Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear-Equality-Constrained Optimization with Rank-Deficient Jacobians

Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, Daniel P. Robinson

To cite this article:
Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, Daniel P. Robinson (2023) A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear-Equality-Constrained Optimization with Rank-Deficient Jacobians. Mathematics of Operations Research

Published online in Articles in Advance 30 Oct 2023

. https://doi.org/10.1287/moor.2021.0154

informs.
https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–37
ISSN 0364-765X (print), ISSN 1526-5471 (online)

A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear-Equality-Constrained Optimization with Rank-Deficient Jacobians

Albert S. Berahas,[a] Frank E. Curtis,[b,*] Michael J. O'Neill,[c] Daniel P. Robinson[b]

[a] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109; [b] Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania 18015; [c] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

*Corresponding author

Contact: albertberahas@gmail.com, https://orcid.org/0000-0002-2371-9398 (ASB); frank.e.curtis@lehigh.edu, https://orcid.org/0000-0001-7214-9187 (FEC); mikeoneill@unc.edu (MJO'N); daniel.p.robinson@lehigh.edu, https://orcid.org/0000-0003-0251-4227 (DPR)

Received: June 23, 2021
Revised: July 8, 2022; June 12, 2023
Accepted: September 19, 2023
Published Online in Articles in Advance: October 30, 2023

MSC2020 Subject Classifications: 49M05, 49M10, 49M37, 65K05, 65K10, 90C15, 90C30, 90C55

https://doi.org/10.1287/moor.2021.0154

Abstract. A sequential quadratic optimization algorithm is proposed for solving smooth nonlinear-equality-constrained optimization problems in which the objective function is defined by an expectation. The algorithmic structure of the proposed method is based on a step decomposition strategy that is known in the literature to be widely effective in practice, wherein each search direction is computed as the sum of a normal step (toward linearized feasibility) and a tangential step (toward objective decrease in the null space of the constraint Jacobian). However, the proposed method is unique from others in the literature in that it both allows the use of stochastic objective gradient estimates and possesses convergence guarantees even in the setting in which the constraint Jacobians may be rank-deficient. The results of numerical experiments demonstrate that the algorithm offers superior performance when compared with popular alternatives.

Funding: This material is based upon work supported by the U.S. National Science Foundation's Division of Computing and Communication Foundations under award [CF-1740796], by the Office of Naval Research under award [N00014-21-1-2532], and by the National Science Foundation under award [2030859] to the Computing Research Association for the CIFellows Project.

Keywords: nonlinear optimization • stochastic optimization • stochastic average approximation • sequential quadratic optimization • rank deficiency

1. Introduction

We propose an algorithm for solving equality-constrained optimization problems in which the objective function is defined by an expectation. Formulations of this type arise throughout science and engineering in important applications such as data-fitting problems, where one aims to determine a model that minimizes the discrepancy between values yielded by the model and corresponding known outputs. An example application area that motivates the particular features of the proposed algorithm is physics-informed machine learning; see Section 5.

Our algorithm is designed for solving problems when the decision variables are restricted to the solution set of a (potentially nonlinear) set of equations. We are particularly interested in such problems when the constraint Jacobian—that is, a matrix of first-order derivatives of the constraint function—may be (nearly) rank-deficient in some or even all iterations during the run of an algorithm, because this can be an unavoidable occurrence in practice that would ruin the convergence properties of any algorithm that is not specifically designed for this setting. The structure of our algorithm follows a step decomposition strategy that is common in the constrained-optimization literature; in particular, our algorithm has roots in the Byrd-Omojokun approach (Omojokun [26]). However, our algorithm is unique from previously proposed algorithms in that it offers convergence guarantees while allowing for the use of stochastic objective gradient information in each iteration. We prove that our algorithm converges to stationarity (in expectation) both in nice cases when the constraints are feasible and convergence to the feasible region can be guaranteed (in expectation) and in more challenging cases such as when the constraints are infeasible and one can guarantee convergence only to an infeasible stationary point. To the best of our knowledge, there exist no other algorithms in the literature that have been designed specifically for this setting, namely, stochastic optimization with equality constraints that may exhibit rank deficiency.

The step decomposition strategy employed by our algorithm makes it similar to the method proposed in Curtis et al. [6], although that method is designed for deterministic optimization only and employs a line search, whereas our approach is designed for stochastic optimization and requires no line searches. Our algorithm builds upon the method for solving equality-constrained optimization problems proposed in Berahas et al. [1]. The method proposed in that article assumes that the singular values of the constraint Jacobians are bounded below by a positive constant throughout the optimization process, which implies that the linear independence constraint qualification (LICQ) holds at all iterates. By contrast, the algorithm proposed in this paper makes no such assumption. Handling the potential lack of full-rank Jacobians necessitates a different algorithmic structure and a distinct approach to proving convergence guarantees; for example, one needs to account for the fact that primal-dual stationarity conditions may not be necessary and/or that the constraints may be infeasible.

Similar to the context in Berahas et al. [1], our algorithm is intended for the highly stochastic regime in which the stochastic gradient estimates might only be unbiased estimators of the gradients of the objective at the algorithm iterates that satisfy a loose variance condition. Indeed, we show that in nice cases—in particular, when the adaptive merit parameter employed in our algorithm eventually settles at a value that is sufficiently small—our algorithm has convergence properties that match those of the algorithm in Berahas et al. [1]. These results parallel those for the stochastic gradient method in the context of unconstrained optimization (Bottou et al. [2], Robbins and Monro [30], Robbins and Siegmund [31]). However, for cases not considered in Berahas et al. [1] when the merit parameter may vanish, we require the stronger assumption that the differences between the stochastic gradient estimators and the corresponding true gradients are bounded. This is appropriate because in such a scenario the algorithm aims to transition from a *stochastic* algorithm for solving a constrained optimization problem to one that offers guarantees on par with a *deterministic* algorithm for minimizing constraint violation. Finally, we show under reasonable assumptions that the probability is zero that the merit parameter settles at too large of a value.

Our algorithm has some similarities but many differences with another recently proposed algorithm, namely, that in Na et al. [22]. That stochastic algorithm is also designed for equality-constrained optimization, but (*i*) like for the algorithm in Berahas et al. [1], for the algorithm in Na et al. [22] the constraint Jacobians are required to have singular values that are bounded below by a positive constant (meaning that the LICQ holds at all algorithm iterates), and (*ii*) the algorithm in Na et al. [22] employs an adaptive line search that may require the algorithm to compute relatively accurate stochastic gradient estimates throughout the optimization process. Our algorithm, on the other hand, does not require the LICQ to hold and is meant for a more stochastic regime, meaning that it does not require a procedure for refining the stochastic gradient estimate within an iteration. Consequently, the convergence guarantees that can be proved for our method and the expectations that one should have about the practical performance of our method are quite distinct from those for the algorithm in Na et al. [22]. (See also Fang et al. [11], which also proposes an algorithm for the highly stochastic regime but again requires the constraint Jacobians to have singular values that are bounded below by a positive constant.)

Besides the methods in Berahas et al. [1] and Na et al. [22], there have been few proposed algorithms that might be used to solve problem of the form (1). Some methods have been proposed that employ stochastic (proximal) gradient strategies applied to minimizing penalty functions derived from constrained problems (Chen et al. [4], Kumar Roy et al. [18], Nandwani et al. [23]), but these do not offer convergence guarantees to stationarity with respect to the original constrained problem. On the other hand, stochastic Frank-Wolfe methods have been proposed (Hazan and Luo [16], Locatello et al. [19], Lu and Freund [20], Ravi et al. [28], Reddi et al. [29], Zhang et al. [35]), but these can be applied only in the context of convex feasible regions. Our algorithm, by contrast, is designed for nonlinear-equality-constrained optimization.

1.1. Notation

The set of real numbers is denoted as $\mathbb{R}$, the set of real numbers greater than (respectively, greater than or equal to) $r \in \mathbb{R}$ is denoted as $\mathbb{R}_{>r}$ (respectively, $\mathbb{R}_{\geq r}$), the set of n-dimensional real vectors is denoted as $\mathbb{R}^n$, the set of m-by-n-dimensional real matrices is denoted as $\mathbb{R}^{m\times n}$, and the set of n-by-n-dimensional real symmetric matrices is denoted as $\mathbb{S}^n$. Given $J \in \mathbb{R}^{m\times n}$, the range space of J^T is denoted as $\text{Range}(J^T)$, and the null space of J is denoted as $\text{Null}(J)$. (By the Fundamental Theorem of Linear Algebra, for any $J \in \mathbb{R}^{m\times n}$, the spaces $\text{Range}(J^T)$ and $\text{Null}(J)$ are orthogonal and $\text{Range}(J^T) + \text{Null}(J) = \mathbb{R}^n$, where in this instance "+" denotes the Minkowski sum operator.) The set of nonnegative integers is denoted as $\mathbb{N} := \{0, 1, 2, \ldots\}$. For any $m \in \mathbb{N}$, let $[m]$ denote the set of integers $\{0, 1, \ldots, m\}$. Correspondingly, to represent a set of vectors $\{v_0, \ldots, v_k\}$, we define $v_{[k]} := \{v_0, \ldots, v_k\}$.

The algorithm that we propose is iterative in the sense that, given a starting point $x_0 \in \mathbb{R}^n$, any run generates a sequence of iterates $\{x_k\}$ with $x_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$, which is itself a realization of a stochastic process $\{X_k\}$ with $X_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$. (We state the algorithm in terms of a particular realization of it, although our analysis considers the stochastic process generated by the algorithm, which we formalize at the beginning of Section 4.) For simplicity of notation, the iteration number is appended as a subscript to other quantities corresponding to each iteration; for example, with a function $c : \mathbb{R}^n \to \mathbb{R}$, its value at x_k is denoted as $c_k := c(x_k)$ for all $k \in \mathbb{N}$.

1.2. Organization

Our problem of interest and basic assumptions about the problem and the behavior of our algorithm are presented in Section 2. Our algorithm is motivated and presented in Section 3. Convergence guarantees for our algorithm are presented in Section 4. The results of numerical experiments are provided in Section 5, and concluding remarks are provided in Section 6.

2. Problem Statement

Our algorithm is designed for solving (potentially nonlinear and/or nonconvex) equality-constrained optimization problems of the form

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

where the functions $f:\mathbb{R}^n\to\mathbb{R}$ and $c:\mathbb{R}^n\to\mathbb{R}^m$ are smooth, ι is a random variable with associated probability space $(\Omega,\mathcal{F},\mathbb{P}_\iota)$, $F:\mathbb{R}^n\times\Omega\to\mathbb{R}$, and $\mathbb{E}_\iota[\cdot]$ denotes expectation taken with respect to $\mathbb{P}_\iota$. We assume that values and first-order derivatives of the constraint functions can be computed but that the objective and its associated first-order derivatives are intractable to compute, and one must instead employ stochastic estimates. (We formalize our assumptions about such stochastic estimates starting with Assumption 2.) Formally, we make the following assumption with respect to (1) and the stochastic process generated by our proposed algorithm. (For further generality, one could relax the following assumption to say that the stated properties hold almost surely, that is, with probability one. However, such a relaxation would require constant reference to probability-one events throughout our analysis without adding substantially to the strength of our results, so for the sake of brevity we do not bother with such generality.)

Assumption 1. *Let $\mathcal{R}\subseteq\mathbb{R}^n$ be an open convex set that contains the iterate sequence $\{X_k\}$. The objective function $f:\mathbb{R}^n\to\mathbb{R}$ is continuously differentiable and bounded over $\mathcal{R}$, and its gradient function $\nabla f:\mathbb{R}^n\to\mathbb{R}^n$ is Lipschitz continuous with constant $L\in\mathbb{R}_{>0}$ (with respect to $\|\cdot\|_2$) and bounded over $\mathcal{R}$. The constraint function $c:\mathbb{R}^n\to\mathbb{R}^m$ (with $m\le n$) is continuously differentiable and bounded over $\mathcal{R}$, and its Jacobian function $J:=\nabla c^T:\mathbb{R}^n\to\mathbb{R}^{m\times n}$ is Lipschitz continuous with constant $\Gamma\in\mathbb{R}_{>0}$ (with respect to $\|\cdot\|_2$) and bounded over $\mathcal{R}$.*

The aspects of Assumption 1 that pertain to the objective function f and constraint function c are typical for the equality-constrained-optimization literature. Notice that we do not assume that the iterate sequence itself is bounded. Under Assumption 1, it follows that there exist positive real numbers $(f_{\inf},f_{\sup},\kappa_{\nabla f},\kappa_c,\kappa_J)\in\mathbb{R}_{>0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ such that

$$f_{\inf}\le f(x)\le f_{\sup},\ \|\nabla f(x)\|_2\le\kappa_{\nabla f},\ \|c(x)\|_2\le\kappa_c,\ \text{and}\ \ \|J(x)\|_2\le\kappa_J \ \text{ for all } \ x\in\mathcal{R}. \tag{2}$$

Given that our proposed algorithm is stochastic, it is admittedly not ideal to have to assume that the objective value, objective gradient, constraint value, and constraint Jacobian are bounded over the set $\mathcal{R}$ containing the iterates. This is a common assumption in the deterministic optimization literature, where it may be justified in the context of an algorithm that is guaranteed to make progress in each iteration, say, with respect to a merit function. However, for a stochastic algorithm such as ours, such a claim may be seen as less than ideal because a stochastic algorithm may be guaranteed to make progress only in expectation in each iteration, meaning that it is possible for the iterates to drift far from desirable regions of the search space during the optimization process.

Our justification for Assumption 1 is twofold. First, any reader who is familiar with analyses of stochastic algorithms for unconstrained optimization—in particular, those analyses that do not require that the objective gradient is bounded over a set containing the iterates—should appreciate that additional challenges present themselves in the context of constrained optimization. For example, whereas in unconstrained optimization one naturally considers the objective f as a measure of progress, in (nonconvex) constrained optimization one needs to employ a merit function for measuring progress, and for practical purposes such a function typically needs to involve a parameter (or parameters) that must be adjusted dynamically by the algorithm. One finds that it is the adaptivity of our merit parameter (see (9) later on) that necessitates the aforementioned boundedness assumptions that we use in our analysis. (Certain exact merit functions, such as that employed in Na et al. [22], might not lead to the same issues as the merit function that we employ. However, we remark that the merit function employed in Na et al. [22] is not a viable option unless the constraint Jacobians have singular values that are bounded below by a positive constant throughout any run of the algorithm.) Our second justification is that we know of no other algorithm that offers convergence guarantees that are as comprehensive as ours (in terms of handling feasible, degenerate, and infeasible settings) under an assumption that is at least as loose as Assumption 1.

Let the Lagrangian $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ corresponding to (1) be given by $\ell(x,y) = f(x) + c(x)^T y$, where $y \in \mathbb{R}^m$ represents a vector of Lagrange multipliers. Under a constraint qualification (such as the LICQ), necessary conditions for first-order stationarity with respect to (1) are given by

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

see, for example, Nocedal and Wright [25]. However, under only Assumption 1, it is possible for (1) to be degenerate, in which case (3) might not be necessary at a solution of (1), or (1) may be infeasible. In the latter case, one aims to design an algorithm that transitions automatically from seeking stationarity with respect to (1) to seeking stationarity with respect to a measure of infeasibility of the constraints. For our purposes, we employ the infeasibility measure $\varphi : \mathbb{R}^n \to \mathbb{R}$ defined by $\varphi(x) = \|c(x)\|_2$. A point $x \in \mathbb{R}^n$ is stationary with respect to φ if and only if either $c(x) = 0$ or both $c(x) \neq 0$ and

$$0 = \nabla \varphi(x) = \frac{J(x)^T c(x)}{\|c(x)\|_2}. \tag{4}$$

3. Algorithm Description

Our algorithm can be characterized as a sequential quadratic optimization (commonly known as SQP) method that employs a step decomposition strategy and chooses step sizes that attempt to ensure sufficient decrease in a merit function in each iteration. We present our complete algorithm in this section, which builds upon this basic characterization to involve various unique aspects that are designed for handling the combination of (i) stochastic gradient estimates and (ii) potential rank deficiency of the constraint Jacobians.

In each iteration $k \in \mathbb{N}$, the algorithm first computes the *normal component* of the search direction toward reducing linearized constraint violation. Conditioned on the event that x_k is reached as the kth iterate, the problem defining this computation, namely,

$$\min_{v \in \mathbb{R}^n} \frac{1}{2}\|c_k + J_k v\|_2^2 \text{ s.t. } \|v\|_2 \leq \omega \|J_k^T c_k\|_2 \tag{5}$$

where $\omega \in \mathbb{R}_{>0}$ is a user-defined parameter, is determined because the constraint function value c_k and constraint Jacobian J_k are determined by x_k. If J_k has full row rank, ω is sufficiently large, and (5) is solved to optimality, then one obtains v_k such that $c_k + J_k v_k = 0$. However, an exact solution of (5) may be expensive to obtain, and—as has been shown for various step decomposition strategies, such as the Byrd-Omojokun approach (Omojokun [26])—the consideration of (5) is viable when J_k might not have full row rank. Fortunately, our algorithm merely requires that the normal component $v_k \in \mathbb{R}^n$ is feasible for problem (5), lies in Range(J_k^T), and satisfies the Cauchy decrease condition

$$\|c_k\|_2 - \|c_k + J_k v_k\|_2 \geq \epsilon_v (\|c_k\|_2 - \|c_k + \alpha_k^C J_k v_k^C\|_2) \tag{6}$$

for some user-defined parameter $\epsilon_v \in (0,1]$. Here, $v_k^C := -J_k^T c_k$ is the steepest descent direction for the objective of problem (5) at $v = 0$, and the step size $\alpha_k^C \in \mathbb{R}$ is the unique solution to the problem to minimize $\frac{1}{2}\|c_k + \alpha^C J_k v_k^C\|_2^2$ over $\alpha^C \in \mathbb{R}_{\geq 0}$ subject to $\alpha^C \leq \omega$ (see, e.g., Nocedal and Wright [25], equations (4.11)–(4.12)). Because this allows one to choose $v_k \leftarrow v_k^C$, the normal component can be computed at low computational cost. For a more accurate solution to (5), one can employ a so-called matrix-free iterative algorithm such as the linear conjugate gradient (CG) method with Steihaug stopping conditions (Steihaug [32]) or GLTR (Gould et al. [13]), each of which is guaranteed to yield a solution satisfying the aforementioned conditions no matter how many iterations (greater than or equal to one) are performed.

After the computation of the normal component, our algorithm computes the *tangential component* of the search direction by minimizing a model of the objective function subject to remaining in the null space of the constraint Jacobian. This ensures that the progress toward linearized feasibility offered by the normal component is not undone by the tangential component when the components are added together. The problem defining the computation of the tangential component is

$$\min_{u \in \mathbb{R}^n} (g_k + H_k v_k)^T u + \frac{1}{2} u^T H_k u \quad \text{s.t.} \quad J_k u = 0, \tag{7}$$

where $g_k \in \mathbb{R}^n$ is a stochastic gradient estimate and $H_k \in \mathbb{S}^n$ yields $u^T H_k u > 0$ for all nonzero $u \in \text{Null}(J_k)$. (Specific additional requirements for $\{g_k\}$ and $\{H_k\}$ that are needed for our convergence guarantees are stated formally throughout our analysis in Section 4.)

Because $H_k \in \mathbb{S}^n$ yields $u^T H_k u > 0$ for all nonzero $u \in \text{Null}(J_k)$, the tangential component u_k that is defined as the unique solution of (7) can be obtained by solving

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k \\ y_k \end{bmatrix} = -\begin{bmatrix} g_k + H_k v_k \\ 0 \end{bmatrix}. \tag{8}$$

Even if the constraint Jacobian J_k does not have full row rank, the linear system (8) is consistent because it represents sufficient optimality conditions of the linearly constrained quadratic optimization problem in (7). (Factorization methods that are popular in the context of solving symmetric indefinite linear systems of equations, such as the Bunch-Kaufman factorization, can fail when the matrix in (8) is singular. However, Krylov subspace methods provide a viable alternative, because for such methods singularity is benign as long as the system is known to be consistent, as is the case for (8).) Under the aforementioned conditions on H_k, the solution component u_k is unique, although y_k might not be unique (if J_k does not have full row rank).

