Mathematical Programming
https://doi.org/10.1007/s10107-023-01981-1

FULL LENGTH PAPER

Series A

# Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization

**Frank E. Curtis**[1] · **Michael J. O'Neill**[2] · **Daniel P. Robinson**[1]

Received: 31 December 2021 / Accepted: 4 May 2023


**Abstract**
A worst-case complexity bound is proved for a sequential quadratic optimization (commonly known as SQP) algorithm that has been designed for solving optimization problems involving a stochastic objective function and deterministic nonlinear equality constraints. Barring additional terms that arise due to the adaptivity of the monotonically nonincreasing merit parameter sequence, the proved complexity bound is comparable to that known for the stochastic gradient algorithm for unconstrained nonconvex optimization. The overall complexity bound, which accounts for the adaptivity of the merit parameter sequence, shows that a result comparable to the unconstrained setting (with additional logarithmic factors) holds with high probability.

**Keywords** Nonlinear optimization · Stochastic optimization · Sequential quadratic optimization · Worst-case complexity

**Mathematics Subject Classification** 49M37 · 62L20 · 65K05 · 65Y20 · 68Q25 · 68W40 · 90C06 · 90C30 · 90C55

## 1 Introduction

We present a worst-case complexity analysis of an algorithm for minimizing a smooth objective function subject to nonlinear equality constraints. (Due to the nature of

✉ Michael J. O'Neill
mikeoneill@unc.edu

Frank E. Curtis
frank.e.curtis@lehigh.edu

Daniel P. Robinson
daniel.p.robinson@lehigh.edu

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

[2] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA

Published online: 07 June 2023

the algorithm, this worst-case complexity analysis holds in terms of iterations, function evaluations, and derivative evaluations.) Problems of this type arise in various important applications throughout science and engineering, including optimal control, PDE-constrained optimization, and resource allocation [3, 4, 20, 29]. However, unlike the vast majority of the literature on equality constrained optimization, the algorithm that we consider has been designed to solve problems in which the objective function is stochastic, in the sense that it is defined by the expectation of a function that has a random variable argument. The algorithm that we consider assumes that evaluations of the objective function and its gradient are intractable to obtain, but that it has access to (unbiased) stochastic gradient estimates.

A few algorithms have been proposed recently for solving problems of this type. These approaches fall into two categories: penalty methods [9, 25, 30] (which includes the class of augmented Lagrangian methods) and sequential quadratic optimization (commonly known as SQP) methods [2, 23]. Penalty methods aim to solve the constrained optimization problem by adding a term to the objective function, weighted by a penalty parameter, that penalizes constraint violation. Unconstrained optimization techniques are then applied to minimize the resulting penalty function, after which the penalty parameter may be modified and the minimization is performed again in an iterative manner until a solution is obtained that (approximately) satisfies the original constraints. Methods of this type perform well in some situations, but in others they perform poorly, e.g., due to ill-conditioning and/or nonsmoothness of the subproblems. Such methods also often suffer due to their sensitivity to the particular scheme used for updating the penalty parameter. See [28] for further commentary on the advantages and disadvantages of penalty methods.

In practice in both deterministic and stochastic optimization contexts, penalty methods are frequently outperformed by SQP methods. Indeed, it is commonly accepted in the deterministic optimization literature that a state-of-the-art algorithm is an SQP method that chooses stepsizes based on a line search applied to a merit function. In this deterministic setting, such an algorithm is intimately connected with applying Newton's method to the first-order primal-dual necessary conditions for optimality of the problem [32].

In this paper, we present a worst-case complexity analysis of the SQP method proposed in [2], which can be seen as an extension of an SQP method from the deterministic to the stochastic setting. A consequence of our analysis is that, in an idealized setting in which (i) one knows a threshold for the merit parameter below which knowledge of the exact gradient would not lead to a merit parameter decrease in any iteration (which, for one thing, is a threshold for the merit function to be *exact* [18]) and (ii) the algorithmic rule that might otherwise modify the merit parameter is skipped, the number of iterations required until the method generates a point at which first-order necessary conditions for optimality hold in expectation with accuracy $\varepsilon \in (0, \infty)$ is $\mathcal{O}(\varepsilon^{-4})$. This is the type of result that one should expect, since this is the same bound proved to hold for a stochastic gradient method employed to solve an unconstrained nonconvex problem [14]. However, *our analysis does not only consider this idealized setting*; we go further and prove a worst-case complexity bound for the algorithm when the merit parameter threshold is unknown and the algorithm adaptively updates a monotonically nonincreasing merit parameter sequence. We prove under reasonable

assumptions that the aforementioned worst-case bound, with additional logarithmic factors, holds with high probability. The high-probability aspect of this result arises due to the uncertainty of the behavior of the adaptive merit parameter sequence that is a consequence of the uncertainty due to the stochastic gradient estimates, and it does not reflect any uncertainty of the behavior of the method during situations in which the merit parameter sequence remains constant. A major challenge in our analysis is accounting for the *transient* behavior of the adaptive merit parameter sequence in the sense that our analysis accounts for the real possibilities that (i) different runs of the algorithm may decrease the merit parameter by varying amounts and (ii) it is not possible to bound the number of iterations until the merit parameter settles on a sufficiently small value. In other words, since our loose assumptions do not allow for one to presume that the algorithm experiences two distinct phases (e.g., one in which—over a bounded number of iterations—the merit parameter is reduced down to a threshold, then a second in which it remains fixed), our analysis must account for the possibility that the merit parameter is transient through any number of iterations, which leads to significant complications that are not present in the context of unconstrained optimization.

To the best of our knowledge, ours is the first worst-case complexity result for an SQP algorithm that operates in the *highly stochastic regime* (where one merely presumes that the stochastic gradient estimates have bounded variance) for solving stochastic optimization problems involving deterministic nonlinear equality constraints. Prior to this work, the only known complexity results for stochastic constrained optimization were for algorithms for solving problems with simple constraint sets that enable projection-based methods [14, 16, 26] and Frank-Wolfe type methods [19]. (One exception is a complexity bound proved for the SQP algorithm proposed in [23], although that result only holds for the idealized setting in which the algorithm has *a priori* knowledge of a threshold for the merit function parameter.) After the initial release of this paper, the article [24] appeared that proposes an SQP method that uses stochastic estimates of the objective gradient and Hessian of the Lagrangian and computes search directions using a sketch-and-project framework. The algorithm in that paper does not use a merit function and requires that the stepsizes remain within prescribed intervals that are non-adaptive to the stochastic gradients. Under standard assumptions, that algorithm offers a worst-case complexity result that is comparable to the one we derive in this paper, as well as an asymptotic convergence rate. Our analysis focuses a great deal on the complications that arise due to the adaptivity of the merit parameter sequence, which essentially means that the algorithm in our consideration is aiming to reduce a merit function that *changes* during the optimization process. Hence, many aspects of our analysis are quite distinct from the analyses that have been presented for stochastic gradient methods in the context of unconstrained optimization or optimization over simple constraint sets, for which the tool for measuring the progress of an algorithm—namely, the objective function itself—remains the same throughout the optimization.

### 1.1 Problem formulation

The algorithm that we consider is designed to solve problems of the form

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

where $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, $\omega$ is a random variable with associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, and $\mathbb{E}$ represents expectation with respect to $\mathbb{P}$. In particular, similar to [2], we make the following assumption.

**Assumption 1** The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below by $f_{\text{low}} \in \mathbb{R}$ and the corresponding gradient function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipschitz continuous with constant $L \in (0, \infty)$. The constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ (where $m \leq n$) and the corresponding Jacobian function $J := \nabla c^\top : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ are bounded, each gradient function $\nabla c_i : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $\gamma_i$ for all $i \in \{1, \ldots, m\}$, and the singular values of $J \equiv \nabla c^\top$ are bounded below and away from zero.

Defining the Lagrangian $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ corresponding to (1) by $\ell(x, y) := f(x) + c(x)^\top y$, first-order primal-dual stationarity conditions for (1), which are necessary for optimality under Assumption 1, are given by

$$0 = \begin{bmatrix} \nabla_x \ell(x, y) \\ \nabla_y \ell(x, y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + \nabla c(x) y \\ c(x) \end{bmatrix}. \tag{2}$$

### 1.2 Notation

We adopt the notation that $\|\cdot\|$ denotes the $\ell_2$-norm for vectors and the vector-induced $\ell_2$-norm for matrices. We denote by $\mathbb{S}^n$ the set of $n \times n$ dimensional real symmetric matrices. The set of nonnegative integers is denoted as $\mathbb{N} := \{0, 1, 2, \ldots, \}$. For any integer $k \in \mathbb{N}$, we use $[k]$ to denote the subset of nonnegative integers up to $k$, namely, $[k] := \{0, \ldots, k\}$. Correspondingly, to represent a set of vectors $\{v_0, \ldots, v_k\}$, we define $v_{[k]} := \{v_0, \ldots, v_k\}$.

Given $\phi : \mathbb{R} \to \mathbb{R}$ and $\varphi : \mathbb{R} \to [0, \infty)$, we write $\phi(\cdot) = \mathcal{O}(\varphi(\cdot))$ to indicate that $|\phi(\cdot)| \leq c\varphi(\cdot)$ for some $c \in (0, \infty)$. Similarly, we write $\phi(\cdot) = \widetilde{\mathcal{O}}(\varphi(\cdot))$ to indicate that $|\phi(\cdot)| \leq c\varphi(\cdot)|\log^{\overline{c}}(\cdot)|$ for some $c \in (0, \infty)$ and $\overline{c} \in (0, \infty)$. In this manner, one finds that $\mathcal{O}(\varphi(\cdot)|\log^{\overline{c}}(\cdot)|) \equiv \widetilde{\mathcal{O}}(\varphi(\cdot))$ for any $\overline{c} \in (0, \infty)$.

The algorithm that we analyze is iterative, generating in each realization a sequence $\{x_k\}$. (See Sect. 4.1 for a complete description of the stochastic process generated by the algorithm.) We also append the iteration number to other quantities corresponding an iteration, e.g., $f_k := f(x_k)$ for all $k \in \mathbb{N}$.

### 1.3 Outline

Section 2 provides a worst-case complexity result for the algorithm from [2] for the deterministic setting, and uses this result and further commentary to provide an

overview of our main result for the stochastic setting. Details of the algorithm for the stochastic setting are presented in Sect. 3, followed by our main result and analysis, which are provided in Sect. 4. Finally, we provide concluding thoughts and mention future directions in Sect. 5.

## 2 Outline of main results

Our algorithm of consideration is derived from Algorithm 3.1 in [2], which is proposed for the stochastic setting. As a precursor, Algorithm 2.1 was proposed in [2] for the deterministic setting, many of the features of which are used in Algorithm 3.1 in [2] for the stochastic setting. In Algorithm 2.1 in [2] for the deterministic setting, the $k$th search direction $d_k \in \mathbb{R}^n$ is computed by solving a subproblem defined by a quadratic approximation of the objective function and an affine approximation of the constraints using derivative information at the current iterate $x_k \in \mathbb{R}^n$. This computation also results in a Lagrange multiplier vector $y_k \in \mathbb{R}^m$. The subsequent iterate is set by $x_{k+1} \leftarrow x_k + \alpha_k d_k$, where $\alpha_k \in (0, \infty)$ is a stepsize determined by a procedure to reduce the merit function $\phi : \mathbb{R}^n \times (0, \infty) \to \mathbb{R}$ defined by $\phi(x, \tau) = \tau f(x) + \|c(x)\|_1$. In particular, based on properties of the search direction $d_k$, a value of the merit parameter $\tau_k \in (0, \tau_{k-1}]$ is set by the algorithm, after which $\alpha_k \in (0, \infty)$ is computed to ensure that $\phi(x_k, \tau_k) - \phi(x_{k+1}, \tau_k)$ is sufficiently positive. This description of the algorithm suffices for providing an outline of our main results; for further details about the algorithm that we analyze, see Sect. 3.

### 2.1 Complexity of the deterministic algorithm

To motivate our main result for the stochastic setting, it is instructive to state a worst-case complexity bound for the deterministic algorithm. Such a result is the following; further details and a proof are provided in Appendix A. The approximate stationarity conditions in (3) have been defined for consistency between the primal and dual stationarity measures; the reason that this choice leads to such consistency is revealed in the analysis in Appendix A.

**Theorem 1** *Consider Algorithm 2.1 in* [2] *and suppose that Assumption 1 holds along with Assumption 2.4 from* [2] *(see also the similar Assumption 2 on page 8 for our stochastic setting). Let* $\tau_{-1} \in \mathbb{R}_{>0}$ *be the initial value of the merit parameter sequence and let* $\tau_{\min} \in (0, \tau_{-1}]$ *be a positive lower bound for the merit parameter sequence (the existence of which follows from Lemma 2.16 in* [2]*). Then, for any* $\varepsilon \in (0, 1)$*, there exists* $(\kappa_1, \kappa_2) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ *such that the algorithm reaches an iterate* $(x_k, y_k) \in \mathbb{R}^n \times \mathbb{R}^m$ *satisfying*

$$\|g_k + J_k^\top y_k\| \leq \varepsilon \quad \text{and} \quad \sqrt{\|c_k\|_1} \leq \varepsilon \tag{3}$$

*in a number of iterations no more than*

$$\left(\frac{\tau_{-1}(f_0 - f_{low}) + \|c_0\|_1}{\min\{\kappa_1, \tau_{\min}\kappa_2\}}\right)\varepsilon^{-2}. \tag{4}$$

Theorem 1 is notable since it shows that Algorithm 2.1 in [2] has the standard worst-case iteration, function-evaluation, and derivative-evaluation complexity for a first-order-derivative-based algorithm for solving equality constrained optimization problems; see [8] for a comparable complexity bound for another "short-step" algorithm for solving equality constrained problems. That said, the $\mathcal{O}(\varepsilon^{-2})$ bound in Theorem 1 is not surprising. After all, such a complexity bound is well-known for gradient-based algorithms for solving unconstrained nonconvex optimization problems as well. Since Algorithm 2.1 in [2] and its corresponding analysis do not exploit the use of exact higher-order derivative information, this complexity bound is on the order of what could be expected for such a method. There are ways for achieving improved worst-case complexity if higher-order derivatives are employed in a constrained-optimization algorithm (see, e.g., [7, 13, 17]), but the use of such higher-order derivatives—especially in the stochastic setting—is outside of our scope.

### 2.2 Preview of the complexity of the stochastic algorithm

Moving to the stochastic setting, there are a few major technical hurdles that need to be addressed, all of which relate to the adaptivity of the merit parameter sequence. In particular, the analysis for the deterministic setting relies heavily on the facts that (i) each step of the algorithm yields a sufficient reduction in the merit function for the current value of the merit parameter, (ii) each such reduction in the merit function can be tied to a first-order primal-dual stationarity measure for the current iterate, and (iii) under Assumption 1, one can be certain of the existence of a positive lower bound for the merit parameter sequence. This lower bound for the merit parameter is referenced directly in the proof of the worst-case bound for the deterministic algorithm; in particular, it is shown (see Lemma 13 and the beginning of the proof of Theorem 4) that the improvement in the merit function from any iterate that is not $\varepsilon$-stationary (see (3)) is at least proportional to $\min\{1, \tau_{\min}\}\varepsilon^{-2}$, even if the current value of the merit parameter is greater than $\tau_{\min}$. Unfortunately, these properties of the steps and merit parameter sequence are not certain in the stochastic setting. For example, as discussed in [2], it is possible—even under Assumption 1—for the merit parameter sequence to vanish or for it to eventually remain constant at a value that is not sufficiently small, and for there to be iterations in which the expected reduction in the merit function cannot be tied to a first-order primal-dual stationarity measure. As a result, we have devised new analytical approaches that confront the fact that $\{\tau_k\}$ is a random process, the ultimate behavior of which is uncertain.

To aid the reader, we provide here an overview and commentary about our ultimate complexity bound; see Corollary 1 on page 27. Our result is proved under Assumption 1 along with others that are introduced in the subsequent sections. For one thing, as is common in SQP methods for deterministic optimization, we assume that the

subproblem defining the search direction in each iteration is defined by a matrix that is positive definite in the null space of the constraint Jacobian; see Assumption 2 on page 8. We also assume, as is common for stochastic gradient methods, that the stochastic gradient estimates are unbiased with variance bounded by $M \in (0, \infty)$, along with some related assumptions; see Assumption 3 on page 14. Furthermore, our analysis conditions on the occurrence of an event that we call $E$ (see (15)); this event captures situations in which, over a total of $k_{\max} + 1 \in \mathbb{N}$ iterations, the merit parameter is reduced at most $s_{\max} \in [k_{\max}]$ times and the merit parameter is bounded below by $\tau_{\min} \in (0, \infty)$. Under these conditions, our main complexity result shows that, within $k_{\max} + 1$ iterations, it holds with probability $1 - \delta \in (0, 1)$ that the algorithm generates $x_{k^*} \in \mathbb{R}^n$ corresponding to which there exists an associated Lagrange multiplier $y_{k^*}^{\text{true}} \in \mathbb{R}^m$ such that

$$\begin{aligned} &\mathbb{E}[\|\nabla f_{k^*} + J_{k^*}^\top y_{k^*}^{\text{true}}\|^2 + \|c_{k^*}\|_1 | E] \\ &= \mathcal{O}\left( \frac{\tau_{-1}(f_0 - f_{\text{low}}) + \|c_0\|_1 + M}{\sqrt{k_{\max} + 1}} \right. \qquad (5a) \\ &\quad \left. + \frac{(\tau_{-1} - \tau_{\min})(s_{\max} \log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max} + 1}} \right). \qquad (5b) \end{aligned}$$

This form of the result is commonly called a convergence rate since it bounds the expected stationarity error from above by a function that decreases with the number of iterations performed, namely, $k_{\max} + 1$. This bound can be used to form a worst-case complexity result. Specifically, the result above and Jensen's inequality imply that, within $k_{\max} + 1$ iterations and as long as $s_{\max} = \mathcal{O}(\log(k_{\max}))$ (more on this below), it holds with probability $1 - \delta$ that the algorithm requires at most $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ iterations to generate $x_{k^*}$ with corresponding $y_{k^*}^{\text{true}}$ such that $\mathbb{E}[\|\nabla f_{k^*} + J_{k^*}^\top y_{k^*}^{\text{true}}\| | E] \leq \varepsilon$ and $\mathbb{E}[\sqrt{\|c_{k^*}\|_1} | E] \leq \varepsilon$. The probability in this result is with respect to the distribution of the stochastic gradients conditioned on the occurrence of the event $E$.

The first three quantities on the right-hand side of the convergence rate, namely, in (5a), representing the initial objective function gap, initial constraint violation, and the variance of the stochastic gradient estimates, mirror the presence of similar terms that appear for comparable results for the stochastic gradient method in an unconstrained or simple-constraint-set setting [14, 16, 26]. The final term in (5b), on the other hand, as well as the fact that the result is stated as a high-probability result, are unique to our setting and arise due to the adaptivity of the merit parameter sequence. If one were to have prior knowledge of $\tau_{\min}$, then one could set $\tau_{-1} = \tau_{\min}$ (and disable the update mechanism for the merit parameter in the algorithm), in which case our analysis would show that the expected stationarity error is bounded above by (5a) (surely, not only with high probability).

In the context of an adaptive merit parameter sequence, the particular form of our complexity result depends on the magnitude of $s_{\max}$ relative to $k_{\max} + 1$, i.e., the bound on the number of times that the merit parameter is decreased relative to the total number of iterations performed. One setting in which our result is relatively straightforward

is when, over all realizations of the algorithm, the differences between the stochastic gradient estimates and the true gradients are bounded deterministically, in which case the merit parameter sequence is provably bounded below, which in turn means that $s_{\max}$ is bounded by a value that is independent from $k_{\max}+1$ (at least as long as $k_{\max}$ is sufficiently large relative to $\tau_{-1}/\tau_{\min}$); this follows from a deterministic lower bound on $\tau_{\min}$ [2, Proposition 3.18] and the fact that whenever the merit parameter is decreased, it is done so by a constant factor. Beyond this setting, for another concrete example of a situation in which $s_{\max}$ is guaranteed to be sufficiently small relative to $k_{\max}$, we prove in Sect. 4.5 that if the distributions of the stochastic gradient estimates are sub-Gaussian, then with probability $1-\delta$ one finds that $s_{\max} = \mathcal{O}(\log(\log(\frac{k_{\max}}{\delta})))$, meaning that our proved convergence rate is not ruined by the term in (5b). There are certainly other special cases in which similar types of relationships between $s_{\max}$ and $k_{\max}$ hold, but for our purposes we simply provide the example in Sect. 4.5.

