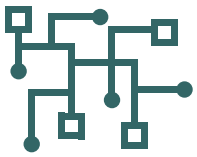

ISE

Industrial and
Systems Engineering

Worst-Case Complexity of an SQP Method for Nonlinear Equality Constrained Stochastic Optimization

Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson

Department of Industrial and Systems Engineering, Lehigh University, USA

COR@L Technical Report 22T-001

LEHIGH
UNIVERSITY

COR@L
COMPUTATIONAL OPTIMIZATION
RESEARCH AT LEHIGH

Worst-Case Complexity of an SQP Method for Nonlinear Equality Constrained Stochastic Optimization

Frank E. Curtis*[1], Michael J. O'Neill†[1], and Daniel P. Robinson‡[1]

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

January 3, 2022

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.

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 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, 15, 22]. 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 [7, 19, 23] (which includes the class of augmented Lagrangian methods) and sequential quadratic optimization (commonly known as SQP) methods

*E-mail: frank.e.curtis@lehigh.edu, supported by NSF Grant CCF-2008484 and ONR Grant N00014-21-1-2532

†E-mail: moneill@lehigh.edu, supported by the CI Fellows Program

‡E-mail: daniel.p.robinson@lehigh.edu, supported by supported by ONR Grant N00014-21-1-2532

[1, 18]. 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.

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 [25].

In this paper, we present a worst-case complexity analysis of the SQP method proposed in [1], 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 one knows a threshold for the merit parameter beyond which the merit function is *exact* [13], 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 arguably the best result that one can expect given this is the same bound proved to hold for a stochastic gradient method employed to solve an unconstrained nonconvex problem [11]. 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 purely due to the uncertainty of the behavior of the adaptive merit parameter sequence, and does not reflect any uncertainty of the behavior of the method during situations in which the merit parameter sequence remains constant.

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 [11, 12] and Frank-Wolfe type methods [14]. (One exception is a complexity bound proved for the SQP algorithm proposed in [18], 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.) 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 $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$, ω is a random variable with associated probability space $(\Omega, \mathcal{F}, P)$, $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, and $\mathbb{E}$ represents expectation with respect to P. In particular, following [1], we make the following assumption about problem (1) and the algorithm that we analyze (stated as Algorithm 1 on page 9), which in any run generates a sequence of iterates $\{x_k\} \subset \mathbb{R}^n$.

Assumption 1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be an open convex set that contains $\{x_k\}$ for all realizations of Algorithm 1. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below by $f_{\text{low}} \in \mathbb{R}$ over $\mathcal{X}$ and the corresponding gradient function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipschitz continuous with constant $L \in (0,\infty)$ over $\mathcal{X}$. 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 over $\mathcal{X}$, each gradient function $\nabla c_i : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant γ_i over $\mathcal{X}$ for all $i \in \{1,\dots,m\}$, and the singular values of $J \equiv \nabla c^\top$ are bounded below and away from zero over $\mathcal{X}$.*

A consequence of Assumption 1 is that there exists $\kappa_g \in \mathbb{R}_{>0}$ such that

$$\|\nabla f(x_k)\| \leq \kappa_g \quad \text{for any } k \in \mathbb{N} \text{ in any realization of Algorithm 1.} \tag{2}$$

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

$$0 = \begin{bmatrix} \nabla_x \ell(x,y) \\ \nabla_y \ell(x,y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + \nabla c(x) y \\ c(x) \end{bmatrix}. \tag{3}$$

1.2 Notation

We adopt the notation that $\|\cdot\|$ denotes the ℓ_2-norm for vectors and the vector-induced ℓ_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 $\mathbb{N} := \{0, 1, 2, \dots, \}$. For any integer $k \in \mathbb{N}$, we use $[k]$ to denote the subset of nonnegative integers up to k, namely, $[k] := \{0, \dots, k\}$. Correspondingly, to represent a set of vectors $\{v_0, \dots, v_k\}$, we define $v_{[k]} := \{v_0, \dots, v_k\}$.

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 $|\phi(\cdot)| \leq c\varphi(\cdot)|\log^{\bar{c}}(\cdot)|$ for some $c \in (0,\infty)$ and $\bar{c} \in (0,\infty)$. In this manner, one finds that $\mathcal{O}(\varphi(\cdot)|\log^{\bar{c}}(\cdot)|) \equiv \widetilde{\mathcal{O}}(\varphi(\cdot))$ for any $\bar{c} \in (0,\infty)$.

Algorithm 1 is iterative, generating in each realization a sequence $\{x_k\}$. (See Section 4.1 for a complete description of the stochastic process generated by the algorithm.) For our analysis, we also append the iteration number to other quantities corresponding to each iteration, e.g., $f_k := f(x_k)$ for all $k \in \mathbb{N}$.

When discussing stochastic quantities, we use capital letters to denote random variables and corresponding lower case letters to denote a realization of a random variable. For example, a stochastic gradient in iteration $k \in \mathbb{N}$ is denoted as G_k, a realization of which is written as g_k.

1.3 Outline

Section 2 provides a worst-case complexity result for the algorithm from [1] 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 Section 3, followed by our main result and analysis, which are provided in Section 4. Finally, we provide concluding thoughts and mention future directions in Section 5.

2 Outline of Main Results

Our algorithm of consideration, namely, Algorithm 3.1 in [1], is derived from Algorithm 2.1 in [1], which is proposed and analyzed for the deterministic setting as a precursor for the stochastic setting (of Algorithm 3.1 in [1]). In the deterministic algorithm, the kth search direction $d_k \in \mathbb{R}^n$ is computed by solving an optimization 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

$$\phi(x, \tau) = \tau f(x) + \|c(x)\|_1. \tag{4}$$

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.

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 6.

Theorem 1. *Consider Algorithm 2.1 in [1] and suppose that Assumption 1 holds along with Assumption 2.4 from [1]. 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 [1]). 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*

$$\|g_k + J_k^\top y_k\| \leq \varepsilon \quad \text{and} \quad \sqrt{\|c_k\|_1} \leq \varepsilon \tag{5}$$

in a number of iterations no more than

$$\left(\frac{\tau_{-1}(f_0 - f_{\text{low}}) + \|c_0\|_1}{\min\{\kappa_1, \tau_{\min}\kappa_2\}} \right) \varepsilon^{-2}. \tag{6}$$

Theorem 1 is not surprising. After all, such a complexity bound of $\mathcal{O}(\varepsilon^{-2})$ is well-known for gradient-based algorithms in unconstrained nonconvex optimization. Since Algorithm 2.1 in [1] 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.

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 14 and the beginning of the proof of Theorem 4) that the improvement in the merit function from any iterate that is not ε-stationary 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 [1], 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 had to devise 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 25. 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 7. 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 12. Furthermore, our analysis conditions on the occurrence of an event that we call E (see (17)); 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

$$\begin{aligned} &\mathbb{E}[\|\nabla f_{k^*} + J_{k^*}^\top y_{k^*}^{\text{true}}\|^2 + \|c_{k^*}\|_1 | E] \\ &= \mathcal{O}\left(\frac{\tau_{-1}(f_0 - f_{\text{low}}) + \|c_0\|_1 + M}{\sqrt{k_{\max} + 1}} \right. && \text{(7a)} \\ &\qquad \left. + \frac{(\tau_{-1} - \tau_{\min})(s_{\max} \log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max} + 1}} \right). && \text{(7b)} \end{aligned}$$

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 al-

gorithm 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 first three quantities on the right-hand side of the convergence rate, namely, in (7a), 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. The final term in (7b), 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 (7a) (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 $s_{\max}$ is bounded by a value that is independent from $k_{\max} + 1$; this follows from a deterministic lower bound on $\tau_{\min}$ [1, 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 Section 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 (7b).