Upon computation of the search direction, our algorithm proceeds toward determining a positive step size. For this purpose, we employ the merit function $\phi : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ defined by

$$\phi(x, \tau) = \tau f(x) + \|c(x)\|_2, \tag{9}$$

where τ is a merit parameter whose value is set dynamically. The function ϕ is an exact penalty function that is common in the literature (Han [14], Han and Mangasarian [15], Powell [27]). For setting the merit parameter value in each iteration, we employ a local model of ϕ denoted as $l : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and defined by

$$l(x, \tau, g, d) = \tau(f(x) + g^T d) + \|c(x) + J(x)d\|_2.$$

Given the search direction vectors v_k, u_k, and $d_k \leftarrow v_k + u_k$, the algorithm sets

$$\tau_k^{\text{trial}} \leftarrow \begin{cases} \infty & \text{if } g_k^T d_k + u_k^T H_k u_k \leq 0 \\ \dfrac{(1-\sigma)(\|c_k\|_2 - \|c_k + J_k d_k\|_2)}{g_k^T d_k + u_k^T H_k u_k} & \text{otherwise,} \end{cases} \tag{10}$$

where $\sigma \in (0,1)$ is user-defined. The merit parameter value is then set as

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

where $\epsilon_\tau \in (0,1)$ is user-defined. This rule ensures that $\{\tau_k\}$ is monotonically nonincreasing, $\tau_k \leq \tau_k^{\text{trial}}$ for all $k \in \mathbb{N}$, and, with the reduction function $\Delta l : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\Delta l(x, \tau, g, d) = l(x, \tau, g, 0) - l(x, \tau, g, d) = -\tau g^T d + \|c(x)\|_2 - \|c(x) + J(x)d\|_2 \tag{12}$$

and the aforementioned conditions on H_k, it ensures the following critical fact:

$$\Delta l(x_k, \tau_k, g_k, d_k) \geq \tau_k u_k^T H_k u_k + \sigma(\|c_k\|_2 - \|c_k + J_k v_k\|_2). \tag{13}$$

Similar to the algorithm in Berahas et al. [1], our algorithm also adaptively sets other parameters that are used for determining an allowable range for the step size in each iteration. (There exist constants that, if known in advance, could be used by the algorithm for determining the allowable range for each step size; see Lemma 2 in our analysis later on. However, to avoid the need to know these problem-dependent constants in advance, our algorithm generates these parameter sequences adaptively, which our analysis shows is sufficient to ensure convergence guarantees.) For distinguishing between search directions that are dominated by the tangential component and others that are dominated by the normal component, the algorithm adaptively defines sequences $\{\chi_k\}$ and $\{\zeta_k\}$. (These sequences were not present in the algorithm in Berahas et al. [1]; they are newly introduced for the needs of our proposed algorithm.) In particular, in iteration $k \in \mathbb{N}$, the algorithm employs the conditions

$$\|u_k\|_2^2 \geq \chi_{k-1}\|v_k\|_2^2 \text{ and } \frac{1}{2} d_k^T H_k d_k < \frac{1}{4}\zeta_{k-1}\|u_k\|_2^2 \tag{14}$$

in order to set

$$(\chi_k, \zeta_k) \leftarrow \begin{cases} ((1+\epsilon_\chi)\chi_{k-1}, (1-\epsilon_\zeta)\zeta_{k-1}) & \text{if (14) holds} \\ (\chi_{k-1}, \zeta_{k-1}) & \text{otherwise,} \end{cases} \tag{15}$$

where $\epsilon_\chi \in \mathbb{R}_{>0}$ and $\epsilon_\zeta \in (0,1)$ are user-defined. It follows from (15) that $\{\chi_k\}$ is monotonically nondecreasing and

$\{\zeta_k\}$ is monotonically nonincreasing. It will be shown in our analysis that $\{\chi_k\}$ is always bounded above uniformly by a positive real number and $\{\zeta_k\}$ is always bounded below uniformly by a positive real number. This means that despite the stochasticity of the algorithm iterates, these sequences have $(\chi_k, \zeta_k) = (\chi_{k-1}, \zeta_{k-1})$ for all sufficiently large $k \in \mathbb{N}$.

Whether $\|u_k\|_2^2 \geq \chi_k \|v_k\|_2^2$ (i.e., the search direction is *tangentially dominated*) or $\|u_k\|_2^2 < \chi_k \|v_k\|_2^2$ (i.e., the search direction is *normally dominated*) influences two aspects of iteration $k \in \mathbb{N}$. First, it influences a value that the algorithm employs to determine the range of allowable step sizes that represents a lower bound for the ratio between the reduction in the model l of the merit function and a quantity involving the squared norm of the search direction. (A similar, but slightly different sequence was employed for the algorithm in Berahas et al. [1].) In iteration $k \in \mathbb{N}$ of our algorithm, the estimated lower bound is set adaptively by first setting

$$\xi_k^{\text{trial}} \leftarrow \begin{cases} \dfrac{\Delta l(x_k, \tau_k, g_k, d_k)}{\tau_k \|d_k\|_2^2} & \text{if } \|u_k\|_2^2 \geq \chi_k \|v_k\|_2^2 \\ \dfrac{\Delta l(x_k, \tau_k, g_k, d_k)}{\|d_k\|_2^2} & \text{otherwise,} \end{cases} \tag{16}$$

then setting

$$\xi_k \leftarrow \begin{cases} \xi_{k-1} & \text{if } \xi_{k-1} \leq \xi_k^{\text{trial}} \\ \min\{(1-\epsilon_\xi)\xi_{k-1}, \xi_k^{\text{trial}}\} & \text{otherwise,} \end{cases} \tag{17}$$

for some user-defined $\epsilon_\xi \in (0,1)$. The procedure in (17) ensures that $\{\xi_k\}$ is monotonically nonincreasing and $\xi_k \leq \xi_k^{\text{trial}}$ for all $k \in \mathbb{N}$. It will be shown in our analysis that $\{\xi_k\}$ is always bounded below uniformly by a positive real number, even though in each iteration it depends on stochastic quantities. (Like for $\{\chi_k\}$ and $\{\zeta_k\}$, there exists a constant that, if known in advance, could be used in place of ξ_k for all $k \in \mathbb{N}$—see Lemma 3—but for ease of employment our algorithm generates $\{\xi_k\}$.) To achieve this property, it is critical that the denominator in (16) is different, depending on whether the search direction is tangentially or normally dominated; see Lemma 3. The second aspect of the algorithm that is affected by whether a search direction is tangentially or normally dominated is a rule for setting the step size; this will be seen in (21) later on.

We are now prepared to present the mechanism by which a positive step size is selected in each iteration $k \in \mathbb{N}$ of our algorithm. We present a strategy that allows for our convergence analysis in Section 4 to be as straightforward as possible. In Section 5, we remark on extensions of this strategy that are included in our software implementation for which our convergence guarantees also hold (as long as some additional cases are considered in one key lemma).

We motivate our strategy by considering an upper bound for the change in the merit function corresponding to the computed search direction, namely, $d_k \leftarrow v_k + u_k$. In particular, under Assumption 1, in iteration $k \in \mathbb{N}$, one has for any nonnegative step size $\alpha \in \mathbb{R}_{\geq 0}$ that

$$\begin{aligned} &\phi(x_k + \alpha d_k, \tau_k) - \phi(x_k, \tau_k) \\ &= \tau_k f(x_k + \alpha d_k) - \tau_k f(x_k) + \|c(x_k + \alpha d_k)\|_2 - \|c_k\|_2 \\ &\leq \alpha \tau_k \nabla f(x_k)^T d_k + \|c_k + \alpha J_k d_k\|_2 - \|c_k\|_2 + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|_2^2 \\ &\leq \alpha \tau_k \nabla f(x_k)^T d_k + |1 - \alpha| \|c_k\|_2 - \|c_k\|_2 + \alpha \|c_k + J_k d_k\|_2 + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|_2^2. \end{aligned} \tag{18}$$

This upper bound is a convex, piecewise quadratic function in α. In a deterministic algorithm in which the gradient $\nabla f(x_k)$ is available, it is common to require that the step size α yields

$$\phi(x_k + \alpha d_k, \tau_k) - \phi(x_k, \tau_k) \leq -\eta \alpha \Delta l(x_k, \tau_k, \nabla f(x_k), d_k), \tag{19}$$

where $\eta \in (0,1)$ is user-defined. However, in our setting, (19) cannot be enforced because our algorithm avoids the evaluation of $\nabla f(x_k)$ and in lieu of it only computes a stochastic gradient g_k. The first main idea of our step size strategy is to determine a step size such that the upper bound in (18) is less than or equal to the right-hand side of (19) when the true gradient $\nabla f(x_k)$ is replaced by its estimate g_k. Because (13), the orthogonality of $v_k \in \text{Range}(J_k^T)$ and $u_k \in \text{Null}(J_k)$, and the properties of the normal step (which, as shown in Lemma 1 later on, include that the left-hand side of (6) is positive whenever $v_k \neq 0$) ensure that $\Delta l(x_k, \tau_k, g_k, d_k) > 0$ whenever $d_k \neq 0$, it follows that a step size satisfying this aforementioned property is given, for any $\beta_k \in (0,1]$, by

$$\alpha_k^{\text{suff}} \leftarrow \min\left\{\frac{2(1-\eta)\beta_k \Delta l(x_k,\tau_k,g_k,d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2}, 1\right\} \in \mathbb{R}_{>0}. \tag{20}$$

The sequence $\{\beta_k\}$ referenced in (20) is chosen with different properties—namely, constant or diminishing—depending on the desired type of convergence guarantee. We discuss details of the possible choices for $\{\beta_k\}$ and the consequences of these choices along with our convergence analysis.

Given that the step size α_k^{suff} in (20) has been set based on a stochastic gradient estimate, a safeguard is needed for our convergence guarantees. For this purpose, the second main idea of our step size selection strategy is to project the trial step size onto an interval that is appropriate, depending on whether the search direction is tangentially dominated or normally dominated. In particular, the step size is chosen as $\alpha_k \leftarrow \text{Proj}_k(\alpha_k^{\text{suff}})$, where

$$\text{Proj}_k(\cdot) := \begin{cases} \text{Proj}\left(\cdot \,\middle|\, \left[\dfrac{2(1-\eta)\beta_k\xi_k\tau_k}{\tau_k L + \Gamma}, \dfrac{2(1-\eta)\beta_k\xi_k\tau_k}{\tau_k L + \Gamma} + \theta\beta_k^2\right]\right) & \text{if } \|u_k\|_2^2 \geq \chi_k\|v_k\|_2^2 \\ \text{Proj}\left(\cdot \,\middle|\, \left[\dfrac{2(1-\eta)\beta_k\xi_k}{\tau_k L + \Gamma}, \dfrac{2(1-\eta)\beta_k\xi_k}{\tau_k L + \Gamma} + \theta\beta_k^2\right]\right) & \text{otherwise.} \end{cases} \tag{21}$$

Here, $\text{Proj}(\cdot|\mathcal{I})$ denotes the projection onto the interval $\mathcal{I} \subset \mathbb{R}$. In our analysis, the rules for $\{\beta_k\}$ (see Lemma 6) ensure that this projection only ever decreases the step size; hence, the overall motivation for the projection is to ensure that the step size is not too large compared with a conservative choice, namely, the lower end of the projection interval. Motivation for the difference in the interval, depending on whether the search direction is tangentially or normally dominated, can be seen Lemma 12 later on, where it is critical that the step size for a normally dominated search direction does not necessarily vanish if/when the merit parameter vanishes, that is, $\{\tau_k\} \searrow 0$.

Overall, our step size selection mechanism can be understood as follows. First, the algorithm adaptively sets the sequences $\{\chi_k\}$, $\{\zeta_k\}$, and $\{\xi_k\}$ in order to estimate bounds that are needed for the step size selection and are known to exist theoretically but cannot be computed directly. By the manner in which these sequences are set, our analysis shows that they remain constant for sufficiently large $k \in \mathbb{N}$. With these values, our step size selection strategy aims to achieve a reduction in the merit function in expectation with safeguards because the computed values are based on stochastic quantities. One finds by the definition of the projection interval in (21) that the step size for a *tangentially dominated search direction* may decrease to zero if $\{\tau_k\} \searrow 0$; this is needed in cases when the problem is degenerate or infeasible and the algorithm wants to avoid long steps in the tangential component that may ruin progress toward minimizing constraint violation. Otherwise, for a *normally dominated search direction*, the step size would remain bounded away from zero if $\beta_k = \beta \in (0,1]$ for all $k \in \mathbb{N}$; that is, it can only decrease to zero if $\{\beta_k\}$ is diminishing. If our algorithm did not make this distinction between the projection intervals for tangentially versus normally dominated search directions, then the algorithm would fail to have desirable convergence guarantees even in the deterministic setting. (In particular, our proof in Appendix A of Theorem 1, which is upcoming in Section 4, would break down.)

Our complete algorithm is stated as Algorithm 1. For the computation of the stochastic gradient estimate g_k and matrix H_k in each iteration, the algorithm statement refers to Assumptions 2 and 3, which are introduced formally for our analysis in the subsequent section.

Algorithm 1 (Stochastic SQP Algorithm)

```
Require $L \in \mathbb{R}_{>0}$, a Lipschitz constant for $\nabla f$; $\Gamma \in \mathbb{R}_{>0}$, a Lipschitz constant for $c$; $\{\beta_k\} \subset (0,1]$; $x_0 \in \mathbb{R}^n$; $\tau_{-1} \in \mathbb{R}_{>0}$; $\chi_{-1} \in \mathbb{R}_{>0}$; $\zeta_{-1} \in \mathbb{R}_{>0}$; $\xi_{-1} \in \mathbb{R}_{>0}$; $\omega \in \mathbb{R}_{>0}$; $\epsilon_v \in (0,1]$; $\sigma \in (0,1)$; $\epsilon_\tau \in (0,1)$; $\epsilon_\chi \in \mathbb{R}_{>0}$; $\epsilon_\zeta \in (0,1)$; $\epsilon_\xi \in (0,1)$; $\eta \in (0,1)$; $\theta \in \mathbb{R}_{\geq 0}$
for $k \in \mathbb{N}$ do
    if $\|J_k^T c_k\|_2 = 0$ and $\|c_k\|_2 > 0$ then
        terminate and return $x_k$ (infeasible stationary point)
    end if
    Compute a stochastic gradient $g_k$ satisfying Assumption 2
    Compute $v_k \in \text{Range}(J_k^T)$ that is feasible for problem (5) and satisfies (6)
    Compute $H_k$ satisfying Assumption 3
    Compute $(u_k, y_k)$ as a solution of (8), and then set $d_k \leftarrow v_k + u_k$
    if $d_k = 0$ then
```

```
        Set $\tau_k^{\rm trial} \leftarrow \infty$ and $\tau_k \leftarrow \tau_{k-1}$
        Set $(\chi_k, \zeta_k) \leftarrow (\chi_{k-1}, \zeta_{k-1})$
        Set $\xi_k^{\rm trial} \leftarrow \infty$ and $\xi_k \leftarrow \xi_{k-1}$
        Set $\alpha_k^{\rm suff} \leftarrow 1$ and $\alpha_k \leftarrow 1$
    else
        Set $\tau_k^{\rm trial}$ by (10) and $\tau_k$ by (11)
        Set $(\chi_k, \zeta_k)$ by (14)–(15)
        Set $\xi_k^{\rm trial}$ by (16) and $\xi_k$ by (17)
        Set $\alpha_k^{\rm suff}$ by (20) and $\alpha_k \leftarrow {\rm Proj}_k(\alpha_k^{\rm suff})$ using (21)
    end if
    Set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
end for
```

4. Convergence Analysis

In this section, we prove convergence guarantees for Algorithm 1. To understand the results that can be expected given our setting and the type of algorithm that we employ, let us first present a set of guarantees that can be proved if Algorithm 1 were to be run with $g_k = \nabla f(x_k)$ and $\beta_k = \beta$ for all $k \in \mathbb{N}$, where $\beta \in \mathbb{R}_{>0}$ is sufficiently small. For such an algorithm, we prove the following theorem in Appendix A. The theorem is consistent with what can be proved for other deterministic algorithms in our context; for example, see theorem 3.3 in Curtis et al. [6]. Given $J_k \in \mathbb{R}^{m\times n}$, we use Z_k to denote a matrix whose columns form an orthonormal basis for ${\rm Null}(J_k)$.

Theorem 1. *Suppose Algorithm 1 is employed to solve (1) such that Assumption 1 holds, $g_k = \nabla f(x_k)$ for all $k \in \mathbb{N}$, there exists $\kappa_H \in \mathbb{R}_{>0}$ such that $\|H_k\|_2 \le \kappa_H$ for all $k \in \mathbb{N}$, there exists $\rho \in \mathbb{R}_{>0}$ such that $u^T H_k u \ge \rho\|u\|_2^2$ for all $u \in {\rm Null}(J_k)$ for all $k \in \mathbb{N}$, and $\beta_k = \beta$ for all $k \in \mathbb{N}$, where*

$$\beta \in (0,1] \text{ and } \frac{2(1-\eta)\beta\xi_{-1}\max\{\tau_{-1},1\}}{\Gamma} \in (0,1]. \tag{22}$$

If there exists $k_J \in \mathbb{N}$ and $\sigma_J \in \mathbb{R}_{>0}$ such that the singular values of J_k are bounded below by σ_J for all $k \ge k_J$, then the merit parameter sequence $\{\tau_k\}$ is bounded below by a positive real number and

$$0 = \lim_{k\to\infty} \left\| \begin{bmatrix} \nabla f(x_k) + J_k^T y_k \\ c_k \end{bmatrix} \right\|_2 = \lim_{k\to\infty} \left\| \begin{bmatrix} Z_k^T \nabla f(x_k) \\ c_k \end{bmatrix} \right\|_2. \tag{23}$$

Otherwise, if such k_J and σ_J do not exist, then it still follows that

$$0 = \lim_{k\to\infty} \|J_k^T c_k\|_2, \tag{24}$$

and if $\{\tau_k\}$ is bounded below by a positive real number, then

$$0 = \lim_{k\to\infty} \|\nabla f(x_k) + J_k^T y_k\|_2 = \lim_{k\to\infty} \|Z_k^T \nabla f(x_k)\|_2. \tag{25}$$

Based on Theorem 1, the following aims—which are all achieved in certain forms in our analyses in Sections 4.1 and 4.2—can be set for Algorithm 1 in the stochastic setting. First, if Algorithm 1 is run and the singular values of the constraint Jacobians happen to remain bounded away from zero beyond some iteration, then (following (23)) one should aim to prove that a primal-dual stationarity measure (recall (3)) vanishes in expectation. This is shown under certain conditions in Corollary 1 (and the subsequent discussion). Otherwise, a (sub)sequence of $\{J_k\}$ tends to singularity, in which case (following (24)) one should at least aim to prove that $\{\|J_k^T c_k\|_2\}$ vanishes in expectation, which would mean that a (sub)sequence of iterates converges in expectation to feasibility or at least stationarity with respect to the constraint infeasibility measure φ (recall (4)). Such a conclusion is offered under certain conditions by combining Corollary 1 and Theorem 3. The remaining aim (paralleling (25)) is that one should aim to prove that even if a (sub)sequence of $\{J_k\}$ tends to singularity, if the merit parameter sequence $\{\tau_k\}$ happens to remain bounded below by a positive real number, then $\{\|Z_k^T \nabla f(x_k)\|_2\}$ vanishes in expectation. This can also be seen to occur under certain conditions in Corollary 1.

In addition, because of its stochastic nature, there are events that one should consider in which the algorithm may exhibit behavior that cannot be exhibited by the deterministic one. One such event is when the merit parameter eventually remains fixed at a value that is not sufficiently small. We show in Section 4.3—with formal results stated and proved in Appendix C—that, under reasonable assumptions, the probability of this event is zero. We complete the picture of the possible behaviors of our algorithm by discussing remaining possible (practically irrelevant) events in Section 4.4.

Let us now commence our analysis of Algorithm 1. For this analysis, we formalize the quantities defined by Algorithm 1 as a stochastic process. This includes quantities that are actually computed by the algorithm as well as others that, for the purposes of our analysis, we refer to as "true" ones. For all $k \in \mathbb{N}$, a "true" quantity is one that would have been computed if the true gradient were used in place of a stochastic gradient in that iteration. Overall, the stochastic process is

$$\{(X_k, G_k, V_k, H_k, U_k, U_k^{\text{true}}, D_k, D_k^{\text{true}}, Y_k, Y_k^{\text{true}}, \mathcal{T}_k^{\text{trial}}, \mathcal{T}_k, \mathcal{T}_k^{\text{trial, true}}, \mathcal{X}_k, \mathcal{Z}_k, \Xi_k^{\text{trial}}, \Xi_k, \mathcal{A}_k^{\text{suff}}, \mathcal{A}_k)\},$$

where, for all $k \in \mathbb{N}$, we denote the primal iterate as X_k, the stochastic gradient estimator as G_k, the normal search direction as V_k, the quadratic-form matrix in (7) as H_k, the tangential search direction as U_k, the "true" tangential search direction as U_k^{true}, the search direction as D_k, the "true" search direction as D_k^{true}, the Lagrange multiplier estimate as Y_k, the "true" Lagrange multiplier estimate as Y_k^{true}, the trial merit parameter as $\mathcal{T}_k^{\text{trial}}$, the merit parameter as $\mathcal{T}_k$, the "true" trial merit parameter as $\mathcal{T}_k^{\text{trial, true}}$, the curvature parameter as $\mathcal{X}_k$, the curvature threshold parameter as $\mathcal{Z}_k$, the trial ratio parameter as Ξ_k^{trial}, the ratio parameter as Ξ_k, the sufficient step size value as defined in (20) as $\mathcal{A}_k^{\text{suff}}$, and the step size as $\mathcal{A}_k$. A realization of the kth element of this process, namely, $(x_k, g_k, v_k, H_k, u_k, u_k^{\text{true}}, d_k, d_k^{\text{true}}, y_k, y_k^{\text{true}}, \tau_k^{\text{trial}}, \tau_k, \tau_k^{\text{trial,true}}, \chi_k, \zeta_k, \xi_k^{\text{trial}}, \xi_k, \alpha_k^{\text{suff}}, \alpha_k)$, includes the algorithmic quantities that have appeared in the prior section.

We remark that we have introduced an abuse of notation with respect to H_k, which now represents the (stochastic) quadratic-form matrix in (7), whereas previously it was a realization of it. However, this should not lead to confusion, because for our purposes of analysis it can be viewed as an element of the stochastic process, not a realization. Similarly, we continue to use J_k to refer to the Jacobian in the kth iteration as well as Z_k to refer to a matrix whose columns form an orthonormal basis for $\text{Null}(J_k)$; these are also now stochastic quantities whose values/properties are defined by X_k. We use C_k to denote the constraint function value in iteration k.

Algorithm 1's behavior is dictated entirely by the initial conditions (i.e., initial point and parameter values) as well as the stochastic gradient estimators; that is, assuming for simplicity that the initial conditions are predetermined, a realization of $\{G_0, \dots, G_{k-1}\}$ determines a realization of

$$\{X_j\}_{j=1}^{k} \text{ and } \{V_k, H_k, U_k, U_k^{\text{true}}, D_k, D_k^{\text{true}}, Y_k, Y_k^{\text{true}}, \mathcal{T}_k^{\text{trial}}, \mathcal{T}_k, \mathcal{T}_k^{\text{trial, true}}, \mathcal{X}_k, \mathcal{Z}_k, \Xi_k^{\text{trial}}, \Xi_k, \mathcal{A}_k^{\text{suff}}, \mathcal{A}_k)\}_{j=0}^{k-1}.$$

Let $\mathcal{G}_0$ be the σ-algebra defined by the initial conditions, and for all $k \in \mathbb{N}$, let $\mathcal{G}_k$ be the σ-algebra generated by the initial conditions and $\{G_0, \dots, G_{k-1}\}$. Thus, $\{\mathcal{G}_k\}$ is a filtration.