## 3 Algorithm

For ease of reference, in this section we present Algorithm 3.1 from [2] (slightly modified, as explained at the end of this section), which is designed to solve problems of the form (1) and is our focus for the remainder of the paper. In the spirit of an SQP method, the algorithm computes a search direction $d_k$ and Lagrange multiplier vector $y_k$ in iteration $k \in \mathbb{N}$ by solving

$$\min_{d\in\mathbb{R}^n} \; f_k + g_k^\top d + \tfrac{1}{2} d^\top H_k d \quad \text{s.t.} \quad c_k + J_k d = 0, \tag{6}$$

where $g_k$ is a stochastic gradient estimate at $x_k$ and $H_k \in \mathbb{S}^n$ is chosen independently from $g_k$. Under Assumption 1 and the following Assumption 2 (that we make throughout the remainder of the paper), the solution of (6) can be obtained from the unique solution of the linear system

$$\begin{bmatrix} H_k & J_k^\top \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ y_k \end{bmatrix} = - \begin{bmatrix} g_k \\ c_k \end{bmatrix}. \tag{7}$$

**Assumption 2** The sequence $\{\|H_k\|\}$ is bounded by $\kappa_H \in \mathbb{R}_{>0}$, where for all $k \in \mathbb{N}$ the matrix $H_k \in \mathbb{S}^n$ is chosen independently from $g_k$. In addition, there exists $\zeta \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the matrix $H_k \in \mathbb{S}^n$ has the property that $u^\top H_k u \geq \zeta \|u\|_2^2$ for all $u \in \mathbb{R}^n$ such that $J_k u = 0$.

After computation of $(d_k, y_k)$, the remainder of the $k$th iteration involves (i) updating the merit parameter, (ii) updating an auxiliary parameter needed for the stepsize computation, and (iii) computing a positive stepsize. These algorithmic components are designed with the aim of yielding a sufficiently positive reduction in a model of the merit function, which in turn is aimed at yielding a sufficiently positive reduction in the merit function itself. The algorithm employs the model

$q : \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{S}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$q(x, \tau, g, H, d) = \tau(f(x) + g^\top d + \tfrac{1}{2}\max\{d^\top H d, 0\}) + \|c(x) + J(x)d\|_1,$$

and the reduction function $\Delta q : \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{S}^n \times \mathbb{R}^n \to \mathbb{R}$, for a given $d \in \mathbb{R}^n$ satisfying $c(x) + J(x)d = 0$, defined by

$$\begin{aligned} \Delta q(x, \tau, g, H, d) &:= q(x, \tau, g, H, 0) - q(x, \tau, g, H, d) \\ &= -\tau\left(g^\top d + \tfrac{1}{2}\max\{d^\top H d, 0\}\right) + \|c(x)\|_1. \end{aligned} \tag{8}$$

Specifically, in order to ensure in iteration $k$ that $\tau_k \le \tau_{k-1}$ and

$$\Delta q(x_k, \tau, g_k, H_k, d_k) \ge \tfrac{1}{2}\tau \max\{d_k^\top H_k d_k, 0\} + \sigma\|c_k\|_1 \ge 0 \tag{9}$$

holds for all $\tau \le \tau_k$, the algorithm sets, for user-defined $\sigma \in (0, 1)$, the value

$$\tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^\top d_k + \max\{d_k^\top H_k d_k, 0\} \le 0 \\ \frac{(1-\sigma)\|c_k\|_1}{g_k^\top d_k + \max\{d_k^\top H_k d_k, 0\}} & \text{otherwise,} \end{cases} \tag{10}$$

and then sets, for some $\epsilon_\tau \in (0, 1)$, the merit parameter value

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

Then, for use in the stepsize computation (as motivated in [2]) it sets

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

for some $\epsilon_\xi \in (0, 1)$, which, for one thing, ensures $\xi_k \le \xi_k^{\text{trial}}$. The last component in the $k$th iteration is to set the stepsize, the magnitude of which is controlled by a prescribed sequence $\{\beta_k\} \subset (0, 1]$, which is employed in the following projection interval that is used in the stepsize computation:

$$\text{Proj}_k(\cdot) := \text{Proj}\left(\cdot \;\middle|\; \left[\frac{\beta_k \xi_k \tau_k}{\tau_k L + \Gamma}, \frac{\beta_k \xi_k \tau_k}{\tau_k L + \Gamma} + \theta\beta_k^2\right]\right),$$

where $\text{Proj}(\cdot \mid \mathcal{B})$ represents the projection operator onto the interval $\mathcal{B} \subset \mathbb{R}$. As in other stochastic-gradient-based methods, the convergence properties of the method depend on properties of $\{\beta_k\}$, which in many analyses is considered to be a constant or diminishing sequence. We establish our complexity result for the case of constant $\{\beta_k\}$ with $\beta_k = \mathcal{O}(1/\sqrt{k_{\max} + 1})$ for all $k \in [k_{\max}]$.

Overall, the algorithm that we consider is stated as Algorithm 1. The only changes from Algorithm 3.1 in [2] are the fixed iteration limit ($k_{\max}$) and the concluding step for producing the return value ($x_{k^*}$). This method of sampling $k^*$ to produce the return value is consistent with other approaches in the literature on complexity analyses for algorithms for solving nonconvex optimization problems; see, e.g., [14]. It amounts to uniform sampling over the iterates when constant $\{\beta_k\}$ is considered, as in our analysis. Finally, we remark that Algorithm 1 presumes knowledge of Lipschitz constants for the objective and constraint gradients, although in practice one might only estimate these values using standard procedures [12].

**Algorithm 1** Stochastic SQP Algorithm

**Require:** $x_0 \in \mathbb{R}^n$; $k_{\max} \in \mathbb{N}$; $\tau_{-1} \in \mathbb{R}_{>0}$; $\epsilon_\tau \in (0,1)$; $\epsilon_\xi \in (0,1)$; $\sigma \in (0,1)$; $\xi_{-1} \in \mathbb{R}_{>0}$; $\{\beta_k\} \subset (0,1]$; $\theta \in \mathbb{R}_{\geq 0}$; $L \in (0,\infty)$, a Lipschitz constant for $\nabla f$; $\Gamma \in [\sum_{i=1}^m \gamma_i, \infty)$, where $\gamma_i \in (0,\infty)$ is a Lipschitz constant for $\nabla c_i$ for all $i \in [m]$

**for all** $k \in [k_{\max}]$ **do**
  Compute $(d_k, y_k)$ as the solution of (7)
  **if** $d_k = 0$ **then**
    Set $\tau_k^{\text{trial}} \leftarrow \infty$, $\tau_k \leftarrow \tau_{k-1}$, $\xi_k^{\text{trial}} \leftarrow \infty$, and $\xi_k \leftarrow \xi_{k-1}$
    Set $\widehat{\alpha}_{k,\text{init}} \leftarrow 1$, $\widetilde{\alpha}_{k,\text{init}} \leftarrow 1$, and $\alpha_k \leftarrow 1$
  **else** (if $d_k \neq 0$)
    Set $\tau_k^{\text{trial}}$ by (10) and $\tau_k$ by (11)
    Set $\xi_k^{\text{trial}}$ and $\xi_k$ by (12)
    Set

$$\widehat{\alpha}_{k,\text{init}} \leftarrow \frac{\beta_k \Delta q(x_k, \tau_k, g_k, H_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \quad \text{and} \quad \widetilde{\alpha}_{k,\text{init}} \leftarrow \widehat{\alpha}_{k,\text{init}} - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

    Set $\widehat{\alpha}_k \leftarrow \text{Proj}_k(\widehat{\alpha}_{k,\text{init}})$ and $\widetilde{\alpha}_k \leftarrow \text{Proj}_k(\widetilde{\alpha}_{k,\text{init}})$, then

$$\alpha_k \leftarrow \begin{cases} \widehat{\alpha}_k & \text{if } \widehat{\alpha}_k < 1 \\ 1 & \text{if } \widetilde{\alpha}_k \leq 1 \leq \widehat{\alpha}_k \\ \widetilde{\alpha}_k & \text{if } \widetilde{\alpha}_k > 1 \end{cases}$$

  **end if**
  Set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
**end for**
Sample $k^* \in [k_{\max}]$, where $\mathbb{P}[k^* = k] = \frac{\beta_k}{\sum_{k=0}^{k_{\max}} \beta_k}$ for all $k \in [k_{\max}]$, then **return** $x_{k^*}$

## 4 Complexity analysis

We begin our complexity analysis by describing the algorithm as a stochastic process (Sect. 4.1), then formalizing the assumptions that we make about the stochastic gradient estimates (Sect. 4.2). We then state, in some cases in a slightly modified form, some key lemmas from [2] that are needed for our analysis (Sect. 4.3). Our generic complexity result, which has been outlined in Sect. 2, is then stated and proved (Sect. 4.4). Consequences and extensions of our generic complexity result are then discussed for some special cases of distributions for the stochastic gradient estimates for which our required assumptions hold with high probability (Sect. 4.5). Finally, we conclude this

section by outlining a form of our generic complexity result that relaxes one of our minor simplifying assumptions (Sect. 4.6).

Similarly as for the convergence analysis in [2], our complexity analysis makes use of orthogonal decompositions of the search directions computed by the algorithm; in particular, for all $k \in \mathbb{N}$, we express $d_k = u_k + v_k$, where $u_k \in \text{Null}(J_k)$ and $v_k \in \text{Range}(J_k^\top)$. We note here that conditioned on the algorithm having reached $x_k$ at iteration $k$, the normal component $v_k$ is *deterministic*, depending only on the constraint value $c_k$ and the Jacobian $J_k$.

In addition to the quantities that are computed explicitly in Algorithm 1, our analysis also refers to the quantities that would have been computed in each iteration $k \in \mathbb{N}$, conditioned on the event that the algorithm has reached $x_k$ as the $k$th iterate, if the true gradient $\nabla f(x_k)$ is used in place of the stochastic gradient $g_k$. These quantities are denoted by a "true" superscript. For example, in iteration $k$, the true search direction and corresponding true Lagrange multiplier estimate are the solution of the linear system

$$\begin{bmatrix} H_k & J_k^\top \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k^{\text{true}} \\ y_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c_k \end{bmatrix}, \tag{13}$$

which may be decomposed as $d_k^{\text{true}} = u_k^{\text{true}} + v_k$, where $u_k^{\text{true}} \in \text{Null}(J_k)$ and $v_k \in \text{Range}(J_k^\top)$. Here, we write $v_k$ (without a superscript) since the normal component of the search direction is defined in a manner that makes it independent of the objective gradient (estimate). Similarly, the true value of the merit parameter that would have been computed is denoted

$$\tau_k^{\text{trial,true}} \leftarrow \begin{cases} \infty & \text{if } \nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} \le 0 \\ \frac{(1-\sigma)\|c_k\|_1}{\nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\}} & \text{otherwise.} \end{cases}$$

This definition of $\tau_k^{\text{trial,true}}$ guarantees that, for any $\tau \le \tau_k^{\text{trial,true}}$, one finds

$$\Delta q(x_k, \tau, \nabla f(x_k), H_k, d_k^{\text{true}}) \ge \tfrac{1}{2}\tau \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} + \sigma \|c_k\|_1. \tag{14}$$

### 4.1 Stochastic process

Henceforth, for the sake of formality, we shall refer in our analysis to the stochastic process generated by Algorithm 1. Specifically, in terms of values that are computed by the algorithm itself, we have the stochastic process

$$\{(X_k, G_k, D_k, Y_k, \mathcal{T}_k, \Xi_k, \mathcal{A}_k)\},$$

where, for all $k \in \mathbb{N}$, the random variables are: the algorithm iterate $X_k$, stochastic gradient estimate $G_k$, search direction $D_k$, Lagrange multiplier estimate $Y_k$, merit parameter $\mathcal{T}_k$, ratio parameter $\Xi_k$, and stepsize $\mathcal{A}_k$. For all $k \in \mathbb{N}$, a realization of the

corresponding element of this process has been denoted $(x_k, g_k, d_k, y_k, \tau_k, \xi_k, \alpha_k)$. Similarly, in terms of "true" values and step decomposition values that are not computed by the algorithm, but are defined for the sake of our analysis, we have the simultaneously generated process

$$\{(V_k, U_k, D_k^{\text{true}}, U_k^{\text{true}}, Y_k^{\text{true}}, \mathcal{T}_k^{\text{trial,true}})\},$$

where, for all $k \in \mathbb{N}$, the random variables are: the normal search direction component $V_k$, the tangential search direction component $U_k$, the true search direction $D_k^{\text{true}}$, the true tangential search direction component $U_k^{\text{true}}$, the true Lagrange multiplier estimate $Y_k^{\text{true}}$, and the true trial merit parameter $\mathcal{T}_k^{\text{trial,true}}$. For all $k \in \mathbb{N}$, a realization of the corresponding element of this process has been denoted $(v_k, u_k, d_k^{\text{true}}, u_k^{\text{true}}, y_k^{\text{true}}, \tau_k^{\text{trial,true}})$. Finally, for the sake of tracking the number of merit parameter updates that occur during runs of the algorithm, we define the stochastic process $\{S_k\}$, where for all $k \in \mathbb{N}$ the random variable $S_k$ represents the number of merit parameter decreases up to the end of the $k$th iteration, i.e., the number of iterations in which $\mathcal{T}_k < \mathcal{T}_{k-1}$. For all $k \in \mathbb{N}$, a realization of $S_k$ is denoted $s_k$.

In any run, the behavior of Algorithm 1 is dictated entirely by the initial conditions and the sequence of stochastic gradient estimates that are generated. Let $\mathcal{G}_k$ denote the $\sigma$-algebra generated by the random variables $\{G_0, \ldots, G_{k-1}\}$, a realization of which (along with all initial conditions of the algorithm, including $X_0 = x_0$) determines the realizations of

$$\{X_j\}_{j=1}^k \text{ and } \{(D_j, Y_j, \mathcal{T}_j, \Xi_j, \mathcal{A}_j, V_j, U_j, D_j^{\text{true}}, U_j^{\text{true}}, Y_j^{\text{true}}, \mathcal{T}_j^{\text{trial,true}}, S_j)\}_{j=0}^{k-1}.$$

For completeness, let $\mathcal{G}_0 = \sigma(x_0)$. As a result, $\{\mathcal{G}_k\}_{k\geq 0}$ is a filtration.

### 4.2 Assumptions

Our analysis presumes certain good behavior of the sequences of merit and ratio parameters that are set adaptively by the algorithm. Formally, given $(k_{\max}, s_{\max}, \tau_{\min}, \xi_{\min}) \in \mathbb{N} \times \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, our main result characterizes the worst-case behavior of Algorithm 1 conditioned on the event denoted as

$$E := E(k_{\max}, s_{\max}, \tau_{\min}, \xi_{\min}), \tag{15}$$

which we define as the event such that

- $\mathcal{T}_k \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\mathcal{T}_k^{\text{trial,true}} \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\Xi_k = \xi_{\min} > 0$ for all $k \in [k_{\max}]$, and
- $|\{k \in [k_{\max}] : \mathcal{T}_k < \mathcal{T}_{k-1}\}| \leq s_{\max}$.

Consideration of this event as a focus for proving a worst-case complexity result for Algorithm 1 is justifiable for the following reasons.

– The condition in $E$ that the ratio parameter sequence $\{\Xi_k\}$ is constant over all iterations is not actually essential for our analysis; rather, it is made for the sake of simplicity. Indeed, in Sect. 4.6, we present an extension of our main result to the setting in which this parameter sequence is not constant. Observe that, as proved in [2, Lemma 3.5], the sequence $\{\Xi_k\}$ is bounded below by a positive real number whose value is deterministic, i.e., it is independent of the sequence of stochastic gradient estimates that are generated by the algorithm. Hence, for the sake of simplicity, we assume for now that $\{\Xi_k\}$ is constant and leave the statement of the more complicated version of our main result to a subsection at the end of our analysis.
– The conditions in $E$ pertaining to the behavior of the merit parameter sequence are not necessarily minor, since stochasticity in the gradient estimates can cause the merit parameter to vanish, even in settings when the merit parameter would remain bounded below in the deterministic algorithm. That said, in Sect. 4.5, we consider a particular setting in which the distributions of the stochastic gradient estimates are sub-Gaussian over any run of the algorithm, in which case we show that the merit parameter remains bounded below with high probability, meaning that our main worst-case complexity bound—which holds with high probability due to the adaptivity of the merit parameter sequence—remains essentially unchanged in this setting when we do not presume upfront that the merit parameter sequence remains bounded above a positive real number.
– The condition in $E$ pertaining to the existence of $s_{\max}$ is not actually an additional requirement beyond the existence of $\tau_{\min}$ in the event. After all, by the construction of Algorithm 1, it follows that when the merit parameter is decreased, it is decreased by at least a constant factor, from which it follows (under the existence of $\tau_{\min}$) that $s_{\max}$ exists and satisfies

$$s_{\max} \leq \min\left\{k_{\max}+1, \left\lceil \frac{\log(\tau_{\min}/\tau_{-1})}{\log(1-\epsilon_\tau)} \right\rceil\right\}. \tag{16}$$

That said, for simplicity and generality in our analysis, we define $s_{\max}$ as a quantity that is decoupled from the above (conservative) inequality.

In summation, while our analysis requires the existence of a lower bound for the merit parameter sequence, and correspondingly an upper bound on the number of potential decreases in the merit parameter, it does *not* presume other convenient behavior of the merit parameter sequence. For example, our analysis does not presume that the merit parameter sequence eventually settles on a sufficiently small value in a number of iterations that can be bounded in a convenient manner. Rather, our analysis respects the stochastic, transient behavior of the merit parameter sequence that one finds in the actual behavior of the algorithm in practice. We have strived to derive a complexity bound that, under relatively loose assumptions in the highly stochastic regime, matches (up to logarithmic factors) in a constrained setting the results that are considered state-of-the-art for the unconstrained setting, even though the algorithm needs to discover for itself, in an adaptive manner, an appropriate balance between the objective function and constraint violation measure. Given our focus on event $E$, we now introduce the

filtration defined by

$$\mathcal{F}_k := \mathcal{G}_k \cap E \quad \text{for all} \quad k \in \mathbb{N}, \tag{17}$$

where, for all $k \in \mathbb{N}$, we use $\mathcal{G}_k \cap E$ to denote the trace $\sigma$-algebra of event $E$ on the $\sigma$-algebra $\mathcal{G}_k$, i.e. $\mathcal{G}_k \cap E = \{G \cap E : G \in \mathcal{G}_k\}$. Here and throughout the remainder of the paper, we let $\mathbb{P}_k[\cdot]$ (respectively, $\mathbb{E}_k[\cdot]$) denote probability (respectively, expectation) conditioned on $\mathcal{F}_k$, which encodes information up to the start of iteration $k \in \mathbb{N}$, i.e., we define

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

Our analysis assumes the following about the stochastic gradient estimates. Such an assumption, namely, that conditioned on the filtration one has that the stochastic gradient is unbiased and has bounded variance, is common in analyses of stochastic optimization methods.

**Assumption 3** There exists $M \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$, one finds

$$\mathbb{E}_k[G_k] = \nabla f(X_k) \quad \text{and} \quad \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2] \leq M. \tag{18}$$

In addition, there exists $M_\tau \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$, one finds

$$\text{either} \quad \mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] = 0 \text{ or} \tag{19a}$$

$$\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2 | \nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] \leq M_\tau. \tag{19b}$$

Observe that (19b) follows from (18) if there exists $p \in (0, 1]$ such that, for any $k \in [k_{\max}]$ with $\mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] > 0$, one finds

$$\mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] \geq p.$$

After all, in this setting with $Z_k$ representing the event that $\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0$ and $\mathcal{T}_k < \mathcal{T}_{k-1}$, and $Z_k^c$ representing the complement of $Z_k$, one finds along with Jensen's inequality that

$$\begin{aligned}
\sqrt{M} &\geq \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \\
&= \mathbb{P}_k[Z_k] \cdot \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2 | Z_k] + \mathbb{P}_k[Z_k^c] \cdot \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2 | Z_k^c] \\
&\geq p \cdot \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2 | Z_k],
\end{aligned}$$

so (19b) holds with $M_\tau = \sqrt{M}/p$. Such a $p$ exists, for example, when the objective of (1) is a finite sum of $N$ terms and each stochastic gradient estimate is computed as a so-called mini-batch estimate through the uniform (random) selection of $b$ indices, in which case the above holds with $p = b/N$.

We make one additional assumption for our analysis, namely, the following.

