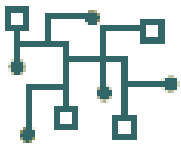

ISE

Industrial and Systems Engineering

Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Function Subject to Deterministic Nonlinear Equality Constraints

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

Department of Industrial and Systems Engineering, Lehigh University

COR@L Technical Report 21T-015

LEHIGH UNIVERSITY

COR@L

COMPUTATIONAL OPTIMIZATION RESEARCH AT LEHIGH

Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Function Subject to Deterministic Nonlinear Equality Constraints

Frank E. Curtis*[1], Daniel P. Robinson†[1], and Baoyu Zhou‡[1]

[1]Department of Industrial and Systems Engineering, Lehigh University

July 6, 2021

Abstract

An algorithm is proposed, analyzed, and tested experimentally for solving stochastic optimization problems in which the decision variables are constrained to satisfy equations defined by deterministic, smooth, and nonlinear functions. 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. A distinguishing feature of the algorithm is that it allows inexact subproblem solutions to be employed, 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 available. Convergence results in expectation are established for the method. Numerical experiments show that it outperforms an alternative algorithm that employs highly accurate subproblem solutions in every iteration.

1 Introduction

In this paper, we consider the design, analysis, and implementation of a stochastic inexact sequential quadratic optimization (SISQO) algorithm for minimizing a stochastic objective function subject to deterministic equality constraints. Specifically, we consider problems that may be written in the form

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

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 $\mathbb{E}[\cdot]$ 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 [32]; (ii) multiple deep learning problems including physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data [41], natural language processing with constraints on output labels [25], image classification, detection, and localization [30], deep reinforcement learning [1], deep network compression [9], and manifold regularized deep learning [22, 37]; (iii) accelerating

*E-mail: frank.e.curtis@lehigh.edu
†E-mail: daniel.p.robinson@lehigh.edu
‡E-mail: baoyu.zhou@lehigh.edu

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-order and full-order models [34]; (iv) multi-stage modeling [33]; (v) portfolio selection [33]; (vi) optimal power flow [36, 38, 40]; and (vii) statistical problems such as maximum likelihood estimation with constraints [8, 14]. (For an overview of the promises and limitations of imposing hard constraints during deep neural network training, see [23].)

Popular algorithmic approaches for solving problems of the form (1) when the objective function f is *deterministic* include penalty methods [11, 13] and sequential quadratic optimization (SQO) methods [29, 39]. 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 [35], or proximal methods [21, 31] (with the appropriateness of the method depending 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. For example, so-called line search SQO methods are considered state-of-the-art for solving deterministic equality constrained optimization problems [17, 16, 28]. 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 can be seen as being 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 methods.) For 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 MINRES [27], to the linear system. 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 few solutions [4, 6, 7, 18, 19].

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 only aware of three papers, namely [2, 3, 24], that present algorithms for solving stochastic optimization problems with deterministic 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 in [24] is a line search method that uses a differentiable exact augmented Lagrangian function as its merit function, whereas [3] (resp., [2]) is an SQO method that uses an ℓ_1-norm (resp., ℓ_2-norm) penalty function as its merit function. All of these methods must factorize a matrix during each iteration, which may not be tractable for large-scale problems. This motivates the work in this paper, which extends the methods in [2, 3] to allow for inexact subproblem solutions, thereby making our approach applicable for solving problem (1) in large-scale settings.

1.1 Contributions

The contributions of this paper pertain to a new algorithm for solving problem (1), which we now summarize. (i) We design a SISQO method for solving the stochastic optimization problem (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, since it is this parameter that balances the emphasis between the objective function and the constraint violation in

the optimization process. (ii) Under mild assumptions that include good behavior of the adaptive merit parameter (which can be justified as explained in the paper), we prove convergence in expectation of our algorithm. (iii) We present numerical results that compare our SISQO algorithm to a stochastic exact SQO algorithm. These experiments show that our SISQO algorithm benefits from our proposed inexactness strategy.

1.2 Notation

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_{\geq p}$ (resp., $\mathbb{R}_{>p}$) denote the set of real numbers greater than or equal to (resp., strictly greater than) $p \in \mathbb{R}$, and $\mathbb{N} := \{0, 1, 2, \ldots\}$ 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 $[p] := \{1, \ldots, p\}$. The ℓ_2-norm is written simply as $\|\cdot\|$.

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 in the kth iteration of the algorithm, and for brevity we define $\nabla f_k := \nabla f(x_k)$, $c_k := c(x_k)$, and $J_k = \nabla c(x_k)^T$. We refer to the range space of J_k^T as $\text{Range}(J_k^T)$ and refer to the null space of J_k as $\text{Null}(J_k)$, and recall that the Fundamental Theorem of Linear Algebra provides that these spaces are orthogonal and $\text{Range}(J_k^T) + \text{Null}(J_k) = \mathbb{R}^n$. Finally, recall (see, e.g., [26]) that a *primal* point $x \in \mathbb{R}^n$ and *dual* point $y \in \mathbb{R}^m$ constitute a first-order stationary point for problem (1) if and only if

$$c(x) = 0 \quad \text{and} \quad \nabla f(x) + \nabla c(x) y = 0. \tag{2}$$

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. Our convergence analysis for the algorithm is presented in Section 3. The results of numerical experiments are presented in Section 4 and concluding remarks are presented in Section 5.

2 SISQO Algorithm

Our proposed algorithm generates a sequence

$$\{(x_k, y_k, v_k, u_k, d_k, \delta_k, \tau_k, \alpha_k)\} \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{>0} \times \mathbb{R}_{>0},$$

where, for all $k \in \mathbb{N}$, (x_k, y_k) is a primal-dual iterate pair, v_k is a *normal* direction that aims to reduced infeasibility by reducing a local derivative-based model of the ℓ_2-norm constraint violation measure, u_k is a *tangential* direction that aims to maintain the reduction in linearized infeasibility achieved by the normal direction while also aiming to reduce the objective function by reducing a stochastic-gradient-based quadratic approximation of the objective, $d_k := v_k + u_k$ is a full primal search direction, δ_k is a dual search direction, τ_k is a merit function parameter, and α_k is a step size that aims to produce $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yielding sufficient reduction in the ℓ_2-norm merit function. (The algorithm also generates sequences of adaptive auxiliary parameters that are introduced throughout our algorithm description.) In the remainder of this section, we discuss each of these quantities in further detail toward our complete algorithm statement, which is provided as Algorithm 1 on page 9.

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 the iterate sequence $\{x_k\}$ generated by any run of our algorithm. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below over $\mathcal{X}$ and its gradient function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $L \in \mathbb{R}_{>0}$ (with respect to the ℓ_2-norm) and bounded over $\mathcal{X}$. The constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ (with $m \leq n$) is continuously*

differentiable and bounded over $\mathcal{X}$ and its Jacobian function $J : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is Lipschitz continuous with constant $\Gamma \in \mathbb{R}_{>0}$ (with respect to the induced ℓ_2-norm) and bounded over $\mathcal{X}$. In addition, for all $x \in \mathcal{X}$, the Jacobian $J(x)$ has singular values that are bounded uniformly below by a positive real number.

Such an assumption is standard in the literature on deterministically constrained optimization. Observe that it does not include an assumption that $\mathcal{X}$ is bounded.

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

$$\phi(x, \tau) = \tau f(x) + \|c(x)\|, \tag{3}$$

where τ is a positive *merit parameter* that is updated adaptively by the algorithm. (The choice of the ℓ_2-norm in ϕ is not essential for our method. Another norm could be used instead. 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 given by

$$l(x, \tau, g, d) = \tau(f(x) + g^T d) + \|c(x) + \nabla c(x)^T d\|,$$

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

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

The merit function, and in particular the model reduction function (4), play critical roles in our inexactness conditions for defining acceptable search directions and in our step size selection scheme, as can be seen in the following subsections.

2.2 Computing a search direction

During the kth iteration, the algorithm computes a normal direction $v_k \in \text{Range}(J_k^T)$ based on the subproblem

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

Instead of solving (5) exactly, the algorithm allows for an inexact solution to be employed by only requiring the computation of $v_k \in \text{Range}(J_k^T)$ satisfying

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

(commonly known as the Cauchy decrease condition), where $\epsilon_c \in (0, 1]$ is a user-defined constant. In (6), $v_k^c := -J_k^T c_k$ 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 $\|c_k + \alpha J_k v_k^c\|$ over $\alpha \in \mathbb{R}$. If $\|c_k\| \neq 0$, then under Assumption 1 it follows that $\|J_k^T c_k\| \neq 0$,

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

otherwise, if $\|c_k\| = 0$, then $\|J_k^T c_k\| = 0$ and it follows that $v_k = 0$ is the unique solution to (5). Popular choices for computing a normal direction satisfying the aforementioned conditions include any of various Krylov subspace methods, such as the linear conjugate gradient (CG) method; see, e.g., [26].

Before describing the algorithm's procedure for computing the tangential direction, let us first introduce assumptions that the algorithm makes related to the stochastic gradients $\{g_k\}$ and symmetric matrices $\{H_k\}$ that it employs.

Assumption 2. *There exists $M_g \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the stochastic gradient g_k has the properties that $\mathbb{E}_k[g_k] = \nabla f_k$ and $\mathbb{E}_k[\|g_k - \nabla f_k\|^2] \leq M_g$, where $\mathbb{E}_k[\cdot]$ denotes expectation with respect to the distribution of ω (recall (1)) conditioned on the event that x_k is the primal iterate in iteration $k \in \mathbb{N}$.*

Combining Assumption 2 with Jensen's Inequality, it holds for all $k \in \mathbb{N}$ that

$$\mathbb{E}_k[\|\nabla f_k - g_k\|] \leq \sqrt{\mathbb{E}_k[\|\nabla f_k - g_k\|^2]} \leq \sqrt{M_g}. \tag{8}$$

Assumption 3. *For all $k \in \mathbb{N}$, the matrix $H_k \in \mathbb{S}^n$ is chosen independently from g_k. In addition, there exist $M_H \in \mathbb{R}_{>0}$ and $\zeta \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, it holds that $\|H_k\| \leq M_H$ and $u^T H_k u \geq \zeta \|u\|^2$ for all $u \in \text{Null}(J_k)$.*

For describing the tangential direction computation as it is performed in our algorithm, let us first describe what would be the computation of a tangential direction in a determinstic variant of our approach. In particular, given (x_k, y_k), ∇f_k, a normal direction $v_k \in \text{Range}(J_k^T)$, and H_k satisfying Assumption 3, consider the subproblem

$$\min_{u \in \mathbb{R}^n} \ (\nabla f_k + H_k v_k)^T u + \tfrac{1}{2} u^T H_k u \ \text{ s.t. } J_k u = 0, \tag{9}$$

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

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k^{\text{true}} \\ \delta_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}. \tag{10}$$

This allows us to define, for purposes of our analysis only, the *true and exact* primal-dual search direction (conditioned on x_k being the kth iterate) as $(d_k^{\text{true}}, \delta_k^{\text{true}})$, where $d_k^{\text{true}} := v_k + u_k^{\text{true}}$. Since our algorithm only has access to a stochastic gradient estimate in each iteration, the corresponding *exact* (but not *true*) primal-dual search direction is given by $(d_{k,*}, \delta_{k,*})$, where $d_{k,*} := v_k + u_{k,*}$ with $(u_{k,*}, \delta_{k,*})$ satisfying

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_{k,*} \\ \delta_{k,*} \end{bmatrix} = - \begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}. \tag{11}$$

Our algorithm, to avoid having to form or factor the matrix in (11) in order to solve the system exactly, computes a tangential direction by computing (u_k, δ_k) through iterative linear algebra techniques applied to the symmetric indefinite system (11). In particular, our algorithm computes (u_k, δ_k) such that the full primal search direction $d_k := v_k + u_k$, dual search direction δ_k, and residual defined by

$$\begin{bmatrix} \rho_k \\ r_k \end{bmatrix} := \begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k \\ \delta_k \end{bmatrix} + \begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix} \tag{12}$$

satisfy at least one of a couple sets of conditions. In the remainder of this subsection, we describe the sets of conditions that the algorithm employs to determine what constitutes an acceptable search direction (and corresponding pair of 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 search 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, e.g., [6, 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 the merit parameter τ be set such that the model reduction condition

$$\Delta l(x_k, \tau, g_k, v_k + u_k) \geq \sigma_u \tau \max\{u_k^T H_k u, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + J_k v_k\|) \tag{13}$$

holds for some user-defined $\sigma_u \in (0,1)$, $\epsilon_u \in (0, \zeta)$, and $\sigma_c \in (0,1)$. (The particular value for the merit parameter τ for which the inequality (13) is required to hold depends on one of two different situations, as described below.)

