An Adaptive Sampling Sequential Quadratic Programming Method for Equality Constrained Stochastic Optimization

Albert S. Berahas*† Raghu Bollapragada‡ Baoyu Zhou§

June 30, 2022

Abstract

This paper presents a methodology for using varying sample sizes in sequential quadratic programming (SQP) methods for solving equality constrained stochastic optimization problems. The first part of the paper deals with the delicate issue of dynamic sample selection in the evaluation of the gradient in conjunction with inexact solutions to the SQP subproblems. Under reasonable assumptions on the quality of the gradient approximations employed and the accuracy of the solutions to the SQP subproblems, we establish global convergence results for the proposed method. Motivated by these results, the second part of the paper describes a practical adaptive inexact stochastic sequential quadratic programming (PAIS-SQP) method. We propose criteria for controlling the sample size and the accuracy in the solutions of the SQP subproblems based on the variance estimates obtained as the optimization progresses. Finally, we demonstrate the practical performance of the method on a subset of the CUTE problems and constrained classification tasks.

1 Introduction

We consider stochastic optimization problems with deterministic equality constraints, and design, analyze and implement a stochastic adaptive sampling algorithm based on the sequential quadratic programming (SQP) paradigm. Optimization problems of this form arise in a plethora of real-world applications, including but not limited to computer vision [32], optimal control [8], network optimization [7], partial differential equation optimization [33], deep learning [36], and reinforcement learning [1].

The majority of the methods developed for solving constrained stochastic optimization problems, i.e., optimization problems with deterministic constraints and stochastic objective functions, are based on the penalty method approach [19, 26, 29, 32]. Such methods

*Corresponding author.
†Dept. of Industrial and Operations Engineering, University of Michigan. (albertberahas@gmail.com)
‡Operations Research and Industrial Engineering Prog., UT Austin. (raghu.bollapragada@utexas.edu)
§Dept. of Industrial and Systems Engineering, Lehigh University. (baoyu.zhou@lehigh.edu)

transform the given constrained problem into an unconstrained problem by penalizing constraint violation in the objective function, and then apply classical unconstrained stochastic optimization methods on the modified objective. While these methods are well studied, recently a number of algorithms endowed with sound theoretical guarantees and superior empirical performance have been proposed [5, 28]. In [5], a stochastic SQP method with adaptive step size selection for the fully stochastic regime is proposed, and, in [28], a stochastic line search SQP method with adaptive gradient accuracy is proposed. Several other extensions of these methods have been developed, e.g., relaxing constraint assumptions [4], employing inexact computations [22] and employing variance reduction [6].

Adaptive sampling is a powerful technique used in stochastic optimization to control the variance in the approximations employed as the optimization progresses. The idea is simple yet powerful; far from the solution(s) inaccurate and cheap (gradient) information can be employed while near the solution(s) accurate (gradient) information is required for both theory and practice. Of course, the key to such methods is the mechanism by which the sample size (or accuracy of the approximation) is selected. In [23, 12], algorithms that increase the samples sizes employed with prescribed (geometric) rules are proposed, and in [12] the authors showed that these methods achieve optimal worst-case first-order complexity for unconstrained problems. Other algorithms utilize gradient approximation tests to control the accuracy in the approximations; e.g., norm test [15, 12], inner product test [9, 10], and others [16, 3, 25]. Adaptive sampling algorithms have also been developed for simulation-based optimization problems [24, 31]. Finally, [2, 35] apply adaptive sampling methods to constrained stochastic optimization problems with convex feasible sets.

1.1 Contributions

In this paper, motivated by the successes of adaptive sampling methods (unconstrained stochastic optimization problems) and SQP methods (constrained deterministic optimization problems), we propose an adaptive sampling stochastic sequential quadratic programming algorithm for solving optimization problems with deterministic constraints and stochastic objective functions. This is far from a trivial task as complication arises when incorporating existing adaptive sampling strategies into the SQP paradigm primarily due to the multi-objective nature of constrained optimization. We propose a novel mechanism to control the accuracy in gradient approximations employed and the search directions calculated that balances the goals of achieving feasibility and reducing the objective function value. The main ingredients of our proposed method are: (1) the extension of the norm condition [15, 12] for equality constrained stochastic optimization problems, and (2) the adaptation of the stochastic SQP method proposed in [5].

We prove convergence guarantees for two different inexactness conditions, predetermined and adaptive. Specifically, under standard stochastic condition on the accuracy of the gradient approximations and deterministic conditions on the quality of the solutions of the linear system, we prove that a measure of first-order stationarity evaluated at the

iterates generated by our proposed algorithm converges to zero in expectation from arbitrary starting points. We also established worst-case sampling complexity results when the sample sizes are controlled using predetermined rates instead of adaptive tests showing that even though sampling work per iteration is increasing, the overall sampling complexity is still $O(\ ^{2(1+\ )})$, where $\ > 1$. Moreover, we then propose a practical variant of our algorithm, inspired by [12, 9]. Finally, we present numerical results on binary classi cation tasks with equality constraints and on equality constrained problems from the CUTE collection that demonstrate the e ciency and e cacy of our proposed practical method.

1.2 Notation

The set of natural numbers is denoted by $\mathbb{N} := \{0, 1, 2, \ldots\}$. The set of real numbers (i.e., scalars) is denoted by $\mathbb{R}$, and $\mathbb{R}_{>r}$ ($\mathbb{R}_{\ r}$) denotes the set of real numbers greater than (greater than or equal to) $r \in \mathbb{R}$. The set of n-dimensional vectors is denoted by $\mathbb{R}^n$, the set of m-by-n matrices is denoted by $\mathbb{R}^{m\ \ n}$, and the set of n-by-n symmetric matrices is denoted by $\mathbb{S}^n$. Our proposed algorithms are iterative, and generate a sequence of iterates $\{x_k\}$ with $x_k \in \mathbb{R}^n$. Let $f_k := f(x_k)$, $g_k := \nabla f(x_k)$, $c_k := c(x_k)$, and $J_k := \nabla c(x_k)^T$ for all $k \in \mathbb{N}$.

1.3 Organization

The paper is organized as follows. In Section 2 we formalize the problem statement and main assumptions. The general algorithmic framework of our proposed method is presented in Section 3. Convergence and complexity guarantees are stated and proven in Section 4. We present a practical adaptive sampling SQP method in Section 5. In Section 6, we demonstrate the empirical performance of of the proposed method on classi cation tasks. Finally, in Section 7, we make some nal remarks and discuss avenues for future research.

2 Problem Statement and Assumptions

We consider the following potentially nonlinear and/or nonconvex equality constrained optimization problem

$$\min_{x \in \mathbb{R}^n} \ f(x) \quad \text{s.t.:} \quad c(x) = 0\ ; \quad \text{with} \quad f(x) = \mathbb{E}[F(x;\ )]; \tag{2.1}$$

where the objective function $f : \mathbb{R}^n \ \to \ \mathbb{R}$ and constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth, $\ $ is a random variable with associated probability space $(\ ; \mathcal{F}; \mathbb{P})$, $F : \mathbb{R}^n \ \to \ \mathbb{R}$, and $\mathbb{E}[\ ]$ denotes the expectation taken with respect to $\mathbb{P}$. Throughout the paper, we assume that the constraint function and its associated derivatives can be computed exactly. With regards to the objective function and its associated derivatives, we assume that these quantities are expensive to compute, but that accurate evaluations can be obtained as required. We formalize the notions of accuracy in subsequent sections of the paper.

We make the following main assumption with regards to (2.1) and the iterates $\{x_k\}$ generated by our proposed algorithms.

Assumption 2.1. Let $X \subseteq \mathbb{R}^n$ be an open convex set containing the sequence $\{x_k\}$ generated by any run of the algorithm. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below over X and its gradient function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $L \in \mathbb{R}_{>0}$ (with respect to $\|\cdot\|_2$) and bounded over X. The constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ (with $m \leq n$) is continuously differentiable and bounded over X and its Jacobian function $J := \nabla c^T : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ is Lipschitz continuous with constant $\in \mathbb{R}_{>0}$ (with respect to $\|\cdot\|_2$) and bounded over X. In addition, for all $x \in X$, the Jacobian $J(x)$ has singular values that are bounded uniformly below by $\in \mathbb{R}_{>0}$.

Remark 2.2. We make the following remarks about Assumption 2.1. Under Assumption 2.1, there exist constants $(f_{\inf}, \ _g, \ _c, \ \ , \ _J) \in \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$,

$$f_{\inf} \leq f_k, \ \|\nabla f(x_k)\|_2 \leq \ _g, \ \|c_k\|_1 \leq \ _c, \ \|J_k\|_2 \leq \ _J, \text{ and } \|(J_k J_k^T)^{-1}\|_2 \leq \ ^2.$$

The components pertaining to the objective and constraint functions are standard assumptions in the equality constrained optimization literature. With regards to the algorithmic components, we do not assume that the iterate sequence itself is bounded, however, we do assume that the objective function and constraints, function values and derivatives, are bounded over the set X containing the iterates. While this assumption is reasonable in the deterministic setting, it is not ideal in the stochastic setting. That being said, it is a common assumption in the equality constrained stochastic optimization literature [5, 28]. The justification is that in this constrained setting, iterates are presumably converging towards a determinsitic feasible region. Furthermore, we assume that the accuracy in the gradient approximations can be controlled, and as such claim that the assumption above is reasonable.

Let $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be the Lagrangian function corresponding to (2.1), $\ell(x, y) = f(x) + c(x)^T y$, where $y \in \mathbb{R}^m$ is the vector of Lagrange multipliers. Under Assumption 2.1, the necessary conditions for first-order stationarity for (2.1) are

$$0 = \begin{bmatrix} \nabla_x \ell(x, y) \\ \nabla_y \ell(x, y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + J(x)^T y \\ c(x) \end{bmatrix}.$$

3 General Algorithmic Framework

Our proposed algorithms can be characterized as adaptive (selection of the step size/accuracy of approximations), inexact (linear system solves), and stochastic (gradient approximation employed) sequential quadratic optimization (SQP) methods. Specifically, given x_k for all

$k \in \mathbb{N}$, a search direction $\bar{d}_k \in \mathbb{R}^n$ is computed by inexactly solving a quadratic optimization subproblem based on a local quadratic model of the objective function constructed using

$$\min_{d \in \mathbb{R}^n} \; f_k + \bar{g}_k^T d + \tfrac{1}{2} d^T H_k d \quad \text{s.t.:} \quad c_k + J_k d = 0; \tag{3.1}$$

where the matrix $H_k \in \mathbb{R}^{n \times n}$ satisfies the assumption below (Assumption 3.1).

Assumption 3.1. The sequence of symmetric matrices $\{H_k\} \subset \mathbb{S}^n$ is bounded in norm by $\beta_H \in \mathbb{R}_{>0}$ such that $\|H_k\|_2 \leq \beta_H$ for all $k \in \mathbb{N}$. In addition, there exists a constant $\zeta \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the matrix H_k has the property that $u^T H_k u \geq \zeta \|u\|_2^2$ for all $u \in \mathbb{R}^n$ such that $J_k u = 0$.

Under Assumptions 2.1 and 3.1, the optimal solution $\bar{d}_k \in \mathbb{R}^n$ of the subproblem (3.1), and an associated displacement in the Lagrange multiplier $\bar{e}_k \in \mathbb{R}^m$, can be obtained by solving the linear system of equations given by

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{e}_k \end{bmatrix} = \begin{bmatrix} -\bar{g}_k + J_k^T y_k \\ -c_k \end{bmatrix}. \tag{3.2}$$

Solving such linear systems can be expensive, and as such, our algorithms employ inexact solutions to the above linear systems given by $(\bar{d}_k, \bar{e}_k) \leftarrow (d_k, e_k)$, i.e.,

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ e_k \end{bmatrix} = \begin{bmatrix} -g_k + J_k^T y_k \\ -c_k \end{bmatrix} + \begin{bmatrix} \delta_k \\ r_k \end{bmatrix}; \tag{3.3}$$

where the tuple (δ_k, r_k) defines the residuals, and $\|(\delta_k, r_k)\|$ can be controlled as required. Our algorithms impose conditions on the norm of the residuals.

Given a pair $(\bar{d}_k, \bar{e}_k)$, our algorithms proceed to compute a positive step size in order to update the primal variables x_k and Lagrange multipliers y_k. The step size selection strategy is similar to that in [5]. To this end, the algorithms employ a merit function $\phi : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$, parameterized by a merit parameter $\tau_k \in \mathbb{R}_{>0}$, defined as

$$\phi(x_k; \tau_k) = \tau_k f_k + \|c_k\|_1. \tag{3.4}$$

The merit parameter is dynamically adjusted as the optimization progresses. We employ an local model $l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ of the merit function defined by

$$l(x_k; \tau_k; \bar{g}_k; \bar{d}_k) = \tau_k (f_k + \bar{g}_k^T \bar{d}_k) + \|c_k + J_k \bar{d}_k\|_1 = \tau_k (f_k + \bar{g}_k^T \bar{d}_k) + \|r_k\|_1;$$

and its associated reduction function $\Delta l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\begin{aligned} \Delta l(x_k; \tau_k; \bar{g}_k; \bar{d}_k) &= l(x_k; \tau_k; \bar{g}_k; 0) - l(x_k; \tau_k; \bar{g}_k; \bar{d}_k) \\ &= -\tau_k \bar{g}_k^T \bar{d}_k + \|c_k\|_1 - \|c_k + J_k \bar{d}_k\|_1 \\ &= -\tau_k \bar{g}_k^T \bar{d}_k + \|c_k\|_1 - \|r_k\|_1; \end{aligned} \tag{3.5}$$

to guide the selection of the merit parameter. Note, $l(x_k; \lambda_k; \bar{g}_k; \bar{d}_k) = \lambda_k \bar{g}_k^T \bar{d}_k + \|c_k\|_1$, where $(\bar{d}_k; \bar{e}_k)$ is the solution of (3.2). The mechanism for updating the merit parameter λ_k is similar to that proposed in [5], and is motivated by state-of-the-art SQP methods [20, 14]. First, a trial merit parameter λ_k^{trial} is computed. Given user-defined parameters $(\eta_1; \eta_2; \eta_b) \in (0;1) \times (0;1) \times \mathbb{R}_{>0}$, if $\|r_k\|_1 \leq (1-\eta_1)\eta_2\|c_k\|_1$ and/or $\|\bar{r}_k\|_1 \leq \eta_b\|c_k\|_1$, we set $\lambda_k^{trial} = 1$. Otherwise, we set

$$\lambda_k^{trial} = \begin{cases} 1 & \text{if } \bar{g}_k^T \bar{d}_k + \max\{\tfrac{1}{2}\bar{d}_k^T H_k \bar{d}_k, \delta_d\|\bar{d}_k\|_2^2\} \leq 0 \\ \dfrac{(1-\eta_1)(1-\eta_2)\|c_k\|_1}{\bar{g}_k^T \bar{d}_k + \max\{\tfrac{1}{2}\bar{d}_k^T H_k \bar{d}_k, \delta_d\|\bar{d}_k\|_2^2\}} & \text{otherwise,} \end{cases} \tag{3.6}$$

where $\delta_d \in (0; \infty)$ is user-defined. Then, the merit parameter value λ_k is updated via

$$\lambda_k = \begin{cases} \lambda_{k-1} & \text{if } \lambda_{k-1} \in [(1-\tau)\lambda_k^{trial}, \lambda_k^{trial}] \\ (1-\tau)\lambda_k^{trial} & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\tau \in (0;1)$ is user-defined. This rule ensures that $\{\lambda_k\}$ is a monotonically non-increasing positive sequence with $\lambda_k \geq (1-\tau)\lambda_k^{trial}$ for all $k \in \mathbb{N}$. Moreover, this rule aims to ensure that the reduction function (3.5) is non-negative and satisfies,

$$l(x_k; \lambda_k; g_k; d_k) \leq \lambda_k\{-\eta_1 \max\{\tfrac{1}{2}d_k^T H_k d_k, \delta_d\|d_k\|_2^2\} + \eta_1 \max\{\|c_k\|_1, \|r_k\|_1\} - \|c_k\|_1\}. \tag{3.8}$$

We show (see Lemma 3.6 below) that this bound indeed holds for the iterates generated by our proposed algorithm.

Finally, the step size $\alpha_k \in \mathbb{R}_{>0}$ is set as follows. At the kth iteration, we set via

$$\alpha_k = \min\left\{\frac{2(1-\eta_2)(\theta-1)\, l(x_k; \lambda_k; g_k; d_k)}{(\lambda_k L_k + \gamma_k)\|d_k\|_2^2}; \alpha_k^{opt}; \alpha_u(2-\theta); 1\right\}; \tag{3.9}$$

where $\theta \in (0;1)$, $\eta_2 \in (0;1]$, $\gamma_2 \in [2;4]$ and $\alpha_u \in \mathbb{R}_{>0}$ are user-defined parameters, and

$$\alpha_k^{opt} = \max\left\{\min\left\{\frac{l(x_k; \lambda_k; g_k; d_k)}{(\lambda_k L_k + \gamma_k)\|d_k\|_2^2}; 1\right\}; \frac{l(x_k; \lambda_k; g_k; d_k) - 2\|c_k\|_1}{(\lambda_k L_k + \gamma_k)\|d_k\|_2^2}\right\}. \tag{3.10}$$

The step-size choice is motivated by that proposed in [5], and adapted for our specific setting. The user-defined parameters θ and η_2 that influence the step size are also related to the accuracy of gradient approximations employed and the accuracy of the solutions of the linear system (3.2), which are introduced and discussed in Section 4.

Our algorithmic framework, Adaptive, Inexact, Stochastic SQP (AIS-SQP), is presented in Algorithm 1.

Remark 3.2. We make the following remarks about Algorithm 1.

- Lines 2 and 3: For ease of exposition, we leave these two steps arbitrary and specify them later in the paper (Section 4). We assume that gradient approximations of arbitrary accuracy can be computed, and that the linear system can be solved iteratively and to arbitrary accuracy. In Section 5, we present a practical adaptive sampling strategy and a mechanism for solving the linear system inexactly.

Algorithm 1 (AIS-SQP) Adaptive, Inexact, Stochastic SQP Algorithm

Require: $x_0 \in \mathbb{R}^n$; $y_0 \in \mathbb{R}^m$; $\{H_k\} \subset \mathbb{S}^n$; $\beta_1 \in \mathbb{R}_{>0}$; $\{\beta_{1,1}, \beta_{1,2}, \gamma\} \subset (0,1)$; $\beta_a \in \mathbb{R}_{>0}$; $\beta_b \in \mathbb{R}_{>0}$; $\kappa \in (0,1]$; $\beta_u \in \mathbb{R}_{>0}$; $\beta_d \in (0, \infty)$; $\psi \in [2,4]$

1: **for all** $k \in \mathbb{N}$ **do**
2: Compute some gradient approximation $g_k \in \mathbb{R}^n$
3: Solve (3.2) iteratively; compute a step (d_k, δ_k) that satisfies (a) or (b):
(a) (3.8) and $\|r_k\|_1 \leq \beta_a \psi^{-1}(\Delta_k, \delta_k, g_k, d_k)$, with $\Delta_k = \beta_{1,1}\Delta_{k-1}$, and, additionally, $g_k^T d_k + \max\{\tfrac{1}{2} d_k^T H_k d_k, \beta_d \|d_k\|_2^2\} \leq 0$ if $\|c_k\|_1 > 0$,
(b) $\|r_k\|_1 \leq \min\{(1-\beta_{1,1})\beta_{1,2}, \beta_{1,1}\beta_a\} \psi^{-1}(\|c_k\|_1)$ and $\|\delta_k\|_1 \leq \beta_b \|c_k\|_1$
4: Update Δ_k via (3.6)–(3.7) only if (b) is satisfied
5: Compute a step size α_k via (3.9)–(3.10)
6: Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$, and $y_{k+1} \leftarrow y_k + \alpha_k \delta_k$
7: **end for**

- Lines 3(a) and 3(b): Lemma 3.5 (below) shows that the algorithm is well defined, and that Line 3 will terminate in either (a) or (b). One could enforce a simpler condition on the solution to the linear system, however, this simpler condition would come at the cost of having to solve the linear system exactly if/when the iterates are feasible. In the special case where $\|c_k\|_1 = 0$, by Assumptions 3.1 and 3.3 (presented below; pertaining to the linear system solutions) and (3.3), Line 3 of Algorithm 1 is guaranteed to terminate in case (a). The additional condition in case (a) is added for technical reasons discussed in Section 4. Intuitively, if the additional condition holds, then sufficient progress can be made in terms of minimizing the objective function, and thus progress can be made without modifying the merit parameter.

- Merit parameter update: If Line 3 terminates due to the condition (a), then Line 4 is skipped. If Line 4 is triggered, then the merit parameter is potentially updated to ensure (3.8) is satisfied.

- Step size selection: The step size selection strategy (3.9)–(3.10) depends on Lipschtiz constants (L and M), or estimates of these quantities. If one knows the Lipschitz constants, one could simply set $L_k = L$ and $M_k = M$ for all $k \in \mathbb{N}$. If such Lipschitz constants are unknown, as is the case more often than not, one can approximate these constants following the approaches proposed in [5, 23, 9]. See Section 5 for details of our practical implementation.

- Comparison to other algorithms: The step computation, merit parameter update, and step size selection mechanisms are similar to those proposed in [5, 22, 4]. However, there are some key differences, primarily due to the fact that in this work we assume that the accuracy in the gradient approximations can be controlled as the optimization progresses. Similar to [22], the linear system is solved inexactly, but with a simpler approach that

does not require an explicit step decomposition.