**Assumption 4** There exists $p_\tau \in (0, 1]$ such that, for all $k \in [k_{\max}]$, one finds

$$\begin{aligned}&\mathbb{P}_k[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\}\\ &\quad \geq \nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}] \geq p_\tau.\end{aligned}$$

Similar to [2, Proposition 3.16], Assumption 4 allows us to prove that, with high probability, the number of iterations in which $\mathcal{T}_k > \mathcal{T}_k^{\text{trial,true}}$ is not too large. In [2, Example 3.17], it was shown that the inequality in this assumption holds with $p_\tau = \frac{1}{2}$ when, conditioned on having reached a realized iterate $x_k$, the stochastic gradient $G_k$ has a Gaussian distribution. We show in Sect. 4.5 that this result can be extended to other settings as well.

### 4.3 Properties of algorithm 1

In this section, we state key preliminary results from [2] that are needed for our analysis. It is important to note that these results are written in [2] in the context of conditioning, for all $k \in \mathbb{N}$, on a particular realization of the algorithm up to the beginning of iteration $k$, which is different from our setting in which we condition on $\mathcal{F}_k$. That said, one finds that these results from [2] carry over, with nearly the same line of arguments, to our setting. After all, for any random variable $X$ that is $\mathcal{F}_k$-measurable and any random variable $Y$ with $\mathbb{E}[|Y|] < \infty$ and $\mathbb{E}[|XY|] < \infty$, [15, Theorem 4.1.14] states that

$$\mathbb{E}[XY|\mathcal{F}_k] = X\mathbb{E}[Y|\mathcal{F}_k]. \tag{20}$$

This property, combined with the arguments found in [2], is sufficient to prove the results of this section, so we state them without proof.

By [2, Lemma 2.10], there exists $\kappa_{uv} \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$, if $\|U_k^{\text{true}}\|^2 \geq \kappa_{uv}\|V_k\|^2$, then $\frac{1}{2}(D_k^{\text{true}})^\top H_k D_k^{\text{true}} \geq \frac{1}{4}\zeta\|U_k^{\text{true}}\|^2$, where $\zeta$ is defined in Assumption 2. Correspondingly, let us define

$$\Psi_k := \begin{cases} \|U_k^{\text{true}}\|^2 + \|c(X_k)\| & \text{if } \|U_k^{\text{true}}\|^2 \geq \kappa_{uv}\|V_k\|^2 \\ \|c(X_k)\| & \text{otherwise.} \end{cases}$$

The following Lemmas 1 and 2 show that there exists a common quantity that both bounds from above the squared-norm of the true search direction plus the constraint violation and bounds from below the reduction in the model of the merit function. Essentially, the combination of these results shows that the search direction offers sufficient decrease relative to its norm. Here, Lemma 1 is stated using a different norm for $c(X_k)$ than in [2, Lemma 2.11]. The result holds in the same manner (with a different value for $\kappa_\Psi$) due to the norm equivalence between $\|\cdot\|$ and $\|\cdot\|_1$ in $\mathbb{R}^m$. We state the result in this manner for consistency with the measure used in our final results.

**Lemma 1** ([2, Lemma 2.11]) *Let Assumptions 1 and 2 hold. Then, there exists* $\kappa_\Psi \in \mathbb{R}_{>0}$ *such that, for all* $k \in [k_{\max}]$*, the true search direction and constraint violation satisfy* $\|D_k^{true}\|^2 + \|c(X_k)\|_1 \leq (\kappa_\Psi + 1)\Psi_k$.

**Lemma 2** ([2, Lemma 2.12]) *Let Assumptions 1 and 2 hold. Then, there exists* $\kappa_q \in \mathbb{R}_{>0}$ *such that, for all* $k \in [k_{\max}]$ *and* $\mathcal{T} \in \mathbb{R}_{>0}$ *with* $\mathcal{T} \leq \mathcal{T}_k^{trial,true}$*, one finds* $\Delta q(X_k, \mathcal{T}, \nabla f(X_k), H_k, D_k^{true}) \geq \kappa_q \mathcal{T} \Psi_k$.

The next lemma shows that the reduction in the merit function is at least a reduction in the model of the merit function defined with respect to the true gradient and true search direction, except for two terms that may be attributed to the noise in the stochastic gradient estimates. Observe that the requirement in the lemma that $\beta_k \Xi_k \mathcal{T}_k / (\mathcal{T}_k L + \Gamma) \in (0, 1]$ for all $k \in [k_{\max}]$ can be enforced in practice, despite the fact that $\{\Xi_k\}$ and $\{\mathcal{T}_k\}$ evolve randomly. After all, since $\{\Xi_k\}$ and $\{\mathcal{T}_k\}$ are monotonically nonincreasing, one need only choose $\beta \in \mathbb{R}_{>0}$ sufficiently small such that $\beta \xi_{-1} \tau_{-1} / (\tau_{-1} L + \Gamma) \in (0, 1]$ to find that $\beta_k = \beta$ for all $k \in [k_{\max}]$ satisfies the requirement.

**Lemma 3** ([2, Lemma 3.7]) *Let Assumptions 1 and 2 hold and suppose that* $\beta_k \Xi_k \mathcal{T}_k / (\mathcal{T}_k L + \Gamma) \in (0, 1]$ *for all* $k \in [k_{\max}]$*. Then, for all* $k \in [k_{\max}]$*, one finds*

$$\begin{aligned} \phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \leq & -\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{true}) \\ & + \tfrac{1}{2} \mathcal{A}_k \beta_k \Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k) \\ & + \mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^\top (D_k - D_k^{true}). \end{aligned} \tag{21}$$

The next two lemmas bound (in expectation) differences and products between stochastic and true quantities. These bounds are critical in the analysis.

**Lemma 4** *Let Assumptions 1, 2, and 3 hold. Then, for all* $k \in [k_{\max}]$*, it follows that* $\mathbb{E}_k[D_k] = D_k^{true}$, $\mathbb{E}_k[U_k] = U_k^{true}$*, and* $\mathbb{E}_k[Y_k] = Y_k^{true}$*. Moreover, there exists* $\kappa_g \in \mathbb{R}_{>0}$ *and* $\kappa_d \in \mathbb{R}_{>0}$ *such that, for all* $k \in [k_{\max}]$*, one finds*

$$\begin{aligned} \|\nabla f(X_k)\| &\leq \kappa_g, \\ \|D_k^{true}\| \leq \kappa_d \|\nabla f(X_k)\| &\leq \kappa_d \kappa_g, \\ \mathbb{E}_k[\|D_k - D_k^{true}\|] \leq \kappa_d \mathbb{E}_k[\|G_k - \nabla f(X_k)\|] &\leq \kappa_d \sqrt{M}, \quad \text{and} \\ \mathbb{E}_k[\|D_k - D_k^{true}\| | \nabla f(X_k)^\top (D_k - D_k^{true}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] &\leq \kappa_d M_\tau. \end{aligned}$$

***Proof*** Except for the last inequality, the result follows from the assumptions and (the proof of) [2, Lemma 3.8]. As for the last inequality, observe as in [2, Lemma 3.8] that $\|D_k - D_k^{\text{true}}\| \leq \kappa_d \|G_k - \nabla f(X_k)\|$, which with (19) gives

$$\begin{aligned} &\mathbb{E}_k[\|D_k - D_k^{\text{true}}\| | \nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] \\ &\quad \leq \kappa_d \mathbb{E}_k[\|G_k - \nabla f(X_k)\| | \nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}] \leq \kappa_d M_\tau, \end{aligned}$$

as desired. □

**Lemma 5** ([2, Lemma 3.9]) *Let Assumptions 1, 2, and 3 hold. Then, for all* $k \in [k_{\max}]$, *it follows that*

$$\nabla f(X_k)^\top D_k^{true} \geq \mathbb{E}_k[G_k^\top D_k] \geq \nabla f(X_k)^\top D_k^{true} - \zeta^{-1} M$$
$$\text{and } \mathbb{E}_k[D_k^\top H_k D_k] \geq (D_k^{true})^\top H_k D_k^{true}.$$

### 4.4 Complexity result

In this section, we present our main complexity results. We derive our results in largely the same manner as the global convergence result in [2], but with two major changes that stem from the need to characterize the behavior of the algorithm in the context of an adaptive merit parameter sequence. At a high level, the two modifications are as follows:

1. We derive, in Lemma 6, an upper bound for the last term in (21), the derivation of which is complicated by the fact that, conditioned on $\mathcal{F}_k$, this term is the product of three correlated random variables: $\mathcal{A}_k$, $\mathcal{T}_k$, and $\nabla f(X_k)^\top (D_k - D_k^{\text{true}})$. A critical aspect of our derived bound is that we isolate a term for the event when $\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0$ and $\mathcal{T}_k < \mathcal{T}_{k-1}$, since this happens to be an event that complicates subsequent aspects of our analysis. In Lemma 9, we prove a high-probability bound on the sum of the probabilities of the occurrences of this event over a run of the algorithm.
2. A critical aspect of the analysis in [2] for the deterministic setting is that one can always tie the reduction in the model of the merit function to a first-order stationarity error measure (with respect to the constrained optimization problem) due to the fact that $\mathcal{T}_k^{\text{trial,true}} \geq \mathcal{T}_k$ for all $k \in \mathbb{N}$. Unfortunately, however, this inequality is not guaranteed to hold in the stochastic setting, which is problematic for our purposes in this paper. To account for this issue, we define an auxiliary sequence $\{\hat{\mathcal{T}}_k\}$ (not generated by the algorithm) such that $\hat{\mathcal{T}}_k := \min\{\mathcal{T}_k, \mathcal{T}_k^{\text{trial,true}}\}$ for all $k \in [k_{\max}]$. In Lemmas 7 and 8, we analyze behaviors of the algorithm with respect to this auxiliary sequence, and in Lemma 9 we provide a high-probability bound on the total number of iterations in which $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_k$ may occur. (More precisely, Lemma 9 considers a superset of the iterations in which this bound may occur, which serves our purposes just as well.)

The first few results in this section consider properties of algorithmic quantities conditioned on $\mathcal{F}_k$. Conditioned on $\mathcal{F}_k$, let us define three events:

- $E_{k,1}$, the event that $\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) \geq 0$;
- $E_{k,2}$, the event that $\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0$ and $\mathcal{T}_k = \mathcal{T}_{k-1}$; and
- $E_{k,3}$, the event that $\nabla f(X_k)^\top (D_k - D_k^{\text{true}}) < 0$ and $\mathcal{T}_k < \mathcal{T}_{k-1}$.

We now derive an upper bound on the final term in (21). In the following lemma, we use, given $k \in [k_{\max}]$, the stepsize values

$$\alpha_{\min,k}^{<} := \frac{\beta_k \xi_{\min} \tau_{\min}}{\tau_{\min} L + \Gamma}, \quad \mathcal{A}_{\min,k}^{=} := \frac{\beta_k \xi_{\min} \mathcal{T}_{k-1}}{\mathcal{T}_{k-1} L + \Gamma}, \tag{22}$$
$$\text{and } \mathcal{A}_{\max,k} := \mathcal{A}_{\min,k}^{=} + \theta \beta_k^2.$$

The first value here represents a lower bound on the smallest stepsize that may be computed in the event that $\mathcal{T}_k < \mathcal{T}_{k-1}$, whereas the second value is the smallest stepsize that may be computed in the event that $\mathcal{T}_k = \mathcal{T}_{k-1}$; it is easily verified that $\alpha^{<}_{\min,k} < \mathcal{A}^{=}_{\min,k}$. Hence, $\mathcal{A}_{\max,k}$ represents an upper bound on the largest stepsize that may be computed.

**Lemma 6** *Suppose that Assumptions 1, 2, and 3 hold, and let $\kappa_d \in \mathbb{R}_{>0}$ be defined by Lemma 4. Then, for all $k \in [k_{\max}]$, and with the stepsizes $(\alpha^{<}_{\min,k}, \mathcal{A}^{=}_{\min,k}, \mathcal{A}_{\max,k})$ defined as in (22), one finds that*

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{true})]\\ &\quad\leq\ (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta_k^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M}.\end{aligned}$$

***Proof*** Consider arbitrary $k \in [k_{\max}]$, and for ease of exposition, let us denote $\mathbb{E}_{k,j} = \mathbb{E}_k[\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,j}]$ for all $j \in \{1, 2, 3\}$. By the Law of Total Expectation, the fact that $0 < \tau_{\min} \leq \mathcal{T}_k \leq \mathcal{T}_{k-1}$ under $E$, (20), and the definitions of $\alpha^{<}_{\min,k}$, $\mathcal{A}^{=}_{\min,k}$, and $\mathcal{A}_{\max,k}$, one finds that

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})]\\ &\quad=\ \mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,1}]\mathbb{P}_k[E_{k,1}]\\ &\qquad+\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,2}]\mathbb{P}_k[E_{k,2}]\\ &\qquad+\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,3}]\mathbb{P}_k[E_{k,3}]\\ &\quad\leq\ \mathcal{A}_{\max,k}\mathcal{T}_{k-1}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + \mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}]\\ &\qquad+\alpha^{<}_{\min,k}\tau_{\min}\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}].\end{aligned}$$

Using this inequality, the Law of Total Expectation, (20), and Lemma 4 ($\mathbb{E}_k[D_k] = D_k^{\text{true}}$), one obtains three upper bounds by adding and subtracting like terms:

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})]\\ &\quad\leq\ \mathcal{A}_{\max,k}\mathcal{T}_{k-1}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + \mathcal{A}_{\max,k}\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}]\\ &\qquad-\theta\beta_k^2\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}]\\ &\qquad+\mathcal{A}_{\max,k}\mathcal{T}_{k-1}\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] + (\alpha^{<}_{\min,k}\tau_{\min} - \mathcal{A}_{\max,k}\mathcal{T}_{k-1})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}]\\ &\quad=\ -\theta\beta_k^2\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] + (\alpha^{<}_{\min,k}\tau_{\min} - \mathcal{A}_{\max,k}\mathcal{T}_{k-1})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}]\\ &\text{and}\quad\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})]\\ &\quad\leq\ \mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + \theta\beta_k^2\mathcal{T}_{k-1}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}]\\ &\qquad+\mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}]\\ &\qquad+\mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1}\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] + (\alpha^{<}_{\min,k}\tau_{\min} - \mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}]\\ &\quad=\ \theta\beta_k^2\mathcal{T}_{k-1}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + (\alpha^{<}_{\min,k}\tau_{\min} - \mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}]\\ &\text{and}\quad\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})]\\ &\quad\leq\ \alpha^{<}_{\min,k}\tau_{\min}\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}]\end{aligned}$$

$$
\begin{aligned}
&+ \alpha^{<}_{\min,k}\tau_{\min}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] + (\mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] \\
&+ \alpha^{<}_{\min,k}\tau_{\min}\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] \\
&= (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] \\
&\quad + (\mathcal{A}^{=}_{\min,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}].
\end{aligned}
$$

From averaging these upper bounds And the definition of $\mathcal{A}_{\max,k}$, one obtains

$$
\begin{aligned}
&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})] \\
&\leq \tfrac{1}{3}((\mathcal{A}^{=}_{\min,k} + 2\theta\beta_k^2)\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] \\
&\quad + \tfrac{1}{3}((\mathcal{A}^{=}_{\min,k} - \theta\beta_k^2)\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] \\
&\quad + \tfrac{1}{3}(2\alpha^{<}_{\min,k}\tau_{\min} - (2\mathcal{A}^{=}_{\min,k} + \theta\beta_k^2)\mathcal{T}_{k-1})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] \\
&= \tfrac{1}{3}((\mathcal{A}^{=}_{\min,k} + 2\theta\beta_k^2)\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})(\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + \mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}]) \\
&\quad - \theta\beta_k^2\mathcal{T}_{k-1}\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] \\
&\quad - \tfrac{1}{3}((2\mathcal{A}^{=}_{\min,k} + \theta\beta_k^2)\mathcal{T}_{k-1} - 2\alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}].
\end{aligned} \tag{23}
$$

This bound can be rewritten as follows. By the Law of Total Expectation, Lemma 4, and $\nabla f(X_k) \in \mathcal{F}_k$, one finds that

$$
\begin{aligned}
&\mathbb{E}_{k,1}\mathbb{P}_k[E_{k,1}] + \mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] \\
&= \mathbb{E}_k[\nabla f(X_k)^\top(D_k - D_k^{\text{true}})] - \mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] = -\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}],
\end{aligned} \tag{24}
$$

and along with Lemma 4 one finds that

$$
\begin{aligned}
-\mathbb{E}_{k,2}\mathbb{P}_k[E_{k,2}] &= -\mathbb{E}_k[\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,2}]\mathbb{P}_k[E_{k,2}] \\
&\leq \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\||E_{k,2}]\mathbb{P}_k[E_{k,2}] \\
&= \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\|] \\
&\quad - \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\||E_{k,1}]\mathbb{P}_k[E_{k,1}] \\
&\quad - \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\||E_{k,3}]\mathbb{P}_k[E_{k,3}] \\
&\leq \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\|] \\
&\leq \kappa_g\mathbb{E}_k[\|D_k - D_k^{\text{true}}\|] \leq \kappa_g\kappa_d\sqrt{M}.
\end{aligned} \tag{25}
$$

In addition, Lemma 4 also yields that

$$
\begin{aligned}
-\mathbb{E}_{k,3}\mathbb{P}_k[E_{k,3}] &= -\mathbb{E}_k[\nabla f(X_k)^\top(D_k - D_k^{\text{true}})|E_{k,3}]\mathbb{P}_k[E_{k,3}] \\
&\leq \mathbb{E}_k[\|\nabla f(X_k)\|\|D_k - D_k^{\text{true}}\||E_{k,3}]\mathbb{P}_k[E_{k,3}] \\
&\leq \kappa_g\kappa_d M_\tau\mathbb{P}_k[E_{k,3}].
\end{aligned} \tag{26}
$$

Combining (23), (24), (25), and (26), the desired result follows. □

Now, for all $k \in [k_{\max}]$, let us define

$$\hat{\mathcal{T}}_k := \min\{\mathcal{T}_k, \mathcal{T}_k^{\text{trial,true}}\}. \tag{27}$$

**Lemma 7** *Suppose that Assumptions 1, 2, and 3 hold and let $\kappa_d \in \mathbb{R}_{>0}$ be defined by Lemma 4. Then, for all $k \in [k_{\max}]$, one finds that*

$$\begin{aligned}
&\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\quad \le \mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{true})] + \tfrac{1}{2}(\mathcal{T}_{k-1} + \tau_{\min})\zeta^{-1}M \\
&\qquad + (\mathcal{T}_{k-1} - \tau_{\min})(\kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2)).
\end{aligned}$$

***Proof*** By the definition of $\Delta q$ in (8), one has that

$$\begin{aligned}
&\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\quad = \mathbb{E}_k[-\hat{\mathcal{T}}_k(G_k^\top D_k + \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\}) + \|c(X_k)\|_1] \\
&\quad = \mathbb{E}_k[-\hat{\mathcal{T}}_k(G_k^\top D_k - \nabla f(X_k)^\top D_k^{\text{true}} + \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\} \\
&\qquad\quad - \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\})] \\
&\qquad + \mathbb{E}_k[-\hat{\mathcal{T}}_k(\nabla f(X_k)^\top D_k^{\text{true}} + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}) + \|c(X_k)\|_1] \\
&\quad = \mathbb{E}_k[\hat{\mathcal{T}}_k(\nabla f(X_k)^\top D_k^{\text{true}} - G_k^\top D_k + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} \\
&\qquad\quad - \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\})] + \mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})].
\end{aligned} \tag{28}$$

Now, for simplicity of notation, define

$$\begin{aligned}
Q_k :=\ & \nabla f(X_k)^\top D_k^{\text{true}} - G_k^\top D_k \\
& + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} - \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\}.
\end{aligned}$$

Let $E_Q$ denote the event that $Q_k \ge 0$ occurs and let $E_Q^c$ denote the event that $Q_k < 0$ occurs. By the Law of Total Expectation and (20), one has that

$$\begin{aligned}
\mathbb{E}_k[\hat{\mathcal{T}}_k Q_k] &= \mathbb{E}_k[\hat{\mathcal{T}}_k Q_k | E_Q]\mathbb{P}_k[E_Q] + \mathbb{E}_k[\hat{\mathcal{T}}_k Q_k | E_Q^c]\mathbb{P}_k[E_Q^c] \\
&\le \mathcal{T}_{k-1}\mathbb{E}_k[Q_k | E_Q]\mathbb{P}_k[E_Q] + \tau_{\min}\mathbb{E}_k[Q_k | E_Q^c]\mathbb{P}_k[E_Q^c].
\end{aligned}$$

