INFORMS Journal on Optimization

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Stochastic Inexact Sequential Quadratic Optimization Algorithm for Nonlinear Equality-Constrained Optimization

Frank E. Curtis, Daniel P. Robinson, Baoyu Zhou

To cite this article:
Frank E. Curtis, Daniel P. Robinson, Baoyu Zhou (2024) A Stochastic Inexact Sequential Quadratic Optimization Algorithm for Nonlinear Equality-Constrained Optimization. INFORMS Journal on Optimization

Published online in Articles in Advance 24 May 2024

. https://doi.org/10.1287/ijoo.2022.0008

https://pubsonline.informs.org/journal/ijoo

INFORMS JOURNAL ON OPTIMIZATION
Articles in Advance, pp. 1–23
ISSN 2575-1484 (print), ISSN 2575-1492 (online)

A Stochastic Inexact Sequential Quadratic Optimization Algorithm for Nonlinear Equality-Constrained Optimization

Frank E. Curtis,[a] Daniel P. Robinson,[a] Baoyu Zhou[b,*]

[a] Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania 18015; [b] School of Computing and Augmented Intelligence, Arizona State University, Tempe, Arizona 85281

*Corresponding author

Contact: frank.e.curtis@lehigh.edu, https://orcid.org/0000-0001-7214-9187 (FEC); daniel.p.robinson@lehigh.edu, https://orcid.org/0000-0003-0251-4227 (DPR); baoyu.zhou@asu.edu, https://orcid.org/0000-0003-3385-5788 (BZ)

Received: March 14, 2022
Revised: March 23, 2023; July 6, 2023; February 12, 2024
Accepted: April 26, 2024
Published Online in Articles in Advance: May 24, 2024

https://doi.org/10.1287/ijoo.2022.0008

Abstract. A stochastic algorithm is proposed, analyzed, and tested experimentally for solving continuous optimization problems with nonlinear equality constraints. It is assumed that constraint function and derivative values can be computed but that only stochastic approximations are available for the objective function and its derivatives. The algorithm is of the sequential quadratic optimization variety. Distinguishing features of the algorithm are that it only employs stochastic objective gradient estimates that satisfy a relatively weak set of assumptions (while using neither objective function values nor estimates of them) and that it allows inexact subproblem solutions to be employed, the latter of which is particularly useful in large-scale settings when the matrices defining the subproblems are too large to form and/or factorize. Conditions are imposed on the inexact subproblem solutions that account for the fact that only stochastic objective gradient estimates are employed. Convergence results are established for the method. Numerical experiments show that the proposed method vastly outperforms a stochastic subgradient method and can outperform an alternative sequential quadratic programming algorithm that employs highly accurate subproblem solutions in every iteration.

Funding: This material is based upon work supported by the National Science Foundation [Awards CCF-1740796 and CCF-2139735] and the Office of Naval Research [Award N00014-21-1-2532].

Keywords: nonlinear optimization • stochastic optimization • sequential quadratic optimization • inexact subproblem solves • iterative linear algebra techniques

1. Introduction

We propose, analyze, and present experimental results with a stochastic inexact sequential quadratic optimization (SISQO) algorithm for minimizing an objective function subject to (s.t.) nonlinear equality constraints. Specifically, our algorithm is designed to solve problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \text{ s.t. } c(x) = 0, \text{ with } f(x) = \mathbb{E}_\omega[F(x,\omega)], \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $c:\mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable, ω is a random variable with probability space $(\Omega, \mathcal{F}, P)$, $F:\mathbb{R}^n \times \Omega \to \mathbb{R}$, and $\mathbb{E}_\omega[\cdot]$ represents expectation taken with respect to the distribution of ω. Problems of this type arise in numerous important application areas. A partial list is the following: (i) learning a deep convolutional neural network for image recognition that imposes properties (e.g., smoothness) of the systems of partial differential equations (PDEs) that the convolutional layers are meant to interpret (Ruthotto and Haber 2020); (ii) multiple deep learning problems (see Márquez-Neila et al. 2017), including physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data (Zhu et al. 2019), natural language processing with constraints on output labels (Nandwani et al. 2019), image classification, detection, and localization (Ravi et al. 2019), deep reinforcement learning (Achiam et al. 2017), deep network compression (Chen et al. 2018), and manifold-regularized deep learning (Tomar and Rose 2014, Kumar Roy et al. 2018); (iii) accelerating the solution of PDE-constrained inverse problems by using a reduced-order model in place of a full-order model coupled with techniques to learn the discrepancy between the reduced- and full-order models (Sheriffdeen et al. 2019); (iv) multistage modeling (Shapiro et al. 2014); (v) portfolio selection (Shapiro et al. 2014); (vi) optimal power flow (Summers et al. 2015); and (vii) statistical problems, such as maximum likelihood estimation with constraints (Geyer 1991, Chatterjee et al. 2016).

Popular algorithmic approaches for solving problems of the form (1) when objective function and derivative values can be computed *deterministically* include penalty methods (Courant 1943, Fletcher 2000) and sequential quadratic optimization (SQO) methods (Wilson 1963, Powell 1978b). Penalty methods, which include popular strategies such as the augmented Lagrangian method and its variants, handle the constraints indirectly by adding a measure of constraint violation to the objective function, perhaps aided by information related to Lagrange multiplier estimates. The resulting unconstrained optimization problem, which can be nonsmooth depending on the choice of the constraint violation measure, may be solved using a host of methods, such as line search, trust region, cubic regularization, subgradient (Shor 2012), or proximal methods (Rockafellar 1976), where the appropriateness of the method depends on whether the constraint violation measure is smooth or nonsmooth. It is often the case that a sequence of such unconstrained problems needs to be solved to obtain appropriate Lagrange multiplier estimates and/or to identify an adequate weighting between the original objective f and the measure of constraint violation so that the original constrained problem can be solved to reasonable accuracy.

SQO methods, on the other hand, handle the constraints directly by employing local derivative-based approximations of the nonlinear constraints in explicit affine constraints in the subproblems employed to compute search directions; see, for example, Gill et al. (2002). For example, so-called line-search-SQO methods are considered state of the art for solving equality constrained optimization problems (Han 1977, Powell 1978a, Han and Mangasarian 1979). During each iteration of such a line-search-SQO method, a symmetric indefinite linear system of equations is solved, followed by a line search on an appropriate merit function to compute the next iterate. Here, the linear system is derived from applying Newton's method to the stationarity conditions for the nonlinear problem, and for this reason in the setting of equality constrained optimization, SQO methods are often referred to as Newton or Newton–SQO methods. For solving large-scale problems, factorizing the matrix in this linear system may be prohibitively expensive, in which case it may be preferable instead to apply an iterative linear system solver, such as minimum residual (MINRES) (Paige and Saunders 1975). This, in turn, opens the door to employing inexact subproblem solutions that may offer a better balance between per-iteration and overall computational costs of the algorithm for solving the original nonlinear problem. Identifying appropriate inexactness conditions that ensure that each search direction is sufficiently accurate so that the SQO algorithm is well posed and converges to a solution under reasonable assumptions is a challenging task with relatively few solutions in the literature (Heinkenschloss and Vicente 2002; Biros and Ghattas 2003; Byrd et al. 2008, 2010; Heinkenschloss and Ridzal 2008).

The success of SQO methods in the deterministic setting motivates us to study their extensions to the *stochastic* setting, which is a very challenging task. We are aware of only a few papers (e.g., Berahas et al. 2021, 2023; Na et al. 2023) that present stochastic algorithms for solving optimization problems with nonlinear equality constraints that offer convergence guarantees with respect to solving the constrained problem (rather than, say, merely a minimizer of a penalty function derived from the constrained problem). The algorithm by Na et al. (2023) is a line-search method that uses a differentiable exact augmented Lagrangian function as its merit function, whereas Berahas et al. (2021) (respectively, Berahas et al. 2023) propose an SQO method that uses an ℓ_1-norm (respectively, ℓ_2-norm) penalty function as its merit function. A similar algorithm but for a different setting is that by Oztoprak et al. (2023), which considers the setting in which only noisy objective and constrained values are accessible. All of these methods must factorize a matrix during each iteration, which may not be tractable when solving large-scale problems. This motivates us to extend the method in Berahas et al. (2021) to allow for using inexact subproblem solutions. Recently, Curtis et al. (2024) analyzed the worst-case complexity of the method by Berahas et al. (2021); this work makes it clear that such an analysis for constrained optimization is highly nontrivial.

1.1. Contributions

The contributions of this paper pertain to a new algorithm for solving Problem (1). (i) We design an SISQO method for solving (1) that is built upon a set of conditions that determine what constitutes an acceptable inexact subproblem solution along with an adaptive step size selection policy. The algorithm employs an ℓ_2-norm merit function, the parameter of which is updated dynamically by a procedure that has been designed with considerable care because it is this parameter that balances the emphasis between the objective function and the constraint violation during the optimization. (ii) Under mild assumptions (that we justify) that include good behavior of the adaptive merit parameter, we prove convergence guarantees for our algorithm. (iii) We present numerical results that compare our SISQO algorithm with a stochastic "exact" SQO algorithm and a stochastic subgradient method that show that our algorithm can outperform alternative techniques. In particular, we show that when all algorithms are given equal computational budgets, SISQO finds points with lower feasibility and Karush–Kuhn–Tucker (KKT) stationarity errors.

1.2. Notation

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_{\geq p}$ (respectively, $\mathbb{R}_{>p}$) denote the set of real numbers greater than or equal to (respectively, strictly greater than) $p \in \mathbb{R}$, and $\mathbb{N} := \{0, 1, 2, \dots\}$ denote the set of natural numbers. Let $\mathbb{R}^n$ denote the set of n-dimensional real vectors, $\mathbb{R}^{m \times n}$ denote the set of m- by n-dimensional real matrices, and $\mathbb{S}^n$ denote the set of n- by n-dimensional symmetric real matrices. For any $p \in \mathbb{N} \setminus \{0\}$, let $[p] := \{1, \dots, p\}$. The ℓ_2-norm is written simply as $\|\cdot\|$.

Any run of our algorithm generates a sequence of iterates $\{x_k\}$, where $x_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$. For all $k \in \mathbb{N}$, we append the subscript k to other quantities defined with respect to the kth iteration, and for brevity, we define $\nabla f_k := \nabla f(x_k)$, $c_k := c(x_k)$, and $\nabla c_k = \nabla c(x_k)$. We refer to the range space of ∇c_k as $\text{Range}(\nabla c_k)$ and refer to the null space of ∇c_k^T as $\text{Null}(\nabla c_k^T)$. The fundamental theorem of linear algebra provides that these spaces are orthogonal and $\text{Range}(\nabla c_k) + \text{Null}(\nabla c_k^T) = \mathbb{R}^n$. Finally, recall (see, e.g., Nocedal and Wright 2006) that a *primal* point $x \in \mathbb{R}^n$ and a *dual* point $y \in \mathbb{R}^m$ constitute a first-order stationary point for Problem (1) if and only if $c(x) = 0$ and $\nabla f(x) + \nabla c(x) y = 0$. These conditions are necessary for x to be a local minimizer when the constraint functions satisfy a constraint qualification as is assumed in the paper; see Assumption 1.

1.3. Organization

Our algorithm is presented in Section 2, with a convergence analysis in Section 3. The results of numerical experiments are presented in Section 4, and concluding remarks are presented in Section 5.

2. SISQO Algorithm

A run of our proposed SISQO algorithm generates a sequence

$$\{(x_k, y_k, g_k, v_k, u_k, d_k, \delta_k, \rho_k, r_k, \tau_k^{\text{trial}}, \tau_k, \xi_k^{\text{trial}}, \xi_k, \alpha_{k,\min}, \alpha_{k,\max}, \alpha_k)\}, \tag{2}$$

where for all $k \in \mathbb{N}$, $(x_k, y_k) \in \mathbb{R}^n \times \mathbb{R}^m$ is a primal-dual iterate pair; $g_k \in \mathbb{R}^n$ is a stochastic gradient estimate; $v_k \in \mathbb{R}^n$ is a *normal* primal direction that aims to reduced infeasibility by reducing a local derivative-based model of the ℓ_2-norm constraint violation measure; $u_k \in \mathbb{R}^n$ is a *tangential* primal direction that aims to maintain the reduction in linearized infeasibility achieved by the normal primal direction while also aiming to reduce the objective by reducing a stochastic gradient-based quadratic approximation of the objective; $d_k := v_k + u_k \in \mathbb{R}^n$ is a full primal direction; $\delta_k \in \mathbb{R}^m$ is a dual direction; $(\rho_k, r_k) \in \mathbb{R}^n \times \mathbb{R}^m$ is a primal-dual linear system residual pair; $(\tau_k^{\text{trial}}, \tau_k) \in \mathbb{R}_{>0} \cup \{\infty\} \times \mathbb{R}_{>0}$ is a pair of trial and employed merit parameter values; $(\xi_k^{\text{trial}}, \xi_k) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ is a pair of trial and employed ratio parameter values; and $(\alpha_{k,\min}, \alpha_{k,\max}, \alpha_k) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ is a tuple of minimum, maximum, and employed step size values, the last of which aims to produce a subsequent iterate $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yielding sufficient reduction in the ℓ_2-norm merit function. We present our algorithm in the context of the generation of a realization of the sequence (2), although our analysis ultimately considers the stochastic process defined by the algorithm, namely

$$\{(X_k, Y_k, G_k, V_k, U_k, D_k, \mathfrak{D}_k, \mathfrak{R}_k, R_k, \mathcal{T}_k^{\text{trial}}, \mathcal{T}_k, \Xi_k^{\text{trial}}, \Xi_k, \mathcal{A}_{k,\min}, \mathcal{A}_{k,\max}, \mathcal{A}_k)\}, \tag{3}$$

of which (2) is a realization. In the rest of this section, we present our algorithm in the context of the generation of a realization (2) toward our complete statement of Algorithm 1.

For the remainder of the paper, we make the following assumption.

Assumption 1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be an open convex set containing $\{x_k\}$ generated by every run of Algorithm 1. The objective $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below over $\mathcal{X}$, and its gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $L \in \mathbb{R}_{>0}$ (with respect to the ℓ_2-norm) and bounded over $\mathcal{X}$. The constraint $c : \mathbb{R}^n \to \mathbb{R}^m$ (with $m \leq n$) is continuously differentiable and bounded over $\mathcal{X}$, and its Jacobian $\nabla c(\cdot)^T : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ is Lipschitz continuous with constant $\Gamma \in \mathbb{R}_{>0}$ (with respect to the vector-induced ℓ_2-norm) and bounded over $\mathcal{X}$. In addition, for all $x \in \mathcal{X}$, the singular values of $\nabla c(x)^T$ are bounded uniformly below by a positive real number.*

The elements of this assumption are standard in the continuous constrained optimization literature. Note that it does not include an assumption that $\mathcal{X}$ is bounded. One could relax the assumption to say that $\mathcal{X}$ contains $\{X_k\}$ almost surely, but because such a relaxation would only make it necessary to remark constantly on the probability-zero event that $\{X_k\} \not\subset \mathcal{X}$ without adding much value to our ultimate results, we employ Assumption 1 as it is stated.

2.1. Merit Function

Motivated by the success of numerous line-search-SQO methods for solving deterministic equality constrained optimization problems, our algorithm employs an exact penalty function as a merit function; in particular, it

employs the ℓ_2-norm merit function $\phi : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$ defined by $\phi(x,\tau) = \tau f(x) + \|c(x)\|$, where τ is a positive *merit parameter* that is updated adaptively. (The choice of the ℓ_2-norm in ϕ is not essential. Another norm could be used. The choice of the ℓ_2-norm merely makes certain calculations simpler for our presentation and analysis.) A model $l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ of the merit function based on $g \approx \nabla f(x)$ and $\nabla c(x)$ is $l(x,\tau,g,d) = \tau(f(x) + g^T d) + \|c(x) + \nabla c(x)^T d\|$, with which we define the model reduction function $\Delta l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$\Delta l(x,\tau,g,d) = l(x,\tau,g,0) - l(x,\tau,g,d) = -\tau g^T d + \|c(x)\| - \|c(x) + \nabla c(x)^T d\|. \tag{4}$$

The merit and model reduction functions play critical roles in our inexactness conditions.