3 Algorithm

For ease of reference, in this section we present Algorithm 3.1 from [1], 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

$$\min_{d \in \mathbb{R}^n} \; f_k + g_k^\top d + \tfrac{1}{2} d^\top H_k d \quad \text{s.t.} \quad c_k + J_k d = 0, \tag{8}$$

where $H_k \in \mathbb{S}^n$ is chosen independently from g_k, and we remind the reader that g_k is a realization of the stochastic gradient G_k. Under Assumption 1 and the following Assumption 2 (that we make throughout the remainder of the paper), the solution of (8) can be obtained from the unique solution of

$$\begin{bmatrix} H_k & J_k^\top \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ y_k \end{bmatrix} = - \begin{bmatrix} g_k \\ c_k \end{bmatrix}. \tag{9}$$

Assumption 2. *The sequence $\{\|H_k\|\}$ is bounded by $\kappa_H \in \mathbb{R}_{>0}$. In addition, there exists $\zeta \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$, 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 kth 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

$$\begin{aligned} \Delta q(x, \tau, g, H, d) &:= q(x, \tau, g, H, 0) - q(x, \tau, g, H, d) \\ &= -\tau(g^\top d + \tfrac{1}{2} \max\{d^\top H d, 0\}) + \|c(x)\|_1. \end{aligned} \tag{10}$$

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

$$\Delta q(x_k, \tau, g_k, H_k, d_k) \geq \tfrac{1}{2}\tau \max\{d_k^\top H_k d_k, 0\} + \sigma \|c_k\|_1 \geq 0 \tag{11}$$

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

$$\tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^\top d_k + \max\{d_k^\top H_k d_k, 0\} \leq 0 \\ \frac{(1-\sigma)\|c_k\|_1}{g_k^\top d_k + \max\{d_k^\top H_k d_k, 0\}} & \text{otherwise,} \end{cases} \tag{12}$$

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

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \leq \tau_k^{\text{trial}} \\ (1 - \epsilon_\tau)\tau_k^{\text{trial}} & \text{otherwise.} \end{cases} \tag{13}$$

Then, for use in the stepsize computation (as motivated in [1]) it sets

$$\xi_k^{\text{trial}} \leftarrow \frac{\Delta q(x_k, \tau_k, g_k, H_k, d_k)}{\tau_k \|d_k\|^2} \quad \text{then} \quad \xi_k \leftarrow \begin{cases} \xi_{k-1} & \text{if } \xi_{k-1} \leq \xi_k^{\text{trial}} \\ (1 - \epsilon_\xi)\xi_k^{\text{trial}} & \text{otherwise} \end{cases} \tag{14}$$

for some $\epsilon_\xi \in (0, 1)$, which, for one thing, ensures $\xi_k \leq \xi_k^{\text{trial}}$. The last component in the kth iteration is to set the stepsize, the magnitude of which is controlled by prescribed $\{\beta_k\} \subset (0, 1]$, 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 $\text{Proj}(\cdot \mid \mathcal{I})$ represents the projection operator onto the interval $\mathcal{I} \subset \mathbb{R}$. 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. In our analysis, we establish our result for the case of $\beta_k = \mathcal{O}(1/\sqrt{k_{\max} + 1})$ 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 [1] are the fixed iteration limit ($k_{\max}$), a slightly stronger decrease requirement for the definition of τ_k in (13) when $\tau_{k-1} > \tau_k^{\text{trial}}$, 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., [11]. 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 [10].

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 ∇f; $\Gamma \in [\sum_{i=1}^m \gamma_i, \infty)$, where $\gamma_i \in (0,\infty)$ is a Lipschitz constant for ∇c_i for all $i \in [m]$

for all $k \in [k_{\max}]$ **do**
  Compute (d_k, y_k) as the solution of (9)
  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 τ_k^{trial} by (12) and τ_k by (13)
    Set ξ_k^{trial} and ξ_k by (14)
    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 (Section 4.1), then formalizing the assumptions that we make about the stochastic gradient estimates (Section 4.2). We then state, in some cases in a slightly modified form, some key lemmas from [1] that are needed for our analysis (Section 4.3). Our generic complexity result, which has been outlined in Section 2, is then stated and proved (Section 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 (Section 4.5). Finally, we conclude this section by outlining a form of our generic complexity result that relaxes one of our minor simplifying assumptions (Section 4.6).

Similarly as for the convergence analysis in [1], 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 kth 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

$$\begin{bmatrix} H_k & J_k^\top \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k^{\text{true}} \\ y_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c_k \end{bmatrix}, \tag{15}$$

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

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

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

$$\Delta q(x_k, \tau, \nabla f(x_k), H_k, d_k^{\text{true}}) \geq \tfrac{1}{2}\tau \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} + \sigma\|c_k\|_1. \tag{16}$$

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 Ξ_k, and stepsize $\mathcal{A}_k$. For all $k \in \mathbb{N}$, a realization of the corresponding element of this process is 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^{true}, the true tangential search direction component U_k^{true}, the true Lagrange multiplier estimate Y_k^{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 is 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 kth 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 $\mathcal{F}_k$ denote the σ-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 $\mathcal{F}_0 = \sigma(x_0)$. As a result, $\{\mathcal{F}_k\}_{k\geq 0}$ is a filtration. When conditioning on a specific realization of Algorithm 1 up to the beginning of iteration $k \in \mathbb{N}$, we condition on $G_{[k-1]} = g_{[k-1]}$, since these stochastic gradients determine x_k. (Recall our notation that $g_{[k-1]}$ represents $\{g_0, \ldots, g_{k-1}\}$.) Similarly, later in our analysis when we condition on $\mathcal{F}_k$, we are conditioning on all realizations of $G_{[k-1]}$ that are measurable with respect to the filtration $\mathcal{F}_k$.

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

$$E := E(k_{\max}, s_{\max}, \tau_{\min}, \xi_{\min}), \tag{17}$$

which we define as the event such that, in every realization of the algorithm, the merit parameters $\{\tau_k\}_{k=0}^{k_{\max}}$ and ratio parameters $\{\xi_k\}_{k=0}^{k_{\max}}$ satisfy

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

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 is constant over all iterations is not actually essential for our analysis; rather, it is made for the sake of simplicity. Indeed, in Section 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 [1, 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. That said, attention to this behavior of the algorithm is justified by arguments made in [2, 1], which under the same kinds of assumptions made in this paper argue that, in any run of the algorithm, the probability is zero that the merit parameter vanishes. Furthermore, in Section 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

$$s_{\max} \leq \min\left\{k_{\max}+1, \left\lceil \frac{\log(\tau_{\min}/\tau_{-1})}{\log(1-\epsilon_\tau)} \right\rceil\right\}. \tag{18}$$

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

Conditioned on E, we assume the following about the stochastic gradient estimates. Such an assumption, namely, that conditioned on the event that a given iterate has been reached the stochastic gradient is unbiased and has bounded variance, is common in analyses of stochastic optimization methods. Here and throughout the remainder of the paper, we let $\mathbb{P}_k[\cdot]$ (respectively, $\mathbb{E}_k[\cdot]$) denote probabilty (respectively, expectation) conditioned on event E and that $G_{[k-1]} = g_{[k-1]}$ for a given $g_{[k-1]}$, i.e., we define

$$\mathbb{P}_k[\cdot] := \mathbb{P}[\cdot | E, G_{[k-1]} = g_{[k-1]}] \text{ and } \mathbb{E}_k[\cdot] := \mathbb{E}[\cdot | E, G_{[k-1]} = g_{[k-1]}].$$

Assumption 3. *There exists $M \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$, one finds that*

$$\mathbb{E}_k[G_k] = \nabla f(x_k) \quad \textit{and} \quad \mathbb{E}_k[\|G_k - \nabla f(x_k)\|_2^2] \leq M. \tag{19}$$

In addition, there exists $M_\tau \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$, one finds that

$$\begin{aligned} &\textit{either} \quad \mathbb{P}_k[\nabla f(x_k)^\top (D_k - d_k^{\text{true}}) < 0, \mathcal{T}_k < \tau_{k-1}] = 0 \\ &\textit{or} \quad \mathbb{E}_k[\|G_k - \nabla f(x_k)\|_2 | \nabla f(x_k)^\top (D_k - d_k^{\text{true}}) < 0, \mathcal{T}_k < \tau_{k-1}] \leq M_\tau. \end{aligned} \tag{20}$$

Observe that the inequality in (19) can imply (20), such as when there exists $p \in (0,1]$ such that, for all $k \in [k_{\max}]$ and $g_{[k-1]}$, one finds that

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

whenever this probability is nonzero. This occurs, 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 $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$, one finds that*

$$\mathbb{P}_k[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\}] \geq p_\tau.$$

Similar to [1, Proposition 3.16], Assumption 4 allows us to prove that, with high probability, the number of iterations in which $\tau_k > \tau_k^{\text{trial,true}}$ is not too large. In [1, Example 3.17], it was shown that the inequality in this assumption holds with $p_\tau = \frac{1}{2}$ when, conditioned on having reached x_k, the stochastic gradient G_k has a Gaussian distribution. We show in Section 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 results from [1] that are needed for our analysis.