Therefore, by the Law of Total Probability, Lemma 5, (20), Jensen's inequality, and convexity of $\max\{\cdot, 0\}$, it follows that

$$\begin{aligned}
\mathbb{E}_k[\hat{\mathcal{T}}_k Q_k] \le\ & \mathcal{T}_{k-1}\mathbb{E}_k[Q_k | E_Q]\mathbb{P}_k[E_Q] + \mathcal{T}_{k-1}\mathbb{E}_k[Q_k | E_Q^c]\mathbb{P}_k[E_Q^c] \\
& + (\tau_{\min} - \mathcal{T}_{k-1})\mathbb{E}_k[Q_k | E_Q^c]\mathbb{P}_k[E_Q^c] \\
=\ & \mathcal{T}_{k-1}\mathbb{E}_k[Q_k] + (\tau_{\min} - \mathcal{T}_{k-1})\mathbb{E}_k[Q_k | E_Q^c]\mathbb{P}_k[E_Q^c] \\
=\ & \mathcal{T}_{k-1}(\nabla f(X_k)^\top D_k^{\text{true}} - \mathbb{E}_k[G_k^\top D_k] \\
& + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} - \tfrac{1}{2}\mathbb{E}_k[\max\{D_k^\top H_k D_k, 0\}])
\end{aligned}$$

$$
\begin{aligned}
&+ (\tau_{\min} - \mathcal{T}_{k-1})\mathbb{E}_k[Q_k|E_Q^c]\mathbb{P}_k[E_Q^c] \\
&\leq \mathcal{T}_{k-1}\zeta^{-1}M + (\tau_{\min} - \mathcal{T}_{k-1})\mathbb{E}_k[Q_k|E_Q^c]\mathbb{P}_k[E_Q^c],
\end{aligned}
$$

and by similar reasoning one finds that

$$
\begin{aligned}
\mathbb{E}_k[\hat{\mathcal{T}}_k Q_k] &\leq \tau_{\min}\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q] + \tau_{\min}\mathbb{E}_k[Q_k|E_Q^c]\mathbb{P}_k[E_Q^c] \\
&\quad + (\mathcal{T}_{k-1} - \tau_{\min})\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q] \\
&= \tau_{\min}\mathbb{E}_k[Q_k] + (\mathcal{T}_{k-1} - \tau_{\min})\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q] \\
&\leq \tau_{\min}\zeta^{-1}M + (\mathcal{T}_{k-1} - \tau_{\min})\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q].
\end{aligned}
$$

Averaging these two upper bounds, one finds that

$$
\begin{aligned}
\mathbb{E}_k[\hat{\mathcal{T}}_k Q_k] \leq \tfrac{1}{2}(\mathcal{T}_{k-1} + \tau_{\min})\zeta^{-1}M + \tfrac{1}{2}(\mathcal{T}_{k-1} - \tau_{\min})\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q] \\
+ \tfrac{1}{2}(\tau_{\min} - \mathcal{T}_{k-1})\mathbb{E}_k[Q_k|E_Q^c]\mathbb{P}_k[E_Q^c].
\end{aligned} \tag{29}
$$

Our goal now is to bound the latter two terms in (29). Toward this end, observe that by the triangle and Cauchy-Schwarz inequalities, the proof of Lemma 4, Assumption 3, (20), Jensen's inequality, and concavity of the square root over $\mathbb{R}_{\geq 0}$,

$$
\begin{aligned}
&\mathbb{E}_k[|\nabla f(X_k)^\top D_k^{\text{true}} - G_k^\top D_k|] \\
&\leq \mathbb{E}_k[|\nabla f(X_k)^\top D_k^{\text{true}} - G_k^\top D_k^{\text{true}}|] + \mathbb{E}_k[|G_k^\top D_k^{\text{true}} - G_k^\top D_k|] \\
&\leq \|D_k^{\text{true}}\|\mathbb{E}_k[\|\nabla f(X_k) - G_k\|] + \mathbb{E}_k[\|G_k\|\|D_k^{\text{true}} - D_k\|] \\
&\leq \kappa_d\kappa_g\sqrt{\mathbb{E}_k[\|\nabla f(X_k) - G_k\|^2]} + \kappa_d\mathbb{E}_k[\|G_k\|\|\nabla f(X_k) - G_k\|] \\
&\leq \kappa_d\kappa_g\sqrt{\mathbb{E}_k[\|\nabla f(X_k) - G_k\|^2]} \\
&\quad + \kappa_d\mathbb{E}_k[(\|G_k - \nabla f(X_k)\| + \|\nabla f(X_k)\|)\|\nabla f(X_k) - G_k\|] \\
&\leq \kappa_d\kappa_g\sqrt{M} + \kappa_d(M + \kappa_g\sqrt{M}) = \kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}).
\end{aligned} \tag{30}
$$

In addition, by the Cauchy-Schwarz inequality, the proof of Lemma 4, Assumption 3, (20), and since $\|a\|^2 \leq 2(\|a-b\|^2 + \|b\|^2)$ for any $(a,b) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$
\begin{aligned}
&\mathbb{E}_k\left[\left|\tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} - \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\}\right|\right] \\
&\leq |\tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}| + \mathbb{E}_k\left[\left|\tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\}\right|\right] \\
&\leq \tfrac{1}{2}\|H_k\|\|D_k^{\text{true}}\|^2 + \tfrac{1}{2}\|H_k\|\mathbb{E}_k[\|D_k\|^2] \\
&\leq \tfrac{1}{2}\|H_k\|\|D_k^{\text{true}}\|^2 + \tfrac{1}{2}\kappa_d^2\|H_k\|\mathbb{E}_k[\|G_k\|^2] \\
&\leq \tfrac{1}{2}\|H_k\|\|D_k^{\text{true}}\|^2 + \kappa_d^2\|H_k\|\mathbb{E}_k[\|G_k - \nabla f(X_k)\|^2 + \|\nabla f(X_k)\|^2] \\
&\leq \tfrac{1}{2}\kappa_H\kappa_d^2\kappa_g^2 + \kappa_H\kappa_d^2(M + \kappa_g^2) = \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2).
\end{aligned} \tag{31}
$$

By the Law of Total Expectation, (30), and (31), it follows that

$$\begin{aligned}\mathbb{E}_k[Q_k|E_Q]\mathbb{P}_k[E_Q] &= \mathbb{E}_k[|Q_k||E_Q]\mathbb{P}_k[E_Q] \\ &= \mathbb{E}_k[|Q_k|] - \mathbb{E}_k[|Q_k||E_Q^c]\mathbb{P}_k[E_Q^c] \\ &\leq \kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2),\end{aligned}$$

and by a similar argument, one finds that

$$\begin{aligned}-\mathbb{E}_k[Q_k|E_Q^c]\mathbb{P}_k[E_Q^c] &= \mathbb{E}_k[|Q_k||E_Q^c]\mathbb{P}_k[E_Q^c] \\ &= \mathbb{E}_k[|Q_k|] - \mathbb{E}_k[|Q_k||E_Q]\mathbb{P}_k[E_Q] \\ &\leq \kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2\left(M + \tfrac{3}{2}\kappa_g^2\right).\end{aligned}$$

The conclusion follows by combining these equations, (28), and (29). □

Our next lemma bounds differences between expected reductions in the model of the merit function that account for cases when $\hat{\mathcal{T}}_k < \mathcal{T}_k$.

**Lemma 8** *Let Assumptions 1, 2, and 3 hold and let $\kappa_d \in \mathbb{R}_{>0}$ be defined by Lemma 4. Then, for all $k \in [k_{\max}]$, with $(\alpha^{<}_{\min,k}, \mathcal{A}^{=}_{\min,k}, \mathcal{A}_{\max,k})$ defined as in (22) and $\hat{\mathcal{T}}_k$ defined in (27), one finds that*

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{true})] - \mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{true})] \\ &\quad\leq\ (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_d\kappa_g^2(1 + \tfrac{1}{2}\kappa_H\kappa_d)\end{aligned}$$

*and*

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)] - \mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\ &\quad\leq\ (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_d(2 + \kappa_H\kappa_d)(M + \kappa_g^2).\end{aligned}$$

***Proof*** Under the stated assumptions and definitions, one finds that

$$\mathcal{A}_k(\mathcal{T}_k - \hat{\mathcal{T}}_k) = \mathcal{A}_k\mathcal{T}_k - \mathcal{A}_k\hat{\mathcal{T}}_k \leq \mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min}.$$

Hence, under the stated assumptions, it follows from the stated lemma and definitions, along with the definition of $\Delta q$ in (8) and equation (20), that

$$\begin{aligned}&\mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] - \mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] \\ &\quad=\ \mathbb{E}_k[\mathcal{A}_k(\mathcal{T}_k - \hat{\mathcal{T}}_k)(\nabla f(X_k)^\top D_k^{\text{true}} + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\})] \\ &\quad\leq\ (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})|\nabla f(X_k)^\top D_k^{\text{true}} + \tfrac{1}{2}\max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}| \\ &\quad\leq\ (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})(\kappa_d\kappa_g^2 + \tfrac{1}{2}\kappa_H\kappa_d^2\kappa_g^2),\end{aligned}$$

and, along with $\|a\|^2 \leq 2(\|a-b\|^2 + \|b\|^2)$ for any $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, one finds

$$\mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)] - \mathbb{E}_k[\mathcal{A}_k\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)]$$

$$
\begin{aligned}
&= \mathbb{E}_k[\mathcal{A}_k(\hat{\mathcal{T}}_k - \mathcal{T}_k)(G_k^\top D_k + \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\})] \\
&\leq (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\mathbb{E}_k[|G_k^\top D_k + \tfrac{1}{2}\max\{D_k^\top H_k D_k, 0\}|] \\
&\leq (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})(\kappa_d + \tfrac{1}{2}\kappa_d^2\kappa_H)\mathbb{E}_k[\|G_k\|^2] \\
&\leq (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})(\kappa_d + \tfrac{1}{2}\kappa_d^2\kappa_H)(2(M + \kappa_g^2)),
\end{aligned}
$$

which together are the desired conclusions. □

Our next lemma is a critical element of our analysis. For any $s \in \mathbb{N}$ and $\delta \in (0, 1)$, let

$$
\hat{\delta} := \frac{\delta}{\sum_{j=0}^{\max\{s_{\max}-1,0\}} \binom{k_{\max}}{j}} \tag{32}
$$

and

$$
\ell(s, \hat{\delta}) := s + \log(1/\hat{\delta}) + \sqrt{\log(1/\hat{\delta})^2 + 2s\log(1/\hat{\delta})}. \tag{33}
$$

This lemma provides a bound on two quantities with high probability, the first of which is the sum of the probabilities of the occurrences of event $E_{k,3}$ over the run of the algorithm.

The following lemma also provides a bound on the cardinality of the random index set,

$$
\mathcal{K}_\tau := \left\{k \in [k_{\max}] : \mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}\right\}. \tag{34}
$$

By the manner in which $\{\mathcal{T}_k\}$, $\{\mathcal{T}_k^{\text{trial,true}}\}$, and $\{\hat{\mathcal{T}}_k\}$ are defined, this set is always a superset of the iterations in which $\hat{\mathcal{T}}_k < \mathcal{T}_k$; hence, by bounding the cardinality of (34), one bounds the cardinality of the set of iterations in which $\hat{\mathcal{T}}_k < \mathcal{T}_k$, which is needed for our main theorem. The reason that we consider the set $\mathcal{K}_\tau$ in (34) is the fact that the event $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$ and its complement are members of $\mathcal{F}_k$; in other words, the occurrence of the event defined by this inequality does not depend on $G_k$.

**Lemma 9** *Suppose Assumptions 1, 2, and 3 hold. Then, for any $\delta \in (0, 1)$,*

$$
\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}_k[E_{k,3}] \leq \ell(s_{\max}, \hat{\delta}) + 1 \,\middle|\, E\right] \geq 1 - \delta. \tag{35}
$$

*In addition, suppose Assumption 4 holds as well. Then, for any $\delta \in (0, 1)$, (35) holds and*

$$
\mathbb{P}\left[|\mathcal{K}_\tau| \leq \left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau} \right\rceil \,\middle|\, E\right] \geq 1 - \delta. \tag{36}
$$

***Proof*** This result is proved in Appendix B. □

Our proof of Lemma 9 uses a novel tree structure for analyzing the behavior of the adaptive merit parameter sequence in the context of an algorithm for solving constrained stochastic optimization problems; see also [1, Appendix B] for the use of such a tree structure in a different context. Starting with the initialization at a root node, each subsequent level of the tree captures different sets of realizations of the algorithm in terms of the number of decreases of the merit parameter and the probability that the parameter will be decreased further in the next iteration. The leaves of the tree represent either *bad* situations when the sum of the probabilities that the merit parameter could decrease exceeds a critical threshold (defined with respect to the function $\ell$ defined in (33)) or *good* situations in which this sum remains below the threshold and the maximum number of merit parameter decreases has occurred and/or the iteration limit has been reached. Essentially, in this manner, *bad* situations are when the algorithm repeatedly has a high probability of decreasing the merit parameter, but does not do so a sufficient number of times. The proof of Lemma 9 ultimately relies on applications of Chernoff's bound—employed with respect to independent random variables that are defined carefully with respect to the non-independent random variables in the stochastic process defined by the algorithm—to show that the probability is small that the algorithm ends at a *bad* leaf node.

We are now prepared to prove a convergence rate result.

**Theorem 2** *Suppose Assumptions 1, 2, 3, and 4 hold, let* $s_{\max} \in \mathbb{N} \setminus \{0\}$*, let* $\kappa_d \in \mathbb{R}_{>0}$ *be defined by Lemma 4, define*

$$A_{\min} := \frac{\xi_{\min}\tau_{\min}}{\tau_{\min}L + \Gamma} \quad \text{and} \quad A_{\max} := \frac{\xi_{-1}\tau_{-1}}{\tau_{-1}L + \Gamma},$$

*suppose that* $\beta_k = \beta$ *for all* $k \in [k_{\max}]$ *where*

$$\beta := \frac{\gamma}{\sqrt{k_{\max}+1}} \quad \text{for some} \quad \gamma \in \left(0, \frac{\min\{1, A_{\min}\}}{A_{\max} + \theta}\right], \tag{37}$$

*define*

$$\begin{aligned}
\overline{M} :=\ & \tfrac{1}{4}(A_{\max} + \theta\beta)(\tau_{-1} + \tau_{\min})\zeta^{-1}M \\
& + \tfrac{1}{2}(A_{\max} + \theta\beta)(\tau_{-1} - \tau_{\min})(\kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2)) \\
& + \theta\tau_{-1}\kappa_g\kappa_d\sqrt{M} \\
\kappa_{E_3} :=\ & ((A_{\max} + \theta\beta)\tau_{-1} - A_{\min}\tau_{\min})\kappa_g\kappa_d M_\tau, \\
\kappa_{\Delta q,1} :=\ & ((A_{\max} + \theta\beta)\tau_{-1} - A_{\min}\tau_{\min})\kappa_d\kappa_g^2(1 + \tfrac{1}{2}\kappa_H\kappa_d) \quad \text{and} \\
\kappa_{\Delta q,2} :=\ & ((A_{\max} + \theta\beta)\tau_{-1} - A_{\min}\tau_{\min})\kappa_d(1 + \tfrac{1}{2}\kappa_H\kappa_d)(M + \kappa_g^2),
\end{aligned}$$

*and, for all* $k \in [k_{\max}]$*, let* $\hat{\mathcal{I}}_k$ *be defined as in (27). Then, for any* $\delta \in (0, 1)$*, it follows with* $K^*$ *having a discrete uniform distribution over* $[k_{\max}]$ *and* $\hat{\delta}$ *and* $\ell$ *defined as in*

(32) *and* (33) *that, with probability at least* $1-\delta$,

$$\begin{aligned}
&\mathbb{E}[\Delta q(X_{K^*}, \hat{\mathcal{T}}_{K^*}, \nabla f(X_{K^*}), H_{K^*}, D_{K^*}^{\text{true}})|E] \\
&\quad \leq 2\left(\frac{\tau_{-1}(f_0 - f_{\min}) + \|c_0\|_1 + \overline{M}\gamma^2 + \kappa_{E_3}\gamma(\ell(s_{\max}, \hat{\delta}) + 1)/\sqrt{k_{\max}+1}}{A_{\min}\gamma\sqrt{k_{\max}+1}}\right) \\
&\qquad + \frac{2(\kappa_{\Delta q,1}\gamma + \kappa_{\Delta q,2}\gamma^2/\sqrt{k_{\max}+1})}{A_{\min}\gamma(k_{\max}+1)}\left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau}\right\rceil. \qquad (38)
\end{aligned}$$

***Proof*** First, consider arbitrary $k \in [k_{\max}]$. By Lemmas 3 and 6 and equation (20), one has that

$$\begin{aligned}
&\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)] \\
&\quad \leq \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k) \\
&\qquad + \mathcal{A}_k\mathcal{T}_k\nabla f(X_k)^\top(D_k - D_k^{\text{true}})] \\
&\quad \leq \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)] \\
&\qquad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M}. \qquad (39)
\end{aligned}$$

Our next aim is to prove that, roughly speaking, one in fact finds that

$$\begin{aligned}
&\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)] \\
&\quad \leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] + \text{“noise.”} \qquad (40)
\end{aligned}$$

Such a bound does not follow directly from (39) since the first term on the right-hand side in (39) involves a model reduction with respect to $\mathcal{T}_k$ (which cannot be tied to a stationarity measure), whereas the first term on the right-hand side of the bound in (40) involves a model reduction with respect to $\hat{\mathcal{T}}_k$ (which can be tied to a stationarity measure). Toward the aim of proving a bound of the form in (40), first observe that it follows with Lemma 7 that

$$\begin{aligned}
&\mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\qquad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\quad \leq -A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] \\
&\qquad + \tfrac{1}{2}(A_{\max}\beta + \theta\beta^2)\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\qquad + ((A_{\max}\beta + \theta\beta^2)\mathcal{T}_{k-1} - A_{\min}\beta\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\quad \leq -(A_{\min} - \tfrac{1}{2}(A_{\max} + \theta\beta)\beta)\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] \\
&\qquad + \tfrac{1}{4}(A_{\max} + \theta\beta)\beta^2(\mathcal{T}_{k-1} + \tau_{\min})\zeta^{-1}M \\
&\qquad + \tfrac{1}{2}(A_{\max} + \theta\beta)\beta^2(\mathcal{T}_{k-1} - \tau_{\min})(\kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2)) \\
&\qquad + ((A_{\max} + \theta\beta)\mathcal{T}_{k-1} - A_{\min}\tau_{\min})\beta\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M}
\end{aligned}$$

$$\begin{aligned} &\leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] \\ &\quad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}] + \overline{M}\beta^2, \end{aligned} \tag{41}$$

where the final inequality follows due to the fact that $(A_{\max} + \theta\beta)\beta \leq A_{\min}$ holds by the definitions of $\beta$, $\gamma$, $A_{\min}$, and $A_{\max}$.