2.2. Computing a Search Direction

During the kth iteration, the algorithm computes a normal direction $v_k \in \text{Range}(\nabla c_k)$ based on

$$\min_{v \in \text{Range}(\nabla c_k)} \frac{1}{2}\|c_k + \nabla c_k^T v\|^2. \tag{5}$$

Instead of solving (5) exactly, the algorithm allows for an inexact solution to be employed by only requiring the computation of $v_k \in \text{Range}(\nabla c_k)$ satisfying the Cauchy decrease condition

$$\|c_k\| - \|c_k + \nabla c_k^T v_k\| \geq \epsilon_c(\|c_k\| - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|), \tag{6}$$

where $\epsilon_c \in (0,1]$ is user defined. In (6), $v_k^c := -\nabla c_k c_k$ is the negative-gradient direction for the objective of (5) at $v = 0$, and α_k^c is the step size along v_k^c that minimizes $\|c_k + \alpha \nabla c_k^T v_k^c\|$ over $\alpha \in \mathbb{R}$. If $c_k \neq 0$, then under Assumption 1, it follows that $\nabla c_k c_k \neq 0$,

$$\alpha_k^c = \|\nabla c_k c_k\|^2 / \|\nabla c_k^T \nabla c_k c_k\|^2 > 0,\ \alpha_k^c v_k^c \neq 0,\ \text{and}\ \ \|c_k\| - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\| > 0; \tag{7}$$

otherwise, if $c_k = 0$, then $\nabla c_k c_k = 0$, and $v_k = 0$ is the unique solution to (5). Popular choices for computing a normal direction satisfying the aforementioned conditions include Krylov subspace methods, such as the linear conjugate gradient (CG) method; see, for example, Nocedal and Wright (2006).

For describing the tangential direction computation, let us first describe what would be the computation in a deterministic variant of our approach. Given (x_k, y_k), ∇f_k, $v_k \in \text{Range}(\nabla c_k)$, and $H_k \in \mathbb{S}^n$ satisfying Assumption 2, consider the quadratic optimization subproblem

$$\min_{u \in \mathbb{R}^n} (\nabla f_k + H_k v_k)^T u + \frac{1}{2} u^T H_k u \ \text{ s.t. } \nabla c_k^T u = 0, \tag{8}$$

which has the unique solution $u_k^{\text{true}} \in \text{Null}(\nabla c_k^T)$ that satisfies, for some $\delta_k^{\text{true}} \in \mathbb{R}^m$,

$$\begin{bmatrix} H_k & \nabla c_k \\ \nabla c_k^T & 0 \end{bmatrix} \begin{bmatrix} u_k^{\text{true}} \\ \delta_k^{\text{true}} \end{bmatrix} = -\begin{bmatrix} \nabla f_k + H_k v_k + \nabla c_k y_k \\ 0 \end{bmatrix}. \tag{9}$$

We make the following assumption pertaining to $\{H_k\}$ throughout the paper.

Assumption 2. *There exists $\zeta \in \mathbb{R}_{>0}$ and $\kappa_H \in \mathbb{R}_{>\zeta}$ such that for all $k \in \mathbb{N}$ in any run of the algorithm, $\|H_k\| \leq \kappa_H$ and $u^T H_k u \geq \zeta \|u\|^2$ for all $u \in \text{Null}(\nabla c_k^T)$.*

The introduction of (8) and (9) allows us to define, for the purposes of our analysis only (i.e., not for actual computation), the *true and exact* primal-dual search direction conditioned on the behavior of the algorithm up to the kth iteration as $(d_k^{\text{true}}, \delta_k^{\text{true}})$, where $d_k^{\text{true}} := v_k + u_k^{\text{true}}$. Because our algorithm only presumes access to a stochastic gradient estimate g_k of ∇f_k, the corresponding *exact*, but not *true*, primal-dual search direction is $(d_{k,*}, \delta_{k,*})$, where $d_{k,*} := v_k + u_{k,*}$ with $(u_{k,*}, \delta_{k,*})$ satisfying

$$\begin{bmatrix} H_k & \nabla c_k \\ \nabla c_k^T & 0 \end{bmatrix} \begin{bmatrix} u_{k,*} \\ \delta_{k,*} \end{bmatrix} = -\begin{bmatrix} g_k + H_k v_k + \nabla c_k y_k \\ 0 \end{bmatrix}. \tag{10}$$

(For the description of our algorithm and our initial analysis, it suffices that $g_k \in \mathbb{R}^n$. Our ultimate required assumption about the stochastic gradient estimators is Assumption 6.)

Our algorithm, to avoid having to form or factorize the matrix in (10), computes a tangential direction by computing (u_k, δ_k) through iterative linear algebra techniques applied to the symmetric indefinite system (10). In

particular, our algorithm computes (u_k, δ_k) such that the full primal search direction $d_k := v_k + u_k$, dual direction δ_k, and residual

$$\begin{bmatrix} \rho_k \\ r_k \end{bmatrix} := \begin{bmatrix} H_k & \nabla c_k \\ \nabla c_k^T & 0 \end{bmatrix} \begin{bmatrix} u_k \\ \delta_k \end{bmatrix} + \begin{bmatrix} g_k + H_k v_k + \nabla c_k y_k \\ 0 \end{bmatrix} \tag{11}$$

satisfy at least one of two sets of conditions. Next, we describe these conditions that the algorithm employs to determine what constitutes an acceptable search direction and corresponding residuals.

In the deterministic setting, line-search-SQO methods commonly combine the search direction with an updating strategy for the merit parameter in a manner that ensures that the computed direction is one of sufficient descent for the merit function. The required descent condition is guaranteed to be satisfied by choosing the merit parameter to be sufficiently small so that the reduction in a model of the merit function (recall (4)) is sufficiently large; see, for example, Byrd et al. (2008, lemma 3.1). Following such an approach, our algorithm requires that (u_k, δ_k) (yielding $d_k := v_k + u_k$) be computed and τ be set such that the model reduction condition

$$\Delta l(x_k, \tau, g_k, v_k + u_k) \geq \sigma_u \tau \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|) \tag{12}$$

holds for some user-defined $\sigma_u \in (0,1)$, $\epsilon_u \in (0, \zeta)$ (see Assumption 2), and $\sigma_c \in (0,1)$. The value for τ for which (12) is required to hold depends on one of two different situations as described next.

Condition (12) plays a central role in the conditions that we require (u_k, δ_k) to satisfy. We define these in the context of *termination tests* (TTs) because they dictate conditions that once satisfied, can cause termination of an iterative linear system solver applied to (10). (The tests are inspired by the *sufficient merit approximation reduction termination tests* developed in Byrd et al. (2008, 2010) and Curtis et al. (2009) for a deterministic SQO method.) Our first termination test states that an inexact solution is acceptable if (12) is satisfied with the current merit parameter value (i.e., $\tau \equiv \tau_k \leftarrow \tau_{k-1}$), the norms of the residuals satisfy certain upper bounds, and either the tangential direction is sufficiently small in norm compared with the normal direction or the tangential direction is one of sufficiently positive curvature for H_k and yields a sufficiently small objective value for (8) (with g_k in place of ∇f_k). The test makes use of a sequence $\{\beta_k\}$ that will also play a critical role in our step size selection scheme that is described in the next subsection.

2.2.1. Termination Test 1. Given $\kappa \in (0,1)$, $\beta_k \in (0,1]$, $\kappa_\rho \in \mathbb{R}_{>0}$, $\kappa_r \in \mathbb{R}_{>0}$, $\kappa_u \in \mathbb{R}_{>0}$, $\epsilon_u \in (0, \zeta)$, $\kappa_v \in \mathbb{R}_{>0}$, $\sigma_u \in (0,1)$, $\sigma_c \in (0,1)$, and $v_k \in \text{Range}(\nabla c_k)$ computed to satisfy (6), the pair (u_k, δ_k) satisfies TT1 if with the pair (ρ_k, r_k) defined in (11), it holds that

$$\|\rho_k\| \leq \kappa \min\left\{ \left\| \begin{bmatrix} g_k + \nabla c_k(y_k + \delta_k) \\ c_k \end{bmatrix} \right\|, \left\| \begin{bmatrix} g_{k-1} + \nabla c_{k-1} y_k \\ c_{k-1} \end{bmatrix} \right\| \right\}; \tag{13}$$

$$\|\rho_k\| \leq \kappa_\rho \beta_k \quad \text{and} \quad \|r_k\| \leq \kappa_r \beta_k; \tag{14}$$

$$\|u_k\| \leq \kappa_u \|v_k\| \text{ or } \left\{ \begin{array}{r} u_k^T H_k u_k \geq \epsilon_u \|u_k\|^2 \text{ and} \\ (g_k + H_k v_k)^T u_k + \frac{1}{2} u_k^T H_k u_k \leq \kappa_v \|v_k\| \end{array} \right\}, \tag{15}$$

and (12) is satisfied with $\tau \equiv \tau_{k-1}$. (In this case, $\tau_k \leftarrow \tau_{k-1}$, so (12) holds with $\tau \equiv \tau_k$.)

TT1 cannot be enforced in every iteration, even in the deterministic setting, because there may exist points in the search space at which all of the conditions required cannot be satisfied simultaneously, even if the linear system (10) is solved accurately. In short, the algorithm needs to allow for the computation of a search direction for which (12) can only be satisfied with a decrease of the merit parameter. That said, the algorithm needs to be careful in terms of the situations in which such a decrease is allowed, or else, the merit parameter sequence may behave in a manner that ruins a convergence guarantee for solving the original constrained optimization problem. For our algorithm, we employ the following termination test for this situation.

2.2.2. Termination Test 2. Given $\kappa \in (0,1)$, $\beta_k \in (0,1]$, $\kappa_\rho \in \mathbb{R}_{>0}$, $\kappa_r \in \mathbb{R}_{>0}$, $\kappa_u \in \mathbb{R}_{>0}$, $\epsilon_u \in (0, \zeta)$, $\kappa_v \in \mathbb{R}_{>0}$, $\sigma_c \in (0,1)$, $\epsilon_r \in (\sigma_c, 1)$, and $v_k \in \text{Range}(\nabla c_k)$ computed to satisfy (6), the pair (u_k, δ_k) satisfies Termination Test 2 if with the pair (ρ_k, r_k) defined in (11), (13)–(15) and

$$\|c_k\| - \|c_k + \nabla c_k^T v_k + r_k\| \geq \epsilon_r(\|c_k\| - \|c_k + \nabla c_k^T v_k\|) > 0 \tag{16}$$

hold. (In this case, for user-defined $\epsilon_\tau \in (0,1)$, the algorithm will set

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \leq \tau_k^{\text{trial}} \\ \min\{(1-\epsilon_\tau)\tau_{k-1}, \tau_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \tag{17}$$

$$\text{where } \tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} \leq 0 \\ \dfrac{\left(1 - \frac{\sigma_\epsilon}{\epsilon_r}\right)(\|c_k\| - \|c_k + \nabla c_k^T v_k + r_k\|)}{g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}} & \text{otherwise,} \end{cases} \tag{18}$$

so (12) is satisfied with $\tau \equiv \tau_k$. See Lemma 3 for a proof.)

In Lemma 1, we show under a loose assumption about the iterative linear system solver that for all $k \in \mathbb{N}$, the algorithm can compute a pair (u_k, δ_k) satisfying at least one of TT1 or TT2. Therefore, the index of each iteration of each realization of our method is contained in one of two index sets:

$$\mathcal{K}_1 := \{k \in \mathbb{N} : (u_k, \delta_k) \text{ satisfies TT1}\} \text{ or } \mathcal{K}_2 := \{k \in \mathbb{N} : (u_k, \delta_k) \text{ satisfies TT2, but not TT1}\}.$$

It is worthwhile to emphasize that in terms of the stochastic process defined by the algorithm, the index sets $\mathcal{K}_1$ and $\mathcal{K}_2$ are random (i.e., they may contain different indices in different runs). This randomness is handled as part of our convergence analysis.

2.3. Computing a Step Size

Upon computation of $d_k \leftarrow v_k + u_k$, our algorithm computes a positive step size α_k to set x_{k+1}. Given positive Lipschitz constants L and Γ (recall Assumption 1), it follows for all $\alpha \in \mathbb{R}_{>0}$ that

$$f(x_k + \alpha d_k) \leq f_k + \alpha \nabla f_k^T d_k + \frac{1}{2} L \alpha^2 \|d_k\|^2 \text{ and } \|c(x_k + \alpha d_k)\| \leq \|c_k + \alpha \nabla c_k^T d_k\| + \frac{1}{2} \Gamma \alpha^2 \|d_k\|^2. \tag{19}$$

Combining these with (4) yields

$$\begin{aligned} \phi(x_k + \alpha d_k, \tau_k) - \phi(x_k, \tau_k) &= \tau_k f(x_k + \alpha d_k) - \tau_k f_k + \|c(x_k + \alpha d_k)\| - \|c_k\| \\ &\leq \alpha \tau_k \nabla f_k^T d_k + (|1-\alpha| - 1)\|c_k\| + \alpha \|c_k + \nabla c_k^T d_k\| + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2 \\ &= -\alpha \Delta l(x_k, \tau_k, \nabla f_k, d_k) + (|1-\alpha| - (1-\alpha))\|c_k\| + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2. \end{aligned} \tag{20}$$

This derivation provides a convex piecewise quadratic upper-bounding function for the change in the merit function corresponding to a step from x_k to $x_k + \alpha d_k$. Given user-defined $\eta \in (0,1)$ and the aforementioned sequence $\{\beta_k\} \subset (0,1]$, our algorithm's step size selection scheme makes use of

$$\alpha_k^{\text{suff}} := \min\left\{\frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2}, 1\right\}. \tag{21}$$

The definition of α_k^{suff} can be motivated as follows. Its value, when $\beta_k = 1$, is the largest in [0, 1] such that for all $\alpha \in [0, \alpha_k^{\text{suff}}]$, the right-hand side of (20) (with ∇f_k replaced by g_k) is less than or equal to $-\eta\alpha\Delta l(x_k, \tau_k, g_k, d_k)$. Such an inequality is representative of one enforced in deterministic line-search-SQO methods. Otherwise, with $\beta_k \in (0,1]$ introduced and not necessarily equal to one, the value of α_k^{suff} can be diminished during the optimization, which allows for step size control as is required for convergence guarantees for certain stochastic gradient-based methods; see, for example, Bottou et al. (2018). The first term inside the min appearing in (21) is important for the convergence guarantees that we prove for our method, but it can behave erratically because of the algorithm's use of stochastic gradients. To account for this, given user-defined $\epsilon_\xi \in (0,1)$, our algorithm defines

$$\xi_k^{\text{trial}} := \frac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|^2} \text{ and } \xi_k := \begin{cases} \xi_{k-1} & \text{if } \xi_{k-1} \leq \xi_k^{\text{trial}} \\ \min\{(1-\epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \tag{22}$$

so that $\xi_k \leq \xi_k^{\text{trial}} = \Delta l(x_k, \tau_k, g_k, d_k)/(\tau_k \|d_k\|^2)$ for all $k \in \mathbb{N}$. Combining this inequality with (21), the monotonically nonincreasing behaviors of $\{\xi_k\}$ and $\{\tau_k\}$, and assuming that $\{\beta_k\}$ satisfies

$$2(1-\eta)\beta_k \xi_{-1} \tau_{-1}/\Gamma \in (0,1] \text{ for all } k \in \mathbb{N}, \tag{23}$$

where ξ_{-1} and τ_{-1} initialize the sequences $\{\xi_k\}$ and $\{\tau_k\}$, respectively, one finds that

$$\alpha_k^{\min} := \frac{2(1-\eta)\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} \leq \min\left\{\frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2}, 1\right\} \equiv \alpha_k^{\text{suff}}. \tag{24}$$

The value $\alpha_k^{\min}$ serves as a minimum value (i.e., a lower bound) for our choice of step size. In our analysis, we show that the sequence $\{\xi_k\}$ is bounded below and away from zero by a positive real number that is common to all runs of the algorithm (see Lemma 9).

