROBERTS, L. and BŰRMEN, Á. (2022). PyCUTEst: an open source Python package of optimization test problems. *Journal of Open Source Software* **7** 4377. https://doi.org/10.21105/joss.04377
[29] GAUVIN, J. (1977). A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Mathematical Programming* **12** 136–138. https://doi.org/10.1007/bf01593777
[30] GOODFELLOW, I., SHLENS, J. and SZEGEDY, C. (2015). Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.
[31] GOULD, N. I. M., ORBAN, D. and TOINT, P. L. (2014). CUTEst: a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications* **60** 545–557. https://doi.org/10.1007/s10589-014-9687-3
[32] GOWER, R. M. and RICHTÁRIK, P. (2015). Randomized Iterative Methods for Linear Systems. *SIAM Journal on Matrix Analysis and Applications* **36** 1660–1690. https://doi.org/10.1137/15m1025487
[33] HANSEN, D., MADDIX, D. C., ALIZADEH, S., GUPTA, G. and MAHONEY, M. W. (2023). Learning Physical Models that Can Respect Conservation Laws. In *Proceedings of International Conference on Machine Learning* **202** 12469–12510. PMLR.
[34] HAO, S. and LIU, Q. (2014). Convergence rates in the law of large numbers for arrays of martingale differences. *Journal of Mathematical Analysis and Applications* **417** 733–773. https://doi.org/10.1016/j.jmaa.2014.03.049
[35] HODGKINSON, L., VAN DER HEIDE, C., ROOSTA, F. and MAHONEY, M. W. (2022). Monotonicity and Double Descent in Uncertainty Estimation with Gaussian Processes Technical Report No. Preprint: arXiv:2210.07612.
[36] HODGKINSON, L., VAN DER HEIDE, C., SALOMONE, R., ROOSTA, F. and MAHONEY, M. W. (2023). The Interpolating Information Criterion for Overparameterized Models. *arXiv preprint arXiv:2307.07785*.
[37] HOFFMAN, A. J. (2003). On Approximate Solutions of Systems of Linear Inequalities. In *Selected Papers of Alan J Hoffman* 174–176. World Scientific. https://doi.org/10.1142/9789812796936_0018
[38] JORGE, N. and STEPHEN, J. W. (2006). *Numerical optimization*. Spinger. https://doi.org/10.1007/0-387-22742-3_18
[39] KARNIADAKIS, G. E., KEVREKIDIS, I. G., LU, L., PERDIKARIS, P., WANG, S. and YANG, L. (2021). Physics-informed machine learning. *Nature Reviews Physics* **3** 422–440. https://doi.org/10.1038/s42254-021-00314-5
[40] KRISHNAPRIYAN, A., GHOLAMI, A., ZHE, S., KIRBY, R. and MAHONEY, M. W. (2021). Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems* **34** 26548–26560.
[41] LE CAM, L. M. and YANG, G. L. (2000). *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4684-0377-0
[42] LELUC, R. and PORTIER, F. (2020). Asymptotic Analysis of Conditioned Stochastic Gradient Descent. *arXiv preprint arXiv:2006.02745*. https://doi.org/arXiv:2006.02745
[43] LIU, H., LI, Z., HALL, D., LIANG, P. and MA, T. (2023). Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training. *arXiv preprint arXiv:2305.14342*. https://doi.org/arXiv:2305.14342
[44] LIU, X. and YUAN, Y. (2011). A Sequential Quadratic Programming Method Without A Penalty Function or a Filter for Nonlinear Equality Constrained Optimization. *SIAM Journal on Optimization* **21** 545–571. https://doi.org/10.1137/080739884
[45] MOU, W., LI, C. J., WAINWRIGHT, M. J., BARTLETT, P. L. and JORDAN, M. I. (2020). On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory* 2947–2997. PMLR.
[46] MURRAY, R., DEMMEL, J., MAHONEY, M. W., ERICHSON, N. B., MELNICHENKO, M., MALIK, O. A., GRIGORI, L., LUSZCZEK, P., DEREZIŃSKI, M., LOPES, M. E., LIANG, T., LUO, H. and DONGARRA, J. (2023). Randomized Numerical Linear Algebra – A Perspective on the Field with an Eye to Software Technical Report No. Preprint: arXiv:2302.11474v2.
[47] NA, S., ANITESCU, M. and KOLAR, M. (2022). An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming* **199** 721–791. https://doi.org/10.1007/s10107-022-01846-z

[48] NA, S., ANITESCU, M. and KOLAR, M. (2023). Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*. https://doi.org/10.1007/s10107-023-01935-7

[49] NA, S., DEREZIŃSKI, M. and MAHONEY, M. W. (2022). Hessian averaging in stochastic Newton methods achieves superlinear convergence. *Mathematical Programming* **201** 473–520. https://doi.org/10.1007/s10107-022-01913-5

[50] NA, S. and KOLAR, M. (2021). High-dimensional index volatility models via Stein's identity. *Bernoulli* **27**. https://doi.org/10.3150/20-bej1238

[51] NA, S. and MAHONEY, M. W. (2022). Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv preprint arXiv:2205.13687*. https://doi.org/arXiv:2205.13687

[52] NA, S., YANG, Z., WANG, Z. and KOLAR, M. (2019). High-dimensional Varying Index Coefficient Models via Stein's Identity. *Journal of Machine Learning Research* **20** 152–1.

[53] NÉGIAR, G., MAHONEY, M. W. and KRISHNAPRIYAN, A. (2023). Learning differentiable solvers for systems with hard constraints. In *The Eleventh International Conference on Learning Representations*.

[54] NEYSHABUR, B., BHOJANAPALLI, S., MCALLESTER, D. and SREBRO, N. (2017). Exploring generalization in deep learning. *Advances in Neural Information Processing Systems* **30**.

[55] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* **30** 838–855. https://doi.org/10.1137/0330046

[56] POWELL, M. J. D. (1978). A fast algorithm for nonlinearly constrained optimization calculations. In *Lecture Notes in Mathematics* 144–157. Springer Berlin Heidelberg. https://doi.org/10.1007/bfb0067703

[57] QIU, S. and KUNGURTSEV, V. (2023). A sequential quadratic programming method for optimization with stochastic objective functions, deterministic inequality constraints and robust subproblems. *arXiv preprint arXiv:2302.07947*. https://doi.org/arXiv:2302.07947

[58] RAISSI, M., PERDIKARIS, P. and KARNIADAKIS, G. E. (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** 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[59] ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* 233–257. Elsevier. https://doi.org/10.1016/b978-0-12-604550-5.50015-8

[60] ROBINSON, S. M. (1976). Stability Theory for Systems of Inequalities, Part II: Differentiable Nonlinear Systems. *SIAM Journal on Numerical Analysis* **13** 497–513. https://doi.org/10.1137/0713043

[61] ROOSTA-KHORASANI, F. and MAHONEY, M. W. (2019). Sub-Sampled Newton Methods. *Mathematical Programming* **174** 293–326. https://doi.org/10.1007/s10107-018-1346-5

[62] RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process Technical Report, Cornell University Operations Research and Industrial Engineering.

[63] SCHITTKOWSKI, K. and YUAN, Y.-X. (2011). Sequential Quadratic Programming Methods. https://doi.org/10.1002/9780470400531.eorms0984

[64] SEABOLD, S. and PERKTOLD, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* **57** 10–25080. Austin, TX.

[65] SEN, P. K. (1979). Asymptotic Properties of Maximum Likelihood Estimators Based on Conditional Specification. *The Annals of Statistics* **7**. https://doi.org/10.1214/aos/1176344785

[66] SHAPIRO, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72** 133–144. https://doi.org/10.1093/biomet/72.1.133

[67] SHAPIRO, A. (2000). On the asymptotics of constrained local M-estimators. *The Annals of Statistics* **28**. https://doi.org/10.1214/aos/1015952006

[68] SU, J., VARGAS, D. V. and SAKURAI, K. (2019). One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* **23** 828–841. https://doi.org/10.1109/tevc.2019.2890858

[69] TIAN, Y. and TAKANE, Y. (2009). The inverse of any two-by-two nonsingular partitioned matrix and three matrix inverse completion problems. *Computers & Mathematics with Applications* **57** 1294–1304. https://doi.org/10.1016/j.camwa.2009.01.025

[70] TOULIS, P. and AIROLDI, E. M. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics* **45**. https://doi.org/10.1214/16-aos1506

[71] ULBRICH, S. (2003). On the superlinear local convergence of a filter-SQP method. *Mathematical Programming* **100**. https://doi.org/10.1007/s10107-003-0491-6

[72] VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press. https://doi.org/10.1017/CBO9780511802256

[73] WANG, S., WANG, H. and PERDIKARIS, P. (2021). Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances* **7**. https://doi.org/10.1126/sciadv.abi8605

[74] XU, M., YE, J. J. and ZHANG, L. (2015). Smoothing SQP Methods for Solving Degenerate Nonsmooth Constrained Optimization Problems with Applications to Bilevel Programs. *SIAM Journal on Optimization* **25** 1388–1410. https://doi.org/10.1137/140971580

[75] XU, P., ROOSTA, F. and MAHONEY, M. W. (2020). Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming* **184** 35–70. https://doi.org/10.1007/s10107-019-01405-z

[76] XU, P., ROOSTA, F. and MAHONEY, M. W. (2020). Second-order optimization for non-convex machine learning: An empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining* 199–207. SIAM. https://doi.org/10.1137/1.9781611976236.23

[77] YAO, Z., GHOLAMI, A., SHEN, S., MUSTAFA, M., KEUTZER, K. and MAHONEY, M. (2021). ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **35** 10665–10673. https://doi.org/10.1609/aaai.v35i12.17275

[78] YAO, Z., XU, P., ROOSTA, F. and MAHONEY, M. W. (2021). Inexact nonconvex newton-type methods. *INFORMS Journal on Optimization* **3** 154–182. https://doi.org/10.1287/ijoo.2019.0043

[79] YUE, M.-C., ZHOU, Z. and SO, A. M.-C. (2019). On the Quadratic Convergence of the Cubic Regularization Method under a Local Error Bound Condition. *SIAM Journal on Optimization* **29** 904–932. https://doi.org/10.1137/18m1167498

[80] ZAFAR, M. B., VALERA, I., GOMEZ-RODRIGUEZ, M. and GUMMADI, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* **20** 2737–2778.

[81] ZHU, L., FENG, K., PU, Z. and MA, W. (2023). Adversarial Diffusion Attacks on Graph-based Traffic Prediction Models. *IEEE Internet of Things Journal* 1–1. https://doi.org/10.1109/jiot.2023.3290401

APPENDIX A: CONSTRAINTS RELAXATION AND DETERMINISTIC ALGORITHM

A.1. Proof for Proposition 1. We first prove the first part of the proposition that the relaxation parameter is non-zero if EGMFCQ holds at the iterate. Define $\mathcal{I}_k := \mathcal{I}(\boldsymbol{x}_k) := \{i \in [d] : (\boldsymbol{x}_k)_i = (\boldsymbol{\ell})_i\}$ and $\mathcal{J}_k := \mathcal{J}(\boldsymbol{x}_k) := \{i \in [d] : (\boldsymbol{x}_k)_i = (\boldsymbol{u})_i\}$, then $\mathcal{I}_k \cap \mathcal{J}_k = \emptyset$ and we denote $\mathcal{A}_k := \mathcal{A}(\boldsymbol{x}_k) := \mathcal{I}(\boldsymbol{x}_k) \cup \mathcal{J}(\boldsymbol{x}_k)$ as the active set of $\boldsymbol{x}_k$. Suppose that $\boldsymbol{z}_k$ is the vector satisfying Equation (2.4) at $\boldsymbol{x}_k$ and we simply let

$$\varepsilon := \min\left\{|(\boldsymbol{x}_k - \boldsymbol{\ell})_i|, |(\boldsymbol{u} - \boldsymbol{x}_k)_i|, |(\boldsymbol{u} - \boldsymbol{\ell})_j| : i \in \mathcal{A}_k^-, j \in \mathcal{A}_k\right\} > 0,$$

and

$$\bar{\boldsymbol{z}}_k = \frac{\varepsilon}{\|\boldsymbol{z}_k\|_2}\boldsymbol{z}_k.$$

Then, it is not difficult to verify that

$$\boldsymbol{\ell} \le \boldsymbol{x}_k + \bar{\boldsymbol{z}}_k \le \boldsymbol{u},$$

and

$$\frac{\varepsilon}{\|\boldsymbol{z}_k\|_2}\boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\bar{\boldsymbol{z}}_k = \boldsymbol{0}, \tag{A.1}$$

which imply that $\bar{\boldsymbol{z}}_k \in \widetilde{\Omega}_k$ with $\theta_k = \frac{\varepsilon}{\|\boldsymbol{z}_k\|_2}$. The following lemma shows that if $\widetilde{\Omega}_k$ with $\bar{\theta} \in (0, 1]$ is feasible and $0 < \bar{\bar{\theta}} \le \bar{\theta}$, then $\widetilde{\Omega}_k$ with $\bar{\bar{\theta}}$ is also feasible. It further indicates that Assumption 1 on the lower-boundedness of the relaxation parameter makes sense.

LEMMA 5. *If* $\{\boldsymbol{p} : \bar{\theta}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}\} \neq \emptyset$ *and* $0 < \bar{\bar{\theta}} \le \bar{\theta}$, *then* $\{\boldsymbol{p} : \bar{\bar{\theta}}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}\} \neq \emptyset$. *Therefore, Assumption 1 makes sense.*

PROOF. Suppose that $\bar{\boldsymbol{p}} \in \{\boldsymbol{p} : \bar{\theta}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}\}$, then $\bar{\theta}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \bar{\boldsymbol{p}} = \boldsymbol{0}$ and thus, $\bar{\bar{\theta}}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \left(\bar{\bar{\theta}}/\bar{\theta} \cdot \bar{\boldsymbol{p}}\right) = \boldsymbol{0}$. Let $\bar{\bar{\boldsymbol{p}}} = \bar{\bar{\theta}}/\bar{\theta} \cdot \bar{\boldsymbol{p}}$, then $\bar{\bar{\theta}}\boldsymbol{c}_k + \boldsymbol{J}_k^\top \bar{\bar{\boldsymbol{p}}} = \boldsymbol{0}$ and $\boldsymbol{\ell} \le \boldsymbol{x}_k + \bar{\bar{\boldsymbol{p}}} \le \boldsymbol{u}$, which complete the proof. □

LEMMA 6 (Theorem 3 in [60]). *If* $\bar{\boldsymbol{x}}$ *satisfies EGMFCQ, then there exists a neighborhood* $\mathcal{B}(\bar{\boldsymbol{x}}; \bar{r}) := \{\boldsymbol{x} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \le \bar{r}\}$ *with some sufficiently small radius* $\bar{r} > 0$, *such that all points in the neighborhood satisfy EGMFCQ.*

LEMMA 7. *Suppose that EGMFCQ holds at* $\bar{\boldsymbol{x}}$, *then EGMFCQ also holds at* $\bar{\boldsymbol{x}}_k$ *when* $\bar{\boldsymbol{x}}_k$ *is sufficiently close to* $\bar{\boldsymbol{x}}$, *for any sequence* $\bar{\boldsymbol{x}}_k \to \bar{\boldsymbol{x}}$, *by Lemma 6. Let* $\bar{\boldsymbol{z}}$ *be the vectors satisfying Condition (2.4) at* $\bar{\boldsymbol{x}}$. *Then we can always find a sequence of vectors* $\{\bar{\boldsymbol{z}}_k\}$ *with* $\bar{\boldsymbol{z}}_k$ *satisfying Condition (2.4) at* $\bar{\boldsymbol{x}}_k$ *such that* $\|\bar{\boldsymbol{z}}_k - \bar{\boldsymbol{z}}\|_2 \to 0$ *as* $\|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\|_2 \to 0$.

PROOF. Since the vector $\bar{\boldsymbol{z}}$ satisfies Condition (2.4) at $\bar{\boldsymbol{x}}$, i.e., $\boldsymbol{c}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{z}} = \boldsymbol{0}$, by the smoothness of $\boldsymbol{c}(\boldsymbol{x})$ and the linear independence of columns of $\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})$, we can find $\bar{\boldsymbol{z}}_k$ such that $\boldsymbol{c}(\bar{\boldsymbol{x}}_k) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}}_k)^\top \bar{\boldsymbol{z}}_k = \boldsymbol{0}$ and $\|\bar{\boldsymbol{z}}_k - \bar{\boldsymbol{z}}\|_2 \to 0$ as $\|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\|_2 \to 0$. Let $\varepsilon := \min\{|(\bar{\boldsymbol{z}})_i| : (\bar{\boldsymbol{x}})_i = (\boldsymbol{\ell})_i \text{ or } (\bar{\boldsymbol{x}})_i = (\boldsymbol{u})_i\}$. Due to the fact that $\mathcal{A}(\bar{\boldsymbol{x}}_k) \subseteq \mathcal{A}(\bar{\boldsymbol{x}})$, we have $(\bar{\boldsymbol{z}}_k)_i > 0$, if $(\bar{\boldsymbol{x}}_k)_i = (\boldsymbol{\ell})_i$ and $(\bar{\boldsymbol{z}}_k)_i < 0$, if $(\bar{\boldsymbol{x}}_k)_i = (\boldsymbol{u})_i$, when $\|\bar{\boldsymbol{z}}_k - \bar{\boldsymbol{z}}\|_2 \le \varepsilon$.

□

LEMMA 8. *Let* θ_k *be selected in* $(0, 1]$ *such that the relaxed feasible region* $\widetilde{\Omega}_k$ *is nonempty with* θ_k *but is empty with* $\min\{1.1\theta_k, 1\}$, *and we can always achieve it based on Lemma 5. If* $\liminf_{k\to\infty} \theta_k = 0$, *then there exists an accumulation point* $\boldsymbol{x}^*$ *of* $\{\boldsymbol{x}_k\}$ *where EGMFCQ does not hold at* $\boldsymbol{x}^*$.

PROOF. Without the loss of generality, we assume that $\lim_{k\to\infty}\theta_k = 0$ and $\lim_{k\to\infty}\boldsymbol{x}_k = \boldsymbol{x}^*$. Let $l_k := \inf\{\|\boldsymbol{z}_k\|_2 : \boldsymbol{z}_k \text{ satisfies Condition (2.4) at } \boldsymbol{x}_k\}$. The construction of $\theta_k = \frac{\varepsilon}{\|\boldsymbol{z}_k\|_2}$ in Equation (A.1) shows that $\limsup_{k\to\infty} l_k = \infty$. Suppose that EGMFCQ holds at $\boldsymbol{x}^*$ and let $l^* = \|\boldsymbol{z}^*\|_2 < \infty$ for some $\boldsymbol{z}^*$ satisfying Condition (2.4) at $\boldsymbol{x}^*$. It is a contradiction to Lemma 7 as $\infty = \limsup_{k\to\infty} l_k \le l^* < \infty$. Therefore, EGMFCQ does not hold at $\boldsymbol{x}^*$. □

EGMFCQ and its multiple variants are common in constrained optimization algorithms, i.e., [11, 74]. According to the above proposition, EGMFCQ makes the relaxed SQP subproblem feasible. Instead of assuming the EGMFCQ at all iterates $\{\boldsymbol{x}_k\}$, which is difficult to verify in real applications, a weaker and more explicit assumption (Assumption 1) is made. Proposition 5 shows the reasonability of Assumption 1. To verify Assumption 1, as shown in the deterministic SQP (RelaxedSQP, Algorithm 1), we first validate and adopt a feasible $\widetilde{\Omega}_k$ with proper relaxation parameters θ_k. If $\widetilde{\Omega}_k$ is not feasible for small θ_k below the predefined tolerance, then $\boldsymbol{x}_k$ is close to a point where EGMFCQ does not hold, by Lemma 1. For completeness, we put the definition of LICQ here.

DEFINITION 2 (Linear independence constraint qualification (LICQ)). *The linear independence constraint qualification (LICQ) is satisfied at a point $\tilde{\boldsymbol{x}}$, if columns of $[\nabla \boldsymbol{c}(\tilde{\boldsymbol{x}}), \boldsymbol{I}_{\mathcal{A}(\tilde{\boldsymbol{x}})}]$ are linearly independent, where $\mathcal{A}(\tilde{\boldsymbol{x}}) := \{i : (\tilde{\boldsymbol{x}})_i = (\boldsymbol{\ell})_i \text{ or } (\tilde{\boldsymbol{x}})_i = (\boldsymbol{u})_i\}$ is the active set of inequality constraints at $\tilde{\boldsymbol{x}}$.*

A.2. EGMFCQ and Boundedness of Lagrangian Multipliers. The following Lemma 9 shows that if the sequence $\{\boldsymbol{x}_k\}$ generated by the algorithm is convergent to a feasible point $\boldsymbol{x}^*$ satisfying EGMFCQ (Definition 1), then the corresponding Lagrangian multipliers of the SQP subproblem are bounded.

