SIAM J. OPTIM.
Vol. 34, No. 2, pp. 2007–2037

FULLY STOCHASTIC TRUST-REGION SEQUENTIAL QUADRATIC PROGRAMMING FOR EQUALITY-CONSTRAINED OPTIMIZATION PROBLEMS*

YUCHEN FANG†, SEN NA‡, MICHAEL W. MAHONEY§, AND MLADEN KOLAR¶

Abstract. We propose a trust-region stochastic sequential quadratic programming algorithm (TR-StoSQP) to solve nonlinear optimization problems with stochastic objectives and deterministic equality constraints. We consider a fully stochastic setting, where at each step a single sample is generated to estimate the objective gradient. The algorithm adaptively selects the trust-region radius and, compared to the existing line-search StoSQP schemes, allows us to utilize indefinite Hessian matrices (i.e., Hessians without modification) in SQP subproblems. As a trust-region method for constrained optimization, our algorithm must address an infeasibility issue—the linearized equality constraints and trust-region constraints may lead to infeasible SQP subproblems. In this regard, we propose an *adaptive relaxation technique* to compute the trial step, consisting of a normal step and a tangential step. To control the lengths of these two steps while ensuring a scale-invariant property, we adaptively decompose the trust-region radius into two segments, based on the proportions of the rescaled feasibility and optimality residuals to the rescaled full KKT residual. The normal step has a closed form, while the tangential step is obtained by solving a trust-region subproblem, to which a solution ensuring the Cauchy reduction is sufficient for our study. We establish a global almost sure convergence guarantee for TR-StoSQP and illustrate its empirical performance on both a subset of problems in the CUTEst test set and constrained logistic regression problems using data from the LIBSVM collection.

Key words. constrained stochastic optimization, nonlinear optimization, trust-region method, sequential quadratic programming

MSC codes. 90-08, 90C15, 90C30, 90C55, 90C90

DOI. 10.1137/22M1537862

1. Introduction.

We consider the following constrained stochastic optimization problem:

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is a stochastic objective with $F(\cdot;\xi)$ being one of its realizations, $c:\mathbb{R}^d\to\mathbb{R}^m$ are deterministic equality constraints, ξ is a random variable following the distribution $\mathcal{P}$, and the expectation $\mathbb{E}[\cdot]$ is taken over the randomness of ξ. Problem (1.1) appears in various applications, including constrained deep neural networks [13],

*Received by the editors November 29, 2022; accepted for publication (in revised form) January 29, 2024; published electronically June 11, 2024.
https://doi.org/10.1137/22M1537862
Funding: This work was partially supported by the DOE, NSF, and ONR, as well as the JPMorgan Chase Faculty Research Award.
†Committee on Computational and Applied Mathematics, University of Chicago, Chicago, IL 60637 USA (ycfang@uchicago.edu).
‡International Computer Science Institute and Department of Statistics, University of California, Berkeley, CA 94720 USA (senna@berkeley.edu).
§International Computer Science Institute, Lawrence Berkeley National Laboratory, and Department of Statistics, University of California, Berkeley, CA 94720 USA (mmahoney@stat.berkeley.edu).
¶Marshall School of Business, University of Southern California, Los Angeles, CA 90089 USA (mkolar@marshall.usc.edu).

constrained maximum likelihood estimation [21], optimal control [7], PDE-constrained optimization [32], and network optimization [5].

There are numerous methods for solving constrained optimization problems with deterministic objectives. Among them, sequential quadratic programming (SQP) methods are one of the leading approaches and are effective for both small and large problems. When the objective is stochastic, some stochastic SQP (StoSQP) methods have been proposed recently [2, 3, 4, 18, 25, 26]. That body of literature considers the following two different setups for modeling the objective.

The first setup is called the random model setup [14], where samples with adaptive batch sizes are generated in each iteration to estimate the objective model (e.g., objective value and gradient). The algorithms under this setup often require the estimated objective model to satisfy certain adaptive accuracy conditions with a fixed probability in each iteration. Under this setup, [26] proposed an StoSQP algorithm for (1.1), which adopts a stochastic line search procedure with an exact augmented Lagrangian merit function to select the stepsize. Subsequently, [25] further enhanced the designs and arguments in [26] and developed an active-set StoSQP method to enable inequality constraints; and [4] considered a finite-sum objective and accelerated StoSQP by applying the SVRG technique [24], which, however, requires one to periodically compute the full objective gradient. Also, [1] introduced a norm test condition for StoSQP to adaptively select the batch sizes.

The second setup is called the fully stochastic setup [20], where a single sample is generated in each iteration to estimate the objective model. Under this setup, a prespecified sequence is often required as an input to assist with the step selection. For example, [3] designed an StoSQP scheme that uses a random projection procedure to select the stepsize. The projection procedure uses a prespecified sequence $\{\beta_k\}$, together with the estimated Lipschitz constants of the objective gradient and constraint Jacobian, to construct a projection interval in each iteration. A random quantity is then computed and projected into the interval to decide the stepsize, which ensures a sufficient reduction on the ℓ_1 merit function. Based on [3], some algorithmic and theoretical improvements have been reported: [2] dealt with rank-deficient Jacobians; [18] solved Newton systems inexactly; [16] analyzed the worst-case sample complexity; and [28] established the local rate and performed statistical inference for the method in [3].

The existing StoSQP algorithms converge globally either in expectation or almost surely, and enjoy promising empirical performance under favorable settings. However, there are three limitations that motivate our study. First, the algorithms are all line-search-based; that is, a search direction is first computed by solving an SQP subproblem, and then a stepsize is selected, either by random projection or by stochastic line search along the direction. However, it is observed that for deterministic problems, computing the search direction and selecting the stepsize jointly, as is done in trust-region methods, can lead to better performance in some cases [29, Chapter 4]. Second, to make SQP subproblems solvable, the existing schemes require the approximation of the Lagrangian Hessian to be positive definite in the null space of constraint Jacobian. Such a condition is common in the SQP literature [8, 29], while it is often achieved by Hessian modification, which excludes promising choices of the Hessian matrices, such as the unperturbed (stochastic) Hessian of the Lagrangian. Third, to show global convergence, the existing literature requires the random merit parameter to be not only stabilized but also sufficiently large (or small, depending on the context) with an unknown threshold. To achieve the latter goal, [25, 26] imposed an adaptive condition on the feasibility error when selecting the merit parameter, while [2, 3, 4, 18] imposed a symmetry condition on the noise distribution. In contrast, deterministic SQP schemes only require the stability of the merit parameter (see [8] and references therein).

In this paper, we consider the fully stochastic setup and design a trust-region stochastic SQP (TR-StoSQP) method to address the above limitations. As a trust-region method, TR-StoSQP computes the search direction and stepsize jointly, and, unlike line-search-based methods, it avoids Hessian modifications in formulating SQP subproblems. Thus, it can explore negative curvature directions of the Hessian. Further, our analysis only relies on the stability of the merit parameter (of the ℓ_2 merit function), which is consistent with deterministic SQP schemes. The design of TR-StoSQP is inspired by a stochastic trust-region method for solving *unconstrained* problems reported in [20], which improves the authors' prior design in [19] from using a linear model to a quadratic model to approximate the objective function. As in [20], our method inputs a user-specified radius-related sequence $\{\beta_k\}$ to generate the trust-region radius at each step. Beyond this similarity, our scheme differs from [20] in several aspects.

First, it is known that trust-region methods for constrained optimization are bothered by the *infeasibility issue*—the linearized constraints and trust-region constraints may have an empty intersection, leading to an infeasible SQP subproblem. While some literature on trust-region SQP has been proposed to address this issue [10, 11, 30, 36], we develop a novel *adaptive relaxation technique* to compute the trial step, which preserves a scale-invariant property and can be further adapted to our stochastic setup. In particular, we decompose the trial step into a normal step and a tangential step. Then, we control the lengths of the two steps by decomposing the trust-region radius into two segments *adaptively,* based on the proportions of the rescaled estimated feasibility and optimality residuals to the rescaled full KKT residual. Compared to the existing relaxation techniques, *our relaxation technique does not require any tuning parameters.* See section 2 for details.

Second, in TR-StoSQP, we properly compute some control parameters using known or estimable quantities. By the computation, we no longer need to tune the other two input parameter sequences as in [20] (i.e., $\{\gamma_{1,k}, \gamma_{2,k}\}$ in their notation), except to tune the input radius-related sequence $\{\beta_k\}$. Further, we use the control parameters to adjust the input sequence $\{\beta_k\}$ when computing the trust-region radius, so that $\{\beta_k\} \subseteq (0, \beta_{\max}]$ with any $\beta_{\max} > 0$ is sufficient for our convergence analysis. Our design simplifies the one in [20], where there are three parameter sequences to tune whose conditions are highly coupled (see [20, Lemma 4.5]). In addition, as the authors stated, [20] rescaled the Hessian matrix based on the input $\{\gamma_{1,k}\}$, which is not ideal (because the rescaling step modifies the curvature information of the Hessian). We have removed this step in our design.

To our knowledge, TR-StoSQP is the first trust-region SQP algorithm for solving constrained optimization problems under a fully stochastic setup. With a stabilized merit parameter, we establish the global convergence property of TR-StoSQP. In particular, we show that (i) when $\beta_k = \beta$ for all $k \geq 0$, the expectation of weighted averaged KKT residuals converges to a neighborhood around zero; (ii) when β_k decays properly such that $\sum \beta_k = \infty$ and $\sum \beta_k^2 < \infty$, the KKT residuals converge to zero almost surely. These results are similar to the ones for unconstrained and constrained problems established under the fully stochastic setup in [2, 3, 18, 20]. However, we have weaker conditions on the objective gradient noise (e.g., we consider a growth condition) and on the sequence β_k (e.g., we only require $\beta_k \leq \beta_{\max}$). See the discussions after Theorems 4.9 and 4.11 for more details. We also note that a recent paper [35] studied a noisy trust-region method for *unconstrained deterministic* optimization. In that method, the value and gradient of the objective are evaluated with bounded deterministic noise. The authors showed that the trust-region iterates visit a neighborhood of the stationarity infinitely often, with the radius proportional to the noise magnitude. Given the significant differences between stochastic and deterministic problems, and between constrained and unconstrained problems, our algorithm design and analysis are quite different from [35]. That said, when studying the stability of the merit parameter, we follow existing literature (e.g., [3, 2, 26]) and also require the bounded gradient noise condition. We implement TR-StoSQP on a subset of problems in the CUTEst test set and on constrained logistic regression problems using data from the LIBSVM collection. Numerical results demonstrate the promising performance of our method.

Notation. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the operator norm for matrices. I denotes the identity matrix, and $\mathbf{0}$ denotes the zero matrix (or vector). Their dimensions are clear from the context. We let $G(\boldsymbol{x}) = \nabla^T c(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ be the Jacobian matrix of the constraints and $P(\boldsymbol{x}) = I - G^T(\boldsymbol{x})[G(\boldsymbol{x})G^T(\boldsymbol{x})]^{-1}G(\boldsymbol{x})$ be the projection matrix to the null space of $G(\boldsymbol{x})$. We use $\bar{g}(\boldsymbol{x}) = \nabla F(\boldsymbol{x};\xi)$ to denote an estimate of $\nabla f(\boldsymbol{x})$ and use $\bar{(\cdot)}$ to denote stochastic quantities.

Structure of the paper. We introduce the adaptive relaxation technique in section 2. We propose the trust-region stochastic SQP (TR-StoSQP) algorithm in section 3 and establish its global convergence guarantee in section 4. Numerical experiments are presented in section 5, and conclusions are presented in section 6. Some additional analyses are provided in Appendix A.

2. Adaptive relaxation for deterministic setup.

The Lagrangian of problem (1.1) is $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^T c(\boldsymbol{x})$, where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the dual vector. Finding a first-order stationary point of (1.1) is equivalent to finding a pair $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ such that

$$\nabla\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{\lambda}^*) = \begin{pmatrix}\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\end{pmatrix} = \begin{pmatrix}\nabla f(\boldsymbol{x}^*) + G^T(\boldsymbol{x}^*)\boldsymbol{\lambda}^*\\ c(\boldsymbol{x}^*)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}.$$

We call $\|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})\|$ the optimality residual, $\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})\|$ (i.e., $\|c(\boldsymbol{x})\|$) the feasibility residual, and $\|\nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})\|$ the KKT residual. Given $\boldsymbol{x}_k$ in the kth iteration, we denote $\nabla f_k = \nabla f(\boldsymbol{x}_k)$, $c_k = c(\boldsymbol{x}_k)$, $G_k = G(\boldsymbol{x}_k)$, etc.

2.1. Preliminaries.

Given the iterate $\boldsymbol{x}_k$ and the trust-region radius Δ_k in the kth iteration, we compute an approximation B_k of the Lagrangian Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_k$ and aim to obtain the trial step $\Delta\boldsymbol{x}_k$ by solving a trust-region SQP subproblem

$$\min_{\Delta\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\Delta\boldsymbol{x}^T B_k \Delta\boldsymbol{x} + \nabla f_k^T \Delta\boldsymbol{x} \quad \text{s.t. } c_k + G_k\Delta\boldsymbol{x} = \mathbf{0},\ \|\Delta\boldsymbol{x}\| \le \Delta_k. \tag{2.1}$$

However, if $\{\Delta\boldsymbol{x}\in\mathbb{R}^d : c_k + G_k\Delta\boldsymbol{x} = \mathbf{0}\} \cap \{\Delta\boldsymbol{x}\in\mathbb{R}^d : \|\Delta\boldsymbol{x}\| \le \Delta_k\} = \emptyset$, then (2.1) does not have a feasible point. This *infeasibility issue* happens when the radius Δ_k is too short. To resolve this issue, one should not enlarge Δ_k, which would make the trust-region constraint useless and violate the spirit of the trust-region scheme. Instead, one should relax the linearized constraint $c_k + G_k\Delta\boldsymbol{x} = \mathbf{0}$.

Before introducing our *adaptive relaxation technique,* we review some classical relaxation techniques. To start, [11] relaxed the linearized constraint by $\|c_k + G_k\Delta\boldsymbol{x}\| \le \theta_k$ with $\theta_k = \|c_k + G_k\Delta\boldsymbol{x}_k^{CP}\|$, where $\Delta\boldsymbol{x}_k^{CP}$ is the Cauchy point (i.e., the best steepest descent step) of the following problem:

$$\min_{\Delta\boldsymbol{x}\in\mathbb{R}^d} \|c_k + G_k\Delta\boldsymbol{x}\| \quad \text{s.t.} \quad \|\Delta\boldsymbol{x}\| \le \Delta_k. \tag{2.2}$$

However, since after the relaxation one has to minimize a quadratic function over the intersection of two ellipsoids $\|c_k + G_k\Delta\boldsymbol{x}\| \le \theta_k$ and $\|\Delta\boldsymbol{x}\| \le \Delta_k$, the resulting SQP subproblem tends to be expensive to solve. See [39] for some insights into the difficulty, and see [40, 41] for the methods for positive definite B_k. Alternatively, [36] relaxed the linearized constraint by $\gamma_k c_k + G_k\Delta\boldsymbol{x} = \boldsymbol{0}$, with $\gamma_k \in (0,1]$ chosen to make the trust-region constraint of (2.1) inactive. However, [36] only showed the existence of an extremely small γ_k, and it did not provide a practical way to choose it. Subsequently, [10] refined the relaxation technique of [36] by a step decomposition. At the kth step, [10] decomposed the trial step $\Delta\boldsymbol{x}_k$ into a normal step $\boldsymbol{w}_k \in \mathrm{im}(G_k^T)$ and a tangential step $\boldsymbol{t}_k \in \ker(G_k)$, denoted as $\Delta\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{t}_k$. By the constraint $\gamma_k c_k + G_k\Delta\boldsymbol{x}_k = \boldsymbol{0}$, the normal step has a closed form as (suppose G_k has full row rank)

$$\boldsymbol{w}_k := \gamma_k\boldsymbol{v}_k := -\gamma_k \cdot G_k^T[G_kG_k^T]^{-1}c_k, \tag{2.3}$$

and the tangential step is expressed as $\boldsymbol{t}_k = Z_k\boldsymbol{u}_k$ for a vector $\boldsymbol{u}_k \in \mathbb{R}^{d-m}$. Here, the columns of $Z_k \in \mathbb{R}^{d\times(d-m)}$ form the bases of $\ker(G_k)$. [10] proposed to choose γ_k such that $\theta\Delta_k \le \|\boldsymbol{w}_k\| \le \Delta_k$ for a tuning parameter $\theta \in (0,1)$, and solve $\boldsymbol{u}_k$ from

$$\min_{\boldsymbol{u}\in\mathbb{R}^{d-m}} \frac{1}{2}\boldsymbol{u}^T Z_k^T B_k Z_k \boldsymbol{u} + (\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k\boldsymbol{u} \quad \text{s.t. } \|\boldsymbol{u}\|^2 \le \Delta_k^2 - \|\boldsymbol{w}_k\|^2. \tag{2.4}$$

Furthermore, [30] combined the techniques of [11, 10]; it solved the normal step $\boldsymbol{w}_k$ from problem (2.2) by replacing the constraint $\|\Delta\boldsymbol{x}\| \le \Delta_k$ with $\|\Delta\boldsymbol{x}\| \le \theta\Delta_k$ for some $\theta \in (0,1)$; and it solved the tangential step $\boldsymbol{t}_k = Z_k\boldsymbol{u}_k$ from problem (2.4). We note that the solution of (2.2) is naturally a normal step (i.e., lies in $\mathrm{im}(G_k^T)$), because any directions in $\ker(G_k)$ do not change the objective in (2.2).