Condition (13) 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*, since they dictate conditions that, once satisfied, can cause termination of an iterative linear system solver applied to (11). (The tests are inspired by the *sufficient merit approximation reduction termination tests* developed in [6, 7, 12] for the deterministic setting.) Our first termination test states that an inexact solution of this linear system is acceptable if the model reduction condition (13) is satisfied with the current merit parameter value (i.e., $\tau \equiv \tau_k \leftarrow \tau_{k-1}$), the norms of the residual vectors satisfy certain upper bounds, and either the tangential direction is sufficiently small in norm compared to the normal direction or the tangential direction is one of sufficiently positive curvature for H_k and yields a sufficiently small objective value for (9) (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.

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 $v_k \in \text{Range}(J_k^T)$ computed to satisfy (6), the pair (u_k, δ_k) satisfies Termination Test 1 if, with the pair (ρ_k, r_k) defined in (12), it holds that

$$\|\rho_k\| \leq \kappa \min\left\{ \left\| \begin{bmatrix} g_k + J_k^T(y_k + \delta_k) \\ c_k \end{bmatrix} \right\|, \left\| \begin{bmatrix} g_{k-1} + J_{k-1}^T y_k \\ c_{k-1} \end{bmatrix} \right\| \right\}; \tag{14}$$

$$\|\rho_k\| \leq \kappa_\rho \beta_k \quad \text{and} \quad \|r_k\| \leq \kappa_r \beta_k; \tag{15}$$

$$\|u_k\| \leq \kappa_u \|v_k\| \quad \text{or} \quad \left\{ \begin{matrix} u_k^T H_k u_k \geq \epsilon_u \|u_k\|^2 \quad \text{and} \\ (g_k + H_k v_k)^T u_k + \frac{1}{2} u_k^T H_k u_k \leq \kappa_v \|v_k\| \end{matrix} \right\}; \tag{16}$$

and (13) is satisfied with $\tau \equiv \tau_{k-1}$. (In this case, the algorithm will set $\tau_k \leftarrow \tau_{k-1}$ so that (13) holds with $\tau \equiv \tau_k$.)

Termination Test 1 cannot be enforced in every iteration in a run of the algorithm, even in the deterministic setting, since there may exist points in the search space at which all of the conditions required in the test cannot be satisfied simultaneously, even if the linear system (11) is solved to arbitrary accuracy. In short, the algorithm needs to allow for the computation of a search direction for which (13) 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 to occur, 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.

Termination Test 2. Given $\kappa \in (0,1)$, $\beta_k \in (0,1]$, $\kappa_\rho \in \mathbb{R}_{>0}$, $\kappa_r \in \mathbb{R}_{>0}$, $\kappa_u \in \mathbb{R}_{>0}$, $\epsilon_u \in (0, \zeta)$, $\kappa_v \in \mathbb{R}_{>0}$, $\epsilon_r \in (\sigma_c, 1)$, and $v_k \in \text{Range}(J_k^T)$ computed to satisfy (6), the pair (u_k, δ_k) satisfies Termination Test 2 if, with the pair (ρ_k, r_k) defined in (12), the conditions (14)–(16) hold along with

$$\|c_k\| - \|c_k + J_k v_k + r_k\| \geq \epsilon_r (\|c_k\| - \|c_k + J_k v_k\|) > 0. \tag{17}$$

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

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \leq \tau_k^{\text{trial}} \\ \min\{(1-\epsilon_\tau)\tau_{k-1}, \tau_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \tag{18}$$

where

$$\tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} \leq 0 \\ \frac{(1-\frac{\sigma_c}{\epsilon_r})(\|c_k\| - \|c_k + J_k v_k + r_k\|)}{g_k^T d_k + \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\}} & \text{otherwise,} \end{cases} \tag{19}$$

so (13) is satisfied with $\tau \equiv \tau_k$. See Lemma 3 for a proof.)

In Lemma 1, we show under a loose assumption about the behavior of the iterative linear system solver and a practical assumption about the algorithm iterates that, for all $k \in \mathbb{N}$, the algorithm can compute a pair (u_k, δ_k) satisfying at least one of Termination Test 1 or 2. Therefore, the index of each iteration of our method is contained in one of two index sets, namely,

$$\begin{aligned}\mathcal{K}_1 &:= \{k \in \mathbb{N} : (u_k, \delta_k) \text{ satisfies Termination Test 1}\} \text{ or}\\ \mathcal{K}_2 &:= \{k \in \mathbb{N} : (u_k, \delta_k) \text{ satisfies Termination Test 2, but not Termination Test 1}\}.\end{aligned}$$

2.3 Computing a step size

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

$$\begin{aligned} f(x_k + \alpha d_k) &\leq f_k + \alpha \nabla f_k^T d_k + \tfrac{1}{2} L \alpha^2 \|d_k\|^2 \\ \text{and} \quad \|c(x_k + \alpha d_k)\| &\leq \|c_k + \alpha J_k d_k\| + \tfrac{1}{2}\Gamma \alpha^2 \|d_k\|^2. \end{aligned} \tag{20}$$

Combining these inequalities with the definitions (3) and (4), the triangle inequality, and the definition of r_k in (12), one finds that

$$\begin{aligned} &\phi(x_k + \alpha d_k, \tau_k) - \phi(x_k, \tau_k) \\ &= \tau_k f(x_k + \alpha d_k) - \tau_k f_k + \|c(x_k + \alpha d_k)\| - \|c_k\| \\ &\leq \alpha \tau_k \nabla f_k^T d_k + \|c_k + \alpha J_k d_k\| - \|c_k\| + \tfrac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2 \\ &\leq \alpha \tau_k \nabla f_k^T d_k + (|1-\alpha| - 1)\|c_k\| + \alpha \|c_k + J_k d_k\| + \tfrac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2 \\ &= -\alpha \Delta l(x_k, \tau_k, \nabla f_k, d_k) + (|1-\alpha| - (1-\alpha))\|c_k\| + \tfrac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|^2. \end{aligned} \tag{21}$$

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 the quantity

$$\alpha_k^{\text{suff}} := \min\left\{ \tfrac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2}, 1 \right\}. \tag{22}$$

The definition of α_k^{suff} can be motivated as follows. Its value, when $\beta_k = 1$, is the largest value on $[0,1]$ such that for all $\alpha \in [0, \alpha_k^{\text{suff}}]$ the right-hand-side of (21) (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 1, the value of α_k^{suff} can be diminished over the course of the optimization process, which allows for step size control as is required for convergence guarantees for certain stochastic-gradient-based methods; see, e.g., [5]. The first term inside the min appearing in (22) is important for the convergence guarantees that we prove for our method, but it can behave erratically due to the algorithm's use of stochastic gradient estimates. To account for this stochasticity, given user-defined $\epsilon_\xi \in (0,1)$, our algorithm defines

$$\xi_k^{\text{trial}} := \tfrac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|^2} \quad \text{and} \quad \xi_k := \begin{cases} \xi_{k-1} & \text{if } \xi_{k-1} \leq \xi_k^{\text{trial}} \\ \min\{(1-\epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \tag{23}$$

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 (22), the monotonically nonincreasing behaviors of $\{\xi_k\}$ and $\{\tau_k\}$, and assuming that the sequence $\{\beta_k\}$ is chosen to satisfy

$$2(1-\eta)\beta_k \xi_{-1}\tau_{-1}/\Gamma \in (0,1] \quad \text{for all} \quad k \in \mathbb{N} \tag{24}$$

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

$$\alpha_k^{\min} := \tfrac{2(1-\eta)\beta_k \xi_k \tau_k}{(\tau_k L + \Gamma)} \leq \min\left\{ \tfrac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2}, 1 \right\} \equiv \alpha_k^{\text{suff}}. \tag{25}$$

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 will also show that even though ξ_k is stochastic for each $k \in \mathbb{N}$, the sequence $\{\xi_k\}$ is bounded away from zero deterministically (see Lemma 9).

Next, let us derive a maximum value (i.e., an upper bound) for our algorithm's choice of step size. Consider the strongly convex function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned}\varphi(\alpha) :=&(\eta-1)\alpha\beta_k\Delta l(x_k,\tau_k,g_k,d_k) + \|c_k + \alpha J_k d_k\| - \|c_k\| \\ &+ \alpha(\|c_k\| - \|c_k + J_k d_k\|) + \tfrac{1}{2}(\tau_k L + \Gamma)\alpha^2\|d_k\|^2.\end{aligned} \tag{26}$$

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 quantity in the third row of (21) (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 as above, one can be motivated as to the fact that our algorithm never allows a step size larger than

$$\alpha_k^{\varphi} := \max\{\alpha \in \mathbb{R}_{\geq 0} : \varphi(\alpha) \leq 0\}. \tag{27}$$

Finally, again to mitigate adverse affects caused by the use of stochastic gradient estimates, our algorithm employs the maximum step size

$$\alpha_k^{\max} := \min\{\alpha_k^{\varphi}, \alpha_k^{\min} + \theta\beta_k^2\}, \tag{28}$$

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 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; in particular, the algorithm allows any y_{k+1} such that

$$\|g_k + J_k^T y_{k+1}\| \leq \|g_k + J_k^T(y_k + \delta_k)\|. \tag{29}$$

Clearly, choosing $y_{k+1} \leftarrow y_k + \delta_k$ is one particular option satisfying (29), although other choices such as least-squares multipliers could also be used.

Algorithm 1 Stochastic Inexact Sequential Quadratic Optimization (SISQO)

Require: initial values $(x_0, y_0, \tau_{-1}, \xi_{-1}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$; Lipschitz constants $(L, \Gamma) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ 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 (24) and H_k satisfying Assumption 3
  compute $v_k \in \text{Range}(J_k^T)$ satisfying (6)
  generate g_k satisfying Assumption 2
  compute (u_k, δ_k) satisfying at least one of Termination Tests 1 or 2
  if Termination Test 1 is satisfied **then**
    set $\tau_k^{\text{trial}} \leftarrow \infty$ and $\tau_k \leftarrow \tau_{k-1}$ $[k \in \mathcal{K}_1]$
  else (Termination Test 2 is satisfied)
    set τ_k^{trial} and τ_k by (18)–(19) $[k \in \mathcal{K}_2]$
  end if
  set $d_k \leftarrow v_k + u_k$
  compute ξ_k and ξ_k^{trial} by (23)
  choose $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$ using the definitions in (25) and (28)
  set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ and choose y_{k+1} satisfying (29)
end for

3 Analysis

Our analysis is presented in three parts. In Section 3.1, we show that Algorithm 1 is well posed. Then, in Section 3.2, we prove general lemmas about the behavior of our algorithm. Finally, in Section 3.3, we prove convergence properties in expectation for the iterate sequence generated by the algorithm under an assumption about the behavior of the merit parameter sequence. The assumption employed can be justified using the same arguments as in [3], as explained in Section 3.3.

3.1 Well-posedness

Our aim in this subsection is to prove that during each iteration of Algorithm 1, each step of the algorithm can be performed in a manner that terminates finitely. Along the way, we also establish useful properties of quantities computed by the algorithm. We make the following reasonable assumption concerning the behavior of the iterative linear system solver employed by the algorithm for the tangential direction and dual step computation.