Before we proceed, we state and prove a few results that hold throughout the paper. For the purposes of the presentation of the analysis in subsequent sections, we introduce the following notation. (Note, these quantities are never explicitly computed in our algorithms.) Given the iterate x_k and multiplier y_k, let the tuple $(d_k; \ _k)$ denote the solution to

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \ _k \end{bmatrix} = \begin{bmatrix} g_k + J_k^T y_k \\ c_k \end{bmatrix}; \tag{3.11}$$

the deterministic counter-part of (3.2), where $\bar{g}_k$ is replaced by the true gradient of the objective function. Moreover, let $f \ _k g$ and $f \ _k g$ be the sequences of merit parameters and step sizes, respectively, computed at x_k for all $k \ 2 \ N$ by the deterministic variant of the algorithm with $\ _{k \ 1} = \ _{k \ 1}$. Finally, we introduce $\ _{min} \ 2 \ R_{> 0}$, a lower bound on the deterministic merit parameter value, whose existence is guaranteed by Assumptions 2.1 and 3.1; see [5, 13]. We make the following additional assumption.

Assumption 3.3. There exist constants $f \ _r; \ g \ R_{> 0}$ such that $kr_k k_1 \ _r$ and $k \ _k k_1$ for all $k \ 2 \ N$. Moreover, for any $k \ 2 \ N$, a sequence of inexact solutions $f(d_{k;t}; \ _{k;t})g_{t \ 2 \ N}$ is generated by some iterative linear system solver (t denotes the iteration counter of the linear system solver), where $\lim_{t \ ! \ 1} \ f(d_{k;t}; \ _{k;t})g = (\tilde{d}_k; \tilde{\ }_k)$ and $(d_k; \ _k) := (\tilde{d}_{k;t}; \ _{k;t})$ for some $t \ 2 \ N$. Furthermore, for technical reasons, we also assume that either $kc_k k_1 \ 6 = 0$ or $g_k \ 2 \ 6 = \text{Range}(J_k^T)$ for all $k \ 2 \ N$.

Remark 3.4. We make the following remarks about Assumption 3.3. Assumption 3.3 pertains to properties of three main components: (i) the iterative solver, (ii) acceptable inexact solutions, and, (iii) the stochastic gradient estimates. First, we require that the iterative solver is able to return the exact solution of (3.2) in the limit. Second, we assume that all acceptable inexact solutions $(d_k; \ _k)$ have norms of residuals ($kr_k k_1$ and $k \ _k k_1$) that are reasonably small and bounded uniformly. Third, we assume that the stochastic gradients $(\bar{g}_k)$ computed do not lie exactly in the range space of the Jacobian of the constraints (J_k^T) for iterates that are feasible. In general, this is not a strong assumption in the stochastic setting. For details about practical linear system solvers see [34] and references therein, and, for details about our implemented linear system solver see Sections 5.2 and 6.

The first result shows that Algorithm 1 is well-defined.

Lemma 3.5. Line 3 of Algorithm 1 terminates finitely.

Proof. We consider two cases: (i) $kc_k k_1 > 0$ and (ii) $kc_k k_1 = 0$. If $kc_k k_1 > 0$, by Assumption 3.3 it follows that $f \max fk r_{k;t} k_1; k \ _{k;t} k_2 gg \ ! \ 0$. Therefore, for sufficiently large $t \ 2 \ N$, $kr_{k;t} k_1 < \min f(1 \ ! \ _1)! \ _2; ! \ _1 ! \ _a \ ^{=2} gkc_k k_1$ and $k \ _{k;t} k_1 < ! \ _b kc_k k_1$ are satisfied, and Line 3 of Algorithm 1 terminates finitely in case (b).

On the other hand, if $\|c_k\|_1 = 0$, by (3.2) and (3.5), it follows that

$$l(x_k; \ _{k\ 1}; g_k; \bar{d}_k) = \ _{k\ 1} g_k^T \bar{d}_k + \|c_k\|_1 \ \ k\ c_k + J_k \bar{d}_k\|_1 = \ _{k\ 1} g_k^T \bar{d}_k:$$

By (3.2) and Assumption 3.3 it follows that $g_k \in \text{Range}(J_k^T)$ and $\|\bar{d}_k\|_2 > 0$. By Assumption 3.1, (3.2), $\|\bar{d}_k\|_2 > 0$ and $_d \in (0; \infty)$, it follows that $\bar{d}_k^T H_k \bar{d}_k > \ _d \|\bar{d}_k\|_2^2 > 0$. Moreover, by (3.2), $g_k^T \bar{d}_k + \bar{d}_k^T H_k \bar{d}_k = 0$. Combining the above and $\{\ _k\} \subset R_{>0}$,

$$\begin{aligned} & l(x_k; \ _{k\ 1}; g_k; \bar{d}_k) \ \ \ _{k\ 1} + \ _1 \max\{-\bar{d}_k^T H_k \bar{d}_k; \ _d \|\bar{d}_k\|_2^2\} \\ = & \ \ \ _{k\ 1} g_k^T \bar{d}_k \ \ \ _{k\ 1} + \ _1 \bar{d}_k^T H_k \bar{d}_k \\ = & \ \ \ _{k\ 1}(g_k^T \bar{d}_k + \bar{d}_k^T H_k \bar{d}_k) + \ _{k\ 1}(1 \ \ \ _1) \bar{d}_k^T H_k \bar{d}_k \\ = & \ \ _{k\ 1}(1 \ \ \ _1) \bar{d}_k^T H_k \bar{d}_k > 0: \end{aligned}$$

Therefore, by $\{d_{k;t}\} \to \bar{d}_k$ and $\|\bar{d}_k\| > 0$ (by Assumption 3.3), for sufficiently large $t \in \mathbb{N}$ and sufficiently small $\|r_{k;t}\|$, it follows that $l(x_k; \ _{k\ 1}; g_k; d_{k;t}) > 0$ and

$$\begin{aligned} l(x_k; \ _{k\ 1}; g_k; d_{k;t}) \ \ & \ _{k\ 1} + \ _1 \max\{-d_{k;t}^T H_k d_{k;t}; \ _d \|d_{k;t}\|_2^2\} \\ & + \ _1 \max\{\|c_k\|_1; \|r_{k;t}\|_1 \ \ \|c_k\|_1\}; \end{aligned}$$

with $l(x_k; \ _{k\ 1}; g_k; d_{k;t}) \ \ \frac{\|r_{k;t}\|}{\ _a \ ^{=2}}$. So Line 3 (Algorithm 1) terminates finitely. □

Next, we prove that (3.8) is satisfied at every iteration of Algorithm 1. As mentioned above, this component of the algorithm is central in the analysis.

Lemma 3.6. The sequence of iterates generated by Algorithm 1 satisfies (3.8).

Proof. If the first condition on Line 3 of Algorithm 1 is triggered, i.e., case (a), then the result holds trivially. Thus, we focus on the case where the second condition is triggered, i.e., case(b). In this case, the residual vectors satisfy $\|r_k\|_1 < \min\{(1 \ \ \ _1) \ _2; \ _1 \ _a \ ^{=2}\} \|c_k\|_1$ and $\|\ _k\|_1 < \ _b \|c_k\|_1$, and, potentially, the merit parameter is updated via (3.6)–(3.7). By the residual conditions, it follows that $\|c_k\|_1 > 0$. (Note, Lemma 3.5 showed that if $\|c_k\| = 0$, Line 3 of Algorithm 1 terminates in case (a).) By (3.5), (3.8) and $\|r_k\|_1 < (1 \ \ \ _1) \ _2 \|c_k\|_1$, to complete the proof, it is equivalent to show that

$$\ _k \ \ g_k^T d_k + \ _1 \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\} \ \ (1 \ \ \ _1)\|c_k\|_1 \ \ \|r_k\|_1: \tag{3.12}$$

Since $\|r_k\|_1 < (1 \ \ \ _1) \ _2 \|c_k\|_1 < (1 \ \ \ _1)\|c_k\|_1$, (3.12) directly holds if $g_k^T d_k + \ _1 \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\} \ 0$. On the other hand, if $g_k^T d_k + \ _1 \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\} > 0$, which also implies $g_k^T \bar{d}_k + \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\} > 0$, by (3.6)–(3.7),

$$\ _k \ \ (1 \ \ \ \ ) \ _k^{\text{trial}} = \frac{(1 \ \ \ \ )(1 \ \ \ _1)(1 \ \ \ _2)\|c_k\|_1}{g_k^T d_k + \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\}} < \frac{(1 \ \ \ _1)\|c_k\|_1 \ \ \|r_k\|_1}{g_k^T d_k + \ _1 \max\{-d_k^T H_k d_k; \ _d \|d_k\|_2^2\}},$$

which implies that (3.12) and (3.8) both hold. □

The next lemma provides an upper bound on the primal residuals (3.3).

Lemma 3.7. For all $k \in \mathbb{N}$, the residual vector $r_k \in \mathbb{R}^m$ (3.3) satisfies, $\|r_k\|_1 \leq \omega_a^{-\alpha} \, l(x_k, \lambda_k, g_k, d_k)$, where $\omega_a \in \mathbb{R}_{>0}$, $\alpha \in (0, 1]$ and $\beta \in [2, 4]$.

Proof. There are two cases to consider with regards to Line 3 of Algorithm 1. If condition (a) is triggered, the result holds trivially. If condition (b) is triggered, by Lemma 3.6 and (3.8), it follows that

$$\|r_k\|_1 < \min\{(1 + \omega_1)\omega_2, \omega_1 \omega_a^{\bar{\beta}} \|g_k\| c_k\|_1 + \omega_1 \omega_a^{\bar{\beta}} \|c_k\|_1 + \omega_a^{\bar{\beta}}\} \, l(x_k, \lambda_k, g_k, d_k).$$

Combining the two cases yields the desired result. □

The next lemma provides an upper bound on the deterministic dual variable update, $\|y_k + \lambda_k\|_1$, required for the analysis in Section 4.1. We should note that we never require to compute λ_k in Algorithm 1.

Lemma 3.8. There exists some constant $\omega_y \in \mathbb{R}_{>0}$, such that for all $k \in \mathbb{N}$, $\|y_k + \lambda_k\|_1 \leq \omega_y$.

Proof. For all $k \in \mathbb{N}$, let $Z_k \in \mathbb{R}^{n \times (n-m)}$ be an orthonormal basis for the null space of the Jacobian of the constraints J_k, i.e., $J_k Z_k = 0$. Then under Assumption 3.1, it follows that $Z_k^T H_k Z_k \succeq \omega_1 I$. By (3.11), it follows that

$$\begin{aligned} d_k &= -J_k^T (J_k J_k^T)^{-1} c_k - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k - H_k J_k^T (J_k J_k^T)^{-1} c_k), \\ \text{and } y_k + \lambda_k &= -(J_k J_k^T)^{-1} J_k (g_k + H_k d_k) \\ &= -(J_k J_k^T)^{-1} J_k (I - H_k Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T) g_k \\ &\quad + (J_k J_k^T)^{-1} J_k H_k (I - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^T (J_k J_k^T)^{-1} c_k. \end{aligned}$$

By the Cauchy–Schwarz inequality, and Assumptions 2.1 and 3.1, it follows that

$$\|y_k + \lambda_k\|_1 \leq \omega_J^4 \omega_H^2 \|c\| + \omega_J^2 \|g\|.$$

Thus, selecting a sufficiently large constant $\omega_y \in \mathbb{R}_{>0}$, completes the proof. □

Finally, we show that the model reduction function based on deterministic quantities, i.e., $l(x_k, \lambda_k, g_k, d_k)$, is non-negative and bounded above. For non-optimal points, one can further show the model reduction function is strictly positive.

Lemma 3.9. There exists some fixed constant $\omega_l \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$, $l(x_k, \lambda_k, g_k, d_k) \in [0, \omega_l)$.

Proof. Notice, that in this lemma, we consider only deterministic quantities. First, we show that $l(x_k; \ _k; g_k; d_k) \quad 0$ for all $k \ 2 \ N$. We consider two cases for the outcome of Line 3 of Algorithm 1. If (3.8) is satis ed with $\ _k = \ _{k \ 1}$, then by (3.5) and (3.11), we have

$$l(x_k; \ _k; g_k; d_k) \quad _k ! \ _1 \max \ d_k^T H_k d_k; \ _d k d_k k_2^2 \ + ! \ _1 k c_k k_1 \quad 0:$$

Otherwise, we have $0 = k c_k + J_k d_k k_1 < (1 \quad ! \ _1) ! \ _2 k c_k k_1$. We consider two subcases. If $g_k^T d_k + \max f d_k^T H_k d_k; \ _d k d_k k_2^2 g \quad 0$, then it follows that $g_k^T d_k \quad 0$, and by $\ _k > 0$, (3.5) and (3.11), we have

$$l(x_k; \ _k; g_k; d_k) = \quad _k g_k^T d_k + k c_k k_1 \quad 0:$$

On the other hand, if $g_k^T d_k + \max f d_k^T H_k d_k; \ _d k d_k k_2^2 g > 0$, by (3.6){(3.7) and the fact that $_k \quad (1 \quad) \ _k^{trial} < \ _k^{trial}$, we have

$$_k g_k^T d_k < (1 \quad ! \ _1)(1 \quad ! \ _2) k c_k k_1 \quad _k \max f d_k^T H_k d_k; \ _d k d_k k_2^2 g: \tag{3.13}$$

Combining (3.5), (3.11) and (3.13), it follows that

$$\begin{aligned} l(x_k; \ _k; g_k; d_k) &= \quad _k g_k^T d_k + k c_k k_1 \\ &> \ _k \max f d_k^T H_k d_k; \ _d k d_k k_2^2 g + (1 \quad (1 \quad ! \ _1)(1 \quad ! \ _2)) k c_k k_1 \quad 0: \end{aligned}$$

Thus, we have shown that $l(x_k; \ _k; g_k; d_k) \quad 0$ for all $k \ 2 \ N$.

Next, we show that $l(x_k; \ _k; g_k; d_k) \quad \ _l$ for all $k \ 2 \ N$. For all $k \ 2 \ N$, let $Z_k \ 2 \ R^{n \ (n \ m)}$ be an orthonormal basis for the null space of the Jacobian of the constraints J_k, i.e., $J_k Z_k = 0$, and Assumption 3.1 further implies that $Z_k^T H_k Z_k \quad I$. From (3.11) we have

$$g_k^T d_k = \quad g_k^T (I \quad Z_k (Z_k^T H_k Z_k) \ ^1 Z_k^T H_k) J_k^T (J_k J_k^T) \ ^1 c_k \quad g_k^T Z_k (Z_k^T H_k Z_k) \ ^1 Z_k^T g_k: \tag{3.14}$$

By Assumption 2.1 and (3.14), we have

$$k g_k^T d_k k_2 \quad k \ g_k^T J_k^T (J_k J_k^T) \ ^1 c_k k_2 + k g_k^T Z_k (Z_k^T H_k Z_k) \ ^1 Z_k^T g_k k_2 \quad _g \ _J \ ^2 \ _c + \ ^2_g \ ^1:$$

Moreover, by (3.5) and (3.11), it follows that for all $k \ 2 \ N$,

$$l(x_k; \ _k; g_k; d_k) = \quad _k g_k^T d_k + k c_k k_1 \quad 1 \quad _g \ _J \ ^2 \ _c + \ ^2_g \ ^1 + \ _c:$$

Selecting a su ciently large constant $\ _l \ 2 \ R_{> 0}$ completes the proof. □

4 Adaptive, Inexact and Stochastic Sequential Quadratic Programming

In this section, we prove under di erent conditions on the gradient and linear system solution accuracies, that Algorithm 1 has convergence properties that match those from the deterministic setting in expectation. First, we consider adaptive error bounds (Section 4.1) and then consider predetermined sublinear errors (Section 4.2).

4.1 Adaptive Iteration-Dependent Errors

In this section, we provide a comprehensive convergence analysis for Algorithm 1 under stochastic conditions on the error in the gradient approximations (Assumption 4.1), and inexact solutions to the SQP subproblems (3.3) (Assumptions 4.2 and 4.9). The following two assumptions are central to the analysis presented in this section.

Assumption 4.1. For all $k \in \mathbb{N}$, the stochastic gradient estimate $\bar{g}_k \in \mathbb{R}^n$ satisfies

$$\mathbb{E}_k\left[\|\bar{g}_k - g_k\|_2^2\right] \leq \sigma_1 \, \mathcal{I}(x_k, \lambda_k, g_k, d_k), \tag{4.1}$$

where $\sigma_1 \in \mathbb{R}_{>0}$, $\omega \in (0,1)$, and $\eta \in [2,4]$. Additionally, for all $k \in \mathbb{N}$, the stochastic gradient estimate $\bar{g}_k \in \mathbb{R}^n$ is an unbiased estimator of the gradient of f at x_k, i.e., $\mathbb{E}_k[\bar{g}_k] = g_k$, where $\mathbb{E}_k[\cdot]$ denotes the expectation with respect to the distribution of $\bar{g}_k$ conditioned on the event that the algorithm has reached $x_k \in \mathbb{R}^n$ in iteration $k \in \mathbb{N}$.

Assumption 4.2. For all $k \in \mathbb{N}$, the search directions $(d_k, \lambda_k) \in \mathbb{R}^n \times \mathbb{R}^m$ in (3.3) (inexact solutions to (3.2)) satisfy

$$\left\| \begin{bmatrix} \tilde{d}_k \\ \tilde{\lambda}_k \end{bmatrix} - \begin{bmatrix} d_k \\ \lambda_k \end{bmatrix} \right\|_2^2 \leq \sigma_2 \, \mathcal{I}(x_k, \lambda_k, g_k, d_k), \tag{4.2}$$

where $\sigma_2 \in \mathbb{R}_{>0}$, $\omega \in (0,1)$, and $\eta \in [2,4]$. Note, $(\tilde{d}_k, \tilde{\lambda}_k)$ and (d_k, λ_k) are the exact and inexact solutions of (3.2), respectively.

Remark 4.3. We note upfront that the conditions given in Assumptions 4.1 and 4.2 are not implementable in our stochastic setting, as the right-hand-side of the inequalities depend on deterministic quantities. That being said, we use these assumptions as this allows us to gain insights into the errors permitted in the algorithm while still retaining strong convergence guarantees, and will guide the development of our practical algorithm. An important choice in conditions (4.1)–(4.2) is the (deterministic) model reduction function (3.5) on the right-hand-side of the inequality. As we show in the analysis, and similar to [5, 4, 22], we use this quantity as a proxy of convergence, and as such it is an appropriate measure of the accuracy in the gradient approximations. Another interesting question pertains to the analogue of (4.1) in the unconstrained setting, i.e., no equality constraints. One can show that for appropriately chosen constants σ_1 and ω, in the unconstrained setting (4.1) is the well-known "norm" condition (in expectation) [15, 12]. With regards to (4.2), under Assumption 3.3 the inequality is well-defined. Finally, we emphasize that the constants ω and η are the same constants that appear in Algorithm 1 and that are used in the step size selection. Thus, the gradient accuracy, linear system solves and the step size selection are inherently connected. The precise permissible ranges of the constants in Assumptions 4.1 and 4.2 are made explicit in subsequent lemmas and theorems.

We prove convergence guarantees for Algorithm 1, where the stochastic gradients employed satisfy Assumption 4.1 and the search directions employed satisfy Assumption 4.2. Before we delve into the analysis, we discuss the behavior of the merit parameter sequence $\{\nu_k\}$, a key component of our algorithmic framework. In the deterministic setting (e.g., Algorithm 1 with $\bar{g}_k = g_k$ and $\bar{d}_k = d_k$ for all $k \in \mathbb{N}$), under Assumption 2.1 the merit parameter sequence is bounded away from zero if $\{\nu_k\}_{\min} > 0$. In the stochastic setting, where the gradients approximations employed satisfy Assumption 4.1, boundedness (away from zero) of the merit parameter cannot be guaranteed. However, under the assumption that the iterates generated by Algorithm 1 converge to a stationary point of (2.1), the gradient approximation eventually become sufficient accurate, as do the solutions to the linear system, and thus it is reasonable to assume that the stochastic merit parameter value remains bounded away from zero. That being said, this is not sufficient to prove strong convergence and complexity guarantees for Algorithm 1 for the whole merit parameter sequence. To this end, we impose an additional technical condition on the gradient approximation and the search direction employed (see Assumption 4.9 later in the text), and, to the best of our knowledge, prove the first complexity guarantees in this setting.

For ease of exposition, we make the following assumption throughout this section that unifies all prior assumptions, and for brevity, we do not remind the reader of this fact within the statement of each result.

Assumption 4.4. There exist universal quantities such that Assumptions 2.1, 3.1, 3.3, 4.1 and 4.2 hold for any realization of Algorithm 1.

Let us now commence our analysis of Algorithm 1 under Assumption 4.4. We build up to our main results through a series of lemmas. Our first set of lemmas show that the stochastic search directions computed by Algorithm 1 are well-behaved. To this end, we invoke the following orthogonal decomposition of the (stochastic) search direction: $d_k = u_k + v_k$ where $u_k \in \text{Null}(J_k)$ and $v_k \in \text{Range}(J_k^T)$ for all $k \in \mathbb{N}$.

Lemma 4.5. There exists $\bar{v} \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the normal component v_k satisfies $\max\{\|v_k\|_2, \|v_k\|_2^2\} \leq \bar{v} \max\{\|c_k\|_2, \|r_k\|_2\}$.