Although the methods in [10, 30] allow one to employ Cauchy points for trust-region subproblems, they lack guidance for selecting the user-specified parameter θ, which controls the lengths of the normal and tangential steps. In fact, an inappropriate parameter θ may make either step conservative and further affect the effectiveness of the algorithm. As we show in (2.8) and (2.9) later, the normal step relates to the reduction of the feasibility residual, while the tangential step relates to the reduction of the optimality residual. We hope the two steps scale properly so that the model reduction achieved by $\Delta\boldsymbol{x}_k$ is large enough. To that end, we propose an adaptive relaxation technique, which is *parameter-free* in step decomposition compared to [36, 10, 30].

2.2. Our adaptive relaxation technique. We introduce our parameter-free relaxation procedure. The same as in [10], we relax the linearized constraint in (2.1) by $\gamma_k c_k + G_k\Delta\boldsymbol{x} = \boldsymbol{0}$, with γ_k defined later, and decompose the trial step by $\Delta\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{t}_k$. The normal step $\boldsymbol{w}_k$ is given by (2.3), and the tangential step is of the form $\boldsymbol{t}_k = Z_k\boldsymbol{u}_k$.

To control the lengths of the two steps while ensuring a scale-invariant property (cf. Remark 2.2), let us define the rescaled optimality vector $\nabla_{\boldsymbol{x}}\mathcal{L}_k^{RS} := \nabla_{\boldsymbol{x}}\mathcal{L}_k/\|B_k\|$, the feasibility vector $c_k^{RS} := c_k/\|G_k\|$, and the KKT vector $\nabla\mathcal{L}_k^{RS} := (\nabla_{\boldsymbol{x}}\mathcal{L}_k^{RS}, c_k^{RS})$. (One alternative choice of the rescaled feasibility vector can be $\boldsymbol{v}_k = G_k^T[G_kG_k^T]^{-1}c_k$.) Then, we *adaptively* decompose the trust-region radius Δ_k into two segments, based on the proportions of the rescaled feasibility and optimality residuals to the rescaled full KKT residual. We let

$$\breve{\Delta}_k = \frac{\|c_k^{RS}\|}{\|\nabla\mathcal{L}_k^{RS}\|} \cdot \Delta_k \qquad \text{and} \qquad \tilde{\Delta}_k = \frac{\|\nabla_{\boldsymbol{x}}\mathcal{L}_k^{RS}\|}{\|\nabla\mathcal{L}_k^{RS}\|} \cdot \Delta_k. \tag{2.5}$$

It is implicitly assumed that $\|B_k\|, \|G_k\|, \|\nabla\mathcal{L}_k\| \neq 0$, which is quite reasonable for SQP methods. We let $\breve{\Delta}_k$ control the length of the normal step $\boldsymbol{w}_k$ and $\tilde{\Delta}_k$ control the length of the tangential step $\boldsymbol{t}_k$. Specifically, we define γ_k as (recall $\boldsymbol{v}_k$ is defined in (2.3))

$$\gamma_k := \min\{\breve{\Delta}_k/\|\boldsymbol{v}_k\|, 1\} \tag{2.6}$$

so that $\|\boldsymbol{w}_k\| = \gamma_k\|\boldsymbol{v}_k\| \leq \breve{\Delta}_k$, and we compute $\boldsymbol{u}_k$ by solving

$$\min_{\boldsymbol{u}\in\mathbb{R}^{d-m}} m(\boldsymbol{u}) := \frac{1}{2}\boldsymbol{u}^T Z_k^T B_k Z_k \boldsymbol{u} + (\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k \boldsymbol{u} \quad \text{s.t.} \quad \|\boldsymbol{u}\| \leq \tilde{\Delta}_k. \tag{2.7}$$

When $\boldsymbol{v}_k = 0$ (i.e., $c_k = \boldsymbol{0}$), there is no need to choose γ_k and we set $\Delta\boldsymbol{x}_k = Z_k\boldsymbol{u}_k$. Problem (2.7) is a trust-region subproblem that appears in unconstrained optimization. In our analysis, we only require a vector $\boldsymbol{u}_k$ that reduces $m(\boldsymbol{u})$ by at least as much as the Cauchy point, which takes the direction of $-Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)$ and minimizes $m(\boldsymbol{u})$ within the trust region [29, Algorithm 4.2]. Such a reduction requirement can be achieved by various methods, including finding the exact solution or applying the dogleg or two-dimensional subspace minimization methods [29].

The following result provides a bound on the reduction in $m(\boldsymbol{u})$ that is different from the standard analysis of the Cauchy point; see, e.g., [29, Lemma 4.3].

Lemma 2.1. *Let $\boldsymbol{u}_k$ be an approximate solution to* (2.7) *that reduces the objective $m(\boldsymbol{u})$ by at least as much as the Cauchy point. For all $k \geq 0$, we have*

$$\begin{aligned} m(\boldsymbol{u}_k) - m(\boldsymbol{0}) &= \frac{1}{2}\boldsymbol{u}_k^T Z_k^T B_k Z_k \boldsymbol{u}_k + (\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k \boldsymbol{u}_k \\ &\leq -\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2. \end{aligned}$$

Proof. Let $\boldsymbol{u}_k^{CP}$ denote the Cauchy point. Since $m(\boldsymbol{u}_k) \leq m(\boldsymbol{u}_k^{CP})$, it suffices to analyze the reduction achieved by $\boldsymbol{u}_k^{CP}$. By the formula of $\boldsymbol{u}_k^{CP}$ in [29, (4.12)], we know that if $\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^3 \leq \tilde{\Delta}_k(\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k Z_k^T B_k Z_k Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)$, then $\boldsymbol{u}_k^{CP} = -\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^2/(\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k Z_k^T B_k Z_k Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k) \cdot Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)$. In this case, using $\|Z_k\| \leq 1$, we have

$$\begin{aligned} m(\boldsymbol{u}_k^{CP}) - m(\boldsymbol{0}) &= \frac{1}{2}(Z_k\boldsymbol{u}_k^{CP})^T B_k Z_k \boldsymbol{u}_k^{CP} + (\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k \boldsymbol{u}_k^{CP} \\ &= -\frac{1}{2}\frac{\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^4}{(\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k Z_k^T B_k Z_k Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)} \leq -\frac{1}{2}\frac{\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^2}{\|B_k\|}. \end{aligned}$$

Otherwise, $\boldsymbol{u}_k^{CP} = -\tilde{\Delta}_k/\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\| \cdot Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)$. In this case, we have

$$\begin{aligned} m(\boldsymbol{u}_k^{CP}) - m(\boldsymbol{0}) &= \frac{1}{2}(Z_k\boldsymbol{u}_k^{CP})^T B_k Z_k \boldsymbol{u}_k^{CP} + (\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k \boldsymbol{u}_k^{CP} \\ &= \frac{(\nabla f_k + B_k\boldsymbol{w}_k)^T Z_k Z_k^T B_k Z_k Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)}{2\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^2}\tilde{\Delta}_k^2 - \|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|\tilde{\Delta}_k \\ &\leq \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 - \|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|\tilde{\Delta}_k. \end{aligned}$$

Combining the above two cases, we have

$$m(\boldsymbol{u}_k^{CP}) - m(\boldsymbol{0}) \leq -\min\left\{-\frac{\|B_k\|\tilde{\Delta}_k^2}{2} + \|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|\tilde{\Delta}_k,\ \frac{\|Z_k^T(\nabla f_k + B_k\boldsymbol{w}_k)\|^2}{2\|B_k\|}\right\}.$$

Using the fact that

$$\begin{aligned}
&-\frac{1}{2}\|B_k\|\tilde{\Delta}_k^2+\|Z_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\|\tilde{\Delta}_k \\
&=-\frac{\|B_k\|}{2}\left(\tilde{\Delta}_k-\frac{\|Z_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\|}{\|B_k\|}\right)^2+\frac{\|Z_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\|^2}{2\|B_k\|} \\
&\leq\frac{\|Z_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\|^2}{2\|B_k\|},
\end{aligned}$$

we complete the proof. ∎

It is easy to see that our relaxation technique indeed results in a trial step that lies in the trust region. We have (noting that $\|Z_k\|\leq 1$)

$$\|\Delta\boldsymbol{x}_k\|^2=\|\boldsymbol{w}_k\|^2+\|\boldsymbol{t}_k\|^2=(\gamma_k\|\boldsymbol{v}_k\|)^2+\|\boldsymbol{u}_k\|^2\overset{(2.6),(2.7)}{\leq}\breve{\Delta}_k^2+\tilde{\Delta}_k^2\overset{(2.5)}{=}\Delta_k^2.$$

Recalling from (2.3) that $\boldsymbol{w}_k=-\gamma_k G_k^T[G_kG_k^T]^{-1}c_k$, we know $c_k+G_k\boldsymbol{w}_k=(1-\gamma_k)c_k$. Thus, we have

$$\|c_k+G_k\Delta\boldsymbol{x}_k\|-\|c_k\|=\|c_k+G_k\boldsymbol{w}_k\|-\|c_k\|=-\gamma_k\|c_k\|\leq 0, \tag{2.8}$$

where the strict inequality holds as long as $c_k\neq\mathbf{0}$. This inequality suggests that the normal step $\boldsymbol{w}_k$ helps to reduce the feasibility residual. Furthermore, when we define the least-squares Lagrangian multiplier as $\boldsymbol{\lambda}_k=-G_k^T[G_kG_k^T]\nabla f_k$, we have $P_k\nabla f_k=\nabla_{\boldsymbol{x}}\mathcal{L}_k$. Noting that $Z_kZ_k^T=P_k$, $P_k^2=P_k$ and $Z_k^TZ_k=I$, we obtain

$$\begin{aligned}
\|Z_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\|^2&=(\nabla f_k+B_k\boldsymbol{w}_k)^TZ_kZ_k^T(\nabla f_k+B_k\boldsymbol{w}_k)\\
&=(\nabla f_k+B_k\boldsymbol{w}_k)^TP_k^2(\nabla f_k+B_k\boldsymbol{w}_k)=\|\nabla_{\boldsymbol{x}}\mathcal{L}_k+P_kB_k\boldsymbol{w}_k\|^2.
\end{aligned}$$

Thus, the conclusion of Lemma 2.1 can be rewritten as

$$m(\boldsymbol{u}_k)-m(\mathbf{0})\leq-\|\nabla_{\boldsymbol{x}}\mathcal{L}_k+P_kB_k\boldsymbol{w}_k\|\tilde{\Delta}_k+\frac{1}{2}\|B_k\|\tilde{\Delta}_k^2, \tag{2.9}$$

indicating that the tangential step relates to the reduction of the optimality residual.

To end this section, we would like to link our relaxation technique with those in [10, 30] in Remark 2.2.

Remark 2.2. In our method, we define *rescaled* residuals $\|\nabla_{\boldsymbol{x}}\mathcal{L}_k^{RS}\|$, $\|c_k^{RS}\|$, $\|\nabla\mathcal{L}_k^{RS}\|$ and adaptively decompose the radius based on the proportions of these rescaled residuals (cf. (2.5)). We have two motivations: (i) the relation of the normal and tangential steps to the feasibility and optimality residuals; (ii) a scale-invariant property. We explain as follows.

Seeing from (2.8) and (2.9), the normal step relates to the reduction of the feasibility residual, while the tangential step relates to the reduction of the optimality residual. When the proportion of the feasibility residual is larger than that of the optimality residual, decreasing the feasibility residual is more important. As a result, we assign a larger trust-region radius to the normal step to achieve a larger reduction in the feasibility residual. Otherwise, we assign a larger radius to the tangential step to achieve a larger reduction in the optimality residual. In comparison, [10, 30] rely on a fixed proportion constant $\theta\in(0,1)$, making their approach less adaptive than ours.

On the other hand, we note that [10, 30] enjoy a nice scale-invariant property: given the radius Δ_k, the trial step $\Delta\boldsymbol{x}_k$ is invariant when the constraints c and/or the objective f are rescaled by a (positive) scalar. Note that if f (or c) is rescaled by a positive scalar, the Lagrangian Hessian (or the constraints Jacobian) will be rescaled by the same scalar. To preserve the invariance property, we decompose Δ_k using the rescaled residuals, as opposed to the original residuals $\|\nabla_{\boldsymbol{x}}\mathcal{L}_k\|$ and $\|c_k\|$; the latter can never be scale-invariant.

In the next section, we move to the fully stochastic setup and utilize the proposed relaxation scheme to design an StoSQP algorithm for (1.1). We will also discuss how to use the relaxation in [10] to design a StoSQP method.

3. A trust-region stochastic SQP algorithm.

From now on, we replace the deterministic gradient $\nabla f(\boldsymbol{x})$ by its stochastic estimate $\bar{g}(\boldsymbol{x}) = \nabla F(\boldsymbol{x};\xi)$. Similar to section 2, we denote $\bar{g}_k = \bar{g}(\boldsymbol{x}_k)$ and define the *estimated* KKT residual as $\|\bar{\nabla}\mathcal{L}_k\| = \|(\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k, c_k)\|$ with $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k = \bar{g}_k + G_k^T\boldsymbol{\lambda}_k$.

We summarize the proposed TR-StoSQP algorithm in Algorithm 3.1 and introduce the algorithm details as follows. In the kth iteration, we are given the iterate $\boldsymbol{x}_k$, two fixed scalars $\zeta > 0$ and $\delta \geq 0$, and the parameters $(\beta_k, L_{\nabla f,k}, L_{G,k}, \bar{\mu}_{k-1})$. Here, $\beta_k \in (0, \beta_{\max}]$ with upper bound $\beta_{\max} > 0$ being the input radius-related parameter; $L_{\nabla f,k}$ and $L_{G,k}$ are the (estimated) Lipschitz constants of $\nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x})$ (in practice, they can be estimated by standard procedures in [17, 3]); and $\bar{\mu}_{k-1}$ is the merit parameter of the ℓ_2 merit function obtained after the $(k-1)$th iteration. With these parameters, we proceed with the following three steps.

Step 1: Compute control parameters. We compute a matrix B_k to approximate the Hessian of the Lagrangian $\nabla^2_{\boldsymbol{x}}\mathcal{L}_k$ and require it to be deterministic conditioning on $\boldsymbol{x}_k$. With $\boldsymbol{v}_k$ defined in (2.3), we then compute several control parameters:

$$\begin{aligned} &\eta_{1,k} = \zeta\cdot\|\boldsymbol{v}_k\|/\|c_k\|, &&\tau_k = L_{\nabla f,k} + L_{G,k}\bar{\mu}_{k-1} + \|B_k\|,\\ &\alpha_k = \frac{\beta_k}{4(\eta_{1,k}\tau_k+\zeta)\beta_{\max}}, &&\eta_{2,k} = \eta_{1,k} - \frac{1}{2}\zeta\eta_{1,k}\alpha_k. \end{aligned} \tag{3.1}$$

We should emphasize that, compared to the existing line-search-based StoSQP methods [2, 3, 4, 25, 26, 28], we do not require B_k to be positive definite in the null space $\ker(G_k)$. This benefit adheres to the trust-region methods, more precisely the

Algorithm 3.1. A Trust Region Stochastic SQP (TR-StoSQP) Algorithm.

Input: Initial iterate $\boldsymbol{x}_0$, radius-related sequence $\{\beta_k\} \subset (0, \beta_{\max}]$, parameters $\rho > 1, \bar{\mu}_{-1}, \zeta > 0$, $\delta \geq 0$, (estimated) Lipschitz constants $\{L_{\nabla f,k}\}, \{L_{G,k}\}$.
for $k = 0, 1, \ldots,$ **do**
  Compute an approximation B_k and control parameters $\eta_{1,k}, \tau_k, \alpha_k, \eta_{2,k}$ as (3.1);
  Sample ξ_g^k and compute $\bar{g}_k$, $\bar{\boldsymbol{\lambda}}_k$, $\bar{\nabla}\mathcal{L}_k$, and the trust-region radius Δ_k as (3.2);
  Decompose Δ_k as (2.5) and compute $\bar{\gamma}_k^{\text{trial}}$ as (2.6) and $\bar{\gamma}_k$ as (3.3);
  Compute $\Delta\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{t}_k$, where $\boldsymbol{w}_k = \bar{\gamma}_k\boldsymbol{v}_k$ and $\boldsymbol{t}_k = Z_k\boldsymbol{u}_k$ is from (2.7);
  Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x}_k$, set $\bar{\mu}_k = \bar{\mu}_{k-1}$, and compute Pred_k as (3.5);
  while (3.6) does not hold **do**
    Set $\bar{\mu}_k = \rho\bar{\mu}_k$;
  end while
end for

existence of the trust-region constraint. Due to this benefit, we can construct different B_k to formulate the StoSQP subproblems. In our experiments in section 5, we will construct B_k by the identity matrix, the symmetric rank-one (SR1) update, the estimated Hessian without modification, and the average of the estimated Hessians.