Next, let us derive a maximum value (i.e., an upper bound) for our algorithm's choice of step size. If $d_k = 0$, then without loss of generality, the algorithm can set $\alpha_k \leftarrow \alpha_{k,\min}$. Otherwise, if $d_k \neq 0$, consider the strongly convex function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned}\varphi(\alpha) := (\eta - 1)\alpha\beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \|c_k + \alpha \nabla c_k^T d_k\| - \|c_k\| \\ + \alpha(\|c_k\| - \|c_k + \nabla c_k^T d_k\|) + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2.\end{aligned} \tag{25}$$

Notice that when $\beta_k = 1$, it holds that $\varphi(\alpha) \leq 0$ for all $\alpha \in \mathbb{R}_{\geq 0}$ if and only if the right-hand side of (20) with ∇f_k replaced by g_k is less than or equal to $-\eta\alpha\Delta l(x_k, \tau_k, g_k, d_k)$. Thus, following a similar argument, one can be motivated as to the fact that our algorithm never allows a step size larger than $\alpha_k^{\varphi} := \max\{\alpha \in \mathbb{R}_{\geq 0} : \varphi(\alpha) \leq 0\}$. Finally, again to mitigate adverse effects caused by the use of stochastic gradients, our algorithm employs the maximum step size

$$\alpha_k^{\max} := \min\{\alpha_k^{\varphi}, \alpha_k^{\min} + \theta\beta_k^2\}, \tag{26}$$

where $\theta \in \mathbb{R}_{\geq 0}$ is user defined. Overall, our algorithm allows any step size with $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$. Lemma 4 in our analysis shows that this interval is nonempty.

2.4. Updating the Primal-Dual Iterate

In the primal space, our algorithm employs the iterate update $x_{k+1} \leftarrow x_k + \alpha_k d_k$. However, in the dual space, it allows additional flexibility. For consistency with the deterministic setting (see, e.g., Curtis et al. 2009, equation 2.19), our algorithm is stated to require y_{k+1} to satisfy

$$\|g_k + \nabla c_k y_{k+1}\| \leq \|g_k + \nabla c_k (y_k + \delta_k)\|. \tag{27}$$

Clearly, choosing $y_{k+1} \leftarrow y_k + \delta_k$ is one particular option satisfying (27), although other choices, such as least-squares multipliers, could also be used.

Algorithm 1 (Stochastic Inexact Sequential Quadratic Optimization (SISQO))

Require: $(x_0, y_0, \tau_{-1}, \xi_{-1}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$; $(L, \Gamma) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ satisfying Assumption 1; $\epsilon_c \in (0, 1]$; $\epsilon_u \in (0, \zeta)$; $\{\sigma_u, \sigma_c, \kappa, \epsilon_\tau, \epsilon_\xi, \eta\} \subset (0, 1)$; $\{\kappa_\rho, \kappa_r, \kappa_u, \kappa_v, \theta\} \subset \mathbb{R}_{>0}$; $\epsilon_r \in (\sigma_c, 1)$

for all $k \in \mathbb{N}$ **do**
  choose $\beta_k \in (0, 1]$ satisfying (23)
  compute $v_k \in \text{Range}(\nabla c_k)$ satisfying (6)
  compute g_k (see Assumption 6)
  compute H_k (see Assumption 2 and Assumption 6)
  compute (u_k, δ_k) satisfying at least one of TT1 or TT2
  if TT1 is satisfied **then**
    set $\tau_k^{\text{trial}} \leftarrow \infty$ and $\tau_k \leftarrow \tau_{k-1}$ ▷ $k \in \mathcal{K}_1$
  else (TT2 is satisfied)
    set τ_k^{trial} and τ_k by (17) and (18) ▷ $k \in \mathcal{K}_2$
  end if
  set $d_k \leftarrow v_k + u_k$
  compute ξ_k and ξ_k^{trial} by (22)
  choose $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$ using the definitions in (24) and (26)
  set $x_{k+1} \leftarrow x_k + \alpha_k d_k$, and choose y_{k+1} satisfying (27)
end for

3. Analysis

Our analysis is presented in three parts. In Section 3.1, we show that Algorithm 1 is well posed, which is followed by Section 3.2, in which a set of lemmas is proved that hold for every run of the algorithm. The analysis in

Sections 3.1 and 3.2 is written generically in terms of a realization of a run of the algorithm. Then, in Section 3.3, we prove convergence properties for the algorithm that are written in terms of the stochastic process defined by the algorithm. This analysis focuses on an event that presumes certain behavior of the merit and ratio parameter sequences.

3.1. Well-Posedness

Our aim in this subsection is to prove that each procedure in each iteration of every run of Algorithm 1 is performed in a manner that terminates finitely under mild assumptions. Along the way, we establish other useful properties. In this subsection and the following one, our analysis merely requires for all $k \in \mathbb{N}$ that $g_k \in \mathbb{R}^n$ and that $H_k \in \mathbb{S}^n$ satisfies Assumption 2.

For the sake of generality, we make the following assumption about the iterative linear system solver employed in line 6 of Algorithm 1, which merely requires that the residual of the linear system solve vanishes asymptotically over any run of the solver. We emphasize, however, that there exist approaches, such as MINRES (Paige and Saunders 1975), that guarantee that an exact solution—which would satisfy our termination tests—can be computed in a number of linear system solver iterations that is bounded uniformly for all linear systems arising throughout any realization of the algorithm. However, we merely make the following assumption because it is all that is required by our analysis, and it offers more flexibility in the choice of linear system solver.

Assumption 3. *For all $k \in \mathbb{N}$ in any run, the iterative linear system solver employed in line 7 of Algorithm 1 to compute (u_k, δ_k) generates a sequence*

$$\{(u_{k,t}, \delta_{k,t}, \rho_{k,t}, r_{k,t})\}_{t\in\mathbb{N}} \text{ with } \begin{bmatrix} \rho_{k,t} \\ r_{k,t} \end{bmatrix} = \begin{bmatrix} H_k & \nabla c_k \\ \nabla c_k^T & 0 \end{bmatrix} \begin{bmatrix} u_{k,t} \\ \delta_{k,t} \end{bmatrix} + \begin{bmatrix} g_k + H_k v_k + \nabla c_k y_k \\ 0 \end{bmatrix} \text{ for all } t \in \mathbb{N} \tag{28}$$

such that $\lim_{t\to\infty} \|(u_{k,t}, \delta_{k,t}, \rho_{k,t}, r_{k,t}) - (u_{k,*}, \delta_{k,*}, 0, 0)\| = 0$, *where* $(u_{k,*}, \delta_{k,*})$ *uniquely solves* (10).

We also make the following assumption concerning the algorithm iterates and corresponding stochastic gradient estimates computed in each iteration.

Assumption 4. *For all $k \in \mathbb{N}$ in any run, it holds that $c_k \neq 0$ or $g_k \notin \text{Range}(\nabla c_k)$.*

We justify Assumption 4 in the following manner. In the deterministic setting, the algorithm encounters a point x_k such that $c_k = 0$ and $\nabla f_k \in \text{Range}(\nabla c_k)$ if and only if there exists y_k such that (x_k, y_k) is first-order stationary for Problem (1). In such a scenario, it is reasonable to require that an exact solution of (10) is computed or at least that a sufficiently accurate solution is computed such that a practical termination condition for (10) is triggered and the algorithm terminates. In the stochastic setting, the algorithm encounters $c_k = 0$ and $g_k \in \text{Range}(\nabla c_k)$ if and only if x_k is *exactly* feasible and the stochastic gradient lies *exactly* in the range space of ∇c_k. Because g_k is a *stochastic* gradient, we contend that it is unlikely that it will lie *exactly* in $\text{Range}(\nabla c_k)$, except in special circumstances. Thus, for simplicity in our analysis, we impose Assumption 4. Note that if Assumption 4 did not hold, then one of the following could be employed in a practical implementation. (i) If a sufficiently accurate solution of (10) satisfies neither TT1 nor TT2, then a new stochastic gradient could be sampled, perhaps following a procedure to ensure that if multiple new stochastic gradients are computed, then each is computed with lower variance, or (ii) random (e.g., Gaussian) noise could be added to g_k for all $k \in \mathbb{N}$ so that Assumption 4 holds almost surely in all iterations, in which case the convergence result that we prove holds almost surely.

We can now show that the search direction computation is well posed. We remark in passing that if one was to employ a linear system solver, such as MINRES, that would produce an exact solution of the linear system within a uniformly bounded number of iterations, then the arguments in the proof of the following lemma would show that the linear system solver computes (u_k, δ_k), satisfying at least one of TT1 or TT2 in a uniformly bounded number of iterations.

Lemma 1. *For all $k \in \mathbb{N}$ in any run, the iterative linear system solver computes (u_k, δ_k) satisfying at least one of TT1 or TT2 in a finite number of iterations.*

Proof. We prove the result by considering two cases.

Case 1. $c_k \neq 0$. For this case, we show that $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ satisfies TT2 for sufficiently large $t \in \mathbb{N}$. Let us first observe that it follows from Assumption 3, Assumption 4, and $\beta_k \in (0, 1]$ that (13) and (14) hold with $(\rho_k, r_k) \equiv (\rho_{k,t}, r_{k,t})$ for all sufficiently large $t \in \mathbb{N}$.

Let us now show that (15) holds for all sufficiently large $t \in \mathbb{N}$. Because $c_k \neq 0$, it follows under Assumption 1 that $v_k \neq 0$. If $u_{k,*} = 0$, then Assumption 3 implies $\{u_{k,t}\} \to u_{k,*} = 0$, in which case it follows from $\kappa_u \in \mathbb{R}_{>0}$ that the former

condition in (15) holds with $u_k \equiv u_{k,t}$ for all sufficiently large $t \in \mathbb{N}$. On the other hand, if $u_{k,*} \neq 0$, then (10) and Assumption 2 imply

$$u_{k,*}^T(g_k + H_k v_k) + \frac{1}{2} u_{k,*}^T H_k u_{k,*} < u_{k,*}^T(g_k + H_k v_k) + u_{k,*}^T H_k u_{k,*} = -u_{k,*}^T \nabla c_k(y_k + \delta_{k,*}) = 0. \tag{29}$$

Combining this inequality with $\epsilon_u \in (0, \zeta)$, $\kappa_v \in \mathbb{R}_{>0}$, $v_k \neq 0$, and Assumptions 2 and 3, it follows that the latter set of conditions in (15) holds with $u_k \equiv u_{k,t}$ for all sufficiently large $t \in \mathbb{N}$.

Finally, let us show that (16) holds for all sufficiently large $t \in \mathbb{N}$, which combined with the previous conclusions, shows that TT2 is satisfied by $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ for all sufficiently large $t \in \mathbb{N}$. By Assumption 3, (6), and $v_k \neq 0$, it follows that $\lim_{t\to\infty}(\|c_k\| - \|c_k + \nabla c_k^T v_k + r_{k,t}\|) = \|c_k\| - \|c_k + \nabla c_k^T v_k\| > 0$, which shows that (16) holds with $r_k \equiv r_{k,t}$ for all sufficiently large $t \in \mathbb{N}$.

Case 2. $c_k = 0$. For this case, we show that $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ satisfies TT1 for all sufficiently large $t \in \mathbb{N}$. First, recall that $c_k = 0$ implies that $v_k = 0$. We also claim that $u_{k,*} \neq 0$. To prove this by contradiction, suppose that $u_{k,*} = 0$. Combining this fact with $v_k = 0$ and (10), it follows that $g_k + \nabla c_k(y_k + \delta_{k,*}) = 0$, which with $c_k = 0$, violates Assumption 4. Thus, $u_{k,*} \neq 0$.

Next, notice that the argument used in the beginning of case (1) still applies in this case, which allows us to conclude that both (13) and (14) hold with $(\rho_k, r_k) = (\rho_{k,t}, r_{k,t})$ for all sufficiently large $t \in \mathbb{N}$. Combining (29) with Assumptions 2 and 3 and $\epsilon_u \in (0, \zeta)$ allows us to deduce that the second set of conditions in (15) holds with $u_k \equiv u_{k,t}$ for all sufficiently large $t \in \mathbb{N}$. Next, from the fact that $v_k = 0$ and (10), it follows that $\nabla c_k^T d_{k,*} = \nabla c_k^T(u_{k,*} + v_k) = 0$, which with Assumption 2 and $\epsilon_u \in (0, \zeta)$, gives $u_{k,*}^T H_k u_{k,*} \geq \zeta \|u_{k,*}\|^2 > \epsilon_u \|u_{k,*}\|^2$, from which we deduce that $\max\{u_{k,*}^T H_k u_{k,*}, \epsilon_u \|u_{k,*}\|^2\} = u_{k,*}^T H_k u_{k,*} \geq \zeta \|u_{k,*}\|^2 > 0$. Combining this inequality with $c_k = 0$, $v_k = 0$, $\nabla c_k^T d_{k,*} = \nabla c_k^T v_k = 0$, (10), and Assumption 2 shows that

$$\begin{aligned}
&\Delta l(x_k, \tau_{k-1}, g_k, d_{k,*}) = -\tau_{k-1} g_k^T d_{k,*} + \|c_k\| - \|c_k + \nabla c_k^T d_{k,*}\| = -\tau_{k-1} g_k^T u_{k,*} \\
&= -\tau_{k-1}(-H_k u_{k,*} - H_k v_k - \nabla c_k(y_k + \delta_{k,*}))^T u_{k,*} = \tau_{k-1} u_{k,*}^T H_k u_{k,*} \\
&> \sigma_u \tau_{k-1} \max\{u_{k,*}^T H_k u_{k,*}, \epsilon_u \|u_{k,*}\|^2\} + \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|) > 0,
\end{aligned}$$

meaning that (12) holds with $\tau \equiv \tau_{k-1}$ for all sufficiently large $t \in \mathbb{N}$. In summary, we have shown that for all sufficiently large $t \in \mathbb{N}$, the pair $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ satisfies TT1. □

Next, we prove that every full primal direction is nonzero.

Lemma 2. *For all $k \in \mathbb{N}$ in any run, it holds that $d_k \neq 0$.*

Proof. By contradiction, suppose that $d_k = 0$. From this fact, $d_k = v_k + u_k$, and (11), it follows that $\rho_k = g_k + \nabla c_k(y_k + \delta_k) + H_k(v_k + u_k) = g_k + \nabla c_k(y_k + \delta_k)$. If $c_k = 0$, then this shows that the inequality in (13) cannot hold, meaning that (u_k, δ_k) satisfies neither TT1 nor TT2, which contradicts Lemma 1. Hence, the only possibility is that $c_k \neq 0$, which we assume for the rest of the proof.

Notice that from $d_k = 0$, $d_k = v_k + u_k$, and $r_k = \nabla c_k^T u_k$, it follows that $\|c_k\| - \|c_k + \nabla c_k^T v_k + r_k\| = \|c_k\| - \|c_k + \nabla c_k^T d_k\| = 0$, meaning that (16) is not satisfied; thus, (u_k, δ_k) does not satisfy TT2. Also, observe from $v_k \neq 0$ (which follows from $c_k \neq 0$ and Assumption 1), $d_k = 0$, and (6) that $\Delta l(x_k, \tau_{k-1}, g_k, d_k) = 0 < \sigma_u \tau_{k-1} \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|)$, meaning that (12) is not satisfied with $\tau = \tau_{k-1}$; thus, (u_k, δ_k) does not satisfy TT1. Overall, we have reached a contradiction to Lemma 1, and because we have reached a contradiction in all cases, the original supposition that $d_k = 0$ cannot be true. □

We now show that our update strategy for the merit parameter ensures that the model reduction condition (12) always holds for $\tau \equiv \tau_k$. We also show another important property of $\{\tau_k\}$.

Lemma 3. *For all $k \in \mathbb{N}$ in any run, the inequality in (12) holds with $\tau \equiv \tau_k$. In addition, for all $k \in \mathbb{N}$ such that $\tau_{k+1} < \tau_k$, it holds that $\tau_{k+1} \leq (1 - \epsilon_\tau)\tau_k$.*

Proof. The desired conclusion follows for $k \in \mathcal{K}_1$ because of the manner in which TT1 is defined and the fact that the algorithm sets $\tau_k \leftarrow \tau_{k-1}$ for all $k \in \mathcal{K}_1$. Hence, let us proceed under the assumption that $k \in \mathcal{K}_2$. The inequality in (12) holds for $\tau \equiv \tau_k$ with $d_k = v_k + u_k$ if and only if

$$\tau_k(g_k^T d_k + \sigma_u \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}) \leq \|c_k\| - \|c_k + \nabla c_k^T d_k\| - \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|).$$

We now proceed to show that this inequality holds by considering two cases.

