

INDUSTRIAL AND SYSTEMS ENGINEERING



Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints

FRANK E. CURTIS AND DANIEL P. ROBINSON

Department of Industrial and Systems Engineering, Lehigh University

BAOYU ZHOU

Booth School of Business, The University of Chicago

COR@L Technical Report 23T-017



Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints

FRANK E. CURTIS^{*1}, DANIEL P. ROBINSON^{†1}, AND BAOYU ZHOU^{‡2}

¹Department of Industrial and Systems Engineering, Lehigh University

²Booth School of Business, The University of Chicago

February 28, 2023

Abstract

A sequential quadratic optimization algorithm for minimizing an objective function defined by an expectation subject to nonlinear inequality and equality constraints is proposed, analyzed, and tested. The context of interest is when it is tractable to evaluate constraint function and derivative values in each iteration, but it is intractable to evaluate the objective function or its derivatives in any iteration, and instead an algorithm can only make use of stochastic objective gradient estimates. Under loose assumptions, including that the gradient estimates are unbiased, the algorithm is proved to possess convergence guarantees in expectation. The results of numerical experiments are presented to demonstrate that the proposed algorithm can outperform an alternative approach that relies on the ability to compute more accurate gradient estimates.

1 Introduction

We propose a sequential quadratic optimization (commonly known as SQP) algorithm for minimizing an objective function defined by an expectation subject to nonlinear inequality and equality constraints. Such optimization problems arise in a plethora of application areas, including, but not limited to, machine learning [28], network optimization [7], resource allocation [26], portfolio optimization [37], risk-averse partial-differential-equation-constrained optimization [27], maximum-likelihood estimation [25], and multi-stage optimization [40].

The design and analysis of deterministic algorithms for solving continuous optimization problems involving inequality and equality constraints has been a well-studied topic for decades. Numerous types of such algorithms, such as penalty methods, interior-point methods, and SQP methods, have been designed to solve such problems. Penalty methods are based on the idea of using unconstrained optimization algorithms to minimize a weighted sum—determined by a penalty parameter—of the objective and a measure of constraint violation; e.g., see [11, 19, 45] for algorithms that make use of nondifferentiable (exact) penalty functions and see [15, 14, 21, 46] for algorithms that make use of differentiable (exact) penalty functions. While they are able to offer convergence guarantees from remote starting points, the numerical performance of penalty methods often suffers from ill-conditioning of the penalty functions and/or sensitivity of the algorithm’s performance on the particular scheme employed for updating the penalty parameter [34]. Interior-point methods [16] are designed to use barrier functions to guide the algorithm along a central path through the interior of the feasible region (or, at least, the interior of a set defined by bounds on a subset of the variables)

^{*}E-mail: frank.e.curtis@lehigh.edu

[†]E-mail: daniel.p.robinson@lehigh.edu

[‡]E-mail: baoyu.zhou@chicagobooth.edu

to a solution [9, 10, 29, 30, 43, 44]. Such algorithms have been shown to be very effective in practice, which is why many state-of-the-art software packages for continuous nonlinear optimization are built on interior-point methods; see, e.g., [10, 42]. Overall, both penalty and interior-point methods involve the use of additional objective terms to handle the presence of inequality constraints.

Alternatively, in this paper, we present, analyze, and demonstrate the numerical performance of an SQP method for solving continuous nonlinear optimization problems. The SQP paradigm is based on the idea of, at each iterate, solving a subproblem (or subproblems) defined based on a local linearization of the constraint function and a local quadratic approximation of the objective. Unlike in the deterministic setting, for which numerous SQP algorithms have been proposed (see, e.g., [18, 20, 24, 34]), there have been few stochastic algorithms proposed for the setting of solving optimization problems with nonlinear constraints. That said, in the past few years, a couple of classes of stochastic SQP methods have been designed for optimization subject to nonlinear *equality* constraints. For example, the article [1] proposes an SQP algorithm that uses stochastic objective gradient estimates for solving such problems that employs an adaptive step size policy based on Lipschitz constants (or estimates of them). For an alternative setting in which one is willing to compute objective value estimates as well, and to refine objective function and gradient estimates within a given iteration until probabilistic conditions of accuracy are satisfied, the article [32] proposes a line-search stochastic SQP method. There have subsequently been multiple extensions of the methods in [1] and [32], as well as work on different, but related algorithmic strategies—still for the setting of only nonlinear equality constraints. There has been work on relaxing constraint qualifications [3], allowing matrix-free and inexact solves of the arising linear systems [12], using a trust-region methodology [17], incorporating noisy (potentially biased) function and gradient estimates [5, 35], employing variance-reduction strategies [2, 4], considering sketch-and-project techniques [33], and analyzing the worst-case complexity (see [13]) of the method proposed in [1].

Unlike the setting of equality constraints only, to our knowledge there has been very little work on the design and analysis of stochastic algorithms for optimization subject to nonlinear (nonconvex) inequality and equality constraints. Three exceptions are: the active-set line-search SQP algorithms proposed in [31] and (very recently) in [38] and the momentum-based augmented Lagrangian method (a penalty method) proposed in [41]. We expect that our proposed SQP algorithm will perform well in comparison to a stochastic-gradient-based penalty method. We demonstrate with numerical experiments that our approach can outperform the algorithm proposed in [31]. We remark in passing that interior-point methods often outperform SQP methods in the deterministic setting, but as far as we are aware there exists no interior-point method designed for the stochastic setting that we consider.

1.1 Contributions

In this paper, we build on the algorithmic strategy and analysis in [1] to propose and analyze an adaptive stochastic SQP algorithm for solving nonlinear optimization problems subject to (deterministic) inequality and equality constraints. This work involves significant advancements beyond that in [1] that are necessary since, unlike in the setting of only having equality constraints, the presence of inequality constraints automatically guarantees that, at a given iterate, the search direction computed in a stochastic SQP method will be a biased estimate of the “true” search direction, i.e., the one that would be computed if the actual gradient of the objective function were available. This necessitates a distinct change in the design of the algorithm as well as distinct alterations to the convergence analysis, since the analysis in [1] relies heavily on the search directions being (conditionally) unbiased estimators of their “true” counterparts. The algorithm from the literature that can be seen as the nearest alternative approach is the algorithm in [31]. However, there are substantial differences between the algorithm and analysis in [31] and those presented in this paper. Like in [32] for the equality-only case, the algorithm in [31] is designed for the setting in which one is willing to refine function and gradient estimates within an iteration until probabilistic conditions of accuracy are satisfied, and in this manner the analysis of that algorithm offers guarantees that are relatively closer to those offered for a deterministic algorithm. By contrast, the algorithm in this paper, like the algorithm in [1], is designed to allow the stochastic gradient estimates to be potentially much less accurate, and in such a context we are satisfied with offering convergence guarantees in expectation. We compare the numerical

performance of our proposed algorithm with that in [31] to demonstrate that there are settings in which our proposed approach has advantages in practice.

1.2 Notation

We use \mathbb{R} to denote the set of real numbers, $\overline{\mathbb{R}}$ to denote the set of extended-real numbers (i.e., $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$), and $\mathbb{R}_{\geq a}$ (resp., $\mathbb{R}_{>a}$) to denote the set of real numbers greater than or equal to (resp., greater than) $a \in \mathbb{R}$. We append a superscript to such a set to denote the space of vectors or matrices whose elements are restricted to the indicated set; e.g., we use \mathbb{R}^n to denote the set of n -dimensional real vectors and $\mathbb{R}^{m \times n}$ to denote the set of m -by- n -dimensional real matrices. We use $\mathbb{N} := \{1, 2, \dots\}$ to denote the set of positive integers and, given $n \in \mathbb{N}$, we use $[n] := \{1, \dots, n\}$ to denote the set of positive integers less than or equal to n . Given $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, we write $a \perp b$ to mean, with a_i and b_i denoting the i th elements of a and b , respectively, that $a_i = 0$ and/or $b_i = 0$ for all $i \in [n]$. Given real symmetric matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$, we write $A \succeq B$ (resp., $A \succ B$) to indicate that $A - B$ is positive semidefinite (resp., positive definite). Given $H \in \mathbb{R}^{n \times n}$ with $H \succ 0$ and $a \in \mathbb{R}^n$, we denote the norm $\|a\|_H := \sqrt{a^T H a}$.

Our problem of interest is defined with respect to a variable $x \in \mathbb{R}^n$ and the algorithm that we propose and analyze is iterative, meaning that, in any run, it generates an iterate sequence that we denote as $\{x_k\}$ with $x_k \in \mathbb{R}^n$ for all generated $k \in \mathbb{N}$, i.e., $\{x_k\} \subset \mathbb{R}^n$. (We use such notation throughout the paper when the elements of sequence are contained within a given set. We say “for all *generated* $k \in \mathbb{N}$ ” since our proposed algorithm might terminate finitely. Whether a subscript is being used to indicate the element of a vector or the index number of a sequence is always made clear by the context. The i th element of an iterate x_k is denoted $[x_k]_i$.) We use subscripts similarly to denote other quantities corresponding to each iteration of the algorithm; e.g., we introduce a merit parameter denoted as $\tau \in \mathbb{R}_{>0}$ whose value in iteration $k \in \mathbb{N}$ is denoted as $\tau_k \in \mathbb{R}_{>0}$, and corresponding to a constraint function c (see problem (1) below) we denote its value at x_k as $c_k := c(x_k)$.

The iteration-dependent quantities mentioned in the previous paragraph—and additional ones introduced in the description of our algorithm—represent realizations of the random variables in a stochastic process generated by the algorithm. Specifically, the behavior of our algorithm is dictated by prescribed initial conditions and a sequence of stochastic objective gradient estimators that we denote by $\{G_k\}$. After proving preliminary results that hold for every run of the algorithm, we present our ultimate convergence theory for our algorithm in terms of a filtration defined in terms of σ -algebras dependent on the initial conditions of the algorithm and $\{G_k\}$.

1.3 Organization

A statement of our problem of interest and preliminary assumptions about its objective and constraint functions, as well as about user-defined quantities in our proposed algorithm, are stated in Section 2. A description of our proposed algorithm is provided in Section 3. Convergence-in-expectation of the algorithm is proved under reasonable assumptions in Section 4. The results of numerical experiments are presented in Section 5 and concluding remarks are given in Section 6.

2 Setting

We formulate our problem of interest as

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

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are continuously differentiable, ω is a random variable with associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$, and $\mathbb{E}_\omega[\cdot]$ denotes expectation taken with respect to the distribution of ω . Our algorithm and analysis extend easily to the setting in which the nonnegativity constraint in (1) is generalized to $l \leq x \leq u$ for some $(l, u) \in \overline{\mathbb{R}}^n \times \overline{\mathbb{R}}^n$ with $l_i \leq u_i$ for all $i \in [n]$; we

merely consider nonnegativity in (1) for the sake of notational simplicity. It is also worth mentioning that any smooth constrained optimization problem can be reformulated as (1) (or at least as such a problem with generalized bound constraints); e.g., inequality constraints $c_{\mathcal{I}}(x) \leq 0$, where $c_{\mathcal{I}} : \mathbb{R}^n \rightarrow \mathbb{R}^{m_{\mathcal{I}}}$ is continuously differentiable, can be reformulated to fit into the form of (1) through the incorporation of slack variables, say $s \in \mathbb{R}^{m_{\mathcal{I}}}$, to have the constraints $c_{\mathcal{I}}(x) + s_{\mathcal{I}} = 0$ and $s_{\mathcal{I}} \geq 0$.

We make the following assumption throughout the remainder of the paper pertaining to the functions in problem (1) and our proposed algorithm. As seen in the following section, our algorithm seeks feasibility and stationarity with respect to (1) by generating an iterate sequence that stays feasible with respect to the bound constraints, meaning that, in any run of the algorithm, $x_k \in \mathbb{R}_{\geq 0}^n$ for all generated $k \in \mathbb{N}$.

Assumption 1. *Let $\mathcal{X} \subset \mathbb{R}^n$ be an open convex set that almost-surely contains the iterate sequence $\{x_k\} \subset \mathbb{R}_{\geq 0}^n$ generated in any realization of a run of the algorithm. The objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and bounded below over \mathcal{X} and the objective gradient function $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is Lipschitz continuous and bounded in norm over \mathcal{X} . Similarly, for all $i \in [m]$, the constraint function $c_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and bounded over \mathcal{X} and the constraint gradient function $\nabla c_i : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is Lipschitz continuous and bounded in norm over \mathcal{X} . Finally, the constraint Jacobian $\nabla c^T : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ has full row rank over \mathcal{X} .*

Under Assumption 1, there exists $f_{\inf} \in \mathbb{R}$ and a tuple of positive constants $(\kappa_{\nabla f}, \kappa_c, \kappa_{\nabla c}, L, \Gamma) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that for all $x \in \mathcal{X}$ one has

$$f(x) \geq f_{\inf}, \quad \|\nabla f(x)\|_2 \leq \kappa_{\nabla f}, \quad \|c(x)\|_2 \leq \kappa_c, \quad \text{and} \quad \|\nabla c(x)\|_2 \leq \kappa_{\nabla c}, \quad (2)$$

and for all $(x, \bar{x}) \in \mathcal{X} \times \mathcal{X}$ one has

$$\|\nabla f(x) - \nabla f(\bar{x})\|_2 \leq L\|x - \bar{x}\|_2 \quad \text{and} \quad \|\nabla c(x)^T - \nabla c(\bar{x})^T\|_2 \leq \Gamma\|x - \bar{x}\|_2. \quad (3)$$

In addition, due to the continuous differentiability of the objective and constraint functions and the full row rank of the constraint Jacobians, it follows that at any (local) minimizer of (1), call it $x \in \mathbb{R}^n$, there exists $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$ such that the following Karush-Kuhn-Tucker (KKT) conditions are satisfied:

$$\nabla f(x) + \nabla c(x)y - z = 0, \quad c(x) = 0, \quad 0 \leq x \perp z \geq 0. \quad (4)$$

We refer to any $x \in \mathbb{R}^n$ such that there exists $(y, z) \in \mathbb{R}^m \times \mathbb{R}^n$ satisfying (4) as a first-order stationary point (or KKT point) with respect to (1).

Since our algorithm generates iterates that are feasible with respect to the bound constraints, but not necessarily with respect to the equality constraints, we need to account for the possible existence of points that are infeasible for (1), but are stationary with respect to the minimization of a constraint violation measure over $\mathbb{R}_{\geq 0}^n$. We refer to a point that is infeasible for (1) as an infeasible stationary point if it is stationary with respect to the minimization of $\frac{1}{2}\|c(x)\|_2^2$ subject to $x \in \mathbb{R}_{\geq 0}^n$, meaning

$$0 \leq x \perp \nabla c(x)c(x) \geq 0. \quad (5)$$

Each iteration of our algorithm requires a stochastic estimate of the gradient of the objective at the current iterate. In a given run at iteration $k \in \mathbb{N}$, the realization of the iterate and gradient estimate is (x_k, g_k) , which later in our analysis we denote as a realization of the pair of random variables (X_k, G_k) . (See Section 4.3 for a complete description of a stochastic process that we analyze.) With respect to the gradient estimators, we make Assumption 2 below. For the prescribed (i.e., not random) sequence $\{\rho_k\} \subset \mathbb{R}_{>0}$ referenced in the assumption, we state precise conditions that it must satisfy in Section 4.3. In the assumption and throughout the remainder of the paper, we use $\mathbb{E}_k[\cdot]$ to denote expectation taken with respect to the distribution of ω conditioned on a trace σ -algebra of an event \mathcal{E} , denoted by \mathcal{F}_k ; see Section 4.3.