Proof. Since $u_k \in \text{Null}(J_k)$ and $v_k \in \text{Range}(J_k^T)$,

$$v_k = J_k^T (J_k J_k^T)^{-1} J_k v_k = J_k^T (J_k J_k^T)^{-1} J_k d_k = J_k^T (J_k J_k^T)^{-1} (r_k - c_k).$$

Thus, by the Cauchy inequality

$$\|v_k\|_2 \leq \|J_k^T (J_k J_k^T)^{-1}\|_2 (\|r_k\|_2 + \|c_k\|_2) \leq 2\|J_k^T (J_k J_k^T)^{-1}\|_2 \max\{\|c_k\|_2, \|r_k\|_2\}.$$

Moreover, it follows that

$$\|v_k\|_2^2 \leq 4\|J_k^T (J_k J_k^T)^{-1}\|_2^2 \max\{\|c_k\|_2, \|r_k\|_2\} \leq \max\{\|c_k\|_2, \|r_k\|_2\}.$$

By Assumption 4.4, we have that $\|c_k\|_2$, $\|r_k\|_2$ and $\|J_k^T (J_k J_k^T)^{-1}\|_2$ are uniformly bounded above for all $k \in \mathbb{N}$, which in turn completes the proof. □

The next lemma shows that if the step d_k is tangentially dominated, i.e., $\|u_k\|_2$ dominates $\|v_k\|_2$, then H_k is sufficiently positive definite along the computed stochastic search direction.

Lemma 4.6. There exists $\theta_{uv} \in \mathbb{R}_{>0}$ such that, for any $k \in \mathbb{N}$, if $\|u_k\|_2^2 \geq \theta_{uv}\|v_k\|_2^2$, then $d_k^T H_k d_k \geq \frac{1}{2}\|u_k\|_2^2$ and $d_k^T H_k d_k \geq \theta_d \|d_k\|_2^2$, where $\bar{H} \in \mathbb{R}_{>0}$ is defined in Assumption 3.1 and $\theta_d \in (0, \frac{1}{2}]$.

Proof. When $\|u_k\| = 0$, to satisfy the condition in the statement, we require $\|u_k\| = \|v_k\| = 0$ which implies $\|d_k\| = 0$, and the statement holds trivially.

When $\|u_k\| > 0$, by Assumption 3.1 it follows that

$$\begin{aligned} d_k^T H_k d_k &= u_k^T H_k u_k + 2 u_k^T H_k v_k + v_k^T H_k v_k \\ &\geq \|u_k\|_2^2 - 2\bar{H}\|u_k\|_2\|v_k\|_2 - \bar{H}\|v_k\|_2^2 \\ &\geq 2\sqrt{\theta_{uv}}\bar{H}\|u_k\|_2^2 - \bar{H}\|u_k\|_2^2 \geq \tfrac{1}{2}\|u_k\|_2^2 \end{aligned}$$

for sufficiently large θ_{uv}. Meanwhile, when $\|u_k\| > 0$, it follows that

$$d_k^T H_k d_k \geq \tfrac{1}{2}\|u_k\|_2^2 \geq \theta_d(1 + \tfrac{1}{\theta_{uv}})\|u_k\|_2^2 \geq \theta_d(\|u_k\|_2^2 + \|v_k\|_2^2) = \theta_d\|d_k\|_2^2$$

for sufficiently large θ_{uv}. □

For $\theta_{uv} \in \mathbb{R}_{>0}$ (defined in Lemma 4.6), let $K_{\bar{u}} := \{k \in \mathbb{N} : \|u_k\|_2^2 \geq \theta_{uv}\|v_k\|_2^2\}$ and $K_{\bar{v}} := \{k \in \mathbb{N} : \|u_k\|_2^2 < \theta_{uv}\|v_k\|_2^2\}$ denote index sets that form a partition of $\mathbb{N}$, and let

$$\rho_k := \begin{cases} \|u_k\|_2^2 + \kappa\|c_k\|_2 & \text{if } k \in K_{\bar{u}}; \\ \max\{\|c_k\|_2, \|r_k\|_2\} & \text{if } k \in K_{\bar{v}}. \end{cases}$$

Our next result shows that the squared norms of the stochastic search directions and the constraint violations for all $k \in \mathbb{N}$ are bounded above by ρ_k.

Lemma 4.7. There exists $\gamma \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the search direction and constraint violation satisfy $\|d_k\|_2^2 \leq \gamma\rho_k$ and $\|d_k\|_2^2 + \kappa\|c_k\|_2 \leq (\gamma + 1)\rho_k$.

Proof. For all $k \in K_{\bar{u}}$,

$$\|d_k\|_2^2 = \|u_k\|_2^2 + \|v_k\|_2^2 \leq 1 + \tfrac{1}{\theta_{uv}}\|u_k\|_2^2 \leq 1 + \tfrac{1}{\theta_{uv}}\rho_k.$$

For all $k \in K_{\bar{v}}$, by Lemma 4.5,

$$\|d_k\|_2^2 = \|u_k\|_2^2 + \|v_k\|_2^2 < (1 + \theta_{uv})\|v_k\|_2^2 \leq (1 + \theta_{uv})\,\bar{v}\,\rho_k.$$

Therefore, we set $\gamma := \max\{1 + \frac{1}{\theta_{uv}}, (1 + \theta_{uv})\,\bar{v}\}$ to satisfy $\|d_k\|_2^2 \leq \gamma\rho_k$. Finally, since $\kappa\|c_k\|_2 \leq \rho_k$ trivially, this concludes the proof. □

The next lemma shows that the stochastic model reduction, $l(x_k; \ _k; g_k; d_k)$, is bounded below by a non-negative quantity.

Lemma 4.8. There exists $\ _l \in R_{>0}$ such that, for all $k \in N$, the reduction in the model of the merit function satisfies, $l(x_k; \ _k; g_k; d_k) \quad \ _l \ _k^2$.

Proof. For all $k \in K_u$, by Lemmas 3.6 and 4.6,

$$\begin{aligned} l(x_k; \ _k; g_k; d_k) \quad & \ _k(-\ _1 \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\} + (\ _1 \max\{\|c_k\|_1; \|r_k\|_1\} \ _k)\|c_k\|_1) \\ & \ _k(-\ _1 \max\{\ _2; \ _d \ \|u_k\|_2^2\} + \ _1\|c_k\|_1) \quad \ _k(-\ _1 \min\{\ _2; \tfrac{1}{\ _1}\} \ _k)\|c_k\|_1. \end{aligned}$$

For all $k \in K_v$, by Lemma 3.6,

$$\begin{aligned} l(x_k; \ _k; g_k; d_k) \quad & \ _k(-\ _1 \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\} + (\ _1 \max\{\|c_k\|_1; \|r_k\|_1\} \ _k)\|c_k\|_1) \\ & (\ _1 \max\{\|c_k\|_1; \|r_k\|_1\} \ _k)\|c_k\|_1) \quad \tfrac{\ _1}{2} \max\{\|c_k\|_1; \|r_k\|_1\} \quad \tfrac{\ _1}{2} \ _k \end{aligned}$$

Setting $\ _l := \ _1 \min\{ =2; 1=2 \ _1\} \in R_{>0}$ completes the proof. □

Next we introduce a technical condition required for the analysis. Specifically, the condition allows us to establish an integral and useful upper bound for the difference between the stochastic and deterministic merit parameters values, and as a result is of vital importance in establishing complexity result.

Assumption 4.9. For all $k \in N$, the stochastic gradient estimate $g_k \in R^n$ and the search direction $d_k \in R^n$ (inexact solution of (3.2)) satisfy

$$\begin{aligned} |(g_k^T d_k + \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\}) \quad & (g_k^T d_k + \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\})| \\ & \ _3 \ ^{=2}|g_k^T d_k + \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\}|; \end{aligned}$$

where $\ _3 \in R_{>0}$, $\in (0;1)$, $\in [2;4]$, and $\ _3 \ ^{=2} \in (0;1)$.

Remark 4.10. We make a few remarks about Assumption 4.9. This is a strong assumption, nonetheless, it is necessary in order to establish the strong non-asymptotic convergence and complexity results presented in this paper. In the unconstrained setting, Assumption 4.9 does not add any additional restrictions. That is, in the unconstrained setting, consider the gradient method, where H_k is the identity matrix, $d_k = \ g_k$, $d_k = \ g_k$. Clearly, $(g_k^T d_k + \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\}) = 0$ and $g_k^T d_k + \max\{d_k^T H_k d_k; \ _d\|d_k\|_2^2\} = 0$, thus no additional restrictions are imposed. Thus, in the unconstrained setting, Assumptions 4.1, 4.2 and 4.9 reduce to the well known "norm-condition". Several difficulties arise in the setting with constraints. The primary difficulty pertains to the fact that the merit parameter is possibly adjusted across iterations making the merit function a moving target. Thus, in order to establish convergence and complexity results for all iterations, as compared to other

papers that only consider the iterations after the merit parameter has stabilized at a sufficiently small constant value, additional control on the permissible differences is required. We should note that if one happened to know a sufficiently small merit parameter value, then Assumption 4.9 would no longer be required for the theory. Finally, again we point out the connection between the accuracy in the gradient approximations, the quality of the solution to the linear system, the step size through user-defined parameters and .

Our next lemma shows that the stochastic merit parameter sequence $\{\bar{\tau}_k\}$ generated by Algorithm 1 is bounded away from zero.

Lemma 4.11. Under Assumptions 4.4 and 4.9, there exists a constant $\bar{\tau}_{\min} \in \mathbb{R}_{>0}$ that following Algorithm 1, $\bar{\tau}_k \geq \bar{\tau}_{\min}$ for all $k \in \mathbb{N}$.

Proof. Under Assumption 4.4 (specifically Assumptions 2.1 and 3.1), it is well established that $\{\tau_k^{\text{trial}}\}$, the sequence of deterministic variant of (3.6), is always positive and bounded away from zero. By Assumption 4.9, it follows that either $(g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\})$ and $(g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\})$ are both positive or both negative.

If $g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\} \leq 0$, by (3.6), $\tau_k^{\text{trial}} = 1$. By Assumption 4.9 it follows that $g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\} \leq 0$ and $\tau_k^{\text{trial}} = 1$.

If $g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\} > 0$, by Assumption 4.9, it follows that

$$0 < (1 - \epsilon_3^{=2}) \leq \frac{g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\}}{g_k^T d_k + \max\{d_k^T H_k d_k, \delta_d \|d_k\|_2^2\}} \leq (1 + \epsilon_3^{=2}).$$

By (3.6), the above inequality implies $\frac{\tau_k^{\text{trial}}}{(1+\epsilon_3^{=2})} \leq \tau_k^{\text{trial}} \leq \frac{\tau_k^{\text{trial}}}{(1-\epsilon_3^{=2})}$.

In both cases considered above, given the fact that $\{\tau_k^{\text{trial}}\}$ is positive and bounded away from zero, one can conclude that there exists a constant $\tau_{\min}^{\text{trial}} \in \mathbb{R}_{>0}$ such that $\tau_k^{\text{trial}} \geq \tau_{\min}^{\text{trial}}$ for all $k \in \mathbb{N}$. Finally, invoking (3.7), completes the proof since $\bar{\tau}_k \geq \min\{\bar{\tau}_1, (1-\epsilon)\tau_{\min}^{\text{trial}}\}$. □

The next lemma provides a useful lower bound for the reduction in the merit function that is proportional to the stochastic search direction computed.

Lemma 4.12. There exist constants $\{\epsilon, \epsilon_{l,d}\} \in \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$, $l(x_k, \bar{\tau}_k; g_k, d_k) \leq -(\bar{\tau}_k L_k + \epsilon)\|d_k\|_2^2 \leq -\epsilon_{l,d}\|d_k\|_2^2$.

Proof. By Lemmas 4.7–4.11, it follows that

$$l(x_k, \bar{\tau}_k; g_k, d_k) \leq -\epsilon_l \bar{\tau}_k \|d_k\|_2^2 = -\frac{\epsilon_l \bar{\tau}_k}{(\bar{\tau}_k L_k + \epsilon)}(\bar{\tau}_k L_k + \epsilon)\|d_k\|_2^2 \leq -\frac{\epsilon_l \bar{\tau}_{\min}}{(\bar{\tau}_{\min} L_k + \epsilon)}(\bar{\tau}_k L_k + \epsilon)\|d_k\|_2^2 \leq -\epsilon_l \bar{\tau}_{\min}\|d_k\|_2^2,$$

where the result follows by choosing appropriate constants and $\epsilon_{l,d}$. □

Remark 4.13. Corollaries of the above lemmas can be derived for the special case under which deterministic quantities are used instead of stochastic, i.e.(3.11) is solved exactly. Specically, under the same logic as in Lemmas 4.7-4.12, it follows that there exist constants $\{\ ;\ _{l;d}\} \quad \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$

$$l(x_k;\ _k;g_k;d_k) \qquad (\ _k L_k + \ _k)kd_kk_2^2 \qquad \ _{l;d}kd_kk_2^2: \tag{4.3}$$

The next lemma provides upper and lower bounds on the adaptive step sizes employed by Algorithm 1.

Lemma 4.14. Let $\ _k$ be defined via (3.9){(3.10). For all $k \in \mathbb{N}$, there exists a constant $\_ \in \mathbb{R}_{>0}$ such that $\_ \qquad \ _k \qquad \ _u^{(2\ =2)}$.

Proof. The upper bound follows from (3.9). To derive the lower bound, we consider different cases. If $\ _k = \min\left\{\ _u^{(2\ =2)};\ _k^{opt};1\right\}$, under the conditions $\in (0;1]$ and $\in [2;4]$, it follows that $\ _k \quad \min\{\ _u;\ _k^{opt};1\}$. Otherwise, if $\ _k = \frac{2(1\ )\ ^{(=2\ 1)}\ l(x_k;\ _k;g_k;d_k)}{(\ _kL_k+\ _k)kd_kk_2^2}$, by Lemma 4.12, $\in (0;1]$ and $\in [2;4]$, it follows that $\ _k \quad 2(1\ )$, where $\in \mathbb{R}_{>0}$ is defined in Lemma 4.12. Combining (3.10) and Lemma 4.12, it follows that $\ _k^{opt} \quad \min\left\{\frac{l(x_k;\ _k;g_k;d_k)}{(\ _kL_k+\ _k)kd_kk^2};1\right\} \quad \min\{\ ;1\}$. Setting $\_ := \min\{\ _u;1;\ ;2(1\ )\ \}$ completes the proof. □

The next result provides an upper bound on the change of merit function value after a step. Central to the proof of this lemma is our step size strategy (3.9)-(3.10).

Lemma 4.15. For all $k \in \mathbb{N}$, it follows that

$$\begin{aligned}(x_k + \ _kd_k;\ _k) \qquad & (x_k;\ _k) \\ & \ _k\ l(x_k;\ _k;g_k;d_k) + (1\ )\ _k^{(=2\ 1)}\ l(x_k;\ _k;g_k;d_k) \\ + \ _k\ _kg_k^T(d_k \quad d_k) + \ _k(\ _k \quad \ _k)g_k^Td_k + \ _kkJ_k(d_k \quad \tilde{d}_k)k_1;\end{aligned}$$

where d_k and d_k are defined in (3.3) and (3.11), respectively, $\ _k$ is updated via (3.6){(3.7) and $\ _k$ is its deterministic counterpart, and $\ _k$ is computed via (3.9)-(3.10).

Proof. By the step size selection strategy (3.9)-(3.10), for all $k \in \mathbb{N}$, we have that $\ _k \quad 1$.

By the triangle inequality, Assumption 2.1, and (3.2), it follows that

$$
\begin{aligned}
& (x_k + \alpha_k d_k; \mu_k) \quad (x_k; \mu_k) \\
= \; & \alpha_k (f(x_k + \alpha_k d_k) \quad f(x_k)) + \alpha_k c(x_k + \alpha_k d_k) \quad \alpha_k c_k \\
& \alpha_k (\alpha_k g_k^T d_k + \tfrac{L}{2} \alpha_k^2 \|d_k\|_2^2) + \alpha_k c_k + \alpha_k J_k d_k \quad \alpha_k c_k \\
& \alpha_k (\alpha_k g_k^T d_k + \tfrac{L}{2} \alpha_k^2 \|d_k\|_2^2) + |1 \quad \alpha_k| \alpha_k c_k + \alpha_k c_k + J_k d_k \quad \alpha_k c_k \\
= \; & \alpha_k (\alpha_k g_k^T d_k \quad \alpha_k c_k) + \tfrac{1}{2} (\alpha_k L +) \alpha_k^2 \|d_k\|_2^2 + \alpha_k c_k + J_k d_k \\
= \; & \alpha_k (\alpha_k g_k^T d_k \quad \alpha_k c_k) + \tfrac{1}{2} (\alpha_k L +) \alpha_k^2 \|d_k\|_2^2 \\
& + \alpha_k \alpha_k g_k^T (d_k \quad d_k) + \alpha_k (\alpha_k) g_k^T d_k + \alpha_k c_k + J_k \tilde{d}_k + J_k (d_k \quad \tilde{d}_k) \\
= \; & \alpha_k l(x_k; \mu_k; g_k; d_k) + \tfrac{1}{2} (\alpha_k L +) \alpha_k^2 \|d_k\|_2^2 \\
& + \alpha_k \alpha_k g_k^T (d_k \quad d_k) + \alpha_k (\alpha_k) g_k^T d_k + \alpha_k J_k (d_k \quad \tilde{d}_k) .
\end{aligned}
$$

By (3.9), we have for all $k \in \mathbb{N}$, that $\alpha_k \leq \frac{2(1 \quad)}{(\alpha_k L +) \|d_k\|_2^2} (\theta = 2 \quad 1) l(x_k; \mu_k; g_k; d_k)$, and

$$
\begin{aligned}
& (x_k + \alpha_k d_k; \mu_k) \quad (x_k; \mu_k) \\
& \quad \alpha_k l(x_k; \mu_k; g_k; d_k) + \tfrac{1}{2} \alpha_k (\alpha_k L +) \frac{2(1 \quad)}{(\alpha_k L +) \|d_k\|_2^2} (\theta = 2 \quad 1) l(x_k; \mu_k; g_k; d_k) \|d_k\|_2^2 \\
& + \alpha_k \alpha_k g_k^T (d_k \quad d_k) + \alpha_k (\alpha_k) g_k^T d_k + \alpha_k J_k (d_k \quad \tilde{d}_k) \\
= \; & \alpha_k l(x_k; \mu_k; g_k; d_k) + (1 \quad) \alpha_k (\theta = 2 \quad 1) l(x_k; \mu_k; g_k; d_k) \\
& + \alpha_k \alpha_k g_k^T (d_k \quad d_k) + \alpha_k (\alpha_k) g_k^T d_k + \alpha_k J_k (d_k \quad \tilde{d}_k) ,
\end{aligned}
$$

which is the desired result. □

We now proceed to state and prove a series of lemmas (Lemmas 4.16–4.19) that provide bounds for the differences between the deterministic and stochastic gradients, search directions, and their inner products.

Lemma 4.16. For all $k \in \mathbb{N}$, $\mathbb{E}_k[\|g_k \quad g_k\|_2] \leq \frac{\sqrt{1 \quad }}{2} \sqrt{\theta} \sqrt{l(x_k; \mu_k; g_k; d_k)}$.

Proof. By Assumption 4.1 and Jensen's inequality, it follows that $\mathbb{E}_k[\|g_k \quad g_k\|_2] \leq \sqrt{\mathbb{E}_k[\|g_k \quad g_k\|_2^2]} \leq \sqrt{\frac{1}{1 \quad }} \sqrt{l(x_k; \mu_k; g_k; d_k)}$. □

Lemma 4.17. For all $k \in \mathbb{N}$, $\mathbb{E}_k[\tilde{d}_k] = d_k$, $\mathbb{E}_k[\tilde{u}_k] = u_k$, and $\mathbb{E}_k[\tilde{\mu}_k] = \mu_k$. Moreover, there exists $\sigma \in \mathbb{R}_{>0}$, such that $\mathbb{E}_k[\|\tilde{d}_k \quad d_k\|_2] \leq \sigma \sqrt{\theta} \sqrt{l(x_k; \mu_k; g_k; d_k)}$.