The control parameters in (3.1) play a critical role in adjusting the input $\{\beta_k\}$ and generating the trust-region radius. Compared to [20], $\{\eta_{1,k}, \eta_{2,k}\}$ (i.e., $\{\gamma_{1,k}, \gamma_{2,k}\}$ in their notation) are no longer inputs and B_k is not rescaled by the parameters.

Step 2: Compute the trust-region radius. We sample a realization ξ_g^k and compute an estimate $\bar{g}_k = \nabla F(\boldsymbol{x}_k; \xi_g^k)$ of ∇f_k. We then compute the least-squares Lagrangian multiplier as $\bar{\boldsymbol{\lambda}}_k = -[G_k G_k^T]^{-1} G_k \bar{g}_k$ and the KKT vector $\bar{\nabla}\mathcal{L}_k$. Furthermore, we define the trust-region radius as

$$\Delta_k = \begin{cases} \eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\| & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in (0, 1/\eta_{1,k}), \\ \alpha_k & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in [1/\eta_{1,k}, 1/\eta_{2,k}], \\ \eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\| & \text{if } \|\bar{\nabla}\mathcal{L}_k\| \in (1/\eta_{2,k}, \infty). \end{cases} \tag{3.2}$$

We provide the following remark to compare (3.2) with the line search scheme in [3].

Remark 3.1. It is interesting to see that the scheme (3.2) enjoys the same flavor as the random-projection-based line search procedure in [3]. In particular, [3] updates $\boldsymbol{x}_k$ by $\alpha_k\tilde{\Delta}\boldsymbol{x}_k$ each step, where $\tilde{\Delta}\boldsymbol{x}_k$ is solved from problem (2.1) (without trust-region constraint) and the stepsize α_k is selected by projecting a random quantity into an interval like $[\beta_k, \beta_k + \beta_k^2]$ (see (3.3) below). By the facts that $\|\tilde{\Delta}\boldsymbol{x}_k\| = \mathcal{O}(\|\bar{\nabla}\mathcal{L}_k\|)$ (i.e., $\tilde{\Delta}\boldsymbol{x}_k$ and $\bar{\nabla}\mathcal{L}_k$ have the same order of magnitude) and $\alpha_k = \mathcal{O}(\beta_k)$, we know $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| = \|\alpha_k\tilde{\Delta}\boldsymbol{x}_k\| = \mathcal{O}(\beta_k\|\bar{\nabla}\mathcal{L}_k\|)$. This order is preserved by our trust-region scheme, since, seeing from (3.1) and (3.2), we have $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| = \|\Delta\boldsymbol{x}_k\| = \mathcal{O}(\beta_k\|\bar{\nabla}\mathcal{L}_k\|)$. Furthermore, the projection in [3] brings some sort of adaptivity to the scheme, as the stepsize α_k has a variability of $\mathcal{O}(\beta_k^2)$. This merit is also preserved by (3.2), noting that $(\eta_{1,k} - \eta_{2,k})\alpha_k = \mathcal{O}(\beta_k^2)$.

We emphasize that (3.2) offers adaptivity to selecting the radius Δ_k based on $\alpha_k (= \mathcal{O}(\beta_k))$. When $\|\bar{\nabla}\mathcal{L}_k\|$ is large, the iterate $\boldsymbol{x}_k$ is likely to be far from the KKT point. Thus, we set $\Delta_k > \alpha_k$ to be more aggressive than α_k. Otherwise, when $\|\bar{\nabla}\mathcal{L}_k\|$ is small, the iterate $\boldsymbol{x}_k$ is likely to be near the KKT point. Thus, we set $\Delta_k < \alpha_k$ to be more conservative than α_k.

Step 3: Compute the trial step and update the merit parameter. With Δ_k from Step 2, we adapt the relaxation technique in section 2.2 to compute the trial step $\Delta\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{t}_k$. In particular, we apply (2.5) to decompose Δ_k, with deterministic residuals replaced by their stochastic estimates. Then, we apply (2.6) to compute the stochastic counterpart of γ_k, denoted as $\bar{\gamma}_k^{\text{trial}}$. Then, we set $\bar{\gamma}_k$ as

$$\bar{\gamma}_k \leftarrow \text{Proj}\left(\bar{\gamma}_k^{\text{trial}} \,\middle|\, \left[0.5\zeta\phi_k\alpha_k, 0.5\zeta\phi_k\alpha_k + \delta\alpha_k^2\right]\right), \tag{3.3}$$

where $\phi_k = \min\{\|B_k\|/\|G_k\|, 1\}$ and $\text{Proj}(a|[b,c])$ is the projection function. It equals a if $a \in [b,c]$, b if $a < b$, and c if $a > c$. The normal step is $\boldsymbol{w}_k = \bar{\gamma}_k\boldsymbol{v}_k$, and the tangential step $\boldsymbol{t}_k = Z_k\boldsymbol{u}_k$ is solved from (2.7), achieving an reduction at least as much as Cauchy reduction. Finally, we update the iterate as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x}_k$, and update the merit parameter $\bar{\mu}_{k-1}$ of the ℓ_2 merit function, defined as

$$\mathcal{L}_{\bar{\mu}}(\boldsymbol{x}) = f(\boldsymbol{x}) + \bar{\mu}\|c(\boldsymbol{x})\|. \tag{3.4}$$

Specifically, we let $\bar{\mu}_k = \bar{\mu}_{k-1}$ and compute the predicted reduction of $\mathcal{L}^k_{\bar{\mu}_k}$ as

$$\text{Pred}_k = \bar{g}_k^T \Delta \boldsymbol{x}_k + \frac{1}{2}\Delta \boldsymbol{x}_k^T B_k \Delta \boldsymbol{x}_k + \bar{\mu}_k(\|c_k + G_k \Delta \boldsymbol{x}_k\| - \|c_k\|). \tag{3.5}$$

The parameter $\bar{\mu}_k$ is then iteratively updated as $\bar{\mu}_k \leftarrow \rho\bar{\mu}_k$ with some $\rho > 1$ until

$$\text{Pred}_k \leq -\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2. \tag{3.6}$$

We now explain some components of Step 3 in the following remarks.

Remark 3.2. The update rule for the merit parameter in (3.6) is well-posed and terminates in a finite number of steps. By (2.8), $\bar{\mu}_k(\|c_k + G_k\Delta\boldsymbol{x}_k\| - \|c_k\|) = -\bar{\gamma}_k\bar{\mu}_k\|c_k\|$. Thus, when $\|c_k\| \neq 0$, Pred_k decreases as $\bar{\mu}_k$ increases and (3.6) is satisfied for a sufficiently large $\bar{\mu}_k$. When $\|c_k\| = 0$, both $\boldsymbol{w}_k$ and $\breve{\Delta}_k$ vanish, and $\text{Pred}_k = m(\boldsymbol{u}_k) - m(\boldsymbol{0})$. Then, (3.6) is satisfied solely by the tangential step, without selecting the merit parameter, as can be seen from (2.9). The choice of the right-hand-side threshold of (3.6) ensures that the trial step achieves a sufficient reduction on the merit function (3.4). In particular, it is known for SQP methods that the predicted reduction of the merit function is characterized by the directional derivative of the merit function along the trial step, which is proportional to $-\|\bar{\nabla}\mathcal{L}_k\|^2$ when the merit parameter $\bar{\mu}_k$ is selected properly (see [3, 26]). This motivates the first term of the threshold. Further, to control the quadratic term $\Delta\boldsymbol{x}_k^T B_k \Delta\boldsymbol{x}_k/2$ in (3.5), we offset the threshold by the second term $\|B_k\|\Delta_k^2/2$, which stems from the positive term in Cauchy reduction (see Lemma 2.1). Overall, as shown in Lemma 4.6, the right-hand side of (3.6) is always negative, meaning that the trial steps leads to a sufficient reduction.

The iterative update $\bar{\mu}_k \leftarrow \rho\bar{\mu}_k$ is not essential, since the threshold of $\bar{\mu}_k$ can be obtained by directly solving (3.6). Then, $\bar{\mu}_k$ can be updated by taking the maximum between $\rho\bar{\mu}_k$ and the threshold. The maximum operation ensures that $\bar{\mu}_k$ is increased by at least a fixed amount, $\rho\bar{\mu}_{-1}$, whenever it is updated. This is important for the stability result of $\bar{\mu}_k$ (see Lemma 4.13).

Remark 3.3. We utilize a projection step (3.3) in the selection of $\bar{\gamma}_k$. The interval with a length of $\delta\alpha_k^2$ provides some sort of flexibility in the selection, similar to [3, 2] and references therein. The motivation behind the projection is to regulate $\bar{\gamma}_k$ using control parameters computed in (3.1). To gain insight into the interval boundary, we consider a small α_k. Combining (2.5), (2.6), and (3.2), we obtain that $\bar{\gamma}_k^{\text{trial}} = \breve{\Delta}_k/\|\boldsymbol{v}_k\| = \mathcal{O}(\Delta_k/\|\bar{\nabla}\mathcal{L}_k^{RS}\|) = \mathcal{O}(\alpha_k)$. As a result, the boundary should scale proportionally with α_k. However, $\mathcal{O}(\cdot)$ hides the ratios between unscaled and scaled residuals, such as $\|\bar{\nabla}\mathcal{L}_k\|/\|\bar{\nabla}\mathcal{L}_k^{RS}\|$. The control parameters are utilized to offer a deterministic lower bound for these ratios. In the end, we can show that (see (4.10))

$$\zeta\phi_k\alpha_k/2 \leq \min\{\breve{\Delta}_k/\|\boldsymbol{v}_k\|, 1\} =: \bar{\gamma}_k^{\text{trial}},$$

which implies $\bar{\gamma}_k \leq \bar{\gamma}_k^{\text{trial}}$ and, consequently, the normal step $\|\boldsymbol{w}_k\| = \bar{\gamma}_k\|\boldsymbol{v}_k\| \leq \breve{\Delta}_k$.

Remark 3.4. In addition to our adaptive relaxation technique, we consider two alternative relaxation approaches for designing StoSQP methods. These approaches only affect the computation of $\Delta\boldsymbol{x}_k$, while the remaining parts of the algorithm remain the same. Thus, these approaches enjoy the same global convergence analysis. The proof of the stability result of the merit parameter $\bar{\mu}_k$ may differ slightly. In this regard, the detailed analysis is provided in Appendix A for the sake of completeness. We empirically investigate the performance of the following methods in section 5.

(i) We compute the same normal step $\boldsymbol{w}_k$, but instead of using (2.7) to compute the tangential step $\boldsymbol{t}_k$, we follow the approach of [10, 30] and use (2.4). In other words, we do not decompose Δ_k as in (2.5) but define $\tilde{\Delta}_k := \sqrt{\Delta_k^2 - \|\boldsymbol{w}_k\|^2}$.

(ii) We follow the approach in [10]. In particular, we decompose Δ_k as $\breve{\Delta}_k := \theta\Delta_k$ and $\tilde{\Delta}_k := \sqrt{\Delta_k^2 - \|\boldsymbol{w}_k\|^2}$ for a prespecified constant $\theta \in (0,1]$; and we apply Algorithm 3.1 to derive the normal and tangential steps with ϕ_k in (3.3) replaced by θ.

We end this section by introducing the randomness in TR-StoSQP. We let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots$ be a filtration of σ-algebras with $\mathcal{F}_{k-1}$ generated by $\{\xi_g^j\}_{j=0}^{k-1}$; thus, $\mathcal{F}_{k-1}$ contains all the randomness before the kth iteration. Let $\mathcal{F}_{-1} = \sigma(\boldsymbol{x}_0)$ be the trivial σ-algebra for consistency. It is easy to see that for all $k \geq 0$, we have

$$\sigma(\boldsymbol{x}_k, \eta_{1,k}, \tau_k, \alpha_k, \eta_{2,k}) \subseteq \mathcal{F}_{k-1} \quad \text{and} \quad \sigma(\Delta\boldsymbol{x}_k, \bar{\boldsymbol{\lambda}}_k, \bar{\mu}_k) \subseteq \mathcal{F}_k.$$

In the next section, we conduct the global analysis of the proposed algorithm.

4. Convergence analysis.

We study the convergence of Algorithm 3.1 by measuring the decrease of the ℓ_2 merit function at each step, that is,

$$\mathcal{L}_{\bar{\mu}_k}^{k+1} - \mathcal{L}_{\bar{\mu}_k}^{k} = f_{k+1} - f_k + \bar{\mu}_k(\|c_{k+1}\| - \|c_k\|).$$

We use $\bar{\mu}_k$ to denote the merit parameter obtained after the **while** loop in line 10 of Algorithm 3.1, so that $\bar{\mu}_k$ satisfies (3.6). Following the analysis of [3, 2, 4, 18], we will first assume $\bar{\mu}_k$ stabilizes (but not necessarily at a large enough value) after a few iterations, and then we will validate the stability of $\bar{\mu}_k$ in section 4.3.

We now state the assumptions for the analysis.

Assumption 4.1. Let $\Omega \subseteq \mathbb{R}^d$ be an open convex set containing the iterates $\{\boldsymbol{x}_k\}$. The function $f(\boldsymbol{x})$ is continuously differentiable and is bounded below by $f_{\inf}$ over Ω. The gradient $\nabla f(\boldsymbol{x})$ is Lipschitz continuous over Ω with constant $L_{\nabla f} > 0$, so that the (estimated) Lipschitz constant $L_{\nabla f,k}$ at $\boldsymbol{x}_k$ satisfies $L_{\nabla f,k} \leq L_{\nabla f}$ for all $k \geq 0$. Similarly, the constraint $c(\boldsymbol{x})$ is continuously differentiable over Ω; its Jacobian $G(\boldsymbol{x})$ is Lipschitz continuous over Ω with constant $L_G > 0$; and $L_{G,k} \leq L_G$ for all $k \geq 0$. We also assume there exist positive constants $\kappa_B, \kappa_c, \kappa_{\nabla f}, \kappa_{1,G}, \kappa_{2,G} > 0$ such that

$$\|B_k\| \leq \kappa_B, \ \ \|c_k\| \leq \kappa_c, \ \ \|\nabla f_k\| \leq \kappa_{\nabla f}, \ \ \kappa_{1,G} \cdot I \preceq G_k G_k^T \preceq \kappa_{2,G} \cdot I \ \ \forall k \geq 0.$$

Assumption 4.1 is standard in the literature on both deterministic and stochastic SQP methods; see, e.g., [10, 22, 31, 3, 2, 4, 18]. In fact, when one uses a while loop to adaptively increase $L_{\nabla f,k}$ and $L_{G,k}$ to enforce the Lipschitz conditions (as done in [3, 17]), one has $L_{\nabla f,k} \leq L_{\nabla f}' := \rho L_{\nabla f}$ for a factor $\rho > 1$ (same for $L_{G,k}$; see [3, Lemma 8]). We unify the Lipschitz constant and upper bound of $L_{\nabla f,k}$ as $L_{\nabla f}$ just for simplicity. In addition, the condition $\kappa_{1,G} \cdot I \preceq G_k G_k^T \preceq \kappa_{2,G} \cdot I$ implies G_k has full row rank; thus, the least-squares dual iterate $\bar{\boldsymbol{\lambda}}_k = -[G_k G_k^T]^{-1} G_k \bar{g}_k$ is well defined.

Next, we assume the stability of $\bar{\mu}_k$. Compared to existing StoSQP literature [3, 2, 4, 18], we do not require the stabilized value to be large enough. We will revisit this assumption in section 4.3.

Assumption 4.2. There exist an (possibly random) iteration threshold $\bar{K} < \infty$ and a deterministic constant $\hat{\mu} > 0$, such that for all $k > \bar{K}$, $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$.

Since $\bar{\mu}_k$ is nondecreasing in TR-StoSQP, we have $\bar{\mu}_k \leq \hat{\mu}$ for all $k \geq 0$. The global analysis only needs to study the convergence behavior of the algorithm after $k \geq \bar{K}+1$ iterations. Next, we impose a condition on the gradient estimate.

Assumption 4.3. There exist constants $M_g \geq 1, M_{g,1} \geq 0$ such that the stochastic gradient estimate $\bar{g}_k$ satisfies $\mathbb{E}_k[\bar{g}_k] = \nabla f_k$ and $\mathbb{E}_k[\|\bar{g}_k - \nabla f_k\|^2] \leq M_g + M_{g,1}(f_k - f_{\inf})$ for all $k \geq 0$, where $\mathbb{E}_k[\cdot]$ denotes $\mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$.