LEMMA 9. *If EGMFCQ is satisfied at $\bar{\boldsymbol{x}}$ which is feasible for both the equality and inequality constraints (i.e., $\boldsymbol{c}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$ and $\boldsymbol{\ell} \le \bar{\boldsymbol{x}} \le \boldsymbol{u}$), then there exists a neighborhood $\mathcal{B}(\bar{\boldsymbol{x}}; r_0) := \{\boldsymbol{x} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \le r_0\}$ with some $r_0 > 0$, such that the Lagrangian multipliers of the SQP subproblems are bounded for all points in $\mathcal{N}(\bar{\boldsymbol{x}}; r_0)$, under Assumptions 1 and 2.*

PROOF. We prove it by contradiction. Suppose that there exist sequences $\{(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{B}}_k, \bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})\}$ with Assumptions 1 and 2, such that $\bar{\boldsymbol{x}}_k \to \bar{\boldsymbol{x}}$, $\left\|(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})\right\|_2 \to \infty$ and $\kappa_1 \mathbf{I} \preceq \bar{\boldsymbol{B}}_k \preceq \kappa_2 \mathbf{I}$, where $\bar{\boldsymbol{p}}_k$ and $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})$ are the solution and the Lagrangian multipliers of the SQP subproblem at $\bar{\boldsymbol{x}}_k$ with corresponding relaxing parameters $\bar{\theta}_k$ satisfying

$$
\begin{aligned}
&\nabla f(\bar{\boldsymbol{x}}_k) + \bar{\boldsymbol{B}}_k \bar{\boldsymbol{p}}_k + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}}_k) \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} + \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} = \boldsymbol{0},\\
&\bar{\theta}_k \boldsymbol{c}(\bar{\boldsymbol{x}}_k) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}}_k)^\top \bar{\boldsymbol{p}}_k = \boldsymbol{0}, \quad \boldsymbol{\ell} \le \bar{\boldsymbol{x}}_k + \bar{\boldsymbol{p}}_k \le \boldsymbol{u},\\
&\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}\top} (\bar{\boldsymbol{x}}_k + \bar{\boldsymbol{p}}_k - \boldsymbol{\ell}) = 0,\\
&\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}\top} (\bar{\boldsymbol{x}}_k + \bar{\boldsymbol{p}}_k - \boldsymbol{u}) = 0,\\
&\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} \ge \boldsymbol{0}, \quad \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} \ge \boldsymbol{0}.
\end{aligned}
\tag{A.2}
$$

Note that the sequence $\{(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}}) / \left\|(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})^\top\right\|_2\}$ is bounded. Without the loss of generality, we assume that $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}}) / \left\|(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})^\top\right\|_2 \to (\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2)$, $\bar{\boldsymbol{p}}_k \to \bar{\boldsymbol{p}}$ and $\bar{\theta}_k = \bar{\theta}$ (due to line 4 in Algorithm 1). Then dividing both two sides of the first equality in Equation (A.2) by $\left\|(\bar{\boldsymbol{\lambda}}_k, \bar{\boldsymbol{\mu}}_{1,k}, \bar{\boldsymbol{\mu}}_{2,k})^\top\right\|_2$ and taking the limit of $k \to \infty$, we have

$$
\nabla \boldsymbol{c}(\bar{\boldsymbol{x}}) \bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\mu}}_1 + \bar{\boldsymbol{\mu}}_2 = \boldsymbol{0}. \tag{A.3}
$$

Moreover, the second equality in Equation (A.2) implies that

$$\bar{\theta}\bar{\boldsymbol{\lambda}}^\top \boldsymbol{c}(\bar{\boldsymbol{x}}) = -\bar{\boldsymbol{\lambda}}^\top \nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{p}}.$$

The third and the fourth equality in Equation (A.2) further shows that

$$\bar{\boldsymbol{\mu}}_1^\top (\bar{\boldsymbol{x}} - \boldsymbol{\ell}) = -\bar{\boldsymbol{\mu}}_1^\top \bar{\boldsymbol{p}} \quad \text{and} \quad \bar{\boldsymbol{\mu}}_2^\top (\bar{\boldsymbol{x}} - \boldsymbol{u}) = -\bar{\boldsymbol{\mu}}_2^\top \bar{\boldsymbol{p}}.$$

Combing with the above four equalities, we have

$$\bar{\theta}\bar{\boldsymbol{\lambda}}^\top \boldsymbol{c}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{\mu}}_1^\top (\boldsymbol{\ell} - \bar{\boldsymbol{x}}) + \bar{\boldsymbol{\mu}}_2^\top (\bar{\boldsymbol{x}} - \boldsymbol{u}) = 0. \tag{A.4}$$

Note that $\bar{\boldsymbol{\mu}}_1 \geq \mathbf{0}$ and $\bar{\boldsymbol{\mu}}_2 \geq \mathbf{0}$, we can deduce from Equation (A.4) and $\boldsymbol{c}(\bar{\boldsymbol{x}}) = \mathbf{0}$ that $(\bar{\boldsymbol{\mu}}_1) > 0$ only if $(\bar{\boldsymbol{x}})_i = (\boldsymbol{\ell})_i$ and $(\bar{\boldsymbol{\mu}}_2) > 0$ only if $(\bar{\boldsymbol{x}})_i = (\boldsymbol{u})_i$. The EGMFCQ condition at $\bar{\boldsymbol{x}}$ (Definition 1) implies that there exists $\boldsymbol{p} \in \mathbb{R}^d$ such that $\boldsymbol{c}(\bar{\boldsymbol{x}}) + \nabla \boldsymbol{c}(\bar{\boldsymbol{x}})^\top \boldsymbol{p} = \mathbf{0}$, $(\boldsymbol{p})_i > 0$ if $(\bar{\boldsymbol{x}})_i = (\boldsymbol{\ell})_i$, and $(\boldsymbol{p})_i < 0$ if $(\bar{\boldsymbol{x}})_i = (\boldsymbol{u})_i$. Then $-\boldsymbol{p}^\top \bar{\boldsymbol{\mu}}_1 + \boldsymbol{p}^\top \bar{\boldsymbol{\mu}}_2 < 0$ if $\bar{\boldsymbol{x}}$ is on the boundary of the box constraints. Multiplying both two sides of Equation (A.3) by $-\bar{\theta}\boldsymbol{p}$, we have $0 = -\bar{\theta}\boldsymbol{p}^\top (\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\mu}}_1 + \bar{\boldsymbol{\mu}}_2) = \bar{\theta}\boldsymbol{c}(\bar{\boldsymbol{x}})^\top \bar{\boldsymbol{\lambda}} + \bar{\theta}\boldsymbol{p}^\top \bar{\boldsymbol{\mu}}_1 - \bar{\theta}\boldsymbol{p}^\top \bar{\boldsymbol{\mu}}_2$. It is a contradiction to Equation (A.4). On the other hand, if $\bar{\boldsymbol{x}}$ is in the interior of the box constraints, $\bar{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{\mu}}_2 = \mathbf{0}$. Together with Equation (A.3), the linear independence of the columns of $\nabla \boldsymbol{c}(\bar{\boldsymbol{x}})$ shows $\bar{\boldsymbol{\lambda}} = \mathbf{0}$, which is a contradiction to the fact that $\|(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2)^\top\|_2 = 1$. □

COROLLARY 2. *If all accumulation points of the sequence $\{\boldsymbol{x}_k\}$ are feasible and satisfy EGMFCQ, then the Lagrangian multipliers of the corresponding SQP subproblems are bounded.*

PROOF. We first show that the Lagrangian multipliers of the corresponding SQP subproblems are bounded at all accumulation points of $\{\boldsymbol{x}_k\}$, denoted as $\mathcal{X}$. Note that the set $\mathcal{X}$ is closed, any accumulation point of $\mathcal{X}$ is also an accumulation point of $\{\boldsymbol{x}_k\}$.

Secondly, by Lemma 9, for a sufficiently large number $M_{\text{Lag}} > 0$ and any point $\boldsymbol{x}_i^* \in \mathcal{X}$, there exists $r_i > 0$ such that the Lagrangian multipliers of the corresponding SQP subproblems are bounded at $\boldsymbol{x}$ for any $\boldsymbol{x} \in \cup_{i=1}^{\infty} \mathcal{B}(\boldsymbol{x}_i^*; r_i)$. There must be a finite subset of $\{\boldsymbol{x}_k\}$, that is outside $\cup_{i=1}^{\infty} \mathcal{B}(\boldsymbol{x}_i^*; r_i)$ (otherwise, we can still find an accumulation point). We complete the proof. □

A.3. Proof for Theorem 1. The proof directly comes from the following lemmas. The first lemma here shows that the directional derivative of the merit function is controlled by the improvement $\Delta q(\boldsymbol{x}, \boldsymbol{p}, \nabla f(\boldsymbol{x}), \boldsymbol{B}, \rho)$.

LEMMA 10. *Under Assumption 2, given $(\boldsymbol{x}, \rho, \theta, \boldsymbol{B}, \boldsymbol{p}) \in \mathbb{R}^n \times \mathbb{R}_{>0} \times (0, 1] \times \mathbb{S}_+^n \times \mathbb{R}^n$ with $\theta \boldsymbol{c}(\boldsymbol{x}) + \nabla \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{p} = \mathbf{0}$, then the directional derivative of $\phi(\boldsymbol{x}, \rho)$ along $\boldsymbol{p}$ satisfies*

$$\begin{aligned} \phi'(\boldsymbol{x}, \rho; \boldsymbol{p}) &= \nabla f(\boldsymbol{x})^\top \boldsymbol{p} - \rho\theta \|\boldsymbol{c}(\boldsymbol{x})\|_2 \\ &\leq \nabla f(\boldsymbol{x})^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}\boldsymbol{p} - \rho\theta \|\boldsymbol{c}(\boldsymbol{x})\|_2 \\ &= -\Delta q(\boldsymbol{x}, \boldsymbol{p}, \nabla f(\boldsymbol{x}), \boldsymbol{B}, \rho). \end{aligned} \tag{A.5}$$

PROOF. We prove it by the definition of the directional derivative. Suppose that $\|\nabla^2 f(\boldsymbol{x})\|_2 \leq M$ for some $M > 0$. First,

$$
\begin{aligned}
&\phi(\boldsymbol{x}+\alpha\boldsymbol{p},\rho)-\phi(\boldsymbol{x},\rho)\\
=&f(\boldsymbol{x}+\alpha\boldsymbol{p})+\rho\|\boldsymbol{c}(\boldsymbol{x}+\alpha\boldsymbol{p})\|_2-f(\boldsymbol{x})-\rho\|\boldsymbol{c}(\boldsymbol{x})\|_2\\
\leq&\alpha\nabla f(\boldsymbol{x})^\top\boldsymbol{p}+\frac{\kappa_{\nabla f}}{2}\alpha^2\|\boldsymbol{p}\|_2^2+\rho\|\boldsymbol{c}(\boldsymbol{x}+\alpha\boldsymbol{p})\|_2-\rho\|\boldsymbol{c}(\boldsymbol{x})\|_2\\
=&\alpha\nabla f(\boldsymbol{x})^\top\boldsymbol{p}+\frac{\kappa_{\nabla f}}{2}\alpha^2\|\boldsymbol{p}\|_2^2+\rho\left(|1-\alpha\theta|-1\right)\|\boldsymbol{c}(\boldsymbol{x})\|_2+\frac{\kappa_{\nabla c}}{2}\alpha^2\|\boldsymbol{p}\|_2^2\\
=&\alpha\left(\nabla f(\boldsymbol{x})^\top\boldsymbol{p}-\rho\theta\|\boldsymbol{c}(\boldsymbol{x})\|_2\right)+\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha^2\|\boldsymbol{p}\|_2^2.
\end{aligned}
\tag{A.6}
$$

On the other side, similarly, we have $\phi(\boldsymbol{x}+\alpha\boldsymbol{p},\rho)-\phi(\boldsymbol{x},\rho)\geq\alpha\left(\nabla f(\boldsymbol{x})^\top\boldsymbol{p}-\rho\theta\|\boldsymbol{c}(\boldsymbol{x})\|_2\right)-\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha^2\|\boldsymbol{p}\|_2^2$. Taking limits for $\alpha\to 0$ and the definition, we have $\phi'(\boldsymbol{x},\rho;\boldsymbol{p})=\nabla f(\boldsymbol{x})^\top\boldsymbol{p}-\rho\theta\|\boldsymbol{c}(\boldsymbol{x})\|_2\leq-\Delta q(\boldsymbol{x},\boldsymbol{p},\nabla f(\boldsymbol{x}),\boldsymbol{B},\rho)$. □

We incorporate a backtracking line search in the algorithm while [6] adopted Lipschitz constant estimation for step size selection. We prove that under mild smoothness conditions, the line search condition will be met after a finite number of search steps. Specifically, the backtracking search loop is guaranteed to terminate within a bounded number of iterations.

LEMMA 11. *The strategies in Equations (2.5) and (2.6) for ρ_k guarantee that $\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\boldsymbol{B}_k;\rho_k)\geq\frac{1}{2}\boldsymbol{p}_k^\top\boldsymbol{B}_k\boldsymbol{p}_k+\sigma\rho_k\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$ for some $\sigma\in(0,1)$. Therefore, combining it with Lemma 10, we have that the backtracking line search condition $\phi(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k,\rho_k)\leq\phi(\boldsymbol{x}_k,\rho_k)-\beta\alpha_k\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)$ always holds for $\alpha_k\leq\frac{(1-\beta)\kappa_1}{\kappa_{\nabla f}+\kappa_{\nabla c}}$.*

PROOF. Equation (A.6) in Lemma 10 shows that $\phi(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k,\rho_k)-\phi(\boldsymbol{x}_k,\rho_k)\leq\alpha_k\left(\nabla f(\boldsymbol{x}_k)^\top\boldsymbol{p}_k-\rho_k\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2\right)+\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha_k^2\|\boldsymbol{p}_k\|_2^2$. Here, we let α_k to be small enough such that

$$
\begin{aligned}
&\alpha_k\left(\nabla f(\boldsymbol{x}_k)^\top\boldsymbol{p}_k-\rho_k\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2\right)+\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha_k^2\|\boldsymbol{p}_k\|_2^2\\
\leq&-\alpha_k\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)+\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha_k^2\|\boldsymbol{p}_k\|_2^2\\
\leq&-\beta\alpha_k\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k),
\end{aligned}
$$

i.e., $\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha_k\|\boldsymbol{p}_k\|_2^2\leq(1-\beta)\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)$. Here, we let $\frac{\kappa_{\nabla f}+\kappa_{\nabla c}}{2}\alpha_k\|\boldsymbol{p}_k\|_2^2\leq\frac{(1-\beta)\kappa_1}{2}\|\boldsymbol{p}_k\|_2^2\leq\frac{1-\beta}{2}\boldsymbol{p}_k^\top\boldsymbol{B}_k\boldsymbol{p}_k$, i.e., $\alpha_k\leq\frac{(1-\beta)\kappa_1}{\kappa_{\nabla f}+\kappa_{\nabla c}}$. In conclusion, the backtracking line search condition holds when $\alpha_k\leq\frac{(1-\beta)\kappa_1}{\kappa_{\nabla f}+\kappa_{\nabla c}}$. □

The next lemma demonstrates that if the Lagrange multipliers are bounded, then the penalty parameter will stabilize. This result is crucial for the global convergence of the algorithm, as convergence is only assured subsequent to the penalty parameter's stabilization. Specifically, once the penalty parameter stabilizes, the merit function's convergence naturally leads to the convergence of the iterates.

LEMMA 12. *Under Assumption 1, $\theta_k\geq\tilde{\tau}\tilde{\theta}$ holds for all $k=0,1,\cdots$. If we further assume that Assumption 2 holds, then the sequence $\{\rho_k\}$ is monotonically increasing and there exists a large enough $\widetilde{K}\in\mathbb{Z}$, such that $\rho_k=\tilde{\rho}>0$ for all $k\geq\widetilde{K}$, where $\tilde{\rho}\leq\frac{(1+\epsilon)M_{Lag}}{(1-\sigma)\tilde{\tau}\tilde{\theta}}$.*

PROOF. Under Assumption 1, it is obvious that $\theta_k \geq \tilde{\tau}\tilde{\theta}$ holds in our algorithm, for all $k = 0, 1, \cdots$. If there does not exist $\tilde{\rho} > 0$ and $\widetilde{K} \in \mathbb{Z}$ such that $\rho_k = \tilde{\rho} > 0$ for $k \geq \widetilde{K}$, according to Equation (2.6), then there is an infinite sequence $\{k_j\} \subseteq \mathbb{Z}_+$ where $\rho_{k_j}^{\text{trial}} > \rho_{k_j-1}$ and $\rho_{k_j} = (1+\epsilon)\rho_{k_j}^{\text{trial}}$. It further implies that $-\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k < 0$ and $\rho_{k_j}^{\text{trial}} = \frac{\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k}{(1-\sigma)\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2}$, by Equation (2.5). The KKT conditions for the relaxed SQP Subproblem (2.7) show that there exist some $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}})$ satisfying

$$
\begin{aligned}
&\nabla f(\boldsymbol{x}_k) + \boldsymbol{B}_k \boldsymbol{p}_k + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}} + \boldsymbol{\mu}_{2,k}^{\text{sub}} = \boldsymbol{0},\\
&\theta_k \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p}_k = \boldsymbol{0},\\
&\boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{p}_k \leq \boldsymbol{u},\\
&\boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{x}_k + \boldsymbol{p}_k - \boldsymbol{\ell}) = 0,\\
&\boldsymbol{\mu}_{2,k}^{\text{sub}\top}(\boldsymbol{x}_k + \boldsymbol{p}_k - \boldsymbol{u}) = 0,\\
&\boldsymbol{\mu}_{1,k}^{\text{sub}} \geq \boldsymbol{0} \text{ and } \boldsymbol{\mu}_{2,k}^{\text{sub}} \geq \boldsymbol{0}.
\end{aligned}
\tag{A.7}
$$

Multiplying both two sides of the first equality by $\boldsymbol{p}_k$, we have

$$
\begin{aligned}
\nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k &= -\boldsymbol{p}_k^\top \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k^{\text{sub}} + \boldsymbol{p}_k^\top \boldsymbol{\mu}_{1,k}^{\text{sub}} - \boldsymbol{p}_k^\top \boldsymbol{\mu}_{2,k}^{\text{sub}}\\
&= \theta_k \boldsymbol{\lambda}_k^{\text{sub}\top} \boldsymbol{c}(\boldsymbol{x}_k) - \boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{x}_k - \boldsymbol{\ell}) + \boldsymbol{\mu}_{2,k}^{\text{sub}\top}(\boldsymbol{x}_k - \boldsymbol{u})\\
&\leq \theta_k \boldsymbol{\lambda}_k^{\text{sub}\top} \boldsymbol{c}(\boldsymbol{x}_k) \leq \|\boldsymbol{\lambda}_k^{\text{sub}}\|_2 \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 \leq M_{\text{Lag}} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2,
\end{aligned}
$$

where the first inequality comes from $\boldsymbol{\mu}_{1,k}^{\text{sub}} \geq \boldsymbol{0}$, $\boldsymbol{\mu}_{2,k}^{\text{sub}} \geq \boldsymbol{0}$, and $\boldsymbol{\ell} \leq \boldsymbol{x}_k \leq \boldsymbol{u}$. Then,

$$
\rho_{k_j-1} < \rho_{k_j}^{\text{trial}} = \frac{\nabla f(\boldsymbol{x}_{k_j})^\top \boldsymbol{p}_{k_j} + \boldsymbol{p}_{k_j}^\top \boldsymbol{B}_{k_j} \boldsymbol{p}_{k_j}}{(1-\sigma)\theta_{k_j} \|\boldsymbol{c}(\boldsymbol{x}_{k_j})\|_2} \leq \frac{M_{\text{Lag}}}{(1-\sigma)\theta_{k_j}} \leq \frac{M_{\text{Lag}}}{(1-\sigma)\tilde{\tau}\tilde{\theta}}. \tag{A.8}
$$

However, $\rho_{k_j} = (1+\epsilon)\rho_{k_j}^{\text{trial}} > (1+\epsilon)\rho_{k_j-1}$ implies that $\rho_{k_j-1} \to \infty$ as $k_j \to \infty$. It is a contradiction. Therefore, there exist $\tilde{\rho} > 0$ and a large enough $\widetilde{K} \in \mathbb{Z}$, such that $\rho_k = \tilde{\rho} > 0$ for all $k \geq \widetilde{K}$. Here, we can also conclude from Equation (A.8) that $\tilde{\rho} \leq \frac{(1+\epsilon)M_{\text{Lag}}}{(1-\sigma)\tilde{\tau}\tilde{\theta}}$. □

PROPOSITION 1. *If we suppose that all accumulation points of the generated sequence* $\{\boldsymbol{x}_k\}$ *satisfies EGMFCQ, then* $\lim_k \rho_k < \infty$.