Assumption 4. *For all $k \in \mathbb{N}$, the iterative linear system solver employed in line 5 generates a sequence $\{(u_{k,t}, \delta_{k,t}, \rho_{k,t}, r_{k,t})\}_{t\in\mathbb{N}}$ satisfying*

$$\begin{bmatrix}\rho_{k,t}\\ r_{k,t}\end{bmatrix} = \begin{bmatrix}H_k & J_k^T\\ J_k & 0\end{bmatrix}\begin{bmatrix}u_{k,t}\\ \delta_{k,t}\end{bmatrix} + \begin{bmatrix}g_k + H_k v_k + J_k^T y_k\\ 0\end{bmatrix} \quad \text{for all} \quad t \in \mathbb{N} \tag{30}$$

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 $(u_{k,*}, \delta_{k,*})$ is the unique solution to the linear system defined in (11).

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

Assumption 5. *For all $k \in \mathbb{N}$, it holds that $c_k \neq 0$ or $g_k \notin \text{Range}(J_k^T)$.*

We justify Assumption 5 in the following manner. In the deterministic setting, the algorithm encounters a point x_k such that $c_k = 0$ and $\nabla f_k \in \text{Range}(J_k^T)$ if and only if there exists y_k such that (2) holds for $(x, y) \equiv (x_k, y_k)$, i.e., the point (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 (11) is computed, or at least a sufficiently accurate solution of the system is computed such that a practical termination condition for (11) is triggered and the algorithm terminates. In the stochastic setting, the algorithm encounters $c_k = 0$ and $g_k \in \text{Range}(J_k^T)$ if and only if x_k is *exactly* feasible and the stochastic gradient lies *exactly* in the range space of J_k^T. Since g_k is a *stochastic* gradient, we contend that it is unlikely that it will lie *exactly* in $\text{Range}(J_k^T)$ except in special circumstances. Thus, for simplicity in our analysis, we impose Assumption 5 throughout this section. (If Assumption 5 were not to hold, then one of the following could be employed in a practical implementation: (i) if a sufficiently accurate solution of (11) does not satisfy either Termination Test 1 or 2, 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 5 holds with probability one in all iterations, in which case the convergence result that we prove will hold with probability one.)

We can now show that the search direction computation is well posed.

Lemma 1. *For all $k \in \mathbb{N}$, the iterative linear system solver computes (u_k, δ_k) satisfying at least one of Termination Test 1 or 2 in a finite number of iterations.*

Proof. We prove the result by considering two cases.

Case 1: $\|c_k\| > 0$. For this case, we show that $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ satisfies Termination Test 2 for sufficiently large $t \in \mathbb{N}$. Let us first observe that it follows from Assumption 4, Assumption 5, and the fact that $\beta_k \in (0, 1]$ that both (14) and (15) 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 (16) holds for all sufficiently large $t \in \mathbb{N}$. Since $\|c_k\| > 0$, it follows under Assumption 1 that $\|v_k\| > 0$. If $\|u_{k,*}\| = 0$, then Assumption 4 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 (16) holds with $u_k \equiv u_{k,t}$ for all sufficiently large $t \in \mathbb{N}$. On the other hand, if $\|u_{k,*}\| > 0$, then (11) and Assumption 3 imply

$$\begin{aligned} u_{k,*}^T(g_k + H_k v_k) + \tfrac{1}{2} u_{k,*}^T H_k u_{k,*} &< u_{k,*}^T(g_k + H_k v_k) + u_{k,*}^T H_k u_{k,*} \\ &= u_{k,*}^T(g_k + H_k v_k + H_k u_{k,*} + J_k^T y_k) \\ &= -u_{k,*}^T J_k^T \delta_{k,*} = -(J_k u_{k,*})^T \delta_{k,*} = 0. \end{aligned} \tag{31}$$

Combining this inequality with the facts that $\epsilon_u \in (0, \zeta)$, $\kappa_v \in \mathbb{R}_{>0}$, and $\|v_k\| > 0$, it follows under Assumptions 3 and 4 that the latter set of conditions in (16) holds with $u_k \equiv u_{k,t}$ for all sufficiently large $t \in \mathbb{N}$.

Finally, let us show that (17) holds for all sufficiently large $t \in \mathbb{N}$, which combined with the previous conclusions shows that Termination Test 2 is satisfied by $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ for all sufficiently large $t \in \mathbb{N}$. By Assumption 4, (6), and the aforementioned fact that $\|v_k\| > 0$, it follows that

$$\lim_{t \to \infty} (\|c_k\| - \|c_k + J_k v_k + r_{k,t}\|) = \|c_k\| - \|c_k + J_k v_k\| > 0,$$

which shows that (17) holds with $r_k \equiv r_{k,t}$ for all sufficiently large $t \in \mathbb{N}$, as desired.

Case 2: $\|c_k\| = 0$. For this case, we show that $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ satisfies Termination Test 1 for all sufficiently large $t \in \mathbb{N}$. First, recall that $\|c_k\| = 0$ implies that $\|v_k\| = 0$. We also claim that $\|u_{k,*}\| > 0$. To prove this by contradiction, suppose that $\|u_{k,*}\| = 0$. Combining this with $v_k = 0$ and (11), it follows that $g_k + J_k^T(y_k + \delta_{k,*}) = 0$, which with $c_k = 0$ violates Assumption 5. Thus, $\|u_{k,*}\| > 0$.

Next, notice that the argument used in the beginning of Case 1 still applies in this case, which allows us to conclude that both (14) and (15) hold with $(\rho_k, r_k) = (\rho_{k,t}, r_{k,t})$ for all sufficiently large $t \in \mathbb{N}$. Also, since $\|u_{k,*}\| > 0$, the first inequality in (31) holds as a strict inequality, i.e., $u_{k,*}^T(g_k + H_k v_k) + \frac{1}{2} u_{k,*}^T H_k u_{k,*} < 0$. Combining this inequality with Assumption 4, Assumption 3, and $\epsilon_u \in (0, \zeta)$ allows us to deduce that the second set of conditions in (16) 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 (11), it follows that $J_k d_{k,*} = J_k(u_{k,*} + v_k) = 0$, which with Assumption 3 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 $\|c_k\| = 0$, $\|v_k\| = 0$, $J_k d_{k,*} = J_k v_k = 0$, (11), and Assumption 3 shows that

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

meaning that the sufficient decrease condition (13) 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 $(u_k, \delta_k) \equiv (u_{k,t}, \delta_{k,t})$ will satisfy Termination Test 1, as desired. □

Next, we prove that every search direction is nonzero in norm.

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

Proof. For a proof by contradiction, suppose that $\|d_k\| = 0$. From this fact, $d_k = v_k + u_k$, and (12), it follows that $\rho_k = g_k + J_k^T(y_k + \delta_k) + H_k(v_k + u_k) = g_k + J_k^T(y_k + \delta_k)$. If $\|c_k\| = 0$, then this value for ρ_k shows that the inequality in (14) cannot hold, meaning that (u_k, δ_k) cannot satisfy Termination Test 1 or 2, which contradicts Lemma 1. Hence, the only possibility is that $\|c_k\| > 0$, which we shall assume for the remainder of the proof.

Notice from $\|d_k\| = 0$, $d_k = v_k + u_k$, and $r_k = J_k u_k$, it follows that $\|c_k\| - \|c_k + J_k v_k + r_k\| = \|c_k\| - \|c_k + J_k d_k\| = 0$, meaning that (17) is not satisfied; thus, (u_k, δ_k) does not satisfy Termination Test

2. Also, observe from $\|v_k\| > 0$ (which follows from $\|c_k\| > 0$ and Assumption 1), $\|d_k\| = 0$, and (6) that $\Delta l(x_k, \tau_k, g_k, d_k) = 0 < \sigma_u \tau_{k-1} \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + J_k v_k\|)$, meaning that (13) is not satisfied with $\tau = \tau_{k-1}$; thus, (u_k, δ_k) does not satisfy Termination Test 1. Overall, we have reached a contradiction to Lemma 1, and since 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 sequence ensures that the model reduction condition (13) always holds for $\tau \equiv \tau_k$. We also show another important property of the sequence $\{\tau_k\}$.

Lemma 3. *For all $k \in \mathbb{N}$, the inequality in* (13) *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 $k \in \mathcal{K}_1$ due to the manner in which Termination Test 1 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 (13) holds for $\tau \equiv \tau_k$ with $d_k = v_k + u_k$ if and only if

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

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 (17), $J_k u_k = r_k$, and $\epsilon_r \in (\sigma_c, 1)$ yields

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

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 (18) yields $\tau_k \leq \tau_k^{\text{trial}}$, which combined with (17), (19), $J_k u_k = r_k$, and $\epsilon_r \in (\sigma_c, 1)$ yields

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

as desired. Moreover, from (18), 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 $k \in \mathbb{N}$, it holds that $0 < \alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$ and $0 < \alpha_k^{\min} \leq \alpha_k^{\max}$. In addition, for all $k \in \mathbb{N}$, it holds that $\varphi(\alpha_k) \leq 0$.*

Proof. It follows from (25) 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 (25) and (28), 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 $k \in \mathbb{N}$. Since $\alpha_k^{\varphi} \geq 0$ 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} in (27) 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 the min in (22). 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 (26) shows that

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

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 (26), $\alpha_k^{\text{suff}} \in (0,1]$, and the triangle inequality that

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

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

Finally, let us show that $\varphi(\alpha_k) \leq 0$ for all $k \in \mathbb{N}$. By (4) and (26), one finds (as previously mentioned) that φ is strongly convex. In addition, one finds that $\varphi(0) = \varphi(\alpha_k^{\varphi}) = 0$, where $\alpha_k^{\varphi} \in \mathbb{R}_{>0}$ due to the first part of this lemma. Along with the fact that $0 < \alpha_k^{\min} \leq \alpha_k \leq \alpha_k^{\max} \leq \alpha_k^{\varphi}$, it follows that $\varphi(\alpha_k) \leq 0$, as desired. □

3.2 General results

Our aim in this subsection is to prove general results about the behavior of quantities generated by Algorithm 1. For our purposes here, we make the following assumption about the dual and residual sequences.

Assumption 6. *The dual iterate sequence $\{y_k\}$ and residual sequence $\{(\rho_k, r_k)\}$ (recall (12)) generated by Algorithm 1 are bounded in norm.*

We note that under Assumption 1 and Assumption 3, this additional assumption is mild; it should hold as long as any reasonable iterative solver is applied to (11) in each iteration of a run of the algorithm.

The next lemma gives a lower bound on $\|c_k\| - \|c_k + J_k v_k\|$ relative to $\|c_k\|$.

Lemma 5. *There exists $\omega_1 \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, it holds that*

$$\|c_k\| - \|c_k + J_k v_k\| \geq \omega_1 \|c_k\|.$$

Proof. This result follows as in [12, Lemma 3.5], but with small straightforward modifications to account for the fact that, in our analysis here, the singular values of $\{J_k\}$ are bounded away from zero as a consequence of Assumption 1. □

The next lemma shows that $\|v_k\|$ is of the same order as $\|c_k\|$.