For the initial results in our analysis, make the following basic assumptions, where for the sake of brevity we define $\mathbb{E}_k[\cdot] := \mathbb{E}_\iota[\cdot|\mathcal{G}_k]$. In subsequent subsections, we make similar assumptions conditioned on different scenarios of the behavior of the algorithm. (In certain cases, we impose stronger conditions on $\{G_k\}$, as needed; see Sections 4.2 and 4.3 and Appendix C.) As will be explained, our initial results—based on Assumptions 2 and 3—will carry over to each subsection.

Assumption 2. *For all $k \in \mathbb{N}$, the stochastic gradient estimator $G_k \in \mathbb{R}^n$ is unbiased in the sense that $\mathbb{E}_k[G_k] = \nabla f(X_k)$. In addition, there exists a positive real number $\nu \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, one has $\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2] \leq \nu$.*

Assumption 3. *For all $k \in \mathbb{N}$, the matrix $H_k \in \mathbb{S}^n$ is $\mathcal{G}_k$-measurable. In addition, there exist positive real numbers $\kappa_H \in \mathbb{R}_{>0}$ and $\rho \in \mathbb{R}_{>0}$ such that the sequence $\{H_k\}$ is bounded in norm by κ_H and, for all $k \in \mathbb{N}$, one has $u^T H_k u \geq \rho\|u\|_2^2$ for all $u \in \text{Null}(J_k)$.*

Note that one can generate a realization g_k in iteration $k \in \mathbb{N}$ by independently drawing b_k realizations of the random variable ι, denoting the *mini-batch* as $\mathcal{B}_k := \{\iota_{k,1}, \dots, \iota_{k,b_k}\}$, and setting

$$g_k \leftarrow \frac{1}{b_k} \sum_{\iota \in \mathcal{B}_k} \nabla f(x_k, \iota). \tag{26}$$

It is a modest assumption about the function f and the sample sizes $\{b_k\}$ to say that a realized sequence $\{g_k\}$ generated in this manner is a realization of $\{G_k\}$ satisfying Assumption 2. As for Assumption 3, the assumptions that the elements of $\{H_k\}$ are bounded in norm and that H_k is sufficiently positive definite in $\text{Null}(J_k)$ for all $k \in \mathbb{N}$ are typical for the constrained optimization literature. In practice, one may choose H_k to be (an approximation of) the Hessian of the Lagrangian at (X_k, Y_{k-1}) if such a matrix can be computed with reasonable effort. A simpler alternative is that H_k can be set to some positive definite diagonal matrix, such as the identity.

If a run terminates finitely, then an infeasible stationary point has been found, and there is nothing else to prove about the behavior of the algorithm. Hence, without loss of generality throughout the remainder of our analysis and discussions, we assume that the algorithm does not terminate finitely; for example, an infinite number of iterates are generated. This is implied by the following.

Assumption 4. *For all $k \in \mathbb{N}$, one finds $\|J_k^T C_k\|_2 > 0$ or $\|C_k\|_2 = 0$.*

We build to our main results through a series of lemmas. Our first lemma has appeared for various deterministic algorithms in the literature. It extends easily to our setting because the normal component computation is determined completely by X_k.

Lemma 1. *There exists $(\kappa_v, \underline{\omega}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ (uniform over all runs) such that*

$$\|C_k\|_2(\|C_k\|_2 - \|C_k + J_k V_k\|_2) \geq \kappa_v \|J_k^T C_k\|_2^2$$
$$\text{and } \ \underline{\omega}\|J_k^T C_k\|_2^2 \leq \|V_k\|_2 \leq \omega \|J_k^T C_k\|_2 \text{ for all } \ k \in \mathbb{N} \text{ with } \|C_k\|_2 > 0.$$

Proof. The proof follows the same logic as for lemmas 3.5 and 3.6 in Curtis et al. [6]. □

Our second lemma shows that the procedure for setting $\{\mathcal{X}_k\}$ and $\{\mathcal{Z}_k\}$ guarantees that these sequences are constant for sufficiently large $k \in \mathbb{N}$. The index at which these sequences become constant is a random quantity, but the pair of constants $(\chi_{\max}, \zeta_{\min}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ that is introduced in the lemma is a pair of bounds that are uniform over all runs.

Lemma 2. *There exists $(\chi_{\max}, \zeta_{\min}) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ (uniform over all runs) such that for some $K_{\chi,\zeta} \in \mathbb{N}$ one finds $(\mathcal{X}_k, \mathcal{Z}_k) = (\mathcal{X}_{K_{\chi,\zeta}}, \mathcal{Z}_{K_{\chi,\zeta}})$ for all $k \in \mathbb{N}$ with $k \geq K_{\chi,\zeta}$, where $(\mathcal{X}_{K_{\chi,\zeta}}, \mathcal{Z}_{K_{\chi,\zeta}}) \in (0, \chi_{\max}] \times [\zeta_{\min}, \infty)$.*

Proof. Consider arbitrary $k \in \mathbb{N}$. If $D_k = 0$, then the algorithm sets $(\mathcal{X}_k, \mathcal{Z}_k) = (\mathcal{X}_{k-1}, \mathcal{Z}_{k-1})$. Otherwise, under Assumption 3, it follows for any $\chi \in \mathbb{R}_{>0}$ that $\|U_k\|_2^2 \geq \chi \|V_k\|_2^2$ implies

$$\begin{aligned}\frac{1}{2} D_k^T H_k D_k &= \frac{1}{2}\, U_k^T H_k U_k + U_k^T H_k V_k + \frac{1}{2} V_k^T H_k V_k \\ &\geq \frac{1}{2}\, \rho \|U_k\|_2^2 - \|U_k\|_2 \|H_k\|_2 \|V_k\|_2 - \frac{1}{2}\|H_k\|_2 \|V_k\|_2^2 \geq \left(\frac{\rho}{2} - \frac{\kappa_H}{\sqrt{\chi}} - \frac{\kappa_H}{2\chi}\right)\|U_k\|_2^2.\end{aligned}$$

Hence, for sufficiently large $\chi \in \mathbb{R}_{>0}$, one finds that $\|U_k\|_2^2 \geq \chi \|V_k\|_2^2$ implies $\frac{1}{2}\, D_k^T H_k D_k \geq \frac{1}{4} \rho \|U_k\|_2^2$. The conclusion follows from this fact and the procedure for setting $(\mathcal{X}_k, \mathcal{Z}_k)$ in (14) and (15). □

We now prove that the sequence $\{\Xi_k\}$ is constant for sufficiently large $k \in \mathbb{N}$. As in the previous lemma, the index at which this sequence becomes constant is a random quantity, but the constant $\xi_{\min} \in \mathbb{R}_{>0}$ represents a uniform bound that holds over all runs.

Lemma 3. *There exists $\xi_{\min} \in \mathbb{R}_{>0}$ (uniform over all runs) such that for some $K_\xi \in \mathbb{N}$ one finds $\Xi_k = \Xi_{K_\xi}$ for all $k \in \mathbb{N}$ with $k \geq K_\xi$, where $\xi_{K_\xi} \in [\xi_{\min}, \infty)$.*

Proof. Consider arbitrary $k \in \mathbb{N}$. If $D_k = 0$, then the algorithm sets $\Xi_k = \Xi_{k-1}$. If $D_k \neq 0$ and $\|U_k\|_2^2 \geq \mathcal{X}_k \|V_k\|_2^2$, then it follows from (12)–(13) and (16)–(17) that either $\Xi_k = \Xi_{k-1}$ or

$$\begin{aligned}\Xi_k \geq (1 - \epsilon_\xi)\Xi_k^{\text{trial}} &= (1 - \epsilon_\xi)\left(\frac{\Delta l(X_k, \mathcal{T}_k, G_k, D_k)}{\mathcal{T}_k \|D_k\|_2^2}\right) \\ &\geq (1 - \epsilon_\xi)\frac{\mathcal{T}_k \rho \|U_k\|_2^2}{\mathcal{T}_k(1 + \mathcal{X}_k^{-1})\|U_k\|_2^2} \geq (1 - \epsilon_\xi)\frac{\rho}{(1 + \chi_{-1}^{-1})}.\end{aligned}$$

If $D_k \neq 0$ and $\|U_k\|_2^2 < \mathcal{X}_k \|V_k\|_2^2$, then (12)–(13), (16)–(17), and Lemmas 1 and 2 imply $\Xi_k = \Xi_{k-1}$ or

$$\begin{aligned}\Xi_k \geq (1 - \epsilon_\xi)\Xi_k^{\text{trial}} &= (1 - \epsilon_\xi)\left(\frac{\Delta l(X_k, \mathcal{T}_k, G_k, D_k)}{\|D_k\|_2^2}\right) \\ &\geq (1 - \epsilon_\xi)\frac{\sigma \kappa_v \kappa_c^{-1} \|J_k^T C_k\|_2^2}{(\mathcal{X}_k + 1)\omega^2 \|J_k^T C_k\|_2^2} \geq (1 - \epsilon_\xi)\frac{\sigma \kappa_v \kappa_c^{-1}}{(\chi_{\max} + 1)\omega^2}.\end{aligned}$$

Combining these results, the desired conclusion follows. □

Our next two lemmas provide useful relationships between true and stochastic quantities. The first result is similar to lemma 3.6 in Berahas et al. [1], although the proof presented here is different in order to handle potential rank deficiency of the constraint Jacobians.

Lemma 4. *For all* $k \in \mathbb{N}$, $\mathbb{E}_k[U_k] = U_k^{\text{true}}$ *and* $\mathbb{E}_k[\|D_k - D_k^{\text{true}}\|_2] \leq \rho^{-1}\sqrt{\nu}$.

Proof. Consider arbitrary $k \in \mathbb{N}$. Under Assumption 3, it follows from (8) that there exist W_k and W_k^{true} such that $U_k = Z_k W_k$ and $U_k^{\text{true}} = Z_k W_k^{\text{true}}$, where $W_k = -(Z_k^T H_k Z_k)^{-1} Z_k^T (G_k + H_k V_k)$ and $W_k^{\text{true}} = -(Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(X_k) + H_k V_k)$. Because $(Z_k^T H_k Z_k)^{-1} Z_k^T$ and Z_k are linear operators, it follows that $\mathbb{E}_k[W_k] = W_k^{\text{true}}$, and hence, $\mathbb{E}_k[U_k] = U_k^{\text{true}}$, as desired. Then, it follows from consistency and sub-multiplicity of the spectral norm, orthonormality of Z_k, Jensen's inequality, concavity of the square root operator, and Assumptions 2 and 3 that

$$\begin{aligned}
\mathbb{E}_k[\|D_k - D_k^{\text{true}}\|_2] = \mathbb{E}_k[\|U_k - U_k^{\text{true}}\|_2] &= \mathbb{E}_k[\|Z_k(W_k - W_k^{\text{true}})\|_2] \\
&= \mathbb{E}_k[\|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (G_k - \nabla f(X_k))\|_2] \\
&\leq \mathbb{E}_k[\|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\|_2 \|G_k - \nabla f(X_k)\|_2] \\
&= \|(Z_k^T H_k Z_k)^{-1}\|_2 \mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \\
&\leq \rho^{-1}\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2] \\
&\leq \rho^{-1}\sqrt{\mathbb{E}_k[\|G_k - \nabla f(X_k)\|_2^2]} \leq \rho^{-1}\sqrt{\nu},
\end{aligned}$$

which is the final desired conclusion. □

Our next result is part of lemma 3.9 in Berahas et al. [1]; we provide a proof for completeness.

Lemma 5. *For all* $k \in \mathbb{N}$, $\nabla f(X_k)^T D_k^{\text{true}} \geq \mathbb{E}_k[G_k^T D_k] \geq \nabla f(X_k)^T D_k^{\text{true}} - \rho^{-1}\nu$.

Proof. Consider arbitrary $k \in \mathbb{N}$. The arguments in the proof of Lemma 4 give

$$\begin{aligned}
G_k^T U_k &= -G_k^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (G_k + H_k V_k) \\
\text{and}\ \ \nabla f(X_k)^T U_k^{\text{true}} &= -\nabla f(X_k)^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(X_k) + H_k V_k).
\end{aligned}$$

On the other hand, under Assumptions 2 and 3, it follows that

$$\rho^{-1}\nu \geq \mathbb{E}_k[\|Z_k^T(G_k - \nabla f(X_k))\|^2_{(Z_k^T H_k Z_k)^{-1}}] \geq 0,$$

where

$$\begin{aligned}
&\mathbb{E}_k[\|Z_k^T(G_k - \nabla f(X_k))\|^2_{(Z_k^T H_k Z_k)^{-1}}] \\
&= \mathbb{E}_k[\|Z_k^T G_k\|^2_{(Z_k^T H_k Z_k)^{-1}}] - 2\mathbb{E}_k[G_k^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T \nabla f(X_k)] + \|Z_k^T \nabla f(X_k)\|^2_{(Z_k^T H_k Z_k)^{-1}} \\
&= \mathbb{E}_k[\|Z_k^T G_k\|^2_{(Z_k^T H_k Z_k)^{-1}}] - \|Z_k^T \nabla f(X_k)\|^2_{(Z_k^T H_k Z_k)^{-1}}.
\end{aligned}$$

Combining the facts above and again using Assumption 2, it follows that

$$\begin{aligned}
\nabla f(X_k)^T D_k^{\text{true}} - \mathbb{E}_k[G_k^T D_k] &= \nabla f(X_k)^T V_k + \nabla f(X_k)^T U_k^{\text{true}} - \mathbb{E}_k[G_k^T V_k + G_k^T U_k] \\
&= \nabla f(X_k)^T U_k^{\text{true}} - \mathbb{E}_k[G_k^T U_k] \\
&= -\nabla f(X_k)^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(X_k) + H_k V_k) \\
&\quad + \mathbb{E}_k[G_k^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (G_k + H_k V_k)] \\
&= -\|Z_k^T \nabla f(X_k)\|^2_{(Z_k^T H_k Z_k)^{-1}} + \mathbb{E}_k[\|Z_k^T G_k\|^2_{(Z_k^T H_k Z_k)^{-1}}] \in [0, \rho^{-1}\nu],
\end{aligned}$$

which gives the desired conclusion. □

In the subsequent subsections, our analysis turns to offering guarantees conditioned on each of a few possible events that can occur, a few of which involve that the merit parameter sequence eventually remains constant. To motivate the fact that such behavior of the merit parameter sequence is indeed possible, we show in Appendix B a set of conditions that guarantee that the merit parameter sequence eventually remains constant. We put this material in an appendix because our analyses in the following sections do not rely on it.

4.1. Constant, Sufficiently Small Merit Parameter

Our goal in this subsection is to prove a convergence guarantee for our algorithm under the following event, defined with respect to given constants $k_{\min} \in \mathbb{N}$, $\chi_{\max} \in \mathbb{R}_{>0}$, $\zeta_{\min} \in \mathbb{R}_{>0}$, $\tau_{\min} \in \mathbb{R}_{>0}$, and $\xi_{\min} \in \mathbb{R}_{>0}$:

$$\begin{aligned} E_{\tau,\text{low}}(k_{\min},\chi_{\max},\zeta_{\min},\tau_{\min},\xi_{\min}) := \{&\text{there exist } K' \in \mathbb{N} \text{with } K' \le k_{\min}, \\ &\mathcal{X}' \in \mathbb{R}_{>0} \text{ with } \mathcal{X}' \le \chi_{\max}, \\ &\mathcal{Z}' \in \mathbb{R}_{>0} \text{ with } \mathcal{Z}' \ge \zeta_{\min}, \\ &\mathcal{T}' \in \mathbb{R}_{>0} \text{ with } \mathcal{T}' \ge \tau_{\min}, \text{ and} \\ &\Xi' \in \mathbb{R}_{>0} \text{ with } \Xi' \ge \xi_{\min} \text{ such that} \\ &\mathcal{X}_k = \mathcal{X}', \mathcal{Z}_k = \mathcal{Z}', \mathcal{T}_k = \mathcal{T}' \le \mathcal{T}_k^{\text{trial, true}}, \\ &\text{and } \Xi_k = \Xi' \text{ for all } k \in \mathbb{N} \text{ with } k \ge K'\}. \end{aligned}$$

The following assumption is made throughout this subsection. One could generalize the assumption to allow different constants in the prior assumptions because they are now assumed to hold conditioned on a particular event, but for the sake of simplicity we assume that the prior assumptions hold with the same constants as previously introduced.

Assumption 5. *For some* $(k_{\min},\chi_{\max},\zeta_{\min},\tau_{\min},\xi_{\min}) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, *the event* $E_{\tau,\text{low}} := E_{\tau,\text{low}}(k_{\min}, \chi_{\max},\zeta_{\min},\tau_{\min},\xi_{\min})$ *occurs, and, conditioned on the occurrence of* $E_{\tau,\text{low}}$, *Assumptions 1–4 hold (with the same constants).*

Recall from Lemmas 2 and 3 that, even without considering $E_{\tau,\text{low}}$, the sequences $\{\mathcal{X}_k\}$, $\{\mathcal{Z}_k\}$, and $\{\Xi_k\}$ are bounded uniformly with respect to the constants $\chi_{\max}$, $\zeta_{\min}$, and $\xi_{\min}$, respectively. Therefore, it is not a stretch in our analysis in this section that the event $E_{\tau,\text{low}}$ includes that these sequences are eventually constant with the stated bounds. All that the event adds with respect to these sequences is that they are constant by iteration $k_{\min}$. Hence, the critical distinction in Assumption 5 is that the merit parameter also becomes constant at a value that is sufficiently small such that $\mathcal{T}_k \le \mathcal{T}_k^{\text{trial, true}}$ for all sufficiently large $k \in \mathbb{N}$. This is the key distinction between the event $E_{\tau,\text{low}}$ and the other events that we consider in upcoming subsections.

With respect to $E_{\tau,\text{low}}$, let us denote the trace σ-algebra of $E_{\tau,\text{low}}$ on $\mathcal{G}_k$ as $\mathcal{F}_k := \mathcal{G}_k \cap E_{\tau,\text{low}}$ for all $k \in \mathbb{N}$. It follows that $\{\mathcal{F}_k\}$ is a filtration, and under Assumption 5, the results of Lemmas 1, 4, and 5 all carry over from the previous section, where in the cases of the latter two lemmas one redefines $\mathbb{E}_k[\cdot] := \mathbb{E}_\iota[\cdot|\mathcal{F}_k]$. We redefine $\mathbb{E}_k$ in this manner for this subsection and also let $\mathbb{P}_k[\cdot] := \mathbb{P}_\iota[\cdot|\mathcal{F}_k]$.

Our next lemma provides a key result that drives our analysis for this subsection. It shows that as long as β_k is sufficiently small for all $k \in \mathbb{N}$ with $k \ge k_{\min}$ (in a manner similar to (22)), the reduction in the merit function in each iteration is at least the sum of two terms: (i) the reduction in the model of the merit function corresponding to the *true* gradient and its associated search direction and (ii) a pair of quantities that can be attributed to the error in the stochastic gradient estimate. For practical purposes, it is important to recognize here that even though the bound stated in (28) depends on stochastic quantities, these are quantities that are known in a run of the algorithm; after all, under Assumption 5, the values Ξ' and $\mathcal{T}'$ are $\mathcal{F}_k$-measurable. Therefore, even prior to iteration $k_{\min}$, the bound can be enforced by ensuring that β_k is chosen sufficiently small such that $2(1-\eta)\beta_k\Xi_k\max\{\mathcal{T}_k,1\}/(\mathcal{T}_kL+\Gamma) \in (0,1]$ for all $k \in \mathbb{N}$.

Lemma 6. *Suppose that* $\{\beta_k\}_{k=k_{\min}}^{\infty}$ *is chosen such that*

$$\beta_k \in (0,1] \text{ and } \frac{2(1-\eta)\beta_k\Xi'\max\{\mathcal{T}',1\}}{\mathcal{T}'L+\Gamma} \in (0,1] \text{ for all } k \in \mathbb{N} \text{ with } k \ge k_{\min}. \tag{28}$$

Then, for all such k, it follows that

$$\begin{aligned} &\phi(X_k,\mathcal{T}') - \phi(X_k + \mathcal{A}_kD_k,\mathcal{T}') \\ &\ge \mathcal{A}_k\Delta l(X_k,\mathcal{T}',\nabla f(X_k),D_k^{\text{true}}) - (1-\eta)\mathcal{A}_k\beta_k\Delta l(X_k,\mathcal{T}',G_k,D_k) - \mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}}). \end{aligned}$$

Proof. Consider arbitrary $k \in \mathbb{N}$ with $k \geq k_{\min}$. From (20)–(21) and the conditions on $\{\beta_k\}$, one finds $\mathcal{A}_k \in (0,1]$. Hence, with (18) and $J_k D_k = J_k D_k^{\text{true}}$ (because $J_k U_k = J_k U_k^{\text{true}} = 0$ by (8)), one has

$$\begin{aligned}
&\phi(X_k, \mathcal{T}') - \phi(X_k + \mathcal{A}_k D_k, \mathcal{T}') \\
\geq\ & -\mathcal{A}_k(\mathcal{T}'\nabla f(X_k)^T D_k - \|C_k\|_2 + \|C_k + J_k D_k\|_2) - \frac{1}{2}(\mathcal{T}'L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 \\
=\ & -\mathcal{A}_k(\mathcal{T}'\nabla f(X_k)^T D_k^{\text{true}} - \|C_k\|_2 + \|C_k + J_k D_k^{\text{true}}\|_2) \\
&-\frac{1}{2}(\mathcal{T}'L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 - \mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}}) \\
=\ & \mathcal{A}_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - \frac{1}{2}(\mathcal{T}'L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 - \mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}}).
\end{aligned} \tag{29}$$

By (20), it follows that $\mathcal{A}_k^{\text{suff}} \leq \frac{2(1-\eta)\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k)}{(\mathcal{T}'L+\Gamma)\|D_k\|_2^2}$. If $\|U_k\|_2^2 \geq \mathcal{X}_k\|V_k\|_2^2$, then (16)–(17) shows $\Xi_k \leq \Xi_k^{\text{trial}} = \frac{\Delta l(X_k, \mathcal{T}', G_k, D_k)}{\mathcal{T}'\|D_k\|_2^2}$ and $\frac{2(1-\eta)\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k)}{(\mathcal{T}'L+\Gamma)\|D_k\|_2^2} \geq \frac{2(1-\eta)\beta_k \Xi_k \mathcal{T}'}{\mathcal{T}'L+\Gamma}$. On the other hand, if $\|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2$, then (16)–(17) shows $\Xi_k \leq \Xi_k^{\text{trial}} = \frac{\Delta l(X_k, \mathcal{T}', G_k, D_k)}{\|D_k\|_2^2}$ and $\frac{2(1-\eta)\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k)}{(\mathcal{T}'L+\Gamma)\|D_k\|_2^2} \geq \frac{2(1-\eta)\beta_k \Xi_k}{\mathcal{T}'L+\Gamma}$. It follows from these facts and the supposition about $\{\beta_k\}$ that the projection in (21) never sets $\mathcal{A}_k > \mathcal{A}_k^{\text{suff}}$. Thus, $\mathcal{A}_k \leq \mathcal{A}_k^{\text{suff}} \leq \frac{2(1-\eta)\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k)}{(\mathcal{T}'L+\Gamma)\|D_k\|_2^2}$. Hence, by (29),

$$\begin{aligned}
&\phi(X_k, \mathcal{T}') - \phi(X_k + \mathcal{A}_k D_k, \mathcal{T}') \\
&\geq \mathcal{A}_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) \\
&\quad -\frac{1}{2}\mathcal{A}_k(\mathcal{T}'L + \Gamma)\left(\frac{2(1-\eta)\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k)}{(\mathcal{T}'L + \Gamma)\|D_k\|_2^2}\right)\|D_k\|_2^2 - \mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}}) \\
&= \mathcal{A}_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - (1-\eta)\mathcal{A}_k\beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k) - \mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}}),
\end{aligned}$$

which completes the proof. □

Our second result in this case offers a critical upper bound on the final term in the conclusion of Lemma 6. The result follows in a similar manner as lemma 3.11 in Berahas et al. [1].