PROOF. Suppose that $\lim_{k\to\infty} \rho_k = \infty$, then we can find a subsequence $\{k_j\} \subseteq \mathbb{Z}_+$ such that $\rho_{k_j} > \rho_{k_j-1}$ and $\rho_k = \rho_{k-1}$ for $k \notin \{k_j\}$. By the fact that

$$
\rho_{k_j}^{\text{trial}} = \frac{\nabla f(\boldsymbol{x}_{k_j})^\top \boldsymbol{p}_{k_j} + \boldsymbol{p}_{k_j}^\top \boldsymbol{B}_{k_j} \boldsymbol{p}_{k_j}}{(1-\sigma)\theta_{k_j} \|\boldsymbol{c}(\boldsymbol{x}_{k_j})\|_2} \leq \frac{M_{\nabla f} M_{\boldsymbol{\ell},\boldsymbol{u}} + \kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}}^2}{(1-\sigma)\tilde{\tau}\tilde{\theta}\|\boldsymbol{c}(\boldsymbol{x}_{k_j})\|_2},
$$

we have $\lim_{j\to\infty} \left\|\boldsymbol{c}(\boldsymbol{x}_{k_j})\right\|_2 = 0$. By Lemmas 9 and 12, it is a contradiction. □

Proposition 1 shows the boundedness of the penalty parameters from the constraint qualification perspective. More specifically, if EGMFCQ holds for all accumulation points of the sequence $\{\boldsymbol{x}_k\}$, then the penalty parameter is guaranteed to be bounded. Given that we update the penalty parameter by multiplying it by a factor greater than one, it follows that the penalty parameter will eventually stabilize.

LEMMA 13. *Under Assumptions 1 and 2, there exist sufficiently large $\widetilde{K} \in \mathbb{Z}_+$ and $\tilde{\rho} > 0$, such that $\rho_k = \tilde{\rho}$ for all $k \geq \widetilde{K}$ and*

$$\phi(\boldsymbol{x}_k, \tilde{\rho}) - \phi(\boldsymbol{x}_{k+1}, \tilde{\rho}) \geq \frac{\beta(1-\beta)\tau\kappa_1\tilde{\rho}\tilde{\tau}\tilde{\theta}\sigma}{\kappa_{\nabla f} + \kappa_{\nabla c}}\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 + \frac{\beta(1-\beta)\kappa_1^2}{2(\kappa_{\nabla f} + \kappa_{\nabla c})}\|\boldsymbol{p}_k\|_2^2. \tag{A.9}$$

PROOF. By Lemma 12, the penalty parameter ρ_k becomes stable when $k \geq \widetilde{K}$ for some sufficiently large $\widetilde{K} \in \mathbb{Z}_+$, i.e., $\rho_k = \tilde{\rho}$ for $k \geq \widetilde{K}$. Next, we only consider the iterates when ρ_k becomes stable. The backtracking line search guarantees that

$$\phi(\boldsymbol{x}_k, \tilde{\rho}) - \phi(\boldsymbol{x}_{k+1}, \tilde{\rho}) \geq \beta\alpha_k \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \tilde{\rho}) \geq \beta\alpha_k \cdot \left(\frac{1}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \sigma\tilde{\rho}\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2\right).$$

By Lemma 11, we have $\alpha_k \geq \frac{\tau(1-\beta)\kappa_1}{\kappa_{\nabla f} + \kappa_{\nabla c}}$ by the backtracking line search. Furthermore, by the positive-definiteness of $\boldsymbol{B}_k$ (i.e., $\boldsymbol{B}_k \succeq \kappa_1 \mathbf{I}$) and the lower-boundedness of θ_k (i.e., $\theta_k \geq \tilde{\tau}\tilde{\theta}$), together with the stabilization of ρ_k (i.e., $\rho_k = \tilde{\rho}$) and the lower-boundedness of α_k (i.e., $\alpha_k \geq \frac{\tau(1-\beta)\kappa_1}{\kappa_{\nabla f} + \kappa_{\nabla c}}$), we complete the proof for Equation (A.9). □

Proof for Theorem 1. It is a direct result of Lemma 13. Here, we only consider the case where the penalty parameter ρ_k becomes stable. By the boundedness of the feasible region (i.e., $\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}$) and smoothness of the objective and the constraints, we have that $\phi(\boldsymbol{x}, \tilde{\rho})$ is (lower and upper) bounded. Then Equation (A.9) implies that

$$\frac{\beta(1-\beta)\tau\kappa_1\tilde{\rho}\tilde{\tau}\tilde{\theta}\sigma}{\kappa_{\nabla f} + \kappa_{\nabla c}} \sum_{k=\widetilde{K}}^{\infty} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 + \frac{\beta(1-\beta)\kappa_1^2}{2(\kappa_{\nabla f} + \kappa_{\nabla c})} \sum_{k=\widetilde{K}}^{\infty} \|\boldsymbol{p}_k\|_2^2 < \infty,$$

which completes the proof for Equation (2.12). Conditions in Equation (A.7) show that $\|\nabla f(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}} + \boldsymbol{\mu}_{2,k}^{\text{sub}}\|_2 = \|\boldsymbol{B}_k\boldsymbol{p}_k\|_2 \leq \kappa_2\|\boldsymbol{p}_k\|_2$, $\|\boldsymbol{\mu}_{1,k}^{\text{sub}} \odot (\boldsymbol{x} - \boldsymbol{\ell})\|_2 \leq \boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{x} - \boldsymbol{\ell}) \leq M_{\text{Lag}}\|\boldsymbol{p}_k\|_2$ and $\|\boldsymbol{\mu}_{2,k}^{\text{sub}} \odot (\boldsymbol{x} - \boldsymbol{u})\|_2 \leq \boldsymbol{\mu}_{1,k}^{\text{sub}\top}(\boldsymbol{u} - \boldsymbol{x}) \leq M_{\text{Lag}}\|\boldsymbol{p}_k\|_2$, then Equation (2.13) is straightforward.

A.4. Proof for Lemma 2. Denote $\mathcal{A}^* = \mathcal{A}(\boldsymbol{x}^*) := \{i : (\boldsymbol{x}^*)_i = (\boldsymbol{\ell})_i \text{ or } (\boldsymbol{x}^*)_i = (\boldsymbol{\ell})_i\}$ the active set of inequality constraints at $\boldsymbol{x}^*$. Denote $\varepsilon = \min\{(\boldsymbol{x}^* - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}^*)_i, i \in \mathcal{A}^{*-}\} > 0$. First, let $\boldsymbol{x}_k$ be sufficiently close to $\boldsymbol{x}^*$ such that $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_\infty \leq \frac{1}{4}\varepsilon$, then $\min\{(\boldsymbol{x}_k - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}_k)_i, i \in \mathcal{A}^{*-}\} \geq \frac{3}{4}\varepsilon$. Since EGMFCQ holds at $\boldsymbol{x}^*$, there exists a vector $\boldsymbol{z}^* \in \mathbb{R}^d$ satisfying Condition (2.4) at $\boldsymbol{x}^*$. The fact that $\boldsymbol{c}(\boldsymbol{x}^*) = \mathbf{0}$ further implies that we can scale the vector $\boldsymbol{z}^*$ by some constants such that

$$\begin{aligned}
&\boldsymbol{c}(\boldsymbol{x}^*) + \nabla \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{z}^* = \mathbf{0},\\
&(\boldsymbol{z}^*)_i > 0, \text{ if } (\boldsymbol{x}^*)_i = (\boldsymbol{\ell})_i,\\
&(\boldsymbol{z}^*)_i < 0, \text{ if } (\boldsymbol{x}^*)_i = (\boldsymbol{u})_i,\\
&\|\boldsymbol{z}^*\|_\infty \leq \varepsilon/2.
\end{aligned}$$

By the smoothness of $\boldsymbol{c}(\boldsymbol{x})$ and the linear independence of columns of $\nabla \boldsymbol{c}(\boldsymbol{x}^*)$, we can find $\boldsymbol{z}_k$ such that $\boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{z}_k = \mathbf{0}$ and $\|\boldsymbol{z}_k - \boldsymbol{z}^*\|_\infty \leq \frac{1}{2}\min\{|(\boldsymbol{z}^*)_i| : (\boldsymbol{x}^*)_i = (\boldsymbol{\ell})_i \text{ or } (\boldsymbol{x}^*)_i = (\boldsymbol{u})_i\} \leq \frac{1}{4}\varepsilon$ as $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 \to 0$. Then $\|\boldsymbol{z}_k\|_\infty \leq \|\boldsymbol{z}_k - \boldsymbol{z}^*\|_\infty + \|\boldsymbol{z}^*\|_\infty \leq \frac{3}{4}\varepsilon$; $(\boldsymbol{z}_k)_i > 0$, if $(\boldsymbol{x}^*)_i = (\boldsymbol{\ell})_i$ and $(\boldsymbol{z}_k)_i < 0$, if $(\boldsymbol{x}^*)_i = (\boldsymbol{u})_i$. Together with the fact that $\min\{(\boldsymbol{x}_k - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}_k)_i, i \in \mathcal{A}^{*-}\} \geq \frac{3}{4}\varepsilon$, we show $\boldsymbol{\ell} \leq \boldsymbol{x}_k + \boldsymbol{z}_k \leq \boldsymbol{u}$. Therefore, $\theta_k = 1$ is always accepted if $\boldsymbol{x}_k$ is sufficiently close to $\boldsymbol{x}^*$.

APPENDIX B: PROOF FOR THEOREM 2 AND 3

B.1. Some Technical Lemmas for Theorem 2. We first show that the adaptivity parameter will stabilize after sufficient iterations.

LEMMA 14. *Under Assumption 3, there exists a constant $\bar{\xi} > 0$ such that $\xi_k = \bar{\xi}$ for all sufficiently large k.*

PROOF. Observe that the sequence $\{\xi_k\}$ is monotonically decreasing and $\xi_k < \xi_{k-1}$ holds if and only if $\xi_k^{\text{trial}} < \xi_{k-1}$ and $\xi_k \leq (1-\epsilon_\xi)\xi_{k-1}$. Suppose $\lim_{k\to\infty} \xi_k = 0$, then it follows that $\liminf_{k\to\infty} \xi_k^{\text{trial}} = 0$. However, the selection of ρ_k guarantees that $\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \bar{\boldsymbol{g}}_k, \boldsymbol{B}_k, \rho_k) \geq \frac{1}{2}\bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k \geq \frac{\kappa_1}{2}\|\bar{\boldsymbol{p}}_k\|_2^2$, implying that $\xi_k^{\text{trial}} \geq \frac{\kappa_1}{2}$. It is a contradiction. Therefore, we conclude that $\lim_{k\to\infty} \xi_k > 0$. □

The following lemma is essential in our subsequent analysis and is extended from Lemma A.3 in [51]. The results investigate the competition and reveal the asymptotic behavior between the two sequences $\{\alpha_k\}$ and $\{\beta_k\}$. Importantly, we observe that when $\{\alpha_k\}$ decays faster than $\{\beta_k\}$, the asymptotic behavior of terms described in the lemma is dominated by the sequence $\{\alpha_k\}$, resulting in the asymptotic normality of the generated iterates with averaged gradient as studied in Section 4.

LEMMA 15 (Lemma A.3 in [51]). *For two sequences $\{\alpha_k\}$ and $\{\beta_k\}$ satisfying $\alpha_k = \iota_1(k+1)^{-b_1}$ and $\beta_k = \iota_2(k+1)^{-b_2}$ with $\iota_1, \iota_2 > 0$ and $b_1, b_2 > 0$, the followings hold*

1. *Define $\chi = 0$ if $0 < b_2 < 1$ and $\chi = -\frac{b_1}{\iota_2}$ if $b_2 = 1$, then*

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

where we require that $\sum_{t=1}^{l} a_t + \chi > 0$. Moreover,

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

for any $b \in \mathbb{R}$ and $e_i \to 0$.

2. *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 16. *For two given sequence $\{\alpha_k\}$ and $\{\beta_k\}$ satisfying $\lim_{k\to\infty} \alpha_k = 0$, $\lim_{k\to\infty} \beta_k = 0$, and $\lim_{k\to\infty} \alpha_k/\beta_k = 0$, then*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] = 0. \tag{B.1}$$

Therefore, there exists a number $M_\sigma > 0$ such that

$$\mathbb{E}\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] \leq M_\sigma^2,$$

under Assumption 3.

PROOF. By the update scheme of $\bar{\boldsymbol{g}}_k$, we have

$$\begin{aligned}
\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k) &= \beta_k \left(\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\right) + (1-\beta_k)\left(\bar{\boldsymbol{g}}_{k-1} - \nabla f(\boldsymbol{x}_{k-1})\right) \\
&\quad + (1-\beta_k)\left(\nabla f(\boldsymbol{x}_{k-1}) - \nabla f(\boldsymbol{x}_k)\right) \\
&= \beta_k \left(\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\right) + (1-\beta_k)\left\{\beta_{k-1}\left(\boldsymbol{g}_{k-1} - \nabla f(\boldsymbol{x}_{k-1})\right) + (1-\beta_{k-1})\left(\bar{\boldsymbol{g}}_{k-2} - \nabla f(\boldsymbol{x}_{k-2})\right)\right. \\
&\quad \left. + (1-\beta_{k-1})\left(\nabla f(\boldsymbol{x}_{k-2}) - \nabla f(\boldsymbol{x}_{k-1})\right)\right\} + (1-\beta_k)\left(\nabla f(\boldsymbol{x}_{k-1}) - \nabla f(\boldsymbol{x}_k)\right) \\
&= \cdots \\
&= \sum_{i=0}^{k} \left(\prod_{j=i+1}^{k} (1-\beta_j)\right) \beta_i \left(\boldsymbol{g}_i - \nabla f(\boldsymbol{x}_i)\right) \\
&\quad + \sum_{i=1}^{k} \left(\prod_{j=i}^{k} (1-\beta_j)\right) \left(\nabla f(\boldsymbol{x}_{i-1}) - \nabla f(\boldsymbol{x}_i)\right) \\
&:= \mathcal{W}_1 + \mathcal{W}_2.
\end{aligned}$$

Here, both $\mathcal{W}_1$ and $\mathcal{W}_2$ are random variables. By Lemma 15 and the fact that

$$\|\mathcal{W}_2\|_2 \leq \sum_{i=1}^{k} \left(\prod_{j=i}^{k} (1-\beta_j)\right) \alpha_{i-1} M_{\boldsymbol{\ell},\boldsymbol{u}},$$

we have $\mathcal{W}_2 \to 0$ as $k \to \infty$ since $\lim_{k\to\infty} \alpha_{i-1}/\beta_i = 0$. It follows that $\lim_{k\to\infty} \mathbb{E}\left[\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\right] = \lim_{k\to\infty} \mathbb{E}\left[\mathcal{W}_1\right] = 0$, since

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{W}_1\|_2^2\right] \\
=&\sum_{i=0}^{k} \left(\prod_{j=i+1}^{k} (1-\beta_j)\right)^2 \beta_i^2 \mathbb{E}\left[\|\boldsymbol{g}_i - \nabla f(\boldsymbol{x}_i)\|_2^2\right] \\
\leq&\sigma_g^2 \sum_{i=0}^{k} \left(\prod_{j=i+1}^{k} (1-\beta_j)\right)^2 \beta_i^2 \to 0 \text{ as } k \to \infty,
\end{aligned}$$

where the last convergence result comes from Lemma 15. Therefore, $\lim_{k\to\infty} \mathbb{E}\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] \leq 2\lim_{k\to\infty} \mathbb{E}\left[\|\mathcal{W}_1\|_2^2\right] + 2\lim_{k\to\infty} \|\mathcal{W}_2\|_2^2 = 0$, which completes the first part of the proof. The second result is straightforward since a convergent sequence must be bounded. □

The above lemma establishes the convergence of the averaged gradient to the exact gradient in expectation, by utilizing the asymptotic behavior of two sequences in Lemma 15. The lemma not only assures us of the asymptotic validity of using $\bar{\boldsymbol{g}}_k$ as a surrogate for $\nabla f(\boldsymbol{x}_k)$, but also offers a bound for their difference, lending confidence in the effectiveness of the algorithm. Following this, the next lemma studies the perturbation robustness property of the quadratic SQP subproblems and implies that the solutions are Lipschitz continuous with respect to the gradients. Consequently, the fact that the averaged gradient is asymptotically convergent to the exact gradient implies that the Debiased-StoSQP is arbitrarily close to the deterministic algorithm after sufficiently many iterations. This constitutes one of the most significant advantages of Debiased-StoSQP, employing averaged gradients, over other fully stochastic algorithms [6, 20].

LEMMA 17. *Suppose Assumptions 1 , 2 and 3 hold, then*

$$\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2 \leq \kappa_1^{-1} \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 , \tag{B.2}$$

and

$$\mathbb{E}_k \|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2 \leq \kappa_1^{-1} \mathbb{E}_k \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 .$$

PROOF. The relaxed SQP subproblem at $\boldsymbol{x}_k$ with the averaged gradient $\bar{\boldsymbol{g}}_k$ can be written as

$$\min_{\boldsymbol{p}\in\widetilde{\Omega}_k} \frac{1}{2} \left\|\boldsymbol{p} + \boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k\right\|_{\boldsymbol{B}_k}^2 ,$$

which is a convex-constrained quadratic problem. The variational inequality implies that

$$\left\langle \boldsymbol{p}_k - \bar{\boldsymbol{p}}_k, -\boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k - \bar{\boldsymbol{p}}_k \right\rangle_{\boldsymbol{B}_k} \leq 0.$$

Since $\boldsymbol{p}_k$ is the solution of the relaxed SQP subproblem at $\boldsymbol{x}_k$ with exact gradient $\nabla f(\boldsymbol{x}_k)$, we similarly have

$$\left\langle \bar{\boldsymbol{p}}_k - \boldsymbol{p}_k, -\boldsymbol{B}_k^{-1}\nabla f(\boldsymbol{x}_k) - \boldsymbol{p}_k \right\rangle_{\boldsymbol{B}_k} \leq 0.$$

Summing up the above two inequalities, we have

$$\begin{aligned} 0 &\geq \left\langle \boldsymbol{p}_k - \bar{\boldsymbol{p}}_k, -\boldsymbol{B}_k^{-1}\bar{\boldsymbol{g}}_k - \bar{\boldsymbol{p}}_k + \boldsymbol{B}_k^{-1}\nabla f(\boldsymbol{x}_k) + \boldsymbol{p}_k \right\rangle_{\boldsymbol{B}_k} \\ &= \|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k}^2 + \langle \bar{\boldsymbol{p}}_k - \boldsymbol{p}_k, \bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\rangle \\ &\geq \|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k}^2 - \|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\|_2 \cdot \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 . \end{aligned} \tag{B.3}$$

Note that $\|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k}^2 \geq \kappa_1 \|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\|_2^2$, combining with Assumption 3, we complete the proof. □

LEMMA 18. *Suppose that Assumptions 2 and 3 hold, then*

$$\mathbb{E}_k \left[\left| (\nabla f(\boldsymbol{x}_k) - \bar{\boldsymbol{g}}_k)^\top \bar{\boldsymbol{p}}_k \right| \right] \leq M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}}, \tag{B.4}$$

$$\mathbb{E}_k \left[\left| \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k \right| \right] \leq M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}} + 2\left(M_{\nabla f} + M_\sigma\right) M_{\boldsymbol{\ell},\boldsymbol{u}}, \tag{B.5}$$

and

$$\mathbb{E}_k \left[\left| \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k - \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k \right| \right] \leq 2\kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}} M_\sigma. \tag{B.6}$$

PROOF. The first relation is straightforward since $\mathbb{E}_k\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2\right] \leq M_\sigma$ and $\|\bar{\boldsymbol{p}}_k\|_2 \leq M_{\boldsymbol{\ell},\boldsymbol{u}}$. By triangle inequalities, we have

$$\begin{aligned} &\mathbb{E}_k \left[\left| \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k - \bar{\boldsymbol{g}}_k^\top \bar{\boldsymbol{p}}_k \right| \right] \\ =&\mathbb{E}_k \left[\left| (\nabla f(\boldsymbol{x}_k) - \bar{\boldsymbol{g}}_k)^\top \boldsymbol{p}_k + \bar{\boldsymbol{g}}_k^\top (\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k) \right| \right] \\ \leq& M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}} + 2\left(M_{\nabla f} + M_\sigma\right) M_{\boldsymbol{\ell},\boldsymbol{u}}, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_k \left[\left| \boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k - \bar{\boldsymbol{p}}_k^\top \boldsymbol{B}_k \bar{\boldsymbol{p}}_k \right| \right] \\ =&\mathbb{E}_k \left[\left| (\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k)^\top \boldsymbol{B}_k (\boldsymbol{p}_k + \bar{\boldsymbol{p}}_k) \right| \right] \\ \leq& 2\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}} \mathbb{E}_k \left[\|\boldsymbol{p}_k - \bar{\boldsymbol{p}}_k\| \right] \leq 2\kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}} M_\sigma. \end{aligned}$$

□

LEMMA 19. *Suppose that Assumptions 1 , 2 and 3 hold, then*

$$\begin{aligned}&\mathbb{E}_k\left[\left|\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)-\Delta q(\boldsymbol{x}_k,\bar{\boldsymbol{p}}_k,\bar{\boldsymbol{g}}_k,\boldsymbol{B}_k;\rho_k)\right|\right]\\ \leq& M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}}+2\left(M_{\nabla f}+M_\sigma\right)M_{\boldsymbol{\ell},\boldsymbol{u}}+\kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}}M_\sigma.\end{aligned}\tag{B.7}$$

PROOF. Note that $\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)=-\nabla f(\boldsymbol{x}_k)^\top\boldsymbol{p}_k-\frac{1}{2}\boldsymbol{p}_k^\top\boldsymbol{B}_k\boldsymbol{p}_k+\rho_k\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$, where the last term $\rho_k\theta_k\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2$ is independent of $\nabla f(\boldsymbol{x}_k)$ and $\boldsymbol{p}_k$. Using results in Lemma 18, we have

$$\begin{aligned}&\mathbb{E}_k\left[\left|\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\rho_k)-\Delta q(\boldsymbol{x}_k,\bar{\boldsymbol{p}}_k,\bar{\boldsymbol{g}}_k,\boldsymbol{B}_k;\rho_k)\right|\right]\\ =&\mathbb{E}_k\left[\left|-\nabla f(\boldsymbol{x}_k)^\top\boldsymbol{p}_k+\bar{\boldsymbol{g}}_k^\top\bar{\boldsymbol{p}}_k-\frac{1}{2}\boldsymbol{p}_k^\top\boldsymbol{B}_k\boldsymbol{p}_k+\frac{1}{2}\bar{\boldsymbol{p}}_k^\top\boldsymbol{B}_k\bar{\boldsymbol{p}}_k\right|\right]\\ \leq&\mathbb{E}_k\left[\left|-\nabla f(\boldsymbol{x}_k)^\top\boldsymbol{p}_k+\bar{\boldsymbol{g}}_k^\top\bar{\boldsymbol{p}}_k\right|\right]+\frac{1}{2}\mathbb{E}_k\left[\left|\boldsymbol{p}_k^\top\boldsymbol{B}_k\boldsymbol{p}_k-\bar{\boldsymbol{p}}_k^\top\boldsymbol{B}_k\bar{\boldsymbol{p}}_k\right|\right]\\ \leq&M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}}+2\left(M_{\nabla f}+M_\sigma\right)M_{\boldsymbol{\ell},\boldsymbol{u}}+\kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}}M_\sigma.\end{aligned}$$

□

LEMMA 20. *In line 10 of Algorithm 2, the step size α_k is selected from the interval $[\alpha_k^{min},\alpha_k^{max}]:=\left[\frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}},\frac{\xi_k\gamma_k}{\kappa_{\nabla f}+\rho_k\kappa_{\nabla c}}+\varrho\gamma_k^2\right]$, then*

$$\mathbb{E}_k\left[\alpha_k\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\right]\leq\varrho\gamma_k^2M_{\nabla f}\kappa_1^{-1}M_\sigma+\alpha_k^{min}\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\right],\tag{B.8}$$

under Assumptions 2 and 3.