Lemma 6. *There exists $\{\omega_2, \omega_3\} \subset \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, it holds that*

$$\omega_2 \|c_k\| \leq \|v_k\| \leq \omega_3 \|c_k\|. \tag{32}$$

Proof. Observe that Assumption 1 ensures the existence of $\lambda_{\min} \in \mathbb{R}_{>0}$ such that $J_k J_k^T \succeq \lambda_{\min} I$ for all $k \in \mathbb{N}$. We now prove each desired inequality. First, consider the former inequality in (32). Since this inequality holds trivially whenever $\|c_k\| = 0$, let us proceed under the assumption that $\|c_k\| > 0$. One finds

$$\begin{aligned}\|c_k\|^2 - \|c_k + \alpha_k^c J_k v_k^c\|^2 &= (\|c_k\| - \|c_k + \alpha_k^c J_k v_k^c\|)(\|c_k\| + \|c_k + \alpha_k^c J_k v_k^c\|)\\ &\leq 2\|c_k\|(\|c_k\| - \|c_k + \alpha_k^c J_k v_k^c\|).\end{aligned}$$

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

$$\begin{aligned}&\|J_k\|\|v_k\| \geq \|J_k v_k\| \geq \|c_k\| - \|c_k + J_k v_k\| \geq \epsilon_c(\|c_k\| - \|c_k + \alpha_k^c J_k v_k^c\|)\\ &\geq \tfrac{\epsilon_c}{2\|c_k\|}(\|c_k\|^2 - \|c_k + \alpha_k^c J_k v_k^c\|^2) = \tfrac{\epsilon_c}{2\|c_k\|}(-2\alpha_k^c c_k^T J_k v_k^c - (\alpha_k^c)^2\|J_k v_k^c\|^2).\end{aligned}$$

Substituting in for the value of α_k^c (recall (7)), then substituting $v_k^c = -J_k^T c_k$ and simplifying shows that $\|J_k\|\|v_k\| \geq (\frac{\epsilon_c}{2\|c_k\|})\alpha_k^c \|J_k^T c_k\|^2$. Again substituting the value of α_k^c and using the definition of $\lambda_{\min}$, it follows that

$$\|J_k\|\|v_k\| \geq \tfrac{\epsilon_c \|J_k^T c_k\|^4}{2\|c_k\|\|J_k J_k^T c_k\|^2} \geq \tfrac{\epsilon_c \lambda_{\min}^2 \|c_k\|^4}{2\|c_k\|\|J_k^T J_k\|^2\|c_k\|^2} = \tfrac{\epsilon_c \lambda_{\min}^2}{2\|J_k^T J_k\|^2}\|c_k\|.$$

It follows from this inequality and Assumption 1 that there exists $\omega_2 \in \mathbb{R}_{>0}$ such that the former inequality in (32) holds, as desired.

Let us now turn to the latter inequality in (32). It follows from the normal direction computation that $\|c_k\| \geq \|c_k + J_k v_k\|$, which by the triangle inequality implies that $\|J_k v_k\| \leq 2\|c_k\|$. Note that since $v_k \in \text{Range}(J_k^T)$, one has $v_k = J_k^T w_k$ where $w_k = (J_k J_k^T)^{-1} J_k v_k$. Putting these facts together shows that

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

which combined with Assumption 1 establishes the existence of a $\omega_3 \in \mathbb{R}_{>0}$ such that the second inequality in (32) holds, as desired. □

The next result gives a bound on the size of the search direction relative to the constraint violation and the size of the normal step.

Lemma 7. *There exists $\omega_4 \in \mathbb{R}_{\geq 2}$ such that, for all $k \in \mathbb{N}$, it holds that*

$$\|d_k\|^2 \leq \omega_4 (\|u_k\|^2 + \|c_k\|).$$

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

$$\begin{aligned}
\|d_k\|^2 &\leq (\|u_k\| + \|v_k\|)^2 = \|u_k\|^2 + \|v_k\|^2 + 2\|u_k\|\|v_k\| \\
&\leq 2(\|u_k\|^2 + \|v_k\|^2) \leq 2(\|u_k\|^2 + \omega_3^2 \|c_k\|^2) \\
&\leq \max\{2, 2\omega_3^2 \|c_k\|\}(\|u_k\|^2 + \|c_k\|).
\end{aligned}$$

The existence of the required $\omega_4 \in \mathbb{R}_{\geq 2}$ now follows from Assumption 1 since $\max\{2, 2\omega_3^2\|c_k\|\}$ is uniformly bounded for all $k \in \mathbb{N}$, 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 $\kappa_l \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$, it holds that*

$$\Delta l(x_k, \tau_k, g_k, v_k + u_k) \geq \kappa_l \tau_k (\|u_k\|^2 + \|c_k\|) \geq \tfrac{\kappa_l \tau_k}{\omega_4} \|d_k\|^2 > 0.$$

Proof. Lemma 3 shows that (13) holds with $\tau \equiv \tau_k$. Combining this fact with Lemma 5 and the monotonically nonincreasing behavior of $\{\tau_k\}$ shows that

$$\begin{aligned}
\Delta l(x_k, \tau_k, g_k, v_k + u_k) &\geq \sigma_u \tau_k \max\{u_k^T H_k u_k, \epsilon_u \|u_k\|^2\} + \sigma_c(\|c_k\| - \|c_k + J_k v_k\|) \\
&\geq \sigma_u \tau_k \epsilon_u \|u_k\|^2 + \sigma_c \omega_1 \|c_k\| \geq \tau_k(\sigma_u \epsilon_u \|u_k\|^2 + \sigma_c \omega_1 \|c_k\|/\tau_{-1}) \\
&\geq \min\{\sigma_u \epsilon_u, \tfrac{\sigma_c \omega_1}{\tau_{-1}}\} \tau_k (\|u_k\|^2 + \|c_k\|),
\end{aligned}$$

which proves the existence of the claimed $\kappa_l \in \mathbb{R}_{>0}$ since σ_u, ϵ_u, σ_c, ω_1, and τ_{-1} are positive real numbers. The remaining inequalites follow from Lemmas 7 and 2. □

We next prove a deterministic uniform lower bound for the sequence $\{\xi_k\}$.

Lemma 9. *There exists $\xi_{\min} \in \mathbb{R}_{>0}$ such that, in any run of the algorithm, 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 (23) and Lemmas 7 and 8 that

$$\xi_k^{\text{trial}} = \tfrac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|^2} \geq \tfrac{\kappa_l \tau_k (\|u_k\|^2 + \|c_k\|)}{\tau_k \omega_4 (\|u_k\|^2 + \|c_k\|)} = \tfrac{\kappa_l}{\omega_4}. \tag{33}$$

Now, consider any iteration such that $\xi_k < \xi_{k-1}$. For such iterations, it follows from (23) and (33) that $\xi_k \geq (1 - \epsilon_\xi)\xi_k^{\text{trial}} \geq (1 - \epsilon_\xi)\kappa_l/\omega_4$. Combining this fact with the initial choice of ξ_{-1} shows that $\xi_k \geq \xi_{\min} := \min\{(1 - \epsilon_\xi)\kappa_l/\omega_4, \xi_{-1}\}$ for all $k \in \mathbb{N}$. Combining this result with the fact that anytime $\xi_k < \xi_{k-1}$ it must hold 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 each iteration.

Lemma 10. *For all $k \in \mathbb{N}$, it holds that*

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

Proof. By Lemma 4, one has that $\varphi(\alpha_k) \leq 0$. Hence, starting with the third row of (21), adding and subtracting the terms $\alpha_k \tau_k \nabla f_k^T d_k^{\text{true}}$, $\alpha_k \|c_k\|$, $\alpha_k \|c_k + J_k d_k^{\text{true}}\|$, 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 $J_k d_k^{\text{true}} = J_k(v_k + u_k^{\text{true}}) = J_k v_k$, one finds that

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

which completes the proof. □

We now 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 a matrix whose columns form an orthonormal basis for $\text{Null}(J_k)$, which implies that $Z_k^T Z_k = I$ and $J_k Z_k = 0$. Under 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 $\text{Range}(J_k^T)$ and $\text{Null}(J_k)$ in the sense that $u_k = J_k^T u_{k,1} + Z_k u_{k,2}$. It follows from (12) that $u_{k,1} = (J_k J_k^T)^{-1} r_k$ and $u_{k,2} = -(Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k + H_k J_k^T (J_k J_k^T)^{-1} r_k - \rho_k)$, with which one can derive:

$$\begin{aligned}
u_k &= J_k^T (J_k J_k^T)^{-1} r_k - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k + H_k J_k^T (J_k J_k^T)^{-1} r_k - \rho_k) \\
u_k^{\text{true}} &= -Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f_k + H_k v_k).
\end{aligned} \tag{34}$$

The corresponding values for δ_k^{true} and δ_k are found to be:

$$\begin{aligned}
\delta_k &= -(J_k J_k^T)^{-1} J_k (g_k + H_k v_k + H_k u_k - \rho_k) - y_k \\
\delta_k^{\text{true}} &= -(J_k J_k^T)^{-1} J_k (\nabla f_k + H_k v_k + H_k u_k^{\text{true}}) - y_k.
\end{aligned} \tag{35}$$

In the proof of the lemma below, we use the fact that

$$\|I - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k\| \leq 1, \tag{36}$$

which can be seen as follows: The nonzero eigenvalues of a matrix product AB are equal to the nonzero eigenvalues of BA when the product is valid, from which it follows that the nonzero eigenvalues of $Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k$ are precisely the eigenvalues of $Z_k^T H_k Z_k (Z_k^T H_k Z_k)^{-1} = I$, which are all equal to one; hence, the bound in (36) holds.

Lemma 11. *There exists $\omega_5 \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, it holds that*

$$\|\mathbb{E}_k[u_k - u_k^{\text{true}}]\| \leq \omega_5 \beta_k \quad \textit{and} \quad \mathbb{E}_k[\|u_k - u_k^{\text{true}}\|] \leq \zeta^{-1}\sqrt{M_g} + \omega_5 \beta_k$$

Proof. It follows from (34) that

$$u_k - u_k^{\text{true}} = J_k^T(J_kJ_k^T)^{-1}r_k - Z_k(Z_k^TH_kZ_k)^{-1}Z_k^T(g_k - \nabla f_k + H_kJ_k^T(J_kJ_k^T)^{-1}r_k - \rho_k),$$

which combined with Assumption 2 shows that

$$\begin{aligned}\mathbb{E}_k[u_k - u_k^{\text{true}}] =&(I - Z_k(Z_k^TH_kZ_k)^{-1}Z_k^TH_k)J_k^T(J_kJ_k^T)^{-1}\mathbb{E}_k[r_k]\\ &+ Z_k(Z_k^TH_kZ_k)^{-1}Z_k^T\mathbb{E}_k[\rho_k].\end{aligned}$$

Combining this equation with the triangle inequality, Assumptions 3 and 1, (15), and (36) ensures the existence of $\omega_5 \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$,

$$\begin{aligned}\|\mathbb{E}_k[u_k - u_k^{\text{true}}]\| &\leq \|J_k^T(J_kJ_k^T)^{-1}\|\|\mathbb{E}_k[r_k]\| + \zeta^{-1}\|\mathbb{E}_k[\rho_k]\|\\ &\leq \|J_k^T(J_kJ_k^T)^{-1}\|\kappa_r\beta_k + \zeta^{-1}\kappa_\rho\beta_k \leq \omega_5\beta_k,\end{aligned}$$

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 above for $u_k - u_k^{\text{true}}$ with the triangle inequality to obtain

$$\begin{aligned}\|u_k - u_k^{\text{true}}\| \leq& \|Z_k(Z_k^TH_kZ_k)^{-1}Z_k^T(g_k - \nabla f_k)\| + \|Z_k(Z_k^TH_kZ_k)^{-1}Z_k^T\rho_k\|\\ &+ \|(I - Z_k(Z_k^TH_kZ_k)^{-1}Z_k^TH_k)J_k^T(J_kJ_k^T)^{-1}r_k\|.\end{aligned}$$

Taking conditional expectation and using Assumption 2, (8), (36), and (15),

$$\begin{aligned}\mathbb{E}_k[\|u_k - u_k^{\text{true}}\|] &\leq \zeta^{-1}\sqrt{M_g} + \zeta^{-1}\mathbb{E}_k[\|\rho_k\|] + \|J_k^T(J_kJ_k^T)^{-1}\|\mathbb{E}_k[\|r_k\|]\\ &\leq \zeta^{-1}\sqrt{M_g} + \zeta^{-1}\kappa_\rho\beta_k + \|J_k^T(J_kJ_k^T)^{-1}\|\kappa_r\beta_k \leq \zeta^{-1}\sqrt{M_g} + \omega_5\beta_k,\end{aligned}$$

where ω_5 is the same value as used above, which completes the proof. □

We now bound the difference (in expectation) between $\nabla f_k^T d_k^{\text{true}}$ and $g_k^T d_k$.