Assumption 2. *For a prescribed $\{\rho_k\} \subset \mathbb{R}_{>0}$, one finds for all $k \in \mathbb{N}$ that*

$$\mathbb{E}_k[G_k] = \nabla f(X_k) \quad \text{and} \quad \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2] \leq \rho_k. \quad (6)$$

One might relax the latter condition in (6) and obtain guarantees that are similar to those that we prove; see, e.g., [36]. We employ (6) for simplicity, since it is sufficient for demonstrating the guarantees that our algorithmic approach can offer.

Each iteration of our algorithm also makes use of a symmetric and positive-definite (SPD) matrix, denoted as $H_k \in \mathbb{R}^{n \times n}$ for iteration $k \in \mathbb{N}$, to define a quadratic term in the subproblem that is solved for computing the search direction. For simplicity, we assume that the sequence $\{H_k\}$ is prescribed, e.g., one may consider $H_k = I$ for all $k \in \mathbb{N}$. More generally, one could consider a more sophisticated scheme such as setting, for all $k \in \mathbb{N}$, the matrix H_k as a stochastic estimate of the Hessian of the objective function and/or a Lagrangian function as long as it is sufficiently positive definite and bounded and the choice is made to be conditionally uncorrelated with the stochastic gradient estimate. However, since considering such a loose requirement would only obfuscate our analysis without adding significant value, we assume for simplicity that $\{H_k\}$ is prescribed and merely satisfies the following.

Assumption 3. *There exists $(\kappa_H, \zeta) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ with $\kappa_H \geq \zeta$ such that, for all $k \in \mathbb{N}$, the SPD matrix $H_k \in \mathbb{R}^{n \times n}$ has $\kappa_H I \succeq H_k \succeq \zeta I$.*

Observe from Assumption 3 that we are not assuming that accurate second-order information is being used by the algorithm. Hence, our convergence guarantees are of the type that may be expected for a first-order-type algorithm, although in situations when it is computationally tractable, one might find better performance if H_k incorporates some (approximate) second-order derivative information.

3 Algorithm

In this section, we present our proposed algorithm. We state the algorithm in terms of a particular realization of it (e.g., denoting the iterate for each $k \in \mathbb{N}$ as x_k), although our subsequent analysis of it (starting in Section 4.3) will be written in terms of the stochastic process that the algorithm defines.

Each iteration of our algorithm proceeds as follows. First, given the current iterate $x_k \in \mathbb{R}_{\geq 0}^n$, the algorithm computes a direction whose purpose is to determine the progress that can be made in terms of reducing a measure of violation of a linearization of the equality constraints subject to the bound constraints. This is done in a manner that regularizes the component of the direction that lies in the null space of the constraint Jacobian. Specifically, the iteration commences by computing a direction $v_k := u_k + \nabla c(x_k)w_k \in \mathbb{R}^n$, where $u_k \in \text{Null}(\nabla c(x_k)^T)$ and $\nabla c(x_k)w_k \in \text{Range}(\nabla c(x_k))$, by solving the quadratic optimization subproblem

$$\begin{aligned} \min_{u \in \mathbb{R}^n, w \in \mathbb{R}^m} \quad & \frac{1}{2} \|c_k + \nabla c(x_k)^T \nabla c(x_k)w\|_2^2 + \frac{1}{2} \mu_k \|u\|_2^2 \\ \text{s.t.} \quad & \nabla c(x_k)^T u = 0 \quad \text{and} \quad x_k + u + \nabla c(x_k)w \geq 0, \end{aligned} \tag{7}$$

where $\mu_k \in \mathbb{R}_{>0}$ is a user-prescribed parameter. Observe that since $x_k \in \mathbb{R}_{\geq 0}^n$, this subproblem is always feasible, and by construction it is convex. Generally, the solution of (7) might not be unique, but in our setting it is unique since $\nabla c(x_k)^T$ has full row rank. In our analysis, we show that the solution of subproblem (7) is given by $(u_k, w_k) = (0, 0)$ if and only if the current iterate x_k is stationary with respect to the minimization of $\frac{1}{2} \|c(x)\|_2^2$ over $x \in \mathbb{R}_{\geq 0}^n$. This means, e.g., that if $c_k \neq 0$, but the solution of (7) is $(u_k, w_k) = (0, 0)$ —which, by the Fundamental Theorem of Linear Algebra, occurs if and only if $v_k = u_k + \nabla c(x_k)w_k = 0$ —then it is reasonable to terminate since x_k is an infeasible stationary point (see (5)), as in our algorithm.

After computing $v_k \in \mathbb{R}^n$ by solving (7) and generating a stochastic objective gradient estimate $g_k \in \mathbb{R}^n$ (see Assumption 2), the algorithm next computes a search direction $d_k \in \mathbb{R}^n$ by solving the quadratic optimization subproblem

$$\min_{d \in \mathbb{R}^n} \quad g_k^T d + \frac{1}{2} d^T H_k d \quad \text{s.t.} \quad \nabla c(x_k)^T d = \nabla c(x_k)^T v_k \quad \text{and} \quad x_k + d \geq 0. \tag{8}$$

By construction, this subproblem is feasible, and under Assumption 3 it is convex. The search direction d_k is designed to achieve the same progress toward linearized feasibility within the nonnegative orthant that

is achieved by v_k , then within the null space of $\nabla c(x_k)^T$ and the nonnegative orthant aims to minimize a (stochastically estimated) local quadratic approximation of the objective function at x_k .

The remainder of the k th iteration proceeds in a similar manner as in [1, 12]. In particular, with the ℓ_2 -norm merit function in mind, namely, $\phi : \mathbb{R}^n \times \mathbb{R}_{>0} \rightarrow \mathbb{R}$ defined by $\phi(x, \tau) = \tau f(x) + \|c(x)\|_2$, the algorithm next sets a value for the merit parameter $\tau_k \in \mathbb{R}_{>0}$. This is done by considering a local model of this merit function, namely, $l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $l(x, \tau, g, d) = \tau(f(x) + g^T d) + \|c(x) + \nabla c(x)^T d\|_2$, and in particular the reduction in this model defined for all $k \in \mathbb{N}$ by

$$\begin{aligned} \Delta l(x_k, \tau_k, g_k, d_k) &:= l(x_k, \tau_k, g_k, 0) - l(x_k, \tau_k, g_k, d_k) \\ &= -\tau_k g_k^T d_k + \|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2, \end{aligned} \quad (9)$$

and setting τ_k such that this reduction is sufficiently large. Specifically, if $d_k \neq 0$, then with user-prescribed $(\epsilon_\tau, \sigma) \in (0, 1) \times (0, 1)$, the algorithm first sets

$$\tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^T d_k + \frac{1}{2} d_k^T H_k d_k \leq 0 \\ \frac{(1-\sigma)(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2)}{g_k^T d_k + \frac{1}{2} d_k^T H_k d_k} & \text{otherwise,} \end{cases} \quad (10)$$

then sets the merit parameter value as

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \leq \tau_k^{\text{trial}} \\ \min\{(1 - \epsilon_\tau)\tau_{k-1}, \tau_k^{\text{trial}}\} & \text{otherwise.} \end{cases} \quad (11)$$

(The value $\tau_0 \in \mathbb{R}_{>0}$ is also prescribed by the user.) On the other hand, if $d_k = 0$, then the algorithm simply sets $\tau_k^{\text{trial}} \leftarrow \infty$ and $\tau_k \leftarrow \tau_{k-1}$. We show in our analysis (see Lemma 8) that this procedure for setting τ_k ensures that $\Delta l(x_k, \tau_k, g_k, d_k)$ is sufficiently large relative to the squared norm of the search direction and the improvement offered toward linearized feasibility. For use in the step size procedure, the algorithm next sets a value $\xi_k \in \mathbb{R}_{>0}$ (referred to as the ratio parameter) that acts as an estimate for a lower bound of the ratio between the model reduction and a multiple of the squared norm of the search direction. Specifically, if $d_k \neq 0$, it sets

$$\xi_k^{\text{trial}} \leftarrow \frac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|_2^2}, \quad \text{then } \xi_k \leftarrow \begin{cases} \xi_{k-1} & \text{if } \xi_{k-1} \leq \xi_k^{\text{trial}} \\ \min\{(1 - \epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \quad (12)$$

where $(\xi_0, \epsilon_\xi) \in \mathbb{R}_{>0} \times (0, 1)$ are user-prescribed parameters; see [1, 12] for further motivation. On the other hand, if $d_k = 0$, then it sets $\xi_k^{\text{trial}} \leftarrow \infty$ and $\xi_k \leftarrow \xi_{k-1}$.

The step size selection procedure, which for all $k \in \mathbb{N}$ chooses the step size $\alpha_k \in \mathbb{R}_{>0}$, can now be summarized as follows. First, suppose that $d_k \neq 0$. With user-prescribed $\eta \in (0, 1)$, $\theta \in \mathbb{R}_{>0}$, and $\{\beta_k\}$ with $\beta_k \in (0, 1]$ for all $k \in \mathbb{N}$ such that

$$\alpha_k^{\min} \leftarrow \frac{2(1-\eta)\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} \in (0, 1] \quad \text{for all } k \in \mathbb{N}, \quad (13)$$

and with the strongly convex function $\varphi_k : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ defined by

$$\begin{aligned} \varphi_k(\alpha) &= (\eta - 1)\alpha\beta_k\Delta l(x_k, \tau_k, g_k, d_k) + \|c_k + \alpha\nabla c(x_k)^T d_k\|_2 - \|c_k\|_2 \\ &\quad + \alpha(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2\|d_k\|_2^2, \end{aligned} \quad (14)$$

the algorithm sets the values

$$\begin{aligned} \alpha_k^\varphi &\leftarrow \max\{\alpha \in \mathbb{R}_{\geq 0} : \varphi_k(\alpha) \leq 0\} \\ \text{and } \alpha_k^{\max} &\leftarrow \min\{1, \alpha_k^\varphi, \alpha_k^{\min} + \theta\beta_k\}. \end{aligned} \quad (15)$$

The algorithm then chooses the step size α_k as any value in $[\alpha_k^{\min}, \alpha_k^{\max}]$. Second, if $d_k = 0$, then the algorithm simply sets all step size values to 1.

A complete statement of our algorithm is given as Algorithm 1.

Algorithm 1 Stochastic SQP

Require: $x_1 \in \mathbb{R}_{\geq 0}^n$; $\{\mu_k\} \subset \mathbb{R}_{>0}$; $\{H_k\} \subset \mathbb{R}^{n \times n}$ satisfying Assumption 3; $\tau_0 \in \mathbb{R}_{>0}$; $\xi_0 \in \mathbb{R}_{>0}$; $\{\sigma, \eta, \epsilon_\tau, \epsilon_\xi\} \subset (0, 1)$; $\{\beta_k\} \subset (0, 1]$ satisfying (13); $\theta \in \mathbb{R}_{>0}$; $\{\rho_k\} \subset \mathbb{R}_{>0}$; Lipschitz constants $L \in \mathbb{R}_{>0}$ and $\Gamma \in \mathbb{R}_{>0}$ (see (3))

```
1: for  $k \in \mathbb{N}$  do
2:   compute  $v_k \in \mathbb{R}^n$  by solving (7)
3:   if  $c_k \neq 0$  and  $v_k = 0$  then terminate and return  $x_k$  (infeasible stationary)
4:   compute  $g_k \in \mathbb{R}^n$  (recall Assumption 2)
5:   compute  $d_k \in \mathbb{R}^n$  by solving (8)
6:   if  $d_k = 0$  then
7:     set  $\tau_k^{\text{trial}} \leftarrow \infty$ ,  $\tau_k \leftarrow \tau_{k-1}$ ,  $\xi_k^{\text{trial}} \leftarrow \infty$ , and  $\xi_k \leftarrow \xi_{k-1}$ 
8:     set  $\alpha_k^{\min} \leftarrow 1$ ,  $\alpha_k^\varphi \leftarrow 1$ ,  $\alpha_k^{\max} \leftarrow 1$ , and  $\alpha_k \leftarrow 1$ 
9:   else
10:    set  $\tau_k^{\text{trial}}$  by (10) and  $\tau_k$  by (11)
11:    set  $\xi_k^{\text{trial}}$  and  $\xi_k$  by (12)
12:    set  $\alpha_k^{\min}$  by (13) and both  $\alpha_k^\varphi$  and  $\alpha_k^{\max}$  by (15)
13:    choose  $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$ 
14:   end if
15:   Set  $x_{k+1} \leftarrow x_k + \alpha_k d_k$ 
16: end for
```

4 Analysis

In this section, we provide theoretical results for Algorithm 1. We begin by introducing common assumptions under which one can establish stationarity measures for problem (1) that are defined by solutions of (7) and/or (8). These stationarity measures allow us to connect our convergence guarantees for Algorithm 1 with stationarity conditions for (1). Then, under Assumptions 1 and 3, we prove generally applicable results pertaining to the behavior of algorithmic quantities in any run of the algorithm. These results reveal that the algorithm is well defined in the sense that any run will either terminate and return an infeasible stationary point or generate an infinite sequence of iterates. We then consider convergence properties of the algorithm in the event that the (monotonically nonincreasing) merit parameter sequence eventually produces values that are sufficiently small, yet bounded away from zero, which, as shown in our analysis, means that the sequence ultimately becomes constant at a sufficiently small value. This analysis, which includes our main convergence results for the algorithm, is provided under Assumption 4 stated on page 16. We follow this analysis with a section on theoretical results related to the occurrence of the event in Assumption 4. As in [1] for the equality-constraints-only setting, this discussion illuminates the fact that while the event in Assumption 4 is not always guaranteed to occur due to the looseness of our assumptions about properties of the stochastic gradient estimates, the event represents likely behavior in practice, which shows that our convergence results about the algorithm are meaningful for real-world situations. We conclude this section with a discussion of the behavior of the algorithm in the deterministic setting, i.e., when the true gradient of the objective is employed in all iterations. This discussion is meant to provide confidence to a user that our algorithm is based on one that has state-of-the-art convergence properties under common assumptions in the deterministic setting.

4.1 Subproblems and Stationarity Measures

We begin by showing that subproblem (7) yields a zero solution if and only if the point defining the subproblem is feasible for problem (1) or an infeasible stationary point.