Let us now combine (39) and (41) to prove a bound of the form in (40) by considering two complementary events. In particular, let $E_{k,\tau}$ be the event that $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$ and let $E_{k,\tau}^c$ be the event that $\mathcal{T}_k^{\text{trial,true}} \geq \mathcal{T}_{k-1}$. Observe that $E_{k,\tau}$ and $E_{k,\tau}^c$ only depend on the history of the algorithm prior to iteration $k$ and thus the $\sigma$-algebras generated by $E_{k,\tau}$ and $E_{k,\tau}^c$ are included in $\mathcal{F}_k$. Therefore, by [15, Theorem 4.1.13], for any random variable $Z$, we have

$$\mathbb{E}_k[Z|E_{k,\tau}] = \mathbb{E}_k[\mathbb{E}_k[Z]|E_{k,\tau}] \text{ and } \mathbb{E}_k[Z|E_{k,\tau}^c] = \mathbb{E}_k[\mathbb{E}_k[Z]|E_{k,\tau}^c]. \tag{42}$$

Hence, we can use Lemma 7—and Lemma 8 as well, which is used below—even if one conditions on the occurrence of $E_{k,\tau}$ or of $E_{k,\tau}^c$. Let us now consider $E_{k,\tau}^c$ and $E_{k,\tau}$ in turn. Conditioning on $E_{k,\tau}^c$, one finds from (39), (41), (42), and the fact that $\mathcal{T}_k^{\text{trial,true}} \geq \mathcal{T}_{k-1} \geq \mathcal{T}_k = \hat{\mathcal{T}}_k$ (by (27)) in $E_{k,\tau}^c$ that

$$\begin{aligned} &\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E_{k,\tau}^c] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)|E_{k,\tau}^c] \\ &\quad = \mathbb{E}_k[\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k)]|E_{k,\tau}^c] \\ &\quad \leq \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\ &\qquad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)|E_{k,\tau}^c] \\ &\qquad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha_{\min,k}^{<}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}|E_{k,\tau}^c] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M} \\ &\quad = \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\ &\qquad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)|E_{k,\tau}^c] \\ &\qquad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha_{\min,k}^{<}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}|E_{k,\tau}^c] + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M} \\ &\quad \leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E_{k,\tau}^c] \\ &\qquad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}|E_{k,\tau}^c] + \overline{M}\beta^2. \end{aligned}$$

On the other hand, one finds from (39), (41), (42), and Lemma 8 that

$$\begin{aligned} &\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E_{k,\tau}] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)|E_{k,\tau}] \\ &\quad \leq \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\ &\qquad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)|E_{k,\tau}] \\ &\qquad + \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\ &\qquad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)|E_{k,\tau}] \\ &\qquad - \mathbb{E}_k[-\mathcal{A}_k \Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \end{aligned}$$

$$
\begin{aligned}
&\quad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)|E_{k,\tau}] \\
&\quad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}|E_{k,\tau}] \\
&\quad + \theta\beta^2\mathcal{T}_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\leq \mathbb{E}_k[-\mathcal{A}_k\Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\
&\quad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \mathcal{T}_k, G_k, H_k, D_k)|E_{k,\tau}] \\
&\quad - \tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E_{k,\tau}] \\
&\quad - \mathbb{E}_k[-\mathcal{A}_k\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) \\
&\quad + \tfrac{1}{2}\mathcal{A}_k\beta\Delta q(X_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)|E_{k,\tau}] \\
&\quad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}|E_{k,\tau}] + \overline{M}\beta^2 \\
&\leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E_{k,\tau}] \\
&\quad + (\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_d\kappa_g^2(1 + \tfrac{1}{2}\kappa_H\kappa_d) \\
&\quad + \tfrac{1}{2}(\mathcal{A}_{\max,k}\mathcal{T}_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_d(2 + \kappa_H\kappa_d)(M + \kappa_g^2)\beta \\
&\quad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}|E_{k,\tau}] + \overline{M}\beta^2 \\
&\leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E_{k,\tau}] \\
&\quad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}|E_{k,\tau}] + \overline{M}\beta^2 + \kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2.
\end{aligned}
$$

Hence, by the laws of total probability and expectation, one finds that

$$
\begin{aligned}
&\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)] \\
&\quad = (\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E_{k,\tau}] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)|E_{k,\tau}])\mathbb{P}_k[E_{k,\tau}] \\
&\qquad + (\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E^c_{k,\tau}] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)|E^c_{k,\tau}])\mathbb{P}_k[E^c_{k,\tau}] \\
&\quad \leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] \\
&\qquad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}] + \overline{M}\beta^2 + (\kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2)\mathbb{P}_k[E_{k,\tau}].
\end{aligned}
$$

Summing this inequality for all $k \in [k_{\max}]$ yields

$$
\begin{aligned}
&\sum_{k=0}^{k_{\max}}(\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)]) \\
&\quad \leq \sum_{k=0}^{k_{\max}}(-\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})]) \\
&\qquad + \sum_{k=0}^{k_{\max}}(\kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}] + (\kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2)\mathbb{P}_k[E_{k,\tau}]) + (k_{\max} + 1)\overline{M}\beta^2.
\end{aligned}
$$

Next, by Lemma 9, it follows that, with probability at least $1 - \delta$, one finds

$$\begin{aligned}
&\sum_{k=0}^{k_{\max}} (\mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(X_k, \mathcal{T}_k)]) \\
&\quad \leq -\tfrac{1}{2} A_{\min} \beta \sum_{k=0}^{k_{\max}} \mathbb{E}_k[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})] + \kappa_{E_3} \beta (\ell(s_{\max}, \hat{\delta}) + 1) \\
&\qquad + (\kappa_{\Delta q,1} \beta + \kappa_{\Delta q,2} \beta^2) \left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau} \right\rceil + (k_{\max} + 1) \overline{M} \beta^2.
\end{aligned}$$

Taking the total expectation (conditioned on $E$) of the above inequality,

$$\begin{aligned}
&\sum_{k=0}^{k_{\max}} (\mathbb{E}[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E] - \mathbb{E}[\phi(X_k, \mathcal{T}_k)|E]) \\
&\quad \leq -\tfrac{1}{2} A_{\min} \beta \sum_{k=0}^{k_{\max}} \mathbb{E}[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E] + \kappa_{E_3} \beta (\ell(s_{\max}, \hat{\delta}) + 1) \\
&\qquad + (\kappa_{\Delta q,1} \beta + \kappa_{\Delta q,2} \beta^2) \left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau} \right\rceil + (k_{\max} + 1) \overline{M} \beta^2. \qquad (43)
\end{aligned}$$

The left-hand side of this inequality satisfies

$$\begin{aligned}
&\sum_{k=0}^{k_{\max}} (\mathbb{E}[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E] - \mathbb{E}[\phi(X_k, \mathcal{T}_k)|E]) \\
&\quad = \sum_{k=0}^{k_{\max}} \Big( \mathbb{E}[\mathcal{T}_k (f(X_k + \mathcal{A}_k D_k) - f_{\min}) + \|c(X_k + \mathcal{A}_k D_k)\|_1 |E] \\
&\qquad - \mathbb{E}[\mathcal{T}_k (f(X_k) - f_{\min}) + \|c(X_k)\|_1 |E] \Big). \qquad (44)
\end{aligned}$$

Since $\{\mathcal{T}_k\}$ is a non-decreasing sequence and $f(X_k) \geq f_{\min}$, one finds

$$-\mathbb{E}[\mathcal{T}_k (f(X_k) - f_{\min}) + \|c(X_k)\|_1 |E] \geq -\mathbb{E}[\mathcal{T}_{k-1} (f(X_k) - f_{\min}) + \|c(X_k)\|_1 |E].$$

Thus, from (44), it follows that

$$\begin{aligned}
&\sum_{k=0}^{k_{\max}} (\mathbb{E}[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E] - \mathbb{E}[\phi(X_k, \mathcal{T}_k)|E]) \\
&\quad \geq \mathbb{E}[\mathcal{T}_{k_{\max}} (f(X_{k_{\max}+1}) - f_{\min}) + \|c(X_{k_{\max}+1})\|_1 |E] - \tau_{-1} (f_0 - f_{\min}) - \|c_0\|_1 \\
&\quad \geq -\tau_{-1} (f_0 - f_{\min}) - \|c_0\|_1.
\end{aligned}$$

Combining this with (43) and dividing by $A_{\min} \sum_{k=0}^{k_{\max}} \beta$, one obtains that

$$\begin{aligned}
&\frac{\beta}{\sum_{k=0}^{k_{\max}} \beta} \sum_{k=0}^{k_{\max}} \mathbb{E}[\Delta q(X_k, \hat{T}_k, \nabla f(X_k), H_k, D_k^{\text{true}}) | E] \\
&\leq 2 \left( \frac{\tau_{-1}(f_0 - f_{\min}) + \|c_0\|_1 + (k_{\max}+1)\overline{M}\beta^2 + \kappa_{E_3}\beta(\ell(s_{\max}, \hat{\delta}) + 1)}{A_{\min} \sum_{k=0}^{k_{\max}} \beta} \right) \\
&+ \frac{2(\kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2)}{A_{\min} \sum_{k=0}^{k_{\max}} \beta} \left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau} \right\rceil.
\end{aligned}$$

Hence, by the definitions of $K^*$ and $\beta$, the desired conclusion follows. □

The following corollary translates the result of the preceding theorem to a result pertaining to a stationary measure of (1); recall (2).

**Corollary 1** *Under the assumptions, conditions, and definitions of Theorem 2, it holds with probability at least* $1 - \delta \in (0, 1)$ *that*

$$\begin{aligned}
&\mathbb{E}\left[ \frac{\|\nabla f(X_{K^*}) + J(X_{K^*})^\top Y_{K^*}^{true}\|^2}{\kappa_H^2} + \|c(X_{K^*})\|_1 \middle| E \right] \\
&\leq 2(\kappa_\Psi + 1) \left( \frac{\tau_{-1}(f_0 - f_{\min}) + \|c_0\|_1 + \overline{M}\gamma^2}{\kappa_q \tau_{\min} A_{\min} \gamma \sqrt{k_{\max}+1}} \right) \\
&+ 2(\kappa_\Psi + 1) \left( \frac{\kappa_{E_3}\gamma(\ell(s_{\max}, \hat{\delta}) + 1)}{\kappa_q \tau_{\min} A_{\min} \gamma \sqrt{k_{\max}+1}} \right) \\
&+ (\kappa_\Psi + 1) \left( \frac{2(\kappa_{\Delta q,1}\gamma + \kappa_{\Delta q,2}\gamma^2/\sqrt{k_{\max}+1})}{\kappa_q \tau_{\min} A_{\min} \gamma (k_{\max}+1)} \right) \left\lceil \frac{\ell(s_{\max}, \hat{\delta}) + 1}{p_\tau} \right\rceil.
\end{aligned}$$

*Hence, the complexity bound described in Sect. 2.2 (see (5)) holds.*

***Proof*** The result follows by Lemma 1, Lemma 2, Theorem 2, and (13), which implies that $\|\nabla f(X_{K^*}) + J(X_{K^*})^\top Y_{K^*}^{true}\| = \|H_{k^*} D_{K^*}^{true}\| \leq \kappa_H \|D_{K^*}^{true}\|$. The worst-case complexity bound in Sect. 2.2 follows by combining this result with the definitions of $\kappa_{E_3}$, $\kappa_{\Delta q,1}$, $\kappa_{\Delta q,2}$, and Lemma 21 in Appendix C. □

This result, as well as that of Theorem 2, is proven under the assumption that $s_{\max} \geq 1$. When $s_{\max} = 0$, this result simplifies to a *deterministic* complexity bound with the terms dependent on $s_{\max}$ and $\delta$ omitted. Under the condition $s_{\max} = 0$, the proof follows by noting that $\mathbb{P}_k[E_{k,3}] = \mathbb{P}_k[E_{k,\tau}] = 0$ for all $k \in [k_{\max}]$ (where $E_{k,\tau}$ is defined in the proof of Theorem 2) along with a similar argument to the proof of Theorem 2.

Again, we remark that this result, when viewed in terms of the squared norm of the gradient of the Lagrangian, matches the worst-case complexity of the stochastic gradient method for the unconstrained setting [14].

### 4.5 Complexity result for symmetric sub-Gaussian distributions

In this section, we show that if each stochastic gradient is unbiased with a symmetric, sub-Gaussian distribution and (for simplicity) the ratio parameter sequence remains constant, then the conditions involved in Assumptions 3 and 4 and the event $E$ in (15) occur with high probability. Specifically, this is shown under Assumptions 1, 2, and Assumption 5 below, where it is important to note that Assumption 5 conditions on elements of $\mathcal{G}_k$, not $\mathcal{F}_k$. Overall, our analysis in this section shows that if one makes Assumptions 1, 2, and Assumption 5, then one can be assured that the conditions required for Assumptions 3 and 4 and the event $E$ occur with high probability, meaning that if, in turn, one makes Assumptions 3 and 4 and assumes that $E$ occurs, then our main results of the prior subsection hold.

**Assumption 5** There exists $M \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$,

$$\begin{aligned} \mathbb{E}[G_k|\mathcal{G}_k] &= \nabla f(X_k) \\ \text{and } \mathbb{E}[\exp(\|G_k - \nabla f(X_k)\|^2/M)|\mathcal{G}_k] &\leq \exp(1), \end{aligned} \tag{45}$$

and the random vectors $G_k - \nabla f(X_k)$ and $\nabla f(X_k) - G_k$ have equal distributions. Finally, for all $k \in [k_{\max}]$, the ratio parameter $\Xi_k$ satisfies $\Xi_k = \xi_{\min}$.

Our first lemma shows that, under Assumptions 1, 2, and 5, an inequality of the form in Assumption 4 holds.

**Lemma 10** *Under Assumptions 1, 2, and 5, it follows for all $k \in [k_{\max}]$ that*

$$\mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(X_k)^\top D_k^{true} + \max\{(D_k^{true})^\top H_k D_k^{true}, 0\}|\mathcal{G}_k] \geq \tfrac{1}{2}.$$

***Proof*** Consider arbitrary $k \in [k_{\max}]$. Let $Z_k$ be a basis for the null space of $J(X_k)$, which under Assumption 1 is a matrix in $\mathbb{R}^{n\times(n-m)}$. Then, let $W_k \in \mathbb{R}^{n-m}$ be such that $U_k = Z_k W_k$, and let $W_k^{\text{true}} \in \mathbb{R}^{n-m}$ be such that $U_k^{\text{true}} = Z_k W_k^{\text{true}}$. By (7), $Z_k W_k = -Z_k(Z_k^\top H_k Z_k)^{-1} Z_k^\top (G_k + H_k V_k)$, so that

$$\begin{aligned} &G_k^\top D_k + D_k^\top H_k D_k \\ &\quad = V_k^\top H_k^{1/2}(I - H_k^{1/2} Z_k (Z_k^\top H_k Z_k)^{-1} Z_k^\top H_k^{1/2})(H_k^{-1/2} G_k + H_k^{1/2} V_k) \end{aligned}$$

and similarly

$$\begin{aligned} &\nabla f(X_k)^\top D_k^{\text{true}} + (D_k^{\text{true}})^\top H_k D_k^{\text{true}} \\ &\quad = V_k^\top H_k^{1/2}(I - H_k^{1/2} Z_k (Z_k^\top H_k Z_k)^{-1} Z_k^\top H_k^{1/2})(H_k^{-1/2} \nabla f(X_k) + H_k^{1/2} V_k). \end{aligned}$$

Hence, when conditioned on $\mathcal{G}_k$, the random variables

$$\begin{aligned} &G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} - \nabla f(X_k)^\top D_k^{\text{true}} - \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} \\ &\quad = V_k^\top H_k^{1/2}(I - H_k^{1/2} Z_k (Z_k^\top H_k Z_k)^{-1} Z_k^\top H_k^{1/2})(H_k^{-1/2}(G_k - \nabla f(X_k))) \end{aligned}$$

and

$$\begin{aligned}&\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} - G_k^\top D_k - \max\{D_k^\top H_k D_k, 0\}\\ &= V_k^\top H_k^{1/2}(I - H_k^{1/2} Z_k (Z_k^\top H_k Z_k)^{-1} Z_k^\top H_k^{1/2})(H_k^{-1/2}(\nabla f(X_k) - G_k))\end{aligned}$$

are equivalent in distribution by Assumption 5. Therefore,

$$\begin{aligned}&\mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} - \nabla f(X_k)^\top D_k^{\text{true}}\\ &\quad - \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} \geq 0 | \mathcal{G}_k]\\ &= \mathbb{P}[\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}\\ &\quad - G_k^\top D_k - \max\{D_k^\top H_k D_k, 0\} \geq 0 | \mathcal{G}_k]\end{aligned}$$

and

$$\begin{aligned}1 = &\mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} - \nabla f(X_k)^\top D_k^{\text{true}}\\ &- \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\} \geq 0 | \mathcal{G}_k]\\ &+ \mathbb{P}[\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}\\ &- G_k^\top D_k - \max\{D_k^\top H_k D_k, 0\} \geq 0 | \mathcal{G}_k]\\ &- \mathbb{P}[\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}\\ &- G_k^\top D_k - \max\{D_k^\top H_k D_k, 0\} = 0 | \mathcal{G}_k],\end{aligned}$$

which combined leads to the desired conclusion. □

Next, we state a result based on well-known properties of sub-Gaussian random variables. This lemma follows in the same manner as [21, Lemma 5].

**Lemma 11** *Suppose Assumption 5 holds. Then, for any* $\delta \in (0, 1)$,

$$\mathbb{P}\left[\max_{k \in [k_{\max}]} \|G_k - \nabla f(X_k)\| \leq \sqrt{M\left(1 + \log\left(\frac{k_{\max}+1}{\delta}\right)\right)}\right] \geq 1 - \delta.$$

We conclude by showing that, under Assumptions 1, 2, and 5, the conditions involved in Assumption 3 and $E$ occur with high probability.

**Lemma 12** *Suppose that Assumptions 1, 2, and 5 hold, let* $\kappa_v \in \mathbb{R}_{>0}$ *be such that* $\max\{\|V_k\|_2, \|V_k\|_2^2\} \leq \kappa_v \|c(X_k)\|_2$ *for all* $k \in \mathbb{N}$ *(whose existence follows from Assumption 1 and* [2, Lemma 2.9]*), let* $\kappa_c$ *be an upper bound for* $\|c(X_k)\|_2$ *for all* $k \in \mathbb{N}$ *(whose existence follows from Assumption 1), and define*

$$\kappa_{\tau_{\min}} := \kappa_v\left(\kappa_g + \sqrt{M\left(1 + \log\left(\frac{k_{\max}+1}{\delta}\right)\right)}\right.$$

$$+ \frac{\kappa_H}{\zeta}\left(\sqrt{M\left(1+\log\left(\frac{k_{\max}+1}{\delta}\right)\right)} + \kappa_g + \zeta + \kappa_H\kappa_v\kappa_c\right)\Bigg).$$

*Then, for any $\delta \in (0,1)$, it follows with probability at least $1-\delta$ that the conditions in Assumption 3 and event $E$ hold with*

$$M_\tau = \sqrt{M\left(1+\log\left(\frac{k_{\max}+1}{\delta}\right)\right)}, \quad \tau_{\min} = \frac{(1-\sigma)(1-\epsilon_\tau)}{\kappa_{\tau_{\min}}},$$
$$\text{and } s_{\max} = \min\left\{k_{\max}+1, \left\lceil \frac{\log\left(\frac{\tau_{-1}\kappa_{\tau_{\min}}}{(1-\sigma)(1-\epsilon_\tau)}\right)}{\log\left(\frac{1}{1-\epsilon_\tau}\right)}\right\rceil\right\}.$$

***Proof*** By Lemma 11, the event considered in that lemma holds with probability at least $1-\delta$. Hence, for the purposes of this proof, suppose that event holds. By Jensen's inequality, convexity of $\exp(\cdot)$, and (45), it follows that

$$\mathbb{E}[\|G_k - \nabla f(x_k)\|^2 | \mathcal{G}_k] \leq M.$$

In addition, it follows from the event in Lemma 11 that (19) holds with $M_\tau$ as stated in the lemma. This accounts for Assumption 3. Now consider event $E$. First, it follows from the arguments of [2, Lemma 2.15 and 2.16] combined with the event in Lemma 11 that $\mathcal{T}_k \geq \tau_{\min}$ and $\mathcal{T}_k^{\text{trial,true}} \geq \tau_{\min}$ for all $k \in [k_{\max}]$ for $\tau_{\min}$ as stated in the lemma. Second, it follows from the stated value of $\tau_{\min}$ and (16) that $|\{k \in [k_{\max}] : \mathcal{T}_k < \mathcal{T}_{k-1}\}| \leq s_{\max}$ for $s_{\max}$ as stated in the lemma. Finally, the desired behavior of $\{\Xi_k\}$ follows from Assumption 5. □

### 4.6 Adaptive ratio parameter

In this section, we state a convergence rate result, which can be translated to a worst-case complexity result, that relaxes the definition of the event $E$ considered in prior sections. In particular, we remove the assumption that $\Xi_k = \xi_{\min} \in (0,\infty)$ for all $k \in [k_{\max}]$. Importantly, it has been proved in [2, Lemma 3.5] that, under our remaining assumptions, there still exists $\xi_{\min} \in (0,\infty)$ such that $\Xi_k \geq \xi_{\min}$ for all $k \in [k_{\max}]$. Therefore, by the manner in which the ratio parameter sequence is set, it follows that there exists a maximum number of $k \in [k_{\max}]$ such that $\Xi_k < \Xi_{k-1}$. Denoting this limit as $r_{\max} \in \mathbb{N}$, it follows (for the same reasons as the bound for $s_{\max}$ in (16)) that

$$r_{\max} \leq \min\left\{k_{\max}+1, \left\lceil \frac{\log(\xi_{\min}/\xi_{-1})}{\log(1-\epsilon_\xi)}\right\rceil\right\}.$$

To formalize our new assumption, we define

$$E_\xi := E(k_{\max}, s_{\max}, r_{\max}, \tau_{\min}, \xi_{\min})$$

as the event such that

- $\mathcal{T}_k \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\mathcal{T}_k^{\text{trial,true}} \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\Xi_k \geq \xi_{\min} > 0$ for all $k \in [k_{\max}]$,
- $|\{k \in [k_{\max}] : \mathcal{T}_k < \mathcal{T}_{k-1}\}| \leq s_{\max}$, and
- $|\{k \in [k_{\max}] : \Xi_k < \Xi_{k-1}\}| \leq r_{\max}$.

Since $\{\Xi_k\}$ is bounded below deterministically, this event is identical to the event $E$ defined in (15), except that one may have $\xi_{-1} > \xi_{\min}$.