Proof. The first statement follows from the facts that, the matrix on the left-hand side of (3.2) is invertible and deterministic under Assumption 4.4 and conditioned on Algorithm 1

having reached iterate x_k at iteration k, and due to the fact that expectation is a linear operator. For the second statement, for any realization of g_k,

$$\begin{bmatrix} \tilde{d}_k \\ \tilde{\ }_k \end{bmatrix} - \begin{bmatrix} d_k \\ \ _k \end{bmatrix} = \begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix}^{-1} \begin{bmatrix} g_k \\ 0 \end{bmatrix} - \begin{bmatrix} g_k \\ 0 \end{bmatrix} \implies \| \tilde{d}_k - d_k \|_2 \leq \ _L \| g_k - g_k \|_2;$$

where (under Assumption 4.4) $_L \in \mathbb{R}_{>0}$ is an upper bound on the norm of the matrix shown above. Thus, it follows by Lemma 4.16 that

$$\mathbb{E}_k\left[\| \tilde{d}_k - d_k \|_2\right] \leq \mathbb{E}_k\left[\ _L \| g_k - g_k \|_2\right] \leq \ _L \sqrt{\frac{1}{\ }} = \ ^2 \sqrt{\frac{}{}}(x_k; \ _k; g_k; d_k).$$

□

Lemma 4.18. For all $k \in \mathbb{N}$,

$$\begin{aligned}
&|g_k^T(d_k - \tilde{d}_k)| \leq \ _{g;d\tilde{d}} \| g_k - g_k \|_2 \sqrt{(x_k; \ _k; g_k; d_k)}; \\
&|g_k^T \tilde{d}_k - g_k^T d_k| \leq \ _{gg;d\tilde{d}} \| g_k - g_k \|_2 \sqrt{(x_k; \ _k; g_k; d_k)} + \ _L \| g_k - g_k \|_2^2; \\
&|g_k^T(d_k - \tilde{d}_k)| \leq \ _{g;d\tilde{d}}^{\ 2} (x_k; \ _k; g_k; d_k) + \ _{g;d\tilde{d}}^{\ 2} (x_k; \ _k; g_k; d_k) \\
&\qquad + \ _L \| g_k - g_k \|_2^2; \\
&|g_k^T(d_k - d_k)| \leq \ _{g;dd} \| g_k - g_k \|_2 \sqrt{(x_k; \ _k; g_k; d_k)} \\
&\qquad + \ _{g;dd}^{\ 2} (x_k; \ _k; g_k; d_k) + \ _{g;dd}^{\ 2} (x_k; \ _k; g_k; d_k); \\
&|g_k^T d_k - g_k^T d_k| \leq \ _{gg;dd} \| g_k - g_k \|_2 \sqrt{(x_k; \ _k; g_k; d_k)} + 2 \ _L \| g_k - g_k \|_2^2 \\
&\qquad + \ _{gg;dd}^{\ 2} (x_k; \ _k; g_k; d_k) + \ _{gg;dd}^{\ 2} (x_k; \ _k; g_k; d_k); \\
&\text{and } \|J_k(d_k - \tilde{d}_k)\|_1 \leq \ _{J;d\tilde{d}}^{\ 2} (x_k; \ _k; g_k; d_k);
\end{aligned}$$

where $_{g;d\tilde{d}} = \sqrt{\frac{H \ L}{\ _{l;d}}} \in \mathbb{R}_{>0}$, $_{gg;d\tilde{d}} = \ _{g;d\tilde{d}} + \sqrt{\frac{1}{\ _{l;d}}} \in \mathbb{R}_{>0}$, $_L \in \mathbb{R}_{>0}$, $_{g;d\tilde{d}} = \sqrt{\frac{H \ ^2}{\ _{l;d}}} \in \mathbb{R}_{>0}$, $_{g;dd} = \ _y \| \ _a \in \mathbb{R}_{>0}$, $_{g;dd} = \sqrt{\frac{H \ ^2}{\ _{l;d}}} \in \mathbb{R}_{>0}$, $_{gg;dd} = \ _y \| \ _a \in \mathbb{R}_{>0}$, $_{gg;dd} = \ _{g;d\tilde{d}} \in \mathbb{R}_{>0}$, $_{gg;dd} = \ _{g;d\tilde{d}} \in \mathbb{R}_{>0}$, and $_{J;d\tilde{d}} = \| \ _a \in \mathbb{R}_{>0}$.

Proof. (First inequality) By Assumption 3.1, and (3.2), (3.11), (4.3),

$$\begin{aligned}
|g_k^T(d_k - \tilde{d}_k)| &= |(g_k + J_k^T(y_k + \ _k))^T(d_k - \tilde{d}_k)| \\
&= |(H_k d_k)^T(d_k - \tilde{d}_k)| \leq \ _H \|d_k\|_2 \|d_k - \tilde{d}_k\|_2 \\
&\leq \ _H \sqrt{\frac{(x_k; \ _k; g_k; d_k)}{\ _{l;d}}} \ _L \| g_k - g_k \|_2;
\end{aligned} \tag{4.4}$$

where the result follows using the definition of $_{g;d\tilde{d}}$.

(Second inequality) By Lemma 4.17, and (3.2), (3.11), (4.3), (4.4),

$$
\begin{aligned}
&|g_k^T \tilde{d}_k - g_k^T d_k| \\
&\le |(g_k - g_k)^T d_k| + |g_k^T(\tilde{d}_k - d_k)| + |(g_k - g_k)^T(\tilde{d}_k - d_k)| \\
&\le \|g_k - g_k\|_2 \|d_k\|_2 + \sigma_{g,dd} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + \|g_k - g_k\|_2 \|\tilde{d}_k - d_k\|_2 \\
&\le \|g_k - g_k\|_2 \sqrt{\frac{l(x_k, \epsilon_k; g_k; d_k)}{l_{,d}}} + \sigma_{g,dd} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + \epsilon_L \|g_k\|_2^2;
\end{aligned}
\tag{4.5}
$$

where the result follows using the definition of $\sigma_{gg,\tilde{d}d}$.

(Third Inequality) The third inequality is proven as follows. By Assumptions 3.1 and 4.2, Lemmas 3.7, 3.8 and 4.17, and (3.2), (3.3), (3.11), (4.3),

$$
\begin{aligned}
|g_k^T(d_k - \tilde{d}_k)| &= |(g_k + J_k^T(y_k + \epsilon_k) - J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&\le |(g_k + J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| + |(J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&\le |(g_k + J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| + |(g_k - g_k)^T(d_k - \tilde{d}_k)| \\
&\quad + |(J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&\le |(H_k d_k)^T(d_k - \tilde{d}_k)| + \epsilon_L \|g_k\|_2^2 + \kappa_{y\epsilon} \epsilon_k \|r_k\|_1 \\
&\le \sqrt{\frac{\|H\|_2}{l_{,d}}} \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + \epsilon_y \epsilon_a \sqrt{l(x_k, \epsilon_k; g_k; d_k)} \\
&\quad + \epsilon_L \|g_k\|_2^2;
\end{aligned}
\tag{4.6}
$$

where the result follows using the definitions of $\sigma_{g,d\tilde{d}}$ and $\sigma_{g,\tilde{d}d}$.

(Fourth inequality) By Assumptions 3.1 and 4.2, Lemmas 3.7 and 3.8, and (3.2), (3.3), (3.11), (4.3), it follows that

$$
\begin{aligned}
|g_k^T(d_k - \tilde{d}_k)| &= |(g_k + J_k^T(y_k + \epsilon_k) - J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&\le |(g_k + J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| + |(J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&\le |(H_k d_k)^T(d_k - \tilde{d}_k)| + \kappa_{y\epsilon} \epsilon_k \|r_k\|_1 \\
&\le \sqrt{\frac{\|H\|_2}{l_{,d}}} \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + \epsilon_y \epsilon_a \sqrt{l(x_k, \epsilon_k; g_k; d_k)}.
\end{aligned}
$$

By the triangle inequality and (4.4),

$$
\begin{aligned}
|g_k^T(d_k - d_k)| &\le |g_k^T(\tilde{d}_k - d_k)| + |g_k^T(d_k - \tilde{d}_k)| \\
&\le \|H\|_2 \sqrt{\frac{l(x_k, \epsilon_k; g_k; d_k)}{l_{,d}}} \epsilon_L \|g_k\|_2 + \sqrt{\frac{\|H\|_2}{l_{,d}}} \sqrt{l(x_k, \epsilon_k; g_k; d_k)} \\
&\quad + \epsilon_y \epsilon_a \sqrt{l(x_k, \epsilon_k; g_k; d_k)}
\end{aligned}
$$

where the result follows using the definition of $\sigma_{g,dd}$ and $\sigma_{g,dd}$.

(Fifth inequality) By (4.5), (4.6),

$$|g_k^T d_k - g_k^T \tilde{d}_k| \le |g_k^T d_k - g_k^T \tilde{d}_k| + |g_k^T (\tilde{d}_k - d_k)|$$
$$\le \sigma_{gg,dd} \|g_k\| \|g_k\|_2 \sqrt{l(x_k, \lambda_k, g_k, d_k) + 2\kappa_L \|g_k\| \|g_k\|_2^2}$$
$$+ \kappa_{g,dd}^{=2} l(x_k, \lambda_k, g_k, d_k) + \kappa_{g,dd}^{=2} l(x_k, \lambda_k, g_k, d_k);$$

where the result follows using the definitions of $\sigma_{gg,dd}$ and $\kappa_{gg,dd}$.

(Sixth inequality) By Lemma 3.7, and (3.2), (3.3),

$$\|J_k(d_k - \tilde{d}_k)\|_1 = \|r_k\|_1 \sqrt{\kappa_a^{=2} l(x_k, \lambda_k, g_k, d_k)};$$

where the result follows using the definition of $\kappa_{J,dd}^{=}$. □

The next lemma provides an upper bound for $|g_k^T d_k|$ with respect to reduction in the model of the merit function, i.e., $l(x_k, \lambda_k, g_k, d_k)$.

Lemma 4.19. For all $k \in \mathbb{N}$, $|g_k^T d_k| \le \kappa_{gd,l} \, l(x_k, \lambda_k, g_k, d_k)$, where $\kappa_{gd,l} = \frac{\kappa_H}{l,d} + \sqrt{\frac{\kappa_y}{\bar{m}_{l,\min}}} \in \mathbb{R}_{>0}$.

Proof. By Assumption 3.1, Lemma 4.8 (deterministic variant), and (3.11), (4.3),

$$|g_k^T d_k| = |d_k^T H_k d_k + d_k^T J_k^T (y_k + \lambda_k)| \le \kappa_H \|d_k\|_2^2 + \kappa_y \|c_k\|_1$$
$$\le \frac{\kappa_H}{l,d} + \sqrt{\frac{\kappa_y}{\bar{m}_{l,\min}}} \, l(x_k, \lambda_k, g_k, d_k);$$

where the result follows using the definition of $\kappa_{gd,l}$. □

Next we state and prove an upper bound on the difference between the deterministic and stochastic merit parameters. When $\|c_k\| = 0$, by (3.6)–(3.7), $\sigma_k = \tilde{\sigma}_k = \sigma_{k-1}$. When $\|c_k\| > 0$, for the ease of exposition, we equivalently reformulate the merit parameter sequence $\{\sigma_k\}$ update (3.6)–(3.7) as

$$\sigma_k : \begin{cases} \sigma_{k-1} & \text{if } g_k^T d_k + \max\{d_k^T H_k d_k, \sigma_d \|d_k\|^2\} \ge \frac{(1-\tau)(1-\rho_1)(1-\rho_2)\|c_k\|_1}{\sigma_{k-1}}, \\ \frac{(1-\tau)(1-\rho_1)(1-\rho_2)\|c_k\|_1}{g_k^T d_k + \max\{d_k^T H_k d_k, \sigma_d \|d_k\|^2\}} & \text{otherwise.} \end{cases} \tag{4.7}$$

(The update formula for the deterministic merit parameter $\tilde{\sigma}_k$ can be defined as above with the stochastic quantities replaced by their deterministic counterparts.) It is clear that if the merit parameter is updated from its previous value, i.e., $\sigma_k \neq \sigma_{k-1}$, then

$$\frac{(1-\tau)(1-\rho_1)(1-\rho_2)\|c_k\|_1}{g_k^T d_k + \max\{d_k^T H_k d_k, \sigma_d \|d_k\|_2^2\}} < \sigma_{k-1}. \tag{4.8}$$

Similarly, given $\tilde{\sigma}_{k-1}$ if the deterministic merit parameter is updated,

$$\frac{(1-\tau)(1-\rho_1)(1-\rho_2)\|c_k\|_1}{g_k^T d_k + \max\{d_k^T H_k d_k, \sigma_d \|d_k\|_2^2\}} < \tilde{\sigma}_{k-1}. \tag{4.9}$$

Lemma 4.20. For all $k \in \mathbb{N}$, $|(_k \; _k) g_k^T d_k| \; ^{=2} \; l(x_k; \; _k; g_k; d_k)$, where $=$ $2 \; _3 \; _1 \; gd; \; _l$ and $\in (0; 1=(2 \; _3)^{2=} \,]$.

Proof. By the merit parameter updating mechanism (3.6)–(3.7), the merit parameter values $(_k$ and $_k)$ are only potentially updated if $\|c_k\|_1 > 0$. We divide the proof into two cases based the outcome of Line 3 in Algorithm 1 (a) or case (b)). Let $h_k = g_k^T d_k + \max\{d_k^T H_k d_k; \; _d\|d_k\|_2^2\}$ and $h_k = g_k^T d_k + \max\{d_k^T H_k d_k; \; _d\|d_k\|_2^2\}$.

If Line 3 in Algorithm 1 terminates in case (a), and $\|c_k\|_1 > 0$, it follows that $h_k \; 0$, and by Assumption 4.9, $h_k \; 0$; otherwise, when $\|c_k\|_1 = 0$, by (3.11) and Assumptions 3.1, it follows that $d_k \in \mathrm{Null}(J_k)$ and $h_k = g_k^T d_k + d_k^T H_k d_k = \; d_k^T J_k^T (y_k + \; _k) = c_k^T (y_k + \; _k) = 0$. Thus, neither the stochastic nor the deterministic merit parameters update, i.e., $_k = \; _k = \; _{k\;1}$. The result holds since $l(x_k; \; _k; g_k; d_k)$ is non-negative.

Next, we consider the case where Line 3 in Algorithm 1 terminates with case (b). We divide merit parameter values $(_k; \; _k)$ into four cases.

Case (i) : Neither $_k$ or $_k$ are updated, i.e., $_k = \; _k = \; _{k\;1}$. In this case, the result holds since $l(x_k; \; _k; g_k; d_k)$ is non-negative.

Case (ii) : Both $_k$ and $_k$ are updated, i.e., (4.8) and (4.9) hold. By (4.7), (4.8) and (4.9), and Assumption 4.9, it follows that

$$| _k \; _k| = (1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1 \frac{|h_k \; h_k|}{h_k h_k}$$
$$\frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1 \; _3 \; ^{=2}}{h_k} \quad _3 \; ^{=2} \quad _1 .$$

Case (iii) : The stochastic merit parameter $_k$ is updated and the deterministic merit parameter $_k$ is not, i.e., $_k \quad _k = \; _{k\;1}$. Since the stochastic merit parameter is updated, $_k^{trial} < 1$ and $h_k > 0$, and by Assumption 4.9 it follows that $h_k > 0$, and $_k^{trial} < 1$. Moreover, since the deterministic merit parameter is not updated,

$$\frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1}{h_k} \qquad _{k\;1} . \tag{4.10}$$

By (4.7), (4.8) and (4.10), and Assumption 4.9, it follows that

$$| _k \quad _k| = \; _{k\;1} \quad \frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1}{h_k}$$
$$\frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1}{h_k} \quad \frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1}{h_k}$$
$$(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1 \frac{|h_k \; h_k|}{h_k h_k}$$
$$\frac{(1 \quad)(1 \; ! \, _1)(1 \; ! \, _2)\|c_k\|_1 \; _3 \; ^{=2}}{h_k} < \; _3 \; ^{=2} \quad _1 .$$

Case (iv) : The deterministic merit parameter $_k$ is updated, while the stochastic merit parameter $_k$ is not, i.e., $_k \quad _k = \; _{k\;1}$. Since the deterministic merit parameter

is updated, $\alpha_k^{\text{trial}} < 1$ and $h_k > 0$, and by Assumption 4.9 it follows that $h_k > 0$ and $\alpha_k^{\text{trial}} < 1$. Moreover, since the stochastic merit parameter is not updated,

$$\frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{h_k} \geq \alpha_k \|d_k\|_1. \tag{4.11}$$

By (4.7), (4.9) and (4.11), and Assumption 4.9, it follows that

$$\begin{aligned}
|\Delta_k - \delta_k| &= \alpha_k \|d_k\|_1 \frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{h_k} \\
&< \frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{h_k} \frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{h_k} \\
&\quad (1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1 \frac{|h_k - h_k|}{h_k h_k} \\
&\quad \frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{h_k} \\
&\quad \frac{(1-\eta)(1-\tau_1)(1-\tau_2)\kappa c_k \|d_k\|_1}{(1-\sigma_3^{-2})h_k} \frac{\sigma_3^{-2}}{1-\sigma_3^{-2}} \leq 1.
\end{aligned}$$

Combing the four cases above and Lemma 4.19 yields the result. □

The next lemma bounds the stochastic model of the reduction of the merit function.

Lemma 4.21. For all $k \in \mathbb{N}$ and $\alpha \in (0, 1/(2\sigma_1 + \sigma_{gg,dd})\sigma_2]$,

$$\begin{aligned}
l(x_k + \alpha d_k; g_k; d_k) &\leq \left(1 + \bar{\sigma}_{l,1}\alpha^2\right) l(x_k, \alpha; g_k; d_k) \\
&\quad + 2\alpha\sigma_1 \sigma_{gg,dd}\|g_k\|\|g_k\|_2 \sqrt{l(x_k, \alpha; g_k; d_k)} + 2\alpha L\|g_k\|\|g_k\|_2^2,
\end{aligned}$$

where $\bar{\sigma}_{l,1} = 2(\sigma + \sigma_1\sigma_{gg,dd} + \sigma_1\sigma_{gg,dd}) \in \mathbb{R}_{>0}$. Additionally, under Assumption 4.1, for all $k \in \mathbb{N}$,

$$\mathbb{E}_k\left[l(x_k + \alpha d_k; g_k; d_k)\right] \leq \left(1 + \bar{\sigma}_{l,1}\alpha^2\right) l(x_k, \alpha; g_k; d_k),$$

where $\bar{\sigma}_{l,1} = 2\sigma + \sigma_1(\sigma_{gg,dd} + \sigma_{gg,dd}\sqrt{\sigma_1} + 2\sigma_{gg,dd}L\sigma_1) \in \mathbb{R}_{>0}$.

Proof. By (3.5), and Lemmas 4.18 and 4.20, it follows that

$$\begin{aligned}
l(x_k + \alpha d_k; g_k; d_k) &= \alpha g_k^T d_k + \kappa c_k \|d_k\|_1 \alpha^2 c_k + J_k d_k \|_1 \\
&= \alpha g_k^T d_k + \kappa c_k \|d_k\|_1 \\
&\quad + (\alpha - \alpha)g_k^T d_k + \alpha(g_k^T d_k - g_k^T d_k) \alpha^2 c_k + J_k d_k \|_1 \\
&\quad l(x_k, \alpha; g_k; d_k) + |(\alpha - \alpha)g_k^T d_k| + |\alpha g_k^T d_k - g_k^T d_k| \\
&\quad l(x_k, \alpha; g_k; d_k) + \sigma^{-2} l(x_k, \alpha; g_k; d_k) \\
&\quad + \alpha\sigma_1 \sigma_{gg,dd}\|g_k\|\|g_k\|_2 \sqrt{l(x_k, \alpha; g_k; d_k)} + 2\alpha L\|g_k\|\|g_k\|_2^2 \\
&\quad + \sigma_{gg,dd}\sigma^{-2} l(x_k, \alpha; g_k; d_k) + \sigma_{gg,dd}\sigma^{-2} l(x_k, \alpha; g_k; d_k) :
\end{aligned}$$

Thus, by choosing $\in (0; 1=(2\ _{1\ gg;d\ d})^{2=}]$,

$$l(x_k;\ _k; g_k; d_k) \quad 1 + \frac{(\ + \ _{1\ gg;d\ d} + \ _{1\ gg;d\ d})^{\ =2}}{1 \quad _{1\ gg;d\ d}^{\ =2}} \quad l(x_k;\ _k; g_k; d_k)$$
$$+ \frac{_{1\ \ gg;\eth d} k g_k \quad g_k k_2 \sqrt{l(x_k;\ _k; g_k; d_k)} + 2\ _L k g_k \quad g_k k_2^2}{1 \quad _{1\ gg;d\ d}^{\ =2}}$$
$$1 + 2(\ + \ _{1\ gg;dd} + \ _{1\ gg;dd})^{\ =2} \quad l(x_k;\ _k; g_k; d_k)$$
$$+ 2\ _{1\ \ gg;\eth d} k g_k \quad g_k k_2 \sqrt{l(x_k;\ _k; g_k; d_k)} + 2\ _L k g_k \quad g_k k_2^2 :$$

The first result follows using the definition of $\bar{}_{l;\ l}$. Furthermore, by Assumption 4.1,

$$E_k\ _h \ l(x_k;\ _k; g_k; d_k)$$
$$E_k \quad 1 + 2(\ + \ _{1\ gg;dd} + \ _{1\ gg;dd})^{\ =2} \quad l(x_k;\ _k; g_k; d_k)$$
$$+2\ _{1\ \ gg;\eth d} k g_k \quad g_k k_2 \sqrt{l(x_k;\ _k; g_k; d_k)} + 2\ _L k g_k \quad g_k k_2^2 \ _i$$
$$1 + 2(\ + \ _{1\ gg;dd} + \ _{1\ gg;dd})^{\ =2} \quad l(x_k;\ _k; g_k; d_k)$$
$$+ 2\ _{1\ \ gg;\eth d} \sqrt{\bar{}_{1}^{\ =2}}\ l(x_k;\ _k; g_k; d_k) + 2\ _L \ _1 \quad l(x_k;\ _k; g_k; d_k)$$
$$1 + 2 \quad + \ _1(\ _{gg;dd} + \ _{gg;dd} + \ _{gg;\eth d} \sqrt{\bar{}_{1}} + 2\ _L \ _1)^{\ =2} \quad l(x_k;\ _k; g_k; d_k);$$

where the second result follows using the definition of $\bar{}_{l;\ l}$. □

The next lemma bounds the difference in the merit function after a step.