Lemma 1. *Suppose that Assumption 1 holds, $x \in \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$, and, given $\mu \in \mathbb{R}_{>0}$, consider the quadratic*

optimization problem (recall (7))

$$\begin{aligned} \min_{u \in \mathbb{R}^n, w \in \mathbb{R}^m} \quad & \frac{1}{2} \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2^2 + \frac{1}{2} \mu \|u\|_2^2 \\ \text{s.t.} \quad & \nabla c(x)^T u = 0 \quad \text{and} \quad x + u + \nabla c(x) w \geq 0. \end{aligned} \quad (16)$$

Then, the unique optimal solution of problem (16) is $(u, w) = (0, 0)$ if and only if x is feasible for problem (1) or an infeasible stationary point (i.e., it satisfies (5)), whereas $(u, w) \neq (0, 0)$ if and only if $\|c(x)\|_2 > \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2$.

Proof. Suppose the conditions of the lemma hold and let (u, w) be the unique optimal solution of (16). Since $x \in \mathbb{R}_{\geq 0}^n$, it follows that $(0, 0)$ is feasible for (16). In addition, necessary and sufficient optimality conditions for (16) are that, corresponding to $(u, w) \in \mathbb{R}^n \times \mathbb{R}^m$, there exists $(\gamma, \delta) \in \mathbb{R}^m \times \mathbb{R}^n$ with

$$\begin{aligned} \nabla c(x)^T \nabla c(x) c(x) + \nabla c(x)^T \nabla c(x) \nabla c(x)^T \nabla c(x) w - \nabla c(x)^T \delta &= 0, \\ \mu u + \nabla c(x) \gamma - \delta &= 0, \quad \nabla c(x)^T u = 0, \quad \text{and} \quad 0 \leq \delta \perp x + u + \nabla c(x) w \geq 0. \end{aligned} \quad (17)$$

If $(u, w) = (0, 0)$, then it follows from (17) that

$$\nabla c(x)^T \nabla c(x) c(x) - \nabla c(x)^T \delta = 0, \quad \nabla c(x) \gamma - \delta = 0, \quad \text{and} \quad 0 \leq \delta \perp x \geq 0. \quad (18)$$

Since $\nabla c(x)^T$ has full row rank, (18) implies $\gamma = (\nabla c(x)^T \nabla c(x))^{-1} \nabla c(x)^T \delta = c(x)$, $\delta = \nabla c(x) c(x)$, and $0 \leq \nabla c(x) c(x) \perp x \geq 0$, which from (5) means that x is either feasible or an infeasible stationary point, as desired. On the other hand, if x is either feasible or an infeasible stationary point, meaning $0 \leq \nabla c(x) c(x) \perp x \geq 0$, then $u = 0$, $w = 0$, $\gamma = c(x)$, and $\delta = \nabla c(x) c(x)$ satisfy (17), and this solution (i.e., $(u, w) = (0, 0)$) is unique since the objective of (16) is strongly convex.

Now let us show that the unique optimal solution of (16) is $(u, w) \neq (0, 0)$ if and only if $\|c(x)\|_2 > \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2$. If $\|c(x)\|_2 > \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2$, then $w \neq 0$ follows trivially, giving the desired conclusion. To prove the reverse implication, let us consider two cases. If $u \neq 0$, then, since $(0, 0)$ is feasible for (16),

$$\frac{1}{2} \|c(x)\|_2^2 \geq \frac{1}{2} \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2^2 + \frac{1}{2} \mu \|u\|_2^2 > \frac{1}{2} \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2^2,$$

as desired. Second, if $u = 0$ and $w \neq 0$, then w is the minimizer of the strongly convex objective $\frac{1}{2} \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2^2$ subject to $x + \nabla c(x) w \geq 0$. Since 0 is feasible for this problem, $w \neq 0$ means that $\frac{1}{2} \|c(x)\|_2^2 > \frac{1}{2} \|c(x) + \nabla c(x)^T \nabla c(x) w\|_2^2$, as desired. \square

We now show that, under common assumptions and given $x_k \in \mathbb{R}_{\geq 0}^n$, the quantity $\|v_k\|_2^2$, where $v_k \in \mathbb{R}^n$ solves subproblem (7), represents a stationarity measure with respect to the problem to minimize $\frac{1}{2} \|c(x)\|_2^2$ subject to $x \in \mathbb{R}_{\geq 0}^n$. (The assumption in the lemma that $\mu_k = \mu \in \mathbb{R}_{> 0}$ for all $k \in \mathbb{N}$ could be relaxed; see Remark 1 at the end of this subsection. We consider this case for the sake of brevity.)

Lemma 2. Suppose that Assumption 1 holds and there exists infinite $\mathcal{S} \subseteq \mathbb{N}$ such that for some sequence $\{x_k\} \subset \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ one finds $\{x_k\}_{k \in \mathcal{S}} \rightarrow x_*$ for some $x_* \in \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ where, with $\mathcal{A}(x) := \{i \in [n] : x_i = 0\}$, $I_{\mathcal{A}(x)}$ denoting the matrix composed of rows of $I \in \mathbb{R}^{n \times n}$ corresponding to indices in $\mathcal{A}(x)$, and $\nabla c(x)_{\mathcal{A}(x)}$ denoting the matrix composed of rows of $\nabla c(x)$ corresponding to indices in $\mathcal{A}(x)$, one finds that

(i) $[\nabla c(x_*) c(x_*)]_i > 0$ for all $i \in \mathcal{A}(x_*)$ and

(ii) the following matrix has full row rank: $\begin{bmatrix} 0 & \nabla c(x_*)^T \\ \nabla c(x_*)_{\mathcal{A}(x_*)} & I_{\mathcal{A}(x_*)} \end{bmatrix}$.

Then, with $\mu_k = \mu \in \mathbb{R}_{> 0}$ for all $k \in \mathbb{N}$, and with (u_k, w_k) solving subproblem (7) and $v_k := u_k + \nabla c(x) w_k$ for all $k \in \mathbb{N}$, it follows that x_* satisfies the stationarity conditions (5) if and only if $\{v_k\}_{k \in \mathcal{S}} \rightarrow 0$.

Proof. Let $\mathcal{A}_* := \mathcal{A}(x_*)$ and $j(x) := \nabla c(x)^T$ and consider the linear system

$$\begin{bmatrix} j(x)j(x)^T j(x)j(x)^T & 0 & 0 & -j(x)_{\mathcal{A}_*} \\ 0 & \mu I & j(x)^T & -I_{\mathcal{A}_*}^T \\ 0 & j(x) & 0 & 0 \\ j(x)_{\mathcal{A}_*}^T & I_{\mathcal{A}_*} & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ u \\ \gamma \\ \delta_{\mathcal{A}_*} \end{bmatrix} = \begin{bmatrix} -j(x)j(x)^T c(x) \\ 0 \\ 0 \\ -x_{\mathcal{A}_*} \end{bmatrix}.$$

Since, under the conditions of the lemma, the matrix in this linear system is nonsingular when $x = x_*$ (e.g., this follows from [39, Theorem 1.5.1] and (ii)), it follows that there exists an open ball \mathcal{B}_* centered at x_* such that, for each $x \in \mathcal{B}_* \cap \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$, this linear system has a unique solution, call it $(w(x), u(x), \gamma(x), \delta_{\mathcal{A}_*}(x))$, and—due to continuity of the left-hand-side matrix and right-hand-side vector with respect to x —this solution varies continuously over $\mathcal{B}_* \cap \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$. If x_* satisfies (5), then it follows that $(0, 0, c(x_*), [j(x_*)^T c(x_*)]_{\mathcal{A}_*})$ (with $[j(x_*)^T c(x_*)]_{\mathcal{A}_*} > 0$) is the unique solution of the system at $x = x_*$, and for all $x \in \mathcal{B}_* \cap \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ the solution of the system in conjunction with $\delta_i = 0$ for all $i \notin \mathcal{A}_*$ satisfies (17), meaning that the components $(u(x), w(x))$ represent the unique optimal solution of problem (16). Hence, with respect to the quantities in the lemma and using Assumption 1, one finds that $\{v_k\}_{k \in \mathcal{S}} \rightarrow 0$, as desired. To prove the reverse inclusion, suppose that $\{v_k\}_{k \in \mathcal{S}} \rightarrow 0$, from which it follows by the Fundamental Theorem of Linear Algebra and (ii) that $\{(u_k, w_k)\}_{k \in \mathcal{S}} \rightarrow 0$. For all $k \in \mathcal{S}$, let $(u_k, w_k, \gamma_k, \delta_k)$ be a primal-dual optimal solution of (7) (satisfying optimality conditions of the form in (17)). One finds under the conditions of the lemma that, for all sufficiently large $k \in \mathcal{S}$, this solution has $[\delta_k]_i = 0$ for all $i \notin \mathcal{A}(x_*)$ whereas $(u_k, w_k, \gamma_k, [\delta_k]_{\mathcal{A}_*})$ solves the linear system above at $x = x_k$. Since, by the arguments above, this solution varies continuously within $\mathcal{B}_* \cap \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$, the fact that $\{x_k\}_{k \in \mathcal{S}} \rightarrow x_*$ implies that x_* satisfies (5), as desired. \square

In fact, under the conditions of the prior lemma, the quantity $\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2$ also represents a stationarity measure for the problem to minimize $\frac{1}{2}\|c(x)\|_2^2$ subject to $x \in \mathbb{R}_{\geq 0}^n$. This is shown in the following lemma.

Lemma 3. *Suppose that Assumption 1 holds, $\mu_k = \mu \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$, and there exists $\lambda \in \mathbb{R}_{>0}$ and infinite $\mathcal{S}_\lambda \subseteq \mathbb{N}$ such that for some $\{x_k\} \subset \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ one finds $\nabla c(x_k)^T \nabla c(x_k) \succeq \lambda I$ for all $k \in \mathcal{S}_\lambda$. Then, there exists $\kappa_{v,2} \in \mathbb{R}_{>0}$ such that*

$$\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2 \geq \kappa_{v,2} \|v_k\|_2^2 \quad \text{for all } k \in \mathcal{S}_\lambda, \quad (19)$$

where $v_k = u_k + \nabla c(x_k) w_k$ with (u_k, w_k) being the unique optimal solution of (7). Consequently, under the conditions of Lemma 2 with \mathcal{S} defined in that lemma, if \mathcal{S}_λ defined as all sufficiently large indices in \mathcal{S} satisfies the conditions above, then it follows that $\{v_k\}_{k \in \mathcal{S}} \rightarrow 0$ if and only if $\{\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2\}_{k \in \mathcal{S}} \rightarrow 0$.

Proof. Consider arbitrary $k \in \mathcal{S}_\lambda$. Under the stated conditions with $j_k := \nabla c(x_k)^T$, Lemma 1 implies $\|c_k + j_k v_k\|_2 \leq \|c_k\|_2$. Hence, by Assumption 1,

$$\begin{aligned} \|c_k\|_2^2 - \|c_k + j_k v_k\|_2^2 &= (\|c_k\|_2 + \|c_k + j_k v_k\|_2)(\|c_k\|_2 - \|c_k + j_k v_k\|_2) \\ &\leq 2\|c_k\|_2(\|c_k\|_2 - \|c_k + j_k v_k\|_2) \leq 2\kappa_c(\|c_k\|_2 - \|c_k + j_k v_k\|_2). \end{aligned} \quad (20)$$

If $v_k = 0$, then (19) follows trivially. Hence, we may proceed under the assumption that $v_k \neq 0$, which by $v_k = u_k + j_k^T w_k$ and the Fundamental Theorem of Linear Algebra means that $u_k \neq 0$ and/or $w_k \neq 0$. If $w_k = 0$, then it follows by construction of (7) that $u_k = 0$ as well. Hence, we may conclude from $v_k \neq 0$ that, in fact, $w_k \neq 0$. Since (u_k, w_k) is the unique optimal solution of (7), it follows that $\alpha_k^* = 1$ is the optimal solution of the strongly convex quadratic optimization problem

$$\min_{\alpha \in [0,1]} \frac{1}{2} \|c_k + \alpha j_k j_k^T w_k\|_2^2 + \frac{1}{2} \mu_k \|\alpha u_k\|_2^2, \quad (21)$$

which further implies (since an optimality condition of (21) is that the derivative of its objective function with respect to α is less than or equal to zero at $\alpha_k^* = 1$) that $-c_k^T j_k j_k^T w_k \geq \|j_k j_k^T w_k\|_2^2 + \mu_k \|u_k\|_2^2$. Consequently, one finds

$$\begin{aligned} \|c_k\|_2^2 - \|c_k + j_k v_k\|_2^2 &= \|c_k\|_2^2 - \|c_k + j_k j_k^T w_k\|_2^2 \\ &= -2c_k^T j_k j_k^T w_k - \|j_k j_k^T w_k\|_2^2 \geq \|j_k j_k^T w_k\|_2^2 + 2\mu_k \|u_k\|_2^2. \end{aligned} \quad (22)$$

With (20) and (22), it follows from Assumption 1, the conditions of the lemma, and since the Cauchy-Schwarz inequality implies $\|w_k\|_2 \geq \|j_k^T w_k\|_2 / \|j_k^T\|_2$ that

$$\begin{aligned}
\|c_k\|_2 - \|c_k + j_k v_k\|_2 &\geq (2\kappa_c)^{-1} (\|c_k\|_2^2 - \|c_k + j_k v_k\|_2^2) \\
&\geq (2\kappa_c)^{-1} (\|j_k j_k^T w_k\|_2^2 + 2\mu_k \|u_k\|_2^2) \\
&\geq (2\kappa_c)^{-1} (\lambda^2 \|w_k\|_2^2 + 2\mu_k \|u_k\|_2^2) \\
&\geq (2\kappa_c)^{-1} \left(\frac{\lambda^2}{\kappa_{\nabla c}^2} \|j_k^T w_k\|_2^2 + 2\mu_k \|u_k\|_2^2 \right) \\
&\geq (2\kappa_c)^{-1} \min\left\{ \frac{\lambda^2}{\kappa_{\nabla c}^2}, 2\mu_k \right\} (\|j_k^T w_k\|_2^2 + \|u_k\|_2^2) \\
&= (2\kappa_c)^{-1} \min\left\{ \frac{\lambda^2}{\kappa_{\nabla c}^2}, 2\mu \right\} \|v_k\|_2^2 =: \kappa_{v,2} \|v_k\|_2^2,
\end{aligned}$$

which gives (19), as desired. \square

Next, we show that if the point defining subproblem (8) is not an infeasible stationary point for problem (1), then the subproblem with $g_k = \nabla f(x_k)$ yields a zero solution if and only if the point defining the subproblem is stationary for (1).

Lemma 4. *Suppose that Assumption 1 holds and, with respect to $x \in \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$, one finds $c(x) = 0$. Given $H \in \mathbb{R}^{n \times n}$ with $H \succ 0$, consider (recall (8))*

$$\min_{d \in \mathbb{R}^n} \nabla f(x)^T d + \frac{1}{2} d^T H d \text{ s.t. } c(x) + \nabla c(x)^T d = 0 \text{ and } x + d \geq 0. \quad (23)$$

Then, one finds that the optimal solution of problem (23) is $d = 0$ if and only if x is a KKT point (i.e., first-order stationary point) for problem (1).