Case 1. $g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} \leq 0$. In this case, the algorithm sets $\tau_k \leftarrow \tau_{k-1}$. Combining this with (16), $\nabla c_k^T u_k = r_k$, and $\epsilon_r \in (\sigma_c, 1)$ yields

$$\begin{aligned}\tau_k(g_k^T d_k + \sigma_u \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}) &\leq \tau_k(g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}) \leq 0\\ &\leq \|c_k\| - \|c_k + \nabla c_k^T d_k\| - \epsilon_r(\|c_k\| - \|c_k + \nabla c_k^T v_k\|)\\ &< \|c_k\| - \|c_k + \nabla c_k^T d_k\| - \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|),\end{aligned}$$

which establishes the desired inequality.

Case 2. $g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} > 0$. The update (17) yields $\tau_k \leq \tau_k^{\text{trial}}$, which combined with (16), (18), $\nabla c_k^T u_k = r_k$, and $\epsilon_r \in (\sigma_c, 1)$ yields

$$\begin{aligned}&\tau_k(g_k^T d_k + \sigma_u \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}) \leq \tau_k(g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\})\\ &\leq \left(1 - \frac{\sigma_c}{\epsilon_r}\right)(\|c_k\| - \|c_k + \nabla c_k^T d_k\|) \leq \|c_k\| - \|c_k + \nabla c_k^T d_k\| - \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|),\end{aligned}$$

as desired. Moreover, from (17), we have $\tau_{k+1} \leq (1 - \epsilon_\tau)\tau_k$ whenever $\tau_{k+1} < \tau_k$. □

We conclude this subsection by showing that the interval defining our step size selection scheme (i.e., $[\alpha_k^{\min}, \alpha_k^{\max}]$) is positive and nonempty for all $k \in \mathbb{N}$. We also show a useful property of the computed step size that is needed in our analysis.

Lemma 4. *For all $k \in \mathbb{N}$ in any run, $0 < \alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$, $0 < \alpha_k^{\min} \leq \alpha_k^{\max}$, and $\varphi(\alpha_k) \leq 0$.*

Proof. It follows from (24) and the fact that $\{\beta_k\}$, $\{\xi_k\}$, and $\{\tau_k\}$ are positive sequences that $\alpha_k^{\min} > 0$ for all $k \in \mathbb{N}$. Hence, considering (24) and (26), to prove that $0 < \alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$ and $0 < \alpha_k^{\min} \leq \alpha_k^{\max}$ for all $k \in \mathbb{N}$, it is sufficient to show that $\alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$ for all $k \in \mathbb{N}$. Consider arbitrary $k \in \mathbb{N}$. Because $\alpha_k^{\varphi} \geq 0$ by construction and $\alpha_k^{\text{suff}} \geq 0$ as a consequence of Lemmas 2 and 3, the inequality holds trivially if $\alpha_k^{\text{suff}} = 0$. Hence, we may proceed under the assumption that $\alpha_k^{\text{suff}} > 0$. Moreover, one finds from the definition of α_k^{φ} that to establish $\alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$, it is sufficient to show that $\varphi(\alpha_k^{\text{suff}}) \leq 0$. We consider two cases based on which term yields the minimum in (21). First, suppose that $\alpha_k^{\text{suff}} = 1 \leq \frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2}$, which with (25), shows that

$$\begin{aligned}\varphi(\alpha_k^{\text{suff}}) &= (\eta - 1)\beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \frac{1}{2}(\tau_k L + \Gamma)\|d_k\|^2\\ &\leq (\eta - 1)\beta_k \Delta l(x_k, \tau_k, g_k, d_k) + (1 - \eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k) = 0,\end{aligned}$$

as desired. Second, suppose $\alpha_k^{\text{suff}} = \frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2} < 1$. For this case, it follows from (25), $\alpha_k^{\text{suff}} \in (0, 1]$, and the triangle inequality that

$$\begin{aligned}\varphi(\alpha_k^{\text{suff}}) &= (\eta - 1)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + (1 - \eta)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k)\\ &\quad + \|c_k + \alpha_k^{\text{suff}} \nabla c_k^T d_k\| - \alpha_k^{\text{suff}} \|c_k + \nabla c_k^T d_k\| + (\alpha_k^{\text{suff}} - 1)\|c_k\|\\ &\leq \|(1 - \alpha_k^{\text{suff}})c_k\| + (\alpha_k^{\text{suff}} - 1)\|c_k\| = 0.\end{aligned}$$

Overall, $\alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$ because in both cases, we proved that $\varphi(\alpha_k^{\text{suff}}) \leq 0$.

Finally, let us show $\varphi(\alpha_k) \leq 0$ for all $k \in \mathbb{N}$. By (4) and (25), one finds (as previously mentioned) that φ is strongly convex. In addition, one finds that $\varphi(0) = \varphi(\alpha_k^{\varphi}) = 0$, where $\alpha_k^{\varphi} \in \mathbb{R}_{>0}$. Along with $0 < \alpha_k^{\min} \leq \alpha_k \leq \alpha_k^{\max} \leq \alpha_k^{\varphi}$, it follows that $\varphi(\alpha_k) \leq 0$, as desired. □

3.2. General Results

In this subsection, we prove general results about the behavior of Algorithm 1. As in the previous subsection, our analysis here merely requires for all $k \in \mathbb{N}$ that $g_k \in \mathbb{R}^n$ and $H_k \in \mathbb{S}^n$ satisfy Assumption 2. The next lemma gives a lower bound on $\|c_k\| - \|c_k + \nabla c_k^T v_k\|$ relative to $\|c_k\|$.

Lemma 5. *There exists $\kappa_1 \in \mathbb{R}_{>0}$ (a constant common to all runs of the algorithm) such that for all $k \in \mathbb{N}$ in any run, one has that $\|c_k\| - \|c_k + \nabla c_k^T v_k\| \geq \kappa_1 \|c_k\|$.*

Proof. This result follows as in Curtis et al. (2009, lemma 3.5) but with small straightforward modifications to account for the fact that in our analysis here, the singular values of $\{\nabla c_k^T\}$ are bounded away from zero uniformly over all runs as part of Assumption 1. □

The next lemma shows that $\|v_k\|$ is bounded below and above proportionally to $\|c_k\|$.

Lemma 6. *There exists* $(\kappa_2, \kappa_3) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ (*constants common to all runs of the algorithm*) *such that for all* $k \in \mathbb{N}$ *in any run, one has that* $\kappa_2\|c_k\| \leq \|v_k\| \leq \kappa_3\|c_k\|$.

Proof. Assumption 1 ensures the existence of $\lambda_{\min} \in \mathbb{R}_{>0}$ such that $\nabla c_k^T \nabla c_k \succeq \lambda_{\min} I$ for all $k \in \mathbb{N}$ in any run. We now prove each desired inequality. First, consider the former inequality. Because this holds trivially when $c_k = 0$, let us proceed under the assumption that $c_k \neq 0$. One finds

$$\|c_k\|^2 - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|^2 = (\|c_k\| - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|)(\|c_k\| + \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|)$$
$$\leq 2\|c_k\|(\|c_k\| - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|).$$

It follows from this inequality, the triangle inequality, and (6) that

$$\|\nabla c_k^T\| \, \|v_k\| \geq \|\nabla c_k^T v_k\| \geq \|c_k\| - \|c_k + \nabla c_k^T v_k\| \geq \epsilon_c(\|c_k\| - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|)$$
$$\geq \frac{\epsilon_c}{2\|c_k\|}(\|c_k\|^2 - \|c_k + \alpha_k^c \nabla c_k^T v_k^c\|^2) = \frac{\epsilon_c}{2\|c_k\|}(-2\alpha_k^c c_k^T \nabla c_k^T v_k^c - (\alpha_k^c)^2 \|\nabla c_k^T v_k^c\|^2).$$

Substituting in for the value of α_k^c (recall (7)) and $v_k^c = -\nabla c_k c_k$ shows that $\|\nabla c_k^T\| \, \|v_k\| \geq \left(\frac{\epsilon_c}{2\|c_k\|}\right)\alpha_k^c \|\nabla c_k c_k\|^2$. Again, substituting the value of α_k^c and using the definition of $\lambda_{\min}$, one finds

$$\|\nabla c_k^T\| \, \|v_k\| \geq \frac{\epsilon_c \|\nabla c_k c_k\|^4}{2\|c_k\| \, \|\nabla c_k^T \nabla c_k c_k\|^2} \geq \frac{\epsilon_c \lambda_{\min}^2 \|c_k\|^4}{2\|c_k\| \, \|\nabla c_k^T \nabla c_k\|^2 \|c_k\|^2} = \frac{\epsilon_c \lambda_{\min}^2}{2\|\nabla c_k^T \nabla c_k\|^2}\|c_k\|.$$

It follows from these inequalities and Assumption 1 that there exists $\kappa_2 \in \mathbb{R}_{>0}$ as claimed.

Let us now turn to the latter inequality. It follows from the normal direction computation that $\|c_k\| \geq \|c_k + \nabla c_k^T v_k\|$, which implies that $\|\nabla c_k^T v_k\| \leq 2\|c_k\|$. Note that because $v_k \in \text{Range}(\nabla c_k)$, one has $v_k = \nabla c_k w_k$, where $w_k = (\nabla c_k^T \nabla c_k)^{-1} \nabla c_k^T v_k$. Putting these facts together shows that

$$\|v_k\| = \|\nabla c_k w_k\| = \|\nabla c_k (\nabla c_k^T \nabla c_k)^{-1} \nabla c_k^T v_k\| \leq \|\nabla c_k\| \, \|(\nabla c_k^T \nabla c_k)^{-1}\| \, \|\nabla c_k^T v_k\| \leq \frac{2\|\nabla c_k\|}{\lambda_{\min}}\|c_k\|,$$

which combined with Assumption 1—namely, that the Jacobian function $\nabla c(\cdot)^T$ is bounded over the set $\mathcal{X}$ containing the iterates—establishes the existence of $\kappa_3 \in \mathbb{R}_{>0}$ as claimed. □

The next result gives a useful bound on the size of the search direction.

Lemma 7. *There exists* $\kappa_4 \in \mathbb{R}_{\geq 2}$ (*a constant common to all runs of the algorithm*) *such that for all* $k \in \mathbb{N}$ *in any run, one finds that* $\|d_k\|^2 \leq \kappa_4(\|u_k\|^2 + \|c_k\|)$.

Proof. Observe that $0 \leq (\|u_k\| - \|v_k\|)^2 = \|u_k\|^2 + \|v_k\|^2 - 2\|u_k\| \, \|v_k\|$. Using this fact, $d_k = v_k + u_k$, the triangle inequality, and Lemma 6, it follows that

$$\|d_k\|^2 \leq (\|u_k\| + \|v_k\|)^2 = \|u_k\|^2 + \|v_k\|^2 + 2\|u_k\| \, \|v_k\|$$
$$\leq 2(\|u_k\|^2 + \|v_k\|^2) \leq 2(\|u_k\|^2 + \kappa_3^2\|c_k\|^2) \leq \max\{2, 2\kappa_3^2\|c_k\|\}(\|u_k\|^2 + \|c_k\|).$$

The existence of the required $\kappa_4 \in \mathbb{R}_{\geq 2}$ now follows from Assumption 1 because $\max\{2, 2\kappa_3^2\|c_k\|\}$ is uniformly bounded for all $k \in \mathbb{N}$ in any run, which completes the proof. □

The next lemma shows that the model reduction $\Delta l(x_k, \tau_k, g_k, v_k + u_k)$ is bounded below by a similar quantity as the upper bound for $\|d_k\|^2$ in the previous lemma.

Lemma 8. *There exists* $\kappa_5 \in \mathbb{R}_{>0}$ (*a constant common to all runs of the algorithm*) *such that for all* $k \in \mathbb{N}$ *in any run, one has that* $\Delta l(x_k, \tau_k, g_k, v_k + u_k) \geq \kappa_5\tau_k(\|u_k\|^2 + \|c_k\|) \geq \frac{\kappa_5\tau_k}{\kappa_4}\|d_k\|^2 > 0$.

Proof. Lemma 3 shows that (12) holds with $\tau \equiv \tau_k$. Combining this fact with Lemma 5 and the monotonically nonincreasing behavior of $\{\tau_k\}$ shows that

$$\Delta l(x_k, \tau_k, g_k, v_k + u_k) \geq \sigma_u \tau_k \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + \nabla c_k^T v_k\|)$$
$$\geq \sigma_u \tau_k \epsilon_u \|u_k\|^2 + \sigma_c \kappa_1 \|c_k\| \geq \min\left\{\sigma_u \epsilon_u, \frac{\sigma_c \kappa_1}{\tau_{-1}}\right\}\tau_k(\|u_k\|^2 + \|c_k\|),$$

which proves the existence of the claimed $\kappa_5 \in \mathbb{R}_{>0}$ because σ_u, ϵ_u, σ_c, κ_1, and τ_{-1} are positive real numbers. The remaining inequalities follow from Lemmas 2 and 7. □

We next prove a lower bound for $\{\xi_k\}$ that is uniform over every run of the algorithm.

Lemma 9. *There exists $\xi_{\min} \in \mathbb{R}_{>0}$ (a constant common to all runs of the algorithm) such that in any run, there exists $k_\xi \in \mathbb{N}$ and $\xi_{k_\xi} \in [\xi_{\min}, \infty)$ such that $\xi_k = \xi_{k_\xi}$ for all $k \geq k_\xi$.*

Proof. For all $k \in \mathbb{N}$, it follows from (22) and Lemmas 7 and 8 that

$$\xi_k^{\text{trial}} = \frac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|^2} \geq \frac{\kappa_5 \tau_k (\|u_k\|^2 + \|c_k\|)}{\tau_k \kappa_4 (\|u_k\|^2 + \|c_k\|)} = \frac{\kappa_5}{\kappa_4}. \tag{30}$$

Now, consider any iteration such that $\xi_k < \xi_{k-1}$. For such iterations, it follows from (22) and (30) that $\xi_k \geq (1-\epsilon_\xi)\xi_k^{\text{trial}} \geq (1-\epsilon_\xi)\kappa_5/\kappa_4$. Combining this fact with the initial choice of ξ_{-1} shows that $\xi_k \geq \xi_{\min} := \min\{(1-\epsilon_\xi)\kappa_5/\kappa_4, \xi_{-1}\}$ for all $k \in \mathbb{N}$. Combining this result with the fact that $\xi_k < \xi_{k-1}$ implies that $\xi_k \leq (1-\epsilon_\xi)\xi_{k-1}$ (it decreases by at least a factor of $1-\epsilon_\xi$) gives the desired result. □

The next lemma gives a bound on the change in the merit function in each iteration.