We assume that the variance of the gradient estimate satisfies a growth condition. This condition is weaker than the usual bounded variance condition assumed in the StoSQP literature [2, 3, 20, 25, 26], which corresponds to $M_{g,1} = 0$. The growth condition is more realistic and was recently investigated for stochastic first-order methods on unconstrained problems [9, 15, 34, 37], while it is less explored for StoSQP methods.

4.1. Fundamental lemmas. The following result establishes the reduction of the ℓ_2 merit function achieved by the trial step.

LEMMA 4.4. *Suppose Assumptions* 4.1 *and* 4.2 *hold. For all* $k \geq \bar{K} + 1$*, we have*

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} \leq & -\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2 \\ & + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \|P_k(\nabla f_k - \bar{g}_k)\|\Delta_k + \frac{1}{2}\tau_k\Delta_k^2.\end{aligned} \tag{4.1}$$

Proof. By the definitions of $\mathcal{L}_{\bar{\mu}_{\bar{K}}}(\boldsymbol{x})$ and Pred_k in (3.4) and (3.5), we have

$$\begin{aligned}&\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - \text{Pred}_k \\ &\quad = f_{k+1} - f_k - \bar{g}_k^T\Delta\boldsymbol{x}_k - \frac{1}{2}\Delta\boldsymbol{x}_k^T B_k\Delta\boldsymbol{x}_k + \bar{\mu}_{\bar{K}}(\|c_{k+1}\| - \|c_k + G_k\Delta\boldsymbol{x}_k\|).\end{aligned}$$

By the Lipschitz continuity of $\nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x})$, we further have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - \text{Pred}_k &\leq (\nabla f_k - \bar{g}_k)^T\Delta\boldsymbol{x}_k + \frac{1}{2}(L_{\nabla f,k} + \|B_k\| + L_{G,k}\bar{\mu}_{\bar{K}})\|\Delta\boldsymbol{x}_k\|^2 \\ &\overset{(3.1)}{=} (\nabla f_k - \bar{g}_k)^T\Delta\boldsymbol{x}_k + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2 \\ &= \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + (\nabla f_k - \bar{g}_k)^T Z_k\boldsymbol{u}_k + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2 \quad (\Delta\boldsymbol{x}_k = \bar{\gamma}_k\boldsymbol{v}_k + Z_k\boldsymbol{u}_k) \\ &\leq \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \|P_k(\nabla f_k - \bar{g}_k)\|\|\boldsymbol{u}_k\| + \frac{1}{2}\tau_k\|\Delta\boldsymbol{x}_k\|^2,\end{aligned}$$

where the last inequality uses $Z_kZ_k^T = P_k$. Combining the above result with the reduction condition in (3.6), and noting that $\|\boldsymbol{u}_k\| \leq \|\Delta\boldsymbol{x}_k\| \leq \Delta_k$, we complete the proof. ∎

Now, we further analyze the right-hand side of (4.1). By taking the expectation conditional on $\boldsymbol{x}_k$, we can show that the term $\bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k$ is upper bounded by a quantity proportional to the expected error of the gradient estimate.

LEMMA 4.5. *Suppose Assumptions* 4.1 *and* 4.3 *hold. For all* $k \geq 0$*, we have*

$$\mathbb{E}_k[\bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k] \leq \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2 \cdot \mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|].$$

Proof. When $\boldsymbol{v}_k = \boldsymbol{0}$, the result holds trivially. We consider $\boldsymbol{v}_k \neq \boldsymbol{0}$. By the design of the projection in (3.3), we know

$$\gamma_{k,\min} := \frac{1}{2}\zeta\phi_k\alpha_k \leq \bar{\gamma}_k \leq \frac{1}{2}\zeta\phi_k\alpha_k + \delta\alpha_k^2 =: \gamma_{k,\max}. \tag{4.2}$$

Note that $\sigma(\gamma_{k,\min},\gamma_{k,\max})\subseteq\mathcal{F}_{k-1}$. Let E_k be the event that $(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\geq 0$, let E_k^c be its complement, and let $\mathbb{P}_k[\cdot]$ denote the probability conditional on $\mathcal{F}_{k-1}$. By the law of total expectation, one finds

$$\begin{aligned}
&\mathbb{E}_k[\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k]\\
&\quad=\mathbb{E}_k[\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\mid E_k]\mathbb{P}_k[E_k]+\mathbb{E}_k[\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\mid E_k^c]\mathbb{P}_k[E_k^c]\\
&\quad\overset{(4.2)}{\leq}\gamma_{k,\max}\mathbb{E}_k[(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\mid E_k]\mathbb{P}_k[E_k]+\gamma_{k,\min}\mathbb{E}_k[(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\mid E_k^c]\mathbb{P}_k[E_k^c]\\
&\quad=(\gamma_{k,\max}-\gamma_{k,\min})\mathbb{E}_k[(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\mid E_k]\mathbb{P}_k[E_k]\quad\text{(by Assumption 4.3)}\\
&\quad\leq(\gamma_{k,\max}-\gamma_{k,\min})\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|\|\boldsymbol{v}_k\|\mid E_k]\mathbb{P}_k[E_k]\\
&\quad\leq(\gamma_{k,\max}-\gamma_{k,\min})\|\boldsymbol{v}_k\|\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|]\\
&\quad\overset{(4.2)}{=}\delta\alpha_k^2\|\boldsymbol{v}_k\|\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|]\overset{(3.1)}{\leq}\frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|].
\end{aligned}$$

Here, the last inequality follows from $\boldsymbol{v}_k=G_k^T[G_kG_k^T]^{-1}c_k$ and Assumption 4.1. ∎

We further simplify the result of (4.1) using the trust-region scheme in (3.2).

LEMMA 4.6. *Suppose Assumptions* 4.1*,* 4.2*, and* 4.3 *hold and* $\{\beta_k\}\subseteq(0,\beta_{\max}]$. *For all* $k\geq\bar{K}+1$*, we have*

$$\begin{aligned}
\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}]\leq\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k}-\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2+\frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|]\\
+(\zeta+\eta_{1,k}\tau_k)\,\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|\nabla f_k-\bar{g}_k\|^2].
\end{aligned}$$

Proof. According to the definition in (3.2), we separate the proof into the following three cases: $\|\bar{\nabla}\mathcal{L}_k\|\in(0,1/\eta_{1,k})$, $\|\bar{\nabla}\mathcal{L}_k\|\in[1/\eta_{1,k},1/\eta_{2,k}]$, and $\|\bar{\nabla}\mathcal{L}_k\|\in(1/\eta_{2,k},\infty)$.

Case 1: $\|\bar{\nabla}\mathcal{L}_k\|\in(0,1/\eta_{1,k})$. We have $\Delta_k=\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|$, and therefore

$$\begin{aligned}
-\|\bar{\nabla}\mathcal{L}_k\|\Delta_k+\frac{1}{2}\|B_k\|\Delta_k^2&=-\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2+\frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}\mathcal{L}_k\|^2\\
&=-\left(1-\frac{1}{2}\eta_{1,k}\alpha_k\|B_k\|\right)\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2.
\end{aligned}$$

Plugging the above expression into (4.1) and applying (3.2), we have

$$\begin{aligned}
\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}-\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k}&\leq-\left(1-\frac{1}{2}\eta_{1,k}\alpha_k\|B_k\|\right)\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2+\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\\
&\quad+\eta_{1,k}\alpha_k\|P_k(\nabla f_k-\bar{g}_k)\|\|\bar{\nabla}\mathcal{L}_k\|+\frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|\bar{\nabla}\mathcal{L}_k\|^2\\
&\leq-\frac{1}{2}\left(1-\eta_{1,k}\alpha_k\|B_k\|-\eta_{1,k}\alpha_k\tau_k\right)\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2\\
&\quad+\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k+\frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k-\bar{g}_k)\|^2\ \text{(by Young's inequality)}\\
&\leq-\left(\frac{1}{2}-\eta_{1,k}\alpha_k\tau_k\right)\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2+\bar{\gamma}_k(\nabla f_k-\bar{g}_k)^T\boldsymbol{v}_k\\
&\quad+\frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k-\bar{g}_k)\|^2\quad\text{(since by (3.1), }\|B_k\|\leq\tau_k).
\end{aligned}\tag{4.3}$$

Case 2: $\|\bar{\nabla}\mathcal{L}_k\| \in [1/\eta_{1,k}, 1/\eta_{2,k}]$. We have $\Delta_k = \alpha_k$, and thus

$$\begin{aligned}-\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2 &= -\|\bar{\nabla}\mathcal{L}_k\|\alpha_k + \frac{1}{2}\|B_k\|\alpha_k^2\\ &\leq -\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}\mathcal{L}_k\|^2,\end{aligned}$$

where the inequality is due to $\eta_{1,k}\|\bar{\nabla}\mathcal{L}_k\| \geq 1 \geq \eta_{2,k}\|\bar{\nabla}\mathcal{L}_k\|$. Plugging the above expression into (4.1), using the relation $\eta_{1,k}\|\bar{\nabla}\mathcal{L}_k\| \geq 1$ again, we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\leq -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k}^2\alpha_k\|B_k\|\right)\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k\\ &\quad + \eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}\mathcal{L}_k\| + \frac{1}{2}\eta_{1,k}^2\alpha_k^2\tau_k\|\bar{\nabla}\mathcal{L}_k\|^2\\ &\leq -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k}^2\alpha_k\|B_k\| - \frac{1}{2}\eta_{1,k} - \frac{1}{2}\eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2\\ &\quad + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 \text{ (by Young's inequality)}\\ &\leq -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k\\ &\quad + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 \quad \text{(since by (3.1), } \|B_k\| \leq \tau_k).\end{aligned} \tag{4.4}$$

Case 3: $\|\bar{\nabla}\mathcal{L}_k\| \in (1/\eta_{2,k}, \infty)$. We have $\Delta_k = \eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|$, and thus

$$\begin{aligned}-\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2 &= -\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\|B_k\|\|\bar{\nabla}\mathcal{L}_k\|^2\\ &= -\left(1 - \frac{1}{2}\eta_{2,k}\alpha_k\|B_k\|\right)\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2.\end{aligned}$$

Plugging into (4.1) and applying (3.2), we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\leq -\left(1 - \frac{1}{2}\eta_{2,k}\alpha_k\|B_k\|\right)\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k\\ &\quad + \eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|\|\bar{\nabla}\mathcal{L}_k\| + \frac{1}{2}\eta_{2,k}^2\alpha_k^2\tau_k\|\bar{\nabla}\mathcal{L}_k\|^2\\ &\leq -\frac{1}{2}\left(1 - \eta_{2,k}\alpha_k\|B_k\| - \eta_{2,k}\alpha_k\tau_k\right)\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2\\ &\quad + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 \text{ (by Young's inequality)}\\ &\leq -\left(\frac{1}{2} - \eta_{2,k}\alpha_k\tau_k\right)\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2 + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k\\ &\quad + \frac{1}{2}\eta_{2,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2 \quad \text{(since by (3.1), } \|B_k\| \leq \tau_k).\end{aligned} \tag{4.5}$$

Using $\eta_{2,k} \leq \eta_{1,k}$ and taking an upper for the results of the three cases in (4.3), (4.4), and (4.5), we have

$$\begin{aligned}\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} &\leq -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\|\bar{\nabla}\mathcal{L}_k\|^2\\ &\quad + \bar{\gamma}_k(\nabla f_k - \bar{g}_k)^T\boldsymbol{v}_k + \frac{1}{2}\eta_{1,k}\alpha_k\|P_k(\nabla f_k - \bar{g}_k)\|^2.\end{aligned}$$

Taking expectation conditional on $\boldsymbol{x}_k$, applying Lemma 4.5, and noting that $\mathbb{E}_k[\|\bar{\nabla}\mathcal{L}_k\|^2] = \|\nabla\mathcal{L}_k\|^2 + \mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2]$, we have

$$\begin{aligned}
\mathbb{E}_k[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}] - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} \leq & -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\mathbb{E}_k[\|\bar{\nabla}\mathcal{L}_k\|^2] \\
& + \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] + \frac{1}{2}\eta_{1,k}\alpha_k\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
= & -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\|\nabla\mathcal{L}_k\|^2 \\
& -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
& + \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] + \frac{1}{2}\eta_{1,k}\alpha_k\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \\
= & -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& + \left(\eta_{1,k} - \eta_{2,k} + \eta_{1,k}^2\alpha_k\tau_k\right)\alpha_k\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2].
\end{aligned}$$

Furthermore, we note that

$$\begin{aligned}
\alpha_k \overset{(3.1)}{\leq} \frac{2}{8\eta_{1,k}\tau_k + 3\zeta} & \Longrightarrow 3\zeta\alpha_k + 8\eta_{1,k}\alpha_k\tau_k \leq 2 \\
& \Longrightarrow \frac{1}{2} + \eta_{1,k}\alpha_k\tau_k \leq \frac{3}{4} - \frac{3}{8}\zeta\alpha_k \\
& \Longrightarrow \frac{1}{2}\eta_{1,k} + \eta_{1,k}^2\alpha_k\tau_k \leq \frac{3}{4}\eta_{1,k}\left(1 - \frac{1}{2}\zeta\alpha_k\right) \overset{(3.1)}{=} \frac{3}{4}\eta_{2,k} \\
& \Longrightarrow -\left(\eta_{2,k} - \frac{1}{2}\eta_{1,k} - \eta_{1,k}^2\alpha_k\tau_k\right) \leq -\frac{1}{4}\eta_{2,k}.
\end{aligned}$$

Combining the above two results and using (3.1), we have

$$\begin{aligned}
\mathbb{E}_k[\mathcal{L}^{k+1}_{\bar{\mu}_{\bar{K}}}] - \mathcal{L}^k_{\bar{\mu}_{\bar{K}}} \leq & -\frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|] \\
& + (\zeta + \eta_{1,k}\tau_k)\,\eta_{1,k}\alpha_k^2\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2].
\end{aligned}$$

The conclusion follows by noting that $\mathbb{E}_k[\|P_k(\nabla f_k - \bar{g}_k)\|^2] \leq \mathbb{E}_k[\|\nabla f_k - \bar{g}_k\|^2]$. ∎

Finally, we present some properties of the control parameters generated in Step 1 of Algorithm 3.1.

LEMMA 4.7. *Let Assumptions* 4.1, 4.2 *hold, and let* $\{\beta_k\} \subseteq (0, \beta_{\max}]$. *For all* $k \geq 0$, *the following hold:*

(a) *there exist constants* $\eta_{\min}, \eta_{\max} > 0$ *such that* $\eta_{\min} \leq \eta_{2,k} \leq \eta_{1,k} \leq \eta_{\max}$;
(b) *there exists a constant* $\tau_{\max} > 0$ *such that* $\tau_k \leq \tau_{\max}$;
(c) *there exist constants* $\alpha_l, \alpha_u > 0$ *such that* $\alpha_k \in [\alpha_l\beta_k, \alpha_u\beta_k]$.

Proof. (a) By (3.1), we see that $\eta_{2,k} \leq \eta_{1,k}$. Further, by Assumption 4.1, we have

$$\begin{aligned}
& \eta_{1,k} \overset{(3.1)}{=} \zeta \cdot \|\boldsymbol{v}_k\|/\|c_k\| \leq \zeta \cdot \|G_k^T[G_kG_k^T]^{-1}\| \leq \zeta/\sqrt{\kappa_{1,G}} =: \eta_{\max}, \\
& \eta_{2,k} \overset{(3.1)}{=} \eta_{1,k}\left(1 - \frac{\zeta\alpha_k}{2}\right) \overset{(3.1)}{\geq} \eta_{1,k}\left(1 - \frac{\zeta}{2}\cdot\frac{1}{4\zeta}\right) \overset{(3.1)}{\geq} \frac{7\zeta\|\boldsymbol{v}_k\|}{8\|c_k\|} \geq \frac{7\zeta}{8\sqrt{\kappa_{2,G}}} =: \eta_{\min}.
\end{aligned}$$

(b) By Assumptions 4.1 and 4.2, we have $L_{\nabla f,k} \le L_{\nabla f}, L_{G,k} \le L_G$, $\|B_k\| \le \kappa_B$, and $\bar{\mu}_k \le \hat{\mu}$. Thus, we let $\tau_{\max} := L_{\nabla f} + L_G\hat{\mu} + \kappa_B$ and the result holds.

(c) We let $\alpha_l := 1/(4\eta_{\max}\tau_{\max}\beta_{\max} + 4\zeta\beta_{\max})$ and $\alpha_u := 1/(4\zeta\beta_{\max})$, and the result holds. ∎

In the next subsection, we use Lemmas 4.6 and 4.7 to show the global convergence of TR-StoSQP. We consider both constant and decaying β_k sequences.