Lemma 7. *It follows for any $k \in \mathbb{N}$ with $k \geq k_{\min}$ that*

$$\mathbb{E}_k[\mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}})] \leq \beta_k^2 \theta \mathcal{T}' \kappa_{\nabla f} \rho^{-1}\sqrt{\nu}.$$

Proof. Consider arbitrary $k \in \mathbb{N}$ with $k \geq k_{\min}$. We prove the desired conclusion under the assumption that the search direction in iteration k is tangentially dominated and then argue that it also holds by a similar argument when this search direction is normally dominated. Let I_k be the event that $\nabla f(X_k)^T(D_k - D_k^{\text{true}}) \geq 0$, and let I_k^c be the complementary event. By the law of total expectation, Assumption 5, and (21), one finds that

$$\begin{aligned}
&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}})] \\
=\ & \mathbb{E}_k[\mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k]\mathbb{P}_k[I_k] + \mathbb{E}_k[\mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k^c]\mathbb{P}_k[I_k^c] \\
\leq\ & \mathcal{A}_{k,\max}\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k]\mathbb{P}_k[I_k] + \mathcal{A}_{k,\min}\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k^c]\mathbb{P}_k[I_k^c],
\end{aligned}$$

where $\mathcal{A}_{k,\min} := \frac{2(1-\eta)\beta_k \Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}$ and $\mathcal{A}_{k,\max} := \frac{2(1-\eta)\beta_k \Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma} + \theta\beta_k^2$ are, respectively, the lower and upper bounds for the step size for the tangentially dominated search direction from (21). Thus, because $\mathbb{E}_k[D_k] = D_k^{\text{true}}$ by Lemma 4, the law of total expectation yields

$$\begin{aligned}
&\mathbb{E}_k[\mathcal{A}_k\mathcal{T}'\nabla f(X_k)^T(D_k - D_k^{\text{true}})] \\
\leq\ & \mathcal{A}_{k,\min}\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k]\mathbb{P}_k[I_k] + \mathcal{A}_{k,\min}\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k^c]\mathbb{P}_k[I_k^c] \\
&+ (\mathcal{A}_{k,\max} - \mathcal{A}_{k,\min})\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k]\mathbb{P}_k[I_k] \\
=\ & (\mathcal{A}_{k,\max} - \mathcal{A}_{k,\min})\mathcal{T}'\mathbb{E}_k[\nabla f(X_k)^T(D_k - D_k^{\text{true}})|I_k]\mathbb{P}_k[I_k].
\end{aligned}$$

Moreover, by the Cauchy-Schwarz inequality and law of total expectation, one finds

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

Combining the above results, (2), Lemma 4, and the fact that $\mathcal{A}_{k,\max} - \mathcal{A}_{k,\min} = \theta\beta_k^2$, the desired conclusion follows for tangentially dominated search directions. Finally, using the same arguments, except with $\mathcal{A}_{k,\min} := \frac{2(1-\eta)\beta_k\Xi'}{\mathcal{T}'L+\Gamma}$ and $\mathcal{A}_{k,\max} := \frac{2(1-\eta)\beta_k\Xi'}{\mathcal{T}'L+\Gamma} + \theta\beta_k^2$, where again $\mathcal{A}_{k,\max} - \mathcal{A}_{k,\min} = \theta\beta_k^2$, the desired conclusion follows for normally dominated search directions as well. □

Our next result in this case bounds the middle term in the conclusion of Lemma 6.

Lemma 8. *It follows for any $k \in \mathbb{N}$ with $k \geq k_{\min}$ that*

$$\mathbb{E}_k[\Delta l(X_k, \mathcal{T}', G_k, D_k)] \leq \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \mathcal{T}'\rho^{-1}\nu.$$

Proof. Consider arbitrary $k \in \mathbb{N}$ with $k \geq k_{\min}$. By Assumption 5, it follows from the model reduction definition (12), Lemma 5, and (8) that

$$\begin{aligned}
\mathbb{E}_k[\Delta l(X_k, \mathcal{T}_k, G_k, D_k)] &= \mathbb{E}_k[-\mathcal{T}'G_k^TD_k + \|C_k\|_2 - \|C_k + J_kD_k\|_2] \\
&\leq -\mathcal{T}'\nabla f(X_k)^TD_k^{\text{true}} + \mathcal{T}'\rho^{-1}\nu + \|C_k\|_2 - \|C_k + J_kD_k^{\text{true}}\|_2 \\
&= \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) + \mathcal{T}'\rho^{-1}\nu,
\end{aligned}$$

as desired. □

We now prove our main theorem of this subsection, where one should read

$$\mathbb{E}[\cdot] := \mathbb{E}[\cdot|\text{Assumption 5 holds}].$$

Again, it should be noted that even though the rules for choosing $\{\beta_k\}$ in the theorem are based on stochastic quantities, the rules are implementable; see our discussion at the end of this subsection.

Theorem 2. *Suppose Assumption 5 holds and $\{\beta_k\}_{k=k_{\min}}^{\infty}$ is chosen such that (28) holds. Define*

$$\begin{aligned}
&\underline{A} := \min\left\{\frac{2(1-\eta)\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}, \frac{2(1-\eta)\Xi'}{\mathcal{T}'L+\Gamma}\right\}, \quad \overline{A} := \max\left\{\frac{2(1-\eta)\Xi'\mathcal{T}'}{\mathcal{T}'L+\Gamma}, \frac{2(1-\eta)\Xi'}{\mathcal{T}'L+\Gamma}\right\}, \\
&\textit{and} \;\; \bar{\mathcal{N}} := \mathcal{T}'\rho^{-1}((1-\eta)(\overline{A}+\theta)\nu + \theta\kappa_{\nabla f}\sqrt{\nu})
\end{aligned}$$

along with

$$\begin{aligned}
&\underline{\alpha} := \min\left\{\frac{2(1-\eta)\xi_{\min}\tau_{\min}}{\tau_{\min}L+\Gamma}, \frac{2(1-\eta)\xi_{\min}}{\tau_{-1}L+\Gamma}\right\}, \quad \overline{\alpha} := \max\left\{\frac{2(1-\eta)\xi_{-1}\tau_{-1}}{\tau_{-1}L+\Gamma}, \frac{2(1-\eta)\xi_{-1}}{\tau_{\min}L+\Gamma}\right\}, \\
&\textit{and} \;\; \overline{\nu} := \tau_{-1}\rho^{-1}((1-\eta)(\overline{\alpha}+\theta)\nu + \theta\kappa_{\nabla f}\sqrt{\nu}).
\end{aligned}$$

If $\beta_k = \beta = \epsilon_\beta\underline{A}/((1-\eta)(\overline{A}+\theta))$ for some $\epsilon_\beta \in (0,1)$ for all $k \geq k_{\min}$, then for all $k \geq k_{\min}$ one finds

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{k-k_{\min}+1}\sum_{j=k_{\min}}^{k}\Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}})\right] \\
&\leq \frac{\left(\frac{\epsilon_\beta\overline{\alpha}}{(1-\eta)(\overline{\alpha}+\theta)}\right)^2\overline{\nu}}{\left(\frac{\epsilon_\beta\underline{\alpha}}{(1-\eta)(\underline{\alpha}+\theta)}\right)\underline{\alpha}} + \frac{\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')] - \phi_{\min}}{(k-k_{\min}+1)\left(\frac{\epsilon_\beta\underline{\alpha}}{(1-\eta)(\underline{\alpha}+\theta)}\right)\underline{\alpha}} \xrightarrow{k\to\infty} \frac{\left(\frac{\epsilon_\beta\overline{\alpha}}{(1-\eta)(\overline{\alpha}+\theta)}\right)^2\overline{\nu}}{\left(\frac{\epsilon_\beta\underline{\alpha}}{(1-\eta)(\underline{\alpha}+\theta)}\right)\underline{\alpha}},
\end{aligned} \tag{30}$$

where $\phi_{\min} \in \mathbb{R}_{>0}$ is a lower bound for $\phi(\cdot, \mathcal{T}')$ over $\mathcal{R}$, whose existence follows under Assumption 5. On the other hand, if $\{\beta_k\}_{k=k_{\min}}^{\infty}$ is determined by iteration $k_{\min}$ such that $\sum_{k=k_{\min}}^{\infty}\beta_k = \infty$, $\sum_{k=k_{\min}}^{\infty}\beta_k^2 < \infty$, and $\beta_k \leq \epsilon_\beta\underline{A}/((1-\eta)(\overline{A}+\theta))$ for *some* $\epsilon_\beta \in (0,1)$ *for all* $k \geq k_{\min}$*, then*

$$\lim_{k \geq k_{\min}, k \to \infty} \mathbb{E}\left[\frac{1}{(\sum_{j=k_{\min}}^{k} \beta_j)} \sum_{j=k_{\min}}^{k} \beta_j \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}})\right] = 0. \tag{31}$$

Proof. Consider arbitrary $k \in \mathbb{N}$ with $k \geq k_{\min}$. From the definitions of $\underline{A}$ and $\overline{A}$, the manner in which the step sizes are set by (21), and the fact that $\beta_k \in (0,1]$, it follows that $\underline{A}\beta_k \leq \mathcal{A}_k \leq (\overline{A} + \theta)\beta_k$. Hence, it follows from Lemmas 6–8 and the conditions of the theorem that

$$\begin{aligned}
&\phi(X_k, \mathcal{T}') - \mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}')] \\
&\geq \mathbb{E}_k[\mathcal{A}_k \Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - (1-\eta)\mathcal{A}_k \beta_k \Delta l(X_k, \mathcal{T}', G_k, D_k) - \mathcal{A}_k \mathcal{T}' \nabla f(X_k)^T (D_k - D_k^{\text{true}})] \\
&\geq \beta_k(\underline{A} - (1-\eta)(\overline{A} + \theta)\beta_k)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - \beta_k^2 \overline{\mathcal{N}}.
\end{aligned}$$

If $\beta_k = \beta = \epsilon_\beta \underline{A}/((1-\eta)(\overline{A}+\theta))$ for some $\epsilon_\beta \in (0,1)$ for all $k \geq k_{\min}$, then

$$\phi(X_k, \mathcal{T}') - \mathbb{E}_k[\phi(X_k + \mathcal{A}_k D_k, \mathcal{T}')] \geq \beta(1-\epsilon_\beta)\underline{A}\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - \beta^2 \overline{\mathcal{N}}.$$

In addition, because for such $\{\beta_k\}_{k=k_{\min}}^{\infty}$ one finds that $\epsilon_\beta \underline{\alpha}/((1-\eta)(\underline{\alpha}+\theta)) \leq \beta \leq \epsilon_\beta \overline{\alpha}/((1-\eta)(\overline{\alpha}+\theta))$ for all $k \geq k_{\min}$, one finds taking total expectation under Assumption 5 that

$$\begin{aligned}
&\mathbb{E}[\phi(X_k, \mathcal{T}') - \phi(X_k + \mathcal{A}_k D_k, \mathcal{T})] \\
&\geq \left(\frac{\epsilon_\beta \underline{\alpha}}{(1-\eta)(\underline{\alpha}+\theta)}\right)\underline{\alpha}\mathbb{E}[\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}})] - \left(\frac{\epsilon_\beta \overline{\alpha}}{(1-\eta)(\overline{\alpha}+\theta)}\right)^2 \overline{\nu} \text{ for all } k \geq k_\tau.
\end{aligned}$$

Summing this inequality for $j \in \{k_{\min}, \ldots, k\}$, it follows under Assumption 5 that

$$\begin{aligned}
&\mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}')] - \phi_{\min} \\
&\geq \mathbb{E}[\phi(X_{k_{\min}}, \mathcal{T}') - \phi(X_{k+1}, \mathcal{T}')] \\
&\geq \left(\frac{\epsilon_\beta \underline{\alpha}}{(1-\eta)(\underline{\alpha}+\theta)}\right)\underline{\alpha}\mathbb{E}\left[\sum_{j=k_{\min}}^{k} \Delta l(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}})\right] - (k - k_{\min} + 1)\left(\frac{\epsilon_\beta \overline{\alpha}}{(1-\eta)(\overline{\alpha}+\theta)}\right)^2 \overline{\nu},
\end{aligned}$$

from which (30) follows. On the other hand, if $\{\beta_k\}_{k=k_{\min}}^{\infty}$ satisfies the latter set of conditions in the theorem, then in a similar manner as above one finds for all $k \geq k_{\min}$ that

$$\mathbb{E}[\phi(X_k, \mathcal{T}') - \phi(X_k + \mathcal{A}_k D_k, \mathcal{T}')] \geq \mathbb{E}[\beta_k(\underline{A} - (1-\eta)(\overline{A}+\theta)\beta_k)\Delta l(X_k, \mathcal{T}', \nabla f(X_k), D_k^{\text{true}}) - \beta_k^2 \overline{\mathcal{N}}].$$

Summing this inequality for $j \in \{k_{\min}, \ldots, k\}$, rearranging terms, and taking limits as $k \to \infty$ yields the desired conclusion under Assumption 5. □

We end this subsection with a corollary in which we connect the result of Theorem 2 to first-order stationarity measures (recall (3)). For this corollary, we require the following lemma.

Lemma 9. *For all* $k \in \mathbb{N}$*, it holds that* $\|U_k^{\text{true}}\|_2 \geq \kappa_H^{-1}\|Z_k^T(\nabla f(X_k) + H_k V_k)\|_2$.

Proof. Consider arbitrary $k \in \mathbb{N}$. As in the proof of Lemma 4, $Z_k^T H_k Z_k W_k^{\text{true}} = -Z_k^T(\nabla f(X_k) + H_k V_k)$, so Assumption 5 (namely, Assumption 3) gives $\|U_k^{\text{true}}\|_2 \geq \kappa_H^{-1}\|Z_k^T(\nabla f(X_k) + H_k V_k)\|_2$. □

Corollary 1. *Under the conditions of Theorem 2, the following holds true.*

(a) *If* $\beta_k = \beta = \epsilon_\beta \underline{A}/((1-\eta)(\overline{A}+\theta))$ *for some* $\epsilon_\beta \in (0,1)$ *for all* $k \geq k_{\min}$*, and then for all* $k \geq k_{\min}$*, one finds that (30) holds with the left-hand side replaced by*

$$\mathbb{E}\left[\frac{1}{k - k_{\min} + 1}\sum_{j=k_{\min}}^{k}\left(\frac{\mathcal{T}'\rho\|Z_j^T(\nabla f(X_j) + H_j V_j)\|_2^2}{\kappa_H^2} + \frac{\kappa_v \sigma \|J_j^T C_j\|_2^2}{\kappa_c}\right)\right].$$

(b) *If* $\{\beta_k\}_{k=k_{\min}}^{\infty}$ *is determined by iteration* $k_{\min}$ *such that* $\sum_{k=k_{\min}}^{\infty} \beta_k = \infty$, $\sum_{k=k_{\min}}^{\infty} \beta_k^2 < \infty$*, and* $\beta_k \leq \epsilon_\beta \underline{A}/((1-\eta)(\overline{A}+\theta))$ *for some* $\epsilon_\beta \in (0,1)$ *for all* $k \geq k_{\min}$*, then*

$$\lim_{k \geq k_{\min}, k \to \infty} \mathbb{E}\left[\frac{1}{(\sum_{j=k_{\min}}^{k} \beta_j)}\sum_{j=k_{\min}}^{k} \beta_j\left(\frac{\mathcal{T}'\rho\|Z_j^T(\nabla f(X_j) + H_j V_j)\|_2^2}{\kappa_H^2} + \frac{\kappa_v \sigma \|J_j^T C_j\|_2^2}{\kappa_c}\right)\right] = 0,$$

from which it follows that

$$\liminf_{k\geq k_{\min},k\to\infty} \mathbb{E}\left[\frac{\mathcal{T}'\rho\|Z_k^T(\nabla f(X_k)+H_kV_k)\|_2^2}{\kappa_H^2}+\frac{\kappa_v\sigma\|J_k^TC_k\|_2^2}{\kappa_c}\right]=0.$$

Proof. For all $k\in\mathbb{N}$, it follows under Assumption 5 that (13) holds with $\nabla f(X_k)$ in place of G_k and U_k^{true} in place of U_k. The result follows from this fact, Theorem 2, and Lemmas 1 and 9. □

Observe that if the singular values of J_k are bounded below by $\sigma_J\in\mathbb{R}_{>0}$ for all $k\in\mathbb{N}$, then (as in the proof of Lemma B.1 in Appendix B) it follows that $\|J_k^TC_k\|_2\geq\sigma_J\|C_k\|_2$ for all $k\in\mathbb{N}$. In this case, the results of Corollary 1 hold with $\sigma_J\|C_k\|_2$ in place of $\|J_k^TC_k\|_2$. Overall, Corollary 1 offers results for the stochastic setting that parallel the limits (23) and (25) for the deterministic setting. The only difference is the presence of $Z_k^TH_kV_k$ in the term involving the reduced gradient $Z_k^T\nabla f(X_k)$ for all $k\in\mathbb{N}$. However, this does not significantly weaken the conclusion. After all, it follows from (5) (see also Lemma 1) that $\|V_k\|_2\leq\omega\|J_k^TC_k\|_2$ for all $k\in\mathbb{N}$. Hence, since Corollary 1 shows that at least a subsequence of $\{\|J_k^TC_k\|_2\}$ tends to vanish in expectation, it follows that $\{\|V_k\|_2\}$ vanishes in expectation along the same subsequence of iterations. This, along with Assumption 3 and the orthonormality of Z_k, shows that $\{\|Z_k^TH_kV_k\|_2\}$ exhibits this same behavior, which means that from the corollary one finds that a subsequence of $\{\|Z_k^T\nabla f(X_k)\|_2\}$ vanishes in expectation.

Let us conclude this subsection with a few additional remarks on how one can set $\{\beta_k\}$ to ensure our theoretical conclusions. One learns from the requirements on $\{\beta_k\}$ in Lemma 6, Theorem 2, and Corollary 1 that, rather than employ a prescribed sequence $\{\beta_k\}$, one should instead prescribe $\{\hat\beta_j\}_{j=0}^\infty\subset(0,1]$ and for each $k\in\mathbb{N}$ set β_k based on whether an adaptive parameter changes its value. In particular, anytime $k\in\mathbb{N}$ sees either $\mathcal{T}_k<\mathcal{T}_{k-1}$, $\mathcal{X}_k>\mathcal{X}_{k-1}$, $\mathcal{Z}_k<\mathcal{Z}_{k-1}$, or $\Xi_k<\Xi_{k-1}$, the algorithm should set $\beta_{k+j}\leftarrow\lambda\hat\beta_j$ for $j=0,1,2,\ldots$ (continuing indefinitely or until $\hat k\in\mathbb{N}$ with $\hat k>k$ sees $\mathcal{T}_{\hat k}<\mathcal{T}_{\hat k-1}$, $\mathcal{X}_{\hat k}>\mathcal{X}_{\hat k-1}$, $\mathcal{Z}_{\hat k}<\mathcal{Z}_{\hat k-1}$, or $\Xi_{\hat k}<\Xi_{\hat k-1}$), where $\lambda\in(0,1)$ is a prescribed value that is chosen sufficiently small such that (28) holds along with either of the conditions in (a) or (b) of Corollary 1. Because such a "reset" of $j\leftarrow 0$ will occur only a finite number of times under event $E_{\tau,\text{low}}$, one of the desirable results in Theorem 2/Corollary 1 can be attained if $\{\hat\beta_j\}$ is chosen as an appropriate constant or diminishing sequence.

4.2. Vanishing Merit Parameter

Let us now consider the behavior of the algorithm when the merit parameter vanishes. Because in such a scenario we are interested in the algorithm's ability to transition to an algorithm with guarantees similar to a *deterministic* method for minimizing constraint violation, we also assume that the noise in the stochastic gradient estimators is bounded. Specifically, for our purposes in this section, we introduce for a given bound $\nu\in\mathbb{R}_{>0}$ the event

$$E_{\tau,\text{zero}}(\nu):=\{\{\mathcal{T}_k\}\searrow 0 \text{ and } \|G_k-\nabla f(X_k)\|_2^2\leq\nu \text{ for all } k\in\mathbb{N}\}$$

and make the following assumption.