Lemma 10. *For all $k \in \mathbb{N}$ in any run, one has that*

$$\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \\
&\leq -\alpha_k \Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) + \alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}}) + (1-\eta)\alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) \\
&\quad + \alpha_k \|c_k + \nabla c_k^T d_k\| - \alpha_k \|c_k + \nabla c_k^T v_k\|.
\end{aligned}$$

Proof. By Lemma 4, one has $\varphi(\alpha_k) \leq 0$. Hence, starting as in (20); adding and subtracting the terms $\alpha_k \tau_k \nabla f_k^T d_k^{\text{true}}$, $\alpha_k \|c_k\|$, $\alpha_k \|c_k + \nabla c_k^T d_k^{\text{true}}\|$, and $\alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k)$; using the definition of $\varphi(\cdot)$; and using the fact that $\nabla c_k^T d_k^{\text{true}} = \nabla c_k^T (v_k + u_k^{\text{true}}) = \nabla c_k^T v_k$, one finds that

$$\begin{aligned}
&\phi(x + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \\
&\leq \alpha_k \tau_k \nabla f_k^T d_k + \|c_k + \alpha_k \nabla c_k^T d_k\| - \|c_k\| + \frac{1}{2}(\tau_k L + \Gamma)\alpha_k^2 \|d_k\|^2 \\
&= -\alpha_k \Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) + \alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}}) + (\alpha_k - 1)\|c_k\| \\
&\quad + \|c_k + \alpha_k \nabla c_k^T d_k\| - \alpha_k \|c_k + \nabla c_k^T d_k^{\text{true}}\| + \frac{1}{2}(\tau_k L + \Gamma)\alpha_k^2 \|d_k\|^2 \\
&\quad - \alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) \\
&\leq -\alpha_k \Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) + \alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}}) + \alpha_k \|c_k + \nabla c_k^T d_k\| \\
&\quad - \alpha_k \|c_k + \nabla c_k^T d_k^{\text{true}}\| - \eta \alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) \\
&= -\alpha_k \Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) + \alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}}) + (1-\eta)\alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k) \\
&\quad + \alpha_k \|c_k + \nabla c_k^T d_k\| - \alpha_k \|c_k + \nabla c_k^T v_k\|,
\end{aligned}$$

which completes the proof. □

3.3. Convergence Analysis

Our goal in this subsection is to prove a convergence result for our algorithm. In general, in a run of the algorithm, one of three possible events can occur with respect to the merit parameter sequence. One possible event is that the merit parameter sequence eventually remains constant at a value that is *sufficiently small*. This is the event that we consider in our analysis here, where the meaning of *sufficiently small* is defined formally in the event $\mathcal{E}$ that is introduced shortly. The other two possible events are that the merit parameter sequence vanishes or eventually remains constant at a value that is too large. As discussed in Berahas et al. (2021, section 3.2.2), the former of these two events does not occur for the algorithm in that paper if the differences between the stochastic gradient estimates and the true gradients of the objective are uniformly bounded in norm; in particular, see Berahas et al. (2021, proposition 3.18). It is straightforward to show that such a conclusion also holds for Algorithm 1 in this paper because the merit

parameter update strategy follows the same kind of approach as for the algorithm in Berahas et al. (2021); in particular, see the consistency between Berahas et al. (2021, equations (3.3) and (3.4)) and in this paper, (17) and (18) as well as Byrd et al. (2008, lemma 4.7), which considers the setting of inexact linear system solutions using inexactness tolerance conditions of the same type as in this paper. Moreover, in Berahas et al. (2021, section 3.2.2), it is shown that the latter type of event (namely, that the merit parameter remains constant at a value that is too large) occurs with probability of zero if one makes a reasonable assumption about the influence of the stochastic gradient estimates on the computed search directions; see also Berahas et al. (2023, section 4.3) for additional discussion of this case in the context of an algorithm that employs a step decomposition approach, like in Algorithm 1. Again, it is straightforward to see that such a conclusion also holds for Algorithm 1 because the merit parameter update strategy is of the same form. Consequently, for our purposes here, we do not consider these latter events because we contend that for practical purposes, one can focus on the first event for the same reasons as in Berahas et al. (2021, 2023).

Our main convergence result for Algorithm 1 considers an assumption that combines all of the assumptions required for our analysis until this point and assumes certain behavior of the merit parameter sequence through an event denoted as $\mathcal{E}$. For this event and the subsequent analysis, recall the stochastic process (3) defined by the algorithm. Consider for each $k \in \mathbb{N}$ the condition

$$\nabla f(X_k)^T D_k^{\text{true}} + \max\{(U_k^{\text{true}})^T H_k U_k^{\text{true}}, \epsilon_u \|U_k^{\text{true}}\|^2\} \leq 0, \tag{31}$$

similar to the one appearing in (18). (For the sake of brevity in our notation, we overload the meaning of H_k; here, it may be a random variable satisfying Assumption 6 introduced shortly, which is consistent with the previously introduced and employed Assumption 2.) With this condition, let us define the following trial value of the merit parameter that would be computed in iteration $k \in \mathbb{N}$ if the algorithm was to employ $\nabla f(X_k)$ in place of G_k and solve (8) exactly:

$$\mathcal{T}_k^{\text{trial,true}} \leftarrow \begin{cases} \infty & \text{if (31) holds,} \\ \dfrac{\left(1 - \frac{\sigma_c}{\epsilon_r}\right)(\|c(X_k)\| - \|c(X_k) + \nabla c(X_k)^T D_k^{\text{true}}\|)}{\nabla f(X_k)^T D_k^{\text{true}} + \max\{(U_k^{\text{true}})^T H_k U_k^{\text{true}}, \epsilon_u \|U_k^{\text{true}}\|^2\}} & \text{otherwise.} \end{cases}$$

(To be clear, the quantity $\mathcal{T}_k^{\text{trial,true}}$ never needs to be computed by our algorithm; it is only used in our analysis.) Using this quantity, we define our event of interest, namely $\mathcal{E}$, as the following.

3.3.1. Event $\mathcal{E}$. For some $(k_{\min}, \tau_{\min}, f_{\sup}) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}$, the event $\mathcal{E} := \mathcal{E}(k_{\min}, \tau_{\min}, f_{\sup})$ occurs if and only if $f(X_{k_{\min}}) \leq f_{\sup}$, and there exists $(K, \mathcal{T}', \Xi') \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $K' \leq k_{\min}$, $\mathcal{T}' \geq \tau_{\min}$, and $\Xi' \geq \xi_{\min}$ (see Lemma 9) such that

$$\mathcal{T}_k = \mathcal{T}' \leq \mathcal{T}_k^{\text{trial,true}} \text{ and } \Xi_k = \Xi' \text{ for all } k \in \mathbb{N} \text{ with } k \geq K'. \tag{32}$$

In other words, event $\mathcal{E}$ is the event in which by iteration $k_{\min}$, the merit and ratio parameter sequences become constant at values at least $\tau_{\min}$ and $\xi_{\min}$, respectively, and the objective value at iteration $k_{\min}$ is bounded. With respect to this event, we make the following assumption for the rest of our analysis, our ultimate focus of which will be on the behavior of the algorithm starting in iteration $k_{\min}$, at which point the adaptive merit and ratio parameters are constant.

Assumption 5. *For some $(k_{\min}, \tau_{\min}, f_{\sup}) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}$, the event $\mathcal{E} := \mathcal{E}(k_{\min}, \tau_{\min}, f_{\sup})$ occurs, and conditioned on the occurrence of $\mathcal{E}$, Assumptions 1, 2, 3, and 4 hold. In addition, along with the restrictions that $\{\beta_k\} \subset (0, 1]$ and (23) holds for all $k \in \mathbb{N}$, the sequence $\{\beta_k\}_{k \geq k_{\min}}$ is chosen in a manner that is $\mathcal{F}_{k_{\min}}$-measurable.*

A few remarks are in order with respect to Assumption 5. First, that the event $\mathcal{E}$ includes that the merit parameter sequence is bounded below can, as previously mentioned, be justified for the same reasons as in Berahas et al. (2021); the additional requirement that it eventually remains *constant* can be justified by Lemma 3, which shows that if the merit parameter is decreased, then it is decreased by at least a constant factor. Note that it is the inequality $\mathcal{T}' \leq \mathcal{T}_k^{\text{trial,true}}$ that represents the aforementioned notion of the merit parameter ultimately being *sufficiently small* for all large k. Second, that $\mathcal{E}$ includes that the ratio parameter sequence is bounded below and eventually remains constant is not a strong assumption; it follows under our prior assumptions (that are carried forward in Assumption 5) because of Lemma 9. That said, the critical aspect here is that $\mathcal{E}$ requires that this sequence has become constant by iteration $k_{\min}$. Third, that $\mathcal{E}$ includes that $f(X_{k_{\min}})$ is bounded above is a relatively weak assumption but

necessary for our purposes of ultimately showing that a sequence of stationarity measures vanishes in expectation. Finally, with respect to $\{\beta_k\}_{k\geq k_{\min}}$, we state in Theorem 1 particular choices satisfying Assumption 5 for which our convergence guarantees hold. Precise strategies for setting these values that are consistent with our convergence guarantees are stated after Theorem 1.

Let $\mathcal{G}_0$ be the σ-algebra defined by the initial conditions of Algorithm 1, and for all $k \in \mathbb{N}$, let $\mathcal{G}_k$ be the σ-algebra generated by the initial conditions and $\{G_0, \ldots, G_{k-1}\}$. In addition, for all $k \in \mathbb{N}$, let the trace σ-algebra of $\mathcal{E}$ on $\mathcal{G}_k$ be $\mathcal{F}_k := \mathcal{G}_k \cap \mathcal{E}$. Hence, $\{\mathcal{F}_k\}$ is a filtration. For brevity, let

$$\mathbb{P}_k[\cdot] := \mathbb{P}_\omega[\cdot \,|\, \mathcal{F}_k] \quad \text{and} \quad \mathbb{E}_k[\cdot] := \mathbb{E}_\omega[\cdot \,|\, \mathcal{F}_k],$$

where $\mathbb{P}_\omega$ denotes probability with respect to the distribution of ω (and as for (1), $\mathbb{E}_\omega$ denotes expectation with respect to the distribution of ω). Observe that conditioned on $\mathcal{E}$, one has that

$$\tau_{\min} \leq \mathcal{T}' \leq \tau_{-1} \text{ and } \xi_{\min} \leq \Xi' \leq \xi_{-1}, \tag{33}$$

and one has that the random variables $\mathcal{T}'$ and Ξ' are $\mathcal{F}_k$-measurable for $k = k_{\min} \geq K'$.

We make Assumption 6 about $\{G_k\}$ and $\{H_k\}$. That the stochastic gradient estimators are unbiased is standard for algorithms based on stochastic approximation. One may be able to relax the so-called bounded-variance assumption introduced here, but we contend that this assumption is sufficient for showing the general type of convergence guarantee that our algorithm offers. Hence, we make a bounded-variance assumption here so as not to obfuscate the other details.

Assumption 6. *There exists $M_g \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$, the gradient estimator G_k has that $\mathbb{E}_k[G_k] = \nabla f(X_k)$ and $\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2] \leq M_g$. In addition, for all $k \in \mathbb{N}$, the matrix H_k (satisfying Assumption 5; i.e., the bounds in Assumption 2) is $\mathcal{F}_k$-measurable.*

Combining Assumption 6 with Jensen's inequality, it holds for all $k \in \mathbb{N}$ that

$$\mathbb{E}_k[\|G_k - \nabla f(X_k)\|] \leq \sqrt{\mathbb{E}_k[\|G_k - \nabla f(X_k)\|^2]} \leq \sqrt{M_g}. \tag{34}$$

We now return to our analysis. First, let us derive bounds on the expected difference between U_k and U_k^{true}. To that end, let us define $Z_k \in \mathbb{R}^{n\times(n-m)}$ as an ($\mathcal{F}_k$-measurable) matrix whose columns form an orthonormal basis for $\text{Null}(\nabla c(X_k)^T)$, which implies that $Z_k^T Z_k = I$ and $\nabla c(X_k)^T Z_k = 0$. Under Assumption 5 (namely, Assumption 1), let $U_{k,1} \in \mathbb{R}^m$ and $U_{k,2} \in \mathbb{R}^{n-m}$ be vectors forming the orthogonal decomposition of U_k into $\text{Range}(\nabla c(X_k))$ and $\text{Null}(\nabla c(X_k)^T)$ in the sense that $U_k = \nabla c(X_k) U_{k,1} + Z_k U_{k,2}$. It follows from (11) that $U_{k,1} = (\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k$ and $U_{k,2} = -(Z_k^T H_k Z_k)^{-1} Z_k^T (G_k + H_k V_k + H_k \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k - \mathfrak{R}_k)$, with which one can derive

$$\begin{aligned} U_k &= \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k \\ &\quad - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (G_k + H_k V_k + H_k \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k - \mathfrak{R}_k) \\ U_k^{\text{true}} &= - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(X_k) + H_k V_k). \end{aligned} \tag{35}$$

The corresponding values for $\mathfrak{D}_k$ and $\mathfrak{D}_k^{\text{true}}$ are found to be

$$\begin{aligned} \mathfrak{D}_k &= -(\nabla c(X_k)^T \nabla c(X_k))^{-1} \nabla c(X_k)^T (G_k + H_k V_k + H_k U_k - \mathfrak{R}_k) - Y_k \\ \mathfrak{D}_k^{\text{true}} &= -(\nabla c(X_k)^T \nabla c(X_k))^{-1} \nabla c(X_k)^T (\nabla f(X_k) + H_k V_k + H_k U_k^{\text{true}}) - Y_k. \end{aligned} \tag{36}$$

In the proof of our next lemma, we use the fact that

$$\|I - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k\| \leq 1, \tag{37}$$

which can be seen as follows. The nonzero eigenvalues of AB are equal to those of BA when the products are valid, meaning that the nonzero eigenvalues of $Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k$ equal those of $Z_k^T H_k Z_k (Z_k^T H_k Z_k)^{-1} = I$, which are all one; hence, the bound in (37) holds.

Lemma 11. *There exists $\kappa_6 \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$ with $k \geq k_{\min}$, one finds*

$$\|\mathbb{E}_k[U_k - U_k^{\text{true}}]\| \leq \kappa_6 \beta_k \text{ and } \mathbb{E}_k[\|U_k - U_k^{\text{true}}\|] \leq \zeta^{-1}\sqrt{M_g} + \kappa_6 \beta_k.$$

Proof. It follows from (35) that

$$\begin{aligned} U_k - U_k^{\text{true}} &= \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k \\ &\quad - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (G_k - \nabla f(X_k) + H_k \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k - \mathfrak{R}_k), \end{aligned}$$

which combined with Assumption 6, shows that

$$\begin{aligned} &\mathbb{E}_k[U_k - U_k^{\text{true}}] \\ &= (I - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k) \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} \mathbb{E}_k[R_k] + Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T \mathbb{E}_k[\mathfrak{R}_k]. \end{aligned}$$

Combining this equation with the triangle inequality, Assumption 5 (specifically, Assumptions 1 and 2), (14), and (37) ensures the existence of $\kappa_6 \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$,

$$\begin{aligned} \|\mathbb{E}_k[U_k - U_k^{\text{true}}]\| &\leq \|\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1}\| \, \|\mathbb{E}_k[R_k]\| + \zeta^{-1} \|\mathbb{E}_k[\mathfrak{R}_k]\| \\ &\leq \|\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1}\| \kappa_r \beta_k + \zeta^{-1} \kappa_\rho \beta_k \leq \kappa_6 \beta_k, \end{aligned}$$

which is the first desired result. Next, to derive the desired bound on $\mathbb{E}_k[\|U_k - U_k^{\text{true}}\|]$, one can combine the expression for $U_k - U_k^{\text{true}}$ with the triangle inequality to obtain

$$\begin{aligned} \|U_k - U_k^{\text{true}}\| &\leq \|Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (G_k - \nabla f(X_k))\| + \|Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T \mathfrak{R}_k\| \\ &\quad + \|(I - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k) \nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k\|. \end{aligned}$$

Taking conditional expectation and using Assumption 6, (34), (37), and (14), one finds

$$\begin{aligned} \mathbb{E}_k[\|U_k - U_k^{\text{true}}\|] &\leq \zeta^{-1}\sqrt{M_g} + \zeta^{-1} \mathbb{E}_k[\|\mathfrak{R}_k\|] + \|\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1}\| \mathbb{E}_k[\|R_k\|] \\ &\leq \zeta^{-1}\sqrt{M_g} + \zeta^{-1} \kappa_\rho \beta_k + \|\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1}\| \kappa_r \beta_k \leq \zeta^{-1}\sqrt{M_g} + \kappa_6 \beta_k, \end{aligned}$$

where κ_6 is the same value as used, which completes the proof. □

We now bound the difference in expectation between $\nabla f(X_k)^T D_k^{\text{true}}$ and $G_k^T D_k$.