Lemma 12. *There exist $(\omega_6, \omega_7) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$,*

$$|\mathbb{E}_k[\nabla f_k^T d_k^{\text{true}} - g_k^T d_k]| \leq \omega_6\beta_k + \omega_7\beta_k\sqrt{M_g} + \zeta^{-1}M_g.$$

Proof. It follows from the triangle inequality and linearity of E_k that

$$\begin{aligned}|\mathbb{E}_k[\nabla f_k^T d_k^{\text{true}} - g_k^T d_k]| &= |\mathbb{E}_k[\nabla f_k^T(d_k^{\text{true}} - d_k) + (\nabla f_k - g_k)^T d_k]|\\ &\leq |\nabla f_k^T\mathbb{E}_k[d_k^{\text{true}} - d_k]| + |\mathbb{E}_k[(\nabla f_k - g_k)^T d_k]|.\end{aligned}$$

For the first term on the right-hand side, it follows by the Cauchy-Schwarz inequality, $d_k^{\text{true}} = v_k + u_k^{\text{true}}$, $d_k = v_k + u_k$, and Lemma 11 that there exists $\omega_6 \in \mathbb{R}_{>0}$ with

$$\begin{aligned}|\nabla f_k^T\mathbb{E}_k[d_k^{\text{true}} - d_k]| &\leq \|\nabla f_k\|\|\mathbb{E}_k[d_k^{\text{true}} - d_k]\|\\ &= \|\nabla f_k\|\|\mathbb{E}_k[u_k^{\text{true}} - u_k]\| \leq \omega_6\beta_k.\end{aligned}$$

For the second term on the right-hand side, first observe from Assumption 2 that $\mathbb{E}_k[(\nabla f_k - g_k)^T v_k] = v_k^T\mathbb{E}_k[\nabla f_k - g_k] = 0$. Combining this fact with (34), the triangle inequality, the Cauchy-Schwarz inequality,

Assumptions 1–3, (36), and (8) shows that there exist $(\bar{\omega}_7, \omega_7) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that

$$\begin{aligned}
&|\mathbb{E}_k[(\nabla f_k - g_k)^T d_k]| \\
&= |\mathbb{E}_k[(\nabla f_k - g_k)^T ((I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^T (J_k J_k^T)^{-1} r_k \\
&\quad - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (g_k - \nabla f_k - \rho_k))]| \\
&\leq |\mathbb{E}_k[(\nabla f_k - g_k)^T (I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^T (J_k J_k^T)^{-1} r_k]| \\
&\quad + |\mathbb{E}_k[(\nabla f_k - g_k)^T Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T \rho_k]| \\
&\quad + |\mathbb{E}_k[(\nabla f_k - g_k)^T Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f_k - g_k)]| \\
&\leq \mathbb{E}_k[\|\nabla f_k - g_k\| \|(I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^T (J_k J_k^T)^{-1}\| \|r_k\|] \\
&\quad + \mathbb{E}_k[\|\nabla f_k - g_k\| \|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\| \|\rho_k\|] + \zeta^{-1} \mathbb{E}_k[\|\nabla f_k - g_k\|^2] \\
&\leq (\bar{\omega}_7 \kappa_r \beta_k + \zeta^{-1} \kappa_\rho \beta_k) \mathbb{E}_k[\|\nabla f_k - g_k\|] + \zeta^{-1} M_g \\
&\leq (\bar{\omega}_7 \kappa_r + \zeta^{-1} \kappa_\rho) \beta_k \sqrt{M_g} + \zeta^{-1} M_g = \omega_7 \beta_k \sqrt{M_g} + \zeta^{-1} M_g.
\end{aligned}$$

Combining the results above gives the desired result. □

We now proceed to bound (in expectation) the last two terms appearing in the right-hand side of the inequality proved in Lemma 10.

Lemma 13. *There exists $\omega_8 \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, it holds that*

$$\mathbb{E}_k[\alpha_k(\|c_k + J_k d_k\| - \|c_k + J_k v_k\|)] \leq \omega_8 \beta_k^2.$$

Proof. From the triangle inequality, (12), (15), the fact that $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$, (28), (25), (24), and the monotonically nonincreasing behavior of $\{\tau_k\}$ and $\{\xi_k\}$, it follows that there exists $\omega_8 \in \mathbb{R}_{>0}$ such that

$$\begin{aligned}
&\mathbb{E}_k[\alpha_k(\|c_k + J_k d_k\| - \|c_k + J_k v_k\|)] \leq \mathbb{E}_k[\alpha_k \|J_k u_k\|] = \mathbb{E}_k[\alpha_k \|r_k\|] \\
&\leq \kappa_r \beta_k \mathbb{E}_k[\alpha_k^{\max}] \leq \kappa_r \beta_k \mathbb{E}_k[\alpha_k^{\min} + \theta \beta_k^2] = \kappa_r \beta_k \mathbb{E}_k[(\tfrac{2(1-\eta)\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} + \theta \beta_k^2)] \\
&\leq (\tfrac{2(1-\eta)\xi_{-1}\tau_{-1}}{\Gamma} + \theta \beta_k) \kappa_r \beta_k^2 \leq \omega_8 \beta_k^2,
\end{aligned}$$

which gives the desired conclusion. □

3.3 Convergence analysis

Our goal now is to prove a convergence result for our algorithm. In general, in a run of the algorithm, one of three possible events can occur. 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 below. 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 [3, Section 3.2.2], the former of these two events does not occur if the differences between the stochastic gradient estimates and the true gradients of the objective remain uniformly bounded in norm, and the latter of these two events occurs with probability zero in a given run of the algorithm if one makes a reasonable assumption about the influence of the stochastic gradient estimates on the computed search directions; see also [2, Section 4.3] for additional discussion of the latter case in the context of an algorithm that employs a step decomposition approach, as does our algorithm. For our purposes here, we do not consider these latter two events since we contend that, for practical purposes, they can be ignored for the same reasons as are claimed in [3].

To define our event of interest, consider for each $k \in \mathbb{N}$ the condition

$$\nabla f_k^T d_k^{\text{true}} + \max\{(u_k^{\text{true}})^T H_k u_k^{\text{true}}, \epsilon_u \|u_k^{\text{true}}\|^2\} \leq 0 \tag{37}$$

(similar to the one appearing in (19)). With this condition, let us define the following trial value of the merit parameter that would be computed in iteration $k \in \mathbb{N}$ (conditioned on x_k being the kth iterate) if the algorithm were to employ ∇f_k in place of g_k and compute an exact solution of the linear system (9):

$$\tau_k^{\text{true,trial}} \leftarrow \begin{cases} \infty & \text{if (37) holds,} \\ \frac{(1-\frac{\sigma_c}{\epsilon_\tau})(\|c_k\|-\|c_k+J_k d_k^{\text{true}}\|)}{\nabla f_k^T d_k^{\text{true}}+\max\{(u_k^{\text{true}})^T H_k u_k^{\text{true}}, \epsilon_u \|u_k^{\text{true}}\|^2\}} & \text{if (37) does not hold.} \end{cases}$$

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

Event $E_{\tau,\text{low}}$. Event $E_{\tau,\text{low}}$ occurs if and only if there exists an iteration number $k_{\tau,\xi} \in \mathbb{N}$ such that, with $\xi_{\min}$ given in Lemma 9, it holds that

$$\tau_k = \tau_{k_{\tau,\xi}} \leq \tau_k^{\text{true,trial}} \quad \text{and} \quad \xi_k = \xi_{k_{\tau,\xi}} \geq \xi_{\min} \quad \text{for all} \quad k \geq k_{\tau,\xi}. \tag{38}$$

For our analysis in this subsection, the following supersedes Assumption 2.

Assumption 7. *There exists $M_g \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the stochastic gradient g_k has the properties that $\mathbb{E}_{k,\tau,\text{low}}[g_k] = \nabla f_k$ and $\mathbb{E}_{k,\tau,\text{low}}[\|g_k - \nabla f_k\|_2^2] \leq M_g$, where $\mathbb{E}_{k,\tau,\text{low}}[\cdot]$ denotes expectation with respect to the distribution of ω conditioned on the event that $E_{\tau,\text{low}}$ occurs and x_k is the primal iterate in iteration $k \in \mathbb{N}$.*

Our results in this subsection focus on $k \in \mathbb{N}$ with $k \geq k_{\tau,\xi} + 1$, at which point, in any run in which Event $E_{\tau,\text{low}}$ occurs, the merit parameter satisfies $\tau_k = \tau_{k_{\tau,\xi}}$ independently from the stochastic gradient g_k that is generated.

Our first 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. *Under Event $E_{\tau,\text{low}}$, there exists $\omega_9 \in \mathbb{R}_{>0}$ such that*

$$\mathbb{E}_{k,\tau,\text{low}}[\alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}})] \leq \omega_9 \beta_k^2 \quad \textit{for all} \quad k \geq k_{\tau,\xi} + 1.$$

Proof. Under Assumption 7, the logic as in the proof of Lemma 11 allows us to conclude that, under $E_{\tau,low}$, it holds for all $k \in \mathbb{N}$ that

$$\|\mathbb{E}_{k,\tau,\text{low}}[u_k - u_k^{\text{true}}]\| \leq \omega_5 \beta_k \quad \text{and} \quad \mathbb{E}_{k,\tau,\text{low}}[\|u_k - u_k^{\text{true}}\|] \leq \zeta^{-1}\sqrt{M_g} + \omega_5 \beta_k. \tag{39}$$

Let E_k be the event that $\nabla f_k^T (d_k - d_k^{\text{true}}) \geq 0$ and let E_k^c be its complementary event. Let $\mathbb{P}_{k,\tau,\text{low}}[\cdot]$ denote probability conditioned on the occurrence of event $E_{\tau,\text{low}}$ and x_k being the kth primal iterate. It now follows from (38), the definition of E_k, the fact that $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$, and the Law of Total Expectation that for all $k \geq k_{\tau,\xi} + 1$

$$\begin{aligned}
&\mathbb{E}_{k,\tau,\text{low}}[\alpha_k \tau_k \nabla f_k^T (d_k - d_k^{\text{true}})] \\
&= \mathbb{E}_{k,\tau,\text{low}}[\alpha_k \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}}) | E_k] \mathbb{P}_{k,\tau,\text{low}}[E_k] \\
&\quad + \mathbb{E}_{k,\tau,\text{low}}[\alpha_k \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}}) | E_k^c] \mathbb{P}_{k,\tau,\text{low}}[E_k^c] \\
&\leq \mathbb{E}_{k,\tau,\text{low}}[\alpha_k^{\max} \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}}) | E_k] \mathbb{P}_{k,\tau,\text{low}}[E_k] \\
&\quad + \mathbb{E}_{k,\tau,\text{low}}[\alpha_k^{\min} \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}}) | E_k^c] \mathbb{P}_{k,\tau,\text{low}}[E_k^c] \\
&= \mathbb{E}_{k,\tau,\text{low}}[(\alpha_k^{\max} - \alpha_k^{\min}) \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}}) | E_k] \mathbb{P}_{k,\tau,\text{low}}[E_k] \\
&\quad + \mathbb{E}_{k,\tau,\text{low}}[\alpha_k^{\min} \tau_{k_{\tau,\xi}} \nabla f_k^T (d_k - d_k^{\text{true}})].
\end{aligned}$$