For the purposes of this section, redefining

$$\mathbb{P}_k[\cdot] := \mathbb{P}[\cdot | E_\xi, \mathcal{G}_k] \text{ and } \mathbb{E}_k[\cdot] := \mathbb{P}[\cdot | E_\xi, \mathcal{G}_k],$$

our analysis of this case considers the following replacement of Assumption 3. Like in Assumption 3, the latter part of the assumption only needs to involve probabilities and expectations conditioned on the event that one or the other parameter decreases, i.e., $\mathcal{T}_k < \mathcal{T}_{k-1}$ and/or $\Xi_k < \Xi_{k-1}$.

**Assumption 6** There exists $M \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$,

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

In addition, there exist $(M_1, M_2, M_3) \in \mathbb{R}^3_{>0}$ such that, for all $k \in [k_{\max}]$,

$$\begin{aligned}
&\text{either } \mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}, \Xi_k = \Xi_{k-1}] = 0 \\
&\quad \text{or } \mathbb{E}_k[\|G_k - \nabla f(X_k)\| | \nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}, \\
&\qquad \Xi_k = \Xi_{k-1}] \leq M_1; \\
&\quad \text{either } \mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k = \mathcal{T}_{k-1}, \Xi_k < \Xi_{k-1}] = 0 \\
&\quad \text{or } \mathbb{E}_k[\|G_k - \nabla f(X_k)\| | \nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k = \mathcal{T}_{k-1}, \\
&\qquad \Xi_k < \Xi_{k-1}] \leq M_2; \\
&\quad \text{and either } \mathbb{P}_k[\nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}, \Xi_k < \Xi_{k-1}] = 0 \\
&\quad \text{or } \mathbb{E}_k[\|G_k - \nabla f(X_k)\| | \nabla f(X_k)^\top (D_k - D_k^{\text{true}})) < 0, \mathcal{T}_k < \mathcal{T}_{k-1}, \\
&\qquad \Xi_k < \Xi_{k-1}] \leq M_3.
\end{aligned}$$

We claim that this assumption holds with high probability under Assumption 5 (without the assumption that $\Xi_k = \xi_{\min}$ for all $k \in [k_{\max}]$), a result that can be derived by modification of the techniques used in Sect. 4.5.

The complexity analysis for this case follows by essentially the same arguments as those used to derive a complexity result under Assumption 3. A slight modification of Lemma 6 is needed to include the three events related to the sign of $\nabla f(x_k)^\top (D_k - D_k^{\text{true}})$ that appear in Assumption 6 (as opposed to the one in Assumption 3), which yields a result in terms of the probabilities of these three events. Then, a slightly modified Lemma 9 and the union bound can be applied two additional times to derive

a complexity result. Since the analysis is a similar, but tedious extension of the results in Sect. 4.4, we simply state the extension of (5) to this case without proof.

**Theorem 3** *Suppose that Assumptions 1, 2, 4, and 6 hold and consider arbitrary* $\delta \in (0, 1)$. *Then, within* $k_{\max} + 1$ *iterations, it holds with probability at least* $1 - \delta$ *that the algorithm generates* $x_{k^*} \in \mathbb{R}^n$ *corresponding to which there exists an associated Lagrange multiplier* $y_{k^*}^{true} \in \mathbb{R}^m$ *such that*

$$\begin{aligned}
&\mathbb{E}[\|\nabla f_{k^*} + J_{k^*}^\top y_{k^*}^{true}\|^2 + \|c_{k^*}\| | E_\xi] \\
&\quad = \mathcal{O}\left( \frac{\tau_{-1}(f_0 - f_{low}) + \|c_0\|_1 + M}{\sqrt{k_{\max} + 1}} \right. \\
&\qquad \left. + \frac{(\tau_{-1}\xi_{-1} - \tau_{\min}\xi_{\min})((s_{\max} + r_{\max})\log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max} + 1}} \right).
\end{aligned}$$

## 5 Conclusion

We proved a worst-case complexity bound (in terms of iterations, function evaluations, and (stochastic) derivative evaluations) for the stochastic sequential quadratic optimization method for solving optimization problems involving a stochastic objective function and deterministic equality constraints proposed in [2]. While key to the practical performance of the algorithm, the adaptivity of the merit parameter introduced a number of theoretical challenges to overcome. Under mostly standard assumptions, we proved that, with high probability, a measure of primal-dual stationarity decays at a rate of $k^{-4}$ (ignoring log factors), which translates into a worst-case complexity bound on par with the stochastic gradient method in the unconstrained setting.

While our analytical approach has been developed for an SQP method that uses an $\ell_1$-norm exact merit function, it may be applicable to a wide variety of algorithmic frameworks for constrained stochastic optimization. For example, our approach may be modified to apply to methods that adaptively update critical parameters at each iteration, such as adaptive penalty methods [5, 6, 22], adaptive augmented Lagrangian methods [11], adaptive barrier methods [27], and penalty-interior point methods [10]. In addition, many constrained optimization algorithms generate (often unconstrained) subproblems defined by an auxiliary parameter sequence that is updated dynamically based off of the solution to the previous subproblem. Algorithms of this type include penalty methods, augmented Lagrangian methods, and interior point methods [28]. In cases when the objective is stochastic, this auxiliary sequence would also be a random process, in which case analyzing the behavior of such a process would be paramount to proving a complexity result for such a method. We believe that the techniques that we have devised for this paper are broadly applicable and foundational for performing complexity analyses of deterministically constrained stochastic optimization methods.

**Funding** Funding was provided by National Science Foundation (Grant Nos. 2030859, CCF-2008484) and Office of Naval Research (Grant No. N00014-21-1-2532).

## A Proof of Theorem 1 (Deterministic Algorithm Complexity)

In this appendix, we prove Theorem 1, which states a worst-case complexity bound for Algorithm 2.1 of [2]. We refer to quantities defined and employed in the analysis in [2]. In particular, in this appendix, for all $k \in \mathbb{N}$, we suppose that $g_k = \nabla f(x_k)$ and $d_k = u_k + v_k$ with $u_k \in \text{Null}(J_k)$ and $v_k \in \text{Range}(J_k^\top)$ is the search direction computed by solving the SQP subproblem with $g_k = \nabla f(x_k)$. As seen in [2], the convergence properties of Algorithm 2.1 in that paper are driven by reductions in a model of the merit function in each iteration. Our first lemma proves a useful lower bound for such a reduction.

**Lemma 13** *Define* $(\kappa_{uv}, \kappa_H, \kappa_v, \tau_{\min}, \zeta, \sigma) \in (0, \infty)^5 \times (0, 1)$ *as in* [2] *and let*

$$\hat{\kappa} := \min\left\{1, \frac{1}{(1+\kappa_{uv})\kappa_v\kappa_H^2}\right\} \quad \text{and} \quad \tilde{\kappa} := \tfrac{1}{4}\zeta\kappa_{uv}\kappa_v\hat{\kappa}. \tag{46}$$

*Then, for any* $\varepsilon \in (0, 1)$, *if* $\|g_k + J_k^\top y_k\| > \varepsilon$ *and/or* $\sqrt{\|c_k\|_1} > \varepsilon$, *then*

$$\Delta q(x_k, \tau_k, g_k, H_k, d_k) \geq \min\left\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\right\}\varepsilon^2. \tag{47}$$

***Proof*** Consider arbitrary $(\varepsilon, k) \in (0, 1) \times \mathbb{N}$ such that $\|g_k + J_k^\top y_k\| > \varepsilon$ and/or $\sqrt{\|c_k\|_1} > \varepsilon$. Let us consider two cases. First, suppose that $\|c_k\|_1 > \hat{\kappa}\varepsilon^2$. Then, by [2, equation (2.9)],

$$\Delta q(x_k, \tau_k, g_k, H_k, d_k) \geq \tfrac{1}{2}\tau_k \max\{d_k^\top H_k d_k, 0\} + \sigma\|c_k\|_1 \geq \sigma\|c_k\|_1 \geq \sigma\hat{\kappa}\varepsilon^2,$$

which implies (47), as desired. Second, suppose that $\|c_k\|_1 \leq \hat{\kappa}\varepsilon^2 \leq \varepsilon^2$, which by the definition of $(\varepsilon, k)$ implies that $\|g_k + J_k^\top y_k\| > \varepsilon$. It follows from this fact that $\|d_k\| > \varepsilon/\kappa_H$; indeed, if $\|d_k\| \leq \varepsilon/\kappa_H$, then by [2, equation (2.6) and Assumption 2.4] one would find

$$\|g_k + J_k^\top y_k\| = \|H_k d_k\| \leq \kappa_H\|d_k\| \leq \varepsilon,$$

which is a contradiction. Hence, $\|d_k\| > \varepsilon/\kappa_H$, and by [2, Lemma 2.9], it follows that $\|v_k\|^2 \leq \kappa_v\|c_k\| \leq \kappa_v\|c_k\|_1 \leq \kappa_v\hat{\kappa}\varepsilon^2$, which combined shows that

$$\varepsilon^2/\kappa_H^2 < \|d_k\|^2 = \|u_k\|^2 + \|v_k\|^2 \leq \|u_k\|^2 + \kappa_v\hat{\kappa}\varepsilon^2.$$

From this fact and the definition of $\hat{\kappa}$, it follows that

$$\begin{aligned}\|u_k\|^2 &> \frac{\varepsilon^2}{\kappa_H^2} - \kappa_v\hat{\kappa}\varepsilon^2 \geq \frac{\varepsilon^2}{\kappa_H^2}\left(1 - \frac{1}{(1+\kappa_{uv})}\right) \\ &= \frac{\kappa_{uv}\varepsilon^2}{(1+\kappa_{uv})\kappa_H^2} \geq \kappa_{uv}\kappa_v\hat{\kappa}\varepsilon^2 \geq \kappa_{uv}\|v_k\|^2,\end{aligned}$$

which along with [2, Lemma 2.10] implies $d_k^\top H_k d_k \geq \frac{1}{2}\zeta \|u_k\|^2 \geq \frac{1}{2}\zeta \kappa_{uv}\kappa_v \hat{\kappa}\varepsilon^2$. Thus,

$$\Delta q(x_k, \tau_k, g_k, H_k, d_k) \geq \tfrac{1}{2}\tau_k \max\{d_k^\top H_k d_k, 0\} + \sigma\|c_k\|_1 \geq \tfrac{1}{4}\tau_{\min}\zeta\kappa_{uv}\kappa_v\hat{\kappa}\varepsilon^2,$$

which implies (47), as desired. □

We now prove Theorem 1, further details of which are provided in the statement below.

**Theorem 4** *Define* $(\tau_{-1}, f_{low}, \alpha_{\min}, \tau_{\min}, \eta, \sigma) \in (0,\infty)^4 \times (0,1)^2$ *as in [2] and* $(\hat{\kappa}, \tilde{\kappa}) \in (0,1] \times (0,\infty)$ *as in (46). Then, for any* $\varepsilon \in (0,1)$*, Theorem 1 holds with (4) given by*

$$\overline{K}_\varepsilon := \left(\frac{\tau_{-1}(f_0 - f_{low}) + \|c_0\|_1}{\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}}\right)\varepsilon^{-2}.$$

***Proof*** To derive a contradiction, suppose (3) does not hold for all $k \in \{0, \ldots, \overline{K}_\varepsilon\}$. Then, along with Lemma 13 and [2, equation (2.10) and Lemma 2.17], it follows for all such $k$ that

$$\begin{aligned}\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) &\leq -\eta\alpha_k \Delta q(x_k, \tau_k, g_k, H_k, d_k)\\ &\leq -\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}\varepsilon^2.\end{aligned}$$

By the definition of $\phi$, this means for all such $k$ that

$$\tau_k f_{k+1} + \|c_{k+1}\|_1 \leq \tau_k f_k + \|c_k\|_1 - \eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}\varepsilon^2.$$

Summing this inequality for all $k \in \{0, \ldots, \overline{K}_\varepsilon\}$, one can deduce that

$$\begin{aligned}&\|c_{\overline{K}_\varepsilon+1}\|_1 - \|c_0\|_1 + \tau_{\overline{K}_\varepsilon} f_{\overline{K}_\varepsilon+1} - \tau_0 f_0 + \sum_{k=1}^{\overline{K}_\varepsilon} f_k(\tau_{k-1} - \tau_k)\\ &\leq -(\overline{K}_\varepsilon + 1)\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}\varepsilon^2.\end{aligned}$$

Since $\{\tau_k\}$ is monotonically nonincreasing, $\|c_{\overline{K}_\varepsilon+1}\|_1 \geq 0$, and $f_k \geq f_{\text{low}}$ for all $k \in \mathbb{N}$,

$$-\|c_0\|_1 + \tau_{\overline{K}_\varepsilon} f_{\text{low}} - \tau_0 f_0 + f_{\text{low}}(\tau_0 - \tau_{\overline{K}_\varepsilon}) \leq -(\overline{K}_\varepsilon + 1)\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}\varepsilon^2.$$

Rearranging this inequality, one arrives at the conclusion that

$$\overline{K}_\varepsilon + 1 \leq \left(\frac{\tau_0(f_0 - f_{\text{low}}) + \|c_0\|_1}{\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}}\right)\varepsilon^{-2} \leq \left(\frac{\tau_{-1}(f_0 - f_{\text{low}}) + \|c_0\|_1}{\eta\alpha_{\min}\min\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\}}\right)\varepsilon^{-2} \equiv \overline{K}_\varepsilon,$$

which is a contradiction. Therefore, one arrives at the desired conclusion that Algorithm 2.1 yields an iterate satisfying (3) in at most $\overline{K}_\varepsilon$ iterations. □

## B Proofs of Lemma 9

In this appendix, we prove Lemma 9. Toward this end, we prove for any $\delta \in (0, 1)$ with $\hat{\delta}$ as defined in (32) and $\ell(s_{\max}, \hat{\delta})$ as defined in (33), one finds

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + 1 \middle| E\right] \geq 1 - \delta. \tag{48}$$

We build to this result, ultimately proved as Theorem 5, with a series of preliminary lemmas.

As our first preliminary result, we state a particular form of Chernoff's bound [31] in the following lemma, which will prove instrumental in deriving (48).

**Lemma 14** *For any $k \in \mathbb{N}$, let $\{Y_0, \ldots, Y_k\}$ be independent Bernoulli random variables. Then, for any $s \in \mathbb{N}$ and $\bar{\delta} \in (0, 1)$, it follows that*

$$\mu := \sum_{j=0}^{k} \mathbb{P}[Y_j = 1] \geq \ell(s, \bar{\delta}) \quad \Longrightarrow \quad \mathbb{P}\left[\sum_{j=0}^{k} Y_j \leq s\right] \leq \bar{\delta}. \tag{49}$$

***Proof*** Suppose that $\mu \geq \ell(s, \bar{\delta})$. By the multiplicative form of Chernoff's bound, it follows for $\rho := 1 - s/\mu$ (which is in the interval $(0, 1)$ by (49)) that

$$\mathbb{P}\left[\sum_{j=0}^{k} Y_j \leq s\right] \leq e^{-\frac{1}{2}\mu\rho^2} = e^{-\frac{1}{2}\mu(1-s/\mu)^2}.$$

Hence, to prove the result, all that remains is to show that $e^{-\frac{1}{2}\mu(1-s/\mu)^2} \leq \bar{\delta}$, i.e., that $-\frac{1}{2}\mu(1 - s/\mu)^2 \leq \log(\bar{\delta})$. Using $\log(\bar{\delta}) = -\log(1/\bar{\delta})$, this inequality is equivalent to

$$0 \leq \tfrac{1}{2}\mu(1 - s/\mu)^2 - \log(1/\bar{\delta}) = \tfrac{1}{2\mu}(\mu - s)^2 - \log(1/\bar{\delta}),$$

which holds if and only if $\mu^2 - 2\mu(s + \log(1/\bar{\delta})) + s^2 \geq 0$. Viewing the left-hand side of this inequality as a convex quadratic function in $\mu$, one finds that the inequality holds as long as $\mu$ is greater than or equal to the positive root of the quadratic, i.e.,

$$s + \log(1/\bar{\delta}) + \sqrt{(s + \log(1/\bar{\delta}))^2 - s^2} = s + \log(1/\bar{\delta}) + \sqrt{\log(1/\bar{\delta})^2 + 2s\log(1/\bar{\delta})}.$$

This holds since $\mu \geq \ell(s, \bar{\delta})$; hence, the result is proved. □

Now, we turn our attention to proving (48). For any realization of a run of the algorithm up to iteration $k \in [k_{\max}]$, let $w_k$ denote the number of times that the merit parameter has been decreased up to the beginning of iteration $k$ and let $\overline{p}_k$ denote the

probability that the merit parameter is decreased during iteration $k$. The *signature* of a realization up to iteration $k \in \mathbb{N}$ is $(\overline{p}_0, \ldots, \overline{p}_k, w_0, \ldots, w_k)$, which encodes all of the pertinent information regarding the behavior of the merit parameter sequence up to the start of iteration $k$.

One could imagine using all possible signatures to define a tree whereby each node contains a subset of all realizations of the algorithm. To construct such tree, one could first consider the root node, which could be denoted by $\tilde{N}(\overline{p}_0, w_0)$, where $\overline{p}_0$ is uniquely defined by the starting conditions of our algorithm and $w_0 = 0$. All realizations of our algorithm follow the same initialization, so $\overline{p}_0$ and $w_0$ would be in the signature of every realization. Now, one could define a node $\tilde{N}(\overline{p}_{[k]}, w_{[k]})$ at depth $k \in [k_{\max}]$ (where the root node has a depth of 0) in the tree as the set of all realizations of our algorithm for which the signature of the realization up to iteration $k$ is $(\overline{p}_0, \ldots, \overline{p}_k, w_0, \ldots, w_k)$. One could then define the edges in the tree by connecting nodes at adjacent levels, where node $\tilde{N}(\overline{p}_{[k]}, w_{[k]})$ is connected to node $\tilde{N}(\bar{p}_{[k]}, \overline{p}_{k+1}, w_{[k]}, w_{k+1})$ for any $\overline{p}_{k+1} \in [0, 1]$ and $w_{k+1} \in \{w_k, w_k + 1\}$.

Unfortunately, the construction described in the previous paragraph may lead to nodes in the tree representing realizations with probability zero occurrence. In order to remedy this, we instead construct a tree where the nodes contain all realizations whose probability signatures fall within specified intervals. To define such intervals, consider arbitrary $B \in \mathbb{N} \setminus \{0\}$ and let us restrict the sequence of values $p_{[k]}$ used to define our nodes as those with

$$p_{[k]} = (p_0, \ldots, p_k) \in \left\{0, \tfrac{1}{B}, \ldots, \tfrac{B-1}{B}\right\}^{k+1}. \tag{50}$$

For $p \in \{0, 1/B, \ldots, (B-1)/B\}$, these define the open probability intervals $\iota(p)$ given by

$$\iota(p) = \begin{cases} \left[p, p + \frac{1}{B}\right) & \text{if } p \in \left\{0, \frac{1}{B}, \ldots, \frac{B-2}{B}\right\}, \\ \left[\frac{B-1}{B}, 1\right] & \text{if } p = \frac{B-1}{B}. \end{cases}$$

Now, we can construct our tree as follows. As before, first consider the root node, which we denote by $N(p_0, w_0)$, where $p_0 \in \{0, 1/B, \ldots, (B-1)/B\}$ is uniquely defined by the starting conditions of our algorithm so that $\mathbb{P}[\mathcal{T}_0 < \tau_{-1} | \mathcal{F}_0] \in \iota(p_0)$ and $w_0 = 0$. All realizations of our algorithm follow the same initialization, so with $\bar{p}_0 = \mathbb{P}[\mathcal{T}_0 < \tau_{-1} | \mathcal{F}_0]$ one finds that $\bar{p}_0 \in \iota(p_0)$ and $w_0$ are in the signature of every realization. We define a node $N(p_{[k]}, w_{[k]})$ at depth $k \in [k_{\max}]$ as the set of all realizations for which the signature of the realization at iteration $k$ exactly matches $w_{[k]}$ and has probabilities that fall within the intervals defined by $p_{[k]}$; i.e., a realization with signature $(\overline{p}_{[k]}, w_{[k]})$ is a member of $N(p_{[k]}, w_{[k]})$ if and only if, for all $j \in [k]$, one finds that $\overline{p}_j \in \iota(p_j)$. The edges in the tree connect nodes in adjacent levels, where $N(p_{[k]}, w_{[k]})$ is connected to $N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1})$ for any $p_{k+1} \in \{0, 1/B, \ldots, (B-1)/B\}$ and $w_{k+1} \in \{w_k, w_k + 1\}$.