Proof. Suppose the conditions of the lemma hold and let d be the optimal solution of (23). Since $x \in \mathbb{R}_{\geq 0}^n$ and $c(x) = 0$, it follows that the zero vector is feasible for (23). In addition, necessary and sufficient optimality conditions for subproblem (23) are that, corresponding to $d \in \mathbb{R}^n$, there exist $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$ such that

$$\nabla f(x) + H d + \nabla c(x) y - z = 0, \quad \nabla c(x)^T d = 0, \quad \text{and} \quad 0 \leq x + d \perp z \geq 0. \quad (24)$$

If $d = 0$, then since $c(x) = 0$ it follows that (x, y, z) satisfies (4), as desired. On the other hand, if x is a KKT point for (1), then there exist $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$ such that (x, y, z) satisfies (4), which in turn means that $d = 0$ along with (y, z) satisfies (24), and this solution is unique since the objective of (23) is strongly convex. \square

We conclude this subsection by showing that, under common assumptions and given $x_k \in \mathbb{R}_{\geq 0}^n$, the quantity $\|d_k\|_2^2$, where $d_k \in \mathbb{R}^n$ solves subproblem (8) with $g_k = \nabla f(x_k)$, represents a stationarity measure with respect to (1). (The assumption in the lemma that $H_k = H$ for some $H \succ 0$ for all $k \in \mathbb{N}$ could be relaxed; see Remark 1 at the end of this subsection. We consider this case for the sake of brevity.)

Lemma 5. *Suppose that Assumption 1 holds and there exists infinite $\mathcal{S} \subseteq \mathbb{N}$ such that for some sequence $\{x_k\} \subset \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ one finds $\{x_k\}_{k \in \mathcal{S}} \rightarrow x_*$ for some $x_* \in \mathcal{X} \cap \mathbb{R}_{\geq 0}^n$ with $c(x_*) = 0$ and, with the notation in Lemma 2, one finds that*

$$(i) \quad -\nabla f(x_*) = \nabla c(x_*) y - I_{\mathcal{A}(x_*)}^T z_{\mathcal{A}(x_*)} \text{ for some } (y, z_{\mathcal{A}(x_*)}) \in \mathbb{R}^m \times \mathbb{R}_{>0}^{|\mathcal{A}(x_*)|} \text{ and}$$

$$(ii) \text{ the following matrix has full row rank: } \begin{bmatrix} \nabla c(x_*)^T \\ I_{\mathcal{A}(x_*)} \end{bmatrix}.$$

Then, with $H_k = H$ for some $H \succ 0$ for all $k \in \mathbb{N}$, and with d_k solving (8) with $g_k = \nabla f(x_k)$ for all $k \in \mathbb{N}$, x_ satisfies (4) if and only if $\{\|d_k\|_2^2\}_{k \in \mathcal{S}} \rightarrow 0$.*

Proof. Letting $\mathcal{A}_* := \mathcal{A}(x_*)$ and considering the linear system of equations

$$\begin{bmatrix} H & \nabla c(x) & -I_{\mathcal{A}_*}^T \\ \nabla c(x)^T & 0 & 0 \\ I_{\mathcal{A}_*} & 0 & 0 \end{bmatrix} \begin{bmatrix} d \\ y \\ z_{\mathcal{A}_*} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \\ -x_{\mathcal{A}_*} \end{bmatrix},$$

the proof follows under the conditions of the lemma using the same line of deduction as the proof of Lemma 2, which we omit for the sake of brevity. \square

Remark 1. One might relax the condition in Lemma 2 that $\mu = \mu_k$ for all $k \in \mathbb{N}$ and similarly relax the condition in Lemma 5 that $H_k = H \succ 0$ for all $k \in \mathbb{N}$, such as by requiring merely that $\{\mu_k\}_{k \in \mathbb{N}}$ and $\{H_k\}_{k \in \mathbb{N}}$ have bounded subsequences that converge to some $\mu \in \mathbb{R}_{>0}$ and $H \succ 0$, respectively. In these cases, the “if and only if” statements would be replaced by an “if” statements, which in fact is all that is needed for our subsequent analysis and discussions. Nevertheless, for brevity in the proofs, we provide the conditions that offer the stronger conclusions in these lemmas.

4.2 General Algorithm Behavior

We now prove generally applicable results that hold for every run of Algorithm 1. Our initial results in this section presume that iteration $k \in \mathbb{N}$ is reached and certain properties hold with respect to algorithmic quantities (e.g., $x_k \in \mathbb{R}_{\geq 0}^n$), although we ultimately prove in Lemma 13 that, in fact, these facts are guaranteed, i.e., they hold for any run for any generated $k \in \mathbb{N}$. It is worthwhile to emphasize that the results in this section merely require that $g_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$, which means, for example, that Assumption 2 is not needed in this section. All results that depend on the properties and effects of the stochastic gradient estimates are found in the subsequent subsection, i.e., Section 4.3.

Our first lemma follows directly from Lemma 1, so it is stated without proof.

Lemma 6. Suppose that Assumption 1 holds. Then, in any run of the algorithm such that iteration $k \in \mathbb{N}$ is reached and $x_k \in \mathbb{R}_{\geq 0}^n$, it holds that $v_k = 0$ if and only if x_k satisfies (5), i.e., x_k is either feasible or an infeasible stationary point, whereas $v_k \neq 0$ if and only if $\|c_k\|_2 > \|c_k + \nabla c(x_k)^T v_k\|_2$.

Our next result shows that, in any iteration in which the current iterate x_k is in the nonnegative orthant and $\tau_{k-1} > 0$, the merit parameter is either kept at the same value or decreased, and, if it is decreased, then it is decreased below a constant fraction times its former value. As in other SQP methods with such a feature, this ensures that if the merit parameter sequence does not vanish (i.e., its limiting value is nonzero), then it eventually remains at a constant positive value; see Lemma 13.

Lemma 7. Suppose that Assumption 1 holds. In any run of the algorithm such that line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, and $\tau_{k-1} \in \mathbb{R}_{>0}$, it holds that $0 < \tau_k \leq \tau_{k-1}$, where if $\tau_k < \tau_{k-1}$, then $\tau_k \leq (1 - \epsilon_\tau)\tau_{k-1}$.

Proof. Consider an arbitrary run in which line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, and $\tau_{k-1} \in \mathbb{R}_{>0}$. Let us show that $0 < \tau_k \leq \tau_{k-1}$, in which case the fact that $\tau_k < \tau_{k-1}$ implies $\tau_k \leq (1 - \epsilon_\tau)\tau_{k-1}$ follows from (11). Toward this end, let us next show that $\tau_k^{\text{trial}} > 0$. By the constraints of (8), (10), and Lemma 6, one finds that $\tau_k^{\text{trial}} > 0$ whenever $\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2 > 0$. Hence, to show that one always finds $\tau_k^{\text{trial}} > 0$, all that remains is to consider the case when $\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2 = 0$. In this case, it follows from Lemma 6 that $v_k = 0$, meaning that $d = 0$ is feasible for (8). This, in turn, means that $g_k^T d_k + \frac{1}{2} d_k^T H_k d_k \leq 0$, so by (10) one finds that $\tau_k^{\text{trial}} = \infty > 0$. Since it has been shown that $\tau_k^{\text{trial}} > 0$, the fact that $0 < \tau_k \leq \tau_{k-1}$ now follows directly from (11), completing the proof. \square

We now show that the model reduction offered by the computed search direction satisfies a lower bound with the properties stated in our algorithm development.

Lemma 8. Suppose that Assumptions 1 and 3 hold. In any run of the algorithm such that line 4 is reached in iteration $k \in \mathbb{N}$, $x_k \in \mathbb{R}_{\geq 0}^n$, and $\tau_k \in \mathbb{R}_{>0}$, one finds with ζ from Assumption 3 that

$$\Delta l(x_k, \tau_k, g_k, d_k) \geq \frac{1}{2} \tau_k \zeta \|d_k\|_2^2 + \sigma(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2), \quad (25)$$

and, if $d_k \neq 0$, then $\Delta l(x_k, \tau_k, g_k, d_k) > 0$.

Proof. Consider an arbitrary run in which line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, and $\tau_k \in \mathbb{R}_{>0}$. By (9) and Assumption 3, (25) is implied by

$$(1 - \sigma)(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) \geq \tau_k(g_k^T d_k + \frac{1}{2} d_k^T H_k d_k). \quad (26)$$

If $g_k^T d_k + \frac{1}{2} d_k^T H_k d_k \leq 0$, then (26) holds due to Lemma 6 and the fact that (8) ensures $\nabla c(x_k)^T v_k = \nabla c(x_k)^T d_k$. On the other hand, if $g_k^T d_k + \frac{1}{2} d_k^T H_k d_k > 0$, then one finds by the update of the merit parameter, namely, (10) and (11), that

$$\tau_k \leq \tau_k^{\text{trial}} = \frac{(1 - \sigma)(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2)}{g_k^T d_k + \frac{1}{2} d_k^T H_k d_k},$$

from which (26) follows again. Finally, that $d_k \neq 0$ implies $\Delta l(x_k, \tau_k, g_k, d_k) > 0$ follows from (25), $\tau_k \in \mathbb{R}_{>0}$, and since $\zeta \in \mathbb{R}_{>0}$ in Assumption 3. \square

Our next result is that, under the same conditions as our previous lemmas and under the assumption that $\xi_{k-1} \in \mathbb{R}_{>0}$, the ratio parameter is either kept at the same value or decreased, and, like the merit parameter, if it is decreased, then it is decreased at least below a constant fraction times its previous value.

Lemma 9. Suppose that Assumptions 1 and 3 hold. In any run of the algorithm such that line 4 is reached in iteration $k \in \mathbb{N}$, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{>0}$, and $\xi_{k-1} \in \mathbb{R}_{>0}$, it holds that $0 < \xi_k \leq \xi_{k-1}$, where if $\xi_k < \xi_{k-1}$, then $\xi_k \leq (1 - \epsilon_\xi) \xi_{k-1}$.

Proof. Consider an arbitrary run in which line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{>0}$, and $\xi_{k-1} \in \mathbb{R}_{>0}$. Let us show that $0 < \xi_k \leq \xi_{k-1}$, in which case the fact that $\xi_k < \xi_{k-1}$ implies $\xi_k \leq (1 - \epsilon_\xi) \xi_{k-1}$ follows from (12). Toward this end, observe that if $d_k = 0$, then the algorithm sets $\xi_k \leftarrow \xi_{k-1} > 0$, which is consistent with the desired conclusion. On the other hand, if $d_k \neq 0$, then by (12), $\tau_k \in \mathbb{R}_{>0}$, Lemma 6, the fact that (8) ensures $\nabla c(x_k)^T v_k = \nabla c(x_k)^T d_k$, and Lemma 8,

$$\xi_k^{\text{trial}} = \frac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|_2^2} \geq \frac{\frac{1}{2} \tau_k \zeta \|d_k\|_2^2}{\tau_k \|d_k\|_2^2} = \frac{1}{2} \zeta > 0. \quad (27)$$

Hence, by (12), the desired conclusion follows. \square

Next, we prove bounds for the step size computed in the algorithm.

Lemma 10. Suppose that Assumptions 1 and 3 hold. In any run of the algorithm such that line 4 is reached in iteration $k \in \mathbb{N}$, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{>0}$, and $\xi_k \in \mathbb{R}_{>0}$, it holds that $0 < \alpha_k^{\min} \leq \alpha_k^{\max} \leq \min\{1, \alpha_k^\varphi\}$, and, so, $x_{k+1} \in \mathbb{R}_{\geq 0}^n$.

Proof. Consider an arbitrary run of the algorithm in which line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{>0}$, and $\xi_k \in \mathbb{R}_{>0}$. Let us show that $0 < \alpha_k^{\min} \leq \alpha_k^{\max} \leq 1$, in which case the fact that $x_{k+1} \in \mathbb{R}_{\geq 0}^n$ follows from $x_k \in \mathbb{R}_{\geq 0}^n$, the fact that the constraints of (8) ensure that $x_k + d_k \in \mathbb{R}_{\geq 0}^n$, and since the step size has $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}] \subset (0, 1]$. Toward this end, observe that if $d_k = 0$, then the algorithm yields $\alpha_k = \alpha_k^{\min} = \alpha_k^{\max} = \alpha_k^\varphi = 1$, so the conclusion follows trivially. Hence, let us assume $d_k \neq 0$. Observe that from (13), the algorithm uses α_k^{\min} with

$$0 < \alpha_k^{\min} = \frac{2(1-\eta)\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} \leq 1. \quad (28)$$

Now observing (15), which shows $\alpha_k^{\max} \leq \min\{1, \alpha_k^\varphi\}$, one finds that all that remains is to prove that $\alpha_k^{\min} \leq \alpha_k^\varphi$. For this purpose, let us introduce

$$\alpha_k^{\text{suff}} := \min \left\{ 1, \frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma) \|d_k\|_2^2} \right\},$$

where $\alpha_k^{\text{suff}} \in (0, 1]$ follows by $\beta_k \in (0, 1]$, Lemma 8, and $d_k \neq 0$. To show that $\alpha_k^{\min} \leq \alpha_k^\varphi$, our aim is to show that $\alpha_k^{\min} \leq \alpha_k^{\text{suff}} \leq \alpha_k^\varphi$. First, from (12), one finds

$$\alpha_k^{\min} = \frac{2(1-\eta)\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} \leq \frac{2(1-\eta)\beta_k \xi_k^{\text{trial}} \tau_k}{\tau_k L + \Gamma} = \frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma) \|d_k\|_2^2}. \quad (29)$$

Combining (28) and (29), one finds that $\alpha_k^{\min} \leq \alpha_k^{\text{suff}}$, as desired. Now, toward proving that $\alpha_k^{\text{suff}} \leq \alpha_k^\varphi$, let us first show that $\varphi_k(\alpha_k^{\text{suff}}) \leq 0$. From the triangle inequality, the fact that $\alpha_k^{\text{suff}} \in (0, 1]$, and (14), it follows that

$$\begin{aligned} & \varphi_k(\alpha_k^{\text{suff}}) \\ &= (\eta - 1)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \|c_k + \alpha_k^{\text{suff}} \nabla c(x_k)^T d_k\|_2 - \|c_k\|_2 \\ & \quad + \alpha_k^{\text{suff}} (\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) + \frac{1}{2}(\tau_k L + \Gamma)(\alpha_k^{\text{suff}})^2 \|d_k\|_2^2 \\ &\leq (\eta - 1)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + (1 - \alpha_k^{\text{suff}}) \|c_k\|_2 + \alpha_k^{\text{suff}} \|c_k + \nabla c(x_k)^T d_k\|_2 \\ & \quad - \|c_k\|_2 + \alpha_k^{\text{suff}} (\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) + \frac{1}{2}(\tau_k L + \Gamma)(\alpha_k^{\text{suff}})^2 \|d_k\|_2^2 \\ &= (\eta - 1)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \frac{1}{2}(\tau_k L + \Gamma)(\alpha_k^{\text{suff}})^2 \|d_k\|_2^2 \\ &\leq (\eta - 1)\alpha_k^{\text{suff}} \beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \frac{1}{2}\alpha_k^{\text{suff}} (\tau_k L + \Gamma) \|d_k\|_2^2 \left(\frac{2(1-\eta)\beta_k \Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma) \|d_k\|_2^2} \right) \\ &= 0. \end{aligned}$$

Therefore, by (15), it follows that $\alpha_k^{\text{suff}} \leq \alpha_k^\varphi$. \square

Our next lemma shows an upper bound on the change in the merit function. In the lemma and throughout the rest of the paper, for any $k \in \mathbb{N}$ such that line 4 is reached we let $d_k^{\text{true}} \in \mathbb{R}^n$ denote the solution of (8) when g_k is replaced by $\nabla f(x_k)$.