Assumption 6. *For some $\nu\in\mathbb{R}_{>0}$, the event $E_{\tau,\text{zero}}:=E_{\tau,\text{zero}}(\nu)$ occurs, and, conditioned on the occurrence of $E_{\tau,\text{zero}}$, Assumptions 1, 3, and 4 hold (with the same constants).*

Recalling Theorem 1 and Theorem B.1 (in Appendix B), one may conclude in general that the merit parameter sequence may vanish for one of two reasons; a (sub)sequence of constraint Jacobians tends toward rank-deficiency or a (sub)sequence of stochastic gradient estimates diverges. Assumption 6 has that the latter event does not occur. (In our remarks in Section 4.4, we discuss the obstacles that arise in proving convergence guarantees when the merit parameter vanishes and the stochastic gradient estimates diverge.) Given our setting of constrained optimization, it is reasonable and consistent with Theorem 1 to have convergence toward stationarity with respect to the constraint violation measure as the primary goal in these circumstances.

As in the prior subsection, with respect to $E_{\tau,\text{zero}}$, we can introduce the trace σ-algebra of $E_{\tau,\text{zero}}$ on $\mathcal{G}_k$ as $\mathcal{F}_k:=\mathcal{G}_k\cap E_{\tau,\text{zero}}$ for all $k\in\mathbb{N}$. Thus, in the same manner as in the previous subsection, we can proceed here by redefining $\mathbb{E}_k[\cdot]:=\mathbb{E}_\iota[\cdot|\mathcal{F}_k]$ and $\mathbb{P}_k[\cdot]:=\mathbb{P}_\iota[\cdot|\mathcal{F}_k]$ for this filtration $\{\mathcal{F}_k\}$.

Our first result in this subsection is an alternative of Lemma 6.

Lemma 10. *Under Assumption 6 and assuming that $\{\beta_k\}$ is chosen such that (28) holds for all $k\in\mathbb{N}$, it follows for all $k\in\mathbb{N}$ that*

$$\begin{aligned}&\|C_k\|_2-\|c(X_k+\mathcal{A}_kD_k)\|_2\\&\geq\mathcal{A}_k(1-(1-\eta)\beta_k)\Delta l(X_k,\mathcal{T}_k,G_k,D_k)-\mathcal{T}_k(f(X_k)-f(X_k+\mathcal{A}_kD_k))-\mathcal{A}_k\mathcal{T}_k(\nabla f(X_k)-G_k)^TD_k.\end{aligned}$$

Proof. Consider arbitrary $k \in \mathbb{N}$. As in the proof of Lemma 6, from (20)–(21) and the supposition about $\{\beta_k\}$, one finds $\alpha_k \in (0, 1]$. Hence, with (18), one has

$$\begin{aligned} &\phi(X_k, \mathcal{T}_k) - \phi(X_k + \mathcal{A}_k D_k, \mathcal{T}_k) \\ &\geq -\mathcal{A}_k(\mathcal{T}_k \nabla f(X_k)^T D_k - \|C_k\|_2 + \|C_k + J_k D_k\|_2) - \frac{1}{2}(\mathcal{T}_k L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 \\ &= \mathcal{A}_k \Delta l(X_k, \mathcal{T}_k, G_k, D_k) - \frac{1}{2}(\mathcal{T}_k L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 - \mathcal{A}_k\mathcal{T}_k(\nabla f(X_k) - G_k)^T D_k. \end{aligned}$$

Following the same arguments as in the proof of Lemma 6, it follows that $-\frac{1}{2}(\mathcal{T}_k L + \Gamma)\mathcal{A}_k^2\|D_k\|_2^2 \geq -(1-\eta)\mathcal{A}_k\beta_k \Delta l(X_k, \mathcal{T}_k, G_k, D_k)$, which combined with the above yields the desired conclusion. □

Our next result yields a bound on the final term in the conclusion of Lemma 10.

Lemma 11. *There exists $\kappa_\beta \in \mathbb{R}_{>0}$ (uniform over all runs) such that*

$$\mathcal{A}_k\mathcal{T}_k(\nabla f(X_k) - G_k)^T D_k \leq \begin{cases} \beta_k \mathcal{T}_k \kappa_\beta & \text{for all } k \in \mathbb{N} \text{ such that } \|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2 \\ \beta_k \mathcal{T}_k \max\{\beta_k, \mathcal{T}_k\}\kappa_\beta & \text{for all } k \in \mathbb{N} \text{ such that } \|U_k\|_2^2 \geq \mathcal{X}_k\|V_k\|_2^2. \end{cases}$$

Proof. The existence of $\kappa_d \in \mathbb{R}_{>0}$ (uniform over all runs) such that $\|D_k\|_2 \leq \kappa_d$ for all $k \in \mathbb{N}$ follows from Assumption 6, the fact that $\|D_k\|_2^2 = \|V_k\|_2^2 + \|U_k\|_2^2$ for all $k \in \mathbb{N}$, Lemma B.2, Lemma 1, and Assumption 1. Now consider arbitrary $k \in \mathbb{N}$. If $(\nabla f(X_k) - G_k)^T D_k < 0$, and then the desired conclusion follows trivially (for any $\kappa_\beta \in \mathbb{R}_{>0}$). Hence, let us proceed under the assumption that $(\nabla f(X_k) - G_k)^T D_k \geq 0$. If $\|U_k\| < \mathcal{X}_k\|V_k\|_2^2$, then it follows from (21), the fact that $0 \leq \mathcal{T}_k$, $\Xi_k \leq \Xi_{-1}$, and $\beta_k \leq 1$ for all $k \in \mathbb{N}$, the Cauchy-Schwarz inequality, and Assumption 6 that

$$\begin{aligned} \mathcal{A}_k\mathcal{T}_k(\nabla f(X_k) - G_k)^T D_k &\leq \left(\frac{2(1-\eta)\beta_k\Xi_k}{\mathcal{T}_k L + \Gamma} + \theta\beta_k^2\right)\mathcal{T}_k\|\nabla f(X_k) - G_k\|_2\|D_k\|_2 \\ &\leq \left(\frac{2(1-\eta)\xi_{-1}}{\Gamma} + \theta\right)\beta_k\mathcal{T}_k\sqrt{\nu}\kappa_d. \end{aligned}$$

On the other hand, if $\|U_k\|_2^2 \geq \mathcal{X}_k\|V_k\|_2^2$, then it follows under the same reasoning that

$$\begin{aligned} \mathcal{A}_k\mathcal{T}_k(\nabla f(X_k) - G_k)^T D_k &\leq \left(\frac{2(1-\eta)\beta_k\Xi_k\mathcal{T}_k}{\mathcal{T}_k L + \Gamma} + \theta\beta_k^2\right)\mathcal{T}_k\|\nabla f(X_k) - G_k\|_2\|D_k\|_2 \\ &\leq \left(\frac{2(1-\eta)\xi_{-1}}{\Gamma} + \theta\right)\beta_k\mathcal{T}_k\max\{\beta_k, \mathcal{T}_k\}\sqrt{\nu}\kappa_d. \end{aligned}$$

Overall, the desired conclusion follows with $\kappa_\beta := \left(\frac{2(1-\eta)\xi_{-1}}{\Gamma} + \theta\right)\sqrt{\nu}\kappa_d$. □

Our third result in this subsection offers a formula for a positive lower bound on the step size that is applicable at points that are not stationary for the constraint infeasibility measure. For this lemma and its subsequent consequences, we define for arbitrary $\gamma \in \mathbb{R}_{>0}$ the subset

$$\mathcal{R}_\gamma := \{x \in \mathbb{R}^n : \|J(x)^T c(x)\|_2 \geq \gamma\}. \tag{32}$$

Lemma 12. *There exists $\underline{\alpha} \in \mathbb{R}_{>0}$ such that $\mathcal{A}_k \geq \underline{\alpha}\beta_k$ for each $k \in \mathbb{N}$ such that $\|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2$. On the other hand, for each $\gamma \in \mathbb{R}_{>0}$, there exists $\epsilon_\gamma \in \mathbb{R}_{>0}$ (proportional to γ^2) such that*

$$X_k \in \mathcal{R}_\gamma \text{ implies } \mathcal{A}_k \geq \min\{\epsilon_\gamma\beta_k, \epsilon_\gamma\beta_k\mathcal{T}_k + \theta\beta_k^2\} \text{ whenever } \|U_k\|_2^2 \geq \mathcal{X}_k\|V_k\|_2^2.$$

Proof. Define $\mathcal{K}_\gamma := \{k \in \mathbb{N} : X_k \in \mathcal{R}_\gamma\}$. By Lemma 1, it follows that $\|V_k\|_2 \geq \underline{\omega}\|J_k^T C_k\|_2^2 \geq \underline{\omega}\gamma^2$ for all $k \in \mathcal{K}_\gamma$. Consequently, by Lemma B.2, it follows that

$$\|U_k\|_2 \leq \frac{\kappa_u}{\underline{\omega}\gamma^2}\|V_k\|_2 \text{ for all } k \in \mathcal{K}_\gamma. \tag{33}$$

It follows from (21) that $\mathcal{A}_k \geq 2(1-\eta)\beta_k\Xi_k/(\mathcal{T}_k L + \Gamma)$ whenever $\|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2$. Otherwise, whenever $\|U_k\|_2^2 \geq \mathcal{X}_k\|V_k\|_2^2$, it follows using the arguments in Lemma 6 and (21) that

$$\mathcal{A}_k = \min\left\{\frac{2(1-\eta)\beta_k\Delta l(X_k, \mathcal{T}_k, G_k, D_k)}{(\mathcal{T}_k L + \Gamma)\|D_k\|_2^2}, \frac{2(1-\eta)\beta_k\Xi_k\mathcal{T}_k}{\mathcal{T}_k L + \Gamma} + \theta\beta_k^2, 1\right\},$$

which along with (13), Lemma 1, (2), and (33) imply that

$$\begin{aligned}
\mathcal{A}_k &\geq \min\left\{\frac{2(1-\eta)\beta_k\sigma(\|C_k\|_2 - \|C_k + J_kV_k\|_2)}{(\mathcal{T}_kL+\Gamma)(\|U_k\|_2^2 + \|V_k\|_2^2)}, \frac{2(1-\eta)\beta_k\Xi_k\mathcal{T}_k}{\mathcal{T}_kL+\Gamma} + \theta\beta_k^2, 1\right\} \\
&\geq \min\left\{\frac{2(1-\eta)\beta_k\sigma\kappa_v\|J_k^TC_k\|^2}{(\mathcal{T}_kL+\Gamma)\left(\frac{\kappa_u^2}{\underline{\omega}^2\gamma^4}+1\right)\omega^2\|C_k\|\|J_k^TC_k\|^2}, \frac{2(1-\eta)\beta_k\Xi_k\mathcal{T}_k}{\mathcal{T}_kL+\Gamma} + \theta\beta_k^2, 1\right\} \\
&\geq \min\left\{\frac{2(1-\eta)\beta_k\sigma\kappa_v\underline{\omega}^2\gamma^4}{(\mathcal{T}_kL+\Gamma)\kappa_c\omega^2(\kappa_u^2+\underline{\omega}^2\gamma^4)}, \frac{2(1-\eta)\beta_k\Xi_k\mathcal{T}_k}{\mathcal{T}_kL+\Gamma} + \theta\beta_k^2, 1\right\}.
\end{aligned}$$

Combining the cases above with Lemma 3 yields the desired conclusion. □

We now prove our main theorem of this subsection, followed by a discussion of its consequences.

Theorem 3. *Suppose that Assumption 6 holds, the sequence $\{\beta_k\}$ is chosen such that (28) holds for all $k \in \mathbb{N}$, and either*
(a) *$\beta_k = \beta \in (0,1)$ for all $k \in \mathbb{N}$ or*
(b) *$\sum_{k=0}^{\infty}\beta_k = \infty$, $\sum_{k=0}^{\infty}\beta_k^2 < \infty$, and either $|\{k \in \mathbb{N} : \|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2\}| = \infty$ or $\sum_{k=0}^{\infty}\beta_k\mathcal{T}_k = \infty$.*
Then, $\liminf_{k\to\infty}\|J_k^TC_k\|_2 = 0$.

Proof. To derive a contradiction, consider the event that there exists (potentially run-dependent) $\gamma \in \mathbb{R}_{>0}$ and $K_\gamma \in \mathbb{N}$ such that $X_k \in \mathcal{R}_\gamma$ for all $k \in \mathbb{N}$ with $k \geq K_\gamma$. Our aim is to show that, under (a) or (b), a contradiction is reached, which will prove the result.

First, suppose that condition (a) holds. By Lemmas 10–12, (2), (13), the fact that $\beta \in (0,1)$, Lemma 1, and Assumption 6 (namely, Assumption 1), there exists (uniform) $\underline{\epsilon}_\gamma \in \mathbb{R}_{>0}$ such that

$$\begin{aligned}
&\|C_k\|_2 - \|C_{k+1}\|_2 \\
&\geq \mathcal{A}_k(1-(1-\eta)\beta)\Delta l(X_k,\mathcal{T}_k,G_k,D_k) - \mathcal{T}_k(f(X_k)-f(X_{k+1})) - \mathcal{A}_k\mathcal{T}_k(\nabla f(X_k)-G_k)^TD_k \\
&\geq \underline{\epsilon}_\gamma\beta\eta\sigma(\|C_k\|_2 - \|C_k+J_kV_k\|_2) - \mathcal{T}_k(f_{\sup}-f_{\inf}) - \beta\mathcal{T}_k\max\{1,\mathcal{T}_k\}\kappa_\beta \\
&\geq \underline{\epsilon}_\gamma\beta\eta\sigma\kappa_v\kappa_c^{-1}\|J_k^TC_k\|_2^2 - \mathcal{T}_k(f_{\sup}-f_{\inf}+\beta\max\{1,\mathcal{T}_k\}\kappa_\beta) \text{ for all } k \geq K_\gamma.
\end{aligned} \tag{34}$$

Because $\|J_k^TC_k\|_2 \geq \gamma$ for all $k \geq K_\gamma$ and $\{\mathcal{T}_k\} \searrow 0$ under Assumption 6, it follows that there exists (run-dependent) $K_\tau \geq K_\gamma$ such that $\mathcal{T}_k(f_{\sup}-f_{\inf}+\beta\max\{1,\mathcal{T}_k\}\kappa_\beta) \leq \frac{1}{2}\underline{\epsilon}_\gamma\beta\eta\sigma\kappa_v\kappa_c^{-1}\|J_k^TC_k\|_2^2$ for all $k \geq K_\tau$. Hence, summing (34) for $j \in \{K_\tau,\ldots,k\}$, it follows with (2) that

$$\kappa_c \geq \|C_{k_\tau}\|_2 - \|C_{k+1}\|_2 \geq \frac{1}{2}\underline{\epsilon}_\gamma\,\beta\eta\sigma\kappa_v\kappa_c^{-1}\sum_{j=K_\tau}^{k}\|J_j^TC_j\|_2^2.$$

It follows from this fact that $\{J_k^TK_k\}_{k\geq k_\tau,k\to\infty} \to 0$, yielding the desired contradiction.

Second, suppose that condition (b) holds. Because $\sum_{k=0}^{\infty}\beta_k^2 < \infty$, it follows that there exists $k_\beta \in \mathbb{N}$ with $k_\beta \geq k_\gamma$ such that $(1-(1-\eta)\beta_k) \geq \eta$ for all $k \geq k_\beta$. Hence, for all $k \geq k_\beta$ with $\|U_k\|_2^2 < \mathcal{X}_k\|V_k\|_2^2$, one finds from Lemmas 10–12, (2), (13), Lemma 1, and Assumption 1 that

$$\begin{aligned}
&\|C_k\|_2 - \|C_{k+1}\|_2 \\
&\geq \mathcal{A}_k(1-(1-\eta)\beta_k)\Delta l(X_k,\mathcal{T}_k,G_k,D_k) - \mathcal{T}_k(f(X_k)-f(X_{k+1})) - \mathcal{A}_k\mathcal{T}_k(\nabla f(X_k)-G_k)^TD_k \\
&\geq \beta_k\underline{\alpha}\eta\sigma\kappa_v\kappa_c^{-1}\|J_k^TC_k\|_2^2 - \mathcal{T}_k(f(X_k)-f_{\inf}) + \mathcal{T}_k(f(X_{k+1})-f_{\inf}) - \beta_k\mathcal{T}_k\kappa_\beta \\
&\geq \beta_k\underline{\alpha}\eta\sigma\kappa_v\kappa_c^{-1}\|J_k^TC_k\|_2^2 - \mathcal{T}_{k-1}(f(X_k)-f_{\inf}) + \mathcal{T}_k(f(X_{k+1})-f_{\inf}) - \beta_k\mathcal{T}_k\kappa_\beta.
\end{aligned}$$

Also, for all sufficiently large $k \geq k_\beta$ with $\|U_k\|_2^2 \geq \mathcal{X}_k \|V_k\|_2^2$, that is, $k \geq \overline{K}_\beta$, where (run-dependent) $\overline{K}_\beta \in \mathbb{N}$ is sufficiently large such that $\overline{K}_\beta \geq k_\beta$ and $\epsilon_\gamma \geq \epsilon_\gamma \mathcal{T}_k + \theta \beta_k$, similar reasoning yields

$$\begin{aligned}
&\|C_k\|_2 - \|C_{k+1}\|_2 \\
&\geq \mathcal{A}_k(1-(1-\eta)\beta_k)\Delta l(X_k, \mathcal{T}_k, G_k, D_k) - \mathcal{T}_k(f(X_k) - f(X_{k+1})) - \mathcal{A}_k \mathcal{T}_k (\nabla f(X_k) - G_k)^T D_k \\
&\geq \beta_k \max\{\beta_k, \mathcal{T}_k\} \min\{\epsilon_\gamma, \theta\} \eta \sigma \kappa_v \kappa_c^{-1} \|J_k^T C_k\|_2^2 \\
&\quad - \mathcal{T}_{k-1}(f(X_k) - f_{\inf}) + \mathcal{T}_k(f(X_{k+1}) - f_{\inf}) - \beta_k \max\{\beta_k, \mathcal{T}_k\} \mathcal{T}_k \kappa_\beta.
\end{aligned}$$

Because $\|J_k^T C_k\|_2 \geq \gamma$ for all $k \geq \overline{K}_\beta \geq k_\beta \geq k_\gamma$ and $\{\mathcal{T}_k\} \searrow 0$ under Assumption 6, there exists (run-dependent) $K_\tau \geq \overline{K}_\beta$ such that $\mathcal{T}_k \kappa_\beta \leq \frac{1}{2} \underline{\alpha} \eta \sigma \kappa_v \kappa_c^{-1} \|J_k^T C_k\|_2^2$ and $\mathcal{T}_k \kappa_\beta \leq \frac{1}{2} \min\{\epsilon_\gamma, \theta\} \eta \sigma \kappa_v \kappa_c^{-1} \|J_k^T C_k\|_2^2$ for all $k \geq K_\tau$. Hence, with (random) $\mathcal{K}_u := \{k \in \mathbb{N} : \|U_k\|_2^2 \geq \mathcal{X}_k \|V_k\|_2^2\}$ and $\mathcal{K}_v := \{k \in \mathbb{N} : \|U_k\|_2^2 < \mathcal{X}_k \|V_k\|_2^2\}$, summing the above for $j \in \{K_\tau, \dots, k\}$ gives

$$\begin{aligned}
\kappa_c \geq \|C_{K_\tau}\|_2 - \|C_{k+1}\|_2 \geq\; & -\mathcal{T}_{k_\tau - 1}(f(X_{K_\tau}) - f_{\inf}) + \mathcal{T}_k(f(X_{k+1}) - f_{\inf}) \\
& + \sum_{j=K_\tau, j \in \mathcal{K}_v}^{k} \beta_j (\underline{\alpha} \eta \sigma \kappa_v \kappa_c^{-1} \|J_j^T C_j\|_2^2 - \mathcal{T}_j \kappa_\beta) \\
& + \sum_{j=K_\tau, j \in \mathcal{K}_u}^{k} \beta_j \max\{\beta_j, \mathcal{T}_j\} (\min\{\epsilon_\gamma, \theta\} \eta \sigma \kappa_v \kappa_c^{-1} \|J_j^T C_j\|_2^2 - \mathcal{T}_k \kappa_\beta) \\
\geq\; & -\mathcal{T}_{K_\tau - 1}(f(X_{K_\tau}) - f_{\inf}) \\
& + \sum_{j=K_\tau, j \in \mathcal{K}_v}^{k} \beta_j \frac{1}{2} \underline{\alpha} \eta \sigma \kappa_v \kappa_c^{-1} \|J_j^T C_j\|_2^2 \\
& + \sum_{j=K_\tau, j \in \mathcal{K}_u}^{k} \beta_j \max\{\beta_j, \mathcal{T}_j\} \frac{1}{2} \min\{\epsilon_\gamma, \theta\} \eta \sigma \kappa_v \kappa_c^{-1} \|J_j^T C_j\|_2^2.
\end{aligned} \tag{35}$$

It follows from this fact and the fact that either $|\mathcal{K}_v| = \infty$ or at least $\sum_{j=K_\tau, j \in \mathcal{K}_u} \beta_j \mathcal{T}_j = \infty$ that $\{J_k^T C_k\}_{k \geq K_\tau} \to 0$, yielding the desired contradiction. □

There is one unfortunate case not covered by Theorem 3, namely, the case when $\{\beta_k\}$ diminishes (as in condition (b)), the search direction is tangentially dominated for all sufficiently large $k \in \mathbb{N}$, and $\sum_{k=0}^\infty \beta_k \mathcal{T}_k < \infty$. One can see in the proof of the theorem why the desired conclusion, namely, that the limit inferior of $\{\|J_k^T C_k\|_2\}$ is zero, does not necessarily follow in this setting. If, after some iteration, all search directions are tangentially dominated and $\sum_{k=0}^\infty \beta_k \mathcal{T}_k < \infty$, then the coefficients on $\|J_k^T C_k\|_2^2$ in (35) are summable, which means that there might not be a subsequence of $\{\|J_k^T C_k\|_2^2\}$ that vanishes. Fortunately, however, this situation is detectable in practice, in the sense that one can detect it using computed quantities. In particular, in a given realization of a run of the algorithm, if β_k is below a user-prescribed small threshold, $\|J_k^T c_k\|_2$ has remained above a user-prescribed threshold in all recent iterations, $\tau_k \leq c_{\tau,\beta} \beta_k$ for some user-prescribed $c_{\tau,\beta} \in \mathbb{R}_{>0}$ in recent iterations, and the algorithm has computed tangentially dominated search directions in all recent iterations, then the algorithm may benefit by triggering a switch to a setting in which $\{\beta_k\}$ is kept constant in future iterations, in which case the desired conclusion follows under condition (a). Such a trigger would be reasonable in a practical implementation in any case when the algorithm is having difficulty minimizing constraint violation, say, because of numerical errors. Also, such a trigger arguably does not conflict much with Section 4.1, because the analysis in that section presumes that $\{\tau_k\}$ remains bounded away from zero, whereas here one has confirmed that $\tau_k \approx 0$.