By [1, 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 ζ is defined in Assumption 2. Correspondingly, let us define

$$\Psi_k := \begin{cases} \|u_k^{\text{true}}\|^2 + \|c_k\| & \text{if } \|u_k^{\text{true}}\|^2 \geq \kappa_{uv}\|v_k\|^2 \\ \|c_k\| & \text{otherwise.} \end{cases}$$

The following lemma is stated using a different norm (for c_k) than in [1, Lemma 2.11]. The result holds in the same manner due to the norm-equivalence between $\|\cdot\|$ and $\|\cdot\|_1$ in $\mathbb{R}^m$. We state the result in this manner for consistency in the measure of constraint violation stated in our final complexity bound.

Lemma 1 ([1, 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 $\|d_k^{\text{true}}\|^2 + \|c_k\|_1 \leq (\kappa_\Psi + 1)\Psi_k$.*

Lemma 2 ([1, 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 any $\tau \leq \tau_k^{\text{trial,true}}$, the true reduction in the merit model satisfies $\Delta q(x_k, \tau, \nabla f(x_k), H_k, d_k^{\text{true}}) \geq \kappa_q \tau \Psi_k$.*

Lemma 3 ([1, Lemma 3.7])**.** *Let Assumptions 1 and 2 hold and suppose that the sequence $\{\beta_k\}$ is chosen such that $\beta_k \xi_k \tau_k/(\tau_k L + \Gamma) \in (0, 1]$ for all $k \in [k_{\max}]$. Then, for all $k \in [k_{\max}]$, it follows that*

$$\begin{aligned} \phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \leq & -\alpha_k \Delta q(x_k, \tau_k, \nabla f(x_k), H_k, d_k^{\text{true}}) \\ & + \tfrac{1}{2}\alpha_k\beta_k \Delta q(x_k, \tau_k, g_k, H_k, d_k) \\ & + \alpha_k \tau_k \nabla f(x_k)^\top (d_k - d_k^{\text{true}}). \end{aligned} \tag{21}$$

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

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

Proof. Except for the final inequality, the result follows directly from [1, Lemma 3.8] (or the proof therein) and (2). As for the final inequality, observe by the arguments in [1, Lemma 3.8] that for any realization of G_k and D_k one finds $\|d_k - d_k^{\text{true}}\| \leq \kappa_d \|g_k - \nabla f(x_k)\|$, which combined with (20) gives

$$\mathbb{E}_k[\|D_k - d_k^{\text{true}}\| | \nabla f(x_k)^\top (D_k - d_k^{\text{true}}) < 0, \mathcal{T}_k < \tau_{k-1}]$$

$$\leq \kappa_d \mathbb{E}_k[\|G_k - \nabla f(x_k)\| |\nabla f(x_k)^\top (D_k - d_k^{\text{true}}) < 0, \mathcal{T}_k < \tau_{k-1}] \leq \kappa_d M_\tau,$$

as desired. □ □

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

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

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 [1], 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 x_k being the kth iterate in a run of the algorithm, 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 $\mathcal{T}_k < \tau_{k-1}$, 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 the entire run of the algorithm.

2. A critical aspect of the analysis in [1] 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 $\tau_k^{\text{trial,true}} \geq \tau_k$ 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 10 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 10 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 the algorithm having generated a certain sequence of stochastic gradient estimates through a given iteration. In particular, given $k \in [k_{\max}]$ and $g_{[k-1]}$, values generated by the algorithm have been determined up to the beginning of iteration k, including x_k, d_k^{true}, and τ_{k-1}. Given these quantities, 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 $\mathcal{T}_k = \tau_{k-1}$; and
- $E_{k,3}$, the event that $\nabla f(x_k)^\top (D_k - d_k^{\text{true}}) < 0$ and $\mathcal{T}_k < \tau_{k-1}$.

We now derive an upper bound on the final term in (21). In the following lemma, we make use, for given $k \in [k_{\max}]$ and $g_{[k-1]}$, of the stepsize values

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

The first value here represents a lower bound on the smallest stepsize that may be computed in the event that $\mathcal{T}_k < \tau_{k-1}$, while the second value is the smallest stepsize that may be computed in the event that $\mathcal{T}_k = \tau_{k-1}$; it is easily verified that $\alpha^{<}_{\min,k} < \alpha^{=}_{\min,k}$. Hence, $\varsigma_{\max,k}$ 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 $g_{[k-1]}$, and with the stepsizes $(\alpha^{<}_{\min,k}, \alpha^{=}_{\min,k}, \varsigma_{\max,k})$ defined as in (22), one finds that*

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

Proof. Consider arbitrary $k \in [k_{\max}]$ and $g_{[k-1]}$, 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 $0 < \tau_{\min} \leq \mathcal{T}_k \leq \tau_{k-1}$ under E, and the definitions of $\alpha^{<}_{\min,k}$, $\alpha^{=}_{\min,k}$, and $\varsigma_{\max,k}$, one finds that

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

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

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

and

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

and

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

Averaging these three upper bounds and the definition of $\varsigma_{\max,k}$, one obtains

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

This bound can be rewritten as follows. By the Law of Total expectation,

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

and along with Lemma 4 and (2) one finds

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

In addition, Lemma 4 and (2) also yield that

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

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

Now, conditioned on given $k \in [k_{\max}]$ and $g_{[k-1]}$, let us define

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

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 $k \in [k_{\max}]$ and $g_{[k-1]}$, one finds that*

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

Proof. By the definition of Δq in (10), one has that

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

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

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

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

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

and by similar reasoning one finds that

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

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

Averaging these two upper bounds, one finds that

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

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, Jensen's inequality, and concavity of the square root over $\mathbb{R}_{\geq 0}$, one finds that

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

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

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

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

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

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 $k \in [k_{\max}]$ and $g_{[k-1]}$, with $(\alpha^{<}_{\min,k}, \alpha^{=}_{\min,k}, \varsigma_{\max,k})$ defined as in (22) and $\hat{\mathcal{T}}_k$ defined in (27), one finds that*

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

and

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

Proof. Under the stated assumptions, it follows from the stated lemma and definitions, along with the definition of Δq in (10), that

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

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

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

which together are the desired conclusions. □ □

Our next two lemmas are critical elements of our analysis. For both lemmas, we define, for any $s \in \mathbb{N}$ and $\delta \in (0,1)$, the quantities

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

and

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

The first of these lemmas bounds, with high probability, the sum of the probabilities of the occurrences of event $E_{k,3}$ over the run of the algorithm.

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

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[E_{k,3}|E,\mathcal{F}_k] \leq \ell(s_{\max},\hat{\delta}) + 1 \middle| E\right] \geq 1-\delta. \tag{34}$$

Proof. This result is proved in Appendix 7. □ □

For our next lemma, let us define the random index set

$$\mathcal{K}_\tau := \{k \in [k_{\max}] : \mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}\}. \tag{35}$$

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 (35), 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 (35) is the fact that, conditioned on the algorithm having generated a particular set of stochastic gradients up to the beginning of iteration k, whether or not $\tau_k^{\text{trial,true}} < \tau_{k-1}$ holds has already been determined; in other words, the occurrence of this inequality does not depend on G_k. This means that G_k is conditionally independent of the event $\mathcal{T}_k^{\text{trial,true}} < \mathcal{T}_{k-1}$ given $g_{[k-1]}$ and E, which is a fact that is exploited in the proof of the lemma.