Lemma 11. *Suppose that Assumptions 1 and 3 hold. In any run of the algorithm such that line 4 is reached in iteration $k \in \mathbb{N}$, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{> 0}$, and $\alpha_k \in (0, \alpha_k^\varphi]$, it holds that*

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

Proof. Consider an arbitrary run of the algorithm in which line 4 of iteration $k \in \mathbb{N}$ is reached, $x_k \in \mathbb{R}_{\geq 0}^n$, $\tau_k \in \mathbb{R}_{> 0}$, and $\alpha_k \in (0, \alpha_k^\varphi]$. By Assumption 1 (which led to (3)), (8) (which implies $c_k + \nabla c(x_k)^T d_k = c_k + \nabla c(x_k)^T d_k^{\text{true}}$), (9), (14), and the fact that $0 < \alpha_k \leq \alpha_k^\varphi$ (which means $\varphi_k(\alpha_k) \leq 0$), it follows that

$$\begin{aligned} & \phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \\ &= \tau_k (f(x_k + \alpha_k d_k) - f_k) + \|c(x_k + \alpha_k d_k)\|_2 - \|c_k\|_2 \\ &\leq \alpha_k \tau_k \nabla f(x_k)^T d_k + \|c_k + \alpha_k \nabla c(x_k)^T d_k\|_2 - \|c_k\|_2 + \frac{1}{2}(\tau_k L + \Gamma) \alpha_k^2 \|d_k\|_2^2 \\ &= -\alpha_k \Delta l(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}}) + \alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}}) + \|c_k + \alpha_k \nabla c(x_k)^T d_k\|_2 \\ & \quad - \|c_k\|_2 + \alpha_k (\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) + \frac{1}{2}(\tau_k L + \Gamma) \alpha_k^2 \|d_k\|_2^2 \\ &\leq -\alpha_k \Delta l(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}}) + \alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\ & \quad + (1 - \eta) \alpha_k \beta_k \Delta l(x_k, \tau_k, g_k, d_k), \end{aligned}$$

which shows the desired conclusion. \square

We now show that each search direction—and, similarly, the search direction that would be computed if the true gradient of the objective function were used in place of the stochastic gradient estimate—can be viewed as a projection of the unconstrained minimizer of the objective of (8) onto the feasible region of (8).

Lemma 12. *Suppose that Assumptions 1 and 3 hold. In any run of the algorithm such that line 4 is reached in iteration $k \in \mathbb{N}$, $x_k \in \mathbb{R}_{\geq 0}$, and with*

$$\mathcal{D}_k := \{d \in \mathbb{R}^n : \nabla c(x_k)^T(d - v_k) = 0, x_k + d \geq 0\} \text{ and } \text{Proj}_k(\bar{d}) := \underset{d \in \mathcal{D}_k}{\operatorname{argmin}} \|d - \bar{d}\|_{H_k}^2,$$

it holds that $d_k = \text{Proj}_k(-H_k^{-1}g_k)$ and $d_k^{\text{true}} = \text{Proj}_k(-H_k^{-1}\nabla f(x_k))$.

Proof. Consider an arbitrary run of the algorithm in which line 4 of iteration $k \in \mathbb{N}$ is reached and $x_k \in \mathbb{R}_{\geq 0}$. The desired conclusion follows from the facts that \mathcal{D}_k is convex and, under Assumption 3, H_k is SPD; in particular, one finds that

$$d_k = \underset{d \in \mathcal{D}_k}{\operatorname{argmin}} g_k^T d + \frac{1}{2} d^T H_k d = \underset{d \in \mathcal{D}_k}{\operatorname{argmin}} \frac{1}{2} \|d + H_k^{-1}g_k\|_{H_k}^2 = \text{Proj}_k(-H_k^{-1}g_k),$$

and similarly with respect to d_k^{true} with g_k replaced by $\nabla f(x_k)$. \square

We are now prepared to prove the following lemma, which shows that the algorithm is well defined and either terminates finitely with an infeasible stationary point or generates an infinite sequence of iterates with certain critical properties of the simultaneously generated algorithmic sequences. The lemma also reveals that the monotonically nonincreasing merit parameter sequence either vanishes or ultimately remains constant, and it reveals that the monotonically nonincreasing ratio parameter sequence ultimately remains constant at a value that is greater than or equal to a positive real number that is defined uniformly across all runs of the algorithm.

Lemma 13. *Suppose that Assumptions 1 and 3 hold. In any run, either the algorithm terminates finitely with an infeasible stationary point or it performs an infinite number of iterations such that, for all $k \in \mathbb{N}$, it holds that*

- (a) $x_k \in \mathbb{R}_{\geq 0}^n$,
- (b) $v_k = 0$ if and only if x_k satisfies (5),
- (c) $v_k \neq 0$ if and only if $\|c_k\|_2 > \|c_k + \nabla c(x_k)^T v_k\|_2$,
- (d) $0 < \tau_k \leq \tau_{k-1} < \infty$,
- (e) $\tau_k < \tau_{k-1}$ if and only if $\tau_k \leq (1 - \epsilon_\tau)\tau_{k-1}$,
- (f) (25) holds,
- (g) $d_k \neq 0$ if and only if $\Delta l(x_k, \tau_k, g_k, d_k) > 0$,
- (h) $0 < \xi_k \leq \xi_{k-1} < \infty$,
- (i) $\xi_k < \xi_{k-1}$ if and only if $\xi_k \leq (1 - \epsilon_\xi)\xi_{k-1}$, and
- (j) $0 < \alpha_k^{\min} \leq \alpha_k^{\max} \leq \min\{1, \alpha_k^\varphi\}$.

In addition, in any run that does not terminate finitely, it holds that

- (k) either $\{\tau_k\} \searrow 0$ or there exists $k_\tau \in \mathbb{N}$ and $\tau_{\min} \in \mathbb{R}_{>0}$ such that $\tau_k = \tau_{\min}$ for all $k \in \mathbb{N}$ with $k \geq k_\tau$, and
- (l) there exist $k_\xi \in \mathbb{N}$ and $\xi_{\min} \in \mathbb{R}_{>0}$ with $\xi_{\min} \geq \frac{1}{2}\zeta(1 - \epsilon_\xi)$ such that $\xi_k = \xi_{\min}$ for all $k \in \mathbb{N}$ with $k \geq k_\xi$.

Proof. Given the initialization of the algorithm, statements (a)–(j) follow by induction from Lemmas 6–10. Statement (k) follows from statements (d) and (e). Finally, to prove statement (l), consider arbitrary $k \in \mathbb{N}$ in a run that does not terminate finitely and note that if $d_k = 0$, then $\xi_k^{\text{trial}} \leftarrow \infty$, and if $d_k \neq 0$, then ξ_k^{trial} satisfies (27), meaning that $\xi_k^{\text{trial}} \geq \frac{1}{2}\zeta$. Consequently, by (12), $\xi_k < \xi_{k-1}$ only if $\xi_{k-1} > \frac{1}{2}\zeta$. This, along with statements (h) and (i), leads to the conclusion. \square

4.3 Convergence Guarantees

We now turn to prove convergence results under Assumption 4 below. Recalling the role of $\frac{1}{2}\zeta(1 - \epsilon_\xi) \in \mathbb{R}_{>0}$ in Lemma 13(l), the assumption focuses on the following event for some $(k_{\min}, \tau_{\min}, f_{\sup}) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}$, where for all generated $k \in \mathbb{N}$ we denote $\tau_k^{\text{true,trial}}$ as the value of τ_k^{trial} that would be computed in iteration k if (8) were solved with $\nabla f(x_k)$ in place of g_k :

$$\begin{aligned} &:= \{ \text{An infinite number of iterations are performed, } f(x_{k_{\min}}) \leq f_{\sup}, \text{ and} \\ &\quad \text{there exist } k' \in \mathbb{N} \text{ with } k' \leq k_{\min}, \tau' \in \mathbb{R}_{>0} \text{ with } \tau' \geq \tau_{\min}, \\ &\quad \text{and } \xi' \in \mathbb{R}_{>0} \text{ with } \xi' \geq \frac{1}{2}\zeta(1 - \epsilon_\xi) \text{ such that} \\ &\quad \tau_k = \tau' \leq \tau_k^{\text{true,trial}} \text{ and } \xi_k = \xi' \text{ for all } k \in \mathbb{N} \text{ with } k \geq k' \}. \end{aligned} \quad (30)$$

The following assumption is made in this subsection. We present a discussion and supporting theoretical results about this assumption in Section 4.4.

Assumption 4. For some $(k_{\min}, \tau_{\min}, f_{\sup}) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}$, the event $\mathcal{E} := \mathcal{E}(k_{\min}, \tau_{\min}, f_{\sup})$ occurs and, conditioned on the occurrence of \mathcal{E} , Assumption 1 holds (with the same constants as previously presented in (2) and (3)).

It is not a shortcoming of our analysis that Assumption 4, through the definition of \mathcal{E} , assumes that (i) an infinite number of iterations are performed, (ii) the objective value is bounded above in iteration k_{\min} , and (iii) $\{\xi_k\}$ ultimately becomes a constant sequence with value at least $\frac{1}{2}\zeta(1 - \epsilon_\xi) \in \mathbb{R}_{>0}$. After all: (i) Lemma 13 shows that the only alternative to an infinite number of iterations being performed is that the algorithm terminates finitely with an infeasible stationary point, in which case there is nothing else to prove; (ii) $f_{\sup} \in \mathbb{R}$ can be arbitrarily large and knowledge of it is not required by the algorithm, so assuming that it exists is a very loose requirement; and (iii) Lemma 13(l) shows that, in any run that does not terminate finitely, $\{\xi_k\}$ is monotonically nonincreasing and bounded below by $\frac{1}{2}\zeta(1 - \epsilon_\xi) \in \mathbb{R}_{>0}$, which is a constant, i.e., it is not run-dependent. Overall, the only important restriction of our analysis in this section is the fact that \mathcal{E} includes the requirement that $\{\tau_k\}$ ultimately becomes constant at a value at least τ_{\min} that is sufficiently small relative to $\{\tau_k^{\text{true,trial}}\}$. This restriction is the subject of Section 4.4.

For the remainder of this subsection, we consider the stochastic process corresponding to the statement of Algorithm 1. Specifically, the sequence

$$\{(x_k, v_k, g_k, d_k, d_k^{\text{true}}, \tau_k^{\text{trial}}, \tau_k^{\text{true,trial}}, \tau_k, \xi_k^{\text{trial}}, \xi_k, \alpha_k^{\min}, \alpha_k^\varphi, \alpha_k^{\max}, \alpha_k)\}$$

generated in any run can be viewed as a realization of the stochastic process

$$\{(X_k, V_k, G_k, D_k, D_k^{\text{true}}, \mathcal{T}_k^{\text{trial}}, \mathcal{T}_k^{\text{true,trial}}, \mathcal{T}_k, \Xi_k^{\text{trial}}, \Xi_k, \mathcal{A}_k^{\min}, \mathcal{A}_k^\varphi, \mathcal{A}_k^{\max}, \mathcal{A}_k)\}.$$

Let \mathcal{G}_1 denote the σ -algebra defined by the initial conditions of the algorithm and, for all $k \in \mathbb{N}$ with $k \geq 2$, let \mathcal{G}_k denote the σ -algebra generated by the initial conditions and the random variables $\{G_1, \dots, G_{k-1}\}$. Then, with respect to the event \mathcal{E} in Assumption 4, denote the trace σ -algebra of \mathcal{E} on \mathcal{G}_k as $\mathcal{F}_k := \mathcal{G}_k \cap \mathcal{E}$ for all $k \in \mathbb{N}$. It follows that $\{\mathcal{F}_k\}$ is a filtration, and we proceed in our analysis under Assumptions 2, 3, and 4 (which subsumes Assumption 1) with the definitions

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

(where P_ω denotes probability taken with respect to the distribution of ω). We also define, with respect to \mathcal{E} , the random variables $K' \leq k_{\min}$, $\mathcal{T}' \geq \tau_{\min}$, and $\Xi' \geq \frac{1}{2}\zeta(1 - \epsilon_\xi)$, which for a given run of the algorithm have the realized values k' , τ' , and ξ' , respectively, defined in (30). Conditioned on \mathcal{E} , one has in any run that

$$\tau_{\min} \leq \mathcal{T}' \leq \tau_0 \quad \text{and} \quad \frac{1}{2}\zeta(1 - \epsilon_\xi) \leq \Xi' \leq \xi_0, \quad (31)$$

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

Our next lemma shows upper bounds on the norm of the difference between the computed search direction and the search direction that would be computed with the true gradient of the objective. (The conclusion of this lemma and the following one would hold even without assuming that the event \mathcal{E} occurs, but in each result we condition on $\mathcal{F}_k := \mathcal{G}_k \cap \mathcal{E}$ for use in our ultimate results under \mathcal{E} .)

Lemma 14. *Suppose that Assumptions 2, 3, and 4 hold. For all $k \in \mathbb{N}$,*

$$\begin{aligned} \|D_k - D_k^{\text{true}}\|_2 &\leq \zeta^{-1} \|G_k - \nabla f(X_k)\|_2 \\ \text{and } \mathbb{E}_k[\|D_k - D_k^{\text{true}}\|_2] &\leq \zeta^{-1} \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \leq \zeta^{-1} \sqrt{\rho_k}. \end{aligned}$$

Proof. Consider arbitrary $k \in \mathbb{N}$ under the stated conditions. Lemma 12 and the obtuse angle lemma for projections [6, Proposition 1.1.9] imply

$$\begin{aligned} (D_k - D_k^{\text{true}})^T H_k (-H_k^{-1} \nabla f(X_k) - D_k^{\text{true}}) &\leq 0 \\ \text{and } (D_k^{\text{true}} - D_k)^T H_k (-H_k^{-1} G_k - D_k) &\leq 0. \end{aligned}$$

Summing these inequalities yields

$$\begin{aligned} 0 &\geq (D_k - D_k^{\text{true}})^T H_k (-H_k^{-1} \nabla f(X_k) - D_k^{\text{true}}) + (D_k^{\text{true}} - D_k)^T H_k (-H_k^{-1} G_k - D_k) \\ &= \|D_k - D_k^{\text{true}}\|_{H_k}^2 - (D_k - D_k^{\text{true}})^T (\nabla f(X_k) - G_k). \end{aligned}$$

Hence, by the Cauchy–Schwarz inequality, it follows that

$$\|D_k - D_k^{\text{true}}\|_{H_k}^2 \leq (D_k - D_k^{\text{true}})^T (\nabla f(X_k) - G_k) \leq \|D_k - D_k^{\text{true}}\|_2 \|\nabla f(X_k) - G_k\|_2,$$

which shows under Assumption 3 that $\|D_k - D_k^{\text{true}}\|_2 \leq \zeta^{-1} \|G_k - \nabla f(X_k)\|_2$, as desired. Then, from this inequality, Assumption 2, and Jensen’s inequality, one has

$$\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \leq \sqrt{\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2]} \leq \sqrt{\rho_k},$$

from which the remainder of the conclusion follows. \square

We now show an upper bound on the expected difference between inner products involving the true and stochastic gradients and the true and stochastic directions.