Lemma 4.22. For all $k \in \mathbb{N}$, there exist $\in \mathbb{R}_{>0}$ such that

$$E_k[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] \quad E_k[(\ _{k+1} \quad \ _k)] \ f_{inf} \quad (\_ \quad) \ l(x_k; g_k; d_k):$$

Proof. By Lemmas 4.14, 4.15, and 4.21, and (3.4), it follows that for all k 2 N

$$
\begin{aligned}
& E_k\,[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] \\
= \ & E_k\,[\ (x_{k+1};\ _{k+1}) \quad (x_{k+1};\ _k) + \ (x_{k+1};\ _k) \quad (x_k;\ _k)] \\
= \ & E_k\,[(\ _{k+1} \quad _k) f_{k+1}] + E_k\,[\ (x_{k+1};\ _k) \quad (x_k;\ _k)] \\
& E_k\,[(\ _{k+1} \quad _k)]\, f_{inf} \\
& \quad E_k \Big[\ _k\ l(x_k;\ _k; g_k; d_k) \quad (1 \quad)\ _k{}^{(\ =2\ \ 1)}\ l(x_k;\ _k; g_k; d_k)\Big] \\
& + E_k\ \ _k\ _k g_k^T (d_k \quad d_k)\ + E_k\ \ _k(\ _k \quad _k) g_k^T d_k\ + E_k \Big[\ _k k J_k (d_k \quad d_k) k_1 \Big] \\
& E_k\,[(\ _{k+1} \quad _k)]\, f_{inf} \quad E_k\,[\ _k\ \ l(x_k;\ _k; g_k; d_k)] \\
& + E_k \Big[(1 \quad)\ _k{}^{(\ =2\ \ 1)}(1 + \ \overline{\ \ }_{l;\ \ l}\ {}^{=2})\ \ l(x_k;\ _k; g_k; d_k)\Big] \\
& + E_k \Big[(1 \quad)2\ \ _u \quad _1\ \ _{gg;dd} k g_k \quad g_k k_2 \sqrt{\ \ l(x_k;\ _k; g_k; d_k)} + 2\ \ _L k g_k \quad g_k k_2^2\Big] \\
& + \ \ _u{}^{(2\ \ \ =2)} E_k\ \ _k j g_k^T (d_k \quad d_k) j \\
& + \ \ _u{}^{(2\ \ \ =2)} E_k\ \ j(\ _k \quad _k) g_k^T d_k j\ + \ \ _u{}^{(2\ \ \ =2)} E_k \Big[\ k J_k (d_k \quad d_k) k_1\Big] :
\end{aligned}
$$

Continuing from the above, by Lemmas 4.14, 4.18, 4.20 and 4.21,

$$
\begin{aligned}
& E_k\,[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] \\
& E_k\,[(\ _{k+1} \quad _k)]\, f_{inf} \\
& \quad E_k \Big[\ _k\ l(x_k;\ _k; g_k; d_k) \quad (1 \quad)\ \overline{\ \ }_{l;\ \ l}\ \ _k{}^{(\ \ \ 1)}\ \ l(x_k;\ _k; g_k; d_k)\Big] \\
& + 2(1 \quad)\ \ _u \quad _1\ \ _{gg;dd} \sqrt{\overline{\ \ }_1}\ {}^{=2} + 2\ \ _L\ \ _1 \quad l(x_k;\ _k; g_k; d_k) \\
& + \ \ _u{}^2\ \ _1\ (\ \ _{g;dd} \sqrt{\overline{\ \ }_1} + \ \ _{g;dd})\ \ l(x_k;\ _k; g_k; d_k) + \ \ _{g;dd} E_k \quad l(x_k;\ _k; g_k; d_k) \\
& + \ \ _u{}^2 \quad l(x_k;\ _k; g_k; d_k) + \ \ _u{}^2\ \ _{J;dd} E_k \quad l(x_k;\ _k; g_k; d_k) \\
& E_k\,[(\ _{k+1} \quad _k)]\, f_{inf} \\
& \quad \ \_\ \ \ l(x_k;\ _k; g_k; d_k) + (1 \quad)\ \overline{\ \ }_{l;\ \ l}\ \ _u{}^{(\ =2+1)}\ \ l(x_k;\ _k; g_k; d_k) \\
& + 2(1 \quad)\ \ _u{}^{(\ =2+1)} \quad _1\ \ _{gg;dd} \sqrt{\overline{\ \ }_1} + 2\ \ _L\ \ _1\ {}^{=2} \quad l(x_k;\ _k; g_k; d_k) \\
& + \ \ _u{}^2\ \ _1\ (\ \ _{g;dd} \sqrt{\overline{\ \ }_1} + \ \ _{g;dd}) + \ \ _{g;dd}\ \ 1 + \ \overline{\ \ }_{l;\ \ l}\ {}^{=2} \quad l(x_k;\ _k; g_k; d_k) \\
& + \ \ _u{}^2 \quad l(x_k;\ _k; g_k; d_k) + \ \ _u{}^2\ \ _{J;dd}\ \ 1 + \ \overline{\ \ }_{l;\ \ l}\ {}^{=2} \quad l(x_k;\ _k; g_k; d_k) \\
& E_k\,[(\ _{k+1} \quad _k)]\, f_{inf} \quad \ \_\ \ \ l(x_k;\ _k; g_k; d_k) + \ \ {}^2 \quad l(x_k;\ _k; g_k; d_k):
\end{aligned}
$$

The result follows by setting

$$:= \quad u \quad (1 \quad) \quad \overline{}_{l;\ l} + 2 \quad 1 \quad gg;dd \quad \sqrt{} \quad 1 + 2 \quad L \quad 1 \quad +$$
$$+ \quad 1\Big(\sqrt{}_{g;dd} \quad 1 + \quad g;dd\Big) + \Big(\quad 1 \quad g;dd + \quad J;dd\Big) \quad 1 + \overline{}_{l;\ l} \quad 2\ R_{>0}:$$

□

Lemma 4.22 provides an upper bound on the change in the merit function across a step. The bound has two terms: the first term is related to the difference in the merit parameter across iterations and the second term is negative, conditioned on being sufficiently small, and proportional to the model of the reduction of the merit function. We are now ready to prove the main theorem of this section.

Theorem 4.23. By choosing $2\ \ 0; \min\ \frac{1}{(2\ \ 3)^{2=}}; \frac{(1\ \ )\ \ \_}{}; \frac{1}{(2\ \ 1\ \ gg;d\ d)^{2=}}$ for any 2 $(0;1)$,

$$\lim_{k!1} E\Big[\sum_{j=0}^{k\ \ 1} l(x_j;\ _j;g_j;d_j)\Big] < 1\ ;$$

from which it follows that, $\lim_{k!1} E[\ l(x_k;\ _k;g_k;d_k)] = 0$.

Proof. By Lemma 4.22 and $2\ (0;(1\ \ )\ \ \_=\ ]$, it follows that

$$\begin{aligned} E_k[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] &\ E_k[(\ _{k+1} \quad \ _k)]\ f_{inf} \quad (\ \_\ \ ) \ l(x_k;\ _k;g_k;d_k) \\ &\ E_k[(\ _{k+1} \quad \ _k)]\ f_{inf} \quad \_ \quad l(x_k;\ _k;g_k;d_k): \end{aligned} \tag{4.12}$$

Applying a telescopic sum to (4.12) and taking the total expectation, it follows that

$$\begin{aligned} 1 \ < \ & \inf \quad (x_0;\ _0) \quad E[\ (x_k;\ _k) \quad (x_0;\ _0)] \\ &= E\Big[\sum_{j=0}^{k\ \ 1} (\ (x_{j+1};\ _{j+1}) \quad (x_j;\ _j))\Big] \\ &\ E\Big[\sum_{j=0}^{k\ \ 1} (\ _{j+1} \quad \ _j) f_{inf}\Big] \quad \_ \quad E\Big[\sum_{j=0}^{k\ \ 1} l(x_j;\ _j;g_j;d_j)\Big] \\ &\ \ _1|f_{inf}| \quad \_ \quad E\Big[\sum_{j=0}^{k\ \ 1} l(x_j;\ _j;g_j;d_j)\Big]; \end{aligned}$$

which completes the proof. □

Theorem 4.23 describes the behavior of the model of the reduction of the merit function evaluated at the iterates generated by the Algorithm 1 in expectation. We connect the result of Theorem 4.23 to feasibility and stationarity measures below.

Corollary 4.24. Under the conditions of Theorem 4.23, Algorithm 1 yields a sequence of iterates $\{(x_k, y_k)\}$ for which

$$\lim_{k \to \infty} \mathbb{E}\left[\|d_k\|_2^2\right] = 0; \quad \lim_{k \to \infty} \mathbb{E}\left[\|c_k\|_2\right] = 0; \quad \text{and} \quad \lim_{k \to \infty} \mathbb{E}\left[\|g_k + J_k^T(y_k + \cdot_k)\|_2\right] = 0.$$

Proof. By the deterministic analogue of Lemma 4.8, $l(x_k; \cdot_k; g_k; d_k) \geq \cdot_l \min_k$. By Theorem 4.23 and the definitions of $\cdot_l$ and $\cdot_{\min}$, we have $\lim_{k \to \infty} \mathbb{E}[\cdot_k] = 0$, and thus, our first two results follow from the deterministic variant of Lemma 4.7. The final result follows from (3.11), i.e., $\|g_k + J_k^T(y_k + \cdot_k)\|_2 = \|H_k d_k\|_2 \leq \|H_k\|_2 \|d_k\|_2 \leq \cdot_H \|d_k\|_2$, which completes the proof. □

Corollary 4.24 shows that in the limit and in expectation the search direction, the constraint violation and a first-order stationarity measure converge to zero.

Remark 4.25. We make a few remarks about the main theoretical results (Theorem 4.23 and Corollary 4.24).

- Comparison to deterministic results: The result in Corollary 4.24 is similar, albeit in expectation, to what can be proven for an exact deterministic SQP method, i.e. $\bar{g}_k = g_k$ and $\bar{d}_k = d_k$, under the same assumptions; see e.g., [5, 13].
- Comparison to [5]: The main difference in the result of Corollary 4.24 and similar results for the stochastic SQP algorithm proposed in [5] pertain to the requirements on the $\{\cdot_k\}$ sequence. In [5] (and other works, e.g., [4, 21, 22]) a diminishing $\{\cdot_k\}$ sequence is required to guarantee convergence, whereas in this work convergence with a constant $\{\cdot_k\}$ sequence is derived due to the variance reduction achieved.
- Comparison to [28]: In [28], a stochastic line search SQP method for equality constrained problems that utilizes an exact differentiable merit function is proposed. Under deterministic conditions on the function and derivative approximations and exact solutions to the linear systems, the authors show convergence analogous to that of a deterministic algorithm. We note that under the same deterministic conditions, similar results can be established for our proposed adaptive sampling algorithm.
- Comparison to [6]: A result analogous to Theorem 4.23 is proven in [6]. In both works this is possible due to variance reduction in the approximations employed; the algorithm proposed in [6] makes use of predictive variance reduction via SVRG gradients, whereas in this work achieves variance reduction via adaptive sampling.

The final result we show in this section is a complexity result for our proposed algorithm, i.e., the number of iterations required to achieved an ϵ-accurate solution in expectation. Specifically, we consider the following complexity metric,

$$\mathbb{E}[\|g_k + J_k^T(y_k + \lambda_k)\|_2] \leq \epsilon_L, \quad \text{and} \quad \mathbb{E}[\|c_k\|_1] \leq \epsilon_c, \tag{4.13}$$

for $\epsilon_L \in (0, 1)$ and $\epsilon_c \in (0, 1)$.

Corollary 4.26. Under the conditions of Theorem 4.23, Algorithm 1 generates an iterate $\{(x_k, y_k)\}$ that satisfies (4.13) in at most

$$K = \frac{\theta_1(f(x_0) - f_{\inf}) + \|c_0\|_1}{\bar{\sigma}_x} \max\{\epsilon_L^{-2}, \epsilon_c^{-1}\} \tag{4.14}$$

iterations. Moreover, if $\epsilon_L = \epsilon$ and $\epsilon_c = \epsilon^2$, then $K = O(\epsilon^{-2})$.

Proof. First we show that if $\mathbb{E}[\|g_k + J_k^T(y_k + \lambda_k)\|_2] > \epsilon_L$ or $\mathbb{E}[\|c_k\|_1] > \epsilon_c$, then

$$\mathbb{E}[\Phi(x_k, \lambda_k; g_k, d_k)] \leq -\bar{\sigma}_x \min\{\epsilon_L^2, \epsilon_c\}, \tag{4.15}$$

where $\bar{\sigma}_x = \min\left\{\bar{\omega}_1, \frac{\min\{\bar{\omega}_1 \bar{\sigma}_d\}}{2\bar{H}}\right\} \in \mathbb{R}_{>0}$. Consider arbitrary $(k, \epsilon_L, \epsilon_c) \in \mathbb{N} \times (0, 1) \times (0, 1)$ for which $\mathbb{E}[\|g_k + J_k^T(y_k + \lambda_k)\|_2] > \epsilon_L$ and/or $\mathbb{E}[\|c_k\|_1] > \epsilon_c$. First, suppose that $\mathbb{E}[\|c_k\|_1] > \epsilon_c$. By the deterministic variant of (3.8),

$$\mathbb{E}[\Phi(x_k, \lambda_k; g_k, d_k)] \leq \mathbb{E}[\eta_k \bar{\omega}_1 \max\{-d_k^T H_k d_k, \bar{\sigma}_d \|d_k\|_2^2\} + \bar{\omega}_1 \|c_k\|_1] \leq \mathbb{E}[\bar{\omega}_1 \|c_k\|_1] < \bar{\omega}_1 \epsilon_c.$$

Next, suppose that $\mathbb{E}[\|g_k + J_k^T(y_k + \lambda_k)\|_2] > \epsilon_L$. By Assumption 3.1 and (3.11),

$$\epsilon_L < \mathbb{E}[\|g_k + J_k^T(y_k + \lambda_k)\|_2] = \mathbb{E}[\|H_k d_k\|_2] \leq \mathbb{E}[\bar{H}\|d_k\|_2],$$

and thus, by the deterministic variant of (3.8), the fact that η_k is bounded below and the definition of $\bar{\sigma}_d$, it follows that

$$\begin{aligned}\mathbb{E}[\Phi(x_k, \lambda_k; g_k, d_k)] &\leq \mathbb{E}[\eta_k \bar{\omega}_1 \max\{-d_k^T H_k d_k, \bar{\sigma}_d \|d_k\|_2^2\} + \bar{\omega}_1 \|c_k\|_1] \\ &\leq \mathbb{E}[\min\{\bar{\omega}_1 \bar{\sigma}_d\} \|d_k\|_2^2] < \frac{\min\{\bar{\omega}_1 \bar{\sigma}_d\}}{2\bar{H}} \epsilon_L^2.\end{aligned}$$

Combining the results of the two cases and using the definition of $\bar{\sigma}_x$ yields (4.15).

If (4.13) is violated, then by Lemma 4.14, (3.4), (4.12) and (4.15), we have

$$\begin{aligned}&\mathbb{E}[\theta_1(f(x_0) - f_{\inf}) + \|c_0\|_1] \leq \mathbb{E}[\theta_0(f(x_0) - f_{\inf}) + \|c_0\|_1] \\ &\leq \mathbb{E}[\theta_0 f(x_0) + \|c_0\|_1 - \eta_k f(x_k) - \eta_k \|c_k\|_1 + (\eta_k - \theta_0) f_{\inf}] \\ &= \mathbb{E}\left[\sum_{j=0}^{k-1} (\Phi(x_j, \lambda_j) - \Phi(x_{j+1}, \lambda_{j+1}) + (\theta_{j+1} - \theta_j) f_{\inf})\right] \\ &\leq \mathbb{E}\left[\sum_{j=0}^{k-1} \eta_j \Phi(x_j, \lambda_j; g_j, d_j)\right] \leq \mathbb{E}\left[\sum_{j=0}^{k-1} -\bar{\sigma}_x \Phi(x_j, \lambda_j; g_j, d_j)\right] \\ &\leq k(-\bar{\sigma}_x \min\{\epsilon_L^2, \epsilon_c\}),\end{aligned}$$

which implies that k is bounded above by (4.14). □

The result of Corollary 4.26 is similar to that of determinsitc SQP methods, albeit in expectation, under the same assumptions; see e.g., [21, Theorem 1]. To the best of our knowledge, this is the first time that such complexity results have been derived.

4.2 Predetermined Sublinear Errors

In this subsection, we consider the case where the sample sizes used in the gradient estimation are monotonically increased at a sublinear rate and the accuracy in the linear system solves is increased at a sublinear rate to achieve convergence to the solution in expectation. The reason we include this result is to emphasize that predetermined sampling strategies and error sequences suffice to provide convergence guarantees without the need for stronger assumptions. Also, this result provides guidance on the total sampling complexity. For brevity, we only introduce the assumptions and present the main theoretical results. All technical lemmas and proofs are deferred to Appendix A.

The two assumptions below are analogues of Assumptions 4.1 and 4.2.

Assumption 4.27. For all $k \in \mathbb{N}$, the stochastic gradient estimate $g_k \in \mathbb{R}^n$ satisfies, $\mathbb{E}_k\|g_k - \nabla g_k\|_2^2 \leq \frac{\nu_1}{(k+1)^\alpha}$, where $\nu_1 \in \mathbb{R}_{>0}$, $\alpha \in (0,1)$, $\beta \in \mathbb{R}_{>1}$ and $\gamma \in [2,4]$. Additionally, for all $k \in \mathbb{N}$, the stochastic gradient estimate $g_k \in \mathbb{R}^n$ is an unbiased estimator of the gradient of f at x_k, i.e., $\mathbb{E}_k[g_k] = \nabla g_k$.

Assumption 4.28. For all $k \in \mathbb{N}$, the search directions $(d_k, \lambda_k) \in \mathbb{R}^n \times \mathbb{R}^m$ in (3.3) (inexact solutions to (3.2)) satisfy, $\left\|\begin{bmatrix} \tilde{d}_k \\ \tilde{\lambda}_k \end{bmatrix} - \begin{bmatrix} d_k \\ \lambda_k \end{bmatrix}\right\|_2^2 \leq \frac{\nu_2}{(k+1)^\alpha}$, where $\nu_2 \in \mathbb{R}_{>0}$, $\alpha \in (0,1)$, $\beta \in \mathbb{R}_{>1}$ and $\gamma \in [2,4]$. Note, $(\tilde{d}_k, \tilde{e}_k)$ and $(\bar{d}_k, \bar{\lambda}_k)$ are the exact and inexact solutions of (3.2), respectively.

Next, we state the main results of this subsection. Lemma 4.29 (analogue of Lemma 4.22) provides a bound on the difference of the merit function across iterations.

Lemma 4.29. For all $k \in \mathbb{N}$, there exist constants $(\underline{\quad}, \underline{\quad}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that

$$\mathbb{E}_k[\phi(x_{k+1}, \lambda_{k+1}) - \phi(x_k, \lambda_k)] + \mathbb{E}_k[\psi(\lambda_{k+1} - \lambda_k)] \leq (\underline{\quad}) l(x_k, \lambda_k; g_k, d_k) + \frac{\nu_1 \nu_2}{(k+1)^\alpha}.$$

(Note, the definitions of the constants are given in Appendix A.)

Lemma 4.29 has an additional as compared to Lemma 4.22. This is due to the fact that we control the gradient error and linear system accuracy at a sublinear rate instead of controlling it relative to the algorithmic progress, as measured in terms of $l(x_k, \lambda_k; g_k, d_k)$.

That being said, the additional term is proportional to ^{2} and whose accumulation is nite in the limit.

Theorem 4.30 and Corollary 4.31 (analogues of Theorem 4.23 and Corollary 4.26, respectively) provide convergence and iteration complexity results, respectively. Corollary 4.31 also provides sample complexity results.

Theorem 4.30. By choosing $2 \;\; 0; \min \left\{ 1 = (2 \;\; 3)^{\frac{2}{}}; (1 \;\;) \frac{}{} = ; 1 = 2 \;\; \frac{1}{ \; g; d d} \frac{2}{} \right\}$ for any 2 $(0; 1)$ and $2 \; R_{>1}$, $\lim_{k!1} \; E \left[\sum_{j=0}^{k-1} \; l(x_j; \; j; g_j; d_j) \right] < 1$, from which it follows that, $\lim_{k!1} \; E[\; l(x_k; \; k; g_k; d_k)] = 0$.

The conclusion of Theorem 4.30 is the same as that of Theorem 4.23, up to constants, and, as a result, Corollary 4.24 holds for Theorem 4.30.

Corollary 4.31. Under the conditions of Theorem 4.30, Algorithm 1 generates an iterate $f(x_k; y_k)g$ that satis es (4.13) in at most $K = O\left(\max \;\; {}_L^{2}; \; {}_c^{1} \right)$ iterations and $W = O\left(\max \;\; {}_L^{2}; \; {}_c^{1} \right)^{(+1)}$, $2 \; R_{>1}$, stochastic gradient evaluations. Moreover, if ${}_L =$ and ${}_c = {}^2$, then $K = O \;\; {}^2$ and $W = O(\;\; {}^{2(+1)})$.

Corollary 4.31 shows that one can achieve the same iteration complexity as the deterministic variant of the algorithm under Assumptions 4.27 and 4.28 (and other assumptions stated earlier) at an increased overall sample complexity.