PROOF. Denote the event $\mathcal{C}_k=\{\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\geq 0\}$, then

$$\begin{aligned}&\mathbb{E}_k\left[\alpha_k\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\right]\\ =&\mathbb{E}_k\left[\alpha_k\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)|\mathcal{C}_k\right]+\mathbb{E}_k\left[\alpha_k\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)|\mathcal{C}_k^c\right]\\ \leq&\alpha_k^{\max}\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)|\mathcal{C}_k\right]+\alpha_k^{\min}\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)|\mathcal{C}_k^c\right]\\ =&\alpha_k^{\min}\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\right]+\left(\alpha_k^{\max}-\alpha_k^{\min}\right)\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)|\mathcal{C}_k\right]\\ \leq&\alpha_k^{\min}\mathbb{E}_k\left[\nabla f(\boldsymbol{x}_k)^\top\left(\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\right)\right]+\varrho\gamma_k^2M_{\nabla f}\kappa_1^{-1}M_\sigma.\end{aligned}$$

□

LEMMA 21. *Under Assumptions 1, 2 and 3, if $\sum_{k=0}^\infty\gamma_k=\infty$, $\sum_{k=0}^\infty\gamma_k^2<\infty$ and $\sum_{k=\bar{K}}^\infty\alpha_k^{min}\mathbb{E}\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\right]<\infty$, then*

$$\sum_{k=\bar{K}}^\infty\alpha_k^{min}\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\bar{\rho})<\infty,\ \textit{almost surely}.\tag{B.9}$$

It further implies that

$$\liminf_{k\to\infty}\Delta q(\boldsymbol{x}_k,\boldsymbol{p}_k,\nabla f(\boldsymbol{x}_k),\boldsymbol{B}_k;\bar{\rho})=0,\ \textit{almost surely}.\tag{B.10}$$

PROOF. We only consider the case when ρ_k and ξ_k becomes stable, i.e., $\rho_k = \bar{\rho}$ and $\xi_k = \bar{\xi}$ for all $k \geq \bar{K}$. It follows from Assumptions 1, 2 and 3 that

(B.11)
$$
\begin{aligned}
&\mathbb{E}_k\left[\phi(\boldsymbol{x}_{k+1}, \bar{\rho}) - \phi(\boldsymbol{x}_k, \bar{\rho})\right] \\
=&\mathbb{E}_k\left[f(\boldsymbol{x}_k + \alpha_k \bar{\boldsymbol{p}}_k) - f(\boldsymbol{x}_k) + \bar{\rho}\left(\|\boldsymbol{c}(\boldsymbol{x}_k + \alpha_k \bar{\boldsymbol{p}}_k)\|_2 - \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2\right)\right] \\
\leq&\mathbb{E}_k\left[\alpha_k \nabla f(\boldsymbol{x}_k)^\top \bar{\boldsymbol{p}}_k + \frac{\kappa_{\nabla f}}{2}\alpha_k^2 \|\bar{\boldsymbol{p}}_k\|_2^2 + \bar{\rho}\left(\left\|\boldsymbol{c}(\boldsymbol{x}_k) + \alpha_k \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \bar{\boldsymbol{p}}_k\right\|_2 - \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 + \frac{\kappa_{\nabla c}}{2}\alpha_k^2 \|\bar{\boldsymbol{p}}_k\|_2^2\right)\right] \\
=&\mathbb{E}_k\left[\alpha_k \nabla f(\boldsymbol{x}_k)^\top \bar{\boldsymbol{p}}_k - \alpha_k \bar{\rho}\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 + \frac{\kappa_{\nabla f} + \bar{\rho}\kappa_{\nabla c}}{2}\alpha_k^2 \|\bar{\boldsymbol{p}}_k\|_2^2\right] \\
=&\mathbb{E}_k\left[\alpha_k \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p}_k + \alpha_k \nabla f(\boldsymbol{x}_k)^\top (\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k) - \alpha_k \bar{\rho}\theta_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 + \frac{\kappa_{\nabla f} + \bar{\rho}\kappa_{\nabla c}}{2}\alpha_k^2 \|\bar{\boldsymbol{p}}_k\|_2^2\right] \\
=&\mathbb{E}_k\left[-\alpha_k \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho}) - \frac{\alpha_k}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \alpha_k \nabla f(\boldsymbol{x}_k)^\top (\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k) + \frac{\kappa_{\nabla f} + \bar{\rho}\kappa_{\nabla c}}{2}\alpha_k^2 \|\bar{\boldsymbol{p}}_k\|_2^2\right] \\
\leq&\mathbb{E}_k\Big[-\alpha_k \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho}) - \frac{\alpha_k}{2}\boldsymbol{p}_k^\top \boldsymbol{B}_k \boldsymbol{p}_k + \alpha_k \nabla f(\boldsymbol{x}_k)^\top (\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k) \\
&\quad + \frac{1}{2}\alpha_k \gamma_k \Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \nabla \bar{f}(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho})\Big], \\
=&\mathbb{E}_k\Big[\left(-\alpha_k + \frac{1}{2}\alpha_k \gamma_k\right)\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho}) + \alpha_k \nabla f(\boldsymbol{x}_k)^\top (\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k) \\
&\quad + \frac{1}{2}\alpha_k \gamma_k \left(\Delta q(\boldsymbol{x}_k, \bar{\boldsymbol{p}}_k, \nabla \bar{f}(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho}) - \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho})\right)\Big],
\end{aligned}
$$

where the last inequality comes from the selection of ξ_k and α_k in lines 10 and 11, respectively. Without the loss of generality, we assume that $\gamma_k \leq 1$ and continue from Equation (B.11),

$$
\begin{aligned}
&\mathbb{E}_k\left[\phi(\boldsymbol{x}_{k+1}, \bar{\rho}) - \phi(\boldsymbol{x}_k, \bar{\rho})\right] \\
\leq&\mathbb{E}_k\left[-\frac{1}{2}\alpha_k^{\min} \Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho})\right] + M_{\nabla f}\alpha_k^{\min}\mathbb{E}_k\left[\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2\right] + \varrho\gamma_k^2 M_{\nabla f}\kappa_1^{-1} M_\sigma \\
&+ \frac{1}{2}\alpha_k^{\max}\gamma_k\left(M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}} + 2\left(M_{\nabla f} + M_\sigma\right) M_{\boldsymbol{\ell},\boldsymbol{u}} + \kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}} M_\sigma\right),
\end{aligned}
$$

where the inequality is due to Lemmas 19 and 20. It further implies that

(B.12)
$$
\begin{aligned}
&\mathbb{E}_k\left[\phi(\boldsymbol{x}_{k+1}, \bar{\rho}) - \min_{\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}} \phi(\boldsymbol{x}, \bar{\rho}) | \mathcal{F}_{k-1}\right] \\
\leq&\phi(\boldsymbol{x}_k, \bar{\rho}) - \min_{\boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}} \phi(\boldsymbol{x}, \bar{\rho}) - \frac{1}{2}\sum_{k=\bar{K}}^{\bar{K}+K} \alpha_k^{\min}\mathbb{E}_k\left[\Delta q(\boldsymbol{x}_k, \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k), \boldsymbol{B}_k, \bar{\rho}) | \mathcal{F}_{k-1}\right] \\
&+ M_{\nabla f}\sum_{k=\bar{K}}^{\bar{K}+K} \alpha_k^{\min}\mathbb{E}_k\left[\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2 | \mathcal{F}_{k-1}\right] + \varrho M_{\nabla f}\kappa_1^{-1} M_\sigma \sum_{k=\bar{K}}^{\bar{K}+K} \gamma_k^2 \\
&+ \frac{1}{2}\left(M_\sigma M_{\boldsymbol{\ell},\boldsymbol{u}} + 2\left(M_{\nabla f} + M_\sigma\right) M_{\boldsymbol{\ell},\boldsymbol{u}} + \kappa_1^{-1}\kappa_2 M_{\boldsymbol{\ell},\boldsymbol{u}} M_\sigma\right) \sum_{k=\bar{K}}^{\bar{K}+K} \alpha_k^{\max}\gamma_k.
\end{aligned}
$$

Since $\sum_{k=0}^{\infty}\gamma_k^2<\infty$, $\alpha_k^{\min}=\mathcal{O}(\gamma_k)$ and $\alpha_k^{\max}=\mathcal{O}\left(\gamma_k+\gamma_k^2\right)$, we have $\sum_{k=0}^{\infty}\alpha_k^{\max}\gamma_k<\infty$. Note that $\mathbb{E}\left[\sum_{k=\bar{K}}^{\infty}\alpha_k^{\min}\mathbb{E}_k\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\,|\mathcal{F}_{k-1}\right]\right]=\sum_{k=\bar{K}}^{\infty}\alpha_k^{\min}\mathbb{E}\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\right]<\infty$ shows that $\sum_{k=\bar{K}}^{\infty}\alpha_k^{\min}\mathbb{E}_k\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\,|\mathcal{F}_{k-1}\right]<\infty$. We conclude from Equation (B.12), the step size $\alpha_k^{\min}=\mathcal{O}(\gamma_k)$, the assumption $\sum_{k=\bar{K}}^{\infty}\alpha_k^{\min}\mathbb{E}_k\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\,|\mathcal{F}_{k-1}\right]<\infty$ and Robbins-Siegmund theorem [59] that Equation (B.9) holds. Moreover, since $\sum_{k=0}^{\infty}\gamma_k=\infty$, together with Equation (B.9), we obtain Equation (B.10). □

B.2. Proof for Theorem 2. Note that although $\{\gamma_k\}$ is the pre-defined sequence in the algorithm, the only difference between α_k and γ_k is a constant. Therefore, $\alpha_k^{\min}=\iota_1(k+1)^{-1}$ implies that $\gamma_k=\iota_3(k+1)^{-1}$ for some $\iota_3>0$. For simplicity, we directly discuss the behavior of the sequence related to $\alpha_k^{\min}$ rather than γ_k. We need the techniques and notations in the proof of Lemma 16, where all conditions are satisfied and $\mathbb{E}\left[\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|_2^2\right]\leq 2\mathbb{E}\left[\|\mathcal{W}_1\|_2^2\right]+2\mathbb{E}\left[\|\mathcal{W}_2\|_2^2\right]$. In details,

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{W}_1\|_2^2\right]=&\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_k)\right)^2\beta_i^2\mathbb{E}\left[\|\boldsymbol{g}_i-\nabla f(\boldsymbol{x}_i)\|_2^2\right]\\ \leq&\sigma_g^2\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\beta_k)\right)^2\beta_i^2=\mathcal{O}(\beta_k),\end{aligned}$$

by utilizing Lemma 15. Similarly, for $\|\mathcal{W}_2\|_2$, we have

$$\|\mathcal{W}_2\|_2\leq M_{\boldsymbol{\ell},\boldsymbol{u}}\sum_{i=1}^{k}\left(\prod_{j=i}^{k}(1-\beta_j)\right)\alpha_{i-1}^{\min}=\mathcal{O}\left(\alpha_k^{\min}/\beta_k\right).$$

Therefore, we conclude that $\mathbb{E}\left[\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|_2\right]\leq\mathcal{O}\left(\beta_k+\alpha_k^{\min}/\beta_k\right)$. Since $\alpha_k^{\min}=\iota_1(k+1)^{-b_1}$, it is not difficult to verify that $\sum_{k=\bar{K}}^{\infty}\alpha_k^{\min}\mathbb{E}\left[\|\bar{\boldsymbol{p}}_k-\boldsymbol{p}_k\|_2\right]<\infty$, if $b_1+\frac{b_2}{2}>1$ and $2b_1-b_2>1$. We equivalently require that $b_1\in(\frac{3}{4},1]$ and $b_2\in(2-2b_1,2b_1-1)$.

B.3. Proof for Theorem 3. We first show that the Problem (3.6) is convex and then the corresponding solution $(\boldsymbol{\lambda}_k^*,\boldsymbol{\mu}_{1,k}^*,\boldsymbol{\mu}_{2,k}^*)$ is well-defined. The stability of quadratic problems in Lemma 23 is an generalization of Lemma 17.

LEMMA 22. *Problem (3.6) is convex, i.e., the Hessian matrix $\nabla^2F(\boldsymbol{\lambda},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2;\boldsymbol{x})$ is positive semi-definite for any $\boldsymbol{x}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.*

PROOF. The direct computation of the Hessian matrix for $F(\boldsymbol{\lambda},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2;\boldsymbol{x})$ is

(B.13)
$$\nabla^2F(\boldsymbol{\lambda},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2;\boldsymbol{x})=\begin{pmatrix}2\nabla\boldsymbol{c}(\boldsymbol{x})^\top\nabla\boldsymbol{c}(\boldsymbol{x}) & -2\nabla\boldsymbol{c}(\boldsymbol{x}) & 2\nabla\boldsymbol{c}(\boldsymbol{x})\\ -2\nabla\boldsymbol{c}(\boldsymbol{x})^\top & 2\boldsymbol{I}+2\mathrm{diag}\left((\boldsymbol{x}-\boldsymbol{\ell})^2\right) & -2\boldsymbol{I}\\ 2\nabla\boldsymbol{c}(\boldsymbol{x})^\top & -2\boldsymbol{I} & 2\boldsymbol{I}+2\mathrm{diag}\left((\boldsymbol{x}-\boldsymbol{u})^2\right)\end{pmatrix}.$$

For any vector $\boldsymbol{w}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3)\in\mathbb{R}^r\times\mathbb{R}^d\times\mathbb{R}^d$,

$$\boldsymbol{w}^\top\nabla^2F(\boldsymbol{\lambda},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2;\boldsymbol{x})\boldsymbol{w}=\|\nabla\boldsymbol{c}(\boldsymbol{x})\boldsymbol{w}_1-\boldsymbol{w}_2+\boldsymbol{w}_3\|_2^2+\|(\boldsymbol{x}-\boldsymbol{\ell})\odot\boldsymbol{w}_2\|_2^2+\|(\boldsymbol{x}-\boldsymbol{u})\odot\boldsymbol{w}_3\|_2^2\geq 0.$$

Therefore, Problem (3.6) is convex. □

LEMMA 23 (Stability of Quadratic Programs, Theorem 2.1 in [21]). *For two constrained strongly convex quadratic problems*

$$\boldsymbol{y}^* \in \min_{\boldsymbol{y}\in\Lambda} \boldsymbol{g}^\top \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{Q}\boldsymbol{y},$$

and

$$\boldsymbol{y}^{**} \in \min_{\boldsymbol{y}\in\Lambda} \boldsymbol{g}'^\top \boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{Q}'\boldsymbol{y},$$

where the feasible region Λ is convex. Suppose that $\max\{\|\boldsymbol{y}^*\|_2, \|\boldsymbol{y}^{**}\|_2\} \le M_y$ for some $M_y > 0$. If $\epsilon = \max\{\|\boldsymbol{g}-\boldsymbol{g}'\|_2, \|\boldsymbol{Q}-\boldsymbol{Q}'\|_2\}$ and both two matrices $\boldsymbol{Q}, \boldsymbol{Q}'$ are positive definite with $\upsilon_1 \boldsymbol{I} \preceq \boldsymbol{Q}, \boldsymbol{Q}' \preceq \upsilon_2 \boldsymbol{I}$, for some $0 < \upsilon_1 \le \upsilon_2$. Then, the following holds

$$\|\boldsymbol{y}^* - \boldsymbol{y}^{**}\|_2 \le \epsilon \upsilon_1^{-1}\left(1 + M_y\right).$$

LEMMA 24. *Under assumptions in Theorem 3, we have*

$$\lim_{k\to\infty} F(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*; \boldsymbol{x}_k) = 0, \text{ almost surely.}$$

PROOF. Denote $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}})$ as Lagrangian multipliers of the relaxed SQP subproblem at $\boldsymbol{x}_k$ with full gradient $\nabla f(\boldsymbol{x}_k)$. It follows that

$$\begin{aligned} F(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*; \boldsymbol{x}_k) &\le F(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}}; \boldsymbol{x}_k) \le \|\boldsymbol{B}_k \boldsymbol{p}_k\|_2^2 + \left\|\boldsymbol{\mu}_{1,k}^{\text{sub}} \odot \boldsymbol{p}_k\right\|_2^2 + \left\|\boldsymbol{\mu}_{2,k}^{\text{sub}} \odot \boldsymbol{p}_k\right\|_2^2 \\ &\le \left(\kappa_2 + 2M_{\text{Lag}}\right)\|\boldsymbol{p}_k\|_2^2, \end{aligned}$$

then

$$\liminf_{k\to\infty} F(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*; \boldsymbol{x}_k) = 0,$$

by $\liminf_{k\to\infty} \|\boldsymbol{p}_k\|_2 = 0$. Suppose that $\limsup_{k\to\infty} F(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*; \boldsymbol{x}_k) > 0$, we can find a sufficiently small number $\varepsilon > 0$ and two infinite sequences $\{m_i\}$ and $\{n_i\}$ with $\bar{K} \le m_i < n_i$, such that

$$F(\boldsymbol{\lambda}_{m_i}^*, \boldsymbol{\mu}_{1,m_i}^*, \boldsymbol{\mu}_{2,m_i}^*; \boldsymbol{x}_{m_i}) > 2\varepsilon, \quad \|\boldsymbol{p}_{n_i}\|_2 \le \sqrt{\frac{\varepsilon}{\kappa_2 + M_{\text{Lag}}}},$$

and

$$\|\boldsymbol{p}_k\|_2 \ge \sqrt{\frac{\varepsilon}{\kappa_2 + M_{\text{Lag}}}}, \text{ for } m_i \le k < n_i.$$

Note that we can always achieve it due to the following derivation

(B.14)

$$\begin{aligned} F(\boldsymbol{\lambda}_k^*, \boldsymbol{\mu}_{1,k}^*, \boldsymbol{\mu}_{2,k}^*; \boldsymbol{x}_k) &= \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}_1 \ge \boldsymbol{0}, \boldsymbol{\mu}_2 \ge \boldsymbol{0}} F(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_k) \\ &\le \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}_1 \ge \boldsymbol{0}, \boldsymbol{\mu}_2 \ge \boldsymbol{0}} \left\{ F(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_k) + \frac{\varepsilon}{6M_{\text{Lag}}^2} \|(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)\|_2^2 \right\} \\ &\le F(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}}; \boldsymbol{x}_k) + \frac{\varepsilon}{6M_{\text{Lag}}^2} \left\|(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}})\right\|_2^2 \\ &\le \|\boldsymbol{B}_k \boldsymbol{p}_k\|_2^2 + \left\|\boldsymbol{\mu}_{1,k}^{\text{sub}} \odot \boldsymbol{p}_k\right\|_2^2 + \left\|\boldsymbol{\mu}_{2,k}^{\text{sub}} \odot \boldsymbol{p}_k\right\|_2^2 + \frac{\varepsilon}{2} \\ &\le \left(\kappa_2 + 2M_{\text{Lag}}\right)\|\boldsymbol{p}_k\|_2^2 + \frac{\varepsilon}{2}. \end{aligned}$$