Lemma 15. *Suppose that Assumptions 2, 3, and 4 hold. For all $k \geq k_{\min}$,*

$$\begin{aligned} |\mathbb{E}_k[G_k^T D_k - \nabla f(X_k)^T D_k^{\text{true}}]| &\leq \zeta^{-1} (\rho_k + \kappa_{\nabla f} \sqrt{\rho_k}) \\ \text{and } \mathbb{E}_k[\Delta l(X_k, \mathcal{T}_k, G_k, D_k)] - \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) &\leq \mathcal{T}' \zeta^{-1} (\rho_k + \kappa_{\nabla f} \sqrt{\rho_k}). \end{aligned}$$

Proof. Consider arbitrary $k \geq k_{\min}$ under the stated conditions. From the triangle and Cauchy–Schwarz inequalities and Lemma 14, it holds that

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

$$\begin{aligned}
&= |\mathbb{E}_k[(G_k - \nabla f(X_k))^T(D_k - D_k^{\text{true}})] + \mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})]| \\
&\leq \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2 \|D_k - D_k^{\text{true}}\|_2] + \|\nabla f(X_k)\|_2 \mathbb{E}_k[\|D_k - D_k^{\text{true}}\|_2] \\
&\leq \zeta^{-1} \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2] + \zeta^{-1} \kappa_{\nabla f} \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \leq \zeta^{-1} \rho_k + \zeta^{-1} \kappa_{\nabla f} \sqrt{\rho_k},
\end{aligned}$$

which gives the first result. Then, for $k \geq k_{\min}$, (9) and the equation above give

$$\begin{aligned}
&\mathbb{E}_k[\Delta l(X_k, \mathcal{T}_k, G_k, D_k)] - \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \\
&= \mathcal{T}' \mathbb{E}_k[\nabla f(X_k)^T D_k^{\text{true}} - G_k^T D_k] \leq \mathcal{T}' \zeta^{-1} (\rho_k + \kappa_{\nabla f} \sqrt{\rho_k}),
\end{aligned}$$

which completes the proof. \square

Our next lemma shows a lower bound on the true model reduction. In the lemma and our subsequent results, we define $J_k := \nabla c(X_k)^T$ for the sake of brevity.

Lemma 16. *Suppose that Assumptions 2, 3, and 4 hold. For all $k \geq k_{\min}$,*

$$\Delta l(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}}) \geq \frac{1}{2} \mathcal{T}' \zeta \|D_k^{\text{true}}\|_2^2 + \sigma(\|c(X_k)\|_2 - \|c(X_k) + J_k D_k^{\text{true}}\|_2) \geq 0.$$

Proof. Consider arbitrary $k \geq k_{\min}$ under the stated conditions. By (9), the fact that $\mathcal{T}_k = \mathcal{T}'$, and Assumption 3, the first desired conclusion is implied by

$$(1 - \sigma)(\|c(X_k)\|_2 - \|c(X_k) + J_k D_k^{\text{true}}\|_2) \geq \mathcal{T}'(\nabla f(X_k)^T D_k^{\text{true}} + \frac{1}{2}(D_k^{\text{true}})^T H_k D_k^{\text{true}}).$$

If $\nabla f(X_k)^T D_k^{\text{true}} + \frac{1}{2}(D_k^{\text{true}})^T H_k D_k^{\text{true}} \leq 0$, then the above holds due to Lemma 13 and the fact that $J_k D_k^{\text{true}} = J_k V_k$; else, $\nabla f(X_k)^T D_k^{\text{true}} + \frac{1}{2}(D_k^{\text{true}})^T H_k D_k^{\text{true}} > 0$, in which case one finds from the conditions of the lemma, (10), and (11) that

$$\mathcal{T}_k = \mathcal{T}' \leq \mathcal{T}_k^{\text{true, trial}} = \frac{(1 - \sigma)(\|c(X_k)\|_2 - \|c(X_k) + J_k D_k^{\text{true}}\|_2)}{\nabla f(X_k)^T D_k^{\text{true}} + \frac{1}{2}(D_k^{\text{true}})^T H_k D_k^{\text{true}}},$$

from which the displayed inequality above follows again. Finally, the remaining desired conclusion follows from the first conclusion, Lemma 13, and $J_k D_k^{\text{true}} = J_k V_k$. \square

Next, we prove a critical upper bound on the expected value of the second term on the right-hand side of the upper bound proved in Lemma 11.

Lemma 17. *Suppose that Assumptions 2, 3, and 4 hold. For all $k \geq k_{\min}$,*

$$\mathbb{E}_k[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})] \leq \left(\frac{2(1-\eta)\Xi' \mathcal{T}'}{\mathcal{T}' L + \Gamma} + \theta \right) \beta_k \mathcal{T}' \kappa_{\nabla f} \zeta^{-1} \sqrt{\rho_k}.$$

Proof. For arbitrary $k \geq k_{\min}$ under the conditions, (13) and (15) yield

$$\mathcal{A}_k^{\min} = \beta_k \mathcal{A}' \quad \text{and} \quad \mathcal{A}_k^{\max} \leq \mathcal{A}_k^{\min} + \theta \beta_k, \quad \text{where} \quad \mathcal{A}' = \frac{2(1-\eta)\Xi' \mathcal{T}'}{\mathcal{T}' L + \Gamma}. \quad (32)$$

Letting \mathcal{P}_k denote the event that $\nabla f(X_k)^T (D_k - D_k^{\text{true}}) \geq 0$ and letting \mathcal{P}_k^c denote the event that $\nabla f(X_k)^T (D_k - D_k^{\text{true}}) < 0$, the law of total expectation and the fact that \mathcal{T}' and Ξ' are \mathcal{F}_k -measurable for $k \geq k_{\min}$ shows that

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

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

The Cauchy-Schwarz inequality and law of total expectation show that

$$\begin{aligned}
&\mathbb{P}_k[\mathcal{P}_k] \cdot \mathbb{E}_k[\nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{P}_k] \\
&\leq \mathbb{P}_k[\mathcal{P}_k] \cdot \mathbb{E}_k[\|\nabla f(X_k)\|_2 \|D_k - D_k^{\text{true}}\|_2 | \mathcal{P}_k] \\
&= \mathbb{E}_k[\|\nabla f(X_k)\|_2 \|D_k - D_k^{\text{true}}\|_2] - \mathbb{P}_k[\mathcal{P}_k^c] \cdot \mathbb{E}_k[\|\nabla f(X_k)\|_2 \|D_k - D_k^{\text{true}}\|_2 | \mathcal{P}_k^c] \\
&\leq \mathbb{E}_k[\|\nabla f(X_k)\|_2 \|D_k - D_k^{\text{true}}\|_2],
\end{aligned}$$

so from above, the Cauchy-Schwarz inequality, Assumption 4, and Lemma 14,

$$\begin{aligned}
\mathbb{E}_k[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})] &\leq (\mathcal{A}_k^{\min} + \theta \beta_k) \mathcal{T}' \|\nabla f(X_k)\|_2 \mathbb{E}_k[\|D_k - D_k^{\text{true}}\|_2] \\
&\leq \left(\frac{2(1-\eta)\Xi' \mathcal{T}'}{\mathcal{T}' L + \Gamma} + \theta \right) \beta_k \mathcal{T}' \kappa_{\nabla f} \zeta^{-1} \sqrt{\rho_k},
\end{aligned}$$

which gives the desired conclusion. \square

We now present, as a lemma, results pertaining to the asymptotic behavior of the model reductions generated by the algorithm. In the subsequent theorem after the lemma, these results will be translated in terms of quantities that, as seen in Section 4.1, can be connected to stationarity measures related to problem (1). We remark that the conditions of the lemma can be satisfied in a run-dependent manner if, every time the merit or ratio parameter is decreased, say in iteration $\hat{k} \in \mathbb{N}$, the sequence $\{\beta_k\}$ is “restarted” such that with $\alpha' = 2(1-\eta)\xi_{\hat{k}}\tau_{\hat{k}}/(\tau_{\hat{k}}L + \Gamma)$ and some (run-independent) $\psi \in (0, 1]$ one chooses $\beta_k = \beta = \psi \frac{\alpha'}{2(1-\eta)(\alpha' + \theta)}$ for part (a) of the lemma and $\beta_k = \frac{1}{k - \hat{k} + 1} \psi \frac{\alpha'}{2(1-\eta)(\alpha' + \theta)}$ for part (b); such a scheme was described in [1] as well. Notice that in this situation, β and $\{\beta_k\}_{k \geq \hat{k}}$ in parts (a) and (b), respectively, are random variables, *but, importantly, they are \mathcal{F}_k -measurable for $k \geq \hat{k}_{\min}$* . Alternatively, one could choose $\{\beta_k\}$ using the same formulas, but with ξ_{\min} and τ_{\min} in place of ξ_k and τ_k , respectively, in the formula for α' , in which case the choices are run-independent. The downside of relying on this latter situation is that it requires knowledge of ξ_{\min} and τ_{\min} , which would not typically be known *a priori*. Hence, we analyze the former scheme, but use run-dependent bounds that, under \mathcal{E} , are defined with respect to ξ_{\min} and τ_{\min} (even though these values are unknown).

We also remark that for case (a) in the following lemma, the sequence $\{\rho_k\}$, which bounds the expected squared error in the stochastic gradient estimates, can be a constant sequence. However, for case (b), the relationship between $\{\rho_k\}$ and $\{\beta_k\}$ means that the expected squared error in the gradient estimates must vanish as $k \rightarrow \infty$. This requirement, which is stronger than the requirement for equality-constraints-only case in [1], is needed to overcome the fact that in the presence of bound constraints the search directions can be biased estimates of their true counterparts.

Lemma 18. *Under Assumptions 2, 3, and 4, suppose that $\{\rho_k\}$ is chosen such that there exists $\iota \in \mathbb{R}_{>0}$ with $\rho_k \leq \iota \beta_k^2$ for all $k \in \mathbb{N}$ with $k \geq k_{\min}$, and define*

$$\begin{aligned}
\alpha'_{\min} &= \frac{2(1-\eta)\xi_{\min}\tau_{\min}}{\tau_{\min}L + \Gamma}, \quad \alpha'_{\max} = \frac{2(1-\eta)\xi_0\tau_0}{\tau_0L + \Gamma}, \\
\text{and } \rho'_{\max} &= (\alpha'_{\max} + \theta)\tau_0\zeta^{-1}(\kappa_{\nabla f}\sqrt{\iota} + (1-\eta)(\iota + \kappa_{\nabla f}\sqrt{\iota})).
\end{aligned}$$

Then, with \mathcal{A}' defined in (32) and $\mathbb{E}[\cdot | \mathcal{E}]$ denoting total expectation over all realizations of the algorithm conditioned on event \mathcal{E} , the following statements hold true.

(a) *if $\beta_k = \beta = \psi \frac{\mathcal{A}'}{2(1-\eta)(\mathcal{A}' + \theta)}$ for some $\psi \in (0, 1]$ for all $k \geq k_{\min}$, then*

$$\limsup_{k \rightarrow \infty} \mathbb{E} \left[\frac{1}{k} \sum_{j=k_{\min}}^{k_{\min}+k-1} \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}}) \middle| \mathcal{E} \right] \leq \frac{\psi(\alpha'_{\max})^2(\alpha'_{\min} + \theta)\rho'_{\max}}{2(1-\eta)(1-\frac{\psi}{2})(\alpha'_{\min})^2(\alpha'_{\max} + \theta)^2};$$

(b) if $\sum_{k=k_{\min}}^{\infty} \beta_k = \infty$, $\sum_{k=k_{\min}}^{\infty} \beta_k^2 < \infty$, and $\beta_k \leq \psi \frac{\mathcal{A}'}{2(1-\eta)(\mathcal{A}'+\theta)}$ for some $\psi \in (0, 1]$ for all $k \geq k_{\min}$, it holds that

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

Proof. For arbitrary $k \geq k_{\min}$ under the conditions, it follows from Lemma 11, Lemma 16 (which shows $\Delta l(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}}) \geq 0$), (32), the fact that $\mathcal{A}_k \geq \mathcal{A}_k^{\min} = \mathcal{A}'\beta_k$, Lemma 17, the fact that $\mathcal{A}_k \leq \mathcal{A}_k^{\max} \leq \mathcal{A}_k^{\min} + \theta\beta_k = (\mathcal{A}' + \theta)\beta_k$, Lemma 14, Lemma 15, and $\beta_k \in (0, 1]$ that

$$\begin{aligned} & \mathbb{E}_k[\phi(X_{k+1}, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k)] = \mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k)] \\ & \leq \mathbb{E}_k[-\mathcal{A}_k \Delta l(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}}) \\ & \quad + \mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}}) + (1 - \eta) \mathcal{A}_k \beta_k \Delta l(X_k, \mathcal{T}_k, G_k, D_k)] \\ & \leq -\mathcal{A}' \beta_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \\ & \quad + (\mathcal{A}' + \theta) \beta_k \mathcal{T}' \kappa_{\nabla f} \zeta^{-1} \sqrt{\rho_k} \\ & \quad + (1 - \eta) (\mathcal{A}' + \theta) \beta_k^2 (\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \mathcal{T}' \zeta^{-1} (\rho_k + \kappa_{\nabla f} \sqrt{\rho_k})) \\ & \leq -\mathcal{A}' \beta_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \\ & \quad + (\mathcal{A}' + \theta) \beta_k \mathcal{T}' \kappa_{\nabla f} \zeta^{-1} \sqrt{\iota} \beta_k \\ & \quad + (1 - \eta) (\mathcal{A}' + \theta) \beta_k^2 (\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \mathcal{T}' \zeta^{-1} (\iota \beta_k^2 + \kappa_{\nabla f} \sqrt{\iota} \beta_k)) \\ & \leq -\beta_k (\mathcal{A}' - (1 - \eta) (\mathcal{A}' + \theta) \beta_k) \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + R' \beta_k^2, \end{aligned}$$

where $R' = (\mathcal{A}' + \theta) \mathcal{T}' \zeta^{-1} (\kappa_{\nabla f} \sqrt{\iota} + (1 - \eta) (\iota + \kappa_{\nabla f} \sqrt{\iota}))$. Now, from Assumption 4 (which subsumes Assumption 1), there exists $\phi_{\min} \in \mathbb{R}$ such that $\phi(X_k, \mathcal{T}') \geq \phi_{\min}$ for all $k \geq k_{\min}$. One also finds that $\alpha'_{\min} \leq \mathcal{A}' \leq \alpha'_{\max}$ due to the monotonicity of $\frac{2(1-\eta)\Xi'\tau}{\tau L + \Gamma}$ with respect to τ . Therefore, under part (a) of the lemma, in which case one finds for $k \geq k_{\min}$ that $\psi \frac{\alpha'_{\min}}{2(1-\eta)(\alpha'_{\min} + \theta)} \leq \beta \leq \psi \frac{\alpha'_{\max}}{2(1-\eta)(\alpha'_{\max} + \theta)}$, it follows from above that