Lemma 10. *Suppose Assumptions 1, 2, 3, and 4 hold and let $\mathcal{K}_\tau$ be defined as in* (35)*. Then, for any $\delta \in (0,1)$, it follows that*

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

Proof. This result is proved in Appendix 7. □

We are now prepared to prove a convergence rate result.

Theorem 2. *Suppose Assumptions 1, 2, 3, and 4 hold, let $s_{\max} \geq 1$, 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

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

define

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

and, for all $k \in [k_{\max}]$ and $g_{[k-1]}$, let $\hat{\mathcal{T}}_k$ 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 ℓ defined as in (32) and (33) that, with probability at least $1-\delta$,

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}_k[-\mathcal{A}_k \Delta q(x_k, \hat{\mathcal{T}}_k, \nabla f(x_k), H_k, d_k^{\text{true}}) + \tfrac{1}{2}\mathcal{A}_k \beta \Delta q(x_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\quad + (\varsigma_{\max,k}\tau_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\tau_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\leq -A_{\min}\beta\mathbb{E}_k[\Delta q(x_k, \hat{\mathcal{T}}_k, \nabla f(x_k), H_k, d_k^{\text{true}})] \\
&\quad + \tfrac{1}{2}(A_{\max}\beta + \theta\beta^2)\beta\mathbb{E}_k[\Delta q(x_k, \hat{\mathcal{T}}_k, G_k, H_k, D_k)] \\
&\quad + ((A_{\max}\beta + \theta\beta^2)\tau_{k-1} - A_{\min}\beta\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\tau_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\leq -(A_{\min} - \tfrac{1}{2}(A_{\max} + \theta\beta)\beta)\beta\mathbb{E}_k[\Delta q(x_k, \hat{\mathcal{T}}_k, \nabla f(x_k), H_k, d_k^{\text{true}})] \\
&\quad + \tfrac{1}{4}(A_{\max} + \theta\beta)\beta^2(\tau_{k-1} + \tau_{\min})\zeta^{-1}M \\
&\quad + \tfrac{1}{2}(A_{\max} + \theta\beta)\beta^2(\tau_{k-1} - \tau_{\min})(\kappa_d\sqrt{M}(2\kappa_g + \sqrt{M}) + \kappa_H\kappa_d^2(M + \tfrac{3}{2}\kappa_g^2)) \\
&\quad + ((A_{\max} + \theta\beta)\tau_{k-1} - A_{\min}\tau_{\min})\beta\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}] + \theta\beta^2\tau_{k-1}\kappa_g\kappa_d\sqrt{M} \\
&\leq -\tfrac{1}{2}A_{\min}\beta\mathbb{E}_k[\Delta q(x_k, \hat{\mathcal{T}}_k, \nabla f(x_k), H_k, d_k^{\text{true}})] \\
&\quad + \kappa_{E_3}\beta\mathbb{P}_k[E_{k,3}] + \overline{M}\beta^2, \qquad (41)
\end{aligned}$$

where the final inequality follows due to the fact that $(A_{\max}+\theta\beta)\beta \leq A_{\min}$ holds by the definitions of β, γ, $A_{\min}$, 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 $\mathcal{T}_k^{\text{trial,true}} < \tau_{k-1}$ and let $E_{k,\tau}^c$ be the complementary event that $\mathcal{T}_k^{\text{trial,true}} \geq \tau_{k-1}$. Observe that whether $E_{k,\tau}$ or $E_{k,\tau}^c$ occurs is determined by the condition that $G_{[k-1]} = g_{k-1}$. Hence, the bound in Lemma 7—and in Lemma 8 as well, which is used below—holds even if one conditions on the occurence of $E_{k,\tau}$ or of $E_{k,\tau}^c$. Consequently, the bounds in (39) and (41) hold even if one also 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), and the fact that $\mathcal{T}_k^{\text{trial,true}} \geq \tau_{k-1} \geq \mathcal{T}_k = \hat{\mathcal{T}}_k$ (by (27)) in $E_{k,\tau}^c$ that

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

On the other hand, one finds from (39), (41), 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}] \\
&\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)|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}}) + \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 - \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)|E_{k,\tau}] \\
&\quad + (\varsigma_{\max,k}\tau_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_g\kappa_d M_\tau \mathbb{P}_k[E_{k,3}|E_{k,\tau}] + \theta\beta^2\tau_{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}}) + \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}}) + \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 + (\varsigma_{\max,k}\tau_{k-1} - \alpha^{<}_{\min,k}\tau_{\min})\kappa_d\kappa_g^2(1 + \tfrac{1}{2}\kappa_H\kappa_d) \\
&\quad + \tfrac{1}{2}(\varsigma_{\max,k}\tau_{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

$$
\mathbb{E}_k[\phi(x_k + \mathcal{A}_k D_k, \mathcal{T}_k)] - \mathbb{E}_k[\phi(x_k, \mathcal{T}_k)]
$$

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

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

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

Let us now turn to analyzing the overall behavior of the algorithm through iterations $k \in [k_{\max}]$. Let $f_{G_{[k_{\max}]}}$ denote the probability density function of $G_{[k_{\max}]}$ and observe that, for all $k \in [k_{\max}]$, one finds $\mathbb{P}_k[E_{k,\tau}] \in \{0, 1\}$ since, when conditioned on E and $\mathcal{F}_k$, the event $E_{k,\tau}$ is independent of G_k. Therefore, by the bound above, the law of total expectation, Lemma 9, Lemma 10, and the union bound, it follows that, with probability at least $1 - \delta$, one finds

$$
\begin{aligned}
&\sum_{k=0}^{k_{\max}} (\mathbb{E}[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k)|E, \mathcal{F}_{k+1}] - \mathbb{E}[\phi(X_k, \mathcal{T}_k)|E, \mathcal{F}_{k+1}]) \\
&= \int_{g_{[k_{\max}]} \in \mathcal{F}_{k_{\max}+1}} \sum_{k=0}^{k_{\max}} \mathbb{E}[\phi(x_k + \mathcal{A}_k D_k, \mathcal{T}_k) - \phi(x_k, \mathcal{T}_k)|E, G_{[k_{\max}]} = g_{[k_{\max}]}] \\
&\qquad\qquad \cdot f_{G_{[k_{\max}]}}(g_{[k_{\max}]}) \mathrm{d}g_{[k_{\max}]} \\
&\leq \int_{g_{[k_{\max}]} \in \mathcal{F}_{k_{\max}+1}} \Bigg(-\tfrac{1}{2} A_{\min}\beta \sum_{k=0}^{k_{\max}} \mathbb{E}[\Delta q(x_k, \hat{\mathcal{T}}_k, \nabla f(x_k), H_k, d_k^{\text{true}})|E, G_{[k_{\max}]} = g_{[k_{\max}]}] \\
&\qquad\qquad + \kappa_{E_3}\beta \sum_{k=0}^{k_{\max}} \mathbb{P}[E_{k,3}|E, G_{[k_{\max}]} = g_{[k_{\max}]}] \\
&\qquad\qquad + (\kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2) \sum_{k=0}^{k_{\max}} \mathbb{P}[E_{k,\tau}|E, G_{[k_{\max}]} = g_{[k_{\max}]}] \Bigg) \\
&\qquad\qquad \cdot f_{G_{[k_{\max}]}}(g_{[k_{\max}]}) \mathrm{d}g_{[k_{\max}]} \\
&\quad + (k_{\max}+1)\overline{M}\beta^2 \\
&= -\tfrac{1}{2} A_{\min}\beta \sum_{k=0}^{k_{\max}} \mathbb{E}[\Delta q(X_k, \hat{\mathcal{T}}_k, \nabla f(X_k), H_k, D_k^{\text{true}})|E, \mathcal{F}_k] \\
&\quad + \kappa_{E_3}\beta \sum_{k=0}^{k_{\max}} \mathbb{P}[E_{k,3}|E, \mathcal{F}_k] + (\kappa_{\Delta q,1}\beta + \kappa_{\Delta q,2}\beta^2)\mathbb{E}[|\mathcal{K}_\tau||E] + (k_{\max}+1)\overline{M}\beta^2
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & -\tfrac{1}{2}A_{\min}\beta\sum_{k=0}^{k_{\max}}\mathbb{E}[\Delta q(X_k,\hat{\mathcal{T}}_k,\nabla f(X_k),H_k,D_k^{\text{true}})|E,\mathcal{F}_k]\\
&+\kappa_{E_3}\beta(\ell(s_{\max},\hat{\delta}/2)+1)\\
&+(\kappa_{\Delta q,1}\beta+\kappa_{\Delta q,2}\beta^2)\left\lceil\frac{\ell(s_{\max},\hat{\delta}/2)+1}{p_\tau}\right\rceil+(k_{\max}+1)\overline{M}\beta^2. \qquad (42)
\end{aligned}
$$