4.2. Global convergence. We first consider constant β_k, i.e., $\beta_k = \beta \in (0, \beta_{\max}]$ for all $k \ge 0$. We show that the expectation of weighted averaged KKT residuals converges to a neighborhood around zero with a radius of the order $\mathcal{O}(\beta)$. When the growth condition parameter $M_{g,1} = 0$ (cf. Assumption 4.3), the weighted average reduces to the uniform average.

LEMMA 4.8. *Suppose Assumptions* 4.1, 4.2, *and* 4.3 *hold and* $\beta_k = \beta \in (0, \beta_{\max}]$ *for all* $k \ge 0$. *For any positive integer* $K > 0$, *we define* $w_k = (1 + \Upsilon M_{g,1}\beta^2)^{\bar{K}+K-k}$, $\bar{K} \le k \le \bar{K}+K$, *with* $\Upsilon := (\zeta\eta_{\max} + \eta_{\max}^2\tau_{\max} + \delta\kappa_c/\sqrt{\kappa_{1,G}})\alpha_u^2$. *We have (cf.* $\mathbb{E}_{\bar{K}+1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{\bar{K}}]$*)*

$$\mathbb{E}_{\bar{K}+1}\left[\frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla\mathcal{L}_k\|^2}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\right] \le \frac{4}{\eta_{\min}\alpha_l\beta} \cdot \frac{w_{\bar{K}}(\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf})}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k} + \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta.$$

Proof. From Lemma 4.6 and Assumption 4.3, we have for any $k \ge \bar{K}+1$,

$$\begin{aligned}\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1}] &\le \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - \frac{1}{4}\eta_{2,k}\alpha_k\|\nabla\mathcal{L}_k\|^2 + (\zeta + \eta_{1,k}\tau_k)\,\eta_{1,k}\alpha_k^2[M_g + M_{g,1}(f_k - f_{\inf})] \\ &\quad + \frac{\delta\kappa_c}{\sqrt{\kappa_{1,G}}}\alpha_k^2\sqrt{M_g + M_{g,1}(f_k - f_{\inf})} \\ &\overset{\text{Lemma 4.7}}{\le} \mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - \frac{1}{4}\eta_{\min}\alpha_l\beta\|\nabla\mathcal{L}_k\|^2 + \Upsilon\beta^2[M_g + M_{g,1}(f_k - f_{\inf})] \quad (\text{by } M_g \ge 1).\end{aligned}$$

Using the fact that $f_k - f_{\inf} \le f_k - f_{\inf} + \bar{\mu}_{\bar{K}}\|c_k\| = \mathcal{L}_{\bar{\mu}_{\bar{K}}}^k - f_{\inf}$, we obtain

$$\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - f_{\inf}] \le \left(1 + \Upsilon M_{g,1}\beta^2\right)(\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - f_{\inf}) - \frac{1}{4}\eta_{\min}\alpha_l\beta\|\nabla\mathcal{L}_k\|^2 + \Upsilon M_g\beta^2.$$

Taking the expectation conditional on $\mathcal{F}_{\bar{K}}$ and rearranging the terms, we have

$$\begin{aligned}\mathbb{E}_{\bar{K}+1}[\|\nabla\mathcal{L}_k\|^2] &\le \frac{4(1 + \Upsilon M_{g,1}\beta^2)}{\eta_{\min}\alpha_l\beta}\mathbb{E}_{\bar{K}+1}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - f_{\inf}] \\ &\quad - \frac{4}{\eta_{\min}\alpha_l\beta}\mathbb{E}_{\bar{K}+1}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - f_{\inf}] + \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta.\end{aligned}$$

Multiplying w_k on both sides and summing over $k = \bar{K}+1, \ldots, \bar{K}+K$, we have

$$\begin{aligned}\mathbb{E}_{\bar{K}+1}\left[\frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla\mathcal{L}_k\|^2}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\right] &= \frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\mathbb{E}_{\bar{K}+1}[\|\nabla\mathcal{L}_k\|^2]}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k} \\ &\le \frac{4}{\eta_{\min}\alpha_l\beta} \cdot \frac{w_{\bar{K}}(\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf}) - \mathbb{E}_{\bar{K}+1}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+K+1} - f_{\inf}]}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k} + \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta,\end{aligned}$$

where the first equality uses the fact that $\bar{K}$ is fixed in the conditional expectation. Noting that $\mathbb{E}_{\bar{K}+1}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+K+1} - f_{\inf}] \ge 0$, we complete the proof. ∎

The following theorem follows from Lemma 4.8.

THEOREM 4.9 (global convergence with constant β_k). *Suppose Assumptions* 4.1, 4.2, *and* 4.3 *hold and* $\beta_k = \beta \in (0, \beta_{\max}]$ *for all* $k \geq 0$. *Let us define* w_k *and* Υ *as in Lemma* 4.8. *We have the following:*

(a) *when* $M_{g,1} = 0$,

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{1}{K}\sum_{k=\bar{K}+1}^{\bar{K}+K} \|\nabla \mathcal{L}_k\|^2\right] \leq \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta;$$

(b) *when* $M_{g,1} > 0$,

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{1}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla \mathcal{L}_k\|^2\right] \leq \frac{4\Upsilon\{M_{g,1}\mathbb{E}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf}] + M_g\}}{\eta_{\min}\alpha_l}\beta.$$

Proof. (a) When $M_{g,1} = 0$, we have $w_k = 1$ for $\bar{K}+1 \leq k \leq \bar{K}+K$. From Lemma 4.8, we have

$$\mathbb{E}_{\bar{K}+1}\left[\frac{1}{K}\sum_{k=\bar{K}+1}^{\bar{K}+K} \|\nabla \mathcal{L}_k\|^2\right] \leq \frac{4}{\eta_{\min}\alpha_l\beta}\cdot\frac{\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf}}{K} + \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta.$$

Letting $K \to \infty$ and using the fact that $\|\nabla\mathcal{L}_k\|^2 \leq \kappa_{\nabla f}^2 + \kappa_c^2$ (cf. Assumption 4.1), we apply Fatou's lemma and have (the lim on the left can be strengthened to lim sup)

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{1}{K}\sum_{k=\bar{K}+1}^{\bar{K}+K} \|\nabla \mathcal{L}_k\|^2\right] \leq \mathbb{E}\left[\limsup_{K\to\infty}\mathbb{E}_{\bar{K}+1}\left[\frac{1}{K}\sum_{k=\bar{K}+1}^{\bar{K}+K} \|\nabla \mathcal{L}_k\|^2\right]\right] \leq \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta.$$

(b) When $M_{g,1} > 0$, we apply Lemma 4.8 and the fact that $\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k = (w_{\bar{K}} - 1)/(\Upsilon M_{g,1}\beta^2)$ and obtain

$$\mathbb{E}_{\bar{K}+1}\left[\frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla \mathcal{L}_k\|^2}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\right] \leq \frac{4\Upsilon M_{g,1}\beta}{\eta_{\min}\alpha_l}\cdot\frac{w_{\bar{K}}(\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf})}{w_{\bar{K}} - 1} + \frac{4\Upsilon M_g}{\eta_{\min}\alpha_l}\beta.$$

Since $w_{\bar{K}}/(w_{\bar{K}} - 1) = (1 + \Upsilon M_{g,1}\beta^2)^K/\{(1 + \Upsilon M_{g,1}\beta^2)^K - 1\} \to 1$ as $K \to \infty$, we apply Fatou's lemma and have (the lim on the left can be strengthened to lim sup)

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla \mathcal{L}_k\|^2}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\right] \leq \mathbb{E}\left[\limsup_{K\to\infty}\mathbb{E}_{\bar{K}+1}\left[\frac{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k\|\nabla \mathcal{L}_k\|^2}{\sum_{k=\bar{K}+1}^{\bar{K}+K} w_k}\right]\right]$$
$$\leq \frac{4\Upsilon\{M_{g,1}\mathbb{E}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{\bar{K}+1} - f_{\inf}] + M_g\}}{\eta_{\min}\alpha_l}\beta.$$

This completes the proof. ∎

From Theorem 4.9, we note that the radius of the local neighborhood is proportional to β. Thus, to decrease the radius, one should choose a smaller β. However, the trust-region radius is also proportional to β (cf. (3.2)); thus, a smaller β may result in a slow convergence. This suggests the existence of a trade-off between the convergence speed and convergence precision.

For constant $\{\beta_k\}$, [2, 3, 18, 20] established global results similar to Theorem 4.9. However, our analysis has two major differences. (i) That line of literature required β to be upper bounded by some complex quantities that may be less than 1, while we do not need such a condition. (ii) Compared to the stochastic trust-region method for unconstrained optimization [20], our local neighborhood radius is proportional to the input β (i.e., we can control the radius by the input), while the one in [20] is independent of β.

Next, we consider decaying β_k. We show in the next lemma that, when $\sum \beta_k = \infty$ and $\sum \beta_k^2 < \infty$, the infimum of KKT residuals converges to zero almost surely. Based on this result, we further show that the KKT residuals converge to zero almost surely.

LEMMA 4.10. *Suppose Assumptions* 4.1, 4.2, *and* 4.3 *hold,* $\{\beta_k\} \subseteq (0, \beta_{\max}]$, *and* $\sum_{k=0}^{\infty} \beta_k = \infty$ *and* $\sum_{k=0}^{\infty} \beta_k^2 < \infty$. *We have*

$$\liminf_{k\to\infty} \|\nabla \mathcal{L}_k\| = 0 \quad \textit{almost surely.}$$

Proof. From the proof of Lemma 4.8, we have for any $k \geq \bar{K}+1$ that

$$\mathbb{E}_k[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k+1} - f_{\inf}] \leq \left(1 + \Upsilon M_{g,1}\beta_k^2\right)\left(\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - f_{\inf}\right) - \frac{1}{4}\eta_{\min}\alpha_l\beta_k\|\nabla\mathcal{L}_k\|^2 + \Upsilon M_g \beta_k^2.$$

Since $\mathcal{L}_{\bar{\mu}}(\boldsymbol{x}) - f_{\inf}$ is bounded below by zero, $\eta_{\min}\alpha_l\beta_k\|\nabla\mathcal{L}_k\|^2 > 0$, and $\sum_{k=\bar{K}+1}^{\infty} \beta_k^2 < \infty$, it immediately follows from the Robbins–Siegmund theorem [33] that

$$\text{(4.6)} \qquad \sup_{k \geq \bar{K}+1} \mathbb{E}_{\bar{K}+1}[\mathcal{L}_{\bar{\mu}_{\bar{K}}}^{k} - f_{\inf}] := M_{\bar{K}} < \infty, \qquad \sum_{k=\bar{K}+1}^{\infty} \beta_k \mathbb{E}_{\bar{K}+1}[\|\nabla\mathcal{L}_k\|^2] < \infty.$$

The latter part suggests that $P[\sum_{k=\bar{K}+1}^{\infty}\beta_k\|\nabla\mathcal{L}_k\|^2 < \infty \mid \mathcal{F}_{\bar{K}}] = 1$. Since the result holds for any $\mathcal{F}_{\bar{K}}$, we have $P[\sum_{k=\bar{K}+1}^{\infty}\beta_k\|\nabla\mathcal{L}_k\|^2 < \infty] = 1$. Noting that $\sum_{k=\bar{K}+1}^{\infty}\beta_k = \infty$ for any run of the algorithm, we complete the proof. ∎

Finally, we establish the global convergence theorem for decaying β_k sequence.

THEOREM 4.11 (global convergence with decaying β_k). *Suppose Assumptions* 4.1, 4.2, *and* 4.3 *hold,* $\{\beta_k\} \subseteq (0, \beta_{\max}]$, *and* $\sum_{k=0}^{\infty} \beta_k = \infty$ *and* $\sum_{k=0}^{\infty} \beta_k^2 < \infty$. *We have*

$$\lim_{k\to\infty} \|\nabla \mathcal{L}_k\| = 0 \quad \textit{almost surely.}$$

Proof. For any run of the algorithm, suppose the statement does not hold; then we have $\limsup_{k\to\infty}\|\nabla\mathcal{L}_k\| \geq 2\epsilon$ for some $\epsilon > 0$. For such a run, let us define the set $\mathcal{K}_\epsilon := \{k \geq \bar{K}+1 : \|\nabla\mathcal{L}_k\| \geq \epsilon\}$. By Lemma 4.10, there exist two infinite index sets $\{m_i\}$, $\{n_i\}$ with $\bar{K} < m_i < n_i$ for all $i \geq 0$, such that

$$\text{(4.7)} \qquad \|\nabla\mathcal{L}_{m_i}\| \geq 2\epsilon, \quad \|\nabla\mathcal{L}_{n_i}\| < \epsilon, \quad \|\nabla\mathcal{L}_k\| \geq \epsilon \text{ for } k \in \{m_i+1, \dots, n_i-1\}.$$

By Assumption 4.1 and the definition $\nabla\mathcal{L}_k = (P_k\nabla f_k, c_k)$, there exists $L_{\nabla\mathcal{L}} > 0$ such that $\|\nabla\mathcal{L}_{k+1} - \nabla\mathcal{L}_k\| \leq L_{\nabla\mathcal{L}}\{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| + \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2\}$. Thus, (4.7) implies

$$\begin{aligned}
\epsilon &\leq \|\nabla\mathcal{L}_{m_i}\| - \|\nabla\mathcal{L}_{n_i}\| \leq \|\nabla\mathcal{L}_{n_i} - \nabla\mathcal{L}_{m_i}\| \leq \sum_{k=m_i}^{n_i-1}\|\nabla\mathcal{L}_{k+1} - \nabla\mathcal{L}_k\| \\
&\leq L_{\nabla\mathcal{L}}\sum_{k=m_i}^{n_i-1}\{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| + \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2\} \leq L_{\nabla\mathcal{L}}\sum_{k=m_i}^{n_i-1}(\Delta_k + \Delta_k^2) \\
&\overset{(3.2)}{\leq} L_{\nabla\mathcal{L}}\sum_{k=m_i}^{n_i-1}(\eta_{\max}\alpha_u\beta_k\|\bar{\nabla}\mathcal{L}_k\| + \eta_{\max}^2\alpha_u^2\beta_k^2\|\bar{\nabla}\mathcal{L}_k\|^2) \quad \text{(also by Lemma 4.7).}
\end{aligned}$$

Since $\|\bar{\nabla}\mathcal{L}_k\| \leq \|\nabla\mathcal{L}_k\| + \|\bar{g}_k - \nabla f_k\|$, $\|\bar{\nabla}\mathcal{L}_k\|^2 \leq 2(\|\nabla\mathcal{L}_k\|^2 + \|\bar{g}_k - \nabla f_k\|^2)$ and $\beta_k \leq \beta_{\max}$, we have

$$\epsilon \leq L_{\nabla\mathcal{L}}\eta_{\max}\alpha_u \sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\| + 2L_{\nabla\mathcal{L}}\eta_{\max}^2\alpha_u^2\beta_{\max}\sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\|^2$$
$$+ L_{\nabla\mathcal{L}}\eta_{\max}\alpha_u \sum_{k=m_i}^{n_i-1}\beta_k\|\bar{g}_k - \nabla f_k\| + 2L_{\nabla\mathcal{L}}\eta_{\max}^2\alpha_u^2\beta_{\max}\sum_{k=m_i}^{n_i-1}\beta_k\|\bar{g}_k - \nabla f_k\|^2.$$

Multiplying ϵ on both sides and using $\|\nabla\mathcal{L}_k\| \geq \epsilon$ for $k \in \{m_i, \ldots, n_i - 1\}$, we have (4.8)

$$\epsilon^2 \leq \{L_{\nabla\mathcal{L}}\eta_{\max}\alpha_u + 2\epsilon L_{\nabla\mathcal{L}}\eta_{\max}^2\alpha_u^2\beta_{\max}\}\sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\|^2$$
$$+ \left\{\epsilon L_{\nabla\mathcal{L}}\eta_{\max}\alpha_u + 2\epsilon L_{\nabla\mathcal{L}}\eta_{\max}^2\alpha_u^2\beta_{\max}\right\}\sum_{k=m_i}^{n_i-1}\beta_k\left(\|\bar{g}_k - \nabla f_k\| + \|\bar{g}_k - \nabla f_k\|^2\right).$$

For the sake of contradiction, we will show that the right-hand side of the above expression converges to zero as $i \to \infty$. By (4.6), we know that $\infty > \sum_{k=\bar{K}+1}^{\infty}\beta_k\|\nabla\mathcal{L}_k\|^2 \geq \sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\|^2$. Thus, $\sum_{k=m_i}^{n_i-1}\beta_k\|\nabla\mathcal{L}_k\|^2 \to 0$ as $i \to \infty$. For the second term, we note that

$$\sum_{i=0}^{\infty}\mathbb{E}_{\bar{K}+1}\left[\sum_{k=m_i}^{n_i-1}\beta_k(\|\bar{g}_k - \nabla f_k\| + \|\bar{g}_k - \nabla f_k\|^2)\right]$$
$$= \sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k\mathbb{E}_{\bar{K}+1}[\|\bar{g}_k - \nabla f_k\| + \|\bar{g}_k - \nabla f_k\|^2]$$
$$\leq 2\sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k(M_g + M_{g,1}\mathbb{E}_{\bar{K}+1}[f_k - f_{\inf}]) \overset{(4.6)}{\leq} 2(M_g + M_{g,1}M_{\bar{K}})\sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k.$$

By the definition of $\mathcal{K}_\epsilon$ and (4.6), we have $\sum_{i=0}^{\infty}\sum_{k=m_i}^{n_i-1}\beta_k \leq \sum_{k\in\mathcal{K}_\epsilon}\beta_k < \infty$. We apply the Borel–Cantelli lemma, integrate out the randomness of $\mathcal{F}_{\bar{K}}$, and have $\sum_{k=m_i}^{n_i-1}\beta_k(\|\bar{g}_k - \nabla f_k\| + \|\bar{g}_k - \nabla f_k\|^2) \to 0$ as $i \to \infty$ almost surely. Thus, the right-hand side of (4.8) converges to zero, which leads to the contradiction and completes the proof. ∎

Our almost sure convergence result matches the ones in [26, 25] established for stochastic line search methods in constrained optimization and matches the one in [20] established for stochastic trust-region methods in unconstrained optimization. Compared to [20] (cf. Assumption 4.4 there), we do not assume the variance of the gradient estimates decays as β_k. Such an assumption violates the flavor of fully stochastic methods, since a batch of samples is required per iteration with the batch size going to infinity. On the contrary, we assume a growth condition (cf. Assumption 4.3), which is weaker than the usual bounded variance condition. We should also mention that if one applies the result of [20, Lemma 4.5], one may be able to show almost sure convergence for decaying β_k without requiring decaying variance as in the context of [20]. However, a new concern arises—one needs to rescale the Hessian matrix at each step, which modifies the curvature information and affects the convergence speed.