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

so by taking total expectation conditioned on the event \mathcal{E} one finds

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

Rearranging terms, observing that $\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}') | \mathcal{E}]$ is bounded above under Assumption 4, and considering the limit superior as $k \rightarrow \infty$, the conclusion of part (a) follows. On the other hand, under the conditions of part (b), it follows in a similar manner that, for any $k \in \mathbb{N}$, one finds

$$\begin{aligned} & \phi_{\min} - \mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}') | \mathcal{E}] \\ & \leq \mathbb{E}[\phi(X_{k_{\min}+k}, \mathcal{T}') - \phi(X_{k_{\min}}, \mathcal{T}') | \mathcal{E}] = \mathbb{E} \left[\sum_{j=k_{\min}}^{k_{\min}+k-1} (\phi(X_{j+1}, \mathcal{T}') - \phi(X_j, \mathcal{T}')) \middle| \mathcal{E} \right] \end{aligned}$$

$$\leq \mathbb{E} \left[\sum_{j=k_{\min}}^{k_{\min}+k-1} (-\beta_j(\mathcal{A}' - (1-\eta)(\mathcal{A}' + \theta)\beta_j) \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}}) + R' \beta_j^2) \middle| \mathcal{E} \right].$$

Taking limits as $k \rightarrow \infty$, the conclusion of part (b) follows. \square

We now present our main convergence theorem for Algorithm 1, which is essentially a translation of Lemma 18 from results about model reductions to results about quantities connected to measures of stationarity for problem (1).

Theorem 1. *Suppose the conditions of Lemma 18 hold. Then,*

(a) *under the conditions of Lemma 18(a), there exists $C \in \mathbb{R}_{>0}$ such that*

$$\limsup_{k \rightarrow \infty} \mathbb{E} \left[\frac{1}{k} \sum_{j=k_{\min}}^{k_{\min}+k-1} (\tfrac{1}{2} \mathcal{T}' \zeta \|D_j^{\text{true}}\|_2^2 + \sigma(\|c(X_j)\|_2 - \|c(X_j) + J_j D_j^{\text{true}}\|_2)) \middle| \mathcal{E} \right] = C;$$

(b) *under the conditions of Lemma 18(b), with $B_k := \sum_{j=k_{\min}}^{k_{\min}+k-1} \beta_j$,*

$$\mathbb{E} \left[\frac{1}{B_k} \sum_{j=k_{\min}}^{k_{\min}+k-1} \beta_j (\tfrac{1}{2} \mathcal{T}' \zeta \|D_j^{\text{true}}\|_2^2 + \sigma(\|c(X_j)\|_2 - \|c(X_j) + J_j D_j^{\text{true}}\|_2)) \middle| \mathcal{E} \right] \xrightarrow{k \rightarrow \infty} 0,$$

which further implies $\liminf_{k \rightarrow \infty} \mathbb{E}[\|D_k^{\text{true}}\|_2^2 + (\|c(X_k)\|_2 - \|c(X_k) + J_k D_k^{\text{true}}\|_2) | \mathcal{E}] = 0$.

Proof. The desired conclusions follow from Lemmas 16 and 18. \square

One might be able to strengthen the conclusion in Theorem 1(b), say to an almost-sure convergence guarantee; see, e.g., [8]. However, we are satisfied with Theorem 1(b), which is sufficient for revealing the favorable properties of Algorithm 1 under Assumptions 2, 3, and 4. Theorem 1(a) shows under Assumptions 2, 3, and 4 that if the latter condition in (6) holds with $\rho_k = \rho$ for some $\rho \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$ and $\{\beta_k\} = \{\beta\}$ is chosen as a (sufficiently small) constant sequence, then the limit superior of the expectation of the average of quantities connected to stationarity measures for problem (1) is bounded above by a constant proportional to β . Intuitively, this shows that the iterates generated by the algorithm ultimately remain in a region in which these stationarity measures are small. On the other hand, Theorem 1(b) shows under Assumption 4 that if $\{\rho_k\}$ and $\{\beta_k\}$ vanish with $\rho_k = \mathcal{O}(\beta_k^2)$, then a subsequence of iterates exist over which the expected values of these stationarity measures vanish. As seen in Lemma 3, if there exists a subsequence of iterates, say indexed by $\mathcal{S} \subseteq \mathbb{N}$, that converges to a point satisfying certain regularity conditions, then $\{\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2\}_{k \in \mathcal{S}} \rightarrow 0$ means that the limit point is stationary with respect to the problem to minimize $\frac{1}{2} \|c(x)\|_2^2$ subject to $x \in \mathbb{R}_{\geq 0}^n$. Similarly, as seen in Lemma 5, if there exists such a subsequence and the limit point is feasible with respect to problem (1), then $\{d_k^{\text{true}}\}_{k \in \mathcal{S}} \rightarrow 0$ means that the limit point is stationary with respect to (1). These situations are not guaranteed to occur, but this discussion shows that Theorem 1 is meaningful.

4.4 Non-vanishing Merit Parameter

Our main convergence result in the previous section, namely, Theorem 1, requires Assumption 4, which in turn requires that the merit parameter sequence ultimately becomes a sufficiently small, positive constant sequence. (Recall the discussion after Assumption 4.) To show that this corresponds to a realistic event for practical purposes, we next show conditions under which one finds that the merit parameter would not vanish.

We begin by showing a generally applicable result about the solution of (7). It is related to that in Lemma 3, but is stronger due to an additional assumption.

Lemma 19. *Suppose the conditions of Lemma 3 hold and there exists $\kappa_w \in [0, 1)$ such that for all generated $k \in \mathbb{N}$ in any run of the algorithm one has $\|c_k + \nabla c(x_k)^T v_k\|_2 \leq \kappa_w \|c_k\|_2$. Then, there exists $\kappa_v \in \mathbb{R}_{>0}$ such that, in any run of the algorithm such that iteration $k \in \mathbb{N}$ is reached, one finds*

$$\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2 \geq \kappa_v \|v_k\|_2. \quad (33)$$

Proof. Consider an arbitrary run of the algorithm in which the conditions of the lemma hold and iteration $k \in \mathbb{N}$ is reached. If $c_k = 0$, then it follows by construction of (7) that $v_k = 0$, in which case (33) follows trivially. Hence, we may proceed under the assumption that $c_k \neq 0$, which by the conditions of the lemma, Assumption 1 (see (2)), and the triangle inequality gives

$$\kappa_{\nabla c} \|v_k\|_2 \geq \|\nabla c(x_k)^T v_k\|_2 \geq \|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2 \geq (1 - \kappa_w) \|c_k\|_2.$$

Consequently, from (21), (22), and a similar derivation as in Lemma 3, one finds

$$\begin{aligned} 2\|c_k\|_2(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2) &\geq \|c_k\|_2^2 - \|c_k + \nabla c(x_k)^T v_k\|_2^2 \\ &\geq \min\left\{\frac{\lambda^2}{\kappa_{\nabla c}^2}, 2\mu\right\} \|v_k\|_2^2 \geq \min\left\{\frac{\lambda^2}{\kappa_{\nabla c}^2}, 2\mu\right\} \left(\frac{1 - \kappa_w}{\kappa_{\nabla c}}\right) \|c_k\|_2 \|v_k\|_2, \end{aligned}$$

from which the desired conclusion in (33) follows. \square

We now show that, under common conditions and when the norm of the stochastic gradient estimate is bounded uniformly, the denominator of the formula for τ_k^{trial} in (10) is bounded proportionally to $\|v_k\|_2$.

Lemma 20. *Suppose that Assumptions 1 and 3 hold, and that there exists $(\lambda, \mu, \kappa_g) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that for all generated $k \in \mathbb{N}$ in any run of the algorithm one has $\nabla c(x_k)^T \nabla c(x_k) \succeq \lambda I$, $\mu_k \geq \mu$, and $\|g_k\|_2 \leq \kappa_g$. Then, there exists $\kappa_{g,H} \in \mathbb{R}_{>0}$ such that, in any run such that iteration $k \in \mathbb{N}$ is reached, one finds*

$$g_k^T d_k + \frac{1}{2} d_k^T H_k d_k \leq \kappa_{g,H} \|v_k\|_2.$$

Proof. Consider an arbitrary run in which the conditions of the lemma hold and iteration $k \in \mathbb{N}$ is reached. By Lemma 13, $(u, w) = (0, 0)$ is feasible for (7), so

$$\begin{aligned} &\max \left\{ \frac{1}{2} \|c_k + \nabla c(x_k)^T \nabla c(x_k) w_k\|_2^2, \frac{1}{2} \mu_k \|u_k\|_2^2 \right\} \\ &\leq \frac{1}{2} \|c_k + \nabla c(x_k)^T \nabla c(x_k) w_k\|_2^2 + \frac{1}{2} \mu_k \|u_k\|_2^2 \leq \frac{1}{2} \|c_k\|_2^2. \end{aligned}$$

Since $\frac{1}{2} \|c_k + \nabla c(x_k)^T \nabla c(x_k) w_k\|_2^2 \leq \frac{1}{2} \|c_k\|_2^2$, it follows that

$$\|\nabla c(x_k)^T \nabla c(x_k) w_k\|_2^2 \leq -2c_k^T \nabla c(x_k)^T \nabla c(x_k) w_k \leq 2\|c_k\|_2 \|\nabla c(x_k)^T \nabla c(x_k) w_k\|_2,$$

which along with Assumption 1 (see (2)) shows that

$$\|\nabla c(x_k) w_k\|_2 \leq \kappa_{\nabla c} \|w_k\|_2 \leq \frac{\kappa_{\nabla c}}{\lambda} \|\nabla c(x_k)^T \nabla c(x_k) w_k\|_2 \leq 2 \frac{\kappa_{\nabla c}}{\lambda} \|c_k\|_2 \leq 2 \frac{\kappa_{\nabla c}}{\lambda} \kappa_c.$$

On the other hand, since $\frac{1}{2} \mu_k \|u_k\|_2^2 \leq \frac{1}{2} \|c_k\|_2^2$, it follows under Assumption 1 that $\|u_k\|_2 \leq \frac{1}{\sqrt{\mu_k}} \|c_k\|_2 \leq \frac{1}{\sqrt{\mu}} \kappa_c$. Therefore, overall, it follows that

$$\|v_k\|_2 = \sqrt{\|\nabla c(x_k) w_k\|_2^2 + \|u_k\|_2^2} \leq \left(\sqrt{4 \left(\frac{\kappa_{\nabla c}}{\lambda} \right)^2 + \frac{1}{\mu}} \right) \kappa_c.$$

Now, since $v_k = \nabla c(x_k) w_k + u_k$ is a feasible solution of (8) while d_k is the optimal solution of (8), it follows under the conditions of the lemma that

$$\begin{aligned} g_k^T d_k + \frac{1}{2} d_k^T H_k d_k &\leq g_k^T v_k + \frac{1}{2} v_k^T H_k v_k \\ &\leq \kappa_g \|v_k\|_2 + \frac{1}{2} \kappa_H \|v_k\|_2^2 \leq \left(\kappa_g + \frac{1}{2} \kappa_H \left(\sqrt{4 \left(\frac{\kappa_{\nabla c}}{\lambda} \right)^2 + \frac{1}{\mu}} \right) \kappa_c \right) \|v_k\|_2, \end{aligned}$$

which leads to the desired conclusion. \square

We now prove conditions under which the merit parameter does not vanish.

Theorem 2. *Suppose that Assumptions 1 and 3 hold, and that there exists $(\lambda, \mu, \kappa_g, \kappa_w) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times [0, 1)$ such that for all generated $k \in \mathbb{N}$ in any run of the algorithm one has $\nabla c(x_k)^T \nabla c(x_k) \succeq \lambda I$, $\mu_k \geq \mu$, $\|g_k\|_2 \leq \kappa_g$, and $\|c_k + \nabla c(x_k)^T v_k\|_2 \leq \kappa_w \|c_k\|_2$. Then, in any run that does not terminate finitely, the latter event in Lemma 13(k) occurs (i.e., $\{\tau_k\}$ does not vanish) with $\tau_{\min} \geq \frac{(1-\sigma)\kappa_w}{\kappa_{g,H}}(1 - \epsilon_\tau)$.*

Proof. Consider arbitrary $k \in \mathbb{N}$ in a run that does not terminate finitely and note that if $d_k = 0$ or $g_k^T d_k + \frac{1}{2} d_k^T H_k d_k \leq 0$, then $\tau_k^{\text{trial}} \leftarrow \infty$, and otherwise τ_k^{trial} is set by (10). Hence, under the conditions of the lemma and by Lemmas 19–20,

$$\tau_k^{\text{trial}} \geq \frac{(1-\sigma)(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2)}{g_k^T d_k + \frac{1}{2} d_k^T H_k d_k} = \frac{(1-\sigma)(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T v_k\|_2)}{g_k^T d_k + \frac{1}{2} d_k^T H_k d_k} \geq \frac{(1-\sigma)\kappa_w}{\kappa_{g,H}} =: \tau_*.$$

Consequently, by the merit parameter update in (11), $\tau_k < \tau_{k-1}$ only if $\tau_{k-1} > \tau_*$. This, along with Lemma 13(d)–(e), leads to the conclusion. \square

Since ∇f is bounded in norm over the set \mathcal{X} in Assumption 1, Theorem 2 shows that, amongst the other stated conditions, if $\|g_k - \nabla f(x_k)\|_2$ is bounded uniformly over all $k \in \mathbb{N}$ in any, then the merit parameter sequence always remains bounded below by a positive number. Under such conditions, the only potentially poor behavior of the merit parameter sequence is that, in a given run, it ultimately remains constant at a value that is too large. We claim that, under certain assumptions about the distribution of the stochastic gradient estimates, this behavior can be shown to occur with probability zero. (We do not prove such a result here, but refer the interested reader to Proposition 3.16 in [1] to see such a result for the equality-constraints-only setting, in which case the behavior of the merit parameter is similar.) On the other hand, if $\|g_k - \nabla f(x_k)\|_2$ is not bounded uniformly in this manner, then it is possible for the merit parameter sequence to vanish unnecessarily. This issue is one that should be noted by a user of the algorithm. In particular, if in a run of the algorithm one chooses $\mu_k \geq \mu$ for some $\mu \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$ and one finds for some $(\lambda, \kappa_w) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ that generated $k \in \mathbb{N}$ yield $\nabla c(x_k)^T \nabla c(x_k) \succeq \lambda I$ and $\|c_k + \nabla c(x_k)^T v_k\|_2 \leq \kappa_w \|c_k\|_2$, yet τ_k has become exceedingly small, then Theorem 2 shows that this must be due to the stochastic gradient estimates tending to become significantly large in norm, in which case the performance of the algorithm may improve with more accurate stochastic gradient estimates.