Here, $F(\boldsymbol{\lambda}^*_{m_i}, \boldsymbol{\mu}^*_{1,m_i}, \boldsymbol{\mu}^*_{2,m_i}; \boldsymbol{x}_{m_i}) > 2\varepsilon$ automatically implies that $\|\boldsymbol{p}_{m_i}\|_2 \geq \sqrt{\frac{3\varepsilon}{2(\kappa_2+M_{\text{Lag}})}}$. Since $\liminf_{k\to\infty} \|\boldsymbol{p}_k\|_2 = 0$, there must exists $n_i > m_i$ such that $\|\boldsymbol{p}_{n_i}\|_2 \leq \sqrt{\frac{\varepsilon}{\kappa_2+M_{\text{Lag}}}}$. Let

$$\widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}) = F(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_k) + \frac{\varepsilon}{6M^2_{\text{Lag}}} \|(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)\|_2^2.$$

It also implies that $\widetilde{F}(\boldsymbol{\lambda}^{**}_{m_i}, \boldsymbol{\mu}^{**}_{1,m_i}, \boldsymbol{\mu}^{**}_{2,m_i}; \boldsymbol{x}_{m_i}) \geq 2\varepsilon$ and $\widetilde{F}(\boldsymbol{\lambda}^{**}_{n_i}, \boldsymbol{\mu}^{**}_{1,n_i}, \boldsymbol{\mu}^{**}_{2,n_i}; \boldsymbol{x}_{n_i}) \leq (\kappa_2 + 2M_{\text{Lag}}) \|\boldsymbol{p}_{n_i}\|_2^2 + \frac{\varepsilon}{2} \leq \frac{3}{2}\varepsilon$, where $(\boldsymbol{\lambda}^{**}_{m_i}, \boldsymbol{\mu}^{**}_{1,m_i}, \boldsymbol{\mu}^{**}_{2,m_i}) \in \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}_1 \geq \mathbf{0}, \boldsymbol{\mu}_2 \geq \mathbf{0}} \widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_{m_i})$ and $(\boldsymbol{\lambda}^{**}_{n_i}, \boldsymbol{\mu}^{**}_{1,n_i}, \boldsymbol{\mu}^{**}_{2,n_i}) \in \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}_1 \geq \mathbf{0}, \boldsymbol{\mu}_2 \geq \mathbf{0}} \widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_{n_i})$. Note that the function $\widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x})$ is strictly positive-definite with

$$\frac{\varepsilon}{6M^2_{\text{Lag}}} \leq \left\| \nabla^2 \widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}) \right\|_2 \leq 2\left(M^2_{\nabla c} + 4M^2_{\nabla c} + 2M^2_{\boldsymbol{\ell},\boldsymbol{u}} + 4\right).$$

For simplicity, we denote $\boldsymbol{w}_k = (\boldsymbol{\lambda}^{**}_k, \boldsymbol{\mu}^{**}_{1,k}, \boldsymbol{\mu}^{**}_{2,k})$, then

$$\frac{\varepsilon}{6M^2_{\text{Lag}}} \|\boldsymbol{w}_k\|_2^2 \leq \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}_1 \geq \mathbf{0}, \boldsymbol{\mu}_2 \geq \mathbf{0}} \widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}) \leq (\kappa_2 + 2M_{\text{Lag}}) \|\boldsymbol{p}_k\|_2^2 + \frac{\varepsilon}{2},$$

and thus

$$\|\boldsymbol{w}_k\|_2 \leq \sqrt{\frac{6M^2_{\text{Lag}}(\kappa_2 + 2M_{\text{Lag}}) M^2_{\boldsymbol{\ell},\boldsymbol{u}}}{\varepsilon}} + 3M_{\text{Lag}}, \tag{B.15}$$

for all $k \in \mathbb{N}$.

We first write $\widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_k)$ into the general quadratic form that

$$\widetilde{F}(\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{x}_k) = \|\nabla f(\boldsymbol{x}_k)\|_2^2 + \boldsymbol{q}_k^\top \boldsymbol{w} + \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}_k \boldsymbol{w},$$

where

$$\boldsymbol{q}_k = \begin{pmatrix} 2\nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \nabla f(\boldsymbol{x}_k) \\ -2\nabla f(\boldsymbol{x}_k) \\ 2\nabla f(\boldsymbol{x}_k) \end{pmatrix}$$

and

$$\boldsymbol{Q}_k = \begin{pmatrix} 2\nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \nabla \boldsymbol{c}(\boldsymbol{x}_k) & -2\nabla \boldsymbol{c}(\boldsymbol{x}_k) & 2\nabla \boldsymbol{c}(\boldsymbol{x}_k) \\ -2\nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top & 2\boldsymbol{I} + 2\text{diag}\left((\boldsymbol{x}_k - \boldsymbol{\ell})^2\right) & -2\boldsymbol{I} \\ 2\nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top & -2\boldsymbol{I} & 2\boldsymbol{I} + 2\text{diag}\left((\boldsymbol{x}_k - \boldsymbol{u})^2\right) \end{pmatrix} + \frac{\varepsilon}{6M^2_{\text{Lag}}}\boldsymbol{I}.$$

The smoothness of the objective $f(\boldsymbol{x})$ and the constraints $\boldsymbol{c}(\boldsymbol{x})$ show that

$$\begin{aligned} \|\boldsymbol{q}_{k+1} - \boldsymbol{q}_k\|_2 &\leq 2\left(\kappa_{\nabla c} M_{\nabla f} + M_{\nabla c}\kappa_{\nabla f} + 2\kappa_{\nabla f}\right) \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2 \\ &\leq 2\left(\kappa_{\nabla c} M_{\nabla f} + M_{\nabla c}\kappa_{\nabla f} + 2\kappa_{\nabla f}\right) M_{\boldsymbol{\ell},\boldsymbol{u}} \alpha_k, \end{aligned} \tag{B.16}$$

and

$$\begin{aligned} \|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_2 &\leq 4\left(M_{\nabla c}\kappa_{\nabla c} + 2\kappa_{\nabla c} + 2M_{\boldsymbol{\ell},\boldsymbol{u}}\right) \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2 \\ &\leq 4\left(M_{\nabla c}\kappa_{\nabla c} + 2\kappa_{\nabla c} + 2M_{\boldsymbol{\ell},\boldsymbol{u}}\right) M_{\boldsymbol{\ell},\boldsymbol{u}} \alpha_k. \end{aligned} \tag{B.17}$$

Then

$$
\begin{aligned}
&\left|\widetilde{F}(\boldsymbol{\lambda}_{k+1}^{**},\boldsymbol{\mu}_{1,k+1}^{**},\boldsymbol{\mu}_{2,k+1}^{**};\boldsymbol{x}_{k+1})-\widetilde{F}(\boldsymbol{\lambda}_{k}^{**},\boldsymbol{\mu}_{1,k}^{**},\boldsymbol{\mu}_{2,k}^{**};\boldsymbol{x}_{k})\right| \\
\leq&\left|\boldsymbol{q}_{k+1}^{\top}\boldsymbol{w}_{k+1}+\frac{1}{2}\boldsymbol{w}_{k+1}^{\top}\boldsymbol{Q}_{k+1}\boldsymbol{w}_{k+1}-\boldsymbol{q}_{k}^{\top}\boldsymbol{w}_{k}-\frac{1}{2}\boldsymbol{w}_{k}^{\top}\boldsymbol{Q}_{k}\boldsymbol{w}_{k}\right|+\left|\|\nabla f(\boldsymbol{x}_{k+1})\|_2^2-\|\nabla f(\boldsymbol{x}_{k})\|_2^2\right| \\
\leq&\left|\boldsymbol{q}_{k+1}^{\top}\boldsymbol{w}_{k+1}+\frac{1}{2}\boldsymbol{w}_{k+1}^{\top}\boldsymbol{Q}_{k+1}\boldsymbol{w}_{k+1}-\boldsymbol{q}_{k}^{\top}\boldsymbol{w}_{k+1}-\frac{1}{2}\boldsymbol{w}_{k+1}^{\top}\boldsymbol{Q}_{k}\boldsymbol{w}_{k+1}\right| \\
&-\left|\boldsymbol{q}_{k}^{\top}\boldsymbol{w}_{k+1}+\frac{1}{2}\boldsymbol{w}_{k+1}^{\top}\boldsymbol{Q}_{k}\boldsymbol{w}_{k+1}-\boldsymbol{q}_{k}^{\top}\boldsymbol{w}_{k}-\frac{1}{2}\boldsymbol{w}_{k}^{\top}\boldsymbol{Q}_{k}\boldsymbol{w}_{k}\right|+\left|\|\nabla f(\boldsymbol{x}_{k+1})\|_2^2-\|\nabla f(\boldsymbol{x}_{k})\|_2^2\right| \\
\leq&\|\boldsymbol{w}_{k+1}\|_2\|\boldsymbol{q}_{k+1}-\boldsymbol{q}_{k}\|_2+\frac{1}{2}\|\boldsymbol{w}_{k+1}\|_2^2\|\boldsymbol{Q}_{k+1}-\boldsymbol{Q}_{k}\|_2 \\
&+\|\boldsymbol{q}_{k}\|_2\|\boldsymbol{w}_{k+1}-\boldsymbol{w}_{k}\|_2+\frac{1}{2}\|\boldsymbol{w}_{k+1}\|_2\|\boldsymbol{Q}_{k}\|_2\|\boldsymbol{w}_{k+1}-\boldsymbol{w}_{k}\|_2 \\
&+\frac{1}{2}\|\boldsymbol{w}_{k}\|_2\|\boldsymbol{Q}_{k}\|_2\|\boldsymbol{w}_{k+1}-\boldsymbol{w}_{k}\|_2+\|\nabla f(\boldsymbol{x}_{k+1})-\nabla f(\boldsymbol{x}_{k})\|_2\left(\|\nabla f(\boldsymbol{x}_{k+1})\|_2+\|\nabla f(\boldsymbol{x}_{k})\|_2\right).
\end{aligned}
$$

Using Lemma 23, Equations (B.15), (B.16) and (B.17), and $\boldsymbol{Q}_k \succeq \frac{\varepsilon}{6M_{\text{Lag}}^2}\boldsymbol{I}$, we have $\|\boldsymbol{w}_{k+1}-\boldsymbol{w}_k\|_2=\mathcal{O}\left(\frac{\alpha_k}{\varepsilon^{3/2}}\right)$, where we omit some universal and uncritical constants. Combining it with Equations (B.15), (B.16) and (B.17), we have

$$
\left|\widetilde{F}(\boldsymbol{\lambda}_{k+1}^{**},\boldsymbol{\mu}_{1,k+1}^{**},\boldsymbol{\mu}_{2,k+1}^{**};\boldsymbol{x}_{k+1})-\widetilde{F}(\boldsymbol{\lambda}_{k}^{**},\boldsymbol{\mu}_{1,k}^{**},\boldsymbol{\mu}_{2,k}^{**};\boldsymbol{x}_{k})\right|\leq M_F\frac{\alpha_k}{\varepsilon^2},
$$

for a universal constant $M_F>0$, where the constant is independent of α_k, k and ε.

Therefore, it follows from the above inequalities and our construction of the sequences $\{m_i\}$ and $\{n_i\}$ that

$$
\begin{aligned}
\frac{1}{2}\varepsilon&\leq\widetilde{F}(\boldsymbol{\lambda}_{m_i}^{*},\boldsymbol{\mu}_{1,m_i}^{*},\boldsymbol{\mu}_{2,m_i}^{*};\boldsymbol{x}_{m_i})-\widetilde{F}(\boldsymbol{\lambda}_{n_i}^{*},\boldsymbol{\mu}_{1,n_i}^{*},\boldsymbol{\mu}_{2,n_i}^{*};\boldsymbol{x}_{n_i}) \\
&\leq\sum_{k=m_i}^{n_i-1}\left|\widetilde{F}(\boldsymbol{\lambda}_{k}^{*},\boldsymbol{\mu}_{1,k}^{*},\boldsymbol{\mu}_{2,k}^{*};\boldsymbol{x}_{k})-\widetilde{F}(\boldsymbol{\lambda}_{k+1}^{*},\boldsymbol{\mu}_{1,k}^{*},\boldsymbol{\mu}_{2,k}^{*};\boldsymbol{x}_{k})\right| \\
&\leq\sum_{k=m_i}^{n_i-1}M_F\frac{\alpha_k}{\varepsilon^2}.
\end{aligned}
\tag{B.18}
$$

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

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

However, $\|\boldsymbol{p}_k\|_2\geq\sqrt{\frac{\varepsilon}{\kappa_2+M_{\text{Lag}}}}$ for $m_i\leq k\leq n_i-1$, which further implies that

$$
\sum_{i=1}^{\infty}\sum_{k=m_i}^{n_i-1}\alpha_k\leq\frac{\kappa_2+M_{\text{Lag}}}{\varepsilon}\sum_{i=1}^{\infty}\sum_{k=m_i}^{n_i-1}\alpha_k\|\boldsymbol{p}_k\|_2^2\leq\frac{\kappa_2+M_{\text{Lag}}}{\varepsilon}\sum_{k=\bar{K}}^{\infty}\alpha_k\|\boldsymbol{p}_k\|_2^2<\infty.
$$

It is a contradiction. Therefore, we complete the proof that $\lim_{k\to\infty}F(\boldsymbol{\lambda}_k^*,\boldsymbol{\mu}_{1,k}^*,\boldsymbol{\mu}_{2,k}^*;\boldsymbol{x}_k)=0$. □

LEMMA 25. *Under assumptions in Theorem 3, we have*

$$
\lim_{k\to\infty}\boldsymbol{c}(\boldsymbol{x}_k)=\boldsymbol{0},\ \textit{almost surely}.
$$

PROOF. The proof scheme is similar. For completeness, we provide the details here. Suppose that $\limsup_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 = 0$ but $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 = 0$. Then we can find a sufficiently small number $\varepsilon > 0$ and two infinite sequences $\{m_i\}$ and $\{n_i\}$ with $\bar{K} \le m_i < n_i$, such that

$$\|\boldsymbol{c}(\boldsymbol{x}_{m_i})\|_2 > 2\varepsilon, \quad \|\boldsymbol{c}(\boldsymbol{x}_{n_i})\|_2 < \varepsilon,$$

and

$$\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 \ge \varepsilon, \text{ for } m_i \le k < n_i.$$

It follows from the definition of the sequence that

$$\begin{aligned}
\varepsilon &\le \|\boldsymbol{c}(\boldsymbol{x}_{m_i})\|_2 - \|\boldsymbol{c}(\boldsymbol{x}_{n_i})\|_2 \\
&\le \sum_{k=m_i}^{n_i-1} \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 - \|\boldsymbol{c}(\boldsymbol{x}_{k+1})\|_2 \\
&\le \sum_{k=m_i}^{n_i-1} \|\boldsymbol{c}(\boldsymbol{x}_k) - \boldsymbol{c}(\boldsymbol{x}_{k+1})\|_2 \\
&\le \kappa_c M_{\boldsymbol{\ell},\boldsymbol{u}} \sum_{k=m_i}^{n_i-1} \alpha_k, \quad \text{for all } i \in \mathbb{N}.
\end{aligned}$$

Multiplying both two sides by ε and by the fact that $\|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 \ge \varepsilon, \text{ for } m_i \le k < n_i$, we have

$$\varepsilon^2 \le \kappa_c M_{\boldsymbol{\ell},\boldsymbol{u}} \sum_{k=m_i}^{n_i-1} \alpha_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2, \quad \text{for all } i \in \mathbb{N},$$

which implies that $\infty \le \sum_{i=1}^{\infty} \sum_{k=m_i}^{n_i-1} \alpha_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 \le \sum_{k=\bar{K}}^{\infty} \alpha_k \|\boldsymbol{c}(\boldsymbol{x}_k)\|_2 < \infty$. It is a contradiction. □

Combining with Lemmas 24 and 25, we finish the proof for Theorem 3.

APPENDIX C: PROOF FOR THEOREM 4

C.1. Proof for Lemma 3. We proceed by prove each of the four conclusions in turn.

C.1.1. *Proof for Conclusion 1.* Note that

$$\boldsymbol{p}_k \in \arg\min_{\boldsymbol{p}\in\Omega_k} \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p},$$

where $\Omega_k = \{\boldsymbol{p} : \boldsymbol{c}(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \le \boldsymbol{x}_k + \boldsymbol{p} \le \boldsymbol{u}\}$, and let

$$\boldsymbol{p}^* \in \arg\min_{\boldsymbol{p}\in\Omega^*} \nabla f(\boldsymbol{x}^*)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p},$$

and

$$\boldsymbol{p}_k^* \in \arg\min_{\boldsymbol{p}\in\Omega^*} \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{B}_k \boldsymbol{p},$$

where $\Omega^* = \{\boldsymbol{p} : \boldsymbol{c}(\boldsymbol{x}^*) + \nabla \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{p} = \boldsymbol{0}\} \cap \{\boldsymbol{p} : \boldsymbol{\ell} \le \boldsymbol{x}^* + \boldsymbol{p} \le \boldsymbol{u}\}$. Lemma 23 shows that $\|\boldsymbol{p}_k^* - \boldsymbol{p}^*\|_2 \le C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$ for some $C > 0$, since both $\boldsymbol{p}_k^*$ and $\boldsymbol{p}^*$ are bounded by $M_{\boldsymbol{\ell},\boldsymbol{u}}$. In the next part, for simplicity, we use the same notation C to denote some universal constants. We slightly rewrite the formulation for $\boldsymbol{p}_k$ and $\boldsymbol{p}_k^*$ that

$$\hat{\boldsymbol{p}}_k \in \arg\min_{\boldsymbol{p}\in\widehat{\Omega}_k} \frac{1}{2}\|\boldsymbol{p}\|_{\boldsymbol{B}_k}^2,$$

and

$$\hat{\boldsymbol{p}}_k^* \in \arg\min_{\boldsymbol{p}\in\widehat{\Omega}^*} \frac{1}{2}\|\boldsymbol{p}\|_{\boldsymbol{B}_k}^2,$$

where $\widehat{\Omega}_k = \left\{\boldsymbol{p} + \boldsymbol{B}_k^{-1}\nabla f(\boldsymbol{x}_k) : \boldsymbol{p} \in \Omega_k\right\}$ and $\widehat{\Omega}^* = \left\{\boldsymbol{p} + \boldsymbol{B}_k^{-1}\nabla f(\boldsymbol{x}_k) : \boldsymbol{p} \in \Omega^*\right\}$. Then $\|\boldsymbol{p}_k - \boldsymbol{p}_k^*\|_2 = \|\hat{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}_k^*\|_2$. By Proposition 3.1 in [21] and results in [37], there exists $\hat{\boldsymbol{p}}^{*\prime} \in \widehat{\Omega}^*$ and $\hat{\boldsymbol{p}}_k' \in \widehat{\Omega}_k$, such that $\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} \le C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$ and $\|\hat{\boldsymbol{p}}_k' - \hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k} \le C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$ for some $C > 0$, then

$$\|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k} \le \left\|\hat{\boldsymbol{p}}^{*\prime}\right\|_{\boldsymbol{B}_k} \le \|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} + C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2,$$

and

$$\|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} \le \left\|\hat{\boldsymbol{p}}_k'\right\|_{\boldsymbol{B}_k} \le \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k} + C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2.$$

Equipped with the above inequalities and the optimality condition that $\langle \hat{\boldsymbol{p}}_k^*, \hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^* \rangle \ge 0$, we have

$$\begin{aligned}
\left\|\hat{\boldsymbol{p}}^{*\prime}\right\|_{\boldsymbol{B}_k}^2 &= \left\|\hat{\boldsymbol{p}}_k^* + \hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2 \\
&= \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 + 2\left\langle \hat{\boldsymbol{p}}_k^*, \hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^* \right\rangle + \left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2 \\
&\ge \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 + \left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2 &\le \left\|\hat{\boldsymbol{p}}^{*\prime}\right\|_{\boldsymbol{B}_k}^2 - \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 \\
&= \left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k + \hat{\boldsymbol{p}}_k\right\|_{\boldsymbol{B}_k}^2 - \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 \\
&\le \left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k\right\|_{\boldsymbol{B}_k}^2 + 2\left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k\right\|_{\boldsymbol{B}_k}\|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} + \|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k}^2 - \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 \\
&\le \left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k\right\|_{\boldsymbol{B}_k}^2 + 2\left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k\right\|_{\boldsymbol{B}_k}\|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} + \left|\|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} - \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}\right|\left(\|\hat{\boldsymbol{p}}_k\|_{\boldsymbol{B}_k} + \|\hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}\right) \\
&\le C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2,
\end{aligned}$$

for some $C > 0$. Thus

$$\begin{aligned}\|\hat{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}_k^*\|_{\boldsymbol{B}_k}^2 &\leq \left\|\hat{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}^{*\prime} + \hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2 \\ &\leq 2\left\|\hat{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}^{*\prime}\right\|_{\boldsymbol{B}_k}^2 + 2\left\|\hat{\boldsymbol{p}}^{*\prime} - \hat{\boldsymbol{p}}_k^*\right\|_{\boldsymbol{B}_k}^2 \\ &\leq C\left\|\boldsymbol{x}_k - \boldsymbol{x}^*\right\|_2, \end{aligned} \tag{C.1}$$

for some $C > 0$. The facts that $\|\boldsymbol{p}_k - \boldsymbol{p}_k^*\|_2 = \|\hat{\boldsymbol{p}}_k - \hat{\boldsymbol{p}}_k^*\|_2 \leq C\sqrt{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2}$ and $\|\boldsymbol{p}_k^* - \boldsymbol{p}^*\|_2 \leq C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$ show that $\|\boldsymbol{p}_k - \boldsymbol{p}^*\|_2 \leq C\sqrt{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2}$ for some $C > 0$. Note that $\boldsymbol{p}^* = \boldsymbol{0}$ for any positive-definite matrix $\boldsymbol{B}_k$ since $\boldsymbol{x}^*$ is a local solution of Problem (1.1). We complete the proof as $\boldsymbol{x}_k \to \boldsymbol{x}^*$ in Assumption 5.