4.3. Merit parameter behavior. In this subsection, we study the behavior of the merit parameter. We revisit Assumption 4.2 and show that it is satisfied, provided $\bar{g}_k$ is upper bounded and $\|B_k\|$ is bounded away from zero. The condition on $\bar{g}_k$ can be satisfied if the gradient noise has a bounded support (e.g., sampling from an empirical distribution). Such an assumption is standard to ensure a stabilized merit parameter for both deterministic and stochastic SQP methods [2, 3, 4, 6, 18, 25, 26]. We should mention that this line of literature only assumed the existence of an upper bound on the gradient noise, which can be unknown. In other words, the bound is not involved in the algorithm design. In comparison, [35] explored a bounded noise condition and incorporated the bound into the design of a trust-region algorithm. Certainly, our almost sure convergence result also differs from the one in [35], which showed the iterates visited a neighborhood of stationarity infinitely often.

Furthermore, a nonvanishing $\|B_k\|$ is a fairly mild condition, naturally satisfied by all the reasonable construction methods that one uses in SQP algorithms (e.g., set B_k as identity, estimated Hessian, averaged Hessian, or quasi-Newton update). However, a nonvanishing spectrum of B_k is technically necessary, due to our radius decomposition with the rescaled residuals (cf. (2.5)). We note that a vanishing spectrum leads to $\breve{\Delta}_k \to 0$, leading to a diminishing normal step $\boldsymbol{w}_k$ even if we have a large feasibility residual. The lower bound on $\|B_k\|$ is removable if we use original unscaled residuals to decompose the radius or use the alternative decomposition technique in Remark 3.4(ii); however, an additional tuning parameter θ to balance the feasibility and optimality residuals is introduced there. We provide the analysis in Appendix A for the sake of completeness.

Assumption 4.12. For all $k \geq 0$, (i) there is a constant $M_1 > 0$ such that $\|\bar{g}_k - \nabla f_k\| \leq M_1$; and (ii) there is a constant $\kappa_B > 0$ such that $1/\kappa_B \leq \|B_k\| \leq \kappa_B$.

LEMMA 4.13. *Suppose Assumptions* 4.1 *and* 4.12 *hold. There exist a (potentially random)* $\bar{K} < \infty$ *and a deterministic constant* $\hat{\mu}$, *such that* $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$, *for all* $k > \bar{K}$.

Proof. It suffices to show that there exists a deterministic threshold $\tilde{\mu} > 0$ independent of k such that (3.6) is satisfied as long as $\bar{\mu}_k \geq \tilde{\mu}$. We have

$$
\begin{aligned}
\text{Pred}_k &\overset{(3.5)}{=} \bar{g}_k^T \Delta\boldsymbol{x}_k + \frac{1}{2}\Delta\boldsymbol{x}_k^T B_k \Delta\boldsymbol{x}_k + \bar{\mu}_k(\|c_k + G_k\Delta\boldsymbol{x}_k\| - \|c_k\|) \\
&\overset{(2.8)}{=} \bar{g}_k^T Z_k\boldsymbol{u}_k + \bar{\gamma}_k(\bar{g}_k - \nabla f_k)^T\boldsymbol{v}_k + \bar{\gamma}_k\nabla f_k^T\boldsymbol{v}_k + \frac{1}{2}\boldsymbol{u}_k^T Z_k^T B_k Z_k\boldsymbol{u}_k + \bar{\gamma}_k\boldsymbol{v}_k^T B_k Z_k\boldsymbol{u}_k \\
&\quad + \frac{1}{2}\bar{\gamma}_k^2\boldsymbol{v}_k^T B_k\boldsymbol{v}_k - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \quad (\text{also use } \Delta\boldsymbol{x}_k = \bar{\gamma}_k\boldsymbol{v}_k + Z_k\boldsymbol{u}_k) \\
&\leq (\bar{g}_k + \bar{\gamma}_k B_k\boldsymbol{v}_k)^T Z_k\boldsymbol{u}_k + \frac{1}{2}\boldsymbol{u}_k^T Z_k^T B_k Z_k\boldsymbol{u}_k + \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| \\
&\quad + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \quad (\text{by Assumptions 4.1, 4.12 and } \bar{\gamma}_k \leq 1).
\end{aligned}
$$

From (2.9), and replacing $\nabla\mathcal{L}_k$ by its stochastic estimate, we have

$$
\begin{aligned}
\text{Pred}_k &\leq -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k + \bar{\gamma}_k P_k B_k\boldsymbol{v}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| \\
&\quad + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \\
&\leq -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| \\
&\quad + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \quad (\text{by triangular inequality and } \|P_k\| \leq 1) \\
&\leq -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\Delta_k + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 \\
&\quad + \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \quad (\text{since } \tilde{\Delta}_k \geq \Delta_k - \breve{\Delta}_k) \\
&= -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\Delta_k - \|c_k\|\Delta_k + \|c_k\|\Delta_k + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k \\
&\quad + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\|
\end{aligned}
$$

$$\leq -\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2 + \|c_k\|\Delta_k + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\Delta_k$$
$$+ \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\|,$$

since $\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \|c_k\| \geq \|\bar{\nabla}\mathcal{L}_k\|$ and $\tilde{\Delta}_k \leq \Delta_k$. Thus, (3.6) holds as long as

$$\bar{\mu}_k\bar{\gamma}_k\|c_k\| \geq \|c_k\|\Delta_k + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\Delta_k$$
$$+ \bar{\gamma}_k(M_1 + \kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{\bar{\gamma}_k}{2}\|B_k\|\|\boldsymbol{v}_k\|^2.$$

Since $\|\boldsymbol{v}_k\| \leq \|c_k\|/\sqrt{\kappa_{1,G}}$ and $\Delta_k \leq \Delta_{\max} := \eta_{\max}\alpha_u\beta_{\max}(\kappa_c + M_1 + \kappa_{\nabla f})$ (cf. Assumption 4.1 and Lemma 4.7), it is sufficient to show

$$\bar{\mu}_k\bar{\gamma}_k\|c_k\| \geq \|c_k\|\Delta_k + \|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k + \bar{\gamma}_k\|c_k\|\left(\frac{\kappa_B\Delta_{\max} + M_1 + \kappa_{\nabla f}}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right). \tag{4.9}$$

Equivalently,

$$\bar{\mu}_k \geq \frac{\Delta_k}{\bar{\gamma}_k} + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k}{\bar{\gamma}_k\|c_k\|} + \left(\frac{\kappa_B\Delta_{\max} + M_1 + \kappa_{\nabla f}}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

We only consider $\|c_k\| > 0$, since (4.9) holds when $\|c_k\| = 0$. By (3.2), we find that

$$\frac{\Delta_k}{\bar{\gamma}_k} + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k}{\bar{\gamma}_k\|c_k\|} \leq \frac{\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}\left(1 + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|G_k\|^{-1}}{\|\bar{\nabla}\mathcal{L}_k^{RS}\|}\right).$$

Noticing that $\|\bar{\nabla}\mathcal{L}_k^{RS}\| \geq \min\{\|B_k\|^{-1}, \|G_k\|^{-1}\}\|\bar{\nabla}\mathcal{L}_k\|$, we find

$$\frac{\Delta_k}{\bar{\gamma}_k} + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k}{\bar{\gamma}_k\|c_k\|} \leq \frac{\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}\left(1 + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\|G_k\|^{-1}}{\min\{\|B_k\|^{-1}, \|G_k\|^{-1}\}\|\bar{\nabla}\mathcal{L}_k\|}\right)$$
$$= \frac{\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}\left(1 + \max\left\{\frac{\|B_k\|}{\|G_k\|}, 1\right\}\frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|}{\|\bar{\nabla}\mathcal{L}_k\|}\right)$$
$$\leq \frac{2\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}\max\left\{\frac{\|B_k\|}{\|G_k\|}, 1\right\}.$$

To analyze $\bar{\gamma}_k$, we notice that $\|\bar{\nabla}\mathcal{L}_k^{RS}\| \leq \max\{\|B_k\|^{-1}, \|G_k\|^{-1}\}\|\bar{\nabla}\mathcal{L}_k\|$. Therefore,

$$\frac{\breve{\Delta}_k}{\|\boldsymbol{v}_k\|} \overset{(2.5)}{=} \frac{\|c_k^{RS}\|\Delta_k}{\|\bar{\nabla}\mathcal{L}_k^{RS}\|\|\boldsymbol{v}_k\|} \overset{(3.2)}{\geq} \frac{\eta_{2,k}\alpha_k\|G_k\|^{-1}\|c_k\|}{\max\{\|B_k\|^{-1}, \|G_k\|^{-1}\}\|\boldsymbol{v}_k\|}$$
$$\geq \frac{\eta_{1,k}\alpha_k\|c_k\|}{2\|\boldsymbol{v}_k\|}\min\left\{\frac{\|B_k\|}{\|G_k\|}, 1\right\} \overset{(3.1)}{=} \zeta\alpha_k\phi_k/2,$$

where the last inequality is due to the fact that (3.1) implies $\zeta\alpha_k \leq 1$, implying $\eta_{2,k} \geq \eta_{1,k}/2$. We therefore have

$$\frac{1}{2}\zeta\alpha_k\phi_k \leq \min\left\{\breve{\Delta}_k/\|\boldsymbol{v}_k\|, 1\right\} =: \bar{\gamma}_k^{\text{trial}}. \tag{4.10}$$

The above display suggests that we only need to consider $\bar{\gamma}_k = \frac{1}{2}\zeta\alpha_k\phi_k$. Noting that $\max\{\|B_k\|/\|G_k\|, 1\} \leq \max\{\kappa_B/\sqrt{\kappa_{1,G}}, 1\}$, $\min\{\|B_k\|/\|G_k\|, 1\} \geq \min\{1/(\kappa_B\sqrt{\kappa_{2,G}}), 1\}$, and $\|\bar{\nabla}\mathcal{L}_k\| \leq \kappa_c + M_1 + \kappa_{\nabla f}$, we obtain that

$$\frac{\Delta_k}{\bar{\gamma}_k} + \frac{\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\breve{\Delta}_k}{\bar{\gamma}_k\|c_k\|} \leq \left[\frac{4\eta_{\max}}{\zeta}(\kappa_c + \kappa_{\nabla f} + M_1)\max\left\{\frac{\kappa_B}{\sqrt{\kappa_{1,G}}}, 1\right\}\right] \cdot \max\{\kappa_B\sqrt{\kappa_{2,G}}, 1\}.$$

Therefore, (3.6) holds as long as

$$\bar{\mu}_k \geq \tilde{\mu} := \left[\frac{4\eta_{\max}}{\zeta}(\kappa_c + \kappa_{\nabla f} + M_1)\max\left\{\frac{\kappa_B}{\sqrt{\kappa_{1,G}}}, 1\right\}\right] \cdot \max\{\kappa_B\sqrt{\kappa_{2,G}}, 1\} + \left(\frac{\kappa_B\Delta_{\max} + M_1 + \kappa_{\nabla f}}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

Since $\bar{\mu}_k$ is increased by at least a factor of ρ for each update, we define $\hat{\mu} := \rho\tilde{\mu}$ and complete the proof. ∎

Compared to existing StoSQP methods, we do not require the stabilized merit parameter to be large enough. The additional requirement of having a large enough stabilized value is critical for existing StoSQP methods. To satisfy this requirement, [26, 25] imposed an adaptive condition on the feasibility error to be satisfied when selecting the merit parameter; and [2, 3, 4, 18] imposed a symmetry condition on the noise distribution. Intuitively, the reduction of the merit function in StoSQP methods should be related to the true KKT residual. In the aforementioned methods, the reduction of the stochastic merit function model is first related to the reduction of the deterministic merit function model and then related to the true KKT residual. However, the relation between the reduction in stochastic and deterministic models is only valid when the merit parameter stabilizes at a sufficiently large value [3, Lemma 3.12]. In contrast, our approach relates the reduction of stochastic model to the squared *estimated* KKT residual $\|\bar{\nabla}\mathcal{L}_k\|^2$ (i.e., (3.6)). After taking the conditional expectation and carefully analyzing the error terms, we can further use the true KKT residual to characterize the improvement of the merit function in each step. In the end, we suppress the condition on a sufficiently large merit parameter.

5. Numerical experiments. We demonstrate the empirical performance of Algorithm 3.1 and compare it to the line-search ℓ_1-StoSQP method designed in [3, Algorithm 3] under the same fully stochastic setup. We describe the algorithmic settings in section 5.1; then we show numerical results on a subset of CUTEst problems [23] in section 5.2; and then we show numerical results on constrained logistic regression problems in section 5.3. The implementation of TR-StoSQP is available at https://github.com/ychenfang/TR-StoSQP.

5.1. Algorithm setups. For both our method and ℓ_1-StoSQP, we try two constant sequences, $\beta_k \in \{0.5, 1\}$, and two decaying sequences, $\beta_k \in \{k^{-0.6}, k^{-0.8}\}$. The sequence $\{\beta_k\}$ is used to select the stepsize in ℓ_1-StoSQP. We use the same input, since, as discussed in Remark 3.1, β_k in two methods shares the same order. For both methods, the Lipschitz constants of the objective gradients and constraint Jacobians are estimated around the initialization and kept constant for subsequent iterations.

We follow [3] to set up the ℓ_1-StoSQP method, where we set $B_k = I$ and solve the SQP subproblems exactly. We set the parameters of TR-StoSQP as $\zeta = 10$, $\delta = 10$, $\bar{\mu}_{-1} = 1$, and $\rho = 1.5$. We use `the IPOPT` solver [38] to solve (2.7) and apply four different Hessian approximations B_k as follows:

(a) Identity (Id). We set $B_k = I$, which is widely used in the literature [2, 3, 25, 26].

(b) Symmetric rank-one (SR1) update. We set $H_{-1} = H_0 = I$ and update H_k as

$$H_k = H_{k-1} + \frac{(\boldsymbol{y}_{k-1} - H_{k-1}\Delta\boldsymbol{x}_{k-1})(\boldsymbol{y}_{k-1} - H_{k-1}\Delta\boldsymbol{x}_{k-1})^T}{(\boldsymbol{y}_{k-1} - H_{k-1}\Delta\boldsymbol{x}_{k-1})^T\Delta\boldsymbol{x}_{k-1}} \quad \forall k \geq 1,$$

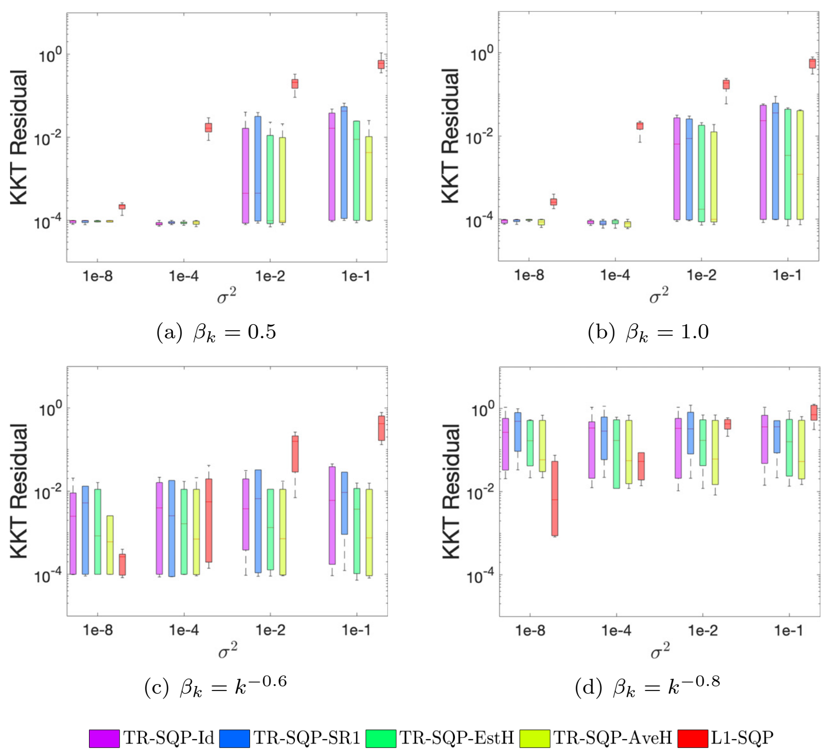
where $\boldsymbol{y}_{k-1} = \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k - \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_{k-1}$ and $\Delta\boldsymbol{x}_{k-1} = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$. Since H_k depends on $\bar{g}_k$, we set $B_k = H_{k-1}$ ($B_0 = H_{-1} = I$) to ensure that $\sigma(B_k) \subseteq \mathcal{F}_{k-1}$.