Combining this with the fact that (28) ensures $\alpha_k^{\max} - \alpha_k^{\min} \leq \theta\beta_k^2$, the Cauchy-Schwarz inequality, the fact that $\alpha_k^{\min} = 2(1-\eta)\beta_k\xi_{k_{\tau,\xi}}\tau_{k_{\tau,\xi}}/(\tau_{k_{\tau,\xi}}L + \Gamma)$ for all $k \geq k_{\tau,\xi} + 1$, and the Law of Total Expectation shows for all $k \geq k_{\tau,\xi} + 1$ that

$$\begin{aligned}\mathbb{E}_{k,\tau,\text{low}}[\alpha_k\tau_k\nabla f_k^T(d_k - d_k^{\text{true}})] &\leq \theta\beta_k^2\tau_{k_{\tau,\xi}}\|\nabla f_k\|\mathbb{E}_{k,\tau,\text{low}}[\|d_k - d_k^{\text{true}}\||E_k]\mathbb{P}_{k,\tau,\text{low}}[E_k] \\ &\quad + \tfrac{2(1-\eta)\beta_k\xi_{k_{\tau,\xi}}\tau_{k_{\tau,\xi}}}{\tau_{k_{\tau,\xi}}L+\Gamma}\tau_{k_{\tau,\xi}}\|\nabla f_k\|\|\mathbb{E}_{k,\tau,\text{low}}[d_k - d_k^{\text{true}}]\| \\ &\leq \theta\beta_k^2\tau_{k_{\tau,\xi}}\|\nabla f_k\|\mathbb{E}_{k,\tau,\text{low}}[\|d_k - d_k^{\text{true}}\|] \\ &\quad + \tfrac{2(1-\eta)\beta_k\xi_{k_{\tau,\xi}}\tau_{k_{\tau,\xi}}}{\tau_{k_{\tau,\xi}}L+\Gamma}\tau_{k_{\tau,\xi}}\|\nabla f_k\|\|\mathbb{E}_{k,\tau,\text{low}}[d_k - d_k^{\text{true}}]\|\end{aligned}$$

Combining this with (39), (24), $\|d_k - d_k^{\text{true}}\| = \|v_k + u_k - (v_k + u_k^{\text{true}})\| = \|u_k - u_k^{\text{true}}\|$, and Assumption 1 shows there exists $\omega_9 \in \mathbb{R}_{>0}$ where, for all $k \geq k_{\tau,\xi} + 1$,

$$\begin{aligned}&\mathbb{E}_{k,\tau,\text{low}}[\alpha_k\tau_k\nabla f_k^T(d_k - d_k^{\text{true}})] \\ &\leq \theta\beta_k^2\tau_{k_{\tau,\xi}}\|\nabla f_k\|(\zeta^{-1}\sqrt{M_g} + \omega_5\beta_k) + \tfrac{2(1-\eta)\beta_k\xi_{k_\xi}\tau_{k_{\tau,\xi}}}{\tau_{k_{\tau,\xi}}L+\Gamma}\tau_{k_{\tau,\xi}}\|\nabla f_k\|\omega_5\beta_k \leq \omega_9\beta_k^2,\end{aligned}$$

which is the desired conclusion. □

We now use the model reduction based on the true step d_k^{true} to build an upper bound on the (expected) reduction in the model based on the step d_k.

Lemma 15. *Under Event $E_{\tau,\text{low}}$, it holds for all $k \geq k_{\tau,\xi} + 1$ that*

$$\begin{aligned}&\mathbb{E}_{k,\tau,\text{low}}[\Delta l(x_k, \tau_k, g_k, d_k)] \\ &\leq \Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + \kappa_r\beta_k + \tau_{k_{\tau,\xi}}(\omega_6\beta_k + \omega_7\beta_k\sqrt{M_g} + \zeta^{-1}M_g).\end{aligned}$$

Proof. Under Assumption 7, the logic as in the proof of Lemma 12 allows us to conclude that, under $E_{\tau,\text{low}}$, it holds for all $k \in \mathbb{N}$ that

$$|\mathbb{E}_{k,\tau,\text{low}}[\nabla f_k^T d_k^{\text{true}} - g_k^T d_k]| \leq \omega_6\beta_k + \omega_7\beta_k\sqrt{M_g} + \zeta^{-1}M_g.$$

It follows from this, (4), the fact that $d_k = v_k + u_k$, the triangle inequality, the fact that c_k, J_k, v_k, ∇f_k, and d_k^{true} are all deterministic conditioned on x_k as the kth primal iterate, (10), and (15) that for all $k \geq k_{\tau,\xi} + 1$

$$\begin{aligned}&\mathbb{E}_{k,\tau,\text{low}}[\Delta l(x_k, \tau_k, g_k, d_k)] = \mathbb{E}_{k,\tau,\text{low}}[-\tau_{k_{\tau,\xi}}g_k^T d_k + \|c_k\| - \|c_k + J_k d_k\|] \\ &\leq \Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + \kappa_r\beta_k + \tau_{k_{\tau,\xi}}(\omega_6\beta_k + \omega_7\beta_k\sqrt{M_g} + \zeta^{-1}M_g),\end{aligned}$$

which is the desired result. □

For the final result of this section, we define

$$\mathbb{E}_{\tau,\text{low}}[\cdot] = \mathbb{E}[\,\cdot \mid \text{Event } E_{\tau,\text{low}} \text{ occurs and Assumption 7 holds}]. \tag{40}$$

In the result, the quantity $\Delta l(x_k, \tau_k, \nabla f_k, d_k^{\text{true}})$ serves as a measure of stationarity with respect to (1); after all, the proof for Lemma 8 shows, with $(\nabla f_k, u_k^{\text{true}}, d_k^{\text{true}})$ in place of (g_k, u_k, d_k), that by Assumption 7 it follows for $k \geq k_{\tau,\xi} + 1$ that

$$\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) \geq \kappa_l\tau_{k_{\tau,\xi}}(\|u_k^{\text{true}}\|^2 + \|c_k\|) \geq \tfrac{\kappa_l\tau_{k_{\tau,\xi}}}{\omega_4}\|d_k^{\text{true}}\|^2 > 0. \tag{41}$$

Thus, if there is an infinite $\mathcal{K} \subseteq \mathbb{N}$ with $\lim_{k\in\mathcal{K},k\to\infty} \Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) = 0$, then it follows from (41) and (6) 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 (10) shows that any limit point of $\{(x_k, y_k + \delta_k^{\text{true}})\}$ is a first-order stationary point for (1). In our stochastic setting, we cannot guarantee that such a limit holds surely. Rather, in the following result, we prove for two different choices of $\{\beta_k\}$ that an expected average of this measure of stationarity exhibits desirable properties. These properties match those ensured by a stochastic gradient method in the unconstrained setting (where $\|\nabla f_k\|^2$ plays the role of the measure of stationarity for the minimization of f).

Theorem 1. *Under Event $E_{\tau,\text{low}}$, let $k_{\tau,\xi}$ be defined as in (38) and define $\overline{A} = \frac{2(1-\eta)\xi_{\min}\tau_{k_{\tau,\xi}}}{\tau_{k_{\tau,\xi}}L+\Gamma}$ and $\overline{M} = (1-\eta)(\overline{A}+\theta)\big(\kappa_r + \tau_{k_{\tau,\xi}}(\omega_6 + \omega_7\sqrt{M_g} + \zeta^{-1}M_g)\big) + \omega_8 + \omega_9$, where $\xi_{\min}$ is defined in Lemma 9. Then, the following results hold:*

(i) If $\beta_k = \beta \in (0, \overline{A}/((1-\eta)(\overline{A}+\theta)))$ for all $k \geq k_{\tau,\xi}+1$, then

$$\begin{aligned} &\mathbb{E}_{\tau,\text{low}}\left[\frac{1}{K}\sum_{j=k_{\tau,\xi}+1}^{k_{\tau,\xi}+K}\Delta l(x_j, \tau_{k_{\tau,\xi}}, \nabla f_j, d_j^{\text{true}})\right] \\ &\leq \tfrac{\beta\overline{M}}{\overline{A}-(1-\eta)(\overline{A}+\theta)\beta} + \tfrac{\mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+1}, \tau_{k_{\tau,\xi}})] - \phi_{\min}}{K\beta(\overline{A}-(1-\eta)(\overline{A}+\theta)\beta)} \xrightarrow{K\to\infty} \tfrac{\beta\overline{M}}{\overline{A}-(1-\eta)(\overline{A}+\theta)\beta} \end{aligned} \tag{42}$$

where $\phi_{\min} \in \mathbb{R}$ is a lower bound of $\phi(\cdot, \tau_{k_{\tau,\xi}})$ over $\mathcal{X}$ (by Assumption 1).

(ii) If $\{\beta_k\}_{k\geq k_{\tau,\xi}+1}$ satisfies $\sum_{k=k_{\tau,\xi}+1}^{\infty}\beta_k = \infty$ and $\sum_{k=k_{\tau,\xi}+1}^{\infty}\beta_k^2 < \infty$, then

$$\lim_{K\to\infty}\mathbb{E}_{\tau,\text{low}}\left[\frac{1}{\sum_{j=k_{\tau,\xi}+1}^{k_{\tau,\xi}+K}\beta_j}\sum_{j=k_{\tau,\xi}+1}^{k_{\tau,\xi}+K}\beta_j\Delta l(x_j, \tau_{k_{\tau,\xi}}, \nabla f_j, d_j^{\text{true}})\right] = 0. \tag{43}$$

Proof. By the definition of $\overline{A}$, the fact that $\{\beta_k\} \subset (0,1]$, and line 13 of Algorithm 1, it follows that $\alpha_k \in [\overline{A}\beta_k, (\overline{A}+\theta)\beta_k]$ for all $k \geq k_{\tau,\xi}+1$. It follows from this fact, $\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) > 0$ (see (41)), Lemmas 10, 14, 8, 13, and 15, and the fact that $\{\beta_k\} \subset (0,1]$ that, for all $k \geq k_{\tau,\xi}+1$, one finds

$$\begin{aligned} &\mathbb{E}_{k,\tau,\text{low}}[\phi(x_k + \alpha_k d_k, \tau_{k_{\tau,\xi}})] - \phi(x_k, \tau_{k_{\tau,\xi}}) \\ &\leq \mathbb{E}_{k,\tau,\text{low}}[-\alpha_k\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + \alpha_k\tau_{k_{\tau,\xi}}\nabla f_k^T(d_k - d_k^{\text{true}})] \\ &\quad + (1-\eta)\mathbb{E}_{k,\tau,\text{low}}[\alpha_k\beta_k\Delta l(x_k, \tau_{k_{\tau,\xi}}, g_k, d_k)] \\ &\quad + \mathbb{E}_{k,\tau,\text{low}}[\alpha_k(\|c_k + J_k d_k\| - \|c_k + J_k v_k\|)] \\ &\leq -\overline{A}\beta_k\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + (\omega_8+\omega_9)\beta_k^2 \\ &\quad + (1-\eta)(\overline{A}+\theta)\beta_k^2\mathbb{E}_{k,\tau,\text{low}}[\Delta l(x_k, \tau_{k_{\tau,\xi}}, g_k, d_k)] \\ &\leq (-\overline{A}\beta_k + (1-\eta)(\overline{A}+\theta)\beta_k^2)\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + \beta_k^2\overline{M} \\ &= -\beta_k\big(\overline{A} - (1-\eta)(\overline{A}+\theta)\beta_k\big)\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}}) + \beta_k^2\overline{M}. \end{aligned} \tag{44}$$

Let us now consider the two cases in the theorem one at a time.

Case (i). By the definition of β, it follows by taking total expectation of (44) (namely, expectation defined in (40)) that for each $k \geq k_{\tau,\xi}+1$ one has

$$\begin{aligned} &\mathbb{E}_{\tau,\text{low}}[\phi(x_k + \alpha_k d_k, \tau_{k_{\tau,\xi}})] - \mathbb{E}_{\tau,\text{low}}[\phi(x_k, \tau_{k_{\tau,\xi}})] \\ &\leq -\beta(\overline{A} - (1-\eta)(\overline{A}+\theta)\beta)\mathbb{E}_{\tau,\text{low}}[\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}})] + \beta^2\overline{M}. \end{aligned}$$

Summing this inequality over $j \in \{k_{\tau,\xi}+1, \ldots, k_{\tau,\xi}+K\}$ shows that

$$\begin{aligned} &\phi_{\min} - \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+1}, \tau_{k_{\tau,\xi}})] \\ &\leq \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+K+1}, \tau_{k_{\tau,\xi}})] - \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+1}, \tau_{k_{\tau,\xi}})] \\ &\leq -\beta(\overline{A} - (1-\eta)(\overline{A}+\theta)\beta)\mathbb{E}_{\tau,\text{low}}\left[\sum_{j=k_{\tau,\xi}+1}^{k_{\tau,\xi}+K}\Delta l(x_j, \tau_{k_{\tau,\xi}}, \nabla f_j, d_j^{\text{true}})\right] + K\beta^2\overline{M}, \end{aligned}$$

which after rearrangement shows that (42) holds, as desired.