Lemma 12. *There exists $(\kappa_7, \kappa_8) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$ with $k \geq k_{\min}$, one finds*

$$|\mathbb{E}_k[\nabla f(X_k)^T D_k^{\text{true}} - G_k^T D_k]| \leq \kappa_7 \beta_k + \kappa_8 \beta_k \sqrt{M_g} + \zeta^{-1} M_g.$$

Proof. It follows from the triangle inequality that

$$\begin{aligned} |\mathbb{E}_k[\nabla f(X_k)^T D_k^{\text{true}} - G_k^T D_k]| &= |\mathbb{E}_k[\nabla f(X_k)^T (D_k^{\text{true}} - D_k) + (\nabla f(X_k) - G_k)^T D_k]| \\ &\leq |\nabla f(X_k)^T \mathbb{E}_k[D_k^{\text{true}} - D_k]| + |\mathbb{E}_k[(\nabla f(X_k) - G_k)^T D_k]|. \end{aligned}$$

For the first term on the right-hand side, the Cauchy–Schwarz inequality, $D_k^{\text{true}} = V_k + U_k^{\text{true}}$, $D_k = V_k + U_k$, Lemma 11, and Assumption 5 (i.e., Assumption 1) imply that there exists $\kappa_7 \in \mathbb{R}_{>0}$ with

$$|\nabla f(X_k)^T \mathbb{E}_k[D_k^{\text{true}} - D_k]| \leq \|\nabla f(X_k)\| \|\mathbb{E}_k[D_k^{\text{true}} - D_k]\| = \|\nabla f(X_k)\| \|\mathbb{E}_k[U_k^{\text{true}} - U_k]\| \leq \kappa_7 \beta_k.$$

Now, for the second term, first observe from Assumption 6 that $\mathbb{E}_k[(\nabla f(X_k) - G_k)^T V_k] = V_k^T \mathbb{E}_k[\nabla f(X_k) - G_k] = 0$. Combining this fact with (35), the Cauchy–Schwarz inequality, Assumption 5 (namely, Assumptions 1 and 2),

(37), and (34) shows that there exists $(\overline{\kappa}_8, \kappa_8) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ giving

$$\begin{aligned}
&|\mathbb{E}_k[(\nabla f(X_k) - G_k)^T D_k]| \\
&= |\mathbb{E}_k[(\nabla f(X_k) - G_k)^T ((I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k)\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k \\
&\quad - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (G_k - \nabla f(X_k) - \mathfrak{R}_k))]| \\
&\le |\mathbb{E}_k[(\nabla f(X_k) - G_k)^T (I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k)\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1} R_k]| \\
&\quad + |\mathbb{E}_k[(\nabla f(X_k) - G_k)^T Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T \mathfrak{R}_k]| \\
&\quad + |\mathbb{E}_k[(\nabla f(X_k) - G_k)^T Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(X_k) - G_k)]| \\
&\le \mathbb{E}_k[\|\nabla f(X_k) - G_k\| \, \|(I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k)\nabla c(X_k)(\nabla c(X_k)^T \nabla c(X_k))^{-1}\| \, \|R_k\|] \\
&\quad + \mathbb{E}_k[\|\nabla f(X_k) - G_k\| \, \|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\| \, \|\mathfrak{R}_k\|] + \zeta^{-1}\mathbb{E}_k[\|\nabla f(X_k) - G_k\|^2] \\
&\le (\overline{\kappa}_8 \kappa_r + \zeta^{-1}\kappa_\rho)\beta_k \sqrt{M_g} + \zeta^{-1} M_g = \kappa_8 \beta_k \sqrt{M_g} + \zeta^{-1} M_g.
\end{aligned}$$

Combining the results gives the desired result. □

We now proceed to bound in expectation the last few terms appearing in the right-hand side of the inequality proved in Lemma 10. The next lemma considers the last pair of terms.

Lemma 13. *There exists $\kappa_9 \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$ with $k \ge k_{\min}$, one finds*

$$\mathbb{E}_k[\mathcal{A}_k(\|c(X_k) + \nabla c(X_k)^T D_k\| - \|c(X_k) + \nabla c(X_k)^T V_k\|)] \le \kappa_9 \beta_k^2.$$

Proof. From (11), (14), the fact that $\mathcal{A}_k \in [\mathcal{A}_k^{\min}, \mathcal{A}_k^{\max}]$, (26), (24), (23), and the monotonically nonincreasing behavior of $\{\mathcal{T}_k\}$ and $\{\Xi_k\}$, it follows that there exists $\kappa_9 \in \mathbb{R}_{>0}$ such that

$$\begin{aligned}
&\mathbb{E}_k[\mathcal{A}_k(\|c(X_k) + \nabla c(X_k)^T D_k\| - \|c(X_k) + \nabla c(X_k)^T V_k\|)] \le \mathbb{E}_k[\mathcal{A}_k \|\nabla c(X_k)^T U_k\|] = \mathbb{E}_k[\mathcal{A}_k \|R_k\|] \\
&\le \kappa_r \beta_k \mathbb{E}_k[\mathcal{A}_k^{\max}] \le \kappa_r \beta_k \mathbb{E}_k[\mathcal{A}_k^{\min} + \theta\beta_k^2] = \kappa_r \beta_k \mathbb{E}_k\left[\left(\frac{2(1-\eta)\beta_k \Xi_k \mathcal{T}_k}{\mathcal{T}_k L + \Gamma} + \theta\beta_k^2\right)\right] \\
&\le \kappa_r \beta_k^2 \left(\frac{2(1-\eta)\xi_{-1}\tau_{-1}}{\Gamma} + \theta\beta_k\right) \le \kappa_9 \beta_k^2,
\end{aligned}$$

which gives the desired conclusion. □

Our next result provides an upper bound in expectation for the second term appearing on the right-hand side of the inequality in Lemma 10.

Lemma 14. *There exists $\kappa_{10} \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$ with $k \ge k_{\min}$, one finds*

$$\mathbb{E}_k[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})] \le \kappa_{10} \beta_k^2.$$

Proof. Let $\mathcal{I}_k$ be the event that $\nabla f(X_k)^T (D_k - D_k^{\text{true}}) \ge 0$, and let $\mathcal{I}_k^c$ be its complementary event. It follows from (32), the definition of $\mathcal{I}_k$, the fact that $\mathcal{A}_k \in [\mathcal{A}_k^{\min}, \mathcal{A}_k^{\max}]$, and the law of total expectation that for all $k \ge k_{\min}$, one finds

$$\begin{aligned}
\mathbb{E}_k[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})] &= \mathbb{E}_k[\mathcal{A}_k \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{I}_k] \mathbb{P}_k[\mathcal{I}_k] \\
&\quad + \mathbb{E}_k[\mathcal{A}_k \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{I}_k^c] \mathbb{P}_k[\mathcal{I}_k^c] \\
&\le \mathbb{E}_k[\mathcal{A}_k^{\max} \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{I}_k] \mathbb{P}_k[\mathcal{I}_k] \\
&\quad + \mathbb{E}_k[\mathcal{A}_k^{\min} \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{I}_k^c] \mathbb{P}_k[\mathcal{I}_k^c] \\
&= \mathbb{E}_k[(\mathcal{A}_k^{\max} - \mathcal{A}_k^{\min}) \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{I}_k] \mathbb{P}_k[\mathcal{I}_k] \\
&\quad + \mathbb{E}_k[\mathcal{A}_k^{\min} \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}})].
\end{aligned}$$

Combining this with the fact that (26) ensures $\mathcal{A}_k^{\max} - \mathcal{A}_k^{\min} \leq \theta\beta_k^2$, that $\mathcal{T}'$ and Ξ' are $\mathcal{F}_k$-measurable for $k \geq k_{\min}$, the Cauchy–Schwarz inequality, that $\mathcal{A}_k^{\min} = 2(1-\eta)\beta_k\Xi'\mathcal{T}'/(\mathcal{T}'L+\Gamma)$ for all $k \geq k_{\min}$, and the law of total expectation shows for all $k \geq k_{\min}$ that

$$\begin{aligned}\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^T(D_k - D_k^{\text{true}})] &\leq \theta\beta_k^2\mathcal{T}'\|\nabla f(X_k)\|\mathbb{E}_k[\|D_k - D_k^{\text{true}}\||\mathcal{I}_k]\mathbb{P}_k[\mathcal{I}_k] \\ &\quad + \frac{2(1-\eta)\beta_k\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}\mathcal{T}'\|\nabla f(X_k)\|\|\mathbb{E}_k[D_k - D_k^{\text{true}}]\| \\ &\leq \theta\beta_k^2\mathcal{T}'\|\nabla f(X_k)\|\mathbb{E}_k[\|D_k - D_k^{\text{true}}\|] \\ &\quad + \frac{2(1-\eta)\beta_k\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}\mathcal{T}'\|\nabla f(X_k)\|\|\mathbb{E}_k[D_k - D_k^{\text{true}}]\|.\end{aligned}$$

Combining this with Lemma 11, (23), $\|D_k - D_k^{\text{true}}\| = \|V_k + U_k - (V_k + U_k^{\text{true}})\| = \|U_k - U_k^{\text{true}}\|$, and Assumption 1 shows that there exists $\kappa_{10} \in \mathbb{R}_{>0}$ such that for all $k \geq k_{\min}$, one finds

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^T(D_k - D_k^{\text{true}})] \\ &\leq \theta\beta_k^2\mathcal{T}'\|\nabla f(X_k)\|(\zeta^{-1}\sqrt{M_g} + \kappa_6\beta_k) + \frac{2(1-\eta)\beta_k\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}\mathcal{T}'\|\nabla f(X_k)\|\kappa_6\beta_k \leq \kappa_{10}\beta_k^2,\end{aligned}$$

which is the desired conclusion. □

We now use the model reduction based on the true step D_k^{true} to show an upper bound on the expected reduction in the model based on the step D_k.

Lemma 15. *For all $k \in \mathbb{N}$ with $k \geq k_{\min}$, one finds*

$$\mathbb{E}_k[\Delta l(X_k, \mathcal{T}_k, G_k, D_k)] \leq \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \kappa_r\beta_k + \mathcal{T}'(\kappa_7\beta_k + \kappa_8\beta_k\sqrt{M_g} + \zeta^{-1}M_g).$$

Proof. It follows from Lemma 12; (4); the fact that $D_k = V_k + U_k$; the fact that $c(X_k)$, $\nabla c(X_k)^T$, V_k, $\nabla f(X_k)$, and D_k^{true} are all $\mathcal{F}_k$-measurable for $k \geq k_{\min}$; (9); and (14) that for all $k \geq k_{\min}$,

$$\begin{aligned}&\mathbb{E}_k[\Delta l(X_k, \mathcal{T}_k, G_k, D_k)] = \mathbb{E}_k[-\mathcal{T}'G_k^TD_k + \|c(X_k)\| - \|c(X_k) + \nabla c(X_k)^TD_k\|] \\ &\leq \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \kappa_r\beta_k + \mathcal{T}'(\kappa_7\beta_k + \kappa_8\beta_k\sqrt{M_g} + \zeta^{-1}M_g),\end{aligned}$$

which is the desired result. □

We now prove our main result. In the result, the quantity $\Delta l(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}})$ serves as a measure of stationarity with respect to (1); after all, the proof for Lemma 8 shows, with $(\nabla f(X_k), U_k^{\text{true}}, D_k^{\text{true}})$ in place of (G_k, U_k, D_k), that by Assumption 5, it follows for $k \geq k_{\min}$ that

$$\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \geq \kappa_5\mathcal{T}'(\|U_k^{\text{true}}\|^2 + \|c(X_k)\|) \geq \frac{\kappa_5\mathcal{T}'}{\kappa_4}\|D_k^{\text{true}}\|^2 \geq 0. \tag{38}$$

Thus, in a run, if there exists infinite $\mathcal{K} \subseteq \mathbb{N}$ with $\lim_{k\in\mathcal{K},k\to\infty}\Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) = 0$, then (38) and Lemma 6 imply that $\lim_{k\in\mathcal{K},k\to\infty}\|c_k\| = \lim_{k\in\mathcal{K},k\to\infty}\|u_k^{\text{true}}\| = \lim_{k\in\mathcal{K},k\to\infty}\|v_k\| = 0$, which combined with (9), shows that any limit of $\{(x_k, y_k + \delta_k^{\text{true}})\}$ is a first-order stationary point for (1). In our stochastic setting, we prove for two different choices of $\{\beta_k\}_{k\geq k_{\min}}$ that an expected average measure of stationarity exhibits desirable properties. These properties match those ensured by a stochastic gradient method in the unconstrained setting (where $\|\nabla f(X_k)\|^2$ is the measure of stationarity).

Theorem 1. *Define*

$$\mathcal{A}' := \frac{2(1-\eta)\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma},\ \alpha'_{\min} := \frac{2(1-\eta)\xi_{\min}\tau_{\min}}{\tau_{\min}L+\Gamma},\ \alpha'_{\max} := \frac{2(1-\eta)\xi_{-1}\tau_{-1}}{\tau_{-1}L+\Gamma}$$

and

$$M_{\max} = (1-\eta)(\alpha'_{\max} + \theta)(\kappa_r + \tau_{-1}(\kappa_7 + \kappa_8\sqrt{M_g} + \zeta^{-1}M_g)) + \kappa_9 + \kappa_{10}.$$

Then, the following results hold.

i. *If* $\beta_k = \beta = \frac{\psi \mathcal{A}'}{(1-\eta)(\mathcal{A}'+\theta)}$ *for some* $\psi \in (0,1)$ *for all* $k \geq k_{\min}$, *then*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{k}\sum_{j=k_{\min}}^{k_{\min}+k-1} \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}}) \middle| \mathcal{E}\right] \\
&\leq \frac{(1-\eta)(\alpha'_{\min}+\theta)(\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}] - \phi_{\min})}{k\psi(1-\psi)(\alpha'_{\min})^2} + \frac{\psi(\alpha'_{\max})^2(\alpha'_{\min}+\theta)M_{\max}}{(1-\eta)(1-\psi)(\alpha'_{\min})^2(\alpha'_{\max}+\theta)^2} \qquad (39) \\
&\overset{k\to\infty}{\to} \frac{\psi(\alpha'_{\max})^2(\alpha'_{\min}+\theta)M_{\max}}{(1-\eta)(1-\psi)(\alpha'_{\min})^2(\alpha'_{\max}+\theta)^2},
\end{aligned}$$

where $\phi_{\min} \in \mathbb{R}$ *is a lower bound for* $\phi(\cdot, \mathcal{T}')$ *over* $\mathcal{X}$ *by Assumption 5 (namely, Assumption 1).*

ii. *If* $\{\beta_k\}_{k \geq k_{\min}}$ *is determined by iteration* $k_{\min}$ *such that* $\sum_{k=k_{\min}}^{\infty} \beta_k = \infty$, $\sum_{k=k_{\min}}^{\infty} \beta_k^2 < \infty$, *and* $\beta_k \leq \frac{\psi \mathcal{A}'}{(1-\eta)(\mathcal{A}'+\theta)}$ *for some* $\psi \in (0,1)$ *for all* $k \geq k_{\min}$, *then*

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{1}{\sum_{j=k_{\min}}^{k_{\min}+k-1} \beta_j} \sum_{j=k_{\min}}^{k_{\min}+k-1} \beta_j \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}}) \middle| \mathcal{E}\right] = 0. \qquad (40)$$

In either case, if in a run, there exists $\mathcal{K} \subseteq \mathbb{N}$ *with* $|\mathcal{K}| = \infty$ *and* $\lim_{k\in\mathcal{K}, k\to\infty} \Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}}) = 0$, *then any limit point of* $\{(x_k, y_k + \delta_k^{\text{true}})\}$ *is a first-order stationary point for* (1).