4.3. Constant, Insufficiently Small Merit Parameter

Our goal now is to consider the event that the algorithm generates a merit parameter sequence that eventually remains constant, but at a value that is too large in the sense that one does not find that $\mathcal{T}_k \leq \mathcal{T}_k^{\text{trial, true}}$ for all sufficiently large $k \in \mathbb{N}$. (Recall that such an inequality holding for all sufficiently large $k \in \mathbb{N}$ was the distinguishing feature of the event $E_{\tau,\text{low}}$ considered in Assumption 5.) Such an event for the algorithm in Berahas et al. [1] is addressed in the proof of proposition 3.16 in that article, where under a reasonable assumption (paralleling (37a) below) it is essentially shown that, conditioned on a certain index set (denoted as $\mathcal{K}_{gd}$ in that paper) having infinite cardinality, the probability is zero of the merit parameter settling on too large of a value. However, unfortunately, such an argument does not address what might be the probability over all possible runs of the algorithm that the merit parameter eventually remains constant at too large of a value in the manner described above. We discuss in this subsection that, under reasonable assumptions, this probability is zero, where a formal theorem and proof are provided in Appendix C. (The distinction being made here is that one cannot prove probability of the event over all runs by conditioning on $\mathcal{K}_{gd}$ in Berahas et al. [1] being infinite.)

For our purposes in this section, we make some mild simplifications. First, as shown in Lemmas 2 and 3, each of the sequences $\{\mathcal{X}_k\}$, $\{\mathcal{Z}_k\}$, and $\{\Xi_k\}$ has a uniform bound that holds over any run of the algorithm. Hence, for simplicity, we shall assume that the initial values of these sequences are chosen such that they are constant over $k \in \mathbb{N}$. (Our discussions in this subsection can be generalized to situations when this is not the case; the conversation merely becomes more cumbersome, which we have chosen to avoid.) Second, it follows as in our analysis of the deterministic instance of our algorithm (recall Theorem 1) that if a subsequence of $\{\mathcal{T}_k^{\text{trial, true}}\}$ converges to zero, then a subsequence of minimum singular values of the constraint Jacobians $\{J_k\}$ vanishes as well. Hence, we shall consider here the event, given $\tau_{\min}^{\text{trial,true}} \in \mathbb{R}_{>0}$, defined as

$$E_\tau(\tau_{\min}^{\text{trial,true}}) := \{\mathcal{X}_k = \chi_{-1},\ \mathcal{Z}_k = \zeta_{-1},\ \Xi_k = \xi_{-1},\ \text{and } \mathcal{T}_k^{\text{trial, true}} \geq \tau_{\min}^{\text{trial,true}} \text{ for all } k \in \mathbb{N}\}.$$

Given this definition, we assume the following throughout this subsection. (We will remark on the consequences of this assumption further in Section 4.4.)

Assumption 7. *For some $\tau_{\min}^{\text{trial,true}} \in \mathbb{R}_{>0}$, the event $E_\tau := E_\tau(\tau_{\min}^{\text{trial,true}})$ occurs, and, conditioned on the occurrence of E_τ, Assumptions 1–4 hold (with the same constants).*

It follows from the definition of E_τ and (11) that if the cardinality of the (random) set of iteration indices $\{k \in \mathbb{N} : \mathcal{T}_k < \mathcal{T}_{k-1}\}$ ever exceeds

$$\bar{s}(\tau_{\min}^{\text{trial,true}}) := \left\lceil \frac{\log(\tau_{\min}^{\text{trial,true}}/\tau_{-1})}{\log(1-\epsilon_\tau)} \right\rceil \in \mathbb{N}, \tag{36}$$

then, for all subsequent $k \in \mathbb{N}$, $\mathcal{T}_{k-1} \leq \tau_{\min}^{\text{trial,true}} \leq \mathcal{T}_k^{\text{trial, true}}$; that is, the merit parameter is not decreased.

As in the prior subsections, with respect to E_τ, we can introduce the trace σ-algebra of E_τ on $\mathcal{G}_k$ as $\mathcal{F}_k := \mathcal{G}_k \cap E_\tau$ for all $k \in \mathbb{N}$. Thus, in the same manner as in the previous subsections, we can proceed here by redefining $\mathbb{E}_k[\cdot] := \mathbb{E}_\iota[\cdot|\mathcal{F}_k]$ and $\mathbb{P}_k[\cdot] := \mathbb{P}_\iota[\cdot|\mathcal{F}_k]$ for this filtration $\{\mathcal{F}_k\}$.

For our analysis of this setting, we prove the following lemma, which parallels proposition 3.16 in Berahas et al. [1]. We remark that this lemma is only the first step toward our main result for this setting, in contrast to Berahas et al. [1], where proposition 3.16 was the end of the discussion.

Lemma 13. *For any $k \in \mathbb{N}$, it follows for any $p \in (0,1]$ that*

$$\mathbb{P}_k[G_k^T D_k + U_k^T H_k U_k \geq \nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}}] \geq p \tag{37a}$$

$$\textit{implies}\ \ \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \geq p. \tag{37b}$$

Proof. Suppose that $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}$ and

$$G_k^T D_k + U_k^T H_k U_k \geq \nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}}. \tag{38}$$

Then, $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}$ and (10) imply that $\mathcal{T}_k^{\text{trial, true}} < \infty$, $\nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}} > 0$, and

$$\mathcal{T}_k^{\text{trial, true}} = \frac{(1-\sigma)(\|C_k\|_2 - \|C_k + J_k D_k^{\text{true}}\|_2)}{\nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}}} < \mathcal{T}_{k-1},$$

from which it follows that

$$(1-\sigma)(\|C_k\|_2 - \|C_k + J_k D_k^{\text{true}}\|_2) < (\nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}})\mathcal{T}_{k-1}. \tag{39}$$

Therefore, (38), (39), and the fact that $J_k D_k^{\text{true}} = J_k D_k$ show that

$$(1-\sigma)(\|C_k\|_2 - \|C_k + J_k D_k\|_2) < (G_k^T D_k + U_k^T H_k U_k)\mathcal{T}_{k-1}.$$

It follows from this inequality and Lemma 1 that $G_k^T D_k + U_k^T H_k U_k > 0$, and with (11) it holds that

$$\mathcal{T}_k \leq \mathcal{T}_k^{\text{trial}} = \frac{(1-\sigma)(\|C_k\|_2 - \|C_k + J_k D_k\|_2)}{G_k^T D_k + U_k^T H_k U_k} < \mathcal{T}_{k-1}.$$

Hence, $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}$ and (38) together imply that $\mathcal{T}_k < \mathcal{T}_{k-1}$. Therefore, because the quantities $\mathcal{T}_k^{\text{trial, true}}$ and $\mathcal{T}_{k-1}$ are $\mathcal{F}_k$-measurable, it follows that

$$\begin{aligned} & \mathbb{P}_k[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \\ \geq\, & \mathbb{P}_k[G_k^T D_k + U_k^T H_k U_k \geq \nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}} | \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \\ =\, & \mathbb{P}_k[G_k^T D_k + U_k^T H_k U_k \geq \nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}}] \geq p, \end{aligned}$$

as desired. □

Using Lemma 13, we prove in Appendix C a formal version of the following informally written theorem. (We remark that example 3.17 in Berahas et al. [1] shows an example in which (37a) holds for all $k \in \mathbb{N}$.)

Theorem 4 (Informal Version of Theorem C.1 in Appendix C). *If Assumption 7 holds and there exists $p \in (0,1]$ such that a condition akin to (37a) holds for all $k \in \mathbb{N}$, then the probability is zero that there exists $\mathcal{K} \subseteq \mathbb{N}$ with $|\mathcal{K}| = \infty$ and $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}$ for all $k \in \mathcal{K}$, so the probability is zero that the merit parameter sequence eventually remains constant at too large of a value.*

The key to our proof of Theorem C.1 is the construction of a tree to characterize the stochastic process generated by the algorithm in a manner that one can employ the multiplicative form of Chernoff's bound to capture the probability of having repeated missed opportunities to decrease the merit parameter when it would have been reduced if the true gradients were computed.

4.4. Complementary Events

Our analyses in Sections 4.1, 4.2, and 4.3 do not cover all possible events. Ignoring events in which the stochastic gradients are biased and/or have unbounded variance, the events that complement those discussed in the prior subsections are the following:

- $E_{\text{unsettled}}$: there exists an infinite sequence of realizations, say, indexed by $j \in \mathbb{N}$, such that for realization j there exists some smallest index k_j such that $\{\chi_k\}_{k=k_j}^{\infty}$, $\{\zeta_k\}_{k=k_j}^{\infty}$, $\{\tau_k\}_{k=k_j}^{\infty}$, and $\{\xi_k\}_{k=k_j}^{\infty}$ are constant and $\tau_k \leq \tau_k^{\text{trial,true}}$ for all $k \in \mathbb{N}$ with $k \geq k_j$, but $\{k_j\} \to \infty$;
- $E_{\tau,\text{zero,bad}}$: $\{\mathcal{T}_k\} \searrow 0$, and for all $\nu \in \mathbb{R}_{>0}$ there exists $K \in \mathbb{N}$ such that $\|G_K - \nabla f(X_K)\|_2^2 > \nu$;
- $E_{\tau,\text{big,bad}}$: $\{\mathcal{T}_k^{\text{trial, true}}\} \searrow 0$ and there exists $\mathcal{T}'_{\text{big}} \in \mathbb{R}_{>0}$ such that $\mathcal{T}_k = \mathcal{T}_{\text{big}}$ for all large $k \in \mathbb{N}$.

The event $E_{\text{unsettled}}$ represents cases when the algorithm is essentially behaving nicely, but there does not exist a uniform iteration number by which the parameter values have *settled*. With respect to this event, we emphasize for many algorithms in the *deterministic* setting that it is not part of standard analysis to prove an upper bound on the number of iterations by which, say, the merit parameter value remains constant. Therefore, it is not a major shortcoming of our analysis that such a bound is not proved for our *stochastic* method whose behavior is not completely determined by the initial conditions. That said, we direct the reader to Curtis et al. [7], which under reasonable assumptions (similar to those made in this paper) shows that within a fixed budget of iterations the probability is high that the merit parameter sequences exhibit desirable properties. That analysis at least partially shows that *unsettled* behavior of the algorithm is not a major practical concern.

The event $E_{\tau,\text{zero,bad}}$ represents cases in which the merit parameter vanishes, whereas the stochastic gradient estimates do not remain in a bounded set. The difficulty of proving a guarantee for this setting can be seen as follows. If the merit parameter vanishes, then this is an indication that less emphasis should be placed on the objective over the course of the optimization process, which may indicate that the constraints are infeasible or degenerate. However, if a subsequence of stochastic gradient estimates diverges at the same time, then each large (in norm) stochastic gradient estimate may suggest that a significant amount of progress can be made in reducing the objective function despite the merit parameter having reached a small value (because it is vanishing). This disrupts the balance that the merit parameter attempts to negotiate between the objective and the constraint violation terms in the merit function. Our analysis of the event $E_{\tau,\text{zero}}$ in Section 4.2 shows that if the stochastic gradient estimates remain bounded, then the algorithm can effectively transition to solving the deterministic problem of minimizing constraint violation. However, it remains an open question whether it is possible to obtain a similar guarantee if/when a subsequence of stochastic gradient estimates diverges. Ultimately, one can argue that scenarios of unbounded noise, such as described here, might be of only theoretical interest rather than real, practical interest. For instance, if f is defined by a (large) finite sum of component functions whose gradients (evaluated at points in a set containing the iterates) are always contained in a ball of uniform radius about the gradient of f, a common scenario in practice, then $E_{\tau,\text{zero,bad}}$ cannot occur.

Now consider the event $E_{\tau,\text{big},\text{bad}}$. We have shown in Section 4.3 that under certain conditions, including if $\{\tau_k^{\text{trial,true}}\}$ is bounded below by $\tau_{\min}^{\text{trial,true}} \in \mathbb{R}_{>0}$, then the probability is zero that the merit parameter remains too large. However, this does not account for situations in which $\{\mathcal{T}_k^{\text{trial, true}}\}$ vanishes, whereas $\{\mathcal{T}_k\}$ does not. Nonetheless, we contend that $E_{\tau,\text{big},\text{bad}}$ can be ignored for practical purposes because the adverse effect that it may have on the algorithm is observable. In particular, if the merit parameter remains fixed at a value that is too large, then the worst that may occur in a realization of a run of the algorithm is that $\{\|J_k^T c_k\|_2\}$ does not vanish. A practical implementation of the algorithm would monitor this quantity in any case (because, by Corollary 1, even in $E_{\tau,\text{low}}$ one only knows that the limit inferior of the expectation of $\{\|J_k^T c_k\|_2\}$ vanishes) and reduce the merit parameter if progress toward reducing constraint violation is inadequate. Hence, $E_{\tau,\text{big},\text{bad}}$ (and in general the event of the merit parameter remaining too large) is an event that at most suggests practical measures of the algorithm that should be employed for $E_{\tau,\text{low}}$ in any case.

5. Numerical Experiments

The goal of our numerical experiments is to compare the empirical performance of our proposed stochastic SQP method (Algorithm 1) on problems from a couple of test set collections, for which we compare against some alternative approaches, and a modern data-fitting problem that exhibits rank deficiency. We implemented our algorithm in Matlab. Our code is publicly available; see https://github.com/frankecurtis/StochasticSQP. We first consider equality-constrained problems from the CUTEst collection (Gould et al. [12]), then consider two types of constrained logistic regression problems with data sets from the LIBSVM collection (Chang and Lin [3]), and finally consider an example in physics-informed machine learning. For the former two sets of experiments, we compare the performance of our method versus a stochastic sub-gradient algorithm (Davis et al. [10]) employed to minimize the exact penalty function (9) and, in one set of our logistic regression experiments where it is applicable, versus a stochastic projected gradient method. These algorithms were chosen because, like our method, they operate in the highly stochastic regime. We do not compare against the aforementioned method from Na et al. [22] because, as previously mentioned, that approach may refine stochastic gradient estimates during each iteration as needed by a line search. Hence, that method offers different types of convergence guarantees and is not applicable in our regime of interest.

In all of our experiments, results are given in terms of feasibility and stationarity errors at the *best* iterate, which is determined as follows. If, for a given problem instance, an algorithm produced an iterate that was sufficiently feasible in the sense that $\|c_k\|_\infty \leq 10^{-6}\max\{1,\|c_0\|_\infty\}$ for some $k \in \mathbb{N}$, then, with the largest $k \in \mathbb{N}$ satisfying this condition, the feasibility error was reported as $\|c_k\|_\infty$, and the stationarity error was reported as $\|\nabla f(x_k) + J_k^T y_k\|_\infty$, where y_k was computed as a least-squares multiplier using the true gradient $\nabla f(x_k)$ and J_k. (The multiplier y_k and corresponding stationarity error are not needed by our algorithm; they are computed merely so that we could record the error for our experimental results.) If, for a given problem instance, an algorithm did not produce a sufficiently feasible iterate, then the feasibility and stationarity errors were computed in the same manner at the least infeasible iterate (with respect to the measure of infeasibility $\|\cdot\|_\infty$).

5.1. Implementation Details

For all methods, Lipschitz constant estimates for the objective gradient and constraint Jacobian—playing the roles of L and Γ, respectively—were computed using differences of gradients near the initial point. Once these values were computed, they were kept constant for all subsequent iterations. This procedure was performed in such a way that, for each problem instance, all algorithms used the same values for these estimates.

As mentioned in Section 3, there are various extensions of our step size selection scheme with which one can prove, with appropriate modifications to our analysis, comparable convergence guarantees as are offered by our algorithm. We included one such extension in our software implementation for our experiments. In particular, in addition to α_k^{suff} in (20), one can directly consider the upper bound in (18) with the gradient $\nabla f(x_k)$ replaced by its estimate g_k, that is,

$$\begin{aligned}&\alpha\tau_k g_k^T d_k + |1-\alpha|\|c_k\|_2 - \|c_k\|_2 + \alpha\|c_k + J_k d_k\|_2 + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2\|d_k\|_2^2\\ &= -\alpha\Delta l(x_k,\tau_k,g_k,d_k) + |1-\alpha|\|c_k\|_2 - (1-\alpha)\|c_k\|_2 + \frac{1}{2}(\tau_k L + \Gamma)\alpha^2\|d_k\|_2^2,\end{aligned}$$

and consider the step size that minimizes this as a function of α (with scale factor β_k), namely,

$$\alpha_k^{\min} := \max\left\{\min\left\{\frac{\beta_k\Delta l(x_k,\tau_k,g_k,d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2}, 1\right\}, \frac{\beta_k\Delta l(x_k,\tau_k,g_k,d_k) - 2\|c_k\|_2}{(\tau_k L + \Gamma)\|d_k\|_2^2}\right\}. \tag{40}$$

(Such a value is used in Berahas et al. [1].) The algorithm can then set a trial step size as any satisfying

$$\alpha_k^{\rm trial} \in [\min\{\alpha_k^{\rm suff}, \alpha_k^{\rm min}\}, \max\{\alpha_k^{\rm suff}, \alpha_k^{\rm min}\}] \tag{41}$$

and set α_k as the projection of this value, rather than $\alpha_k^{\rm suff}$, for all $k \in \mathbb{N}$. (The projection interval in (21) should be modified, specifically with each instance of $2(1-\eta)$ replaced by $\min\{2(1-\eta), 1\}$, to account for the fact that the lower value in (41) may be smaller than $\alpha_k^{\rm suff}$. A similar modification is needed in the analysis, specifically in the requirements for $\{\beta_k\}$ in Lemma 6.)

One can also consider rules that allow even larger step sizes to be taken. For example, rather than consider the upper bound offered by the last expression in (18), one can consider any step size that ensures that the penultimate expression in (18) is less than or equal to the right-hand side of (19) with $\nabla f(x_k)$ replaced by g_k. Such a value can be found with a one-dimensional search over α with negligible computational cost. Our analysis can be extended to account for this option as well. However, for our experimental purposes here, we do not consider such an approach.

For our stochastic SQP method, we set $H_k \leftarrow I$ and $\alpha_k^{\rm trial} \leftarrow \max\{\alpha_k^{\rm suff}, \alpha_k^{\rm min}\}$ for all $k \in \mathbb{N}$. Other parameters were set as $\tau_{-1} \leftarrow 1$, $\chi_{-1} \leftarrow 10^{-3}$, $\zeta_{-1} \leftarrow 10^3$, $\xi_{-1} \leftarrow 1$, $\omega \leftarrow 10^2$, $\epsilon_v \leftarrow 1$, $\sigma \leftarrow 1/2$, $\epsilon_\tau \leftarrow 10^{-2}$, $\epsilon_\chi \leftarrow 10^{-2}$, $\epsilon_\zeta \leftarrow 10^{-2}$, $\epsilon_\xi \leftarrow 10^{-2}$, $\eta \leftarrow 1/2$, and $\theta \leftarrow 10^4$. For the stochastic sub-gradient method, the merit parameter value and step size were tuned for each problem instance, and for the stochastic projected gradient method, the step size was tuned for each problem instance; details are given in the following subsections. In all experiments, both the stochastic sub-gradient and stochastic projected gradient method were given many more iterations to find each of their best iterates for a problem instance; this is reasonable because the search direction computation for our method is more expensive than for the other methods. Again, further details are given below.

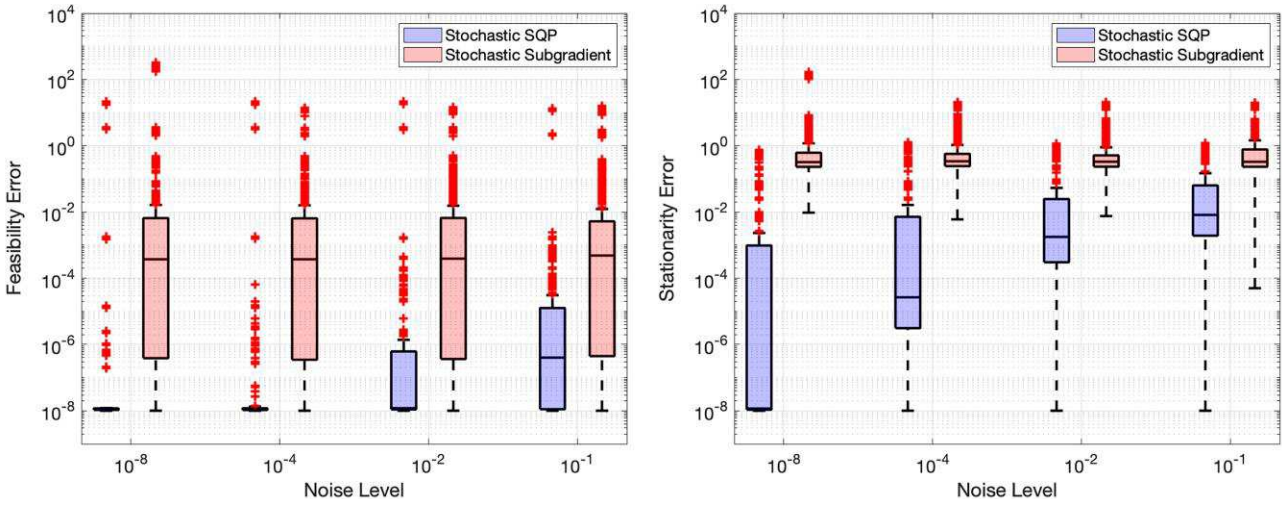
5.2. CUTEst Problems

In our first set of experiments, we consider equality-constrained problems from the CUTEst collection. Specifically, of the 136 such problems in the collection, we selected those for which (i) f is not a constant function and (ii) $n + m + 1 \leq 1000$. This selection resulted in a set of 67 problems. In order to consider the context in which the LICQ does not hold, for each problem we duplicated the last constraint. (This does not affect the feasible region nor the set of stationary points but ensures that the problem instances are degenerate.) Each problem comes with an initial point, which we used in our experiments. To make each problem stochastic, we added noise to each gradient computation. Specifically, for each run of an algorithm, we fixed a *noise level* as $\epsilon_N \in \{10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}\}$ and in each iteration set the stochastic gradient estimate as $g_k \leftarrow \mathcal{N}(\nabla f(x_k), \epsilon_N I)$. For each problem and noise level, we ran 10 instances with different random seeds. This led to a total of 670 runs of each algorithm for each noise level.