(c) Estimated Hessian (EstH). We set $B_0 = I$ and $B_k = \bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_{k-1}$ for all $k \geq 1$, where $\bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_{k-1}$ is estimated using the same sample used to estimate $\bar{g}_{k-1}$.

(d) Averaged Hessian (AveH). We set $B_0 = I$, set $B_k = \sum_{i=k-100}^{k-1} \bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i/100$ for $k \geq 100$, and set $B_k = \sum_{i=0}^{k-1} \bar{\nabla}^2_{\boldsymbol{x}}\mathcal{L}_i/k$ for $0 < k < 100$. This Hessian approximation is inspired by [27], where the authors showed that the Hessian averaging is helpful for denoising the noise in the stochastic Hessian estimates.

5.2. CUTEst. We select problems from the CUTEst set that have a nonconstant objective with only equality constraints, satisfy $d < 1000$, and do not report singularity on $G_k G_k^T$ during the iteration process, resulting in 47 problems in total. The initial iterate is provided by the CUTEst package. At each step, the estimate $\bar{g}_k$ is drawn from $\mathcal{N}(\nabla f_k, \sigma^2(I + \mathbf{1}\mathbf{1}^T))$, where $\mathbf{1}$ denotes the d-dimensional all one vector and σ^2 denotes the noise level varying within $\{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$. When the approximation EstH or AveH is used, the estimate $(\bar{\nabla}^2 f_k)_{i,j}$ (same for the (j,i) entry) is drawn from $\mathcal{N}((\nabla^2 f_k)_{i,j}, \sigma^2)$ with the same σ^2 used for estimating the gradient. We set the iteration budget to 10^5 and, for each setup of β_k and σ^2, average the KKT residuals over five runs. We stop the iteration of both methods if $\|\nabla\mathcal{L}_k\| \leq 10^{-4}$ or $k \geq 10^5$.

We report the KKT residuals of ℓ_1-StoSQP and TR-StoSQP with different Hessian approximations in Figure 1. We observe that for both constant β_k and decaying β_k with a high noise level, TR-StoSQP consistently outperforms ℓ_1-StoSQP. We note that ℓ_1-StoSQP performs better than TR-StoSQP for decaying β_k with a low noise



(a) $\beta_k = 0.5$ (b) $\beta_k = 1.0$

(c) $\beta_k = k^{-0.6}$ (d) $\beta_k = k^{-0.8}$

FIG. 1. *KKT residual boxplots for CUTEst problems. For each σ^2, there are five boxes. The first four boxes correspond to the proposed TR-StoSQP method with four different choices of B_k, while the last box corresponds to the ℓ_1-StoSQP method.*

level (e.g., $\sigma^2 = 10^{-8}$). However, in that case, TR-StoSQP is not sensitive to the noise level σ^2, while the performance of ℓ_1-StoSQP deteriorates rapidly as σ^2 increases. We think that the robustness against noise is a benefit brought by the trust-region constraint, which properly regularizes the SQP subproblem when σ^2 is large. Furthermore, among the four choices of Hessian approximations, TR-StoSQP generally performs the best with the averaged Hessian and the second best with the estimated Hessian. Compared to the identity and SR1 update, the estimated Hessian provides a better approximation to the true Hessian (especially when σ^2 is small); the averaged Hessian further reduces the noise that leads to a better performance (especially when σ^2 is large).

We observe that when σ^2 is large, or σ^2 is small but β_k is constant, TR-StoSQP outperforms ℓ_1-StoSQP, even when the identity Hessian is used. However, for decaying β_k and small σ^2, the performance of TR-StoSQP is less competitive. This disparity in performance could arise from the difference in trial step computation. In line-search methods, even though the search direction is determined by solving a Newton system, it can still be decomposed orthogonally into a normal direction $\boldsymbol{w}_k \in \mathrm{im}(G_k^T)$ and a tangential direction $\boldsymbol{t}_k \in \ker(G_k)$ (see [3] for details). The direction of $\boldsymbol{w}_k$ is consistent between trust-region and line-search methods, represented as $\boldsymbol{v}_k := -G_k^T[G_kG_k^T]^{-1}c_k$. However, the directions of the tangential step are different. In trust-region methods, the tangential step is determined by (2.7) using $\boldsymbol{w}_k = \bar{\gamma}_k\boldsymbol{v}_k$ with $\bar{\gamma}_k$ chosen based on (2.6) and (3.3). In contrast, in line-search methods, the tangential direction effectively comes from (2.7) using $\boldsymbol{w}_k = \boldsymbol{v}_k$ without the trust-region constraint. In stochastic optimization, most iterations satisfy $\bar{\gamma}_k < 1$. Therefore, the directions of the tangential step might differ in trust-region methods and line-search methods, even if the identity Hessians are used and the iterates are near an optimal point. Also, the trust-region constraint serves as a regularization that is potentially helpful for large noise scenarios (σ^2 is large or β_k is constant). We should emphasize that the difference in trial step direction is due to different mechanisms of trust-region methods and line-search methods (trust-region methods compute the search direction and stepsize simultaneously, while line-search methods compute them separately) and the fully stochastic setup (the noise does not gradually vanish), but not due to our algorithm design.

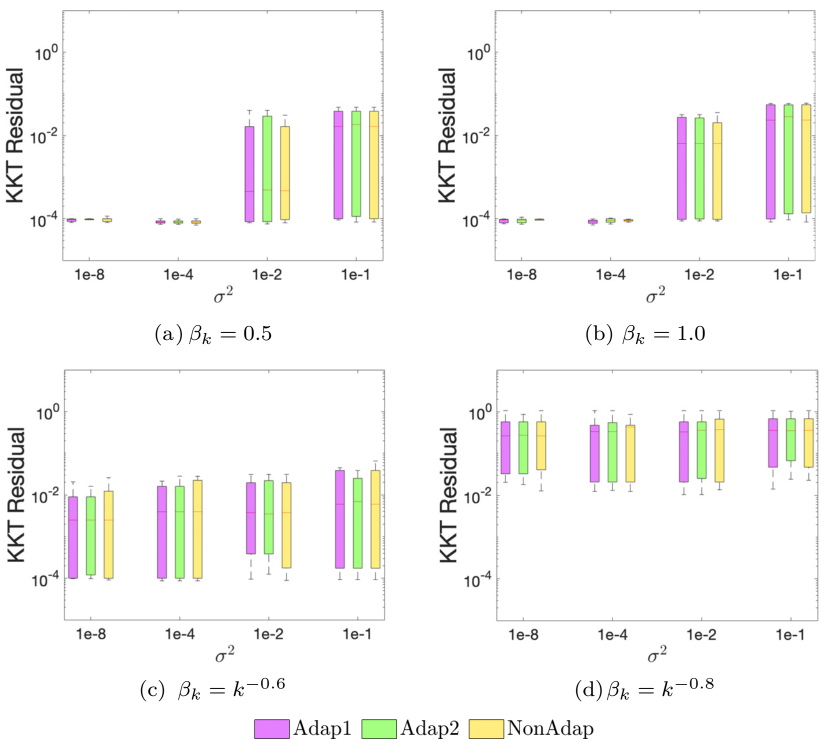
We then investigate the adaptivity of the radius selection scheme in (3.2). As explained in Remark 3.1, the radius Δ_k can be set larger or smaller than $\alpha_k = \mathcal{O}(\beta_k)$, depending on the magnitude of the estimated KKT residual. In Table 1, we report the proportions of the three cases in (3.2): $\Delta_k < \alpha_k$, $\Delta_k = \alpha_k$, and $\Delta_k > \alpha_k$. We average the proportions over five runs of all 47 problems in each setup. From Table 1, we have the following three observations. (i) Case 2 has a near zero proportion for all setups. This phenomenon is due to the fact that $\eta_{1,k} - \eta_{2,k} = \mathcal{O}(\beta_k)$. For constant β_k, this value is small, and thus a few iterations are in Case 2. For decaying β_k, this value even converges to zero, and thus almost no iterations are in Case 2. (ii) Case 3 is triggered quite frequently if β_k decays rapidly. This phenomenon suggests that the adaptive scheme can generate aggressive steps, even if we input a conservative radius-related sequence β_k. (iii) The proportion of Case 1 dominates the other two cases in most of the setups. This is reasonable, since Case 1 is always triggered when the iterates are near a KKT point.

In Remark 3.4, we provide two alternative relaxation techniques to compute the trial step. Figure 2 reports the KKT residuals for these methods. We use `Adap1` to denote TR-StoSQP with our adaptive relaxation technique; `Adap2` to denote TR-StoSQP with the technique in Remark 3.4(i), where the radius of the tangential

TABLE 1

Proportions of the three cases in (3.2) *(%). We highlight the proportion of Case* 3 *if the value is higher than* 25%.

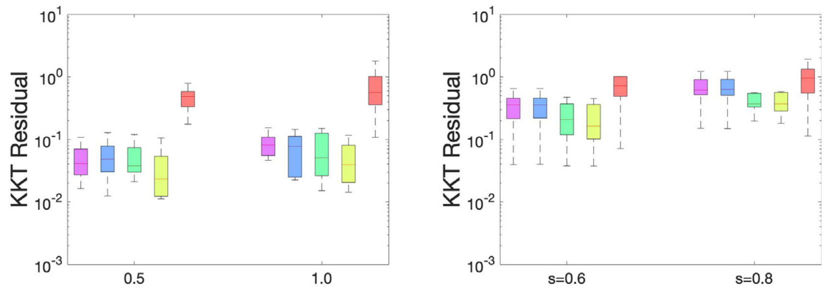
		$\sigma^2 = 10^{-8}$			$\sigma^2 = 10^{-4}$			$\sigma^2 = 10^{-2}$			$\sigma^2 = 10^{-1}$		
β_k	B_k	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
0.5	Id	90.3	0.1	9.6	91.3	0.2	8.5	95.0	0.1	4.9	54.7	0.9	**44.4**
	SR1	93.8	0.1	6.1	92.7	0.1	7.2	94.6	0.1	5.7	56.2	1.1	**42.7**
	EstH	92.2	0.1	7.7	94.8	0.1	5.1	84.8	0.2	15.0	71.1	0.5	**28.4**
	AveH	92.5	0.1	7.4	94.1	0.1	5.8	88.2	0.2	11.6	64.2	0.4	**35.4**
1.0	Id	92.0	0.1	7.9	93.7	0.1	6.2	95.4	0.2	4.4	57.1	1.2	**41.7**
	SR1	94.0	0.2	5.8	96.1	0.1	3.8	97.7	0.2	2.1	64.2	1.2	**34.6**
	EstH	92.4	0.1	7.5	93.8	0.1	6.1	87.5	0.4	12.1	72.8	0.5	**26.7**
	AveH	92.4	0.2	7.4	93.9	0.3	5.8	85.5	0.3	14.2	67.1	0.6	**32.3**
$k^{-0.6}$	Id	97.2	0.0	2.8	96.8	0.0	3.2	93.4	0.0	6.6	51.8	0.0	**48.2**
	SR1	98.3	0.0	1.7	97.1	0.0	2.9	93.2	0.0	6.8	51.5	0.0	**48.5**
	EstH	97.9	0.0	2.1	95.8	0.0	4.2	86.6	0.0	13.4	69.1	0.0	**30.9**
	AveH	97.4	0.0	2.6	96.1	0.0	3.9	86.8	0.0	13.2	65.5	0.0	**34.8**
$k^{-0.8}$	Id	70.6	0.0	**29.4**	68.1	0.0	**31.9**	66.4	0.0	**33.6**	45.8	0.0	**54.2**
	SR1	56.1	0.0	**43.9**	65.7	0.0	**34.3**	66.6	0.0	**33.4**	39.9	0.0	**60.1**
	EstH	67.5	0.0	**32.5**	65.2	0.0	**34.8**	62.0	0.0	**38.0**	54.7	0.0	**45.3**
	AveH	67.9	0.0	**32.1**	66.7	0.0	**33.3**	65.9	0.0	**34.1**	51.4	0.0	**48.6**



(a) $\beta_k = 0.5$ (b) $\beta_k = 1.0$

(c) $\beta_k = k^{-0.6}$ (d) $\beta_k = k^{-0.8}$

FIG. 2. *KKT residual boxplots for CUTEst problems with different relaxation techniques. The Hessian approximation B_k is set as identity matrix. For each σ^2, there are three boxes. The first box corresponds to the proposed adaptive relaxation technique. The second box corresponds to the adaptive technique in Remark* 3.4(i)*. The last box corresponds to the nonadaptive technique in Remark* 3.4(ii).



(a) Constant β_k (b) Decaying $\beta_k = k^{-s}$

TR-SQP-Id TR-SQP-SR1 TR-SQP-EstH TR-SQP-AveH L1-SQP

Fig. 3. *KKT residual boxplots for constrained logistic regression problems. For each setup of β_k, there are five boxes. The first four boxes correspond to the proposed TR-StoSQP method with four different choices of B_k, while the last box corresponds to the ℓ_1-StoSQP method.*

step is controlled by $\tilde{\Delta}_k := \sqrt{\Delta_k^2 - \|\boldsymbol{w}_k\|^2}$; and `NonAdap` to denote TR-StoSQP with the technique in Remark 3.4(ii), where the prespecified parameter is set as $\theta = 0.8$. The remaining algorithm setups follow from TR-StoSQP and $B_k = I$. We observe that the three techniques have comparable performance for most combinations of β_k and σ^2, while `Adap1` is slightly better than the other two techniques in some cases. The results suggest that our adaptive relaxation technique, as well as its variant in Remark 3.4(i), is at least as good as the conventional technique (the nonadaptive technique in Remark 3.4(ii)) in practice, but it requires no effort in tuning parameters.

5.3. Constrained logistic regression. We consider equality-constrained logistic regression of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \log\left\{1 + \exp\left(-y_i \cdot \langle \boldsymbol{z}_i, \boldsymbol{x}\rangle\right)\right\} \qquad \text{s.t. } A\boldsymbol{x} = \boldsymbol{b},$$

where $\boldsymbol{z}_i \in \mathbb{R}^d$ is the sample point, $y_i \in \{-1, 1\}$ is the label, and $A \in \mathbb{R}^{m\times d}$ and $\boldsymbol{b} \in \mathbb{R}^m$ form the deterministic constraints. We implement eight datasets from LIBSVM [12]: `austrilian`, `breast-cancer`, `diabetes`, `heart`, `ionosphere`, `sonar`, `splice`, and `svmguide3`. For each dataset, we set $m = 5$ and generate random A and $\boldsymbol{b}$ by drawing each element from a standard normal distribution. We ensure that A has full row rank in all problems. For both algorithms and all problems, the initial iterate is set to be all one vector of appropriate dimension. In each iteration, we select one sample at random to estimate the objective gradient (and Hessian if EstH or AveH is used). A budget of 20 epochs—the number of passes over the dataset—is used for both algorithms and all problems. We stop the iteration if $\|\nabla\mathcal{L}_k\| \le 10^{-4}$ or the epoch budget is consumed.

We report the average of the KKT residuals over five runs in Figure 3. From the figure, we observe that TR-StoSQP with all four choices of B_k consistently outperforms ℓ_1-StoSQP when $\beta_k = 0.5$, 1.0, and $k^{-0.6}$. When $\beta_k = k^{-0.8}$, TR-StoSQP enjoys a better performance by using the estimated Hessian or averaged Hessian. This experiment further illustrates the promising performance of our method.