The left-hand side of this inequality satisfies

$$
\begin{aligned}
&\sum_{k=0}^{k_{\max}}(\mathbb{E}[\phi(X_k+\mathcal{A}_kD_k,\mathcal{T}_k)|E,\mathcal{F}_{k+1}]-\mathbb{E}[\phi(X_k,\mathcal{T}_k)|E,\mathcal{F}_{k+1}])\\
=\ &\sum_{k=0}^{k_{\max}}\Bigg(\mathbb{E}[\mathcal{T}_k(f(X_k+\mathcal{A}_kD_k)-f_{\min})+\|c(X_k+\mathcal{A}_kD_k)\|_1|E,\mathcal{F}_{k+1}]\\
&\qquad-\mathbb{E}[\mathcal{T}_k(f(X_k)-f_{\min})+\|c(X_k)\|_1|E,\mathcal{F}_{k+1}]\Bigg). \qquad (43)
\end{aligned}
$$

By the merit parameter updating strategy and the definition of the filtration (which guarantees that $\mathcal{T}_{k-1}$ and X_k are fully determined by $G_{[k-1]}$, and are therefore conditionally independent of $\mathcal{F}_{k+1}$) that, for all $k\in[k_{\max}]$, one finds

$$
\begin{aligned}
&-\mathbb{E}[\mathcal{T}_k(f(X_k)-f_{\min})+\|c(X_k)\|_1|E,\mathcal{F}_{k+1}]\\
\geq\ &-\mathbb{E}[\mathcal{T}_{k-1}(f(X_k)-f_{\min})+\|c(X_k)\|_1|E,\mathcal{F}_{k+1}]\\
=\ &-\mathbb{E}[\mathcal{T}_{k-1}(f(X_k)-f_{\min})+\|c(X_k)\|_1|E,\mathcal{F}_k].
\end{aligned}
$$

Thus, from (43), it follows that

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

Combining this with (42), one obtains that

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

Hence, by the definitions of K^* and β, 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 (3).

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

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

Hence, the complexity bound described in Section 2.2 (see (7)*) holds.*

Proof. The result follows by Lemma 1, Lemma 2, Theorem 2, and (15), which implies that $\|\nabla f(X_{K^*}) + J_{K^*}^\top Y_{K^*}^{true}\| = \|H_{k^*} D_{K^*}^{true}\| \leq \kappa_H \|D_{K^*}^{true}\|$. The worst-case complexity bound described in Section 2.2 follows by combining this result with the definitions of κ_{E_3}, $\kappa_{\Delta q,1}$, $\kappa_{\Delta q,2}$, and Lemma 22 in Appendix 7. □ □

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 δ ommitted. 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 [11].

4.5 Complexity Result for Symmetric Sub-Gaussian Distributions

In this section, we show that Assumption 3, Assumption 4, and the event E in (17) occur with high probability when each stochastic gradient is unbiased and has a symmetric, sub-Gaussian distribution and the ratio parameter sequence remains constant. For these purposes, we make the following assumption.

Assumption 5. *There exists* $M \in \mathbb{R}_{>0}$ *such that, for all* $k \in [k_{\max}]$ *and any realization* $g_{[k-1]}$ *of* $G_{[k-1]}$*, one finds that*

$$\begin{aligned}
\mathbb{E}[G_k | G_{[k-1]} = g_{[k-1]}] &= \nabla f(x_k) \\
\text{and } \ \mathbb{E}[\exp(\|G_k - \nabla f(x_k)\|^2 / M) | G_{[k-1]} = g_{[k-1]}] &\leq \exp(1),
\end{aligned} \tag{44}$$

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* Ξ_k *satisfies* $\Xi_k = \xi_{\min}$*.*

Our first lemma shows that, under Assumption 5, Assumption 4 holds.

Lemma 11. *Under Assumptions 1, 2, and 5, it follows for all $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$ that*

$$\mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} | g_{[k-1]}] \geq \tfrac{1}{2}.$$

Proof. Consider arbitrary $k \in [k_{\max}]$. Let Z_k be a basis for the null space of J_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 a random vector 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 (9), $Z_k W_k = -Z_k(Z_k^\top H_k Z_k)^{-1} Z_k^\top (G_k + H_k v_k)$, so that

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

and similarly

$$\begin{aligned} &\nabla f(x_k)^\top d_k^{\text{true}} + (d_k^{\text{true}})^\top H_k d_k^{\text{true}} \\ &= v_k^\top H_k^{1/2}(I - H_k^{1/2} Z_k (Z_k^\top H_k Z_k)^{-1} Z_k^\top H_k^{1/2})(H_k^{-1/2} \nabla f(x_k) + H_k^{1/2} v_k). \end{aligned}$$

Hence, the random variables

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

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,

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

and

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

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 [16, Lemma 5].

Lemma 12. *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 this subsection by showing that, under Assumption 5, both Assumption 3 and event E occur with high probability.

Lemma 13. *Suppose that Assumptions 1, 2, and 5 hold, let κ_v be defined as in [1, Lemma 2.9], let κ_c be an upper bound for $\|c_k\|_2$ for all $k \in [k_{\max}]$ (the existence of which follows under Assumption 1), and define*

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

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

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

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

Proof. By Lemma 12, 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 (44), it follows that

$$\mathbb{E}[\|G_k - \nabla f(x_k)\|^2 | G_{[k-1]} = g_{[k-1]}] \leq M.$$

In addition, it follows from the event in Lemma 12 that (20) holds with M_τ as stated in the lemma. This accounts for Assumption 3. Now consider event E. First, it follows from the arguments of [1, Lemma 2.15 and 2.16] combined with the event in Lemma 12 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 (18) 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 [1, 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 (18)) 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, in every realization of the algorithm, the merit parameters $\{\tau_k\}_{k=0}^{k_{\max}}$ and ratio parameters $\{\xi_k\}_{k=0}^{k_{\max}}$ satisfy

- $\tau_k \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\tau_k^{\text{trial,true}} \geq \tau_{\min} > 0$ for all $k \in [k_{\max}]$,
- $\xi_k \geq \xi_{\min} > 0$ for all $k \in [k_{\max}]$,
- $|\{k \in [k_{\max}] : \tau_k < \tau_{k-1}\}| \leq s_{\max}$, 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 (17), except that one may have $\xi_0 > \xi_{\min}$.

For the purposes of this section, redefining

$$\mathbb{P}_k[\cdot] := \mathbb{P}[\cdot | E_\xi, G_{[k-1]} = g_{[k-1]}] \quad \text{and} \quad \mathbb{E}_k[\cdot] := \mathbb{P}[\cdot | E_\xi, G_{[k-1]} = g_{[k-1]}],$$

our analysis of this case considers the following replacement of Assumption 3.