Notationally, since the behavior of the algorithm up to iteration $k \in \mathbb{N}$ is determined by the initial conditions and stochastic gradients in $G_{[k-1]}$, we write

$$G_{[k-1]} \in N(p_{[k]}, w_{[k]})$$

to denote the event that the signature of the algorithm up to $k$ is a member of $N(p_{[k]}, w_{[k]})$. The initial condition, denoted for consistency as $G_{[-1]} \in N(p_0, w_0)$, occurs with probability one. Based on the description above, the nodes of our tree satisfy: For any node at a depth of $k \geq 2$, the event $G_{[k-1]} \in N(p_{[k]}, w_{[k]})$ occurs if and only if

$$\begin{aligned} & \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] \in \iota(p_k), \\ S_{k-1} := & \sum_{i=0}^{k-1} \mathcal{I}[\mathcal{T}_i < \mathcal{T}_{i-1}] = w_k, \\ & \text{and } \; G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}), \end{aligned} \tag{51}$$

where $\mathcal{I}[\cdot]$ denotes the indicator function of any given event.

Let us now define certain important sets of nodes in the tree. First, let

$$\mathcal{L}_{\text{good}} := \left\{ N(p_{[k]}, w_{[k]}) : \left( \sum_{i=0}^{k} p_i \leq \ell(s_{\max}, \hat{\delta}) + 1 \right) \wedge (w_k = s_{\max} \vee k = k_{\max}) \right\}$$

be the set of nodes at which the sum of the elements of $p_{[k]}$ is sufficiently small and either $w_k$ has reached $s_{\max}$ or $k$ has reached $k_{\max}$. A node in this set is of interest since, due to the iteration and/or merit parameter decrease limit having been reached, the probability is zero that a certain "bad" event can occur over all realizations with signatures that are members of the node; see Lemma 15 on page 38. Second, let

$$\mathcal{L}_{\text{bad}} := \left\{ N(p_{[k]}, w_{[k]}) : \sum_{i=0}^{k} p_i > \ell(s_{\max}, \hat{\delta}) + 1 \right\}$$

be the nodes in the complement of $\mathcal{L}_{\text{good}}$ at which the sum of the elements of $p_{[k]}$ has exceeded the threshold $\ell(s_{\max}, \hat{\delta}) + 1$. A node in this set is of interest since, due to this threshold having been exceeded, all realizations with signatures that are members of this node correspond to poor behavior of the algorithm (and there is no need to consider the behavior of the algorithm beyond this point). Going forward, we restrict attention to the tree defined by the root node and all paths from the root node that terminate at a node contained in $\mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$. It is clear from this restriction and the definitions of $\mathcal{L}_{\text{good}}$ and $\mathcal{L}_{\text{bad}}$ that this tree is finite with the elements of $\mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$ being leaves.

Let us now define relationships between nodes. The parent of a node is defined as

$$P(N(p_{[k]}, w_{[k]})) = N(p_{[k-1]}, w_{[k-1]}).$$

On the other hand, the children of node $N(p_{[k]}, w_{[k]})$ are defined as

$$C(N(p_{[k]}, w_{[k]})) = \begin{cases} \{N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1})\} & \text{if } N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}} \\ \emptyset & \text{otherwise.} \end{cases}$$

This ensures that paths down the tree terminate at nodes in $\mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$, making these nodes the leaves of the tree. For convenience in the remainder of our discussions, let $C(\emptyset) = \emptyset$.

We define the height of node $N(p_{[k]}, w_{[k]})$ as the length of the longest path from $N(p_{[k]}, w_{[k]})$ to a leaf node, i.e., the height is denoted as

$$h(N(p_{[k]}, w_{[k]})) := \Big(\min\{j \in \mathbb{N} \setminus \{0\} : C^j(N(p_{[k]}, w_{[k]})) = \emptyset\}\Big) - 1,$$

where $C^j(N(p_{[k]}, w_{[k]}))$ is shorthand for applying the mapping $C(\cdot)$ consecutively $j$ times. From this definition, $h(N(p_{[k]}, w_{[k]})) = 0$ for all $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$.

Next, let us define two more sets of nodes that will be useful later. Let $C_{\text{dec}}(N(p_{[k]}, w_{[k]}))$ denote the set of children of $N(p_{[k]}, w_{[k]})$ such that the merit parameter decreases and let $C^c_{\text{dec}}(N(p_{[k]}, w_{[k]}))$ denote set of children of $N(p_{[k]}, w_{[k]})$ such that it does not decrease, so

$$\begin{aligned} C_{\text{dec}}(N(p_{[k]}, w_{[k]})) := \{&N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) : \\ &(N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) \in C(N(p_{[k]}, w_{[k]}))) \\ &\wedge (w_{k+1} = w_k + 1)\} \end{aligned} \tag{52}$$

and

$$\begin{aligned} C^c_{\text{dec}}(N(p_{[k]}, w_{[k]})) := \{&N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) : \\ &(N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) \in C(N(p_{[k]}, w_{[k]}))) \\ &\wedge (w_{k+1} = w_k)\}. \end{aligned} \tag{53}$$

Finally, let us define the event $E_{\text{bad},B}$ as the event that for some $j \in [k_{\max}]$ one finds

$$\left(\sum_{i=0}^{j} \mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] > \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1\right). \tag{54}$$

With respect to our goal of proving (48), the event $E_{\text{bad},B}$ is of interest since it is the event that the given probabilities accumulated up to iteration $j \in [k_{\max}]$ (and beyond) exceed the threshold found in (48) plus a factor that is inversely proportional to $B$.

Let us now prove some properties of the leaf nodes.

**Lemma 15** *For any $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$ with $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{good}$, one finds*

$$\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{bad,B} | E] = 0.$$

*On the other hand, for all $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$ with $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{bad}$, one finds*

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{bad,B} | E] \\ &\quad \leq \hat{\delta} \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right].\end{aligned}$$

***Proof*** Consider an arbitrary index $k \in [k_{\max}]$ and an arbitrary pair $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{good}}$. By the definition of $\mathcal{L}_{\text{good}}$, it follows that

$$\sum_{i=0}^{k} p_i \leq \ell(s_{\max}, \hat{\delta}) + 1. \tag{55}$$

Since the maximum depth of a node is $k_{\max}$ and the gap between the discrete values in (50) is $\frac{1}{B}$, it follows along with (55) that

$$\begin{aligned}&\mathbb{P}\left[\sum_{i=0}^{k} \mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] > \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E, G_{[k-1]} \in N(p_{[k]}, w_{[k]})\right] \\ &\quad \leq \mathbb{P}\left[\sum_{i=0}^{k} \left(p_i + \tfrac{1}{B}\right) > \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E, G_{[k-1]} \in N(p_{[k]}, w_{[k]})\right] \\ &\quad \leq \mathbb{P}\left[\ell(s_{\max}, \hat{\delta}) + \tfrac{k+1}{B} + 1 > \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E, G_{[k-1]} \in N(p_{[k]}, w_{[k]})\right] = 0.\end{aligned}$$

Therefore, for any $j \in \{1, \ldots, k\}$, one finds from conditional probability that

$$\begin{aligned}&\mathbb{P}\left[54 \text{ holds} \wedge G_{[j-1]} \in N(p_{[j]}, w_{[j]}) | E\right] \\ &\quad = \mathbb{P}\left[\sum_{i=0}^{j} \mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] > \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E, G_{[j-1]} \in N(p_{[j]}, w_{[j]})\right] \\ &\qquad \cdot \mathbb{P}\left[G_{[j-1]} \in N(p_{[j]}, w_{[j]}) | E\right] = 0.\end{aligned}$$

In addition, (54) cannot hold for $j = 0$ since $\ell(s_{\max}, \hat{\delta}) + 1 > 1$. Hence, along with the conclusion above, it follows that $E_{\text{bad},B}$ does not occur when a signature up to iteration $j \in \{1, \ldots, k\}$ falls into a node along any path from the root node to $N(p_{[k]}, w_{[k]})$. Now, by the definition of $\mathcal{L}_{\text{good}}$, at least one of $w_k = s_{\max}$ or $k = k_{\max}$ holds. Let us consider each case in turn. If $k = k_{\max}$, then it follows by the preceding arguments that

$$\mathbb{P}\left[\sum_{i=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \leq \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E, G_{[k-1]} \in N(p_{[k]}, w_{[k]})\right] = 1.$$

Otherwise, if $w_k = s_{\max}$, then it follows by the definition of the event $E$ that $\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] = 0$ for all $i \in \{k, \ldots, k_{\max}\}$, and therefore the equation above again follows. Overall, it follows that $\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k-1]}) \wedge E_{\text{bad},B} | E] = 0$, as desired.

Next, we remark that

$$\mathbb{P}[G_{[-1]} \in N(p_0, w_0) \wedge E_{\text{bad},B} | E] = 0$$

since

$$\mathbb{P}[\mathcal{T}_0 < \tau_{-1} | \mathcal{F}_0] \leq p_0 + \frac{1}{B} < \ell(s_{\max}, \hat{\delta}) + \frac{k_{\max} + 1}{B} + 1.$$

Thus, consider arbitrary $k \in \mathbb{N} \setminus \{0\}$ and $(p_{[k]}, w_{[k]})$ with $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{bad}}$. One finds

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad = \mathbb{P}[E_{\text{bad},B} | E, G_{[k-1]} \in N(p_{[k]}, w_{[k]})] \cdot \mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) | E] \\
&\quad \leq \mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) | E].
\end{aligned}$$

Hence, using the initial condition that $G_{[-1]} \in N(p_0, w_0)$, it follows that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad \leq \mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) | E] = \mathbb{P}\big[51 \text{ holds} \big| E\big] \\
&\quad = \mathbb{P}\big[\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] \in \iota(p_k) \big| E, S_{k-1} = w_k, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\big] \\
&\qquad \cdot \mathbb{P}\big[S_{k-1} = w_k \wedge G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}) \big| E\big] \\
&\quad = \mathbb{P}\big[\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] \in \iota(p_k) \big| E, S_{k-1} = w_k, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\big] \\
&\qquad \cdot \mathbb{P}\big[S_{k-1} = w_k \big| E, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\big] \mathbb{P}\big[G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}) \big| E\big] \\
&\quad = \mathbb{P}[G_{-1} \in N(p_0, w_0)] \\
&\qquad \cdot \prod_{i=1}^{k} \Big( \mathbb{P}\big[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \big| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\big] \\
&\qquad \cdot \mathbb{P}\big[S_{i-1} = w_i \big| E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\big] \Big) \\
&\quad = \prod_{i=1}^{k} \Big( \mathbb{P}\big[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \big| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\big] \\
&\qquad \cdot \mathbb{P}\big[S_{i-1} = w_i \big| E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\big] \Big).
\end{aligned} \tag{56}$$

Our goal is to bound (56). Toward this end, define

$$\mathcal{I}_{\text{dec}} := \{i \in \{1, \ldots, k\} : w_i = w_{i-1} + 1\} \text{ and } \mathcal{I}_{\text{dec}}^c := \{i \in \{1, \ldots, k\} : w_i = w_{i-1}\},$$

which by the definition of $w_{[k]}$ form a partition of $\{1, \ldots, k\}$. For any $i \in \mathcal{I}_{\text{dec}}$,

$$\mathbb{P}[S_{i-1} = w_i | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})]$$

$$= \mathbb{P}[\mathcal{T}_{i-1} < \mathcal{T}_{i-2} | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})] \leq p_{i-1} + \tfrac{1}{B}.$$

On the other hand, for any $i \in \mathcal{I}_{\text{dec}}^c$,

$$\begin{aligned}
&\mathbb{P}[S_{i-1} = w_i | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})] \\
&\quad = \mathbb{P}[\mathcal{T}_{i-1} = \mathcal{T}_{i-2} | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})] \\
&\quad = 1 - \mathbb{P}[\mathcal{T}_{i-1} < \mathcal{T}_{i-2} | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})] \leq 1 - p_{i-1}.
\end{aligned}$$

Thus, it follows that the latter term in (56) satisfies

$$\begin{aligned}
&\prod_{i=1}^{k} \mathbb{P}[S_{i-1} = w_i | E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})] \\
&\quad \leq \left( \prod_{i \in \mathcal{I}_{\text{dec}}} (p_{i-1} + \tfrac{1}{B}) \right) \left( \prod_{i \in \mathcal{I}_{\text{dec}}^c} (1 - p_{i-1}) \right).
\end{aligned}$$

Now let us bound this term. By the definition of $\mathcal{L}_{\text{bad}}$, one finds that

$$\sum_{i=0}^{k} p_i > \ell(s_{\max}, \hat{\delta}) + 1 \implies \sum_{i=0}^{k-1} p_i > \ell(s_{\max}, \hat{\delta}). \tag{57}$$

In addition, by the definition of $s_{\max}$, it follows that $w_k \leq s_{\max}$ for all $k \in [k_{\max}]$, from which it follows that $|\mathcal{I}_{\text{dec}}| \leq s_{\max}$. Now, let $\{Z_0, \ldots, Z_{k-1}\}$ be independent Bernoulli random variables such that, for all $i \in \{0, \ldots, k-1\}$, one has

$$\mathbb{P}[Z_i = 1] = \begin{cases} p_i + \frac{1}{B} & \text{if } i+1 \in \mathcal{I}_{\text{dec}} \\ p_i & \text{if } i+1 \in \mathcal{I}_{\text{dec}}^c. \end{cases}$$

By (57), it follows from the definition of these random variables that $\sum_{i=0}^{k-1} \mathbb{P}[Z_i = 1] \geq \ell(s_{\max}, \hat{\delta})$. Then, it follows by Lemma 14 and the preceding argument that

$$\begin{aligned}
&\prod_{i \in \mathcal{I}_{\text{dec}}} \left( p_{i-1} + \frac{1}{B} \right) \prod_{i \in \mathcal{I}_{\text{dec}}^c} (1 - p_{i-1}) \\
&\quad = \mathbb{P}\left[ (Z_{i-1} = 1 \text{ for all } i \in \mathcal{I}_{\text{dec}}) \wedge (Z_{i-1} = 0 \text{for all } i \in \mathcal{I}_{\text{dec}}^c) \right] \\
&\quad = \mathbb{P}\left[ \left( \sum_{i=0}^{k-1} Z_i \leq s_{\max} \right) \wedge (Z_{i-1} = 1 \text{ for all } i \in \mathcal{I}_{\text{dec}}) \wedge (Z_{i-1} = 0 \text{ for all } i \in \mathcal{I}_{\text{dec}}^c) \right] \\
&\quad \leq \mathbb{P}\left[ \sum_{i=0}^{k-1} Z_i \leq s_{\max} \right] \leq \hat{\delta}.
\end{aligned}$$

Combining this with (56), the desired conclusion follows. □

Next, we present a lemma about nodes in the sets defined in (52) and (53). The lemma essentially states that a certain probability of interest, defined as the product of probabilities along a path to a child node, can be reduced to a product of probabilities to the child's parent node by partitioning the childen into those at which a merit parameter decrease has occurred and children at which a merit parameter decrease has not occurred.

**Lemma 16** *For all $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$, one finds that*

$$\sum_{\{(p_{k+1},w_{k+1}):N(p_{[k+1]},w_{[k+1]})\in C_{dec}(N(p_{[k]},w_{[k]}))\}} \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]$$
$$= \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]$$

*and, similarly, one finds that*

$$\sum_{\{(p_{k+1},w_{k+1}):N(p_{[k+1]},w_{[k+1]})\in C^c_{dec}(N(p_{[k]},w_{[k]}))\}} \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]$$
$$= \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right],$$

*where by the definitions of $C_{dec}$ and $C^c_{dec}$ it follows that the value of $w_{k+1}$ in the sum in the former equation is one greater than the value of $w_{k+1}$ in the sum in the latter equation.*

***Proof*** One finds that

$$\sum_{\{(p_{k+1},w_{k+1}):N(p_{[k+1]},w_{[k+1]})\in C_{\text{dec}}(N(p_{[k]},w_{[k]}))\}} \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]$$
$$= \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|\mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]$$
$$\cdot \sum_{\{(p_{k+1},w_{k+1}):N(p_{[k+1]},w_{[k+1]})\in C_{\text{dec}}(N(p_{[k]},w_{[k]}))\}}$$

$$\mathbb{P}\left[\mathbb{P}[\mathcal{T}_{k+1} < \mathcal{T}_k | \mathcal{F}_{k+1}] \in \iota(p_{k+1}) \middle| E, S_k = w_{k+1}, G_{[k-1]} \in N(p_{[k]}, w_{[k]})\right].$$

The desired conclusion follows since, by the definition of $C_{\text{dec}}(N(p_{[k]}, w_{[k]}))$, all elements in the latter sum have $S_k = w_{k+1} = w_k + 1$, meaning that the sum is exhaustive over all possible outcomes with the same conditions, and hence the sum is 1.

The proof of the second desired conclusion follows in the same manner with $C^c_{\text{dec}}$ in place of $C_{\text{dec}}$ and $S_k = w_{k+1} = w_k$ in place of $S_k = w_{k+1} = w_k + 1$. □

Next, we derive a result for certain nodes containing realizations with $w_k = s_{\max} - 1$.

**Lemma 17** *For any $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$ such that $w_k = s_{\max} - 1$ and $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{good}$, it follows that*

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{bad,B} | E] \\ &\leq \hat{\delta} \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right].\end{aligned} \tag{58}$$

***Proof*** By the supposition that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, it follows that any $(p_{[k]}, w_{[k]})$ with $h(N(p_{[k]}, w_{[k]})) = 0$ has $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{bad}}$, in which case the desired conclusion follows from Lemma 15. With this base case being established, we now prove the result by induction. Suppose that the result holds for all $(p_{[k]}, w_{[k]})$ such that $w_k = s_{\max} - 1$, $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, and $h(N(p_{[k]}, w_{[k]})) \leq j$ for some $j \in \mathbb{N}$. Our goal is to show that the same statement holds with $j$ replaced by $j+1$. For this purpose, consider arbitrary $(p_{[k]}, w_{[k]})$ such that $w_k = s_{\max} - 1$, $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, and $h(N(p_{[k]}, w_{[k]})) = j + 1$. Observe that by the definition of the child operators $C$, $C_{\text{dec}}$, and $C^c_{\text{dec}}$, it follows that

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\ &\quad = \sum_{\{(p_{k+1}, w_{k+1}) : N(p_{[k+1]}, w_{[k+1]}) \in C(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E] \\ &\quad = \sum_{\{(p_{k+1}, w_{k+1}) : N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E] \\ &\quad + \sum_{\{(p_{k+1}, w_{k+1}) : N(p_{[k+1]}, w_{[k+1]}) \in C^c_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E].\end{aligned}$$

Since $w_k = s_{\max} - 1$, it follows from the definition of $C_{\text{dec}}$ that for any $(p_{k+1}, w_{k+1})$ with $N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}(N(p_{[k]}, w_{[k]}))$, one finds that $w_{k+1} = w_k + 1 = s_{\max}$. By the definition of $s_{\max}$, this implies that $\mathbb{P}[\mathcal{T}_{k+1} < \mathcal{T}_k | \mathcal{F}_{k+1}] = 0$, so $p_{k+1} = 0$. In addition, since $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{bad}}$ since $C(N(p_{[k]}, w_{[k]})) \neq \emptyset$, it follows that $\sum_{i=0}^{k+1} p_{k+1} \leq \ell(s_{\max}, \hat{\delta}) + 1$, meaning $N(p_{[k+1]}, w_{[k+1]}) \in \mathcal{L}_{\text{good}}$. Consequently, from above and Lemma 15, one finds

$$\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E]$$

$$= \sum_{\{(p_{k+1}, w_{k+1}) : N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}^c(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E].$$

Since $h(N(p_{[k]}, w_{[k]})) = j + 1$, it follows that $h(N(p_{[k+1]}, w_{[k+1]})) \leq j$ for any $(p_{[k+1]}, w_{[k+1]})$ with $h(N(p_{[k+1]}, w_{[k+1]})) \in C_{\text{dec}}^c(N(p_{[k]}, w_{[k]}))$. Therefore, by the induction hypothesis and the result of Lemma 16, it follows that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad = \sum_{\{(p_{k+1}, w_{k+1}) : N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}^c(N(p_{[k]}, w_{[k]}))\}} \\
&\quad \hat{\delta} \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \\
&\quad \leq \hat{\delta} \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right],
\end{aligned}$$

which completes the proof. □

Using the preceding lemma as a base case, we now perform induction on the difference $s_{\max} - w_k$ to prove a similar result for arbitrary $s_{\max}$.