5 Practical AIS-SQP

In this section, we present our proposed practical adaptive inexact stochastic SQP method (PAIS-SQP) and describe the sample size selection mechanism, iterative linear system solver and early termination conditions employed.

5.1 Sample Size Selection

We describe the mechanism by which the sample size is selected at every iteration. Condition (4.1) involves computing population variances and deterministic quantities which are not available in our setting, and possibly requires solving multiple linear systems. That being said, one can approximate these quantities with sample variances and sampled stochastic counterparts of the deterministic quantities required following the ideas proposed in [12, 9].

Condition (4.1) is approximated as follows. Let $g_k \; 2 \; R^n$ be de ned as

$$g_k := \frac{1}{jS_kj} \sum_{i 2 S_k} r \; F(x_k; \; i); \tag{5.1}$$

where S_k is a set consisting of indices drawn at random from the distribution of , and $_i$ is a realization of . The left-and-side of (4.1) can be expressed as

$$E_k \quad kg_k \quad r \quad f(x_k)k_2^2 \;=\; \frac{E_k[kr \;\; F(x_k; \;) r \quad f(x_k)k_2^2]}{jS_kj}: \tag{5.2}$$

Computing the population variance on the right-hand-side of (5.2) is prohibitively expensive in our setting, and thus we approximate it with the sample variance (see left-hand-side of (5.3)). Moreover, the right-hand-side of (4.1) is approximated with its stochastic counter-part. This results in the following approximation to condition (4.1),

$$\frac{Var_{i2S_k}[r \;\; F(x_k; \;_i)]}{jS_kj} \qquad _1 \qquad l(x_k; \;_k; g_k; d_k); \tag{5.3}$$

where $Var_{i2S_k}[r \;\; F(x_k; \;_i)] = \frac{1}{jS_kj \;\; 1} \sum_{i2S_k} kr \;\; F(x_k; \;_i) \quad g_k k_2^2$, used in PAIS-SQP.

In our practical algorithm (Algorithm 2), if inequality (5.3) is not satis ed (given the set S_k), we increase the sample size to a size that (at least with high probability) will satisfy (4.1). The heuristic we propose to do this is as follows. Suppose we wish to nd a larger sample $\hat{S}_k$ ($j\hat{S}_kj > jS_kj$), and let us assume that the change in sample size is gradual enough that for any x_k, such that $Var_{i2S_k}[r \;\; F(x_k; \;_i)] \quad Var_{i2\hat{S}_k}[r \;\; F(x_k; \;_i)]$ and $l(x_k; \;_k; g_k; d_k) \quad l(x_k; \hat{}_k; \hat{g}_k; \hat{d}_k)$, where $\hat{}_k; \hat{g}_k; \hat{d}_k$ are stochastic realizations of these quantities computed based on the sample $\hat{S}_k$. Under this assumption, it is clear that (5.3) is satis ed if

$$j\hat{S}_kj \qquad l \; \frac{Var_{i2S_k}[r \;\; F(x_k; \;_i)]}{_1 \qquad l(x_k; \;_k; g_k; d_k)} \; m : \tag{5.4}$$

This is the nal level of approximation where we check (5.4) to set the sample size for the next iteration, i.e., $S_{k+1} = \hat{S}_k$. The ideas above are used in Algorithm 2.

5.2 Inexact Linear System Solutions

Our proposed practical algorithm makes use of an iterative solver with early termination tests to solve the Newton-SQP linear system (3.3). We use the minimum residual (MINRES) method [18, 30], with early termination conditions, for solving the system inexactly, but note that other iterative algorithms could also be used. For all $k \; 2 \; N$, let $f(d_{k;t}; \;_{k;t}; \;_{k;t}; r_{k;t})g_{t2N}$ denote the steps (and residuals) generated in iteration $t \; 2 \; N$ of MINRES, and $(d_k; r_k; \;_k; r_k) \quad (d_{k;t^0}; r_{k;t^0}; \;_{k;t^0}; r_{k;t^0})$ where t^0 is the last MINRES iteration. The MINRES method was terminated for the minimum t such that either condition (a)

$$(3.8) \;\; \text{and} \;\; kr_kk_1 \quad !_a \;^{=2} \; l(x_k; \;_k; g_k; d_k) \;\; \text{with} \;\; _k = \;_{k \;\; 1} \tag{5.5}$$

or, condition (b)

$$kr_kk_1 < \min f(1 \quad !_1)!_2; !_1!_a \;^{=2} gkc_kk_1 \;\; \text{and} \;\; k \;_kk_1 < !_bkc_kk_1 \tag{5.6}$$

hold. These conditions are inspired by the theory, but relaxed for practicality. Speci cally, the additional condition in case (a) is not checked and neither is Assumption 4.9.

5.3 PAIS-SQP

In this section, we present our practical algorithm PAIS-SQP.

Algorithm 2 (PAIS-SQP) Practical, Adaptive, Inexact, Stochastic SQP Algorithm

Require: $x_0 \in \mathbb{R}^n$; $y_0 \in \mathbb{R}^m$; $\{H_k\} \subset \mathbb{S}^n$; $\hat{S}_{-1} \subset \mathbb{N}$; $\{\beta_1, \beta_2, \gamma, \eta\} \subset (0,1)$; $\beta_a \in \mathbb{R}_{>0}$; $\beta_b \in \mathbb{R}_{>0}$; $\theta \in (0,1]$; $\alpha_u \in \mathbb{R}_{>0}$; $\alpha_d \in (0, \frac{1}{2})$; $\rho \in [2,4]$

1: **for all** $k \in \mathbb{N}$ **do**
2: Compute g_k via (5.1) with $S_k = \hat{S}_{k-1}$
3: Solve (3.2) iteratively using MINRES;
Compute a step (d_k, λ_k) that satisfies either (5.5) or (5.6)
4: Update σ_k via (3.6)–(3.7)
5: Compute a step size α_k via (3.9)–(3.10)
6: Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$, and $y_{k+1} \leftarrow y_k + \alpha_k \lambda_k$
7: Choose a sample $\hat{S}_k$ such that $|\hat{S}_k| = |S_k|$
8: If condition (5.3) is not satisfied, augment $\hat{S}_k$ using formula (5.4)
9: **end for**

6 Numerical Results

The main goal of this section is to illustrate the efficiency of our proposed practical algorithm (Algorithm 2). We characterize efficiency in terms of two metrics that capture the major costs in solving (2.1). The first metric is the number of objective function gradient evaluations (or accessed data points in the context of the machine learning problems presented below). The second metric is the number of iterations of an iterative solver used to solve the linear system (3.2). To illustrate the efficiency of Algorithm 2, we present results on two classes of problems, constrained classification problems that arise in machine learning (Section 6.2) and standard CUTE collection of nonlinear optimization problems (Section 6.3), and compare exact and inexact variants (all implementation details are given in Section 6.1).

In both Sections 6.2 and 6.3, results are given in terms of feasibility and stationarity errors defined as, $\|c_k\|_1$ and $\|g_k + J_k^T y_k\|_1$, respectively, where the vector $y_k \in \mathbb{R}^m$ is computed as a least-squares multiplier of the above using the true gradient g_k. Moreover, in both sections we terminate all algorithms solely due to iteration, sampled gradient evaluation or linear system iteration budgets.

6.1 Implementation Details

We compare different variants of Algorithm 2. Specifically, we compare variants with different levels of accuracy in the gradient approximations employed, as well as variants with and without the early termination conditions. Precise characterization of the accuracy

levels are given in Sections 6.2 and 6.3. For all variants MINRES was used to solve the linear systems, and all variants employed the same adaptive step size selection strategy described by (3.9){(3.10). For all problems estimates of L and $$ were computed using gradient differences around the initial point, and kept constant throughout the course of optimization. This procedure was performed in such a way that, for each problem instance, all algorithms used the same values for these estimates. For all algorithms, $H_k = I$ for all $k \in \mathbb{N}$, $_1 = 1$, $= 1$, $_u = 10^2$, $= 1/2$, $!_1 = 1/2$, $!_2 = 1/2$, $!_a = 10^2$, $!_b = 10^2$, $= 10^{-4}$, and $= 2$. For PAIS-SQP, we additionally set $_1 = 0.99$. (Note, $_2$ and $_3$ are not required by the PAIS-SQP algorithm.) The initial sample size $|\mathcal{S}_1|$ is defined in Sections 6.2 and 6.3. For variants that solve the linear system without the early termination conditions, the termination tolerance used in MINRES was $_{\text{MINRES}} = 10^{-8}$, in order to obtain accurate solutions.

6.2 Constrained Logistic Regression

In our first set of experiments, we consider constrained logistic regression problems,

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-y_i (X_i^T x)}\right) \quad \text{s.t.:} \quad Ax = b_1; \quad \|x\|_2^2 = b_2; \tag{6.1}$$

where $X \in \mathbb{R}^{n \times N}$ is the data matrix, X_i is the ith column of matrix X, $y \in \{-1, 1\}^N$ contains corresponding label data, $A \in \mathbb{R}^{m \times n}$, $b_1 \in \mathbb{R}^m$ and $b_2 \in \mathbb{R}$. We present results on two data sets from the LIBSVM collection [17]; australian and mushroom (Results on more data sets from the LIBSVM collection are presented in Appendix B). For the linear constraints, the data was generated as follows: the entries of the matrix A and the vector b_1 were drawn from a standard normal distribution (for each data set the same A and b_1 were used for all methods). For the ℓ_2-norm constraint, $b_2 = 1$. For all problems and algorithms, the initial primal iterate was set to the vector of all ones of appropriate dimension, and the initial dual variables y_0 were set as the least squares multipliers.

For each data set, we consider exact and inexact (linear system solutions) variants, i.e., variants with and without the early termination conditions, and three different sample sizes ($|\mathcal{S}_k| = |\mathcal{S}| \in \{2, 128, N\}$ for all $k \in \mathbb{N}$) for a total of 6 variants, and compare against PAIS-SQP. A budget of 50 epochs was given to every method. The results for the two data sets are presented in Figures 1 and 2. For every method, we report the feasibility and stationarity errors in terms of iterations, epochs (gradient evaluations) and linear system iterations. The results indicate that our proposed practical inexact SQP method PAIS-SQP strikes a good balances between reducing constraint violation while attempting to find a point that satisfies approximate first-order stationarity across all three evaluation metrics. In Figure 3 we show the sample size and step size selected by the different variants. While the sample sizes increase relatively quickly, there are significant savings that can be achieved by employing inexact information. The step size figures illustrate that the adaptive step size mechanism (3.9){(3.10) is stable.

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

Figure 1: australian : Feasibility & stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1).

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

Figure 2: mushroom Feasibility/stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1).

(a) Step Size vs. Iterations (b) Batch Size vs. Iterations (c) Step Size vs. Iterations (d) Batch Size vs. Iterations

Figure 3: australian ((a), (b)); mushroom ((c), (d)): Step sizes and batch sizes versus iterations.

6.3 CUTE Problems

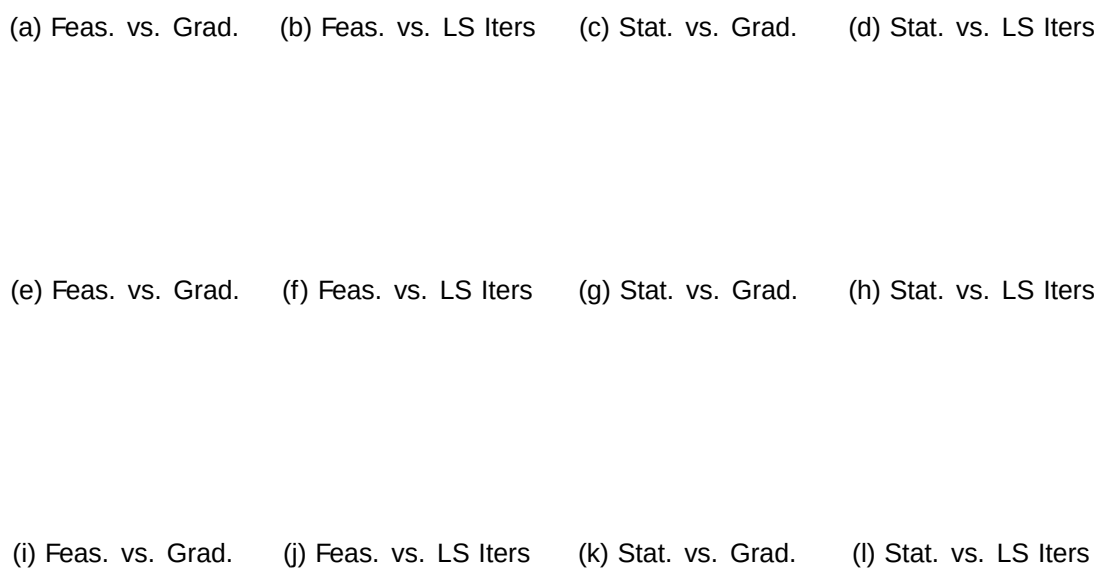
Next, we consider equality constrained problems from the CUTE collection of nonlinear optimization problems [11]. Specifically, of the 123 such problems in the collection we considered 49 problems. (We only used those for which: (i) LICQ held for all iterations of all algorithms, (ii) f is not a constant function, and (iii) $n + m \leq 1000$.) We used the prescribed starting point for all problems and all algorithms. The CUTE problems are deterministic, so we added noise to the gradient computations to make the problems stochastic. Specifically, we consider additive noise where the gradient was computed as

$$g_k = \frac{1}{|S_k|} \sum_{i \in S_k} (\nabla f_i(x_k) + \mathcal{N}(0, \sigma_{N,i} I)),$$

where $\sigma_{N,i} = 10^{-1}$ for all i, and S_k is prescribed by the variant and determines the level of noise.

For each problem, we again consider variants that compute exact and inexact (early termination conditions) linear system solutions. We compare PAIS-SQP to non-adaptive sampling variants with ($|S_k| = |S| \in \{2, 128, 1024\}$ for all $k \in \mathbb{N}$), and limit the maximum sample size employed by PAIS-SQP to 1024. For each problem, we ran 10 instances with different random seeds. This led to a total of 490 runs of each algorithm for each noise level. We terminated the methods on the following budget: $1024 \cdot 10^3$ gradient evaluations or $1024 \cdot 10^2$ linear system iterations (whichever comes first).

The results of these experiments are reported in Figure 4 in the form of performance profiles [27]. We present results in terms of feasibility and stationarity with respect to gradient evaluations and linear system iterations. The performance profiles were constructed as follows. For each problem, method and seed, the iterate used in the performance profile x_{pp} was chosen as: either the point with minimum $\|g_k + J_k^T y_k\|_1$ among all points with $\|c_k\|_1 \leq 10^{-6}$, or if no such point exists, then the point with minimum $\|c_k\|_1$. Following [27], for the two metrics an algorithm was deemed to have solved a given problem for a given seed if $m(x_0) - m(x_{pp}) \geq (1 - \tau_{pp})(m(x_0) - m(x_b))$, where $m(x_l)$ is $\|g_l + J_l^T y_l\|_1$ (for stationarity) and $\|c_l\|_1$ (for feasibility), respectively, $m(x_b)$ denotes the best possible value of either metric for each problem and seed, and tolerance $\tau_{pp} \in (0, 1)$. Overall, across all tolerances and metrics, the PAIS-SQP method appears to be the most robust (as seen by the right-most points on the figures). The ability of PAIS-SQP to make sufficient progress with inexact information in the intial stages of the optimization, combined with its ability to increase accuracy of the approximations employed, as needed, as the optimization progresses allows the algorithm to balance convergence and cost. As a result, PAIS-SQP is efficient and robust in term of all metrics.



(a) Feas. vs. Grad. (b) Feas. vs. LS Iters (c) Stat. vs. Grad. (d) Stat. vs. LS Iters

(e) Feas. vs. Grad. (f) Feas. vs. LS Iters (g) Stat. vs. Grad. (h) Stat. vs. LS Iters

(i) Feas. vs. Grad. (j) Feas. vs. LS Iters (k) Stat. vs. Grad. (l) Stat. vs. LS Iters

Figure 4: CUTE Performance profiles for exact and inexact variants of Algorithm 2 on CUTE collection. First row accuracy $_{pp} = 10^{-1}$; Second row accuracy $_{pp} = 10^{-3}$; Third row accuracy $_{pp} = 10^{-5}$.

7 Final Remarks

In this paper, we have designed and analyzed a stochastic SQP algorithm AIS-SQP for solving optimization problems involving deterministic nonlinear equality constraints and a stochastic objective function. At each iteration, the AIS-SQP method computes a stochastic approximation of the gradient of the objective function, computes a step by solving a stochastic Newton-SQP linear system inexactly, potentially updates the merit parameter, and adaptively selects a step size and updates the iterate. Our algorithm is adaptive in several ways. We have proposed accuracy conditions for the stochastic gradient approximation and the quality of the linear system solutions. Moreover, we have proposed adaptive update rules for the merit and step size parameters. Our algorithmic development and analysis has revealed an intrinsic relationship between the accuracy of stochastic objective gradient realizations, the quality of inexact solutions to the Newton-SQP linear systems, and the adaptive step sizes selected. That is, higher accuracy in the gradient estimation

and linear system solution can potentially lead to the acceptance of larger step sizes as in the deterministic counterparts.

We have proved that our algorithm generates a sequence of iterates whose first-order stationarity measure converges to zero in expectation. While similar results have been established in the literature, e.g., [5, 21], these works only consider asymptotic regimes after which the penalty parameter is sufficiently small and has stabilized. In this work, we have analyzed the complete behavior of the algorithm across all potential merit parameter changes, under an additional assumption, and have provided iteration complexity analysis for AIS-SQP, which matches that of deterministic SQP methods [21], in expectation. We have also established sublinear (gradient) sample complexity results of the proposed algorithm when the gradient and linear system accuracies are controlled at predetermined sublinear rates.

Inspired by AIS-SQP, we have developed, implemented and tested a practical variant PAIS-SQPof the adaptive stochastic SQP method. Our results on two different sets of experiments, constrained logistic regression and standard nonlinear optimization test problems, suggest that our practical algorithm strikes a good balance between minimizing constraint violation while also minimizing the objective function in terms of importance evaluation metrics such as iterations, gradient evaluations and linear system iterations.

Acknowledgements

We would like to thank Professor Frank E. Curtis for all his useful suggestions and feedback. Moreover, we would like to thank the Office of Naval Research for their support of this project (award number: N00014-21-1-2532).

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In International Conference on Machine Learning, pages 22{31. PMLR, 2017.

[2] Florian Beiser, Brendan Keith, Simon Urbainczyk, and Barbara Wohlmuth. Adaptive sampling strategies for risk-averse stochastic optimization with constraints. arXiv preprint arXiv:2012.03844, 2020.

[3] Albert S Berahas, Liyuan Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. SIAM Journal on Optimization , 31(2):1489{1518, 2021.

[4] 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-de cient jacobians. arXiv preprint arXiv:2106.13015 , 2021.

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

[6] 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.

[7] Dimitri Bertsekas. Network optimization: continuous and discrete models Athena Scienti c, 1998.

[8] John T Betts. Practical methods for optimal control and estimation using nonlinear programming. SIAM, 2010.

[9] Raghu Bollapragada, Richard Byrd, and Jorge Nocedal. Adaptive sampling strategies for stochastic optimization. SIAM Journal on Optimization , 28(4):3312{3343, 2018.

[10] Raghu Bollapragada, Jorge Nocedal, Dheevatsa Mudigere, Hao-Jun Shi, and Ping Tak Peter Tang. A progressive batching l-bfgs method for machine learning. In International Conference on Machine Learning, pages 620{629. PMLR, 2018.

[11] Ingrid Bongartz, Andrew R Conn, Nick Gould, and Ph L Toint. Cute: Constrained and unconstrained testing environment. ACM Transactions on Mathematical Software (TOMS) , 21(1):123{160, 1995.

[12] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. Mathematical programming, 134(1):127{155, 2012.

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

[14] Richard H Byrd, Frank E Curtis, and Jorge Nocedal. An inexact newton method for nonconvex equality constrained optimization. Mathematical programming, 122(2):273{ 299, 2010.

[15] Richard G Carter. On the global convergence of trust region algorithms using inexact gradient information. SIAM Journal on Numerical Analysis , 28(1):251{265, 1991.

[16] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Mathematical Programming, 169(2):337{375, 2018.

[17] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):1{27, 2011.

[18] Sou-Cheng T Choi, Christopher C Paige, and Michael A Saunders. Minres-qlp: A krylov subspace method for inde nite or singular symmetric systems. SIAM Journal on Scienti c Computing , 33(4):1810{1836, 2011.

[19] Andrew Cotter, Maya Gupta, and Jan Pfeifer. A light touch for heavily constrained sgd. In Conference on Learning Theory, pages 729{771. PMLR, 2016.

[20] Frank E Curtis, Jorge Nocedal, and Andreas Wächter. A matrix-free algorithm for equality constrained optimization problems with rank-de cient jacobians. SIAM Journal on Optimization , 20(3):1224{1249, 2010.

[21] 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.

[22] Frank E Curtis, Daniel P Robinson, and Baoyu Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. arXiv preprint arXiv:2107.03512 , 2021.

[23] Michael P Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data tting. SIAM Journal on Scienti c Computing , 34(3):A1380{A1405, 2012.

[24] Fatemeh S Hashemi, Soumyadip Ghosh, and Raghu Pasupathy. On adaptive sampling rules for stochastic recursions. In Proceedings of the Winter Simulation Conference 2014, pages 3959{3970. IEEE, 2014.