Assumption 6. *There exists $M \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$, one finds that*

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

In addition, there exist $M_1 \in \mathbb{R}_{>0}$, $M_2 \in \mathbb{R}_{>0}$, and $M_3 \in \mathbb{R}_{>0}$ such that, for all $k \in [k_{\max}]$ and any realization $g_{[k-1]}$ of $G_{[k-1]}$, one finds that

$$\begin{aligned}
&\textit{either} \ \ \mathbb{P}_k[\nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k < \tau_{k-1}, \Xi_k = \xi_{k-1}] = 0 \\
\textit{or} \ \ &\mathbb{E}_k[\|G_k - \nabla f(x_k)\| | \nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k < \tau_{k-1}, \Xi_k = \xi_{k-1}] \leq \kappa_1; \\
&\textit{either} \ \ \mathbb{P}_k[\nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k = \tau_{k-1}, \Xi_k < \xi_{k-1}] = 0 \\
\textit{or} \ \ &\mathbb{E}_k[\|G_k - \nabla f(x_k)\| | \nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k = \tau_{k-1}, \Xi_k < \xi_{k-1}] \leq \kappa_2; \\
&\textit{and either} \ \ \mathbb{P}_k[\nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k < \tau_{k-1}, \Xi_k < \xi_{k-1}] = 0 \\
\textit{or} \ \ &\mathbb{E}_k[\|G_k - \nabla f(x_k)\| | \nabla f(x_k)^\top (D_k - d_k^{\text{true}})) < 0, \mathcal{T}_k < \tau_{k-1}, \Xi_k < \xi_{k-1}] \leq \kappa_3.
\end{aligned}$$

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 Section 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 straightforward, but tedious extension of the results in Section 4.4, we simply state the extension of (7) 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*

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

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 [1]. 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 the merit function in (4), 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, 17], adaptive augmented Lagrangian methods [9], adaptive barrier methods [20], and penalty-interior point methods [8]. In addition, many constrained optimization algorithms generate (often unconstrained) subproblems defined by an auxilliary 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 [21]. In cases when the objective is stochastic, this auxilliary 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.

References

[1] A. S. Berahas, F. E. Curtis, D. P. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

[2] Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear Equality Constrained Optimization with Rank-Deficient Jacobians. arXiv 2106.13015, 2021.

[3] Dimitri P Bertsekas. *Network optimization: continuous and discrete models*. Athena Scientific Belmont, 1998.

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

[5] Richard H Byrd, Gabriel Lopez-Calva, and Jorge Nocedal. A line search exact penalty method using steering rules. *Mathematical Programming*, 133(1):39–73, 2012.

[6] Richard H Byrd, Jorge Nocedal, and Richard A Waltz. Steering exact penalty methods for nonlinear programming. *Optimization Methods and Software*, 23(2):197–213, 2008.

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

[8] Frank E Curtis. A penalty-interior-point algorithm for nonlinear constrained optimization. *Mathematical Programming Computation*, 4(2):181–209, 2012.

[9] Frank E Curtis, Hao Jiang, and Daniel P Robinson. An adaptive augmented lagrangian method for large-scale constrained optimization. *Mathematical Programming*, 152(1):201–245, 2015.

[10] Frank E Curtis and Daniel P Robinson. Exploiting negative curvature in deterministic and stochastic optimization. *Mathematical Programming*, 176(1):69–94, 2019.

[11] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[12] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[13] S. P. Han and O. L. Mangasarian. Exact penalty functions in nonlinear programming. *Math. Programming*, 17(3):251–269, 1979.

[14] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271. PMLR, 2016.

[15] FS Kupfer and Ekkehard W Sachs. Numerical solution of a nonlinear parabolic control problem by a reduced sqp method. *Computational Optimization and Applications*, 1(1):113–135, 1992.

[16] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint 2007.14294*, 2020.

[17] Marcel Mongeau and Annick Sartenaer. Automatic decrease of the penalty parameter in exact penalty function methods. *European Journal of Operational Research*, 83(3):686–699, 1995.

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

[19] Yatin Nandwani, Abhishek Pathak, Mausam, and Parag Singla. A primal dual formulation for deep learning with constraints. In *NeurIPS*, 2019.

[20] Jorge Nocedal, Andreas Wächter, and Richard A Waltz. Adaptive barrier update strategies for nonlinear interior methods. *SIAM Journal on Optimization*, 19(4):1674–1693, 2009.

[21] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

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

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

[24] David Stirzaker. *Elementary probability*. Cambridge University Press, 2003.

[25] Robert B Wilson. A simplicial algorithm for concave programming. *Ph. D. Dissertation, Graduate School of Bussiness Administration*, 1963.

6 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 [1]. We refer to quantities defined and employed in the analysis in [1]. 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 [1], 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 14. *Define* $(\kappa_{uv}, \kappa_H, \kappa_v, \tau_{\min}, \zeta, \sigma) \in (0,\infty)^5 \times (0,1)$ *as in [1] and let*

$$\hat{\kappa} := \min\left\{1, \frac{1}{(1+\kappa_{uv})\kappa_v \kappa_H^2}\right\} \quad \textit{and} \quad \tilde{\kappa} := \tfrac{1}{4}\zeta\kappa_{uv}\kappa_v\hat{\kappa}. \tag{45}$$

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

$$\Delta q(x_k, \tau_k, g_k, H_k, d_k) \geq \min\left\{\sigma\hat{\kappa}, \tau_{\min}\tilde{\kappa}\right\}\varepsilon^2. \tag{46}$$

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 [1, 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 (46), as desired. Second, suppose that $\|c_k\|_1 \leq \hat{\kappa}\varepsilon^2 \leq \varepsilon^2$, which by the definition of (ε, 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 [1, 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 [1, 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

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

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

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

which implies (46), as desired. □ □

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

Theorem 4. *Define* $(\tau_{-1}, f_{\text{low}}, \alpha_{\min}, \tau_{\min}, \eta, \sigma) \in (0,\infty)^4 \times (0,1)^2$ *as in [1] and* $(\hat{\kappa}, \tilde{\kappa}) \in (0,1] \times (0,\infty)$ *as in* (45)*. Then, for any* $\varepsilon \in (0,1)$*, Theorem 1 holds with* (6) *given by*

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

Proof. To derive a contradiction, suppose (5) does not hold for all $k \in \{0, \ldots, \overline{K}_\varepsilon\}$. Then, along with Lemma 14 and [1, equation (2.10) and Lemma 2.17], it follows for all such k that

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

By the definition of ϕ, 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

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

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 (5) in at most $\overline{K}_\varepsilon$ iterations. □ □

7 Proofs of Lemmas 9 and 10

In this appendix, we prove Lemmas 9 and 10. 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

$$\mathbb{P}\left[\sum_{i=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | E, G_{[i-1]}] \leq \ell(s_{\max}, \hat{\delta}) + 1 \middle| E\right] \geq 1 - \delta. \tag{47}$$

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

As our first preliminary result, we state a particular form of Chernoff's bound in the following lemma, which will prove instrumental in deriving (47).

Lemma 15. *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*

$$\mu := \sum_{j=0}^{k} \mathbb{P}[Y_j = 1] \geq \ell(s, \bar{\delta}) \quad \Longrightarrow \quad \mathbb{P}\left[\sum_{j=0}^{k} Y_j \leq s\right] \leq \bar{\delta}. \tag{48}$$

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 (48)) 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 μ, one finds that the inequality holds as long as μ 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 (47). 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 $\bar{p}_k$ 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 $(\bar{p}_0, \ldots, \bar{p}_k, w_0, \ldots, w_k)$, 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 $\tilde{N}(\bar{p}_0, w_0)$, where $\bar{p}_0$ is uniquely defined by the starting conditions

of our algorithm and $w_0 = 0$. All realizations of our algorithm follow the same initialization, so $\bar{p}_0$ and w_0 would be in the signature of every realization. Now, one could define a node $\tilde{N}(\bar{p}_{[k]}, w_{[k]})$ 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 $(\bar{p}_0, \dots, \bar{p}_k, w_0, \dots, w_k)$. One could then define the edges in the tree by connecting nodes at adjacent levels, where node $\tilde{N}(\bar{p}_{[k]}, w_{[k]})$ is connected to node $\tilde{N}(\bar{p}_{[k]}, \bar{p}_{k+1}, w_{[k]}, w_{k+1})$ for any $\bar{p}_{k+1} \in [0,1]$ and $w_{k+1} \in \{w_k, w_k + 1, \dots\}$.