**Lemma 18** *For any $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$ with $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{good}$, it follows that*

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{bad,B} | E] \\
&\quad \leq \hat{\delta} \cdot \sum_{j=0}^{\min\{s_{\max} - w_k - 1, h(N(p_{[k]}, w_{[k]}))\}} \binom{h(N(p_{[k]}, w_{[k]}))}{j} \\
&\qquad \cdot \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right].
\end{aligned}$$

***Proof*** For all $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$ and $h(N(p_{[k]}, w_{[k]})) = 0$, it follows that $N(p_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{bad}}$. The result holds in this case according to Lemma 15 since one finds that $\sum_{j=0}^{\min\{s_{\max} - w_k - 1, h(N(p_{[k]}, w_{[k]}))\}} \binom{h(N(p_{[k]}, w_{[k]}))}{j} = \binom{0}{0} = 1$. On the other hand, for all $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$ and $s_{\max} - w_k = 1$, the result follows from Lemma 17. Hence, to prove the remainder of the result by induction, one may assume that it holds for all $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, $h(N(p_{[k]}, w_{[k]})) \leq t$ for some $t \in \mathbb{N}$, and $s_{\max} - w_k = r$ for some $r \in \mathbb{N} \backslash \{0\}$, and show that it holds for all $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, $h(N(p_{[k]}, w_{[k]})) = t + 1$, and $s_{\max} - w_k = r$. (Notice that the base cases above show that the result holds for $t = 0$ and any $r \in \mathbb{N} \setminus \{0\}$, as well as for any $t \in \mathbb{N}$ and $r = 1$. Hence, one may complete the induction over the index pairs by showing that if it holds for $(t, r)$, then it holds for $(t + 1, r)$, as claimed above.)

Consider arbitrary $(p_{[k]}, w_{[k]})$ such that $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, $h(N(p_{[k]}, w_{[k]})) = t+1$, and $s_{\max} - w_k = r$. By the definitions of $C$, $C_{\text{dec}}$, and $C^c_{\text{dec}}$, it follows that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad = \sum_{\{(p_{[k+1]}, w_{[k+1]}) : N(p_{[k+1]}, w_{[k+1]}) \in C(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E] \\
&\quad = \sum_{\{(p_{[k+1]}, w_{[k+1]}) : N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E] \\
&\qquad + \sum_{\{(p_{[k+1]}, w_{[k+1]}) : N(p_{[k+1]}, w_{[k+1]}) \in C^c_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(p_{[k+1]}, w_{[k+1]}) \wedge E_{\text{bad},B} | E].
\end{aligned}$$

Further by the definition of $C_{\text{dec}}$, it follows that $w_{k+1} = w_k + 1$ (thus $s_{\max} - w_{k+1} = r-1$) for all terms in the former sum on the right-hand side, whereas by the definition of $C^c_{\text{dec}}$ it follows that $w_{k+1} = w_k$ (thus $s_{\max} - w_{k+1} = r$) for all terms in the latter sum on the right-hand side. Moreover, from $h(N(p_{[k]}, w_{[k]})) = t+1$, it follows that $h(N(p_{[k+1]}, w_{[k+1]})) \leq t$ for all terms on the right-hand side. Therefore, by the induction hypothesis, it follows that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad \leq \sum_{\{(p_{[k+1]}, w_{[k+1]}) : N(p_{[k+1]}, w_{[k+1]}) \in C_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \hat{\delta} \sum_{j=0}^{\min\{r-2,t\}} \binom{t}{j} \\
&\qquad \cdot \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \\
&\qquad + \sum_{\{(p_{[k+1]}, w_{[k+1]}) : N(p_{[k+1]}, w_{[k+1]}) \in C^c_{\text{dec}}(N(p_{[k]}, w_{[k]}))\}} \hat{\delta} \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j} \\
&\qquad \cdot \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-2]})\right],
\end{aligned}$$

which by Lemma 16 implies that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\quad \leq \hat{\delta} \left( \sum_{j=0}^{\min\{r-2,t\}} \binom{t}{j} + \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j} \right) \\
&\qquad \cdot \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | \mathcal{F}_i] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right].
\end{aligned} \tag{59}$$

To complete the proof, we need only consider two cases on the relationship between $t$ and $r$. First, if $t \leq r-2$, then Pascal's rule implies that

$$\begin{aligned}\sum_{j=0}^{\min\{r-2,t\}} \binom{t}{j} + \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j} &= 2\sum_{j=0}^{t} \binom{t}{j} \\ &= \binom{t}{t} + \binom{t}{0} + \sum_{j=1}^{t} \left( \binom{t}{j} + \binom{t}{j-1} \right) \\ &= \binom{t+1}{t+1} + \binom{t+1}{0} + \sum_{j=1}^{t} \binom{t+1}{j} \\ &= \sum_{j=0}^{t+1} \binom{t+1}{j} = \sum_{j=0}^{h(N_{p[k],w[k]})} \binom{h(N_{p[k],w[k]})}{j}.\end{aligned}$$

Since $t \leq r-2$, it follows that $h(N_{p[k],w[k]}) = t+1 \leq r-1 = s_{\max} - w_k - 1$, which combined with (59) proves the result in this case. Second, if $t > r-2$, then similarly

$$\begin{aligned}\sum_{j=0}^{\min\{r-2,t\}} \binom{t}{j} + \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j} &= \sum_{j=0}^{r-2} \binom{t}{j} + \sum_{j=0}^{r-1} \binom{t}{j} \\ &= \binom{t}{0} + \sum_{j=1}^{r-1} \left( \binom{t}{j} + \binom{t}{j-1} \right) \\ &= \binom{t+1}{0} + \sum_{j=1}^{r-1} \binom{t+1}{j} \\ &= \sum_{j=0}^{r-1} \binom{t+1}{j} = \sum_{j=0}^{s_{\max}-w_k-1} \binom{h(N_{p[k],w[k]})}{j}.\end{aligned}$$

Since $t > r-2$, $h(N_{p[k],w[k]}) = t+1 > r-1 = s_{\max} - w_{k-1} - 1$, which combined with (59) proves the result for this case as well. □

We now prove our first main result of this section.

**Theorem 5** *For any $\delta \in (0,1)$ with $\hat{\delta}$ as defined in (32) and $\ell(s_{\max}, \hat{\delta})$ as defined in (33), one finds that (48) holds.*

***Proof*** First, consider the case where $s_{\max} = 0$. Then, by the definition of $s_{\max}$,

$$\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] = 0,$$

for all $k = [k_{\max}]$, so the result holds trivially.

Now, let $s_{\max} \in \mathbb{N} \setminus \{0\}$. By construction of our tree and the definitions of $\mathcal{L}_{\text{good}}$ and $\mathcal{L}_{\text{bad}}$, one finds that $h(N(p_0, w_0)) \leq k_{\max}$. In addition, by the definition of $s_{\max}$,

$s_{\max} - 1 < k_{\max}$, so $\min\{s_{\max} - w_0 - 1, h(N(p_0, w_0))\} = s_{\max} - 1 \geq 0$. Consider arbitrary $B \in \mathbb{N} \setminus \{0\}$ (see (54)). By Lemma 18 and (32),

$$\mathbb{P}[E_{\text{bad},B}|E] = \mathbb{P}[G_{[-1]} \in N(p_0, w_0) \wedge E_{\text{bad},B}|E] \leq \hat{\delta} \sum_{j=0}^{\min\{s_{\max}-1, k_{\max}\}} \binom{k_{\max}}{j} = \delta.$$

Therefore, by the definition of $E_{\text{bad},B}$ (see (54)), it follows that

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1 \middle| E\right] \geq 1 - \delta.$$

Now, let us define the event $E_{\text{good},B}$ for $B \in \mathbb{N}\setminus\{0\}$ as the event that

$$\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + \tfrac{k_{\max}+1}{B} + 1.$$

One sees that $E_{\text{good},B} \supseteq E_{\text{good},B+1}$ for all such $B$. Therefore, by the properties of a decreasing sequence of events (see, for example [31, Section 1.5]), it follows that

$$\begin{aligned}
&\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + 1 \middle| E\right] \\
&= \mathbb{P}\left[\lim_{B\to\infty} \left(\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + \frac{k_{\max}+1}{B} + 1\right) \middle| E\right] \\
&= \lim_{B\to\infty} \mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k] \leq \ell(s_{\max}, \hat{\delta}) + \frac{k_{\max}+1}{B} + 1 \middle| E\right] \geq 1 - \delta,
\end{aligned}$$

as desired. □

Next, we present some preliminary results that are required to prove the second statement of Lemma 9. Recall the random index set $\mathcal{K}_\tau$ defined in (34). Our next lemma shows a property about any iteration $k \in [k_{\max}]$ in which $k \in \mathcal{K}_\tau$.

**Lemma 19** *Let Assumption 4 hold. Then, one finds that*

$$\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1}|\mathcal{F}_k, k \in \mathcal{K}_\tau] \geq p_\tau.$$

***Proof*** In any iteration where $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$, it follows that $\mathcal{T}_k^{\text{trial,true}} < \infty$, so

$$\mathcal{T}_k^{\text{trial,true}} = \frac{(1-\sigma)\|c(X_k)\|_1}{\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}}$$

and thus

$$(1-\sigma)\|c(X_k)\|_1 < (\nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\})\mathcal{T}_{k-1}.$$

By the definition of $\mathcal{T}_k$, if

$$G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(X_k)^\top D_k^{\text{true}} + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}$$

in an iteration such that $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$, then

$$(1-\sigma)\|c(X_k)\|_1 < (G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\})\mathcal{T}_{k-1},$$

meaning that $\mathcal{T}_k < \mathcal{T}_{k-1}$. Observe that the event $k \in \mathcal{K}_\tau$ only depends on the history of the algorithm prior to iteration $k$ and thus the $\sigma$-algebra generated by $k \in \mathcal{K}_\tau$ is included in $\mathcal{F}_k$. Therefore, by [15, Theorem 4.1.13], for any random variable $Z$, we have

$$\mathbb{E}_k[Z|k \in \mathcal{K}_\tau] = \mathbb{E}_k[\mathbb{E}_k[Z]|k \in \mathcal{K}_\tau].$$

Therefore, with $\mathbf{1}(\tilde{E})$ denoting the indicator of event $\tilde{E}$, it follows from Assumption 4 that

$$\begin{aligned}
&\mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1}|k \in \mathcal{K}_\tau] \\
&\quad \geq \mathbb{E}_k[\mathbf{1}(G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(X_k)^\top D_k^{\text{true}} \\
&\qquad + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\})|k \in \mathcal{K}_\tau] \\
&\quad = \mathbb{E}_k[\mathbb{E}_k[\mathbf{1}(G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(X_k)^\top D_k^{\text{true}} \\
&\qquad + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\})]|k \in \mathcal{K}_\tau] \\
&\quad = \mathbb{E}_k[\mathbb{P}_k[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(X_k)^\top D_k^{\text{true}} \\
&\qquad + \max\{(D_k^{\text{true}})^\top H_k D_k^{\text{true}}, 0\}]|k \in \mathcal{K}_\tau] \\
&\quad \geq \mathbb{E}_k[p_\tau|k \in \mathcal{K}_\tau] = p_\tau,
\end{aligned}$$

as desired. □

The previous lemma guarantees that in any iteration in which $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$, the probability is at least $p_\tau$ that the merit parameter decreases. By the scheme for setting $\mathcal{T}_k$,

$$\mathbb{P}_k[\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_k|\mathcal{T}_k^{\text{trial,true}} \geq \mathcal{T}_{k-1}] = 0, \tag{60}$$

so one must have $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$ in any iteration when $\hat{\mathcal{T}}_k < \mathcal{T}_k$. Thus, we can obtain a bound on the number of iterations at which $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_k$ by bounding the number of iterations at which $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$. Now we prove a result relating $|\mathcal{K}_\tau|$ to the probabilities of decreasing the merit parameter over all iterations.

**Lemma 20** *One finds that*

$$\sum_{k=0}^{k_{\max}} \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1}] \geq |\mathcal{K}_\tau| p_\tau.$$

***Proof*** By the law of total probability,

$$\begin{aligned}
&\sum_{k=0}^{k_{\max}} \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1}] \\
&\quad = \sum_{k=0}^{k_{\max}} \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | k \in \mathcal{K}_\tau] \mathbb{P}_k[k \in \mathcal{K}_\tau] + \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | k \in \mathcal{K}_\tau^c] \mathbb{P}_k[k \in \mathcal{K}_\tau^c] \\
&\quad \geq \sum_{k=0}^{k_{\max}} \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | k \in \mathcal{K}_\tau] \mathbb{P}_k[k \in \mathcal{K}_\tau].
\end{aligned}$$

Now, similar to the previous lemma, we note that $k \in \mathcal{K}_\tau$ only depends on the history of the algorithm prior to iteration $k$. Denoting the indicator of the event $k \in \mathcal{K}_\tau$ as $\mathbf{1}(k \in \mathcal{K}_\tau)$, it follows by Lemma 19 that

$$\begin{aligned}
\sum_{k=0}^{k_{\max}} \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | k \in \mathcal{K}_\tau] \mathbb{P}_k[k \in \mathcal{K}_\tau] &\geq p_\tau \sum_{k=0}^{k_{\max}} \mathbb{E}_k[\mathbf{1}(k \in \mathcal{K}_\tau)] \\
&= p_\tau \sum_{k=0}^{k_{\max}} \mathbf{1}(k \in \mathcal{K}_\tau) \\
&= p_\tau \, |\mathcal{K}_\tau| \,,
\end{aligned}$$

where the second to last equality follows by $\mathbf{1}(k \in \mathcal{K}_\tau) \in \mathcal{F}_k$. □

Now, we are prepared to prove Lemma 9.

***Proof*** (Lemma 9) For the first statement, observe that, for any $k \in [k_{\max}]$, by the defintion of $E_{k,3}$, the event $\mathcal{T}_k < \mathcal{T}_{k-1}$ must occur whenever $E_{k,3}$ occurs. Therefore, for any $k \in [k_{\max}]$, one finds

$$\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] \geq \mathbb{P}[E_{k,3} | \mathcal{F}_k].$$

Equation (35) then follows directly from Theorem 5.

Now, consider the second statement of Lemma 9. The proof follows by combining Theorem 5 and Lemma 20 with the preceeding argument. □

## C Lemma Required for the Proof of Corollary 1

This appendix provides the following lemma, which shows that the order notation result in (5a) and (5b) holds, as required in the proof of Corollary 1.

**Lemma 21** *Let $\delta \in (0, 1)$, $\hat{\delta}$ be defined in (32), $s_{\max} \in \mathbb{N} \setminus \{0\}$ and $\ell(s_{\max}, \hat{\delta})$ be defined in (33). Then,*

$$\ell(s_{\max}, \hat{\delta}) = \mathcal{O}\left(s_{\max} \log(k_{\max}) + \log(1/\delta)\right).$$

***Proof*** Since $s_{\max} \in \mathbb{N} \setminus \{0\}$, it follows that

$$\begin{aligned}\sum_{j=0}^{\max\{s_{\max}-1,0\}} \binom{k_{\max}}{j} &= \sum_{j=0}^{s_{\max}-1} \frac{(k_{\max})!}{j!(k_{\max}-j)!} \leq \sum_{j=0}^{s_{\max}-1} \frac{(k_{\max})!}{(k_{\max}-j)!} \\ &= 1 + \sum_{j=1}^{s_{\max}-1} \prod_{i=k_{\max}+1-j}^{k_{\max}} i \leq 1 + \sum_{j=1}^{s_{\max}-1} (k_{\max})^j \\ &\leq s_{\max}(k_{\max})^{s_{\max}-1}.\end{aligned}$$

Then, by the definitions of $\ell(s_{\max}, \hat{\delta})$ and $\hat{\delta}$, it follows that

$$\begin{aligned}\ell(s_{\max}, \hat{\delta}) &= \mathcal{O}\left(s_{\max} + \log(1/\hat{\delta})\right) \\ &= \mathcal{O}\left(s_{\max} + \log(s_{\max}) + (s_{\max}-1)\log(k_{\max}) + \log(1/\delta)\right) \\ &= \mathcal{O}\left(s_{\max} \log(k_{\max}) + \log(1/\delta)\right),\end{aligned}$$

as desired. □

## References


1. Berahas, A.S., Curtis, F.E., O'Neill, M.J., Robinson, D.P.: A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear Equality Constrained Optimization with Rank-Deficient Jacobians. arXiv:2106.13015 (2021)
2. Berahas, A.S., Curtis, F.E., Robinson, D.P., Zhou, B.: Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. SIAM J. Optim. **31**, 1352–1379 (2021)
3. Bertsekas, D.P.: Network optimization: continuous and discrete models, Athena Scientific Belmont (1998)
4. Betts, J.T.: Practical methods for optimal control and estimation using nonlinear programming, SIAM (2010)
5. Byrd, R.H., Lopez-Calva, G., Nocedal, J.: A line search exact penalty method using steering rules. Math. Program. **133**, 39–73 (2012)
6. Byrd, R.H., Nocedal, J., Waltz, R.A.: Steering exact penalty methods for nonlinear programming. Optim. Methods Softw. **23**, 197–213 (2008)
7. Cartis, C., Gould, N.I.M., Toint, P.L.: An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. IMA J. Numer. Anal. **32**, 1662–1695 (2012). https://doi.org/10.1093/imanum/drr035
8. Cartis, C., Gould, N.I.M., Toint, P.L.: On the complexity of finding first-order critical points in constrained nonlinear optimization. Math. Program. **144**, 93–106 (2014)

9. Chen, C., Tung, F., Vedula, N., Mori, G.: Constraint-aware deep neural network compression. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 400–415 (2018)
10. Curtis, F.E.: A penalty-interior-point algorithm for nonlinear constrained optimization. Math. Program. Comput. **4**, 181–209 (2012)
11. Curtis, F.E., Jiang, H., Robinson, D.P.: An adaptive augmented lagrangian method for large-scale constrained optimization. Math. Program. **152**, 201–245 (2015)
12. Curtis, F.E., Robinson, D.P.: Exploiting negative curvature in deterministic and stochastic optimization. Math. Program. **176**, 69–94 (2019)
13. Curtis, F.E., Robinson, D.P., Samadi, M.: Complexity analysis of a trust funnel algorithm for equality constrained optimization. SIAM J. Optim. **28**, 1533–1563 (2018)
14. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**, 207–239 (2019)
15. Durrett, R.: Probability: Theory and Examples, vol. 49, Cambridge University Press (2019)
16. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. **155**, 267–305 (2016)
17. Grapiglia, G.N., Yuan, Y.-X.: On the complexity of an augmented Lagrangian method for nonconvex optimization. IMA J. Numer. Anal. **41**, 1546–1568 (2020)
18. Han, S.P., Mangasarian, O.L.: Exact penalty functions in nonlinear programming. Math. Program. **17**, 251–269 (1979). https://doi.org/10.1007/BF01588250
19. Hazan, E., Luo, H.: Variance-reduced and projection-free stochastic optimization. In International Conference on Machine Learning, PMLR, pp. 1263–1271 (2016)
20. Kupfer, F., Sachs, E.W.: Numerical solution of a nonlinear parabolic control problem by a reduced sqp method. Comput. Optim. Appl. **1**, 113–135 (1992)
21. Li, X., Orabona, F.: A high probability analysis of adaptive SGD with momentum. arXiv: 2007.14294 (2020)
22. Mongeau, M., Sartenaer, A.: Automatic decrease of the penalty parameter in exact penalty function methods. Eur. J. Oper. Res. **83**, 686–699 (1995)
23. Na, S., Anitescu, M., Kolar, M.: An Adaptive Stochastic Sequential Quadratic Programming with Differentiable Exact Augmented Lagrangians. arXiv preprint arXiv:2102.05320 (2021)
24. Na, S., Mahoney, M.W.: Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. arXiv:2205.13687 (2022)
25. Nandwani, Y., Pathak, A., Mausam, P.: A primal dual formulation for deep learning with constraints, in NeurIPS, Singla (2019)
26. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**, 1574–1609 (2009)
27. Nocedal, J., Wächter, A., Waltz, R.A.: Adaptive barrier update strategies for nonlinear interior methods. SIAM J. Optim. **19**, 1674–1693 (2009)
28. Nocedal, J., Wright, S.J.: Numerical optimization. Springer Science & Business Media (2006)
29. Rees, T., Dollar, H.S., Wathen, A.J.: Optimal solvers for pde-constrained optimization. SIAM J. Sci. Comput. **32**, 271–298 (2010)
30. Roy, S. K., Mhammedi, Z., Harandi, M.: Geometry aware constrained optimization techniques for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4460–4469 (2018)
31. Stirzaker, D.: Elementary probability. Cambridge University Press (2003)
32. Wilson, R.B.: A Simplicial Algorithm for Concave Programming, Ph.D. Dissertation, Graduate School of Business Administration (1963)