We set a budget of 1,000 iterations for our stochastic SQP algorithm and a more generous budget of 10,000 iterations for the stochastic sub-gradient method. We followed the same strategy as in Berahas et al. [1] to tune the merit parameter τ for the stochastic sub-gradient method but also tuned the step sizes through the sequence $\{\beta_k\}$. Specifically, for each problem instance, we ran the stochastic sub-gradient method for 11 different values of τ and four different values of β, namely, $\tau \in \{10^{-10}, 10^{-9}, \ldots, 10^0\}$ and $\beta \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$, set the step size as $\frac{\beta\tau}{\tau L + \Gamma}$, and selected the combination of τ and β for that problem instance that led to the best iterate overall. (We found through this process that the selected (τ, β) pairs were relatively evenly distributed over their ranges, meaning that this extensive tuning effort was useful to obtain better results for the stochastic sub-gradient method.) For our stochastic SQP method, we set $\beta_k \leftarrow 1$ for all $k \in \mathbb{N}$. Overall, between the additional iterations allowed in each run of the stochastic sub-gradient method, the different merit parameter values tested, and the different step sizes tested, the stochastic sub-gradient method was given 440 times the number of iterations that were given to our stochastic SQP method for each problem.

The results of this experiment are reported in the form of box plots in Figure 1. One finds that the best iterates from our stochastic SQP algorithm generally correspond to much lower feasibility and stationarity errors for all noise levels. The stationarity errors for our method degrade as the noise level increases, but this is not surprising because these experiments are run with $\{\beta_k\}$ being a constant sequence. It is interesting, however, that our algorithm typically finds iterates that are sufficiently feasible, even for relatively high noise levels. This shows that our approach handles the deterministic constraints well despite the stochasticity of the objective gradient estimates. Finally, we remark that for these experiments our algorithm found $\tau_{k-1} \leq \tau_k^{\rm trial,\,true}$ to hold in roughly 98% of all iterations for all runs (across all noise levels), and it found this inequality to hold in the last 50 iterations in

Figure 1. (Color online) Box plots for feasibility errors (left) and stationarity errors (right) when our stochastic SQP method and a stochastic sub-gradient method are employed to solve equality-constrained problems from the CUTEst collection.



100% of all runs. This provides evidence for our claim that the merit parameter not reaching a sufficiently small value is not an issue of practical concern.

5.3. Constrained Logistic Regression

In our next sets of experiments, we consider equality-constrained logistic regression problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-y_i(X_i^Tx)}\right) \text{ s.t. } Ax=b, \|x\|_2^2=1, \tag{42}$$

where $X\in\mathbb{R}^{n\times N}$ contains feature data for N data points (with X_i representing the ith column of X), $y\in\{-1,1\}^N$ contains corresponding label data, $A\in\mathbb{R}^{(m+1)\times n}$, and $b\in\mathbb{R}^{m+1}$. For instances of (X,y), we consider 11 binary classification data sets from the LIBSVM collection (Chang and Lin [3]); specifically, we consider all of the data sets for which $12\le n\le 1000$ and $256\le N\le 100000$. (For datasets with multiple versions, e.g., the {a1a, ..., a9a} data sets, we consider only the largest version.) The names of the data sets that we used and their sizes are given in Table 1. For the linear constraints, we generated random A and b for each problem. Specifically, the first $m=10$ rows of A and the first m entries in b were set as random values, with each entry being drawn from a standard normal distribution. Then, to ensure that the LICQ was not satisfied (at any algorithm iterate), we duplicated the last constraint, making $m+1$ linear constraints overall. For all problems and algorithms, the initial iterate was set to the vector of all ones of appropriate dimension.

For one set of experiments, we consider problems of the form (42), except without the norm constraint. For this set of experiments, the performance of all three algorithms—stochastic SQP, sub-gradient, and projected gradient—are compared. For each data set, we considered two noise levels, where the level is dictated by the mini-batch size of

Table 1. Names and sizes of data sets.

Data set	Dimension (n)	Data points (N)
a9a	123	32,561
australian	14	690
heart	13	270
ijcnn1	22	49,990
ionosphere	34	351
madelon	500	2,000
mushrooms	112	8,124
phising	68	11,055
sonar	60	208
splice	60	1,000
w8a	300	49,749

Source. Chang and Lin [3].

each stochastic gradient estimate (recall (26)). For the mini-batch sizes, we employed $b_k \in \{16, 128\}$ for all problems. For each data set and mini-batch size, we ran five instances with different random seeds.

A budget of five epochs (i.e., number of effective passes over the data set) was used for all methods. For our stochastic SQP method, we used $\beta_k = 10^{-1}$ for all $k \in \mathbb{N}$. For the stochastic sub-gradient method, the merit parameter and step size were tuned like in Section 5.2 over the sets $\beta \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ and $\tau \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. For the stochastic projected gradient method, the step size was tuned using the formula $\frac{\beta}{L}$ over $\beta \in \{10^{-8}, 10^{-7}, \ldots, 10^{1}, 10^{2}\}$. Overall, this meant that the stochastic sub-gradient and stochastic projected gradient methods were effectively run for 16 and 11 times the number of epochs, respectively, that were allowed for our method.

The results for this experiment are reported in Table 2. For every data set and mini-batch size, we report the average feasibility and stationarity errors for the best iterates of each run along with a 95% confidence interval. The results show that our method consistently outperforms the two alternative approaches despite the fact that each of the other methods was tuned with various choices of the merit and/or step size parameter. For a second set of experiments, we consider problems of the form (42) with the norm constraint. The settings for the experiment were the same as above, except that the stochastic projected gradient method is not considered. The results are stated in Table 3. Again, our method regularly outperforms the stochastic sub-gradient method in terms of the best iterates found. For the experiments without the norm constraint, our algorithm found $\tau_{k-1} \leq \tau_k^{\text{trial,true}}$ to hold in roughly 98% of all iterations for all runs, and it found this inequality to hold in all iterations in the last epoch in 100% of all runs. With the norm constraint, our algorithm found $\tau_{k-1} \leq \tau_k^{\text{trial,true}}$ to hold in roughly 97% of all iterations for all runs, and it found this inequality to hold in all iterations in the last epoch in 99% of all runs.

5.4. Physics-Informed Machine Learning

In our final set of experiments, we employ our stochastic SQP method to solve an example problem in physics-informed machine learning; see, for example, Lu et al. [21] and Négiar et al. [24] for other uses of "hard" constraints in this application area. Motivated by a problem in chemical engineering, we consider the problem to train a neural network to learn the solution of the following ODE, where $p : \mathbb{R} \to \mathbb{R}^4$, and $r \in \mathbb{R}^5$ is a given vector of rate constants:

$$\dot{p}_1 = -(r_1 + r_2 + r_4)p_1 + r_3 p_3 + r_5 p_4,\ \dot{p}_2 = 2r_1 p_1,\ \dot{p}_3 = r_2 p_1 - r_3 p_3, \text{ and } \dot{p}_4 = r_4 p_1 - r_5 p_4.$$

Table 2. Average feasibility and stationarity errors, along with 95% confidence intervals, when our stochastic SQP method, a stochastic subgradient method, and a stochastic projected gradient method are employed to solve logistic regression problems with linear constraints (only).

		Stochastic subgradient		Stochastic projected gradient	Stochastic SQP	
Data set	Batch	Feasibility	Stationarity	Stationarity	Feasibility	Stationarity
a9a	16	$8.30e-03 \pm 2.32e-03$	$1.64e-01 \pm 3.55e-03$	$3.64e-02 \pm 2.95e-03$	**$1.22e-15 \pm 2.18e-16$**	**$9.99e-03 \pm 6.92e-03$**
a9a	128	$1.16e-02 \pm 4.60e-05$	$1.69e-01 \pm 2.51e-02$	$1.69e-02 \pm 2.79e-03$	**$1.64e-15 \pm 4.00e-16$**	**$7.33e-03 \pm 4.68e-05$**
australian	16	$7.94e-02 \pm 1.60e-05$	$7.94e-02 \pm 1.60e-05$	$9.17e-02 \pm 4.32e-04$	**$5.72e-06 \pm 1.56e-06$**	**$2.67e-02 \pm 6.43e-04$**
australian	128	$5.02e-01 \pm 7.04e-05$	$5.02e-01 \pm 7.04e-05$	$1.11e-02 \pm 7.19e-05$	**$6.58e-05 \pm 7.90e-07$**	**$5.50e-02 \pm 1.08e-03$**
heart	16	$3.66e-01 \pm 4.37e-03$	$3.28e+01 \pm 7.02e+00$	**$3.17e+01 \pm 6.72e+00$**	**$8.83e-03 \pm 2.77e-03$**	$3.39e+01 \pm 9.85e+00$
heart	128	$1.52e+00 \pm 4.96e-02$	**$1.23e+01 \pm 1.40e+01$**	$3.29e+01 \pm 3.21e+00$	**$1.26e-01 \pm 7.86e-04$**	$3.24e+01 \pm 1.76e+00$
ijccn1	16	$3.58e-03 \pm 2.00e-05$	$4.70e-02 \pm 6.45e-07$	$7.41e-02 \pm 3.33e-07$	**$3.03e-15 \pm 6.20e-16$**	**$1.93e-03 \pm 4.07e-06$**
ijccn1	128	$3.90e-02 \pm 4.01e-06$	$5.17e-02 \pm 1.65e-07$	$3.88e-02 \pm 6.15e-07$	**$2.16e-09 \pm 2.62e-09$**	**$1.70e-02 \pm 5.19e-05$**
ionosphere	16	$5.41e-01 \pm 8.80e-05$	$5.41e-01 \pm 8.80e-05$	$9.77e-01 \pm 8.55e-03$	**$9.61e-07 \pm 2.77e-09$**	**$4.17e-02 \pm 1.08e-03$**
ionosphere	128	$5.76e+00 \pm 3.76e-05$	$5.76e+00 \pm 3.76e-05$	$5.98e+00 \pm 3.21e-03$	**$1.31e-05 \pm 1.14e-09$**	**$1.55e-01 \pm 2.61e-03$**
madelon	16	$3.06e-02 \pm 1.85e-02$	$5.46e+01 \pm 1.25e+01$	$2.11e+01 \pm 2.72e+00$	**$2.88e-08 \pm 5.51e-08$**	**$1.09e+01 \pm 3.00e+00$**
madelon	128	$1.87e+00 \pm 7.62e-01$	$2.21e+01 \pm 1.55e+01$	**$2.16e+01 \pm 4.17e+00$**	**$5.81e-01 \pm 1.63e-02$**	$4.81e+01 \pm 4.75e+00$
mushrooms	16	$2.19e-01 \pm 6.55e-04$	$2.19e-01 \pm 6.55e-04$	$7.31e-03 \pm 3.21e-06$	**$2.08e-15 \pm 3.28e-16$**	**$5.95e-03 \pm 3.21e-05$**
mushrooms	128	$4.73e-01 \pm 4.37e-05$	$4.73e-01 \pm 4.37e-05$	$3.31e-02 \pm 7.13e-05$	**$1.66e-09 \pm 6.20e-14$**	**$3.28e-02 \pm 9.15e-04$**
phishing	16	$2.67e-02 \pm 2.76e-07$	$3.47e-02 \pm 1.39e-09$	**$2.20e-05 \pm 9.29e-06$**	**$4.26e-15 \pm 1.27e-15$**	$3.37e-03 \pm 1.27e-06$
phishing	128	$3.06e-01 \pm 1.13e-06$	$3.06e-01 \pm 1.13e-06$	$2.29e-01 \pm 8.88e-03$	**$1.83e-15 \pm 4.99e-16$**	**$2.20e-02 \pm 7.29e-03$**
sonar	16	$1.33e+00 \pm 1.08e-04$	$1.33e+00 \pm 1.08e-04$	$6.13e-01 \pm 2.22e-03$	**$7.02e-07 \pm 1.60e-07$**	**$2.34e-02 \pm 2.03e-04$**
sonar	128	$1.33e+01 \pm 1.48e-04$	$1.33e+01 \pm 1.48e-04$	$6.46e-02 \pm 4.73e-03$	**$2.07e-06 \pm 6.70e-10$**	**$2.98e-02 \pm 1.71e-03$**
splice	16	$2.56e-03 \pm 3.39e-04$	$4.56e-01 \pm 3.55e-02$	$9.65e-01 \pm 3.19e-03$	**$7.49e-14 \pm 1.03e-13$**	**$2.19e-02 \pm 4.33e-03$**
splice	128	$3.14e-01 \pm 1.09e-04$	$4.83e-01 \pm 4.65e-05$	$1.23e+00 \pm 9.44e-05$	**$3.54e-08 \pm 5.74e-09$**	**$1.07e-02 \pm 3.16e-04$**
w8a	16	$2.38e-02 \pm 1.75e-03$	$1.47e-01 \pm 1.89e-06$	$9.85e-04 \pm 3.31e-05$	**$7.35e-15 \pm 6.98e-16$**	**$6.07e-05 \pm 6.46e-05$**
w8a	128	$1.79e-02 \pm 1.25e-03$	$1.49e-01 \pm 4.64e-03$	$3.41e-02 \pm 7.43e-03$	**$5.96e-15 \pm 5.67e-16$**	**$1.20e-03 \pm 1.85e-03$**

Note. The results for the best-performing algorithm are shown in bold.

Table 3. Average feasibility and stationarity errors, along with 95% confidence intervals, when our stochastic SQP method and a stochastic sub-gradient method are employed to solve logistic regression problems with linear constraints and a squared ℓ_2-norm constraint.

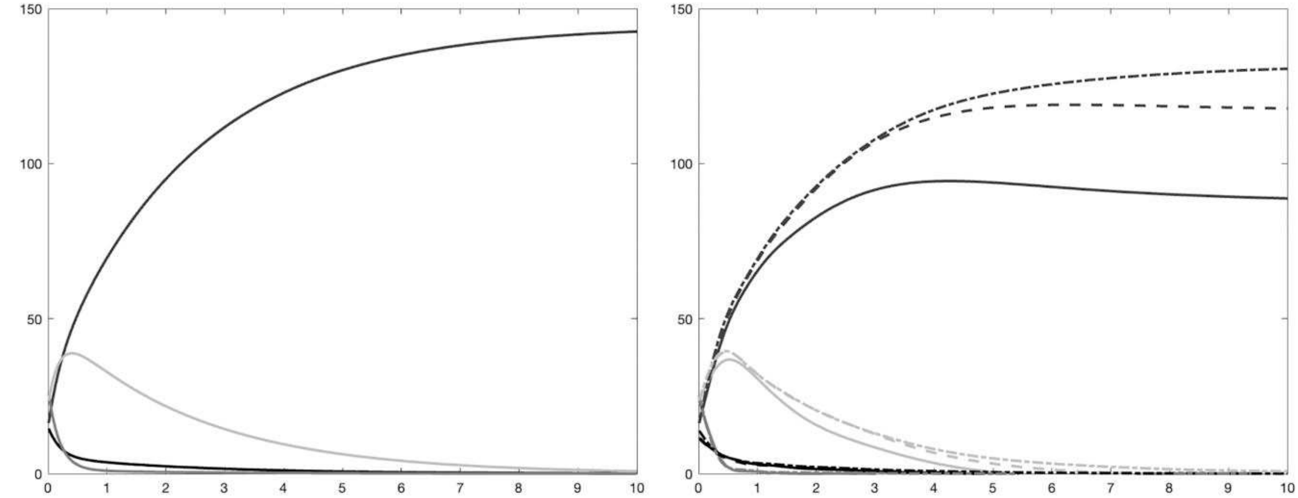
		Stochastic sub-gradient		Stochastic SQP	
Data set	Batch	Feasibility	Stationarity	Feasibility	Stationarity
`a9a`	16	$4.62e-03 \pm 3.27e-04$	$1.24e-01 \pm 7.52e-02$	$\mathbf{5.52e-05 \pm 5.04e-09}$	$\mathbf{6.07e-03 \pm 2.32e-05}$
`a9a`	128	$4.27e-03 \pm 3.92e-04$	$1.90e-01 \pm 3.03e-03$	$\mathbf{6.38e-05 \pm 1.12e-08}$	$\mathbf{4.40e-03 \pm 1.41e-05}$
`australian`	16	$1.51e-01 \pm 1.07e-05$	$1.51e-01 \pm 1.07e-05$	$\mathbf{1.52e-04 \pm 5.58e-06}$	$\mathbf{5.65e-03 \pm 3.73e-05}$
`australian`	128	$3.96e-01 \pm 1.87e-04$	$3.96e-01 \pm 1.87e-04$	$\mathbf{3.83e-04 \pm 5.45e-05}$	$\mathbf{1.68e-02 \pm 3.29e-03}$
`heart`	16	$1.57e+00 \pm 5.76e-01$	$2.86e+01 \pm 1.00e+01$	$\mathbf{9.29e-01 \pm 3.47e-02}$	$\mathbf{2.65e+01 \pm 1.81e+01}$
`heart`	128	$\mathbf{1.33e+00 \pm 6.69e-01}$	$\mathbf{1.69e+01 \pm 2.23e+00}$	$1.88e+00 \pm 1.42e-01$	$2.93e+00 \pm 1.26e+00$
`ijcnn1`	16	$5.36e-02 \pm 9.37e-07$	$5.36e-02 \pm 9.37e-07$	$\mathbf{3.70e-02 \pm 9.24e-05}$	$\mathbf{4.60e-02 \pm 8.32e-03}$
`ijcnn1`	128	$5.41e-02 \pm 1.04e-06$	$5.41e-02 \pm 1.04e-06$	$\mathbf{3.64e-02 \pm 1.06e-04}$	$\mathbf{3.64e-02 \pm 1.06e-04}$
`ionosphere`	16	$3.35e-01 \pm 1.06e-03$	$3.35e-01 \pm 1.06e-03$	$\mathbf{5.79e-03 \pm 1.44e-04}$	$\mathbf{1.21e-02 \pm 4.96e-03}$
`ionosphere`	128	$8.70e-01 \pm 1.43e-03$	$8.70e-01 \pm 1.43e-03$	$\mathbf{5.92e-03 \pm 2.18e-05}$	$\mathbf{4.31e-02 \pm 3.52e-04}$
`madelon`	16	$2.66e+00 \pm 6.84e-01$	$3.86e+01 \pm 3.28e+01$	$\mathbf{3.74e-01 \pm 8.55e-02}$	$\mathbf{4.70e-01 \pm 3.27e-02}$
`madelon`	128	$\mathbf{2.21e+01 \pm 4.90e-01}$	$\mathbf{4.77e+01 \pm 4.84e+00}$	$7.21e+01 \pm 5.28e+00$	$7.21e+01 \pm 5.28e+00$
`mushrooms`	16	$1.01e-01 \pm 5.79e-05$	$1.55e-01 \pm 8.22e-06$	$\mathbf{4.06e-04 \pm 8.76e-09}$	$\mathbf{4.65e-03 \pm 3.65e-05}$
`mushrooms`	128	$9.72e-01 \pm 9.94e-06$	$9.72e-01 \pm 9.94e-06$	$\mathbf{6.96e-04 \pm 1.52e-09}$	$\mathbf{3.34e-03 \pm 2.35e-07}$
`phishing`	16	$1.30e-01 \pm 1.61e-06$	$1.30e-01 \pm 1.61e-06$	$\mathbf{3.65e-05 \pm 2.44e-08}$	$\mathbf{8.17e-03 \pm 2.43e-05}$
`phishing`	128	$1.53e-01 \pm 3.37e-08$	$1.53e-01 \pm 3.37e-08$	$\mathbf{1.26e-04 \pm 3.30e-09}$	$\mathbf{8.45e-04 \pm 2.73e-07}$
`sonar`	16	$6.45e-01 \pm 5.62e-04$	$6.45e-01 \pm 5.62e-04$	$\mathbf{3.38e-03 \pm 8.81e-06}$	$\mathbf{1.48e-02 \pm 2.58e-04}$
`sonar`	128	$5.04e+00 \pm 4.44e-03$	$5.04e+00 \pm 4.44e-03$	$\mathbf{5.71e-03 \pm 8.61e-06}$	$\mathbf{2.16e-02 \pm 8.48e-05}$
`splice`	16	$\mathbf{1.96e-03 \pm 1.78e-04}$	$4.94e-01 \pm 7.35e-03$	$3.96e-03 \pm 7.12e-07$	$\mathbf{1.03e-02 \pm 1.14e-05}$
`splice`	128	$1.40e+00 \pm 7.90e-05$	$1.40e+00 \pm 7.90e-05$	$\mathbf{5.52e-03 \pm 3.72e-06}$	$\mathbf{1.04e-02 \pm 1.06e-04}$
`w8a`	16	$1.32e-02 \pm 6.83e-04$	$1.15e-01 \pm 1.33e-02$	$\mathbf{2.15e-04 \pm 2.24e-09}$	$\mathbf{1.83e-03 \pm 8.90e-07}$
`w8a`	128	$5.35e-02 \pm 7.79e-02$	$1.33e-01 \pm 1.74e-07$	$\mathbf{1.67e-04 \pm 6.01e-09}$	$\mathbf{1.00e-03 \pm 1.01e-06}$

Note. The results for the best-performing algorithm are shown in bold.