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

$$p_{[k]} = (p_0, \dots, p_k) \in \left\{0, \tfrac{1}{B}, \dots, \tfrac{B-1}{B}\right\}^{k+1}. \tag{49}$$

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

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

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 $p_0 \in \{0, 1/B, \dots, (B-1)/B\}$ is uniquely defined by the starting conditions of our algorithm so that $\mathbb{P}[\mathcal{T}_0 < \tau_{-1} | E] \in \iota(p_0)$ and $w_0 = 0$. All realizations of our algorithm follow the same initialization, so with $\bar{p}_0 = \mathbb{P}[\mathcal{T}_0 < \tau_{-1} | E]$ 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 $(\bar{p}_{[k]}, w_{[k]})$ is a member of $N(p_{[k]}, w_{[k]})$ if and only if, for all $j \in [k]$, one finds that $\bar{p}_j \in \iota(p_j)$. 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 $p_{k+1} \in \{0, 1/B, \dots, (B-1)/B\}$ and $w_{k+1} \in \{w_k, w_k + 1, \dots\}$.

Notationally, since the behavior of a realization of the algorithm up to iteration $k \in \mathbb{N}$ is completely determined by the initial conditions and the realization of $G_{[k-1]}$, we say that a realization described by $G_{[k-1]}$ belongs in node $N(p_{[k]}, w_{[k]})$ by writing that

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

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

$$\begin{aligned} &\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | E, G_{[k-1]}] \in \iota(p_k), \\ &S_{k-1} := \sum_{i=0}^{k-1} \mathcal{I}[\mathcal{T}_i < \mathcal{T}_{i-1}] = w_k, \\ &\qquad \text{and} \quad G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}). \end{aligned} \tag{50}$$

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 $k_{\max}$. 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 $\ell(s_{\max}, \hat{\delta}) + 1$. 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 $C_{dec}(N(p_{[k]}, w_{[k]}))$ 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

$$\begin{aligned} C_{\text{dec}}(N(p_{[k]}, w_{[k]})) := \{&N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) : \\ &(N(p_{[k]}, p_{k+1}, w_{[k]}, w_{k+1}) \in C(N(p_{[k]}, w_{[k]}))) \\ &\wedge (w_{k+1} = w_k + 1)\} \end{aligned} \tag{51}$$

and

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

$$\wedge\,(w_{k+1}=w_k)\}. \tag{52}$$

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

$$\left(\sum_{i=0}^{j}\mathbb{P}[\mathcal{T}_i<\mathcal{T}_{i-1}|E,G_{[i-1]}]>\ell(s_{\max},\hat{\delta})+\tfrac{k_{\max}+1}{B}+1\right). \tag{53}$$

With respect to our goal of proving (47), the event $E_{\mathrm{bad},B}$ 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 (47) plus a factor that is inversely proportional to B.

Let us now prove some properties of the leaf nodes.

Lemma 16. *For any $k\in[k_{\max}]$ and $(p_{[k]},w_{[k]})$ with $N(p_{[k]},w_{[k]})\in\mathcal{L}_{\mathrm{good}}$, one finds*

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

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

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

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

$$\sum_{i=0}^{k}p_i\leq\ell(s_{\max},\hat{\delta})+1. \tag{54}$$

Since the maximum depth of a node is $k_{\max}$, it follows from (54) that

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

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

$$\begin{aligned}&\mathbb{P}\left[G_{[j-1]}\in N(p_{[j]},w_{[j]})\wedge(53)\text{ holds}|E\right]\\&=\mathbb{P}\left[\sum_{i=0}^{j}\mathbb{P}[\mathcal{T}_i<\mathcal{T}_{i-1}|E,G_{[i-1]}]>\ell(s_{\max},\hat{\delta})+\tfrac{k_{\max}+1}{B}+1\Big|E,G_{[j-1]}\in N(p_{[j]},w_{[j]})\right]\\&\quad\cdot\mathbb{P}\left[G_{[j-1]}\in N(p_{[j]},w_{[j]})|E\right]=0.\end{aligned}$$

In addition, (53) cannot hold for $j=0$ since $\ell(s_{\max},\hat{\delta})+1>1$. Hence, along with the conclusion above, it follows that $E_{\mathrm{bad},B}$ does not occur in any realization whose signature up to iteration

$j \in \{1, \ldots, k\}$ falls into a node along any path from the root note 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

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

Otherwise, if $w_k = s_{\max}$, then it follows by Assumption 3 that $\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | E, G_{[i-1]}] = 0$ 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.

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

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

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

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\leq \mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) | E] = \mathbb{P}\left[(50) \text{ holds} \big| E\right] \\
&= \mathbb{P}\left[\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | E, G_{[k-1]}] \in \iota(p_k) \big| E, S_{k-1} = w_k, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\right] \\
&\quad \cdot \mathbb{P}\left[S_{k-1} = w_k \wedge G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}) \big| E\right] \\
&= \mathbb{P}\left[\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | E, G_{[k-1]}] \in \iota(p_k) \big| E, S_{k-1} = w_k, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\right] \\
&\quad \cdot \mathbb{P}\left[S_{k-1} = w_k \big| E, G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]})\right] \mathbb{P}\left[G_{[k-2]} \in N(p_{[k-1]}, w_{[k-1]}) \big| E\right] \\
&= \mathbb{P}[G_{-1} \in N(p_0, w_0)] \\
&\quad \cdot \prod_{i=1}^{k} \Big(\mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | E, G_{[i-1]}] \in \iota(p_i) \big| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \\
&\qquad \cdot \mathbb{P}\left[S_{i-1} = w_i \big| E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \Big) \\
&= \prod_{i=1}^{k} \Big(\mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | E, G_{[i-1]}] \in \iota(p_i) \big| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \\
&\quad \cdot \mathbb{P}\left[S_{i-1} = w_i \big| E, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right] \Big).
\end{aligned} \tag{55}$$

Our goal is to bound (55). 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 $i \in \mathcal{I}_{\text{dec}}$,

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

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

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

Thus, it follows that the latter term in (55) satisfies

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

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

$$\sum_{i=0}^{k} p_i > \ell(s_{\max}, \hat{\delta}) + 1 \implies \sum_{i=0}^{k-1} p_i > \ell(s_{\max}, \hat{\delta}). \tag{56}$$

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 $|\mathcal{I}_{\mathrm{dec}}| \leq s_{\max}$. 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

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

By (56), 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 15 and the preceding argument that

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

Combining this with (55), the desired conclusion follows. □ □

Next, we present a lemma about nodes in the sets defined in (51) and (52). 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 17. *For all $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$, one finds that*

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

and, similarly, one finds that

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

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

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

The desired conclusion follows since, by the definition of $C_{\mathrm{dec}}(N(p_{[k]}, w_{[k]}))$, all elements in the latter sum have $S_k = w_{k+1} = w_k + 1$, meaning that the sum on the right-hand side is the sum of all conditional probabilities 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_{dec} in place of C_{dec} and $S_k = w_{k+1} = w_k$ in place of $S_k = w_k = w_{k-1} + 1$. □ □

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

Lemma 18. *For any $k \in [k_{\max}]$ and $(p_{[k]}, w_{[k]})$ such that $w_k = s_{\max} - 1$ and $N(p_{[k]}, w_{[k]}) \notin \mathcal{L}_{\text{good}}$, it follows that*

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\text{bad},B} | E] \\
&\leq \hat{\delta} \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1} | E, G_{[i-1]}] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right]. \qquad (57)
\end{aligned}$$

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 16. 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_{dec}, and C_{dec}^c, it follows that

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

Since $w_k = s_{\max} - 1$, it follows from the definition of C_{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 $\mathbb{P}[\mathcal{T}_{k+1} < \mathcal{T}_k | E, G_{[k]}] = 0$, 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 16, one finds

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

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 17, it follows that

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

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

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

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 16 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 18. 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 $r \in \mathbb{N} \setminus \{0\}$, 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 $s_{\max} - w_k = r$.