[25] Billy Jin, Katya Scheinberg, and Miaolan Xie. High probability complexity bounds for line search based on stochastic oracles. arXiv preprint arXiv:2106.06454 , 2021.

[26] Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, and Jinfeng Yi. Stochastic gradient descent with only one projection. Advances in neural information processing systems, 25:494{502, 2012.

[27] Jorge J Mor e and Stefan M Wild. Benchmarking derivative-free optimization algorithms. SIAM Journal on Optimization , 20(1):172{191, 2009.

[28] Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with di erentiable exact augmented lagrangians. arXiv preprint arXiv:2102.05320, 2021.

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

[30] Christopher C Paige and Michael A Saunders. Solution of sparse inde nite systems of linear equations. SIAM journal on numerical analysis, 12(4):617{629, 1975.

[31] Raghu Pasupathy, Peter Glynn, Soumyadip Ghosh, and Fatemeh S Hashemi. On sampling rates in simulation-based recursions. SIAM Journal on Optimization , 28(1):45{73, 2018.

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

[33] Tyrone Rees, H Sue Dollar, and Andrew J Wathen. Optimal solvers for pde-constrained optimization. SIAM Journal on Scienti c Computing , 32(1):271{298, 2010.

[34] Lloyd N Trefethen and David Bau III. Numerical linear algebra, volume 50. Siam, 1997.

[35] Yuchen Xie, Raghu Bollapragada, Richard Byrd, and Jorge Nocedal. Constrained and composite optimization via adaptive sampling methods. arXiv preprint arXiv:2012.15411, 2020.

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

A Technical Results (Section 4.2)

Here, we present the complete theoretical results for Algorithm 1 under Assumptions 4.27 and 4.28. We present all the technical lemmas required to establish the main theoretical results presented in Section 4.2. We unify the assumptions in the following assumption.

Assumption A.1. There exist universal quantities such that Assumptions 2.1, 3.1, 3.3, 4.27 and 4.28 hold for any realization of Algorithm 1.

Lemmas 4.5-4.8 from Section 4 hold without change. We unify them in the lemma below for completeness of the analysis under Assumption A.1.

Lemma A.2. Under Assumption A.1, for all $k \in \mathbb{N}$, (3.2) has a unique solution. In addition, for the same constants $(\bar{\gamma}_v, \bar{\gamma}_{uv}, \ldots, \bar{\gamma}_l) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ that appear in the respective Lemmas (listed in parenthesis below), the following statements hold true for all $k \in \mathbb{N}$:

a. (Lemma 4.5) The normal component v_k is bounded as $\max\{\|v_k\|_2, \|v_k\|_2^2\} \leq \bar{\gamma}_v \max\{\|c_k\|_2, \|r_k\|_2\}$.

b. (Lemma 4.6) If $\|u_k\|_2^2 \geq \bar{\gamma}_{uv}\|v_k\|_2^2$, then $d_k^T H_k d_k \geq \frac{1}{2}\|u_k\|_2^2$ and $d_k^T H_k d_k \geq \bar{\gamma}_d\|d_k\|_2^2$ where $\bar{\gamma}_d \in (0, \frac{\bar{\gamma}}{2})$ is an user-defined parameter in Algorithm 1.

c. (Lemma 4.7) The search direction satisfies $\|d_k\|_2^2 \leq \cdots \sigma_k$ and $\|d_k\|_2^2 + \|c_k\|_2 \leq (\cdots + 1)\sigma_k$.

d. (Lemma 4.8) The model reduction satisfies $l(x_k, \sigma_k, g_k, d_k) \geq \bar{\gamma}_l \sigma_k$.

Additionally, if Assumption 4.9 holds, then the following results also hold, whose proofs are from the corresponding lemmas in parentheses.

Lemma A.3. Under Assumptions A.1 and 4.9, the following statements hold true for all $k \in \mathbb{N}$ with the same constants appearing in the respective Lemmas (listed in parenthesis below).

a. (Lemma 4.11) There exists a constant $\sigma_{\min} \in \mathbb{R}_{>0}$ such that Algorithm 1 generates a sequence of $\{\sigma_k\}$, where $\sigma_k \geq \sigma_{\min}$.

b. (Lemma 4.12) There exist constants $\{\ldots, \bar{\gamma}_{l,d}\} \subset \mathbb{R}_{>0}$ such that for all $k \in \mathbb{N}$, $l(x_k, \sigma_k, g_k, d_k) \geq (\cdots \bar{L}_k + \cdots)\|d_k\|_2^2 \geq \bar{\gamma}_{l,d}\|d_k\|_2^2$.

c. (Lemma 4.14) For all $k \in \mathbb{N}$, there exists a constant $\bar{\ldots} \in \mathbb{R}_{>0}$ such that, $\ldots \leq \bar{\ldots}\sigma_k^{(2-\cdots)/2}$.

d. (Lemma 4.15) Finally, for all $k \in \mathbb{N}$, it follows that

$$\begin{aligned}
&\Phi(x_k + \cdots d_k, \sigma_k) - \Phi(x_k, \sigma_k) \\
&\leq \cdots \sigma_k\, l(x_k, \sigma_k, g_k, d_k) + (1 - \cdots)\sigma_k^{(\cdots/2 - 1)}\, l(x_k, \sigma_k, g_k, d_k) \\
&+ \cdots \sigma_k g_k^T(d_k - \hat{d}_k) + \cdots(\cdots - \sigma_k) g_k^T d_k + \cdots \sigma_k \|J_k(d_k - \hat{d}_k)\|_1.
\end{aligned} \tag{A.1}$$

We now proceed to state and prove a series of Lemmas that are analogues to those proven in Section 4.1. For brevity, in the rest of this appendix, we always assume that Assumptions A.1 and 4.9 hold without explicitly stating this fact. We begin by bounding the variance in the stochastic gradient approximations and the variance in the search directions computed in Lemmas A.4 and A.5.

Lemma A.4. For all $k \in \mathbb{N}$, $\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2] \leq \frac{\sqrt{\nu_1}\,\sigma}{(k+1)^{1/2}}$.

Proof. By Assumption 4.27 and Jensen's inequality, we have

$$\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2] \leq \sqrt{\mathbb{E}_k\left[\|g_k - \nabla f(x_k)\|_2^2\right]} \leq \frac{\sqrt{\nu_1}\,\sigma}{(k+1)^{1/2}};$$

which completes the proof. □

Lemma A.5. For all k, $\mathbb{E}_k[\tilde{d}_k] = d_k$, $\mathbb{E}_k[\tilde{u}_k] = u_k$, and $\mathbb{E}_k[\tilde{\gamma}_k] = \gamma_k$. Moreover, there exists $L \in \mathbb{R}_{>0}$, independent of k, such that $\|\tilde{d}_k - d_k\|_2 \leq L\|g_k - \nabla f(x_k)\|_2$ and $\mathbb{E}_k\left[\|\tilde{d}_k - d_k\|_2\right] \leq \frac{\sigma}{(k+1)^{1/2}}$, where $\sigma = L\sqrt{\nu_1} \in \mathbb{R}_{>0}$.

Proof. The first two statements follow the same arguments as in the proof of Lemma 4.17. By (3.2) and Lemma A.4,

$$\mathbb{E}_k\left[\|\tilde{d}_k - d_k\|_2\right] \leq \mathbb{E}_k[L\|g_k - \nabla f(x_k)\|_2] = L\,\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2] \leq L\frac{\sqrt{\nu_1}}{(k+1)^{1/2}};$$

which proves the last statement. □

Similar to Lemma 4.18, the next lemma provides bounds on the differences between stochastic and deterministic gradient approximations and exact and inexact search directions.

Lemma A.6. For all $k \in \mathbb{N}$,

$$
\begin{aligned}
|g_k^T(d_k - \tilde{d}_k)| &\le \sigma_{g,\tilde{d}}^{p-} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)}; \\
|g_k^T \tilde{d}_k - g_k^T d_k| &\le \sigma_{g,\tilde{d}}^{p-} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + L \kappa_{g_k} \|g_k\|_2^2; \\
|g_k^T(d_k - \tilde{d}_k)| &\le \sigma_{g,\tilde{d}}^{p-} = 2 \frac{\sqrt{l(x_k, \epsilon_k; g_k; d_k)}}{(k+1)} + \sigma_{g,\tilde{d}}^{p-} \frac{\sqrt{2} = 2}{(k+1)} \kappa_{g_k} \|g_k\|_2 \\
&\quad + \sigma_{g,\tilde{d}} = 2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)}; \\
|g_k^T(d_k - d_k)| &\le \sigma_{g,\tilde{d}}^{p-} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} \\
&\quad + \sigma_{g,\tilde{d}}^{p-} = 2 \frac{\sqrt{l(x_k, \epsilon_k; g_k; d_k)}}{(k+1)} + \sigma_{g,\tilde{d}} = 2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)}; \\
|g_k^T d_k - g_k^T d_k| &\le \sigma_{gg,\tilde{d}\tilde{d}}^{p-} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + L \kappa_{g_k} \|g_k\|_2^2 \\
&\quad + \sigma_{g,\tilde{d}}^{p-} = 2 \frac{\sqrt{l(x_k, \epsilon_k; g_k; d_k)}}{(k+1)} + \sigma_{g,\tilde{d}}^{p-} \frac{\sqrt{2} = 2}{(k+1)} \kappa_{g_k} \|g_k\|_2 \\
&\quad + \sigma_{g,\tilde{d}} = 2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)}; \\
\text{and } \|J_k(d_k - \tilde{d}_k)\|_1 &\le \sigma_{J,\tilde{d}} = 2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)};
\end{aligned}
$$

where $\sigma_{g,\tilde{d}}^{p-} = \frac{H L}{\sqrt{2} \sigma_{l,d}} \in \mathbb{R}_{>0}$, $\sigma_{gg,\tilde{d}\tilde{d}}^{p-} = \sigma_{g,\tilde{d}}^{p-} + \frac{1}{\sqrt{2} \sigma_{l,d}} \in \mathbb{R}_{>0}$, $\sigma_{g,\tilde{d}}^{p-} = H \sqrt{\frac{2}{\sigma_{l,d}}} \in \mathbb{R}_{>0}$, $\sigma_{g,\tilde{d}} = \gamma \| a \in \mathbb{R}_{>0}$, $\sigma_{g,\tilde{d}}^{p-} = H \sqrt{\frac{2}{\sigma_{l,d}}} \in \mathbb{R}_{>0}$, $\sigma_{g,\tilde{d}} = \gamma \| a \in \mathbb{R}_{>0}$, and $\sigma_{J,\tilde{d}} = \| a \in \mathbb{R}_{>0}$.

Proof. (First inequality) By Assumption 3.1, Lemma A.5 and (3.2), (3.11), (4.3),

$$
\begin{aligned}
|g_k^T(d_k - \tilde{d}_k)| &= |(g_k + J_k^T(y_k + \epsilon_k))^T(d_k - \tilde{d}_k)| \\
&= |(H_k d_k)^T(d_k - \tilde{d}_k)| \le H \|d_k\|_2 \|d_k - \tilde{d}_k\|_2 \\
&\quad H \sqrt{\frac{l(x_k, \epsilon_k; g_k; d_k)}{\sigma_{l,d}}} L \kappa_{g_k} \|g_k\|_2;
\end{aligned}
\tag{A.2}
$$

where the result follows using the definition of $\sigma_{g,\tilde{d}}^{p-}$.

(Second inequality) By the Cauchy–Schwarz inequality, the triangle inequality, Assumption 3.1, (3.2), (3.11), (4.3) and (A.2), and Lemma A.5, it follows that

$$
\begin{aligned}
&|g_k^T \tilde{d}_k - g_k^T d_k| \\
&\le |(g_k - g_k)^T d_k| + |g_k^T(\tilde{d}_k - d_k)| + |(g_k - g_k)^T(\tilde{d}_k - d_k)| \\
&\le \|g_k - g_k\|_2 \|d_k\|_2 + |g_k^T(\tilde{d}_k - d_k)| + \|g_k - g_k\|_2 \|\tilde{d}_k - d_k\|_2 \\
&\le \sigma_{gg,\tilde{d}\tilde{d}}^{p-} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + |g_k^T(\tilde{d}_k - d_k)| + \kappa_{g_k} \|g_k\|_2 \|\tilde{d}_k - d_k\|_2 \\
&\le \sigma_{g,\tilde{d}}^{p-} = 2 \frac{\sqrt{l(x_k, \epsilon_k; g_k; d_k)}}{(k+1)} + \sigma_{g,\tilde{d}}^{p-} \frac{\sqrt{2} = 2}{(k+1)} \kappa_{g_k} \|g_k\|_2 \\
&\quad + \sigma_{g,\tilde{d}} = 2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + \kappa_{g_k} \|g_k\|_2^2 \\
&\le \sigma_{g,\tilde{d}}^{p-} + \frac{1}{\sqrt{2} \sigma_{l,d}} \kappa_{g_k} \|g_k\|_2 \sqrt{l(x_k, \epsilon_k; g_k; d_k)} + L \kappa_{g_k} \|g_k\|_2^2;
\end{aligned}
\tag{A.3}
$$

where the result follows using the definitions of $\sigma_{gg,\tilde{d}\tilde{d}}^{p-}$.

(Third inequality) By Assumption 3.1, Lemmas 3.7, 3.8 and A.5, (3.2), (3.3), (3.11), and (4.3), we have for all $k \in \mathcal{N}$ that

$$\begin{aligned}
\|g_k^T(d_k - \tilde{d}_k)\| &= \|(g_k + J_k^T(y_k + \cdot_k) - J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| \\
&\quad \|(g_k + J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| + \|(J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| \\
&\quad \|(g_k + J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| + \|(g_k - g_k)^T(d_k - \tilde{d}_k)\| \\
&\quad + \|(J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| \\
&= \|(H_k d_k)^T(d_k - \tilde{d}_k)\| + \|(g_k - g_k)^T(d_k - \tilde{d}_k)\| + \|(y_k + \cdot_k)^T r_k\| \\
&\quad \|(H_k d_k)^T(d_k - \tilde{d}_k)\| + \|g_k - g_k\|_2 \|d_k - \tilde{d}_k\|_2 + \|y_k + \cdot_k\|_1 \cdot r_k\|_1 \\
&\quad \sqrt{\frac{l(x_k; \cdot_k; g_k; d_k)}{l;d}} \frac{\sqrt{2}p}{\sqrt{(k+1)}} + \frac{\sqrt{2}p}{\sqrt{(k+1)}} \|g_k - g_k\|_2 \\
&\quad + \cdot_y ! \cdot_a \sqrt{2} l(x_k; \cdot_k; g_k; d_k);
\end{aligned} \tag{A.4}$$

where the result follows using the definitions of $\cdot_{g;d\tilde{d};p-}$ and $\cdot_{g;d\tilde{d}}$.

(Fourth inequality) By Assumption 3.1, Lemmas 3.7 and 3.8, (3.2), (3.3), (3.11) and (4.3), it follows that

$$\begin{aligned}
\|g_k^T(d_k - \tilde{d}_k)\| &= \|(g_k + J_k^T(y_k + \cdot_k) - J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| \\
&\quad \|(g_k + J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| + \|(J_k^T(y_k + \cdot_k))^T(d_k - \tilde{d}_k)\| \\
&\quad \|(H_k d_k)^T(d_k - \tilde{d}_k)\| + \|y_k + \cdot_k\|_1 \cdot r_k\|_1 \\
&\quad \sqrt{\frac{l(x_k; \cdot_k; g_k; d_k)}{l;d}} \sqrt{2} \frac{p}{\sqrt{(k+1)}} + \cdot_y ! \cdot_a \sqrt{2} l(x_k; \cdot_k; g_k; d_k):
\end{aligned}$$

By the triangle inequality and (A.2),

$$\begin{aligned}
\|g_k^T(d_k - d_k)\| &\le \|g_k^T(\tilde{d}_k - d_k)\| + \|g_k^T(d_k - \tilde{d}_k)\| \\
&\quad \cdot_{g;d\tilde{d};p-} \|g_k - g_k\|_2 \sqrt{l(x_k; \cdot_k; g_k; d_k)} \\
&\quad + \sqrt{\frac{2}{l;d}} \sqrt{2} \frac{p}{\sqrt{(k+1)}} + \cdot_y ! \cdot_a \sqrt{2} l(x_k; \cdot_k; g_k; d_k);
\end{aligned}$$

where the result follows using the definition of $\cdot_{g;dd;p-}$ and $\cdot_{g;dd}$.

(Fifth inequality) By (A.3), (A.4),

$$\begin{aligned}
\|g_k^T d_k - g_k^T d_k\| &\le \|g_k^T d_k - g_k^T \tilde{d}_k\| + \|g_k^T(\tilde{d}_k - d_k)\| \\
&\quad \cdot_{gg;d\tilde{d};p-} \|g_k - g_k\|_2 \sqrt{l(x_k; \cdot_k; g_k; d_k)} + \cdot_L \|g_k - g_k\|_2^2 \\
&\quad + \cdot_{g;d\tilde{d};p-} \sqrt{2} \frac{p}{\sqrt{(k+1)}} + \frac{\sqrt{p}}{\sqrt{2}} \sqrt{2} \|g_k - g_k\|_2 \\
&\quad + \cdot_{g;d\tilde{d}} \sqrt{2} l(x_k; \cdot_k; g_k; d_k):
\end{aligned}$$

(Sixth inequality) By Lemma 3.7, and (3.2), (3.3),

$$\|J_k(d_k - \hat{d}_k)\|_1 = \|r_k\|_1 \quad ! \; a \quad ^{=2} \; l(x_k; \; _k; g_k; d_k);$$

where the result follows using the de nition of $_{J;\hat{d}}$. □

The next lemma provides a useful upper bound on the di erence between the deterministic and stochastic merit parameters.

Lemma A.7. For all $k \; 2 \; N$, $|(\; _k \quad _k)g_k^T d_k| \quad ^{=2} \; l(x_k; \; _k; g_k; d_k)$, where $= 2 \; _3 \quad _1 \; _{g;\hat{d}; \; l}$ and $2 \; 0; \frac{1}{(2 \; _3)^{2=}}$.

Proof. The proof of this statement is identical to the proof of Lemma 4.20. □

Similar to Lemma 4.21, the next lemma bounds the stochastic model reduction function with respect to its deterministic counterpart (with additional terms).

Lemma A.8. For all $k \; 2 \; N$, by choosing $2 \; 0; \frac{1}{(2 \; _1 \; _{g;\hat{d}})^{2=}}$, it follows that

$$\begin{aligned} l(x_k; \; _k; g_k; d_k) \quad & (1 + \; _{l; \; l} \quad ^{=2}) \; l(x_k; \; _k; g_k; d_k) \\ & + 2 \quad _1 \quad _L \|g_k\| \quad g_k\|_2^2 + \; _{g;\hat{d}; \; l} - \frac{^{=2}\sqrt{l(x_k; \; _k; g_k; d_k)}}{\sqrt{(k+1)}} \\ & + \; _{gg;\hat{d}; \; l} - \|g_k\| \quad g_k\|_2 \sqrt{l(x_k; \; _k; g_k; d_k)} + \frac{\sqrt{2} \; ^{=2} \|g_k\| \quad g_k\|_2}{\sqrt{(k+1)}} \; ; \end{aligned}$$

where $_{l; \; l} = 2(\quad + \; _1 \; _{g;\hat{d}}) \; 2 \; R_{>0}$. Additionally, under Assumption 4.1, for all $k \; 2 \; N$

$$\begin{aligned} E_k \quad l(x_k; \; _k; g_k; d_k) \quad & (1 + \; _{l; \; l} \quad ^{=2}) \; l(x_k; \; _k; g_k; d_k) \\ & + \; _{l; \; l; \; l} - \quad ^{=2} \frac{\sqrt{l(x_k; \; _k; g_k; d_k)}}{\sqrt{(k+1)}} + \; _{l; \; l; \; l} \quad \frac{1}{(k+1)} \; ; \end{aligned}$$

where $_{l; \; l} = \; _{l; \; l} \sqrt{2} \; R_{>0}$, $_{l; \; l; \; l} - \; = 2 \quad _1(\; _{g;\hat{d}; \; l} - + \; _{gg;\hat{d}; \; l} - \sqrt{\quad _1}) \; 2 \; R_{>0}$, and $_{l; \; l; \; l} \; = 2 \quad _1(\; _L \; _1 + \sqrt{\quad _1 \; 2}) \; 2 \; R_{>0}$.

Proof. By (3.5), and Lemmas A.6 and A.7, it follows that

$$\begin{aligned} l(x_k; \; _k; g_k; d_k) = \quad & \; _k g_k^T d_k + \|c_k\|_1 \quad \|c_k\| + \|J_k d_k\|_1 \\ & \; _k g_k^T d_k + \|c_k\|_1 + |(\; _k \quad _k)g_k^T d_k| + \; _k|(g_k^T d_k \quad g_k^T d_k)| \\ & l(x_k; \; _k; g_k; d_k) + |(\; _k \quad _k)g_k^T d_k| + \; _1|g_k^T d_k \quad g_k^T d_k| \\ & l(x_k; \; _k; g_k; d_k) + \quad ^{=2} \; l(x_k; \; _k; g_k; d_k) \\ & + \quad _1 \; _{gg;\hat{d}; \; l} - \|g_k\| \quad g_k\|_2 \sqrt{l(x_k; \; _k; g_k; d_k)} + \; _L \|g_k\| \quad g_k\|_2^2 \\ & + \; _{g;\hat{d}; \; l} - \frac{^{=2}\sqrt{l(x_k; \; _k; g_k; d_k)}}{\sqrt{(k+1)}} + \frac{\sqrt{2} \; ^{=2} \|g_k\| \quad g_k\|_2}{\sqrt{(k+1)}} + \; _{g;\hat{d}} \quad ^{=2} \; l(x_k; \; _k; g_k; d_k) \quad : \end{aligned}$$

The first result follows by re-arranging the above, invoking the restriction on , and the definition of $\bar{\ }_{l;\ l}$.