C.1.2. *Proof for Conclusion 2.* We revisit the definition of $\bar{\boldsymbol{g}}_k$ and have that

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

Here,

$$\|\mathcal{W}_{2,k}\|_2 \leq \sum_{i=1}^{k}\left(\prod_{j=i}^{k}(1-\beta_j)\right)\alpha_{i-1}M_{\boldsymbol{\ell},\boldsymbol{u}},$$

then $\mathcal{W}_{2,k} \to 0$ as $k \to \infty$ since $\lim_{k\to\infty}\alpha_{i-1}/\beta_i = 0$. We apply Theorem 4.4 in [34] with $\gamma = 2$, $\phi(k) = 1$, $\alpha = -1/2$, and $X_{k,h} = \left(\prod_{h'=h+1}^{k}(1-\beta_{h'})\right)\beta_h/(\sqrt{\beta_k}k^{1/2+\epsilon})\left(\boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h)\right)$ for any sufficiently small $\varepsilon > 0$, as well as the Borel-Cantelli Lemma that the martingale difference array satisfies $\|\mathcal{W}_{1,k}\|_2 \to 0$, almost surely.

C.1.3. *Proof for Condition 3.* We will show that there exists a sufficiently small $\varepsilon^* > 0$ such that if $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon^*$, $\mathcal{I}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{I}(\boldsymbol{x}_k + \boldsymbol{p}_k)$ and $\mathcal{J}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{J}(\boldsymbol{x}_k + \boldsymbol{p}_k)$ hold. The almost sure convergence of $\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)$ implies that $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon^*$ holds for some sufficiently large $k \geq K^*$. Let $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}})$ and $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}, \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})$ be the Lagrangian multipliers of the relaxed SQP subproblem with the full gradient $\nabla f(\boldsymbol{x}_k)$ and the stochastic averaged gradient $\bar{\boldsymbol{g}}_k$, respectively. Let $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$ be the Lagrangian multiplier for Problem (1.1) at $\boldsymbol{x}^*$ and denote $\epsilon = \min\{\{(\boldsymbol{\mu}_1^*)_i : i \in \mathcal{I}(\boldsymbol{x}^*)\} \cup \{(\boldsymbol{\mu}_2^*)_i : i \in \mathcal{J}(\boldsymbol{x}^*)\}\} > 0$ due to the strictly complementary slackness condition. Since $\boldsymbol{p}_k$ is the optimal solution of the strongly convex quadratic SQP subproblem, the KKT condition shows that $\nabla f(\boldsymbol{x}_k) + \boldsymbol{B}_k\boldsymbol{p}_k + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}} + \boldsymbol{\mu}_{2,k}^{\text{sub}} = \boldsymbol{0}$. Taking $k \to \infty$ ($\boldsymbol{x}_k \to \boldsymbol{x}^*$ and $\boldsymbol{p}_k \to$

$\mathbf{0}$), it follows from the LICQ at $\boldsymbol{x}^*$ that $(\boldsymbol{\lambda}_k^{\text{sub}}, \boldsymbol{\mu}_{1,k}^{\text{sub}}, \boldsymbol{\mu}_{2,k}^{\text{sub}}) \to (\boldsymbol{\lambda}^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$. So there exists sufficiently large $K^* > 0$ such that $(\boldsymbol{\mu}_{1,k}^{\text{sub}})_i > \frac{3}{4}\epsilon$ for all $i \in \mathcal{I}(\boldsymbol{x}^*)$ and $(\boldsymbol{\mu}_{2,k}^{\text{sub}})_i > \frac{3}{4}\epsilon$ for all $i \in \mathcal{J}(\boldsymbol{x}^*)$. Therefore, $\boldsymbol{x}_k + \boldsymbol{p}_k$ has the same active and inactive set as $\boldsymbol{x}^*$, i.e., $\mathcal{I}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{I}(\boldsymbol{x}^*)$ and $\mathcal{J}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{J}(\boldsymbol{x}^*)$. Denote $\epsilon' = \max\{(\boldsymbol{x}^* - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}^*)_i : i \notin \mathcal{I}(\boldsymbol{x}^*) \text{ and } i \notin \mathcal{J}(\boldsymbol{x}^*)\}$. When K^* is sufficiently large, we have $\max\{(\boldsymbol{x}_k + \boldsymbol{p}_k - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}_k - \boldsymbol{p}_k)_i : i \notin \mathcal{I}(\boldsymbol{x}^*) \text{ and } i \notin \mathcal{J}(\boldsymbol{x}^*)\} \geq \frac{3}{4}\epsilon'$. Lemma 23 shows that $\|\bar{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2 \leq (1 + M_{\boldsymbol{\ell},\boldsymbol{u}})\kappa_1^{-1}\varepsilon^*$ under the assumption that $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon^*$ when $k \geq K^*$. If ε^* is sufficiently small such that $(1 + M_{\boldsymbol{\ell},\boldsymbol{u}})\kappa_1^{-1}\varepsilon^* \leq \frac{1}{4}\epsilon'$, then $\max\{(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}_k - \bar{\boldsymbol{p}}_k)_i : i \notin \mathcal{I}(\boldsymbol{x}^*) \text{ and } i \notin \mathcal{J}(\boldsymbol{x}^*)\} \geq \frac{1}{2}\epsilon'$.

The LICQ condition implies that columns of $[\nabla \boldsymbol{c}(\boldsymbol{x}^*), [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}]$ are linearly independent and $[\nabla \boldsymbol{c}(\boldsymbol{x}^*), [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}]^\top [\nabla \boldsymbol{c}(\boldsymbol{x}^*), [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}] \succeq \kappa_0 \boldsymbol{I}$ for some $\kappa_0 > 0$. By the smoothness of $\boldsymbol{c}(\boldsymbol{x})$, there exists sufficiently large K^* such that $[\nabla \boldsymbol{c}(\boldsymbol{x}_k), [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}]^\top [\nabla \boldsymbol{c}(\boldsymbol{x}_k), [-\boldsymbol{I}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{I}]_{\mathcal{J}(\boldsymbol{x}^*)}] \succeq \frac{1}{2}\kappa_0 \boldsymbol{I}$ for all $k \geq K^*$. The KKT condition of the SQP subproblem at $\boldsymbol{x}_k$ with $\bar{\boldsymbol{g}}_k$ shows that $\bar{\boldsymbol{g}}_k + \boldsymbol{B}_k\bar{\boldsymbol{p}}_k + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} + \bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} = \mathbf{0}$. Since $\max\{(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k - \boldsymbol{\ell})_i, (\boldsymbol{u} - \boldsymbol{x}_k - \bar{\boldsymbol{p}}_k)_i : i \notin \mathcal{I}(\boldsymbol{x}^*) \text{ and } i \notin \mathcal{J}(\boldsymbol{x}^*)\} \geq \frac{1}{2}\epsilon'$, $(\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}})_i = (\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})_i = 0$ for $i \notin \mathcal{I}(\boldsymbol{x}^*)$ and $i \notin \mathcal{J}(\boldsymbol{x}^*)$. Therefore,

$$
\begin{aligned}
&\left\|\nabla \boldsymbol{c}(\boldsymbol{x}_k)\left(\bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}_k^{\text{sub}}\right) - \left(\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}}\right) + \left(\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} - \boldsymbol{\mu}_{2,k}^{\text{sub}}\right)\right\|_2 \\
\leq &\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 + \|\boldsymbol{B}_k\bar{\boldsymbol{p}}_k - \boldsymbol{B}_k\boldsymbol{p}_k\|_2 \\
\leq &\left(1 + (1 + M_{\boldsymbol{\ell},\boldsymbol{u}})\kappa_1^{-1}\kappa_2\right)\varepsilon^*,
\end{aligned}
$$

and

$$
\left\|\begin{pmatrix} \bar{\boldsymbol{\lambda}}_k^{\text{sub}} - \boldsymbol{\lambda}_k^{\text{sub}} \\ \left[\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}} - \boldsymbol{\mu}_{1,k}^{\text{sub}}\right]_{\mathcal{I}(\boldsymbol{x}^*)} \\ \left[\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}} - \boldsymbol{\mu}_{2,k}^{\text{sub}}\right]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2 \leq 2\kappa_0^{-1}\left(M_{\nabla c} + 2\right)\left(1 + (1 + M_{\boldsymbol{\ell},\boldsymbol{u}})\kappa_1^{-1}\kappa_2\right)\varepsilon^*.
$$

We let ε^* to be small enough such that the right-hand side of the above inequality is less than $\frac{1}{4}\epsilon$, i.e., $2\kappa_0^{-1}\left(M_{\nabla c} + 2\right)\left(1 + (1 + M_{\boldsymbol{\ell},\boldsymbol{u}})\kappa_1^{-1}\kappa_2\right)\varepsilon^* \leq \frac{1}{4}\epsilon$. Then, together with $(\boldsymbol{\mu}_{1,k}^{\text{sub}})_i > \frac{3}{4}\epsilon$ for $i \in \mathcal{I}(\boldsymbol{x}^*)$ and $(\boldsymbol{\mu}_{2,k}^{\text{sub}})_i > \frac{3}{4}\epsilon$ for $i \in \mathcal{J}(\boldsymbol{x}^*)$, we have $(\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}})_i > \frac{1}{2}\epsilon$ for $i \in \mathcal{I}(\boldsymbol{x}^*)$ and $(\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}})_i > \frac{1}{2}\epsilon$ for $i \in \mathcal{J}(\boldsymbol{x}^*)$. It implies that both $\boldsymbol{x}_k + \boldsymbol{p}_k$ and $\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k$ can correctly identify the active and inactive sets of constraints at $\boldsymbol{x}^*$. Therefore, $\mathcal{I}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{I}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{I}(\boldsymbol{x}^*)$ and $\mathcal{J}(\boldsymbol{x}_k + \bar{\boldsymbol{p}}_k) = \mathcal{J}(\boldsymbol{x}_k + \boldsymbol{p}_k) = \mathcal{J}(\boldsymbol{x}^*)$.

C.1.4. *Proof for Conclusion 4.* Equipped with the fact that $\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k) \to \mathbf{0}$ almost surely, the condition $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon^*$ always holds when k is sufficiently large. By the proof in the previous section, we know that $(\bar{\boldsymbol{\lambda}}_k^{\text{sub}}, [\bar{\boldsymbol{\mu}}_{1,k}^{\text{sub}}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\bar{\boldsymbol{\mu}}_{2,k}^{\text{sub}}]_{\mathcal{J}(\boldsymbol{x}^*)}) \to (\boldsymbol{\lambda}^*, [\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)})$. The update scheme for dual variables in Step 6 shows the following recursion

$$
\boldsymbol{\lambda}_{k+1} = \prod_{j=K^*}^{k}(1 - \alpha_j)\,\boldsymbol{\lambda}_{K^*} + \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}(1 - \alpha_j)\,\alpha_i\bar{\boldsymbol{\lambda}}_i^{\text{sub}},
$$

then $\boldsymbol{\lambda}_k \to \boldsymbol{\lambda}^*$ almost surely. Similar convergence results hold for $([\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)}) \to ([\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}, [\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)})$ and for dual variables indexed on inactive sets $\mathcal{I}^-(\boldsymbol{x}^*)$ and $\mathcal{J}^-(\boldsymbol{x}^*)$.

C.2. Proof for Lemma 4. The definition of the averaged Hessian matrix $\boldsymbol{B}_k$ shows that

(C.2)
$$\begin{aligned}\|\boldsymbol{B}_k - \boldsymbol{B}^*\|_2 \leq & \left\| \frac{1}{k}\sum_{i=1}^{k} \nabla^2 f(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f(\boldsymbol{x}_i) \right\|_2 \\ & + \frac{1}{k}\sum_{i=1}^{k} \left\| \nabla^2 f(\boldsymbol{x}_i) + \sum_{j=1}^{r} (\boldsymbol{\lambda}_i)_j \nabla^2 c_j(\boldsymbol{x}_i) - \nabla^2 f(\boldsymbol{x}^*) - \sum_{j=1}^{r} (\boldsymbol{\lambda}^*)_j \nabla^2 c_j(\boldsymbol{x}^*) \right\|_2 \\ & + \|\boldsymbol{\Delta}_k\|_2 \\ \leq & \left\| \frac{1}{k}\sum_{i=1}^{k} \nabla^2 f(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f(\boldsymbol{x}_i) \right\|_2 + \frac{\Upsilon_{\nabla^2\mathcal{L}}}{k} \sum_{i=1}^{k} \left\| \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^* \end{pmatrix} \right\|_2 + \|\boldsymbol{\Delta}_k\|_2 ,\end{aligned}$$

for some $\Upsilon_{\nabla^2\mathcal{L}} > 0$ due to the compactness of iterates and smoothness of $\nabla^2 f(\boldsymbol{x})$ and $\nabla^2 c(\boldsymbol{x})$. The first term converges to 0 almost surely by the strong law of large number, while the second term converges to 0 almost surely by the Stolz-Cesaro theorem. Since $\boldsymbol{\Delta}_k$ acts as a regularization term for the positive definiteness of $\boldsymbol{B}_k$ and $\boldsymbol{B}^*$ is positive definite, we deduce that $\boldsymbol{\Delta}_k = \boldsymbol{0}$ when k is sufficiently large. Moreover,

$$\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2 \leq \|\boldsymbol{B}_k - \boldsymbol{B}^*\|_2 + \kappa_{\nabla c} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$$

implies that $\boldsymbol{H}_k \to \boldsymbol{H}^*$ almost surely.

C.3. Proof for Theorem 4.

LEMMA 26. *When $\boldsymbol{H}_k$ is sufficiently close to $\boldsymbol{H}^*$, there exists a constant $\Upsilon_L > 0$, such that*

$$\left\| \boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1} \right\|_2 \leq \Upsilon_L \|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2 . \tag{C.3}$$

Then

$$\left\| \boldsymbol{H}_k^{-1} \right\|_2 , \left\| (\boldsymbol{H}^*)^{-1} \right\|_2 \leq \Upsilon_H ,$$

for some $\Upsilon_H > 0$. Moreover, $\boldsymbol{H}_k^{-1} \to (\boldsymbol{H}^*)^{-1}$ almost surely.

PROOF. First, we build the relationship between $\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1}$ and $\boldsymbol{H}_k - \boldsymbol{H}^*$. Note that

$$\begin{aligned} \boldsymbol{0} &= (\boldsymbol{H}^*)^{-1} \boldsymbol{H}^* - \boldsymbol{H}_k^{-1} \boldsymbol{H}_k \\ &= (\boldsymbol{H}^*)^{-1} \boldsymbol{H}^* - (\boldsymbol{H}^*)^{-1} \boldsymbol{H}_k + (\boldsymbol{H}^*)^{-1} \boldsymbol{H}_k - \boldsymbol{H}_k^{-1} \boldsymbol{H}_k \\ &= (\boldsymbol{H}^*)^{-1} (\boldsymbol{H}^* - \boldsymbol{H}_k) + \left((\boldsymbol{H}^*)^{-1} - \boldsymbol{H}_k^{-1} \right) \boldsymbol{H}_k , \end{aligned} \tag{C.4}$$

then

$$\left\| \boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1} \right\|_2 \leq \frac{\left\| (\boldsymbol{H}^*)^{-1} \right\|_2 \|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2}{\lambda_{\min}(\boldsymbol{H}_k)} \leq \Upsilon_L \|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2 , \tag{C.5}$$

for some $\Upsilon_L > 0$, since we can assume that $\lambda_{\min}(\boldsymbol{H}_k) > \frac{1}{2}\lambda_{\min}(\boldsymbol{H}^*)$ without the loss of generality, when $\boldsymbol{H}_k$ is sufficiently close to $\boldsymbol{H}^*$. The boundedness of $\left\| \boldsymbol{H}_k^{-1} \right\|_2$ is a direct result of Equation (C.3). The almost sure convergence of $\boldsymbol{H}_k^{-1} \to (\boldsymbol{H}^*)^{-1}$ is straightforward from Equation (C.3) and Lemma 4. □

LEMMA 27. *Algorithm 3 generates a sequence* $\{(\boldsymbol{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k})\}$ *satisfying*

$$\begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} = \mathcal{Q}_{1,k} + \mathcal{Q}_{2,k} + \mathcal{Q}_{3,k},$$

and

$$\begin{pmatrix} [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}^-(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}^-(\boldsymbol{x}^*)} \end{pmatrix} = \prod_{i=K^*}^{k} (1 - \alpha_i^{\min}) \begin{pmatrix} [\boldsymbol{\mu}_{1,K^*} - \boldsymbol{\mu}_1^*]_{\mathcal{I}^-(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,K^*} - \boldsymbol{\mu}_2^*]_{\mathcal{J}^-(\boldsymbol{x}^*)} \end{pmatrix},$$

where

$$\mathcal{Q}_{1,k} = \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1 - \alpha_j^{\min}) \alpha_i^{\min} \boldsymbol{\phi}_i,$$

$$\mathcal{Q}_{2,k} = \sum_{i=K^*}^{k} \left(\prod_{j=i+1}^{k} (1 - \alpha_j^{\min}) \right) (\alpha_i - \alpha_i^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta \boldsymbol{\lambda}_k \\ [\Delta \boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta \boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix},$$

$$\mathcal{Q}_{3,k} = \prod_{i=K^*}^{k} (1 - \alpha_i^{\min}) \begin{pmatrix} \boldsymbol{x}_{K^*} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{K^*} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,K^*} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,K^*} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1 - \alpha_j^{\min}) \alpha_i^{\min} \boldsymbol{\delta}_i,$$

and

$$\boldsymbol{\phi}_i = -\boldsymbol{H}_i^{-1} \begin{pmatrix} \bar{\boldsymbol{g}}_i - \nabla f(\boldsymbol{x}_i) \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix},$$

$$\boldsymbol{\delta}_i = -(\boldsymbol{H}^*)^{-1} \boldsymbol{\psi}_i - \left(\boldsymbol{H}_i^{-1} - (\boldsymbol{H}^*)^{-1} \right) \begin{pmatrix} \nabla f(\boldsymbol{x}_i) + \nabla \boldsymbol{c}(\boldsymbol{x}_i) \boldsymbol{\lambda}_i - \boldsymbol{\mu}_{1,i} + \boldsymbol{\mu}_{2,i} \\ \boldsymbol{c}(\boldsymbol{x}_i) \\ [\boldsymbol{\ell} - \boldsymbol{x}_i]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_i - \boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix},$$

$$\boldsymbol{\psi}_i = \begin{pmatrix} \nabla f(\boldsymbol{x}_i) + \nabla \boldsymbol{c}(\boldsymbol{x}_i) \boldsymbol{\lambda}_i - \boldsymbol{\mu}_{1,i} + \boldsymbol{\mu}_{2,i} \\ \boldsymbol{c}(\boldsymbol{x}_i) \\ [\boldsymbol{\ell} - \boldsymbol{x}_i]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_i - \boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} - \boldsymbol{H}^* \begin{pmatrix} \boldsymbol{x}_i - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,i} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}.$$