4.5 Deterministic Algorithm

We conclude this section with a statement of a convergence result that we claim to hold for Algorithm 1 if it were to be run with $g_k = \nabla f(x_k)$ for all $k \in \mathbb{N}$. Due to space considerations, we do not provide a proof of the result, although we offer the proposition for reference for the reader and claim that it holds from results proved in this paper for the stochastic setting as well as other similar results for SQP methods for deterministic continuous nonlinear optimization.

Proposition 1. *Suppose Assumptions 1 and 3 hold and Algorithm 1 is run with $g_k = \nabla f(x_k)$ for all $k \in \mathbb{N}$. If for all large $k \in \mathbb{N}$ there exists $\kappa_w \in [0, 1)$ such that $\|c_k + \nabla c(x_k)^T v_k\|_2 \leq \kappa_w \|c_k\|_2$, then $\{x_k\} \subset \mathbb{R}_{\geq 0}^n$, $\{\tau_k\}$ is bounded away from zero, and, with $y_k \in \mathbb{R}^m$ and $z_k \in \mathbb{R}_{\geq 0}^n$ defined as the optimal multipliers corresponding to the solution of subproblem (8) for all $k \in \mathbb{N}$, it follows that*

$$\left\{ \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k) y_k - z_k \\ c_k \\ x_k^T z_k \end{bmatrix} \right\| \right\} \rightarrow 0.$$

Otherwise, $\{x_k\} \subset \mathbb{R}_{\geq 0}^n$, $\{\min\{\nabla c(x_k) c_k, 0\}\} \rightarrow 0$, and $\{|x_k^T \nabla c(x_k) c_k|\} \rightarrow 0$, and if the sequence $\{\tau_k\}$ is bounded away from zero, then

$$\left\{ \left\| \begin{bmatrix} \nabla f(x_k) + \nabla c(x_k) y_k - z_k \\ x_k^T z_k \end{bmatrix} \right\| \right\} \rightarrow 0.$$

5 Numerical Results

In this section, we provide results demonstrating the performance of a MATLAB implementation of Algorithm 1 when solving a subset of problems from CUTEst [22], where Gurobi is used to solve the arising subproblems [23]. The purpose of these experiments is to compare this performance against that of the Julia implementation provided by the authors of [31, Algorithm 1]. From all inequality-constrained problems in CUTEst, we selected those such that (i) $m \leq n \leq 1000$, (ii) $f(x_k) \geq -10^{20}$ for all $k \in \mathbb{N}$ in all runs of our algorithm, and (iii) Gurobi did not report any errors. This resulted in a set of 323 test problems.

For each test problem, both codes used the same initial iterate and generated stochastic gradient estimates in the same manner. Specifically, for all $k \in \mathbb{N}$ in each run, the codes set $g_k = \mathcal{N}(\nabla f(x_k), \epsilon_g(I + ee^T))$, where e is the all-ones vector and $\epsilon_g \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$ was fixed for each run (see below). If a problem had only inequality constraints, i.e., $m = 0$, then our code explicitly computed α_k^φ (as defined in (15)) and set $\alpha_k \leftarrow \alpha_k^{\max}$ for all $k \in \mathbb{N}$. Otherwise, the code set $\alpha_k \leftarrow \min\{1, (1.1)^{t_k} \alpha_k^{\min}, \alpha_k^{\min} + \theta \beta_k\}$, where $t_k \leftarrow \max\{t \in \mathbb{N} : \varphi_k((1.1)^t \alpha_k^{\min}) \leq 0\}$. This guarantees that $\alpha_k \in [\alpha_k^{\min}, \alpha_k^{\max}]$ for all $k \in \mathbb{N}$. The other user-defined parameters of Algorithm 1 were selected as $\sigma = \tau_0 = 0.1$, $\eta = 0.5$, $\xi_0 = 1$, $\epsilon_\tau = \epsilon_\xi = 10^{-2}$, $\theta = 10^4$, $\mu_k = \max\{10^{-8}, 10^{-4} \|c_k\|_2^2\}$, $\beta_k = 1$, and $H_k = I$ for all $k \in \mathbb{N}$. The Lipschitz constants L and Γ were estimated every 100 iterations by differences of stochastic gradients at ten samples around the current iterate. Meanwhile, we ran the Julia code for [31, Algorithm 1] with the **AdapGD** option and its default parameter settings as described in [31, Section 4]. Each code terminated as soon as 10^4 stochastic gradient samples were evaluated or a 12-hour CPU time limit was reached.

Let $\text{FeasErr}(x)$ be the ∞ -norm constraint violation at x and let $\text{KKTErr}(x, y, z)$ be the ∞ -norm violation of the KKT conditions (recall (4)) at a primal-dual iterate (x, y, z) . Each run of Algorithm 1 generates $\{x_k\} \subset \mathbb{R}^n$. For each $k \in \mathbb{N}$, let $y_k^{\text{true}} \in \mathbb{R}^m$ and $z_k^{\text{true}} \in \mathbb{R}^n$ denote the optimal Lagrange multipliers corresponding to the equality and inequality constraints when (8) is solved with $g_k = \nabla f(x_k)$. For each run of Algorithm 1, we determined the best iterate as $x_{k_{\text{best}}}$ where

$$k_{\text{best}} = \begin{cases} \arg \min_{k \in \mathbb{N}} \text{FeasErr}(x_k) & \text{if } \text{FeasErr}(x_k) > 10^{-4} \text{ for all } k \in \mathbb{N}, \\ \arg \min_{k \in \mathbb{N}} \{\text{KKTErr}(x_k, y_k^{\text{true}}, z_k^{\text{true}}) : \text{FeasErr}(x_k) \leq 10^{-4}\} & \text{otherwise.} \end{cases}$$

We determined the best iterate in a run of [31, Algorithm 1] using the same formula with the sequence of iterates and Lagrange multiplier estimates that are computed as part of the algorithm. Our results for four noise levels, provided in Figure 1 below, are presented in terms of $\text{FeasErr}(x_{k_{\text{best}}})$ as the feasibility error and $\text{KKTErr}(x_{k_{\text{best}}}, y_{k_{\text{best}}}^{\text{true}}, z_{k_{\text{best}}}^{\text{true}})$ as the KKT error for each run of both algorithms.

Since the Julia code for [31, Algorithm 1] is only set up to solve CUTEst problems without simple bound constraints, the results in Figure 1 are presented in two parts. For the 57 problems for which both algorithms were set up to run, the first two box plots show the best feasibility and KKT errors achieved by both codes, where each problem is run 5 times each (since the behaviors of the algorithms are stochastic). In the third box plot, we report the best feasibility and KKT errors obtained by our Matlab code on the remaining 266 ($= 323 - 57$) problems, again with five runs for each problem. Overall, one finds that the performance of our algorithm is comparatively good in this experimental set-up. The best feasibility and KKT errors are relatively low for our algorithm, although the errors increase with the noise level, as may be expected. Experiments with diminishing step sizes also showed favorable performance for our algorithm; these results are omitted due to page limit restrictions.

6 Conclusion

We have proposed, analyzed, and tested an algorithm for solving continuous optimization problems. The algorithm requires that constraint function and derivative values can be computed in each iteration, but does not require exact objective function and derivative values; rather, the algorithm merely requires that a stochastic objective gradient estimate is computed to satisfy relatively loose assumptions in each iteration. The theoretical convergence guarantees of the algorithm require knowledge of Lipschitz constants for the

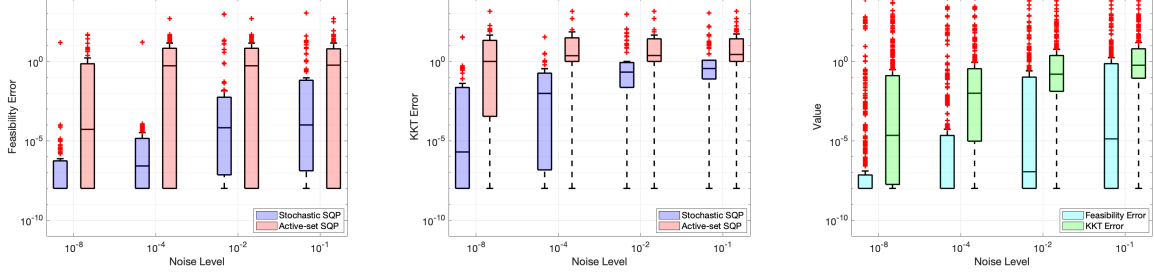


Figure 1: Box plots comparing the best feasibility errors (left) and KKT errors (middle) of a Matlab implementation of Algorithm 1 (“Stochastic SQP”) and the Julia implementation provided by the authors of [31, Algorithm 1] (“Active-st SQP”) when solving 57 CUTEst problems without simple bound constraints. Box plots of the best feasibility and KKT errors (combined, right) of the implementation of Algorithm 1 when solving the other 266 CUTEst problems from the test set.

objective gradient and constraint Jacobian, although in practice these constants can be estimated. Our numerical experiments show that our proposed algorithm can outperform an alternative algorithm that relies on the ability to compute more accurate gradient estimates. We have provided comments throughout the paper on how the assumptions that are required for our theoretical convergence guarantees might be loosened further.

Acknowledgements

The authors are grateful to Sen Na for providing consultation about the Julia implementation provided by the authors of [31, Algorithm 1]. This material is based upon work supported by the U.S. NSF under award CCF-2139735 and by the Office of Naval Research under award N00014-21-1-2532.

References

- [1] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.
- [2] Albert S Berahas, Raghu Bollapragada, and Baoyu Zhou. An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization. *arXiv preprint arXiv:2206.00712*, 2022.
- [3] 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.
- [4] Albert S Berahas, Jiahao Shi, Zihong Yi, and Baoyu Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *arXiv preprint arXiv:2204.04161*, 2022.
- [5] Albert S Berahas, Miaolan Xie, and Baoyu Zhou. A sequential quadratic programming method with high probability complexity bounds for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2301.00477*, 2023.
- [6] Dimitri Bertsekas. *Convex Optimization Theory*, volume 1. Athena Scientific, 2009.
- [7] Dimitri P. Bertsekas. *Network optimization: continuous and discrete models*, volume 8. Athena Scientific, 1998.

- [8] Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.
- [9] Richard H Byrd, Jean Charles Gilbert, and Jorge Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Math. Prog.*, 89(1):149–185, 2000.
- [10] Richard H Byrd, Mary E Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.
- [11] Andrew R Conn. Constrained optimization using a nondifferentiable penalty function. *SIAM Journal on Numerical Analysis*, 10(4):760–784, 1973.
- [12] F. E. Curtis, D. P. Robinson, and B. Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*, 2021.
- [13] Frank E Curtis, Michael J O’Neill, and Daniel P Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2112.14799*, 2021.
- [14] G Di Pillo and L Grippo. Exact penalty functions in constrained optimization. *SIAM Journal on control and optimization*, 27(6):1333–1360, 1989.
- [15] Gianni Di Pillo and Luigi Grippo. A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints. *SIAM Journal on Control and Optimization*, 23(1):72–84, 1985.
- [16] II Dikin. Iterative solution of problems of linear and quadratic programming. In *Doklady Akademii Nauk*, volume 174, pages 747–748. Russian Academy of Sciences, 1967.
- [17] Yuchen Fang, Sen Na, Michael W. Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *arXiv preprint 2211.15943*, 2022.
- [18] Roger Fletcher. An exact penalty function for nonlinear programming with inequalities. *Mathematical Programming*, 5(1):129–150, 1973.
- [19] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.
- [20] Philip E Gill, Walter Murray, and Michael A Saunders. Snopt: An SQP algorithm for large-scale constrained optimization. *SIAM review*, 47(1):99–131, 2005.
- [21] Torkel Glad and Elijah Polak. A multiplier method with automatic limitation of penalty growth. *Mathematical Programming*, 17(1):140–155, 1979.
- [22] Nicholas IM 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(3):545–557, 2015.
- [23] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.
- [24] Shih-Ping Han. Superlinearly convergent variable metric algorithms for general nonlinear programming problems. *Mathematical Programming*, 11(1):263–282, 1976.
- [25] Richard J Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.
- [26] Toshihide Ibaraki and Naoki Katoh. *Resource allocation problems: algorithmic approaches*. MIT press, 1988.

- [27] Drew P Kouri and Thomas M Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM Journal on Optimization*, 26(1):365–396, 2016.
- [28] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.
- [29] L Lasdon, A Waren, and R Rice. An interior penalty method for inequality constrained optimal control problems. *IEEE Transactions on Automatic Control*, 12(4):388–395, 1967.
- [30] Robert McGill. Optimum control, inequality state constraints, and the generalized newton-raphson algorithm. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 3(2):291–298, 1965.
- [31] Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *arXiv preprint arXiv:2109.11502*, 2021.
- [32] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, pages 1–71, 2022.
- [33] Sen Na and Michael W Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv preprint arXiv:2205.13687*, 2022.
- [34] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, 2006.
- [35] Figen Oztoprak, Richard Byrd, and Jorge Nocedal. Constrained optimization in the presence of noise. *arXiv preprint arXiv:2110.04355*, 2021.
- [36] Vivak Patel and Shushu Zhang. Stochastic gradient descent on nonconvex functions with general noise models. *arXiv preprint 2104.00423*, 2021.
- [37] Andre F Perold. Large-scale portfolio optimization. *Management science*, 30(10):1143–1160, 1984.
- [38] Songqiang Qiu and Vyacheslav Kungurtsev. A sequential quadratic programming method for optimization with stochastic objective functions, deterministic inequality constraints and robust subproblems. *arXiv preprint arXiv:2302.07947*, 2023.
- [39] Daniel P. Robinson. *Primal-dual methods for nonlinear optimization (thesis)*. University of California, San Diego, 2007.
- [40] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.
- [41] Qiankun Shi, Xiao Wang, and Hao Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. 2022.
- [42] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
- [43] Stephen J Wright. *Primal-dual interior-point methods*. SIAM, 1997.
- [44] Hiroshi Yamashita. A globally convergent primal-dual interior point method for constrained optimization. *Optimization Methods and Software*, 10(2):443–469, 1998.
- [45] Willard I Zangwill. Non-linear programming via penalty functions. *Management science*, 13(5):344–358, 1967.
- [46] Victor M. Zavala and Mihai Anitescu. Scalable nonlinear programming via exact differentiable penalty functions and trust-region newton methods. *SIAM Journal on Optimization*, 24(1):528–558, 2014.