6. Conclusion. We designed a trust-region stochastic SQP (TR-StoSQP) algorithm to solve nonlinear optimization problems with stochastic objective and deterministic equality constraints. We developed an adaptive relaxation technique to address the infeasibility issue that arises when trust-region methods are applied to constrained problems. With a stabilized merit parameter, TR-StoSQP converges in two regimes. (i) When $\beta_k = \beta$ for all $k \geq 0$, the expectation of weighted averaged KKT residuals converges to a neighborhood around zero. (ii) When β_k satisfies $\sum \beta_k = \infty$ and $\sum \beta_k^2 < \infty$, the KKT residuals converge to zero almost surely. We also showed that the merit parameter is ensured to stabilize, provided the gradient estimates are bounded. Our numerical experiments on a subset of problems of the CUTEst set and constrained logistic regression problems showed promising performance of the proposed method.

There are still several interesting future directions. First, it is of interest to design trust-region StoSQP algorithms when the Jacobians of constraints are rank-deficient. Second, how to establish global convergence without the assumption of bounded noise remains an open question. Removing that assumption may require a deeper understanding of the merit function and randomness in estimation. Finally, it is of interest to devise a method that uses second-order information efficiently. To fully exploit second-order derivatives, the method should move the trial steps along the negative curvature appropriately.

Appendix A. Additional analysis of the behavior of the merit parameter. In this appendix, we further investigate the stability behavior of the merit parameter when using the alternative two approaches in Remark 3.4 to decompose the radius. As mentioned, for both approaches, the global convergence analysis directly follows from section 4.2.

We first show that for the method in Remark 3.4(i), the merit parameter will stabilize under Assumption 4.12.

LEMMA A.1. *Suppose Assumptions* 4.1 *and* 4.12 *hold and the relaxation technique in Remark* 3.4(i) *is employed. Then, there exist a (potentially random)* $\bar{K} < \infty$ *and a deterministic constant* $\hat{\mu}$, *such that* $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$ *for all* $k > \bar{K}$.

Proof. Similar to Lemma 4.13, we only show that there exists a deterministic threshold $\tilde{\mu} > 0$ independent of k such that (3.6) is satisfied as long as $\bar{\mu}_k \geq \tilde{\mu}$. Using the same derivation as Lemma 4.13, we have

$$
\begin{aligned}
\mathrm{Pred}_k \leq & -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\tilde{\Delta}_k + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \bar{\gamma}_k(M_1+\kappa_{\nabla f})\|\boldsymbol{v}_k\| \\
& + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \\
\leq & -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\Delta_k + \bar{\gamma}_k\|\boldsymbol{v}_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 \\
& + \bar{\gamma}_k(M_1+\kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 \\
& - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \quad (\text{since } \tilde{\Delta}_k \geq \Delta_k - \bar{\gamma}_k\|\boldsymbol{v}_k\|) \\
= & -\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\|\Delta_k - \|c_k\|\Delta_k + \|c_k\|\Delta_k + \bar{\gamma}_k\|\boldsymbol{v}_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\tilde{\Delta}_k \\
& + \frac{1}{2}\|B_k\|\tilde{\Delta}_k^2 + \bar{\gamma}_k(M_1+\kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\| \\
\leq & -\|\bar{\nabla}\mathcal{L}_k\|\Delta_k + \frac{1}{2}\|B_k\|\Delta_k^2 + \|c_k\|\Delta_k + \bar{\gamma}_k\|\boldsymbol{v}_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\Delta_k \\
& + \bar{\gamma}_k(M_1+\kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{1}{2}\bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|^2 - \bar{\mu}_k\bar{\gamma}_k\|c_k\|,
\end{aligned}
$$

since $\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \|c_k\| \geq \|\bar{\nabla}\mathcal{L}_k\|$ and $\tilde{\Delta}_k \leq \Delta_k$. Thus, (3.6) holds as long as

$$\bar{\mu}_k\bar{\gamma}_k\|c_k\| \geq \|c_k\|\Delta_k + \bar{\gamma}_k\|\boldsymbol{v}_k\|\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| + \bar{\gamma}_k\|B_k\|\|\boldsymbol{v}_k\|\Delta_k + \bar{\gamma}_k(M_1+\kappa_{\nabla f})\|\boldsymbol{v}_k\| + \frac{\bar{\gamma}_k}{2}\|B_k\|\|\boldsymbol{v}_k\|^2.$$

Since $\|\boldsymbol{v}_k\| \leq \|c_k\|/\sqrt{\kappa_{1,G}}$, $\|\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_k\| \leq \|\nabla_{\boldsymbol{x}}\mathcal{L}_k\| + \|\nabla f_k - \bar{g}_k\| \leq \kappa_{\nabla f} + M_1$, and $\Delta_k \leq \Delta_{\max}$, it is sufficient to show

$$\bar{\mu}_k\bar{\gamma}_k\|c_k\| \geq \|c_k\|\Delta_k + \bar{\gamma}_k\|c_k\|\left(\frac{\kappa_B\Delta_{\max} + 2(M_1+\kappa_{\nabla f})}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

Equivalently,

$$\bar{\mu}_k \geq \frac{\Delta_k}{\bar{\gamma}_k} + \left(\frac{\kappa_B\Delta_{\max} + 2(M_1+\kappa_{\nabla f})}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

Here, we only consider $\|c_k\| \neq 0$, since the result trivially holds when $\|c_k\| = 0$. From (3.2), we find that

$$\frac{\Delta_k}{\bar{\gamma}_k} \leq \frac{\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}.$$

By (4.10), $\bar{\gamma}_k \geq \frac{1}{2}\zeta\phi_k\alpha_k = \frac{1}{2}\zeta\min\{\|B_k\|/\|G_k\|, 1\}\alpha_k$. Noting that $\min\{\|B_k\|/\|G_k\|, 1\} \geq \min\{1/(\kappa_B\sqrt{\kappa_{2,G}}), 1\}$ and $\|\bar{\nabla}\mathcal{L}_k\| \leq \kappa_c + M_1 + \kappa_{\nabla f}$, we obtain

$$\frac{\Delta_k}{\bar{\gamma}_k} \leq \frac{2\eta_{\max}}{\zeta}(\kappa_c + \kappa_{\nabla f} + M_1)\cdot\max\{\kappa_B\sqrt{\kappa_{2,G}}, 1\}.$$

Therefore, (3.6) holds as long as

$$\bar{\mu}_k \geq \tilde{\mu} := \frac{2\eta_{\max}}{\zeta}(\kappa_c + \kappa_{\nabla f} + M_1)\cdot\max\{\kappa_B\sqrt{\kappa_{2,G}}, 1\} + \left(\frac{\kappa_B\Delta_{\max} + 2(M_1+\kappa_{\nabla f})}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

Since $\bar{\mu}_k$ is increased by at least a factor of ρ for each update, we define $\hat{\mu} := \rho\tilde{\mu}$ and complete the proof. ∎

We then show that for the method in Remark 3.4(ii), the merit parameter will stabilize just under Assumption 4.12(i). However, a tuning parameter $\theta \in (0,1)$ is involved to control the length of the normal step.

LEMMA A.2. *Suppose Assumptions* 4.1 *and* 4.12(i) *hold and the relaxation technique in Remark* 3.4(ii) *is employed. Then, there exist a (potentially random)* $\bar{K} < \infty$ *and a deterministic constant* $\hat{\mu}$, *such that* $\bar{\mu}_k = \bar{\mu}_{\bar{K}} \leq \hat{\mu}$ *for all* $k > \bar{K}$.

Proof. Similar to Lemma 4.13, we only show that there exists a deterministic threshold $\tilde{\mu} > 0$ independent of k such that (3.6) is satisfied as long as $\bar{\mu}_k \geq \tilde{\mu}$. Using the same derivation as Lemma A.1, we only need to show

$$\bar{\mu}_k \geq \frac{\Delta_k}{\bar{\gamma}_k} + \left(\frac{\kappa_B\Delta_{\max} + 2(M_1+\kappa_{\nabla f})}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right)$$

holds for $\bar{\mu}_k$ larger than a deterministic threshold for $\|c_k\| \neq 0$. Since all $k \geq 0$,

$$\frac{\Delta_k}{\bar{\gamma}_k} \leq \frac{\eta_{1,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\bar{\gamma}_k}.$$

By the projection technique of choosing $\bar{\gamma}_k$ and the fact that $\eta_{2,k} \geq \eta_{1,k}/2$, we have

$$\frac{\breve{\Delta}_k}{\|\boldsymbol{v}_k\|} = \frac{\theta\Delta_k}{\|\boldsymbol{v}_k\|} \overset{(3.2)}{\geq} \frac{\theta\eta_{2,k}\alpha_k\|\bar{\nabla}\mathcal{L}_k\|}{\|\boldsymbol{v}_k\|} \geq \frac{\theta\eta_{1,k}\alpha_k\|c_k\|}{2\|\boldsymbol{v}_k\|} \overset{(3.1)}{=} \frac{\theta\zeta\alpha_k}{2}.$$

Further, since $\theta\zeta\alpha_k/2 \leq 1$, we know $\theta\zeta\alpha_k/2 \leq \bar{\gamma}_k^{\text{trial}}$, implying $\bar{\gamma}_k \geq \theta\zeta\alpha_k/2$. Thus,

$$\frac{\Delta_k}{\bar{\gamma}_k} \leq \frac{2\eta_{\max}}{\zeta\theta}(\kappa_c + \kappa_{\nabla f} + M_1).$$

Therefore, (3.6) holds as long as

$$\bar{\mu}_k \geq \tilde{\mu} := \frac{2\eta_{\max}}{\zeta\theta}(\kappa_c + \kappa_{\nabla f} + M_1) + \left(\frac{\kappa_B\Delta_{\max} + 2(M_1 + \kappa_{\nabla f})}{\sqrt{\kappa_{1,G}}} + \frac{\kappa_B\kappa_c}{2\kappa_{1,G}}\right).$$

Since $\bar{\mu}_k$ is increased by at least a factor of ρ for each update, we define $\hat{\mu} := \rho\tilde{\mu}$ and complete the proof. ∎

REFERENCES

[1] A. S. Berahas, R. Bollapragada, and B. Zhou, *An Adaptive Sampling Sequential Quadratic Programming Method for Equality Constrained Stochastic Optimization*, preprint, https://arxiv.org/abs/2206.00712, 2022.

[2] A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson, *A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient Jacobians*, Math. Oper. Res., (2023), https://doi.org/10.1287/moor.2021.0154.

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou, *Sequential quadratic optimization for nonlinear equality constrained stochastic optimization*, SIAM J. Optim., 31 (2021), pp. 1352–1379, https://doi.org/10.1137/20m1354556.

[4] A. S. Berahas, J. Shi, Z. Yi, and B. Zhou, *Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction*, Comput. Optim. Appl., 86 (2023), pp. 79–116.

[5] D. Bertsekas, *Network Optimization: Continuous and Discrete Models*, Vol. 8, Athena Scientific, Belmont, MA, 1998, http://www.athenasc.com/netbook.html.

[6] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Elsevier, Amsterdam, 1982, https://doi.org/10.1016/c2013-0-10366-2.

[7] J. R. Birge, *State-of-the-art-survey—stochastic programming: Computation and applications*, INFORMS J. Comput., 9 (1997), pp. 111–133, https://doi.org/10.1287/ijoc.9.2.111.

[8] P. T. Boggs and J. W. Tolle, *Sequential quadratic programming*, Acta Numer., 4 (1995), pp. 1–51, https://doi.org/10.1017/s0962492900002518.

[9] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16m1080173.

[10] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *A trust region algorithm for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1152–1170, https://doi.org/10.1137/0724076.

[11] M. R. Celis, J. Dennis, Jr., and R. A. Tapia, *A Trust Region Strategy for Equality Constrained Optimization*, Technical report, Department of Mathematical Sciences, Rice University, Houston, TX, 1984, https://apps.dtic.mil/sti/citations/ADA454933.

[12] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011), pp. 1–27, https://doi.org/10.1145/1961189.1961199.

[13] C. Chen, F. Tung, N. Vedula, and G. Mori, *Constraint-aware deep neural network compression*, in Computer Vision – ECCV 2018, Springer, Cham, 2018, pp. 409–424, https://doi.org/10.1007/978-3-030-01237-3_25.

[14] R. Chen, M. Menickelly, and K. Scheinberg, *Stochastic optimization using a trust-region method and random models*, Math. Program., 169 (2017), pp. 447–487, https://doi.org/10.1007/s10107-017-1141-8.

[15] Y.-L. Chen, S. Na, and M. Kolar, *Convergence analysis of accelerated stochastic gradient descent under the growth condition*, Math. Oper. Res., (2023), https://doi.org/10.1287/moor.2021.0293.

[16] F. E. Curtis, M. J. O'Neill, and D. P. Robinson, *Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization*, Math. Program., 205 (2024), pp. 431–483.
[17] F. E. Curtis and D. P. Robinson, *Exploiting negative curvature in deterministic and stochastic optimization*, Math. Program., 176 (2018), pp. 69–94, https://doi.org/10.1007/s10107-018-1335-8.
[18] F. E. Curtis, D. P. Robinson, and B. Zhou, *Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Function Subject to Deterministic Nonlinear Equality Constraints*, preprint, https://arxiv.org/abs/2107.03512, 2021.
[19] F. E. Curtis, K. Scheinberg, and R. Shi, *A stochastic trust region algorithm based on careful step normalization*, INFORMS J. Optim., 1 (2019), pp. 200–220, https://doi.org/10.1287/ijoo.2018.0010.
[20] F. E. Curtis and R. Shi, *A fully stochastic second-order trust region method*, Optim. Methods Softw., 37 (2020), pp. 844–877, https://doi.org/10.1080/10556788.2020.1852403.
[21] J. Dupacova and R. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, Ann. Statist., 16 (1988), pp. 1517–1549, https://doi.org/10.1214/aos/1176351052.
[22] M. El-Alem, *A global convergence theory for the Celis-Dennis-Tapia trust-region algorithm for constrained optimization*, SIAM J. Numer. Anal., 28 (1991), pp. 266–290, https://doi.org/10.1137/0728015.
[23] N. I. M. Gould, D. Orban, and P. L. Toint, *CUTEst: A constrained and unconstrained testing environment with safe threads for mathematical optimization*, Comput. Optim. Appl., 60 (2014), pp. 545–557, https://doi.org/10.1007/s10589-014-9687-3.
[24] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Adv. Neural Inf. Process. Syst., 26 (2013), https://proceedings.neurips.cc/paper/2013/hash/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Abstract.html.
[25] S. Na, M. Anitescu, and M. Kolar, *Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming*, Math. Program., 202 (2023), pp. 279–353.
[26] S. Na, M. Anitescu, and M. Kolar, *An adaptive stochastic sequential quadratic programming with differentiable exact augmented Lagrangians*, Math. Program., 199 (2023), pp. 721–791, https://doi.org/10.1007/s10107-022-01846-z.
[27] S. Na, M. Derezinski, and M. W. Mahoney, *Hessian averaging in stochastic Newton methods achieves superlinear convergence*, Math. Program., 201 (2023), pp. 473–520.
[28] S. Na and M. W. Mahoney, *Asymptotic Convergence Rate and Statistical Inference for Stochastic Sequential Quadratic Programming*, preprint, https://arxiv.org/abs/2205.13687, 2022.
[29] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.
[30] E. O. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, University of Colorado, Boulder, CO, 1989, https://dl.acm.org/doi/10.5555/76519.
[31] M. J. D. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization*, Math. Programming, 49 (1990), pp. 189–211, https://doi.org/10.1007/bf01588787.
[32] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 271–298, https://doi.org/10.1137/080727154.
[33] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing Methods in Statistics, Elsevier, Amsterdam, 1971, pp. 233–257, https://doi.org/10.1016/b978-0-12-604550-5.50015-8.
[34] S. U. Stich, *Unified Optimal Analysis of the (stochastic) Gradient Method*, preprint, https://arxiv.org/abs/1907.04232, 2019.
[35] S. Sun and J. Nocedal, *A trust region method for noisy unconstrained optimization*, Math. Program., 202 (2023), pp. 445–472, https://doi.org/10.1007/s10107-023-01941-9.
[36] A. Vardi, *A trust region algorithm for equality constrained minimization: Convergence properties and implementation*, SIAM J. Numer. Anal., 22 (1985), pp. 575–591, https://doi.org/10.1137/0722035.
[37] S. Vaswani, F. Bach, and M. Schmidt, *Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron*, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 1195–1204, https://proceedings.mlr.press/v89/vaswani19a.html.
[38] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2005), pp. 25–57, https://doi.org/10.1007/s10107-004-0559-y.

[39] Y. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63, https://doi.org/10.1007/bf01580852.
[40] Y. Yuan, *A dual algorithm for minimizing a quadratic function with two quadratic constraints*, J. Comput. Math., 9 (1991), pp. 348–359, http://www.jstor.org/stable/45340300.
[41] Y. Zhang, *Computing a Celis-Dennis-Tapia trust-region step for equality constrained optimization*, Math. Programming, 55 (1992), pp. 109–124, https://doi.org/10.1007/bf01581194.