PROOF. By the update scheme of Algorithm 3, we have

$$\begin{pmatrix} \boldsymbol{x}_{k+1}-\boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1}-\boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1}-\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1}-\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{k}-\boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k}-\boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k}-\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k}-\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + (\alpha_k - \alpha_k^{\min} + \alpha_k^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ [\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{x}_{k}-\boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k}-\boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k}-\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k}-\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + (\alpha_k - \alpha_k^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ [\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} - \alpha_k^{\min} \boldsymbol{H}_k^{-1} \begin{pmatrix} \nabla f(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{1,k} + \boldsymbol{\mu}_{2,k} \\ \boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{\ell}-\boldsymbol{x}_k]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_k-\boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}$$

$$+ \alpha_k^{\min}\boldsymbol{\phi}_k$$

$$= \begin{pmatrix} \boldsymbol{x}_{k}-\boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k}-\boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k}-\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k}-\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + (\alpha_k - \alpha_k^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ [\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + \alpha_k^{\min}\boldsymbol{\phi}_k$$

$$- \alpha_k^{\min} \left(\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1} \right) \begin{pmatrix} \nabla f(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{1,k} + \boldsymbol{\mu}_{2,k} \\ \boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{\ell}-\boldsymbol{x}_k]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_k-\boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}$$

$$- \alpha_k^{\min} (\boldsymbol{H}^*)^{-1} \begin{pmatrix} \nabla f(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{1,k} + \boldsymbol{\mu}_{2,k} \\ \boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{\ell}-\boldsymbol{x}_k]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_k-\boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}$$

$$= (1-\alpha_k^{\min}) \begin{pmatrix} \boldsymbol{x}_{k}-\boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k}-\boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k}-\boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k}-\boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + (\alpha_k - \alpha_k^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ [\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + \alpha_k^{\min}\boldsymbol{\phi}_k$$

$$- \alpha_k^{\min} \left(\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1} \right) \begin{pmatrix} \nabla f(\boldsymbol{x}_k) + \nabla \boldsymbol{c}(\boldsymbol{x}_k)\boldsymbol{\lambda}_k - \boldsymbol{\mu}_{1,k} + \boldsymbol{\mu}_{2,k} \\ \boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{\ell}-\boldsymbol{x}_k]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{x}_k-\boldsymbol{u}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}$$

$$- \alpha_k^{\min} (\boldsymbol{H}^*)^{-1} \boldsymbol{\psi}_k.$$

We then obtain the result by applying the above equation recursively. □

LEMMA 28. *Under Assumptions 5 and 6, then*

$$\|\mathcal{Q}_{2,k}\|_2 = \mathcal{O}\left(\alpha_k^{\min}\right).$$

PROOF. Under the boundedness of the generated iterates and the almost sure convergence that $\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon^*$, we have that the iterates in Equation (4.3) are bounded for all $k \geq K^*$, i.e.,

$$\left\| \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta\boldsymbol{\lambda}_k \\ [\Delta\boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta\boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} \right\|_2 \leq M_\Delta,$$

for some $M_\Delta > 0$, due to the LICQ condition. Recall that

$$\mathcal{Q}_{2,k} = \sum_{i=K^*}^{k} \left(\prod_{j=i+1}^{k} (1-\alpha_j^{\min}) \right) (\alpha_i - \alpha_i^{\min}) \begin{pmatrix} \bar{\boldsymbol{p}}_k \\ \Delta \boldsymbol{\lambda}_k \\ [\Delta \boldsymbol{\mu}_{1,k}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\Delta \boldsymbol{\mu}_{2,k}]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix},$$

which shows that

$$\|\mathcal{Q}_{2,k}\|_2 = \mathcal{O}\left(\alpha_k^{\min}\right),$$

since $|\alpha_i - \alpha_i^{\min}| \le (\iota_0/\iota_1^2)(\alpha_k^{\min})^2$ and Lemma 15. □

LEMMA 29. *Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then*

$$\mathbb{E}\left[\|\mathcal{Q}_{1,k}\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right). \tag{C.6}$$

PROOF. By the definition of $\bar{\boldsymbol{g}}_i - \nabla f(\boldsymbol{x}_i)$, we have

$$\begin{aligned} \begin{pmatrix} \bar{\boldsymbol{g}}_i - \nabla f(\boldsymbol{x}_i) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} &= \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'}) \right) \beta_h \begin{pmatrix} \boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \\ &\quad + \sum_{h=K^*}^{i} \left(\prod_{h'=h}^{i} (1-\beta_{h'}) \right) \begin{pmatrix} \nabla f(\boldsymbol{x}_{h-1}) - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \\ &:= \mathcal{W}_{1,i} + \mathcal{W}_{2,i}. \end{aligned}$$

(C.7)

$$\begin{aligned} &\mathbb{E}\left[\|\mathcal{Q}_{1,k}\|_2^2\right] \\ \le &\Upsilon_H^2 \mathbb{E}\left[\left(\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \|\mathcal{W}_{1,i} + \mathcal{W}_{2,i}\|_2 \right)^2 \right] \\ \le &\Upsilon_H^2 \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \sum_{i'=K^*}^{k} \prod_{j'=i'+1}^{k} \left(1-\alpha_{j'}^{\min}\right) \alpha_{i'}^{\min} \mathbb{E}\left[\|\mathcal{W}_{1,i} + \mathcal{W}_{2,i}\|_2 \|\mathcal{W}_{1,i'} + \mathcal{W}_{2,i'}\|_2\right] \\ \le &\Upsilon_H^2 \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \sum_{i'=K^*}^{k} \prod_{j'=i'+1}^{k} \left(1-\alpha_{j'}^{\min}\right) \alpha_{i'}^{\min} \sqrt{\mathbb{E}\left[\|\mathcal{W}_{1,i} + \mathcal{W}_{2,i}\|_2^2\right]} \sqrt{\mathbb{E}\left[\|\mathcal{W}_{1,i'} + \mathcal{W}_{2,i'}\|_2^2\right]} \\ \le &\Upsilon_H^2 \left(\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \sqrt{\mathbb{E}\left[\|\mathcal{W}_{1,i} + \mathcal{W}_{2,i}\|_2^2\right]} \right)^2. \end{aligned}$$

Note that $\mathbb{E}\left[\|\mathcal{W}_{1,i}\|_2^2\right] \le M_\sigma \sum_{h=K^*}^{i} \prod_{h'=h+1}^{i} (1-\beta_{h'})^2 \beta_h^2 \le 2M_\sigma \beta_i$ and $\|\mathcal{W}_{2,i}\|_2^2 \le M_{\boldsymbol{\ell},\boldsymbol{u}}^2 \left(\sum_{h=K^*}^{i} \prod_{h'=h+1}^{i} (1-\beta_{h'}) \alpha_h \right)^2 \le 2M_{\boldsymbol{\ell},\boldsymbol{u}}^2 \alpha_i^2/\beta_i^2 = \mathcal{O}\left(\beta_i + (\alpha_i^{\min})^2/\beta_i^2\right)$, for i sufficiently large, then $\mathbb{E}\left[\|\mathcal{Q}_{1,k}\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right)$. Here, we require that $\iota_1 > b_2$ if $b_1 = 1$, using Lemma 15. □

LEMMA 30. *Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then*

$$\mathbb{E}\left[\left\|\mathcal{Q}_{3,k}\right\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right),$$

and

$$\mathbb{E}\left[\left\|\begin{pmatrix} \boldsymbol{x}_k - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right).$$

PROOF. Recall the definition of $\mathcal{Q}_{3,k}$ that

$$\mathcal{Q}_{3,k} = \prod_{i=K^*}^{k}(1-\alpha_i^{\min})\begin{pmatrix} \boldsymbol{x}_{K^*} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{K^*} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,K^*} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,K^*} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} + \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}(1-\alpha_j^{\min})\alpha_i^{\min}\boldsymbol{\delta}_i,$$

we have the following recursion

$$\mathcal{Q}_{3,k+1} = \left(1-\alpha_{k+1}^{\min}\right)\mathcal{Q}_{3,k} + \alpha_{k+1}^{\min}\boldsymbol{\delta}_{k+1}. \tag{C.8}$$

Here,

$$\begin{aligned}
\|\boldsymbol{\delta}_{k+1}\|_2 &\le \left\|(\boldsymbol{H}^*)^{-1}\right\|_2\|\boldsymbol{\psi}_{k+1}\|_2 + \left\|\boldsymbol{H}_{k+1}^{-1} - (\boldsymbol{H}^*)^{-1}\right\|_2\|\nabla\mathcal{L}_{k+1}\|_2 \\
&\le \kappa_{\nabla\mathcal{L}}\Upsilon_H\left\|\begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2^2 + \kappa_{\nabla\mathcal{L}}\Upsilon_L\|\boldsymbol{H}_{k+1} - \boldsymbol{H}^*\|_2\left\|\begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2 \\
&:= \varepsilon_{k+1}\left\|\begin{pmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k+1} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k+1} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2 \le \varepsilon_{k+1}\left(\|\mathcal{Q}_{1,k}\|_2 + \|\mathcal{Q}_{2,k}\|_2 + \|\mathcal{Q}_{3,k}\|_2\right),
\end{aligned}$$

where we define

$$\varepsilon_k := \kappa_{\nabla\mathcal{H}}\Upsilon_L\left\|\begin{pmatrix} \boldsymbol{x}_k - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}\right\|_2 + \kappa_{\nabla\mathcal{L}}\Upsilon_L\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2.$$

Then, for any $a \in (0,1)$ and there exists the corresponding threshold $K_a \ge K^*$ such that $\varepsilon_{k+1} \le a$ and

$$\|\mathcal{Q}_{3,k+1}\|_2 \le \left(1-(1-a)\alpha_{k+1}^{\min}\right)\|\mathcal{Q}_{3,k}\|_2 + a\alpha_{k+1}^{\min}\cdot\left(\|\mathcal{Q}_{1,k}\|_2 + \|\mathcal{Q}_{2,k}\|_2\right),$$

for all $k \geq K_a$, as $\varepsilon_k \to 0$ almost surely. We then develop the recursion for $\|\mathcal{Q}_{3,k+1}\|_2$ that

(C.9)

$$
\begin{aligned}
&\|\mathcal{Q}_{3,k+1}\|_2 \\
&\leq \left(1-(1-a)\alpha_{k+1}^{\min}\right)\|\mathcal{Q}_{3,k}\|_2 + a\alpha_{k+1}^{\min}\cdot\left(\|\mathcal{Q}_{1,k}\|_2 + \|\mathcal{Q}_{2,k}\|_2\right) \\
&\leq \left(1-(1-a)\alpha_{k+1}^{\min}\right)\left(1-(1-a)\alpha_{k}^{\min}\right)\|\mathcal{Q}_{3,k-1}\|_2 \\
&\quad + \left(1-(1-a)\alpha_{k+1}^{\min}\right)a\alpha_{k}^{\min}\cdot\left(\|\mathcal{Q}_{1,k-1}\|_2 + \|\mathcal{Q}_{2,k-1}\|_2\right) \\
&\quad + a\alpha_{k+1}^{\min}\cdot\left(\|\mathcal{Q}_{1,k}\|_2 + \|\mathcal{Q}_{2,k}\|_2\right) \\
&\leq \cdots \\
&\leq \prod_{j=K_a+1}^{k+1}\left(1-(1-a)\alpha_j^{\min}\right)\|\mathcal{Q}_{3,K_a}\|_2 + \sum_{i=K_a+1}^{k+1}\left(\prod_{j=i+1}^{k+1}\left(1-(1-a)\alpha_j^{\min}\right)\right)a\alpha_i^{\min}\cdot\left(\|\mathcal{Q}_{1,i-1}\|_2 + \|\mathcal{Q}_{2,i-1}\|_2\right),
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathcal{Q}_{3,k}\|_2^2\right] \\
&\leq 2\left(\prod_{j=K_a+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\|\mathcal{Q}_{3,K_a}\|_2\right)^2 + 2\sum_{i=K_a+1}^{k}\left(\prod_{j=i+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\right)a\alpha_i^{\min} \\
&\quad\cdot\sum_{i'=K_a+1}^{k}\left(\prod_{j'=i'+1}^{k}\left(1-(1-a)\alpha_{j'}^{\min}\right)\right)a\alpha_{i'}^{\min}\cdot\mathbb{E}\left[\left(\|\mathcal{Q}_{1,i-1}\|_2 + \|\mathcal{Q}_{2,i-1}\|_2\right)\left(\|\mathcal{Q}_{1,i'-1}\|_2 + \|\mathcal{Q}_{2,i'-1}\|_2\right)\right] \\
&\leq 2\left(\prod_{j=K_a+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\|\mathcal{Q}_{3,K_a}\|_2\right)^2 + 2\sum_{i=K_a+1}^{k}\left(\prod_{j=i+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\right)a\alpha_i^{\min} \\
&\quad\cdot\sum_{i'=K_a+1}^{k}\left(\prod_{j'=i'+1}^{k}\left(1-(1-a)\alpha_{j'}^{\min}\right)\right)a\alpha_{i'}^{\min}\cdot\sqrt{\mathbb{E}\left[\left(\|\mathcal{Q}_{1,i-1}\|_2 + \|\mathcal{Q}_{2,i-1}\|_2\right)^2\right]}\sqrt{\mathbb{E}\left[\left(\|\mathcal{Q}_{1,i'-1}\|_2 + \|\mathcal{Q}_{2,i'-1}\|_2\right.\right.} \\
&\leq 2\left(\prod_{j=K_a+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\|\mathcal{Q}_{3,K_a}\|_2\right)^2 \\
&\quad + 2\left(\sum_{i=K_a+1}^{k}\left(\prod_{j=i+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\right)a\alpha_i^{\min}\sqrt{\mathbb{E}\left[\left(\|\mathcal{Q}_{1,i-1}\|_2 + \|\mathcal{Q}_{2,i-1}\|_2\right)^2\right]}\right)^2.
\end{aligned}
$$

Here, the fact that $\mathbb{E}\left[\left(\|\mathcal{Q}_{1,i}\|_2 + \|\mathcal{Q}_{2,i}\|_2\right)^2\right] \leq 2\mathbb{E}\left[\|\mathcal{Q}_{1,i}\|_2^2 + \|\mathcal{Q}_{2,i}\|_2^2\right] = \mathcal{O}\left(\beta_i + (\alpha_i^{\min})^2/\beta_i^2\right)$ implies $\mathbb{E}\left[\|\mathcal{Q}_{3,k}\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right)$. The second relation comes from the fact that $\mathbb{E}\left[\|\mathcal{Q}_{1,k}\|_2^2\right]$, $\mathbb{E}\left[\|\mathcal{Q}_{2,k}\|_2^2\right]$ and $\mathbb{E}\left[\|\mathcal{Q}_{3,k}\|_2^2\right]$ are at least of the order $\mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right)$. Here, we require that $\iota_1 > \frac{b_2}{2(1-a)}$ if $b_1 = 1$. Since $a \in (0,1)$ can be arbitrarily close to 0, we know $\frac{b_2}{2(1-a)} \leq b_2 < 1$ if $a \leq \frac{1}{2}$, and thus the condition is automatically satisfied for $a \leq \frac{1}{2}$ and $\iota_1 > b_2$. □

LEMMA 31. *Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then we have*

$$\mathbb{E}\left[\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right) \tag{C.10}$$

and

$$\mathbb{E}\left[\left\|\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1}\right\|_2^2\right] = \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right). \tag{C.11}$$

PROOF. We revisit the result in Lemma 26 that $\left\|\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1}\right\|_2 \leq \Upsilon_L \|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2$. Then, in the left part of the proof, we mainly show the first equality.

(C.12)

$$\begin{aligned}\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2 \leq & \left\|\frac{1}{k+1}\sum_{i=0}^{k}\left(\nabla^2 f(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f(\boldsymbol{x}_i)\right)\right\|_2 + \frac{\kappa_{\nabla^2 f}}{k+1}\sum_{i=0}^{k}\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^*\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*\\ [\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}\\ [\boldsymbol{\mu}_{2,i} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)}\end{pmatrix}\right\|_2 \\ & + \kappa_{\nabla c}\left\|\begin{pmatrix}\boldsymbol{x}_k - \boldsymbol{x}^*\\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*\\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}\\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)}\end{pmatrix}\right\|_2.\end{aligned}$$

Note that $\boldsymbol{\Delta}_k$ is the modification to the positive-definiteness of $\boldsymbol{B}_k$. If $\boldsymbol{\Delta}_k$ is the matrix with the smallest ℓ_2-norm such that $\boldsymbol{B}_k$ is positive definite, then $\|\boldsymbol{\Delta}_k\|_2 \leq \left\|\boldsymbol{B}_k - \nabla^2 f(\boldsymbol{x}^*) - \sum_{i=1}^{r}(\boldsymbol{\lambda}^*)_i \nabla^2 c_i(\boldsymbol{x}^*)\right\|_2$. Here, the strong law of large number shows that

(C.13)

$$\left\|\frac{1}{k+1}\sum_{i=0}^{k}\left(\nabla^2 f(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f(\boldsymbol{x}_i)\right)\right\|_2 = o\left(\sqrt{\frac{(\log k)^{1+\nu}}{k}}\right) = \mathcal{O}\left(\sqrt{\beta_k} + \alpha_k^{\min}/\beta_k\right), \text{ almost surely,}$$

for any $\nu > 0$. It further shows that $\boldsymbol{H}_k$ (resp. $\boldsymbol{B}_k$) converges to $\boldsymbol{H}^*$ (resp. $\boldsymbol{B}^*$) almost surely. Then

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{H}_k - \boldsymbol{H}^*\|_2^2\right] \leq & 3\mathbb{E}\left[\left\|\frac{1}{k+1}\sum_{i=0}^{k}\left(\nabla^2 f(\boldsymbol{x}_i;\zeta_i) - \nabla^2 f(\boldsymbol{x}_i)\right)\right\|_2^2\right] \\ & + \frac{3\kappa_{\nabla^2 f}^2}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\left\|\begin{pmatrix}\boldsymbol{x}_i - \boldsymbol{x}^*\\ \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*\\ [\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}\\ [\boldsymbol{\mu}_{2,i} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)}\end{pmatrix}\right\|_2^2\right] + 3\kappa_{\nabla c}^2\mathbb{E}\left[\left\|\begin{pmatrix}\boldsymbol{x}_k - \boldsymbol{x}^*\\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*\\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)}\\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)}\end{pmatrix}\right\|_2^2\right] \\ = & \mathcal{O}\left(\beta_k + (\alpha_k^{\min})^2/\beta_k^2\right).\end{aligned}$$

□

LEMMA 32. *Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then*

$$\mathbb{E}\left[\|\mathcal{W}_{2,k}\|_2^2\right] = o\left(\alpha_k^{\min}\right).$$

PROOF. For simplicity, we denote

$$\boldsymbol{v}_k = \begin{pmatrix} -\bar{\boldsymbol{g}}_k - \boldsymbol{\lambda}_k \nabla \boldsymbol{c}(\boldsymbol{x}_k) + \boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{2,k} \\ -\boldsymbol{c}(\boldsymbol{x}_k) \\ [\boldsymbol{x}_k - \boldsymbol{\ell}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{u} - \boldsymbol{x}_k]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix}, \, and \, \boldsymbol{v}^* = \begin{pmatrix} -\nabla f(\boldsymbol{x}^*) - \boldsymbol{\lambda}^* \nabla \boldsymbol{c}(\boldsymbol{x}^*) + \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^* \\ -\boldsymbol{c}(\boldsymbol{x}^*) \\ [\boldsymbol{x}^* - \boldsymbol{\ell}]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{u} - \boldsymbol{x}^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} = \mathbf{0}.$$

Then, there exist some $\kappa_v > 0$ such that

$$\|\boldsymbol{v}_k - \boldsymbol{v}^*\|_2 \leq \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2 + \kappa_v \left\| \begin{pmatrix} \boldsymbol{x}_k - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} \right\|_2.$$

We further have

$$\begin{aligned} \mathbb{E}\left[\|\bar{\boldsymbol{p}}_k\|_2^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{H}_k^{-1}\boldsymbol{v}_k\right\|_2^2\right] = \mathbb{E}\left[\left\|\boldsymbol{H}_k^{-1}\boldsymbol{v}_k - \boldsymbol{H}_k^{-1}\boldsymbol{v}^*\right\|_2^2\right] \\ &\leq \Upsilon_H^2 \mathbb{E}\left[\|\boldsymbol{v}_k - \boldsymbol{v}^*\|_2^2\right] \\ &\leq 2\Upsilon_H^2 \left(\mathbb{E}\left[\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|_2^2\right] + \kappa_v^2 \mathbb{E}\left[\left\| \begin{pmatrix} \boldsymbol{x}_k - \boldsymbol{x}^* \\ \boldsymbol{\lambda}_k - \boldsymbol{\lambda}^* \\ [\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_1^*]_{\mathcal{I}(\boldsymbol{x}^*)} \\ [\boldsymbol{\mu}_{2,k} - \boldsymbol{\mu}_2^*]_{\mathcal{J}(\boldsymbol{x}^*)} \end{pmatrix} \right\|_2^2 \right] \right) \\ &= \mathcal{O}\left(\beta_k\right). \end{aligned} \tag{C.14}$$

Then

$$\mathbb{E}\left[\|\mathcal{W}_{2,k}\|_2^2\right] \leq \left(\sum_{h=K^*}^{k} \prod_{h'=h+1}^{k} (1-\beta_{h'})\, \alpha_{h-1} \sqrt{\mathbb{E}\left[\|\bar{\boldsymbol{p}}_{h-1}\|_2^2\right]} \right)^2 = \mathcal{O}\left((\alpha_k^{\min})^2/\beta_k\right) = o\left(\alpha_k^{\min}\right).$$

After putting back N times, we have

$$\mathbb{E}\left[\|\mathcal{W}_{1,k}\|_2^2\right] = \mathcal{O}\left(\beta_k\right),$$

and

$$\mathbb{E}\left[\|\mathcal{W}_{2,k}\|_2^2\right] = \mathcal{O}\left(\beta_k \sum_{i=1}^{N} \left(\alpha_k/\beta_k\right)^{2i} + \left(\alpha_k/\beta_k\right)^{2(N+1)} \right).$$