Case (ii). Given the definition of $\{\beta_k\}$, let us assume without loss of generality that $\beta_k \leq \overline{A}/\big(2(1-\eta)(\overline{A}+\theta)\big)$ for all $k \geq k_{\tau,\xi}+1$, which implies that $\overline{A}-(1-\eta)(\overline{A}+\theta)\beta_k \geq \frac{1}{2}\overline{A}$ for all $k \geq k_{\tau,\xi}+1$. Using this fact, taking total expectation of (44) (namely, expectation defined in (40)), and using (41) it holds that

$$\begin{aligned} &\mathbb{E}_{\tau,\text{low}}[\phi(x_k+\alpha_k d_k, \tau_{k_{\tau,\xi}})] - \mathbb{E}_{\tau,\text{low}}[\phi(x_k, \tau_{k_{\tau,\xi}})] \\ \leq\ & -\tfrac{1}{2}\beta_k \overline{A}\mathbb{E}_{\tau,\text{low}}[\Delta l(x_k, \tau_{k_{\tau,\xi}}, \nabla f_k, d_k^{\text{true}})] + \beta_k^2 \overline{M}. \end{aligned}$$

Summing this inequality over $j \in \{k_{\tau,\xi}+1, \ldots, k_{\tau,\xi}+K\}$ shows that

$$\begin{aligned} &\phi_{\min} - \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+1}, \tau_{k_{\tau,\xi}})] \\ &\leq \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+K+1}, \tau_{k_{\tau,\xi}})] - \mathbb{E}_{\tau,\text{low}}[\phi(x_{k_{\tau,\xi}+1}, \tau_{k_{\tau,\xi}})] \\ &\leq -\tfrac{1}{2}\overline{A}\mathbb{E}_{\tau,\text{low}}\left[\sum_{j=k_{\tau,\xi}+1}^{k_{\tau,\xi}+K} \beta_j \Delta l(x_j, \tau_{k_{\tau,\xi}}, \nabla f_j, d_j^{\text{true}})\right] + \overline{M}\sum_{j=k_{\tau\xi}+1}^{k_{\tau,\xi}+K} \beta_j^2, \end{aligned}$$

which after rearrangement and taking limits proves that (43) holds. □

4 Numerical Results

In this section, we demonstrate the performance of a Matlab implementation of Algorithm 1 for solving (i) a subset of the CUTEst collection of test problems [15] and (ii) two optimal control problems from [20]. 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 the conjugate gradient (CG) method to $J_k^T J_k v = -J_k^T c_k$. 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 $\|J_k^T J_k v_{k,t} + J_k^T c_k\| \leq \max\{0.1\|J_k^T c_k\|, 10^{-10}\}$. The properties of the CG method as a Krylov subspace method ensure that $v_{k,t} \in \text{Range}(J_k^T)$ for all $t \in \mathbb{N}$ (in exact arithmetic); hence, $v_k \in \text{Range}(J_k^T)$.

To obtain the tangential direction u_k and associated dual search direction δ_k, we applied the minimum residual (MINRES) method, namely, the implementation from [10, 27], to the linear system

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u \\ \delta \end{bmatrix} = -\begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}. \tag{45}$$

(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 $\kappa \in (0,1)$,

$$\left\| \begin{bmatrix} \rho_{k,t} \\ r_{k,t} \end{bmatrix} \right\|_\infty \leq \max\left\{ \kappa \left\| \begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix} \right\|_\infty, 10^{-12} \right\} \tag{46}$$

and Termination Test 1 and/or 2 holds. (Recall the definition of $(\rho_{k,t}, r_{k,t})$ in (30).) The choice of $\kappa \in (0,1)$ is discussed along with each set of experiments.

4.2 Choosing the step size

Algorithm 1 (see line 13) 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 the inequalities $\alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq 1$ and $\alpha_k^{\text{suff}} \leq \alpha_k^{\varphi}$ (see Lemma 4)

hold, we take advantage of this flexibility in choosing the step size by defining

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

where t_k is the largest value of $t \in \mathbb{N}$ such that

$$(1.1)^t\alpha_k^{\text{suff}} \le \min\{\alpha_k^{\varphi}, \alpha_k^{\min} + \theta\beta_k^2\} \equiv \alpha_k^{\max} \text{ and } (1.1)^{t-1}\alpha_k^{\text{suff}} < 1.$$

When $\alpha_k^{\text{suff}} < 1$, this strategy allows for the possibility that step sizes larger than $\min\{\alpha_k^{\text{suff}}, \alpha_k^{\min} + \theta\beta_k^2\}$ be taken (namely, they can be as large as $\min\{\alpha_k^{\varphi}, \alpha_k^{\min} + \theta\beta_k^2\}$). This led to better performance while still having a rule that satisfies the requirements of our analysis. We do not explicitly compute α_k^{φ} in our code. Instead, we can verify directly whether $(1.1)^t\alpha_k^{\text{suff}} \le \alpha_k^{\varphi}$ (as needed above) since this is ensured by checking whether $\varphi((1.1)^t\alpha_k^{\text{suff}}) \le 0$, which is easily checked by the code.

4.3 Algorithm variants tested

To test the utility of using inexact subproblem solutions in Algorithm 1, we consider two algorithm variants that we refer to as `SISQO` and `SISQO_exact`. The variant `SISQO` is Algorithm 1 with inexact solutions computed as described in Section 4.1 with a relatively *large* value for κ in (46). On the other hand, the variant `SISQO_exact` is identical to `SISQO` with the exception that it uses a relatively *small* value for κ in (46). 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. For this reason, we allow both variants to use the same computation for the normal step, thus allowing any numerical gains to be directly attributed to the inexact tangential step computation. Although other variants could be tested (e.g., allowing the normal step computation to differ as well) we prefer the approach described above since it limits the variation attributable to the different calculations in the SISQO framework.

4.4 Metrics used for comparison

Our metrics of interest are feasibility and stationarity. Specifically, for any run of `SISQO`, we terminate with $x_{\texttt{SISQO}} \leftarrow x_k$, where $k \in \mathbb{N}$ is the first iteration such that $\|c(x_k)\|_\infty \le 10^{-6}$ and $\|\nabla f_k + J_k^T y_{k,\text{ls}}\|_\infty \le 10^{-2}$, where $y_{k,\text{ls}}$ is the least-square multiplier at x_k. (The computations of ∇f_k and $y_{k,\text{ls}}$ are not required by our algorithm in general; they were computed in our experiments merely for the purpose of being able to determine an accurate measure of stationarity at x_k.) This allows us to associate with each run of `SISQO` the two measures

$$\text{error}_{\text{feasibility}}(\texttt{SISQO}) = \|c(x_{\texttt{SISQO}})\|_\infty \text{ and}$$
$$\text{error}_{\text{stationarity}}(\texttt{SISQO}) = \|\nabla f(x_{\texttt{SISQO}}) + J(x_{\texttt{SISQO}})^T y_{\texttt{SISQO}}\|_\infty,$$

where $y_{\texttt{SISQO}} \in \mathbb{R}^m$ is the least-square multiplier at $x_{\texttt{SISQO}}$. We use the total number of MINRES iterations performed by `SISQO` as a budget for the number of MINRES iterations performed by `SISQO_exact`; no other termination condition is used for `SISQO_exact`. Upon termination of `SISQO_exact`, we define $x_{\texttt{SISQO\_exact}}$ in the following manner: If an iterate is computed with $\|c(x_k)\|_\infty \le 10^{-6}$, then $x_{\texttt{SISQO\_exact}}$ is chosen as the iterate with smallest stationarity measure among those satisfying this tolerance for the feasibility measure; otherwise, $x_{\texttt{SISQO\_exact}}$ is chosen as the iterate with the smallest feasibility measure. In any case, once $x_{\texttt{SISQO\_exact}}$ is determined, we proceed to compute the least-square multiplier $y_{\texttt{SISQO\_exact}}$ at $x_{\texttt{SISQO\_exact}}$, then define

$$\text{error}_{\text{feasibility}}(\texttt{SISQO\_exact}) = \|c(x_{\texttt{SISQO\_exact}})\|_\infty \text{ and}$$
$$\text{error}_{\text{stationarity}}(\texttt{SISQO\_exact}) = \|\nabla f(x_{\texttt{SISQO\_exact}}) + J(x_{\texttt{SISQO\_exact}})^T y_{\texttt{SISQO\_exact}}\|_\infty.$$

These are the metrics that we use in the next two subsections.

4.5 Results on the CUTEst problems

In the CUTEst collection [15], there are a total of 138 equality constrained problems with $m \leq n$. From these problems, we selected those such that (i) $(n+m) \in [500, 10000]$, (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 LICQ 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 problems from the CUTEst collection 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 to estimate Lipschitz constants and using a (modified) Hessian of the Lagrangian in each search direction computation, as explained below. However, we introduced noise into the computation of the objective function gradients. 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 2 would be given by $M_g = \{10^{-8}, 10^{-4}, 10^{-2}\}$, corresponding to the values for ϵ_N.

In terms of algorithm parameters, we set $\kappa = 0.1$ for `SISQO` and $\kappa = 10^{-7}$ for `SISQO_exact`. All of the remaining parameters were set identically for the two variants with the following values: $\tau_{-1} = \sigma_c = \eta = \kappa_v = \kappa_u = 0.1$, $\xi_{-1} = \epsilon_c = 1$, $\epsilon_\tau = \epsilon_\xi = 0.01$, $\kappa_\rho = \kappa_r = 100$, $\epsilon_r = 1 - 10^{-4}$, $\zeta = 10^{-8}$, $\epsilon_u = 5 \times 10^{-9}$, $\sigma_u = 1 - 10^{-12}$, $\theta = 10^4$, and $\beta_k = 1$ for all $k \in \mathbb{N}$. During each iteration $k \in \mathbb{N}$, we randomly generated a sample point near x_k, then 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.

For this collection of problems, we employed an iterative Hessian modification strategy as proposed in [7, 12]. Specifically, for all $k \in \mathbb{N}$, the matrix H_k is initialized to the true Hessian of the Lagrangian, but may be set ultimately as

$$H_k \leftarrow \iota_k \nabla^2_{xx}\big(f(x) + c(x)^T y\big)|_{(x,y)=(x_k,y_k)} + (1-\iota_k) I$$

with $\iota_k = 10^{-j_k}$, where j_k is the smallest element in $\{0, \ldots, 10\}$ such that a modification is not triggered. If a modification is triggered at $\iota_k = 10^{-10}$, then the algorithm sets $H_k \leftarrow I$ to guarantee that no further modifications are required.

For each test problem, we ran both `SISQO` and `SISQO_exact` five times, and for each computed the resulting feasibility and stationarity errors as described in Section 4.4. The results are shown in the form of box plots in Figure 1.