Proof. By the definition of $\mathcal{A}'$, $\{\beta_k\} \subset (0,1]$, and line 15 of Algorithm 1, it follows that $\mathcal{A}_k \in [\mathcal{A}'\beta_k, (\mathcal{A}'+\theta)\beta_k]$ for all $k \geq k_{\min}$. It follows from this fact; $\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \geq 0$ (see (38)); Lemmas 10, 14, 8, 13, and 15; and the fact that $\{\beta_k\} \subset (0,1]$ that for all $k \geq k_{\min}$, one finds

$$\begin{aligned}
&\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \phi(X_k, \mathcal{T}_k) \\
&\leq \mathbb{E}_k[-\mathcal{A}_k \Delta l(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}}) + \mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})] \\
&\quad + (1-\eta)\mathbb{E}_k[\mathcal{A}_k \beta_k \Delta l(X_k, \mathcal{T}_k, G_k, D_k)] \\
&\quad + \mathbb{E}_k[\mathcal{A}_k(\|c(X_k) + \nabla c(X_k)^T D_k\| - \|c(X_k) + \nabla c(X_k)^T V_k\|)] \\
&\leq -\mathcal{A}'\beta_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + (\kappa_9 + \kappa_{10})\beta_k^2 \\
&\quad + (1-\eta)(\mathcal{A}'+\theta)\beta_k^2 \mathbb{E}_k[\Delta l(X_k, \mathcal{T}', G_k, D_k)] \\
&\leq -\beta_k(\mathcal{A}' - (1-\eta)(\mathcal{A}'+\theta)\beta_k)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \beta_k^2 M', \qquad (41)
\end{aligned}$$

where $M' := (1-\eta)(\mathcal{A}'+\theta)(\kappa_r + \mathcal{T}'(\kappa_7 + \kappa_8\sqrt{M_g} + \zeta^{-1}M_g)) + \kappa_9 + \kappa_{10}$. Observe that under Assumption 5, one has that $\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}]$ is bounded, $M' \leq M_{\max}$, and $\alpha'_{\min} \leq \mathcal{A}' \leq \alpha'_{\max}$ because of the monotonicity of $\frac{2(1-\eta)\xi\tau}{\tau L + \Gamma}$ with respect to both ξ and τ. Consider now the theorem's two cases.

Case i. By the definition of β, it follows that $\frac{\psi\alpha'_{\min}}{(1-\eta)(\alpha'_{\min}+\theta)} \leq \beta \leq \frac{\psi\alpha'_{\max}}{(1-\eta)(\alpha'_{\max}+\theta)}$ for all $k \geq k_{\min}$. Hence, along with (41), it follows for all $k \geq k_{\min}$ that

$$\begin{aligned}
&\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}')] - \phi(X_k, \mathcal{T}') \\
&\leq -\beta(\mathcal{A}' - (1-\eta)(\mathcal{A}'+\theta)\beta)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \beta^2 M' \\
&\leq -\left(\frac{\psi(1-\psi)(\alpha'_{\min})^2}{(1-\eta)(\alpha'_{\min}+\theta)}\right)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \left(\frac{\psi\alpha'_{\max}}{(1-\eta)(\alpha'_{\max}+\theta)}\right)^2 M_{\max}.
\end{aligned}$$

Then, taking total expectation conditioned only on the event $\mathcal{E}$, it follows for $k \in \mathbb{N}$ that

$$\begin{aligned}
&\phi_{\min} - \mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}] \\
&\leq \mathbb{E}[\phi(X_{k_{\min}+k}, \mathcal{T}') - \phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}] \\
&\leq -\left(\frac{\psi(1-\psi)(\alpha'_{\min})^2}{(1-\eta)(\alpha'_{\min}+\theta)}\right)\mathbb{E}\left[\sum_{j=k_{\min}}^{k_{\min}+k-1} \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}}) \middle| \mathcal{E}\right] + k\left(\frac{\psi\alpha'_{\max}}{(1-\eta)(\alpha'_{\max}+\theta)}\right)^2 M_{\max}.
\end{aligned}$$

After rearrangement, one finds that (39) holds, where the limit as $k \to \infty$ holds because of the aforementioned fact that $\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}]$ is bounded under Assumption 5.

Case ii. By the conditions on $\{\beta_k\}_{k \geq k_{\min}}$, it follows in a similar manner as in case (i) that

$$\begin{aligned}
&\phi_{\min} - \mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}] \\
&\leq \mathbb{E}[\phi(X_{k_{\min}+K}, \mathcal{T}') - \phi(X_{k_{\min}}, \mathcal{T}')|\mathcal{E}] \\
&\leq \mathbb{E}\left[\sum_{j=k_{\min}}^{k_{\min}+k-1} (-\beta_k(\mathcal{A}' - (1-\eta)(\mathcal{A}'+\theta)\beta_k)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \beta_k^2 M') \middle| \mathcal{E}\right],
\end{aligned}$$

which after rearrangement and taking limits as $k \to \infty$, proves that (40) holds.

The final conclusion of the theorem follows by the arguments provided before the theorem. □

We close this section by remarking that as described in Berahas et al. (2021), the elements of $\{\beta_k\}$ can be chosen to satisfy Assumption 5 and the conditions of Theorem 1. Specifically, to obtain the convergence guarantees in case (i) of Theorem 1, the algorithm can set $\beta_k \leftarrow \min\{1, \frac{\Gamma}{2(1-\eta)\xi_{-1}\tau_{-1}}, \frac{\psi\alpha'_k}{(1-\eta)(\alpha'_k+\theta)}\}$, where $\alpha'_k \leftarrow \frac{2(1-\eta)\xi_k\tau_k}{\tau_k L+\Gamma}$ for all $k \in \mathbb{N}$, which clearly ensures that $\beta_k \in (0,1]$ and that (23) holds. In addition, assuming that event $\mathcal{E}$ occurs, the value α'_k for sufficiently large k becomes the realization of $\mathcal{A}'$ stated in the theorem, in which case $\beta_k = \beta$ for all subsequent k satisfies the condition stated in the theorem. To obtain the convergence guarantees in case (ii) of Theorem 1, the algorithm can "reset" a diminishing sequence after each iteration, in which the merit parameter and/or the ratio parameter are decreased. Specifically, in any iteration, say $\hat{k} \in \mathbb{N}$, in which the merit parameter and/or the ratio parameter were reduced from the prior iteration, one can set $\beta_k \leftarrow c_{\hat{k}}/(k-\hat{k}+1)$ for all $k \geq \hat{k}$, where $c_{\hat{k}} := \min\{1, \frac{\Gamma}{2(1-\eta)\xi_{-1}\tau_{-1}}, \frac{\psi\alpha'_{\hat{k}}}{(1-\eta)(\alpha'_{\hat{k}}+\theta)}\}$ with $\frac{\alpha'_{\hat{k}} \leftarrow 2(1-\eta)\xi_{\hat{k}}\tau_{\hat{k}}}{\tau_{\hat{k}}L+\Gamma}$. If the value for $\hat{k}$ is reset after every time the merit parameter and/or the ratio parameter are decreased, under event $\mathcal{E}$, the value for $\hat{k}$ eventually will not be reset, meaning that in subsequent iterates, $\{\beta_k\}$ will satisfy the conditions of the theorem.

4. Numerical Results

In this section, we demonstrate the performance of a MATLAB implementation of Algorithm 1 for solving (i) a subset of the constrained and unconstrained testing environment with safe threads (CUTEst) set (Gould et al. 2015) and (ii) two optimal control problems from Hintermüller et al. (2003). The goal of our testing is to demonstrate the computational benefits of using inexact subproblem solutions obtained based on our termination tests from Section 2.2.

4.1. Iterative Solvers

To obtain the normal direction v_k as an inexact solution of (5), we applied CG to $\nabla c_k \nabla c_k^T v = -\nabla c_k c_k$. Denoting the tth CG iterate as $v_{k,t}$, where $v_{k,0} = 0$, the method sets $v_k \leftarrow v_{k,t}$, where t is the first CG iteration such that $\|\nabla c_k \nabla c_k^T v_{k,t} + \nabla c_k c_k\| \leq 10^{-8} \max\{\|\nabla c_k c_k\|, 1\}$. The properties of CG as a Krylov subspace method ensure that $v_{k,t} \in \text{Range}(\nabla c_k)$ for all $t \in \mathbb{N}$; hence, $v_k \in \text{Range}(\nabla c_k)$.

To obtain the tangential direction u_k and associated dual search direction δ_k, we applied the MINRES method (Paige and Saunders 1975, Choi et al. 2011) to (10). (We discuss our choice of H_k along with each set of experiments.) Letting $(u_{k,t}, \delta_{k,t})$ denote the tth MINRES iterate, where $(u_{k,0}, \delta_{k,0}) = (0,0)$, the method sets $(u_k, \delta_k) \leftarrow (u_{k,t}, \delta_{k,t})$, where t is the first MINRES iteration such that for some $\kappa \in (0,1)$ (recalling the definition of $(\rho_{k,t}, r_{k,t})$ in (28)),

$$\left\| \begin{bmatrix} \rho_{k,t} \\ r_{k,t} \end{bmatrix} \right\|_\infty \leq \max\{\kappa \|g_k + H_k v_k\|_\infty, 10^{-12}\}, \tag{42}$$

and TT1 and/or TT2 hold. The choice of $\kappa \in (0,1)$ is discussed with each experiment.

4.2. Choosing the Step Size

Algorithm 1 (see line 15) stipulates that the step size α_k chosen for the kth iteration satisfies $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$. Keeping in mind that $\alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq \min\{\alpha_k^{\varphi}, 1\}$ (see Lemma 4), we take advantage of this flexibility in choosing the step size by defining

$$\alpha_k \leftarrow \begin{cases} 1 & \text{if } \alpha_k^{\min} = 1 \\ \min\{\alpha_k^{\min} + \theta\beta_k^2, (1.1)^{t_k+1}\alpha_k^{\min}\} & \text{if } \alpha_k^{\min} < 1 \text{ and } \min\{\alpha_k^{\min} + \theta\beta_k^2, (1.1)^{t_k+1}\alpha_k^{\min}\} \leq \alpha_k^{\varphi} \\ (1.1)^{t_k}\alpha_k^{\min} & \text{otherwise,} \end{cases}$$

where t_k is the largest value of $t \in \mathbb{N}$ such that $(1.1)^t\alpha_k^{\min} \leq \min\{\alpha_k^{\varphi}, \alpha_k^{\min} + \theta\beta_k^2, 1\}$. We do not explicitly compute α_k^{φ} in our code. Instead, we can verify whether $(1.1)^t\alpha_k^{\min} \leq \alpha_k^{\varphi}$ (as needed) because it is equivalent to verifying whether $\varphi((1.1)^t\alpha_k^{\min}) \leq 0$, which is computable.

4.3. Algorithm Variants Tested

To test the utility of using inexact subproblem solutions in Algorithm 1, we consider two algorithm variants: `SISQO` and `SISQO_exact`. `SISQO` is Algorithm 1 with inexact solutions computed as described in Section 4.1 with a relatively *large* value for κ in (42). On the other hand, `SISQO_exact` is identical to `SISQO` with the exception that it uses a relatively *small* value for κ in (42). (Because of the similarities of the algorithms, `SISQO_exact` acts as a proxy for the stochastic sequential quadratic programming (SQP) algorithm from Berahas et al. (2021), although because it employs iterative linear algebra techniques, we are able to compare `SISQO_exact` with `SISQO` more readily.) We specify the values of $\kappa \in (0,1)$ used along with each of our tests in Sections 4.5 and 4.6. Our reason for comparing these two variants is to focus attention on the numerical gains obtained as a result of using inexact subproblem solutions. Both variants use the same computation for the normal step, so the performance difference can be attributed directly to the inexact tangential step computation.

Additionally, we compare `SISQO` with a stochastic subgradient method employed to minimize the merit function ϕ directly (for various fixed values of τ). We refer to our implementation of this algorithm as `Subgrad`. Because H_k is a diagonal matrix for all $k \in \mathbb{N}$ in all of our experiments, *one* CG or MINRES iteration is comparable computationally with *two* iterations of `Subgrad`.

4.4. Metrics Used for Comparison

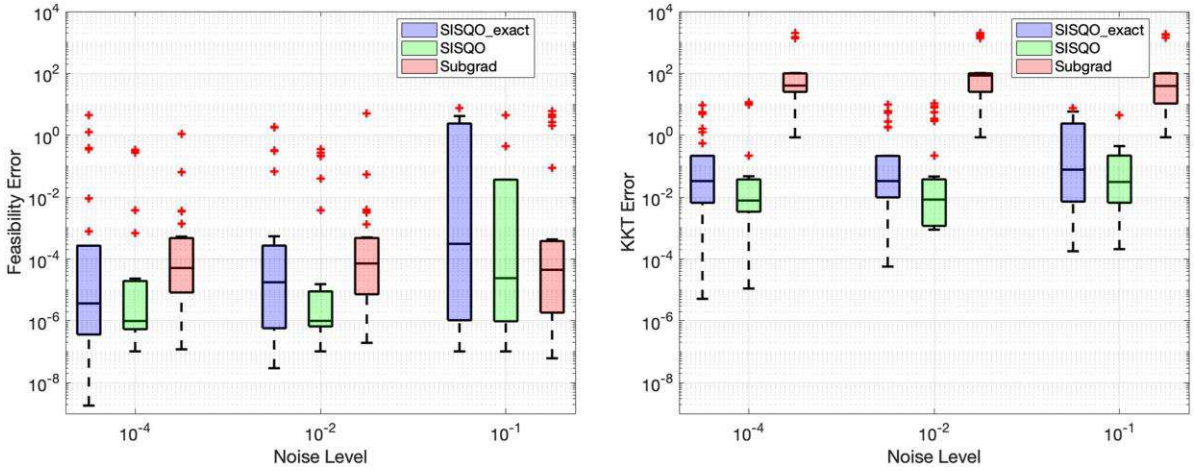
Our metrics of interest are infeasibility and stationarity. Given any iterate x_k in a run of `SISQO`, we consider the termination conditions $\|c(x_k)\|_\infty \leq 10^{-6}$ and $\|\nabla f_k + \nabla c_k y_{k,\text{ls}}\|_\infty \leq 10^{-2}$, where $y_{k,\text{ls}}$ is the least-squares multiplier at x_k. If an iterate satisfying these conditions is found in the first 1,000 iterations, then `SISQO` terminates and returns $x_{\texttt{SISQO}} \leftarrow x_k$. Otherwise, `SISQO` terminates after the 1,000th iteration and sets $k' \leftarrow \arg\min_{i \in \{0\} \cup [1,000]} \|c(x_i)\|_\infty$ (so $x_{k'}$ is the most feasible iterate found). If $\|c(x_{k'})\|_\infty > 10^{-6}$, then it returns $x_{\texttt{SISQO}} \leftarrow x_{k'}$; otherwise, it returns $x_{\texttt{SISQO}} \leftarrow x_{k''}$, where $k'' = \arg\min_{\{i \in \{0\} \cup [1,000] : \|c(x_i)\|_\infty \leq 10^{-6}\}} \|\nabla f_i + J_i^T y_{i,\text{ls}}\|_\infty$. This allows us to associate with each run of `SISQO` the two measures $\text{err}_{\text{feasibility}} = \|c(x_{\texttt{SISQO}})\|_\infty$ and $\text{err}_{\text{stationarity}} = \|\nabla f(x_{\texttt{SISQO}}) + J(x_{\texttt{SISQO}})^T y_{\texttt{SISQO}}\|_\infty$, where $y_{\texttt{SISQO}} \in \mathbb{R}^m$ is the least-squares multiplier at $x_{\texttt{SISQO}}$. We use the total number of CG and MINRES iterations performed by `SISQO` as a budget for the total number of CG and MINRES iterations performed by `SISQO_exact`; no other termination condition is used for `SISQO_exact`. Upon termination of `SISQO_exact`, we define $x_{\texttt{exact}}$—the iterate with which we define the feasibility and stationarity errors—using the same strategy as for setting $x_{\texttt{SISQO}}$. Finally, we ran `Subgrad` for multiple instances of τ. (Further details on the choices of τ and the iteration budget for `Subgrad` are given with each experiment.) Upon termination of `Subgrad`, we define x_{subgrad}—the iterate with which we define the feasibility and stationarity errors—using the same strategy as for setting $x_{\texttt{SISQO}}$. In all cases, we define the KKT error as the maximum of the feasibility and stationarity errors.

4.5. Results on the CUTEst Problems

In the CUTEst set (Gould et al. 2015), there are 138 equality constrained problems with $m \leq n$. From these, we selected those such that (i) $(n+m) \in [500, 10,000]$, (ii) the objective function is not constant, (iii) the objective function remained above -10^{50} over the sequences of iterates generated by runs of our algorithm, and (iv) the linear independence constraint qualification was satisfied at all iterates encountered in each run of our algorithm. This process of elimination resulted in the following 11 test problems: `ELEC`, `LCH`, `LUKVLE1`, `LUKVLE3`, `LUKVLE4`, `LUKVLE6`, `LUKVLE7`, `LUKVLE9`, `LUKVLE10`, `LUKVLE13`, and `ORTHREGC`.