Observe that $\beta_k \sum_{i=1}^{N} \left(\alpha_k^{\min}/\beta_k\right)^{2i} = o\left(\alpha_k^{\min}\right)$ and N can be any arbitrarily large integer, then

$$\mathbb{E}\left[\|\mathcal{W}_{2,k}\|_2^2\right] = o\left(\alpha_k^{\min}\right),$$

under the condition that $b_2 < b_1$. □

LEMMA 33. *Denote*

$$\mathcal{E}_{1,k}^* = \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min}) \alpha_i^{\min} \left((\boldsymbol{H}^*)^{-1} - \boldsymbol{H}_i^{-1} \right) \begin{pmatrix} \bar{\boldsymbol{g}}_i - \nabla f(\boldsymbol{x}_i) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then

$$\mathbb{E}\left[\left\|\mathcal{E}_{1,k}^*\right\|_2\right] = \mathcal{O}\left(\beta_k\right), \tag{C.15}$$

and

$$\mathbb{E}\left[\left\|\mathcal{Q}_{3,k}\right\|_2\right] = \mathcal{O}\left(\beta_k\right). \tag{C.16}$$

PROOF.

$$\begin{aligned}\mathbb{E}\left[\left\|\mathcal{E}_{1,k}^*\right\|_2\right] &\leq \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}\left(1-\alpha_j^{\min}\right)\alpha_i^{\min}\sqrt{\mathbb{E}\left[\left\|\boldsymbol{H}_k^{-1}-\left(\boldsymbol{H}^*\right)^{-1}\right\|_2^2\right]}\sqrt{\mathbb{E}\left[\left\|\mathcal{W}_{1,i}+\mathcal{W}_{2,i}\right\|_2^2\right]}\\ &= \mathcal{O}\left(\beta_k\right).\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left\|\mathcal{Q}_{3,k}\right\|_2\right] = &\prod_{j=K_a+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\left\|\mathcal{Q}_{3,K_a}\right\|_2\\ &+\sum_{i=K_a+1}^{k}\left(\prod_{j=i+1}^{k}\left(1-(1-a)\alpha_j^{\min}\right)\right)\alpha_i^{\min}\cdot\sqrt{\mathbb{E}\left[\varepsilon_i^2\right]}\sqrt{\mathbb{E}\left[\left(\left\|\mathcal{Q}_{1,i-1}\right\|_2+\left\|\mathcal{Q}_{2,i-1}\right\|_2\right)^2\right]}.\end{aligned}$$

Here,

$$\mathbb{E}\left[\varepsilon_i^2\right] = \mathcal{O}\left(\mathbb{E}\left[\left\|\mathcal{Q}_{1,i}\right\|_2^2+\left\|\mathcal{Q}_{2,i}\right\|_2^2+\left\|\mathcal{Q}_{3,i}\right\|_2^2+\left\|\boldsymbol{H}_{i+1}-\boldsymbol{H}^*\right\|_2^2\right]\right) = \mathcal{O}\left(\beta_i\right)$$

and

$$\mathbb{E}\left[\left(\left\|\mathcal{Q}_{1,i}\right\|_2+\left\|\mathcal{Q}_{2,i}\right\|_2\right)^2\right] = \mathcal{O}\left(\beta_i\right)$$

complete the first part of the proof. □

LEMMA 34. *Let*

$$\mathcal{Q}_{1,k} = \mathcal{Q}_{1,k}^* + \mathcal{E}_{1,k}^*,$$

where

$$\begin{aligned}\mathcal{Q}_{1,k}^* &= \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}\left(1-\alpha_j^{\min}\right)\alpha_i^{\min}\left(-\boldsymbol{H}^*\right)^{-1}\begin{pmatrix}\bar{\boldsymbol{g}}_i-\nabla f(\boldsymbol{x}_i)\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}\\ &:= \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}\left(1-\alpha_j^{\min}\right)\alpha_i^{\min}\left(-\boldsymbol{H}^*\right)^{-1}\left(\mathcal{W}_{1,i}+\mathcal{W}_{2,i}\right).\end{aligned}$$

and

$$\mathcal{E}_{1,k}^* = \sum_{i=K^*}^{k}\prod_{j=i+1}^{k}\left(1-\alpha_j^{\min}\right)\alpha_i^{\min}\left(\left(-\boldsymbol{H}_i\right)^{-1}-\left(-\boldsymbol{H}^*\right)^{-1}\right)\begin{pmatrix}\bar{\boldsymbol{g}}_i-\nabla f(\boldsymbol{x}_i)\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}.$$

Under Assumptions 5 and 6, and suppose that $\iota_1 > b_2$ if $b_1 = 1$, then

$$\frac{1}{\sqrt{\alpha_k^{min}}}\mathcal{Q}_{1,k}^* \to \mathcal{N}\left(\mathbf{0}, \Theta\boldsymbol{\Omega}^*\right). \tag{C.17}$$

PROOF. Let

(C.18)

$$\begin{aligned}
\mathcal{Q}_{1,k}^{**} :=& \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} (-\boldsymbol{H}^*)^{-1} \mathcal{W}_{1,i} \\
=& \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'}) \right) \beta_h (-\boldsymbol{H}^*)^{-1} \begin{pmatrix} \boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \\
=& \sum_{h=K^*}^{k} \sum_{i=h}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \prod_{h'=h+1}^{i} (1-\beta_{h'})\beta_h (-\boldsymbol{H}^*)^{-1} \begin{pmatrix} \boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \\
:=& \sum_{h=K^*}^{k} a_{h,k} (-\boldsymbol{H}^*)^{-1} \begin{pmatrix} \boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} := \sum_{h=K^*}^{k} \boldsymbol{s}_{h,k},
\end{aligned}$$

where $a_{h,k} = \sum_{i=h}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right) \alpha_i^{\min} \prod_{h'=h+1}^{i} (1-\beta_{h'})\beta_h$ and $\boldsymbol{s}_{h,k}$ are independent for different h. The asymptotic normality can be implied by the central limit theorem for the martingale difference array. Before that, we first verify the corresponding conditions. We first denote

$$\boldsymbol{\phi}_h^* = (-\boldsymbol{H}^*)^{-1} \begin{pmatrix} \boldsymbol{g}_h - \nabla f(\boldsymbol{x}_h) \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

and note that $\mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] \to \boldsymbol{\Omega}^*$ as $h \to \infty$ almost surely, since the smoothness of $f(\boldsymbol{x}, \xi)$ shows that

$$\begin{aligned}
\boldsymbol{\Lambda}_i :=& \mathbb{E}\left[\boldsymbol{g}_i \boldsymbol{g}_i^\top - \nabla f(\boldsymbol{x}_i) \nabla f(\boldsymbol{x}_i)^\top | \mathcal{F}_{i-1}\right] - \mathbb{E}\left[\nabla f(\boldsymbol{x}^*;\xi) \nabla f(\boldsymbol{x}^*;\xi)^\top - \nabla f(\boldsymbol{x}^*) \nabla f(\boldsymbol{x}^*)^\top\right] \\
=& \mathbb{E}\left[\nabla f(\boldsymbol{x}_i;\xi) \nabla f(\boldsymbol{x}_i;\xi)^\top - \nabla f(\boldsymbol{x}^*;\xi) \nabla f(\boldsymbol{x}_i;\xi)^\top | \mathcal{F}_{i-1}\right] + \mathbb{E}\left[\nabla f(\boldsymbol{x}^*;\xi) \nabla f(\boldsymbol{x}_i;\xi)^\top - \nabla f(\boldsymbol{x}^*;\xi) \nabla f(\boldsymbol{x}^*;\xi)^\top | \mathcal{F}\right. \\
&+ \nabla f(\boldsymbol{x}_i) \nabla f(\boldsymbol{x}_i)^\top - \nabla f(\boldsymbol{x}^*) \nabla f(\boldsymbol{x}_i)^\top + \nabla f(\boldsymbol{x}^*) \nabla f(\boldsymbol{x}_i)^\top - \nabla f(\boldsymbol{x}^*) \nabla f(\boldsymbol{x}^*)^\top \\
\leq& \kappa_{\nabla f} \left\|\boldsymbol{x}_i - \boldsymbol{x}^*\right\|_2 \left(\sqrt{\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_i;\xi)\right\|_2^2 | \mathcal{F}_{i-1}\right]} + \sqrt{\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}^*;\xi)\right\|_2^2\right]} + \left\|\nabla f(\boldsymbol{x}_i)\right\|_2 + \left\|\nabla f(\boldsymbol{x}^*)\right\|_2 \right) \\
\leq& 4\kappa_{\nabla f} M_{\nabla f} \left\|\boldsymbol{x}_i - \boldsymbol{x}^*\right\|_2 \to 0, \text{ as } i \to \infty.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\sum_{h=K^*}^{k} \mathbb{E}\left[a_{h,k}^2 \boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] \\
=&\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} \cdot \sum_{i'=K^*}^{k} \prod_{j'=i'+1}^{k} (1-\alpha_{j'}^{\min})\alpha_{i'}^{\min} \sum_{h=K^*}^{\min\{i,i'\}} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right)\left(\prod_{h'=h+1}^{i'} (1-\beta_{h'})\right)\beta_h^2 \\
&\cdot \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] \\
=&2\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} \cdot \sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{k} (1-\alpha_{j'}^{\min})\alpha_{i'}^{\min} \sum_{h=K^*}^{i'} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right)\left(\prod_{h'=h+1}^{i'} (1-\beta_{h'})\right)\beta_h^2 \\
&\cdot \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] - \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})^2 (\alpha_i^{\min})^2 \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right)^2 \beta_h^2 \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] \\
=&2\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right)^2 \alpha_i^{\min} \cdot \sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{i} \left(1-\alpha_{j'}^{\min}\right)(1-\beta_{j'})\,\alpha_{i'}^{\min} \sum_{h=K^*}^{i'} \left(\prod_{h'=h+1}^{i'} (1-\beta_{h'})\right)^2 \beta_h^2 \\
&\cdot \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] - \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})^2 (\alpha_i^{\min})^2 \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right)^2 \beta_h^2 \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right].
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\lim_{i'\to\infty} \beta_{i}^{-1} \sum_{h=K^*}^{i'} \left(\prod_{h'=h+1}^{i'} (1-\beta_{h'})\right)^2 \beta_h^2 \mathbb{E}\left[\boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] = \frac{1}{2}\boldsymbol{\Omega}^*, \\
&\lim_{i\to\infty} (\alpha_i^{\min})^{-1} \sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{i} \left(1-\alpha_{j'}^{\min}\right)(1-\beta_{j'})\,\alpha_{i'}^{\min}\beta_{i'} = 1, \\
&\lim_{k\to\infty} (\alpha_k^{\min})^{-1} \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} \left(1-\alpha_j^{\min}\right)^2 (\alpha_i^{\min})^2 = \Theta := \begin{cases} 1/2, & \text{if } b_1 < 1, \\ 1/\left(2-\frac{1}{\iota_1}\right), & \text{if } b_1 = 1, \end{cases} \\
&\lim_{i\to\infty} (\alpha_k^{\min})^{-1} \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})^2 (\alpha_i^{\min})^2 \beta_i = 0,
\end{aligned}
\tag{C.19}
$$

where we require that $\iota_1 > \frac{1}{2}$ if $b_1 = 1$. Therefore,

$$
\lim_{k\to\infty} (\alpha_k^{\min})^{-1} \sum_{h=K^*}^{k} \mathbb{E}\left[a_{h,k}^2 \boldsymbol{\phi}_h^* \boldsymbol{\phi}_h^{*\top} | \mathcal{F}_{h-1}\right] = \Theta\boldsymbol{\Omega}^*.
$$

We then verify the Lindeberg condition. It is equivalent to showing that

$$
\begin{aligned}
&\lim_{k\to\infty} \frac{1}{\alpha_k^{\min}} \sum_{h=K^*}^{k} a_{h,k}^2 \mathbb{E}\left[\|\boldsymbol{\phi}_h^*\|_2^2 \cdot \mathbf{1}_{\|a_{h,k}\boldsymbol{\phi}_h^*\|_2 \geq \epsilon(\alpha_k^{\min})^{1/2}} | \mathcal{F}_{h-1}\right] \\
\leq &\lim_{k\to\infty} \frac{1}{\epsilon(\alpha_k^{\min})^{3/2}} \sum_{h=K^*}^{k} a_{h,k}^3 \mathbb{E}\left[\|\boldsymbol{\phi}_h^*\|_2^3 | \mathcal{F}_{h-1}\right] \leq \lim_{k\to\infty} \frac{\Upsilon_\phi}{\epsilon(\alpha_k^{\min})^{3/2}} \sum_{h=K^*}^{k} a_{h,k}^3 = 0.
\end{aligned}
\tag{C.20}
$$

Suppose that $X_1, X_2, \cdots, X_k, \cdots$ are i.i.d. 1-dimensional random variables with zero mean and unit 3-moment, i.e., $\mathbb{E}\left[X_i^3\right] = 1$ for all $i \in \mathbb{N}$, then $\sum_{h=K^*}^{k} a_{h,k}^3 = \mathbb{E}\left[\left(\sum_{h=K^*}^{k} a_{h,k} X_h\right)^3\right]$. The equivalent form

$$\sum_{h=K^*}^{k} a_{h,k} X_h = \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min}) \alpha_i^{\min} \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right) \beta_h X_h$$

further shows

$$\begin{aligned}
&\sum_{h=K^*}^{k} a_{h,k}^3 = \mathbb{E}\left[\left(\sum_{h=K^*}^{k} a_{h,k} X_h\right)^3\right] \\
=&\mathbb{E}\left[\left(\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min}) \alpha_i^{\min} \sum_{h=K^*}^{i} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right) \beta_h X_h\right)^3\right] \\
\leq& 6 \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min}) \alpha_i^{\min} \cdot \sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{k} (1-\alpha_{j'}^{\min}) \alpha_{i'}^{\min} \cdot \sum_{i''=K^*}^{i'} \prod_{j''=i''+1}^{k} (1-\alpha_{j''}^{\min}) \alpha_{i''}^{\min} \\
&\cdot \sum_{h=K^*}^{i''} \left(\prod_{h'=h+1}^{i} (1-\beta_{h'})\right) \left(\prod_{h'=h+1}^{i'} (1-\beta_{h'})\right) \left(\prod_{h'=h+1}^{i''} (1-\beta_{h'})\right) \beta_h^3 \mathbb{E}\left[X_h^3\right] \\
=& 6 \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})^3 \alpha_i^{\min} \cdot \sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{i} (1-\alpha_{j'}^{\min})^2 (1-\beta_{j'}) \alpha_{i'}^{\min} \cdot \sum_{i''=K^*}^{i'} \prod_{j''=i''+1}^{i'} (1-\alpha_{j''}^{\min}) (1-\beta_{j''})^2 \alpha_{i''}^{\min} \\
&\cdot \sum_{h=K^*}^{i''} \left(\prod_{h'=h+1}^{i''} (1-\beta_{h'})\right)^3 \beta_h^3.
\end{aligned}$$

Similarly, note that

$$\begin{aligned}
&\sum_{h=K^*}^{i''} \left(\prod_{h'=h+1}^{i''} (1-\beta_{h'})\right)^3 \beta_h^3 = \frac{1}{3} = \mathcal{O}\left(\beta_{i''}^2\right) \\
&\sum_{i''=K^*}^{i'} \prod_{j''=i''+1}^{i'} (1-\alpha_{j''}^{\min}) (1-\beta_{j''})^2 \alpha_{i''}^{\min} \beta_{i''}^2 = \mathcal{O}\left(\alpha_{i'}^{\min} \beta_{i'}\right), \\
&\sum_{i'=K^*}^{i} \prod_{j'=i'+1}^{i} (1-\alpha_{j'}^{\min})^2 (1-\beta_{j'}) (\alpha_{i'}^{\min})^2 \beta_{i'} = \mathcal{O}\left((\alpha_i^{\min})^2\right), \\
&\sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})^3 (\alpha_i^{\min})^3 = \mathcal{O}\left((\alpha_k^{\min})^2\right),
\end{aligned} \tag{C.21}$$

where we require that $\iota_1 > b_2$ if $b_1 = 1$. The above results imply that $\sum_{h=K^*}^{k} a_{h,k}^3 = \mathcal{O}\left((\alpha_k^{\min})^2\right)$, thus the Lindeberg condition is satisfied. By the central limit theorem for martingale difference array (also called Lévy's theorem), we deduce that

$$\frac{1}{\sqrt{\alpha_k^{\min}}} \mathcal{Q}_{1,k}^{**} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Theta \mathbf{\Omega}^*\right).$$

Denote

$$\mathcal{E}_{1,k}^{**} = \mathcal{Q}_{1,k}^{*} - \mathcal{Q}_{1,k}^{**} := \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} (-\boldsymbol{H}^*)^{-1} \mathcal{W}_{2,i},$$

according to Lemma, we have

$$\mathbb{E}\left[\left\|\mathcal{E}_{1,k}^{**}\right\|_2\right] \leq \Upsilon_H \sum_{i=K^*}^{k} \prod_{j=i+1}^{k} (1-\alpha_j^{\min})\alpha_i^{\min} \sqrt{\mathbb{E}\left[\left\|\mathcal{W}_{2,i}\right\|_2^2\right]} = o\left(\sqrt{\alpha_k^{\min}}\right),$$

where we require that $\iota_1 > b_2$ if $b_1 = 1$. By Slutsky's theorem,

$$\frac{1}{\sqrt{\alpha_k^{\min}}} \mathcal{Q}_{1,k}^{*} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Theta\boldsymbol{\Omega}^*\right).$$

□

Proof for Theorem 4: it is a direct result from Lemmas 28, 33 and 34.

C.4. Proof for Theorem 5. The second relation is implied by the first one because the proof in Lemma 31 and the almost sure convergence of primal-dual iterates jointly show that $\left\|\boldsymbol{H}_k^{-1} - (\boldsymbol{H}^*)^{-1}\right\|_2 \to 0$, almost surely. We are left to show the first relation. Note that

$$\begin{aligned}
\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}^*\|_2 = & \left\| \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \boldsymbol{g}_i^\top - \mathbb{E}\left[\nabla f(\boldsymbol{x}^*;\zeta) \nabla f(\boldsymbol{x}^*;\zeta)^\top\right] \right\|_2 \\
& + \left\| \left(\frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \right) \left(\frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \right)^\top - \nabla f(\boldsymbol{x}^*) \nabla f(\boldsymbol{x}^*)^\top \right\|_2 .
\end{aligned}$$

$$\begin{aligned}
& \left\| \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \boldsymbol{g}_i^\top - \mathbb{E}\left[\nabla f(\boldsymbol{x}^*;\zeta) \nabla f(\boldsymbol{x}^*;\zeta)^\top\right] \right\|_2 \\
= & \left\| \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \boldsymbol{g}_i^\top - \mathbb{E}\left[\nabla f(\boldsymbol{x}_i;\zeta) \nabla f(\boldsymbol{x}_i;\zeta)^\top | \mathcal{F}_{i-1}\right] \right\|_2 \\
& + \left\| \frac{1}{k+1} \sum_{i=0}^{k} \mathbb{E}\left[\nabla f(\boldsymbol{x}_i;\zeta) \nabla f(\boldsymbol{x}_i;\zeta)^\top | \mathcal{F}_{i-1} - \nabla f(\boldsymbol{x}^*;\zeta) \nabla f(\boldsymbol{x}^*;\zeta)^\top\right] \right\|_2
\end{aligned}$$

The strong law of large number shows that

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i \boldsymbol{g}_i^\top - \mathbb{E}\left[\nabla f(\boldsymbol{x}_i;\zeta) \nabla f(\boldsymbol{x}_i;\zeta)^\top | \mathcal{F}_{i-1}\right] \right\|_2 = o\left(\sqrt{\frac{(\log k)^{1+\nu}}{k}}\right),$$

for any $\nu > 0$, almost surely. The almost sure convergence of iterates (i.e., $\boldsymbol{x}_k \to \boldsymbol{x}^*$) implies that

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} \mathbb{E}\left[\nabla f(\boldsymbol{x}_i;\zeta) \nabla f(\boldsymbol{x}_i;\zeta)^\top - \nabla f(\boldsymbol{x}^*;\zeta) \nabla f(\boldsymbol{x}^*;\zeta)^\top | \mathcal{F}_{i-1}\right] \right\|_2 \to 0,$$

almost surely. Similarly, for the second term

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} \boldsymbol{g}_i - \nabla f(\boldsymbol{x}^*) \right\|_2 \leq \left\| \frac{1}{k+1} \sum_{i=0}^{k} (\boldsymbol{g}_i - \nabla f(\boldsymbol{x}_k)) \right\|_2 + \left\| \frac{1}{k+1} \sum_{i=0}^{k} (\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*)) \right\|_2,$$

the strong law of large number also shows

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} (\boldsymbol{g}_i - \nabla f(\boldsymbol{x}_k)) \right\|_2 = o\left(\sqrt{\frac{(\log k)^{1+\nu}}{k}} \right),$$

for any $\nu > 0$ almost surely, and

$$\left\| \frac{1}{k+1} \sum_{i=0}^{k} (\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*)) \right\|_2 \to 0,$$

almost surely. Therefore, we complete the proof.