Taking the conditional expectation, by Assumption 4.27 and Lemma A.4,

$$\begin{aligned} E_k\quad l(x_k;\ _k;g_k;d_k) \qquad & 1+2(\quad +\ _{1\ g;dd})^{\ =2}\quad l(x_k;\ _k;g_k;d_k) \\ & +2\quad _1\ \frac{L\ 1}{(k+1)}+\ _{g;dd;}{}^{p\ -}\frac{^{=2}p\ \overline{l(x_k;\ _k;g_k;d_k)}}{p\ \overline{(k+1)}} \\ & +\ _{gg;dd;}{}^{p\ -}\frac{p\ \overline{\ }}{1}\frac{^{=2}p\ \overline{l(x_k;\ _k;g_k;d_k)}}{p\ \overline{(k+1)}}+\frac{p\ \overline{\ }}{(k+1)}\quad : \end{aligned}$$

Using the definitions of $\bar{\ }_{l;\ l}$, $\bar{\ }_{l;\ l}$, $\bar{\ }_{l;\ l;}{}^{p\ -}$, and $\bar{\ }_{l;\ l;}$, completes the proof. □

Finally, we restate and prove the theoretical results stated in Section 4.2. Specifically, we state and prove Lemma 4.29, Theorem 4.30 and Corollary 4.31.

Lemma A.9. (Lemma 4.29) For all $k \in \mathbb{N}$,

$$\begin{aligned} E_k[\ (x_{k+1};\ _{k+1})\quad (x_k;\ _k)] \quad E_k[(\ _{k+1}\quad _k)]\ f_{inf}\quad (\_\quad)\quad l(x_k;\ _k;g_k;d_k) \\ +\ ^2\frac{;}{(k+1)}: \end{aligned}$$

where $\quad =\quad +\frac{;^{\ p\ -}}{2}\ 2\ R_{>0}$ and $_{;}\ =\ _{;}\ +\frac{;^{\ p\ -}}{2}\ 2\ R_{>0}$, and

$$\begin{aligned} & =\ _u\ (1\quad)\ \bar{\ }_{l;\ l}+\quad +(\ _{1\ g;dd}+\ _{J;dd})(1+\ \bar{\ }_{l;\ l})\ \ 2\ R_{>0}; \\ _{;}\ & =\ _u\ 2\ _1(1\quad)(\ _L\ _1+\frac{p\ \overline{\ }}{1\ 2})+(\ _{1\ g;dd}+\ _{J;dd})\ \bar{\ }_{l;\ l;}\quad 2\ R_{>0}; \\ _{;}{}^{p\ -} & =\ _u\ 2\ _1(1\quad)\ _{g;dd;}{}^{p\ -}+\ _1\ _{gg;dd;}{}^{p\ -} \\ & \quad +\ _1\ \frac{p\ \overline{\ }}{1}\ _{g;dd;}{}^{p\ -}+\ _{g;dd;}{}^{p\ -}+(\ _{1\ g;dd}+\ _{J;dd})\ \bar{\ }_{l;\ l;}{}^{p\ -}\ \ 2\ R_{>0}: \end{aligned}$$

Proof. By Lemmas A.3, A.7, A.6, A.8 and 4.14, and (3.4), it follows that for all $k \in \mathbb{N}$

$$
\begin{aligned}
& E_k[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] \\
= \ & E_k[(\ _{k+1} \quad _k) f_{k+1}] + E_k[\ (x_{k+1};\ _k) \quad (x_k;\ _k)] \\
& E_k[(\ _{k+1} \quad _k)]\, f_{\inf} \\
& \quad E_k \quad _k \ l(x_k;\ _k; g_k; d_k) \quad (1 \quad) \ _k \ ^{(\ =2\ \ 1)} \ l(x_k;\ _k; g_k; d_k) \\
& + E_k \quad _k \ _k g_k^T (d_k \quad d_k) + E_k \quad _k (\ _k \quad _k) g_k^T d_k + E_k \quad _k k J_k (d_k \quad \tilde{d}_k) k_1 \\
& E_k[(\ _{k+1} \quad _k)]\, f_{\inf} \quad E_k[\ _k \ l(x_k;\ _k; g_k; d_k)] \\
& + E_k \ (1 \quad) \ _k \ ^{(\ =2\ \ 1)} \ (1 + \ \overline{}_{l;\ l} \ ^{=2}) \ l(x_k;\ _k; g_k; d_k) \\
& \quad + 2 \quad _1 \quad _L k g_k \quad g_k k_2^2 + \ _{g;dd;} p - \frac{^{=2} p \ \overline{l(x_k;\ _k; g_k; d_k)}}{(k+1)} \\
& \quad + \ _{gg;dd;} p - k g_k \quad g_k k_2 \ \sqrt{l(x_k;\ _k; g_k; d_k)} + \frac{p \ \overline{}_2 \ p^{=2} k g_k \quad g_k k_2}{(k+1)} \\
& + \ _u \ ^{(2 \ \ =2)} \ _1 E_k \ j g_k^T (d_k \quad d_k) j \ + \ _u \ ^{(2 \ \ =2)} E_k \ j(\ _k \quad _k) g_k^T d_k j \\
& + \ _u \ ^{(2 \ \ =2)} E_k \ k J_k (d_k \quad \tilde{d}_k) k_1 \\
& E_k[(\ _{k+1} \quad _k)]\, f_{\inf} \quad E_k[\ _k \ l(x_k;\ _k; g_k; d_k)] \\
& + (1 \quad) \ \overline{}_{l;\ l} \ _u \ ^{(1+\ =2)} \ l(x_k;\ _k; g_k; d_k) \\
& + 2 \quad _1 (1 \quad) \ _u \quad _L \frac{1}{(k+1)} + \ _{g;dd;} p - \frac{^{=2} p \ \overline{l(x_k;\ _k; g_k; d_k)}}{(k+1)} \\
& \quad + \ _{gg;dd;} p - p \frac{p \ \overline{}_1 \ ^{=2}}{(k+1)} \ \sqrt{l(x_k;\ _k; g_k; d_k)} + \frac{p \ \overline{}_{1 \ \ 2}}{(k+1)} \\
& + \ _u \ ^{(2 \ \ =2)} \ _1 \quad _{g;dd;} p - \frac{p \ \overline{}_1 \ ^{=2} p \ \overline{l(x_k;\ _k; g_k; d_k)}}{(k+1)} + \ _{g;dd;} p - \frac{^{=2} p \ \overline{l(x_k;\ _k; g_k; d_k)}}{(k+1)} \\
& \quad + \ _{g;dd} \ ^{=2} E_k \quad l(x_k;\ _k; g_k; d_k)
\end{aligned}
$$

Continuing from the above, by Lemmas A.8 and 4.14, it follows that for all $k \in \mathbb{N}$

$$
\begin{aligned}
&\mathbb{E}_k\left[\Phi(x_{k+1}, \theta_{k+1}) - \Phi(x_k, \theta_k)\right] \\
&\mathbb{E}_k\left[\Phi(\theta_{k+1} - \theta_k)\right] \leq f_{\inf} \\
&\quad - \alpha I(x_k, \theta_k, g_k, d_k) + (1-\alpha)\bar{\lambda}_{i,j} u^{(1+\gamma/2)} I(x_k, \theta_k, g_k, d_k) \\
&\quad + 2\alpha_1(1-\alpha) u^{(1+\gamma/2)} \|g_{\mathcal{D}}\|_p + \sqrt{\alpha_1 \|g_{\mathcal{D}}\|_p} \sqrt{\frac{I(x_k, \theta_k, g_k, d_k)}{(k+1)}} \\
&\quad + \frac{L\alpha_1 + \sqrt{\alpha_1 \alpha_2}}{(k+1)^{\gamma/2}} \\
&\quad + u^2 \sqrt{\alpha_1 \|g_{\mathcal{D}}\|_p} \sqrt{\frac{I(x_k, \theta_k, g_k, d_k)}{(k+1)}} + u^2 I(x_k, \theta_k, g_k, d_k) \\
&\quad + u^2(\alpha_1 \|g_{\mathcal{D}}\| + \|J_{\mathcal{D}}\|)(1 + \bar{\lambda}_{i,j}^{\gamma/2}) I(x_k, \theta_k, g_k, d_k) \\
&\quad + \bar{\lambda}_{i,j} \|I\|_p \sqrt{\frac{I(x_k, \theta_k, g_k, d_k)}{(k+1)}} + \bar{\lambda}_{i,j} \|I\| \frac{1}{(k+1)} \\
&\mathbb{E}_k\left[\Phi(\theta_{k+1} - \theta_k)\right] \leq f_{\inf} - \alpha I(x_k, \theta_k, g_k, d_k) \\
&\quad + u(1-\alpha)\bar{\lambda}_{i,j} + u^2(\alpha_1 \|g_{\mathcal{D}}\| + \|J_{\mathcal{D}}\|)(1 + \bar{\lambda}_{i,j}^{\gamma/2}) I(x_k, \theta_k, g_k, d_k) \\
&\quad + u^2 2\alpha_1(1-\alpha) \|g_{\mathcal{D}}\|_p + \sqrt{\alpha_1 \|g_{\mathcal{D}}\|_p} + \alpha_1 \sqrt{\alpha_1 \|g_{\mathcal{D}}\|_p} + \|g_{\mathcal{D}}\|_p \\
&\quad + (\alpha_1 \|g_{\mathcal{D}}\| + \|J_{\mathcal{D}}\|)\bar{\lambda}_{i,j}^{\gamma/2} \sqrt{\frac{I(x_k, \theta_k, g_k, d_k)}{(k+1)}} \\
&\quad + u^2 2\alpha_1(1-\alpha)(L\alpha_1 + \sqrt{\alpha_1 \alpha_2}) + (\alpha_1 \|g_{\mathcal{D}}\| + \|J_{\mathcal{D}}\|)\bar{\lambda}_{i,j}^{\gamma/2} \frac{1}{(k+1)} \\
&= \mathbb{E}_k\left[\Phi(\theta_{k+1} - \theta_k)\right] \leq f_{\inf} - \alpha I(x_k, \theta_k, g_k, d_k) + u^2 I(x_k, \theta_k, g_k, d_k) \\
&\quad + \sigma^{\gamma/2} \sqrt{\frac{I(x_k, \theta_k, g_k, d_k)}{(k+1)}} + \sigma \frac{1}{(k+1)} \\
&\mathbb{E}_k\left[\Phi(\theta_{k+1} - \theta_k)\right] \leq f_{\inf} - \alpha I(x_k, \theta_k, g_k, d_k) + \frac{\sigma^{\gamma/2}}{2} u^2 I(x_k, \theta_k, g_k, d_k) \\
&\quad + \sigma + \frac{\sigma^{\gamma/2}}{2} u^2 \frac{1}{(k+1)};
\end{aligned}
$$

where the result follows using the definitions of α, $\sigma^{\gamma/2}$, σ, and σ. □

We are now ready to prove the main theorem of this section.

Theorem A.10. (Theorem 4.30) For some $\alpha \in (0,1)$ and $\gamma > 1$, and choosing $u \in \left(0, \min\left\{\frac{1}{(2\alpha\sigma)^{2/\gamma}}, \frac{(1-\alpha)\bar{\lambda}}{\sigma}, \frac{1}{(2\alpha_1 \|g_{\mathcal{D}}\|)^{2/\gamma}}\right\}\right)$, it follows that

$$
\lim_{k \to \infty} \mathbb{E}\left[\sum_{j=0}^{k} I(x_j, \theta_j, g_j, d_j)\right] < \infty;
$$

from which it follows that $\lim_{k!1} \quad E[\ \ l(x_k;\ _k;g_k;d_k)] = 0$.

Proof. By Lemma A.9 and $\quad 2 \quad 0;\frac{(1 \quad)}{\quad} =$, it follows that

$$\begin{aligned} & E_k[\ (x_{k+1};\ _{k+1}) \quad (x_k;\ _k)] \\ & E_k[(\ _{k+1} \quad _k)] f_{inf} \quad (\_ \quad) \ \ l(x_k;\ _k;g_k;d_k) + \ ^2\frac{;}{(k+1)} \\ & E_k[(\ _{k+1} \quad _k)] f_{inf} \quad \_ \quad l(x_k;\ _k;g_k;d_k) + \ ^2\frac{;}{(k+1)}: \end{aligned} \tag{A.5}$$

Applying a telescopic sum to (A.5) and taking the total expectation, it follows that

$$\begin{aligned} 1 \quad < \quad & _{inf} \quad (x_0;\ _0) \quad E[\ (x_k;\ _k) \quad (x_0;\ _0)] \\ = \ & E\Big[\sum_{j=0}^{k\ 1} (\ (x_{j+1};\ _{j+1}) \quad (x_j;\ _j))\Big] \\ & E\Big[\sum_{j=0}^{k\ 1} (\ _{j+1} \quad _j) f_{inf}\Big] \quad \_ \quad \sum_{j=0}^{k\ 1} l(x_j;\ _j;g_j;d_j) + \sum_{j=0}^{k\ 1} \ ^2\frac{;}{(j+1)}\Big] \\ & _1 j f_{inf} j \quad \_ \quad E\Big[\sum_{j=0}^{k\ 1} l(x_j;\ _j;g_j;d_j)\Big] + \ ^2\sum_{j=0}^{k\ 1}\frac{;}{(j+1)}: \end{aligned}$$

Finally, using the fact that $\sum_{j=0}^{k\ 1} \ ^2\frac{;}{(j+1)} < 1$ for any > 1, we may complete the proof. □

Corollary A.11. Under Assumptions A.1 and 4.9, and the conditions of Theorem A.10, Algorithm 1 yields a sequence of iterates $f(x_k;y_k)g$ for which

$$\lim_{k!1} E\ \ kd_k k_2^2 = 0; \quad \lim_{k!1} E[kc_k k_2] = 0; \quad \text{and} \quad \lim_{k!1} E\ \ kg_k + J_k^T(y_k + \ _k)k_2 = 0:$$

Proof. The proof of this corollary follows the exact same arguments as the proof of Corollary 4.24 (Section 4). □

The nal result we show in this section is a complexity result for our proposed algorithm, i.e., the number of iterations and the total number of stochastic gradient evaluations required to achieved an -accurate solution in expectation. Speci cally, to measure the complexity of our algorithm, we consider the minimum number of iterations, K , and the minimum total number of stochastic gradient evaluations, W , required to achieve the following approximate stationary measure

$$E[kg_k + J_k^T(y_k + \ _k)k_2] \quad _L \quad \text{and} \quad E[kc_k k_1] \quad _c; \tag{A.6}$$

for $_L \ 2\ (0;1)$ and $_c \ 2\ (0;1)$.

Corollary A.12. (Corollary 4.31) Under the conditions of Theorem 4.30, Algorithm 1 generates an iterate $\{(x_k, y_k)\}$ that satisfies (A.6) in at most $K = O\left(\max\left\{\frac{L^2}{\epsilon}, \frac{c}{\epsilon}\right\}^{\frac{1}{\alpha}}\right)$ iterations and $W = O\left(\max\left\{\frac{L^2}{\epsilon}, \frac{c}{\epsilon}\right\}^{\frac{(\beta+1)}{\alpha}}\right)$, $\alpha \in (1, 2]$, $\beta > -1$, stochastic gradient evaluations. Moreover, if $L = \alpha$ and $c = \epsilon^2$, then $K = O\left(\epsilon^{-\frac{2}{\alpha}}\right)$ and $W = O\left(\epsilon^{-\frac{2(\beta+1)}{\alpha}}\right)$.

Proof. Using the same logic as the proof of Corollary 4.26, we know that if (A.6) is violated, then for all $k \in \{0, \ldots, K-1\}$

$$\mathbb{E}\left[\Delta_l(x_k, \eta_k, g_k, d_k)\right] \leq -\eta_x \min\left\{\frac{\epsilon^2}{L}, c\epsilon\right\}, \tag{A.7}$$

where $\eta_x = \min\left\{\frac{1}{L}, \frac{\min\{|g_k^T d_k|, 1\}}{d_k^T H_k d_k}\right\} \in \mathbb{R}_{>0}$. Re-arranging terms in (A.5), and using (3.4) and (A.7) to do a telescoping, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\alpha_1(f(x_0) - f_{inf}) + \alpha\beta c_0 k_1\right] \leq \mathbb{E}\left[\alpha_0(f(x_0) - f_{inf}) + \alpha\beta c_0 k_1\right] \\
&\mathbb{E}\left[\alpha_0 f(x_0) + \alpha\beta c_0 k_1 - \alpha_K f(x_K) - \alpha\beta c_K k_1 + (\alpha_K - \alpha_0) f_{inf}\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K-1} \left(\alpha_k f(x_k) + \alpha\beta c_k k_1 - \alpha_{k+1} f(x_{k+1}) - \alpha\beta c_{k+1} k_1 + (\alpha_{k+1} - \alpha_k) f_{inf}\right)\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K-1} \left(\Delta_l(x_k, \eta_k) - \Delta_l(x_{k+1}, \eta_{k+1}) + (\alpha_{k+1} - \alpha_k) f_{inf}\right)\right] \\
&\mathbb{E}\left[\sum_{k=0}^{K-1} -\Delta_l(x_k, \eta_k, g_k, d_k) \leq \epsilon^2 \frac{\gamma}{(k+1)}\right] \\
&\leq -\eta_x \min\left\{\frac{\epsilon^2}{L}, c\epsilon\right\} \epsilon^2 \frac{\gamma}{(k+1)} \\
&\leq -\eta_x \min\left\{\frac{\epsilon^2}{L}, c\epsilon\right\} K \epsilon^2 \gamma \sum_{k=0}^{K-1} \frac{1}{(k+1)}.
\end{aligned}$$

It further implies that K is bounded as

$$K \leq \frac{\alpha_1(f(x_0) - f_{inf}) + \alpha\beta c_0 k_1 + \epsilon^2 \gamma \sum_{k=0}^{K-1} \frac{1}{(k+1)}}{\eta_x \min\left\{\frac{\epsilon^2}{L}, c\epsilon\right\}} = O(\min\{\epsilon^{-2}, \epsilon^{-1}\}) \tag{A.8}$$

under the condition that $\gamma \in \mathbb{R}_{>1}$. Next, we analyze the sampling complexity. Suppose, we use $|S_k|$ samples to estimate the stochastic gradient g_k, then,

$$\mathbb{E}_k \left\| g_k - \nabla f_k \right\|_2^2 = \frac{\sigma^2}{|S_k|}$$

where σ^2 is the population variance. Therefore, the minimum number of samples required to satisfy Assumption 4.27 is

$$|S_k| = \frac{\sigma^2 (k+1)^{\alpha}}{\epsilon^{\alpha}}.$$

By (A.8), it follows that the total number of stochastic gradient evaluations required to satisfy (A.6) can be expressed as

$$W = \sum_{k=0}^{K(\epsilon)-1} |S_k| = \sum_{k=0}^{K(\epsilon)-1} \frac{\sigma^2(k+1)}{\epsilon_1} = \frac{\sigma^2}{\epsilon_1} \sum_{k=0}^{K(\epsilon)-1} (k+1) \ .$$

Using Faulhaber's formula, sum of β-th power of first k positive integers is a polynomial function of $k+1$. Therefore, there exists some constant $\alpha_k \in (0, 1)$ such that $\sum_{k=0}^{K(\epsilon)-1}(k+1) \leq \alpha_k K(\epsilon)^{\beta+1}$. Therefore,

$$W \leq \frac{\sigma^2 \alpha_k}{\epsilon_1} K^{\beta+1}$$

Substituting $K = O(\epsilon^2)$ yields the desired result. □

B Additional Numerical Results: Constrained Logistic Regression

In this section, we provide numerical results corresponding to the binary classification data sets given in Table 1 from the [17] collection.

Table 1: Binary classification data set details. For more information see [17].

data set	dimension (n)	datapoints (N)
australian	14	690
ionosphere	34	351
mushrooms	112	8,124
sonar	60	208
splice	60	1,000

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

(g) Step Size vs. Iterations (h) Batch Size vs. Iterations

Figure 5: australian : First & Second Row: Feasibility & stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1). Last Row: Step sizes and Batch sizes versus iterations.

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

(g) Step Size vs. Iterations (h) Batch Size vs. Iterations

Figure 6: ionosphere : First & Second Row: Feasibility & stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1). Last Row: Step sizes and Batch sizes versus iterations.

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

(g) Step Size vs. Iterations (h) Batch Size vs. Iterations

Figure 7: mushroom First & Second Row: Feasibility & stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1). Last Row: Step sizes and Batch sizes versus iterations.

(a) Feasibility vs. Iterations (b) Feasibility vs. Epochs (c) Feasibility vs. LS Iters

(d) Stationarity vs. Iterations (e) Stationarity vs. Epochs (f) Stationarity vs. LS Iters

(g) Step Size vs. Iterations (h) Batch Size vs. Iterations

Figure 8: sonar: First & Second Row: Feasibility & stationarity errors versus iterations/epochs/linear system iterations for exact and inexact variants of Algorithm 2 on (6.1). Last Row: Step sizes and Batch sizes versus iterations.