The function and derivative evaluations from CUTEst are deterministic, and for the purpose of these experiments, we exploited this fact to compute values as needed by our algorithm, including using function evaluations

Figure 1. (Color online) Box Plots of CUTEst Problems for Feasibility (Left Panel) and KKT (Right Panel) Errors



Note. Subgrad, stochastic subgradient method.

to estimate Lipschitz constants. However, we introduced noise into the computation of the objective gradients for each application of our stochastic algorithm. In particular, we generated stochastic gradients as $g_k = \mathcal{N}(\nabla f_k, \frac{\epsilon_N^2}{n} I)$, where for testing purposes, we considered the three noise levels $\epsilon_N \in \{10^{-4}, 10^{-2}, 10^{-1}\}$. This particular choice for defining the stochastic gradients ensured that an appropriate value for M_g as indicated in Assumption 6 would be given by $M_g = \{10^{-8}, 10^{-4}, 10^{-2}\}$, corresponding to the values for ϵ_N.

We set $\kappa = 0.1$ for `SISQO` and $\kappa = 10^{-7}$ for `SISQO_exact`. All of the remaining parameters were set identically: $\tau_{-1} = \sigma = \kappa_v = 0.1$, $\eta = 0.5$, $\epsilon_{uv} = 1{,}000$, $\xi_{-1} = \epsilon_c = 1$, $\epsilon_\tau = \epsilon_\xi = 0.01$, $\kappa_\rho = \kappa_r = 100$, $\epsilon_2 = 0.9$, $\kappa_u = 10^{-8}$, $\chi = 1 - 10^{-8}$, $\theta = 10^4$, and $\beta_k = 1$ for all $k \in \mathbb{N}$. For all $k \in \mathbb{N}$, we randomly generated a sample point near x_k, and then, we estimated L_k and Γ_k using finite differences of the objective gradients and constraint Jacobians between x_k and the sampled point. These values were used in place of L and Γ, respectively, in our step size selection. Here, $H_k = I$ for all $k \in \mathbb{N}$ in all runs.

For each test problem, we ran `SISQO`, `SISQO_exact`, and `Subgrad` with five different random seeds. As previously justified, the iteration budget for `Subgrad` was set to be twice the total numbers of CG and MINRES iterations used by `SISQO`. Also, for `Subgrad`, we ran the algorithm with the 11 merit parameter values in $\tau \in \{10^0, 10^{-1}, \ldots, 10^{-10}\}$ with step sizes set as $\alpha_k = \frac{\tau}{\tau L_k + \Gamma_k}$ for all $k \in \mathbb{N}$, and then, we selected the best iterate over all of these runs. We computed the feasibility and KKT errors for all algorithms as described in Section 4.4; see Figure 1.

From Figure 1, one finds that `SISQO` performs better than `SISQO_exact` and `Subgrad` in terms of both feasibility and KKT errors. `SISQO` achieves smaller errors for smaller noise levels, which may be expected because of the fact that these experiments are run with constant $\{\beta_k\}$.

4.6. Results on Optimal Control Problems

In our second set of experiments, we considered two optimal control problems motivated by those in Hintermüller et al. (2003). In particular, we modified the problems to have equality constraints only and finite sum objective functions. Specifically, given a domain $\Xi \in \mathbb{R}^2$, a constant $N \in \mathbb{N}_{>0}$, reference functions $\overline{w}_{ij} \in L^2(\Xi)$ and $\overline{z}_{ij} \in L^2(\Xi)$ for $(i,j) \in \{1,\ldots,N\} \times \{1,\ldots,N\}$, and a regularization parameter $\lambda \in \mathbb{R}_{>0}$, we first considered the problem

$$\min_{w,z} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{2} \|w - \overline{w}_{ij}\|_{L^2(\Xi)}^2 + \frac{\lambda}{2} \|z - \overline{z}_{ij}\|_{L^2(\Xi)}^2 \right) \text{s.t.} -\Delta w = z \text{ in } \Xi \text{ and } w = 0 \text{ on } \partial\Xi. \tag{43}$$

Second, with the same notation but $\overline{z}_{ij} \in L^2(\partial\Xi)$, we also considered

$$\min_{w,z} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{2} \|w - \overline{w}_{ij}\|_{L^2(\Xi)}^2 + \frac{\lambda}{2} \|z - \overline{z}_{ij}\|_{L^2(\partial\Xi)}^2 \right) \text{s.t.} -\Delta w + w = 0 \text{ in } \Xi \text{ and } \frac{\partial w}{\partial p} = z \text{ on } \partial\Xi, \tag{44}$$

where p represents the unit outer normal to Ξ along $\partial\Xi$. As reference functions for both problems, we chose $\overline{z}_{ij} = 0$ and $\overline{w}_{ij}(x_1, x_2) = \sin\left(\left(4 + \frac{\epsilon_N}{\epsilon_S}\left(i - \frac{N+1}{2}\right)\right)x_1\right) + \cos\left(\left(3 + \frac{\epsilon_N}{\epsilon_S}\left(j - \frac{N+1}{2}\right)\right)x_2\right)$ for all $(i,j) \in \{1,\ldots,N\} \times \{1,\ldots,N\}$ for some

Table 1. Numerical Results for Problems (43) and (44) Averaged over 10 Independent Runs

		Problem (43)			Problem (44)		
Strategy	ϵ_N	Feasibility error	KKT error	C + M iter. (iter.)	Feasibility error	KKT error	C + M iter. (iter.)
`SISQO`	10^{-4}	6.30×10^{-7}	2.08×10^{-6}	61,225.8 (6)	7.96×10^{-7}	7.72×10^{-6}	96,684.4 (9)
`SISQO_exact`	10^{-4}	5.90×10^{-7}	1.76×10^{0}	61,225.8 (6)	3.91×10^{-6}	8.29×10^{-1}	96,684.4 (8)
`Subgrad`	10^{-4}	4.98×10^{1}	4.98×10^{1}	0 (61,225.8)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96,684.4)
`SISQO`	10^{-2}	6.37×10^{-7}	2.10×10^{-4}	60,113 (6)	7.80×10^{-7}	1.86×10^{-4}	96,103.4 (9)
`SISQO_exact`	10^{-2}	5.82×10^{-7}	1.76×10^{0}	60,113 (6)	1.44×10^{-6}	8.29×10^{-1}	96,103.4 (8.8)
`Subgrad`	10^{-2}	4.98×10^{1}	4.98×10^{1}	0 (60,113)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96,103.4)
`SISQO`	10^{-1}	6.81×10^{-7}	2.09×10^{-3}	58,901.2 (6)	8.12×10^{-7}	1.68×10^{-3}	96,914.6 (9.2)
`SISQO_exact`	10^{-1}	5.85×10^{-7}	1.76×10^{0}	58,901.2 (6)	1.33×10^{-6}	8.29×10^{-1}	96,914.6 (8.8)
`Subgrad`	10^{-1}	4.98×10^{1}	4.98×10^{1}	0 (58,901.2)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96,914.6)

Note. C + M iter., CG and MINRES iteration; iter., iteration; Subgrad, stochastic subgradient method.

$(\epsilon_S, \epsilon_N) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. We selected the following values for the constants: $N = 3$, $\lambda = 10^{-5}$, $\epsilon_S = 50$, and $\epsilon_N \in \{10^{-4}, 10^{-2}, 10^{-1}\}$. Because the objective functions of (43) and (44) are finite sums, to generate stochastic gradients as unbiased estimates of the true gradient, we first uniformly generated random $(i, j) \in \{1, \ldots, N\} \times \{1, \ldots, N\}$, and then, we computed the gradient corresponding to the (i, j)th term in the objective function. We note that with the choice of parameters, it follows that an appropriate value for M_g in Assumption 6 is given by $M_g \approx \{10^{-8}, 10^{-4}, 10^{-2}\}$ to correspond, respectively, to the values for ϵ_N.

Because the optimal control problems have a quadratic objective function and linear constraints, we used the exact second derivative matrix $H_k = \text{diag}(I, \lambda I)$ for all $k \in \mathbb{N}$. For this choice, the curvature condition on H_k in Assumption 2 is trivially satisfied.

In terms of algorithm parameters, we set $\kappa = 10^{-4}$ for `SISQO` and $\kappa = 10^{-7}$ for `SISQO_exact`. All of the remaining parameters were set identically for the two variants in the same manner as in the previous section with the following exceptions: $\tau_{-1} = 10^{-4}$, $L = 1$, and $\Gamma = 0$, where the latter choice is valid because the objectives are quadratic and the constraints are linear.

For each of the two optimal control problems in (43) and (44), we ran `SISQO`, `SISQO_exact`, and `Subgrad` with five different random seeds, and then, we computed their average feasibility and KKT errors as described in Section 4.4. We observed that $\{\tau_k\}$ was constant in all runs of `SISQO` and `SISQO_exact`. Therefore, we ran `Subgrad` with only three merit parameter values, namely $\tau \in \{10^{-2}, 10^{-4}, 10^{-6}\}$, and we choose step sizes as $\alpha_k \leftarrow \frac{\tau}{\tau L + \Gamma}$ for all $k \in \mathbb{N}$. (In these experiments, the budget for `Subgrad` iterations was set to the total numbers of CG and MINRES iterations used by `SISQO` because the constraint function evaluations, required in each iteration of `Subgrad`, are as expensive computationally as each CG and MINRES iteration.) In Table 1, we report average feasibility and KKT errors as well as the average number of iterations performed by Algorithm 1 before termination ("iter.") and the number of CG and MINRES iterations ("C + M iter."), with the latter discussed in Section 4.1. The results are given in Table 1. One can observe that `SISQO` performs better than the others in terms of average feasibility and KKT errors.

5. Conclusion

We have proposed, analyzed, and tested an *inexact* stochastic SQP algorithm for solving stochastic optimization problems involving deterministic, smooth, nonlinear equality constraints. We proved a convergence guarantee (in expectation) for our algorithm that is comparable with that proved for the *exact* stochastic SQP method recently presented by Berahas et al. (2021), which in turn, is comparable with that known for the stochastic gradient in unconstrained settings (Bottou et al. 2018). Our MATLAB implementation, `SISQO`, illustrated the benefits of allowing inexact step computation for solving problems from the CUTEst set (Gould et al. 2015) and two optimal control problems.

References

Achiam J, Held D, Tamar A, Abbeel P (2017) Constrained policy optimization. Precup D, Teh YW, eds. *Proc. 34th Internat. Conf. Machine Learn.*, vol. 70 (PMLR, New York), 22–31.

Berahas AS, Curtis FE, O'Neill MJ, Robinson DP (2023) A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. *Math. Oper. Res.*, ePub ahead of print October 30, https://doi.org/10.1287/moor.2021.0154.

Berahas AS, Curtis FE, Robinson DP, Zhou B (2021) Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM J. Optim.* 31(2):1352–1379.

Biros G, Ghattas O (2003) Inexactness issues in the Lagrange–Newton–Krylov–Schur method for PDE-constrained optimization. Biegler LT, Ghattas O, Heinkenschloss M, Van Bloemen Waanders B, eds. *Large-Scale PDE-Constrained Optimization* (Springer, New York), 93–114.

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

Byrd RH, Curtis FE, Nocedal J (2008) An inexact SQP method for equality constrained optimization. *SIAM J. Optim.* 19(1):351–369.

Byrd RH, Curtis FE, Nocedal J (2010) An inexact Newton method for nonconvex equality constrained optimization. *Math. Programming* 122(2):273–299.

Chatterjee N, Chen YH, Maas P, Carroll RJ (2016) Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.* 111(513):107–117.

Chen C, Tung F, Vedula N, Mori G (2018) Constraint-aware deep neural network compression. Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018. ECCV 2018* (Springer, Berlin, Heidelberg), 409–424.

Choi SCT, Paige CC, Saunders MA (2011) MINRES-QLP: A Krylov subspace method for indefinite or singular symmetric systems. *SIAM J. Sci. Comput.* 33(4):1810–1836.

Courant R (1943) Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.* 49(1):1–23.

Curtis FE, Nocedal J, Wächter A (2009) A matrix-free algorithm for equality constrained optimization problems with rank-deficient Jacobians. *SIAM J. Optim.* 20(3):1224–1249.

Curtis FE, O'Neill MJ, Robinson DP (2024) Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Math. Programming* 205(1):431–483.

Fletcher R (2000) *Practical Methods of Optimization* (John Wiley & Sons, New York).

Geyer CJ (1991) Constrained maximum likelihood exemplified by isotonic convex logistic regression. *J. Amer. Statist. Assoc.* 86(415):717–724.

Gill PE, Murray W, Saunders MA (2002) SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM J. Optim.* 12(4):979–1006.

Gould NIM, Orban D, Toint PL (2015) CUTEst: A constrained and unconstrained testing environment with safe threads for mathematical optimization. *Comput. Optim. Appl.* 60:545–557.

Han SP (1977) A globally convergent method for nonlinear programming. *J. Optim. Theory Appl.* 22(3):297–309.

Han SP, Mangasarian OL (1979) Exact penalty functions in nonlinear programming. *Math. Programming* 17(1):251–269.

Heinkenschloss M, Ridzal D (2008) An inexact trust-region SQP method with applications to PDE-constrained optimization. Kunisch K, Of G, Steinbach O, eds. *Numer. Math. Advanced Appl.* (Springer, Berlin, Heidelberg), 613–620.

Heinkenschloss M, Vicente LN (2002) Analysis of inexact trust-region SQP algorithms. *SIAM J. Optim.* 12(2):283–302.

Hintermüller M, Ito K, Kunisch K (2003) The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.* 13(3):865–888.

Kumar Roy S, Mhammedi Z, Harandi M (2018) Geometry aware constrained optimization techniques for deep learning. *Proc. IEEE Conf. CVPR* (IEEE, Piscataway, NJ), 4460–4469.

Márquez-Neila P, Salzmann M, Fua P (2017) Imposing hard constraints on deep networks: Promises and limitations. Preprint, submitted June 7, https://arxiv.org/abs/1706.02025.

Na S, Anitescu M, Kolar M (2023) An adaptive stochastic sequential quadratic programming with differentiable exact augmented Lagrangians. *Math. Programming* 199(1):721–791.

Nandwani Y, Pathak A, Singla P (2019) A primal-dual formulation for deep learning with constraints. *Adv. Neural Inform. Processing Systems* 1091:12157–12168.

Nocedal J, Wright SJ (2006) *Numerical Optimization*, 2nd ed., Springer Series in Operations Research (Springer, New York).

Oztoprak F, Byrd R, Nocedal J (2023) Constrained optimization in the presence of noise. *SIAM J. Optim.* 33(3):2118–2136.

Paige CC, Saunders MA (1975) Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.* 12(4):617–629.

Powell MJD (1978a) A fast algorithm for nonlinearly constrained optimization calculations. Watson GA, eds. *Numerical Analysis*, Lecture Notes in Mathematics, vol. 630 (Springer, Berlin, Heidelberg), 144–157.

Powell MJD (1978b) Algorithms for nonlinear constraints that use Lagrangian functions. *Math. Programming* 14(1):224–248.

Ravi SN, Dinh T, Lokhande VS, Singh V (2019) Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. *Proc. Conf. AAAI Artificial Intelligence* (AAAI Press, Palo Alto, CA), 4772–4779.

Rockafellar RT (1976) Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* 14(5):877–898.

Ruthotto L, Haber E (2020) Deep neural networks motivated by partial differential equations. *J. Math. Imaging Vision* 62:352–364.

Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).

Sheriffdeen S, Ragusa JC, Morel JE, Adams ML, Bui-Thanh T (2019) Accelerating PDE-constrained inverse solutions with deep learning and reduced order models. Preprint, submitted December 17, https://arxiv.org/abs/1912.08864.

Shor NZ (2012) *Minimization Methods for Non-Differentiable Functions*, vol. 3 (Springer Science & Business Media, New York).

Summers T, Warrington J, Morari M, Lygeros J (2015) Stochastic optimal power flow based on conditional value at risk and distributional robustness. *Internat. J. Electr. Power Energy Systems* 72:116–125.

Tomar VS, Rose RC (2014) Manifold regularized deep neural networks. INTERSPEECH (ISCA, Singapore), 348–352.

Wilson RB (1963) A simplicial algorithm for concave programming. PhD thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA.

Zhu Y, Zabaras N, Koutsourelakis PS, Perdikaris P (2019) Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* 394:56–81.