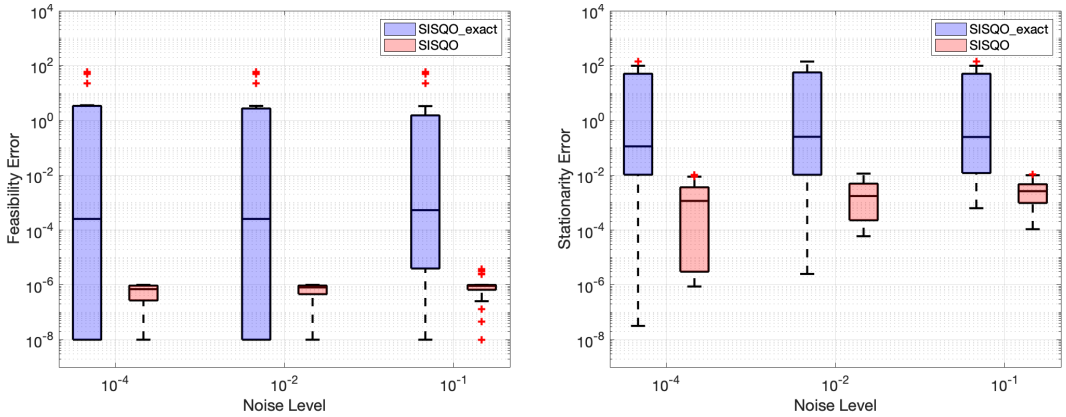

Figure 1: Box plots on CUTEst problems for feasibility (left) and stationarity (right).

From Figure 1, one finds that `SISQO` performs better than `SISQO_exact` in terms of both feasibility and

stationarity errors. Also, in general, SISQO achieves smaller feasibility and stationarity errors for smaller noise levels, which may be expected due to 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 [20]. 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 $\bar{z}_{ij} \in L^2(\Xi)$ for $(i,j) \in \{1,\dots,N\} \times \{1,\dots,N\}$, and a regularization parameter $\lambda \in \mathbb{R}_{>0}$, we first considered the problem

$$\begin{aligned} \min_{w,z} \ & \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\tfrac{1}{2} \|w - \overline{w}_{ij}\|^2_{L^2(\Xi)} + \tfrac{\lambda}{2} \|z - \bar{z}_{ij}\|^2_{L^2(\Xi)} \right) \\ \text{s.t.} \ & -\Delta w = z \text{ in } \Xi, \text{ and } w = 0 \text{ on } \partial\Xi. \end{aligned} \tag{47}$$

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

$$\begin{aligned} \min_{w,z} \ & \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\tfrac{1}{2} \|w - \overline{w}_{ij}\|^2_{L^2(\Xi)} + \tfrac{\lambda}{2} \|z - \bar{z}_{ij}\|^2_{L^2(\partial\Xi)} \right) \\ \text{s.t.} \ & -\Delta w + w = 0 \text{ in } \Xi, \text{ and } \tfrac{\partial w}{\partial p} = z \text{ on } \partial\Xi, \end{aligned} \tag{48}$$

where p represents the unit outer normal to Ξ along $\partial\Xi$. As reference functions for both problems, we chose for all $(i,j) \in \{1,\dots,N\} \times \{1,\dots,N\}$ the following:

$$\bar{z}_{ij} = 0 \text{ and } \overline{w}_{ij}(x_1,x_2) = \sin((4 + \tfrac{\epsilon_N}{\epsilon_S}(i - \tfrac{N+1}{2}))x_1) + \cos((3 + \tfrac{\epsilon_N}{\epsilon_S}(j - \tfrac{N+1}{2}))x_2) \tag{49}$$

for some $(\epsilon_S, \epsilon_N) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. We selected the following values for the above constants: $N = 3$, $\lambda = 10^{-5}$, $\epsilon_S = \sqrt{15}$, and $\epsilon_N \in \{10^{-4}, 10^{-2}, 10^{-1}\}$. Since the objective functions of (47) and (48) are finite sums, to generate stochastic gradients as unbiased estimates of the true gradient, we first uniformly generated random $(i,j) \in \{1,\dots,N\} \times \{1,\dots,N\}$, then computed the gradient corresponding to the (i,j)th term in the objective function. We note that with the above choice of parameters, it follows that an appropriate value for M_g in Assumption 2 is given by $M_g \approx \{10^{-8}, 10^{-4}, 10^{-2}\}$ to correspond, respectively, to the above values for ϵ_N.

Since 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 3 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: $\tau_{-1} = 10^{-4}$, $\eta = 0.5$, and $L_k = 1$ and $\Gamma_k = 0$ for all $k \in \mathbb{N}$, which are valid choices since the objective functions are quadratic and the constraints are linear.

For each of the two optimal control problems in (47) and (48), we ran both SISQO and SISQO_exact ten times, then computed their average feasibility and stationarity errors as described in Section 4.4. In Table 1 and Table 2, we report these average values as well as the average number of iterations performed by Algorithm 1 before termination ("iterations") and number of MINRES iterations ("MINRES iterations"), with the latter discussed in Section 4.1. The results are given in Table 1 and Table 2 for problem (47) and problem (48), respectively. One can observe that despite performing more "outer" iterations on average, SISQO outperforms SISQO_exact due to the fact that it requires fewer overall linear system solver iterations on average in order to attain better average feasibility and stationarity 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

strategy	ϵ_N	feasibility error	stationarity error	MINRES iterations	iterations
`SISQO`	10^{-4}	2.41×10^{-7}	1.76×10^{-5}	55117	8.9
`SISQO_exact`	10^{-4}	3.86×10^{-5}	4.05×10^{-5}	55117	6.9
`SISQO`	10^{-2}	4.14×10^{-7}	2.09×10^{-3}	60894	8.8
`SISQO_exact`	10^{-2}	3.46×10^{-5}	1.95×10^{-3}	60894	6.8
`SISQO`	10^{-1}	3.43×10^{-7}	5.15×10^{-3}	93634	12.3
`SISQO_exact`	10^{-1}	2.36×10^{-6}	1.68×10^{-2}	93634	10

Table 1: Numerical results for problem (47) averaged over ten independent runs.

strategy	ϵ_N	feasibility error	stationarity error	MINRES iterations	iterations
`SISQO`	10^{-4}	3.29×10^{-7}	2.35×10^{-5}	91478	9.9
`SISQO_exact`	10^{-4}	5.44×10^{-4}	5.46×10^{-4}	91478	7.1
`SISQO`	10^{-2}	2.90×10^{-7}	2.07×10^{-3}	99921	10
`SISQO_exact`	10^{-2}	5.71×10^{-5}	2.37×10^{-3}	99921	7.6
`SISQO`	10^{-1}	1.68×10^{-7}	3.88×10^{-4}	158825	14.5
`SISQO_exact`	10^{-1}	1.31×10^{-5}	2.58×10^{-2}	158825	11.1

Table 2: Numerical results for problem (48) averaged over ten independent runs.

guarantee (in expectation) for our algorithm that is comparable to that proved for the *exact* stochastic SQP method recently presented in [3], which in turn is comparable to that known for the stochastic gradient in unconstrained settings [5]. Our Matlab implementation, `SISQO`, illustrated the benefits of allowing inexact step computation for solving problems from the CUTEst collection [15] as well as two optimal control problems.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31, 2017.

[2] Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient Jacobians. *arXiv preprint arXiv:2106.13015*, 2021.

[3] Albert S. Berahas, Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

[4] G. Biros and O. Ghattas. Inexactness Issues in the Lagrange-Newton-Krylov-Schur Method for PDE-constrained Optimization. In L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. Van Bloemen Waanders, editors, *Large-Scale PDE-Constrained Optimization*, pages 93–114, New York, NY, USA, 2003. Springer.

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[6] Richard H. Byrd, Frank E. Curtis, and Jorge Nocedal. An inexact SQP method for equality constrained optimization. *SIAM Journal on Optimization*, 19(1):351–369, 2008.

[7] Richard H Byrd, Frank E Curtis, and Jorge Nocedal. An inexact Newton method for nonconvex equality constrained optimization. *Mathematical programming*, 122(2):273, 2010.

[8] Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

[9] Changan Chen, Frederick Tung, Naveen Vedula, and Greg Mori. Constraint-aware deep neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 400–415, 2018.

[10] Sou-Cheng T Choi, Christopher C Paige, and Michael A Saunders. MINRES-QLP: A Krylov subspace method for indefinite or singular symmetric systems. *SIAM Journal on Scientific Computing*, 33(4):1810–1836, 2011.

[11] R Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49(1):1–23, 1943.

[12] Frank E Curtis, Jorge Nocedal, and Andreas Wächter. A matrix-free algorithm for equality constrained optimization problems with rank-deficient Jacobians. *SIAM Journal on Optimization*, 20(3):1224–1249, 2009.

[13] Roger Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, 2013.

[14] Charles J. Geyer. Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, 86(415):717–724, 1991.

[15] Nicholas I. M. Gould, Dominique Orban, and Philippe L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60:545–557, 2015.

[16] S-P Han and Olvi L Mangasarian. Exact penalty functions in nonlinear programming. *Mathematical programming*, 17(1):251–269, 1979.

[17] Shih-Ping Han. A globally convergent method for nonlinear programming. *Journal of optimization theory and applications*, 22(3):297–309, 1977.

[18] M. Heinkenschloss and D. Ridzal. An Inexact Trust-Region SQP Method with Applications to PDE-Constrained Optimization. In K. Kunisch, G. Of, and O. Steinbach, editors, *Numerical Mathematics and Advanced Applications: Proceedings of ENUMATH 2007, the 7th European Conference on Numerical Mathematics and Advanced Applications, Graz, Austria*, pages 613–620. Springer, 2008.

[19] M. Heinkenschloss and L. N. Vicente. Analysis of Inexact Trust-Region SQP Algorithms. *SIAM Journal on Optimization*, 12(2):283–302, 2002.

[20] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003.

[21] Alexander Kaplan and Rainer Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

[22] Soumava Kumar Roy, Zakaria Mhammedi, and Mehrtash Harandi. Geometry aware constrained optimization techniques for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4469, 2018.

[23] Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Imposing hard constraints on deep networks: Promises and limitations. *arXiv preprint arXiv:1706.02025*, 2017.

[24] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented Lagrangians. *arXiv preprint arXiv:2102.05320*, 2021.

[25] Yatin Nandwani, Abhishek Pathak, and Parag Singla. A primal-dual formulation for deep learning with constraints. In *Advances in Neural Information Processing Systems*, pages 12157–12168, 2019.

[26] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, NY, USA, second edition, 2006.

[27] Christopher C Paige and Michael A Saunders. Solution of sparse indefinite systems of linear equations. *SIAM journal on numerical analysis*, 12(4):617–629, 1975.

[28] M. J. D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical Analysis*, Lecture Notes in Mathematics, pages 144–157. Springer, Berlin, 1978.

[29] Michael JD Powell. Algorithms for nonlinear constraints that use Lagrangian functions. *Mathematical programming*, 14(1):224–248, 1978.

[30] Sathya N Ravi, Tuan Dinh, Vishnu Suresh Lokhande, and Vikas Singh. Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4772–4779, 2019.

[31] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[32] Lars Ruthotto and Eldad Haber. Deep Neural Networks Motivated by Partial Differential Equations. *Journal of Mathematical Imaging and Vision*, 62:352–364, 2020.

[33] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[34] Sheroze Sheriffdeen, Jean C. Ragusa, Jim E. Morel, Marvin L. Adams, and Tan Bui-Thanh. Accelerating PDE-constrained Inverse Solutions with Deep Learning and Reduced Order Models. arXiv 1912.08864, 2019.

[35] Naum Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*, volume 3. Springer Science & Business Media, 2012.

[36] Tyler Summers, Joseph Warrington, Manfred Morari, and John Lygeros. Stochastic optimal power flow based on conditional value at risk and distributional robustness. *International Journal of Electrical Power & Energy Systems*, 72:116–125, 2015.

[37] Vikrant Singh Tomar and Richard C Rose. Manifold regularized deep neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[38] Maria Vrakopoulou, Johanna L. Mathieu, and Göran Andersson. Stochastic optimal power flow with uncertain reserves from demand response. In *2014 47th Hawaii International Conference on System Sciences*, pages 2353–2362. IEEE, 2014.

[39] R. B. Wilson. *A Simplicial Algorithm for Concave Programming*. Ph.D. Thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA, USA, 1963.

[40] Allen J. Wood, Bruce F Wollenberg, and Gerald B Sheblé. *Power generation, operation, and control*. John Wiley & Sons, 2013.

[41] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394:56–81, 2019.