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_{dec}, and C_{dec}^c, it follows that

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

Further by the definition of C_{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_{dec}^c 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

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

$$+ \sum_{\{(p_{[k+1]},w_{[k+1]}):N(p_{[k+1]},w_{[k+1]})\in C^c_{\mathrm{dec}}(N(p_{[k]},w_{[k]}))\}} \hat{\delta} \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j}$$
$$\cdot \prod_{i=1}^{k+1} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|E, G_{[i-1]}] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-2]})\right],$$

which by Lemma 17 implies that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(p_{[k]}, w_{[k]}) \wedge E_{\mathrm{bad},B}|E] \\
&\leq \hat{\delta} \left(\sum_{j=0}^{\min\{r-2,t\}} \binom{t}{j} + \sum_{j=0}^{\min\{r-1,t\}} \binom{t}{j} \right) \\
&\quad \cdot \prod_{i=1}^{k} \mathbb{P}\left[\mathbb{P}[\mathcal{T}_i < \mathcal{T}_{i-1}|E, G_{[i-1]}] \in \iota(p_i) \middle| E, S_{i-1} = w_i, G_{[i-2]} \in N(p_{[i-1]}, w_{[i-1]})\right].
\end{aligned} \tag{58}$$

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 (58) 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 (58) 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 (47) holds.*

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

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

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

Now, let $s_{\max} \in \mathbb{N}_{>0}$. 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 (53)). By Lemma 19 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 (53)), it follows that

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

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

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

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 [24, Section 1.5]), it follows that

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

as desired. □ □

Now, we are prepared to prove Lemma 9.

Lemma 9. 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

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

The result then follows directly from Theorem 5. □ □

Now, we turn our attention to Lemma 10. Let a realization of the random index set $\mathcal{K}_\tau$ defined in (35) be denoted by k_τ. Our next lemma shows an important property about any iteration $k \in [k_{\max}]$ in which $k \in k_\tau$ for a given realization k_τ.

Lemma 20. *For any $k \in [k_{\max}]$ and $g_{[k-1]}$ such that $k \in k_\tau$ for some realization k_τ of the random index set $\mathcal{K}_\tau$, one finds that $\mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}, k \in k_\tau] \geq p_\tau$.*

Proof. In any iteration during which $\tau_k^{\text{trial,true}} < \tau_{k-1}$, it follows that $\tau_k^{\text{trial,true}} < \infty$, so

$$\tau_k^{\text{trial,true}} = \frac{(1-\sigma)\|c_k\|_1}{\nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\}}$$

and thus

$$(1-\sigma)\|c_k\|_1 < (\nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\})\tau_{k-1}.$$

By the definition of τ_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 $\tau_k^{\text{trial,true}} < \tau_{k-1}$, then

$$(1-\sigma)\|c_k\|_1 < (g_k^\top d_k + \max\{d_k^\top H_k d_k, 0\})\tau_{k-1},$$

meaning that $\tau_k < \tau_{k-1}$. Noting that the event $k \in k_\tau$ is conditionally independent of G_k given E and $g_{[k-1]}$, it follows from Assumption 4 that

$$\begin{aligned}
&\mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}, k \in k_\tau] \\
&\geq \mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} | E, g_{[k-1]}, k \in k_\tau] \\
&= \mathbb{P}[G_k^\top D_k + \max\{D_k^\top H_k D_k, 0\} \geq \nabla f(x_k)^\top d_k^{\text{true}} + \max\{(d_k^{\text{true}})^\top H_k d_k^{\text{true}}, 0\} | E, g_{[k-1]}] \geq p_\tau,
\end{aligned}$$

as desired. □ □

The previous lemma guarantees that in any iteration in which $\tau_k^{\text{trial,true}} < \tau_{k-1}$, the probability is at least p_τ that the merit parameter decreases. By scheme for setting τ_k,

$$\mathbb{P}[\tau_k^{\text{trial,true}} < \tau_k | E, g_{[k-1]}, \tau_k^{\text{trial,true}} \geq \tau_{k-1}] = 0, \tag{59}$$

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

Lemma 21. *One finds that*

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

Proof. Consider arbitrary $g_{[k_{\max}]}$ and the corresponding realization k_τ of the random index set $\mathcal{K}_\tau$. By Lemma 20, it follows for all $k \in [k_{\max}]$ that

$$\begin{aligned}
\mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}] &\geq \mathbb{P}[\mathcal{T}_k < \tau_{k-1} \wedge k \in k_\tau | E, g_{[k-1]}] \\
&= \mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}, k \in k_\tau] \cdot \mathbb{P}[k \in k_\tau | E, g_{[k-1]}] \\
&= \mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}, k \in k_\tau] \cdot \mathcal{I}[k \in k_\tau] \\
&\geq \mathcal{I}[k \in k_\tau] p_\tau,
\end{aligned}$$

where $\mathcal{I}[k \in k_\tau]$ is the indicator function for the event $k \in k_\tau$ and the second equality follows due to the fact that the event $k \in k_\tau$ is deterministically known when conditioned on $g_{[k-1]}$. Summing this inequality over $k \in [k_{\max}]$, one finds that

$$\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}] \geq \sum_{k=0}^{k_{\max}} \mathcal{I}[k \in k_\tau] p_\tau = |k_\tau| p_\tau.$$

Letting $f_{G_{[k_{\max}]}}$ denote the probability density function of $G_{[k_{\max}]}$, the fact that the bound above holds *deterministically* for any realization $g_{[k_{\max}]}$ that

$$\begin{aligned}
&\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | E, \mathcal{F}_{k-1}] \geq |\mathcal{K}_\tau| p_\tau \;\middle|\; E \right] \\
&= \int_{g_{[k_{\max}]} \in \mathcal{F}_{k_{\max}+1}} \mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathbb{P}[\mathcal{T}_k < \tau_{k-1} | E, g_{[k-1]}] \geq |k_\tau| p_\tau \;\middle|\; E, g_{[k_{\max}]} \right] f_{G_{[k_{\max}]}}(g_{[k-1]}) \mathrm{d} g_{[k_{\max}]} \\
&= \int_{g_{[k_{\max}]} \in \mathcal{F}_{k_{\max}+1}} 1 \cdot f_{G_{[k_{\max}]}}(g_{[k-1]}) \mathrm{d} g_{[k_{\max}]} = 1.
\end{aligned}$$

Therefore, the desired result holds as well. □ □

We now claim that Lemma 10 follows.

Lemma 10. The proof follows by combining Theorem 5 and Lemma 21. □ □

We conclude this appendix by showing that the order notation result in (7a) and (7b) holds, as required in the proof of Corollary 1.

Lemma 22. *Let $\delta \in (0,1)$, $\hat{\delta}$ be defined in (32), $s_{\max} \in \mathbb{N}_{>0}$ 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 $s_{\max} \in \mathbb{N}_{>0}$, it follows

$$\begin{aligned}
\sum_{j=0}^{\max\{s_{\max}-1,0\}} \binom{k_{\max}}{j} &= \sum_{j=0}^{s_{\max}-1} \frac{(k_{\max})!}{j!(k_{\max}-j)!} \\
&\leq \sum_{j=0}^{s_{\max}-1} \frac{(k_{\max})!}{(k_{\max}-j)!} \\
&= 1 + \sum_{j=1}^{s_{\max}-1} \prod_{i=k_{\max}+1-j}^{k_{\max}} i \\
&\leq 1 + \sum_{j=1}^{s_{\max}-1} (k_{\max})^j \\
&\leq 1 + (s_{\max}-2)(k_{\max})^{s_{\max}-1} \\
&\leq (s_{\max}-1)(k_{\max})^{s_{\max}-1}.
\end{aligned}$$

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

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

as desired. □ □