In our test problem, we employed $r = (4.283, 1.191, 5.743, 10.219, 1.535)$ and the initial condition $p(0) = q_0 := (14.546, 16.335, 25.947, 23.525)$. We designed our experiment with a neural network with an input layer of dimension 1 (time), a fully connected hidden layer with 2,048 nodes and tanh activation, and an output layer of dimension 4 (i.e., the dimension of the codomain of p). In the standard fashion of training a PINN (Cuomo et al. [5], Karniadakis et al. [17]), this neural network is appended with an additional layer encoding the known ODE. Before running our experiment, we pretrained the neural network to moderate accuracy using a stochastic gradient method to minimize the combined mean squared error of the residual of the ODE system over $t \in \{0, 10^{-3}, 2 \times 10^{-3}, \dots, 10\}$ (namely, 1,001 terms corresponding to input times to the neural network) and residual of the initial condition.

Our experiment involved training the network further to achieve higher accuracy, that is, a lower residual of the ODE system over the input times. For this, we ran two stochastic optimization methods starting with the parameters of the pretrained network. In one run, we continued using an unconstrained training paradigm (namely, the stochastic gradient method) to minimize mean squared error, as in the pretraining. In a second run, we ran StochasticSQP, where rather than including the initial conditions in the objective function, we imposed them as an explicit constraint, namely, that given the input $t = 0$ a forward pass over the network should yield $p(0) = q_0$. In addition, corresponding to the 20 input times in $\{0, 0.5, 1, \dots, 9, 9.5\}$, we included a mass-balance constraint $\sum_{i=1}^{4} w_i \dot{p}_i = 0$, where the weights $w = (1, 0.5, 1, 1)$ come from the application. (Our StochasticSQP code was used in both cases so that algorithmic quantities such as step sizes would be comparable for both the unconstrained and constrained optimization runs.) We observed that because of the structure of the problem, the constraint Jacobians were nearly rank-deficient throughout the optimization, as evidenced by the trust-region constraint in the normal component subproblem (5) regularly being active (with $\omega = 10^4$ for this experiment only). That said, by imposing the initial and mass-balance conditions as constraints, training more than 60,000 iterations improved the accuracy by 95.2% (i.e., the result of the second run), as opposed to improving the accuracy by 83.5%, which was achieved by running an additional 60,000 iterations of the unconstrained method (i.e., the result of the first run); see Figure 2. Much remains to be explored in the use of hard constraints for physics-informed machine learning and in particular the use of our proposed stochastic SQP method in such a context. However, this experiment both illustrates a type of real-world application for which methods such as ours might

Figure 2. True ODE solution (left) and predicted ODE solutions (right) yielded by the pretrained network (solid lines), by the network trained with additional iterations of the stochastic gradient method (dashed lines), and by the network trained with StochasticSQP iterations (dashed-dotted lines).



Notes. Most importantly, StochasticSQP employed to solve a constrained optimization problem outperforms the stochastic gradient method employed to solve an unconstrained problem in the prediction of p_2, which dominates the residual errors. The better performance of StochasticSQP continues if additional iterations of both solvers are performed.

be useful and motivates further investigations along these lines, such as in situations when the ODE itself is unknown, and one aims to train a neural network to discover the solution of a differential-equation system based on empirical observations, say, subject to constraints defined in terms of boundary conditions and/or known conservation laws.

6. Conclusion

We have proposed, analyzed, and tested a stochastic SQP method for solving equality-constrained optimization problems in which the objective function is defined by an expectation. Our algorithm is specifically designed for cases when the LICQ does not necessarily hold in every iteration. The convergence guarantees that we have proved for our method consider situations when the merit parameter sequence eventually remains fixed at a value that is sufficiently small (in which case the algorithm drives expected stationarity measures for the constrained optimization problem to zero), situations when the merit parameter vanishes (in which case the algorithm can drive a stationarity measure for minimizing constraint violation to zero, which might be all that is possible because the original optimization problem may be degenerate and/or infeasible), and situations when the merit parameter remains fixed at a value that is too large (which we show under modest assumptions is a probability-zero event). Numerical experiments demonstrate that our algorithm consistently outperforms alternative approaches in the highly stochastic regime.

We close by remarking on how our algorithm and analysis may be useful for the context of inequality-constrained optimization as well. An SQP algorithm with features similar to ours for solving inequality-constrained problems has been proposed and analyzed in Curtis et al. [8]. The analysis in that paper assumed that the constraint Jacobians have full row rank, but that work might be extendable to the rank-deficient setting with tools from our paper. Our step-decomposition technique is also of interest in the context of interior-point methods as well, even when rank deficiency of the constraint Jacobians (of nonlinear constraint functions) is not necessarily an issue. After all, as is well known, interior-point methods that compute search directions to satisfy a local linear model of the constraints can fail to converge (Wächter and Biegler [34]), and one manner in which this failure can be avoided is through the use of a step-decomposition technique. Therefore, our tools may be useful for the context of a stochastic interior-point method, say, in combination with the ideas in Curtis et al. [9].

Acknowledgments

The authors thank the editors and reviewers for their helpful comments and suggestions, and they thank Srinivas Rangarajan and Siddharth Prabhu (Lehigh University) for providing the motivating example in Section 5.4.

Appendix A: Deterministic Analysis

In this Appendix, we prove that Theorem 1 holds, where in particular we consider the context when $g_k = \nabla f(x_k)$ and $\beta_k = \beta$ satisfy (22) for all $k \in \mathbb{N}$. For this purpose, we introduce a second termination condition in Algorithm 1. In particular, after line 8, we terminate the algorithm if both $\|g_k + J_k^T y_k\|_2 = 0$ and $\|c_k\|_2 = 0$. In this manner, if the algorithm terminates finitely, then it returns an infeasible stationary point (recall (4)) or primal-dual stationary point for problem (1), and there is nothing left to prove. Hence, without loss of generality, we proceed under the assumption that the algorithm runs for all $k \in \mathbb{N}$. For our purposes in this section, let us refer to this as Assumption 4′.

Throughout our analysis in this Appendix, we simply refer to the tangential direction as u_k, the full search direction as $d_k = v_k + u_k$, etc., even though it is assumed throughout this Appendix that these are the true quantities computed using the true gradient $\nabla f(x_k)$ for all $k \in \mathbb{N}$.

It follows in this context that both Lemmas 1 and 2 hold. In addition, Lemma 3 holds, where, in the proof, the case that $d_k = 0$ can be ignored because of the following lemma.

Lemma A.1. *For all $k \in \mathbb{N}$, one finds that $d_k = v_k + u_k \neq 0$.*

Proof. For all $k \in \mathbb{N}$, $v_k \in \text{Range}(J_k^T)$ and $u_k \in \text{Null}(J_k)$ imply that $d_k = v_k + u_k = 0$ if and only if $v_k = 0$ and $u_k = 0$. Under Assumption 4′, it follows for all $k \in \mathbb{N}$ that $\|J_k^T c_k\|_2 > 0$ or $\|c_k\|_2 = 0$. If $\|J_k^T c_k\|_2 > 0$, then Lemma 1 implies that $v_k \neq 0$, and the desired conclusion follows. Hence, we may proceed under the assumption that $\|c_k\|_2 = 0$. In this case, it follows under Assumption 3 that $g_k + J_k^T y_k = 0$ if and only if $u_k = 0$, which by Assumption 4′ means that $u_k \neq 0$. □

We now prove a lower bound on the reduction in the merit function that occurs in each iteration. This is a special case of Lemmas 6 and 10 for the deterministic setting.

Lemma A.2. *For all $k \in \mathbb{N}$, it holds that $\phi(x_k, \tau_k) - \phi(x_k + \alpha_k d_k, \tau_k) \geq \eta \alpha_k \Delta l(x_k, \tau_k, g_k, d_k)$.*

Proof. For all $k \in \mathbb{N}$, it follows by the definition of α_k^{suff} that (recall (19))

$$\phi(x_k + \alpha d_k, \tau_k) - \phi(x_k, \tau_k) \leq -\eta \alpha \Delta l(x_k, \tau_k, g_k, d_k) \text{ for all } \alpha \in [0, \alpha_k^{\text{suff}}].$$

If $\|u_k\|_2^2 \geq \chi_k \|v_k\|_2^2$, then the only way that $\alpha_k > \alpha_k^{\text{suff}}$ is if

$$\frac{2(1-\eta)\beta\xi_k\tau_k}{\tau_k L + \Gamma} > \min\left\{\frac{2(1-\eta)\beta\Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2}, 1\right\}.$$

By (22), the left-hand side of this inequality is less than 1, meaning that $\alpha_k > \alpha_k^{\text{suff}}$ only if

$$\frac{2(1-\eta)\beta\xi_k\tau_k}{\tau_k L + \Gamma} > \frac{2(1-\eta)\beta\Delta l(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \Longleftrightarrow \xi_k\tau_k > \frac{\Delta l(x_k, \tau_k, g_k, d_k)}{\|d_k\|_2^2}.$$

However, this is not true because $\xi_k \leq \xi_k^{\text{trial}}$ for all $k \in \mathbb{N}$. Following a similar argument for the case when $\|u_k\|_2^2 < \chi_k \|v_k\|_2^2$, the desired conclusion follows. □

For our purposes going forward, let us define the shifted merit function $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ by

$$\tilde{\phi}(x, \tau) = \tau(f(x) - f_{\text{inf}}) + \|c(x)\|_2.$$

Lemma A.3. *For all $k \in \mathbb{N}$, it holds that $\tilde{\phi}(x_k, \tau_k) - \tilde{\phi}(x_{k+1}, \tau_{k+1}) \geq \eta \alpha_k \Delta l(x_k, \tau_k, g_k, d_k)$.*

Proof. For arbitrary $k \in \mathbb{N}$, it follows from Lemma A.2 that

$$\begin{aligned}\tau_{k+1}(f(x_k + \alpha_k d_k) - f_{\text{inf}}) + \|c(x_k + \alpha_k d_k)\|_2 &\leq \tau_k(f(x_k + \alpha_k d_k) - f_{\text{inf}}) + \|c(x_k + \alpha_k d_k)\|_2 \\ &\leq \tau_k(f(x_k) - f_{\text{inf}}) + \|c_k\|_2 - \eta\alpha_k \Delta l(x_k, \tau_k, g_k, d_k),\end{aligned}$$

from which the desired conclusion follows. □

We now prove our first main result of this Appendix.

Lemma A.4. *The sequence $\{\|J_k^T c_k\|_2\}$ vanishes. Moreover, if there exist $k_J \in \mathbb{N}$ and $\sigma_J \in \mathbb{R}_{>0}$ such that the singular values of J_k are bounded below by σ_J for all $k \geq k_J$, then $\{\|c_k\|_2\}$ vanishes.*

Proof. Let $\gamma \in \mathbb{R}_{>0}$ be arbitrary. Our aim is to prove that the number of iterations with $x_k \in \mathcal{X}_\gamma$ (recall (32)) is finite. Because γ has been chosen arbitrarily in $\mathbb{R}_{>0}$, the conclusion will follow. By Lemma 12 and the fact that $\{\beta_k\}$ is chosen as a constant sequence, it follows that there exists $\underline{\alpha} \in \mathbb{R}_{>0}$ such that $\alpha_k \geq \underline{\alpha}$ for all $k \in \mathcal{K}_\gamma$ (regardless of whether the search direction is tangentially or normally dominated). Hence, using Lemmas 1 and A.3, it follows that

$$\tilde{\phi}(x_k, \tau_k) - \tilde{\phi}(x_{k+1}, \tau_{k+1}) \geq \eta\underline{\alpha}\Delta l(x_k, \tau_k, g_k, d_k) \geq \eta\underline{\alpha}\sigma(\|c_k\|_2 - \|c_k + J_k v_k\|_2) \geq \eta\underline{\alpha}\sigma\kappa_v\kappa_c^{-1}\gamma^2.$$

Hence, the desired conclusion follows because $\{\tilde{\phi}(x_k, \tau_k)\}$ is monotonically nonincreasing by Lemma A.3 and is bounded below under Assumption 1. □

We now show a consequence of the merit parameter eventually remaining constant.

Lemma A.5. *If there exists $k_\tau \in \mathbb{N}$ and $\tau_{\min} \in \mathbb{R}_{>0}$ such that $\tau_k = \tau_{\min}$ for all $k \geq k_\tau$, then*

$$0 = \lim_{k\to\infty} \|u_k\|_2 = \lim_{k\to\infty} \|d_k\|_2 = \lim_{k\to\infty} \|g_k + J_k^T y_k\|_2 = \lim_{k\to\infty} \|Z_k^T g_k\|_2.$$

Proof. Under Assumption 1 and the conditions of the lemma, Lemmas 12 and A.3 imply that $\Delta l(x_k, \tau_k, g_k, d_k)\} \to 0$, which with (13) and Lemma 1 implies that $\{\|u_k\|_2\} \to 0$, $\{\|v_k\|_2\} \to 0$, and $\{\|J_k^T c_k\|_2\} \to 0$. The remainder of the conclusion follows from Assumption 3 and (8). □

The proof of Theorem 1 can now be completed.

Proof of Theorem 1. The result follows from Lemmas A.4 and A.5 along with Theorem B.1 (considered in the deterministic setting, namely, in the case that $\kappa_g = 0$ in that theorem).

Appendix B: Bounded Merit Parameter

Our goal in this section is to prove that there exist conditions under which the merit parameter sequence remains bounded below by a positive constant, which by the manner that the merit parameter sequence is determined means that there are conditions under which the sequence eventually remains constant. As seen in Theorem 1, it is worthwhile to consider such an occurrence regardless of the properties of the sequence of constraint Jacobians. That said, one might be able to prove only that it occurs when the constraint Jacobians are (eventually) bounded away from singularity over a run of the algorithm.

Our first lemma for these purposes proves that if the constraint Jacobians are bounded away from singularity by a uniform constant, then the normal components of the search directions satisfy a useful upper bound. (One could prove such a property for all sufficiently large iteration indices if the Jacobians are eventually bounded away from singularity, but for simplicity we consider the case when they are bounded away for all $k \in \mathbb{N}$.) The proof is essentially the same as that of lemma 3.15 in Curtis et al. [6], but we provide it for completeness. In the lemma, the constant $\sigma_J \in \mathbb{R}_{>0}$ is presumed to be uniform over all realizations of a run of the algorithm.

Lemma B.1. *Consider $\sigma_J \in \mathbb{R}_{>0}$. If the singular values of J_k are bounded below by σ_J for all $k \in \mathbb{N}$, then there exists $\kappa_\omega \in \mathbb{R}_{>0}$ (uniform over all runs) such that*

$$\|V_k\|_2 \leq \kappa_\omega(\|C_k\|_2 - \|C_k + J_k V_k\|_2) \textit{ for all } k \in \mathbb{N}.$$

Proof. Under the conditions of the lemma, $\|J_k^T C_k\|_2 \geq \sigma_J \|C_k\|_2$ for all $k \in \mathbb{N}$. Hence, along with Lemma 1, it follows that $\|C_k\|_2(\|C_k\|_2 - \|C_k + J_k V_k\|_2) \geq \kappa_v \|J_k^T C_k\|_2^2 \geq \kappa_v \sigma_J^2 \|C_k\|_2^2$ for all $k \in \mathbb{N}$. Combining this again with Lemma 1, it follows with the Cauchy-Schwarz inequality and (2) that

$$\|V_k\|_2 \leq \omega \|J_k^T\|_2 \|C_k\|_2 \leq \frac{\omega \kappa_J}{\kappa_v \sigma_J^2}(\|C_k\|_2 - \|C_k + J_k V_k\|_2) \text{ for all } k \in \mathbb{N},$$

from which the desired conclusion follows. □

We now prove that if the differences between the stochastic gradient estimators and the true gradients are bounded, then the sequence of tangential components is bounded. As in the previous lemma, the constant $\kappa_g \in \mathbb{R}_{>0}$ is presumed to be a uniform bound.

Lemma B.2. *Consider $\kappa_g \in \mathbb{R}_{>0}$. If the sequence $\{\|G_k - \nabla f(X_k)\|_2\}$ is bounded by κ_g, then there exists $\kappa_u \in \mathbb{R}_{>0}$ (uniform over all runs) such that $\|U_k\|_2 \leq \kappa_u$ for all $k \in \mathbb{N}$.*

Proof. Under Assumption 1, $\{\|\nabla f(X_k)\|_2\}$ is bounded; recall (2). Hence, under the conditions of the lemma, $\{\|G_k\|_2\}$ is bounded. The first block of (8) yields $U_k^T H_k U_k = -U_k^T(G_k + H_k V_k)$, which under Assumption 3 yields $\rho \|U_k\|_2^2 \leq -U_k^T G_k - U_k^T H_k V_k \leq (\|G_k\|_2 + \|H_k\|_2 \|V_k\|_2)\|U_k\|_2$. Hence, the conclusion follows from these facts, Assumption 1, Assumption 3, and Lemma 1. □

By combining the preceding two lemmas, the following theorem indicates certain circumstances under which the sequence of merit parameters will eventually remain constant. We remark that it is possible in a run of the algorithm for the merit parameter sequence to remain constant (eventually) even if the conditions of the theorem do not hold, which is why our analyses in the main body of the paper do not presume that these conditions hold. That said, to prove that there exist settings in which the merit parameter is guaranteed to remain constant eventually, we offer the theorem. In the theorem, it is important to note that the constant $\tau_{\min} \in \mathbb{R}_{>0}$ is uniform over all runs despite the fact that the iteration in which the merit parameter experiences its last decrease and the final value at which it settles might be run-dependent.

Theorem B.1. *Consider $\sigma_J \in \mathbb{R}_{>0}$ and $\kappa_g \in \mathbb{R}_{>0}$. If the singular values of J_k are bounded below by σ_J for all $k \in \mathbb{N}$ and $\{\|G_k - \nabla f(X_k)\|_2\}$ is bounded by κ_g, then there exists $\tau_{\min} \in \mathbb{R}_{>0}$ (uniform over all runs) such that for some $K_\tau \in \mathbb{N}$ and $\mathcal{T}' \in \mathbb{R}_{>0}$, one finds*

$$\mathcal{T}_k = \mathcal{T}' \geq \tau_{\min} \text{ for all } k \in \mathbb{N} \text{ with } k \geq K_\tau.$$

Proof. The algorithm terminates if $\|J_k^T C_k\|_2 = 0$ while $\|C_k\|_2 > 0$. Let us show that if $\|C_k\|_2 = 0$, then the algorithm sets $\mathcal{T}_k \leftarrow \mathcal{T}_{k-1}$. Indeed, $\|C_k\|_2 = 0$ implies $V_k = 0$ by Lemma 1. If $U_k = 0$ as well, then $D_k = 0$, and the algorithm explicitly sets $\mathcal{T}_k \leftarrow \mathcal{T}_{k-1}$. Otherwise, if $V_k = 0$ and $U_k \neq 0$, then (8) yields $0 = G_k^T U_k + U_k^T H_k U_k = G_k^T D_k + U_k^T H_k U_k$, in which case (10)–(11) again yield $\mathcal{T}_k \leftarrow \mathcal{T}_{k-1}$. Overall, it follows that $\mathcal{T}_k < \mathcal{T}_{k-1}$ if and only if one finds $\|J_k^T C_k\|_2 > 0$, $G_k^T D_k + U_k^T H_k U_k > 0$, and $\mathcal{T}_{k-1}(G_k^T D_k + U_k^T H_k U_k) > (1-\sigma)(\|C_k\|_2 - \|C_k + J_k V_k\|_2)$. On the other hand, from the first equation in (8), the Cauchy-Schwarz inequality, (2), and Lemmas B.1 and B.2, it holds that

$$\begin{aligned} G_k^T D_k + U_k^T H_k U_k = (G_k - H_k U_k)^T V_k &= (G_k - \nabla f(X_k) + \nabla f(X_k) - H_k U_k)^T V_k \\ &\leq (\kappa_g + \kappa_{\nabla f} + \kappa_H \kappa_u)\|V_k\|_2 \\ &\leq (\kappa_g + \kappa_{\nabla f} + \kappa_H \kappa_u)\kappa_\omega(\|C_k\|_2 - \|C_k + J_k V_k\|_2). \end{aligned}$$

Combining these facts, the desired conclusion follows. □

Appendix C: Insufficiently Small Merit Parameter

In this Appendix, we prove a formal version of Theorem 4, which is stated at the end of this Appendix as Theorem C.1. In the process of proving this main result (Theorem C.1), we prove a set of lemmas about the behavior of the merit parameter sequence after a finite number of iterations. Specifically, we consider the behavior of Algorithm 1 when it is terminated at iteration $k_{\max} \in \mathbb{N}$. With this consideration, we define a tree with a depth bounded by $k_{\max}$, which will be integral to our arguments in this section. The proof of Theorem C.1 ultimately considers the behavior of the algorithm as $k_{\max} \to \infty$.

Let $\mathcal{I}[\cdot]$ denote the indicator function of an event, and for all $k \in [k_{\max}]$, define

$$Q_k := \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \text{ and } W_k := \sum_{i=0}^{k-1} \mathcal{I}[\mathcal{T}_i < \mathcal{T}_{i-1}].$$

Accordingly, for any realization of a run of Algorithm 1 and any $k \in [k_{\max}]$, the realization (q_k, w_k) of (Q_k, W_k) is determined at the beginning of iteration k. The *signature* of a realization up to iteration k is $(q_0, \ldots, q_k, w_0, \ldots, w_k)$, which encodes all of the pertinent information regarding the behavior of the merit parameter sequence and these indicators up to the start of iteration k.

We use the set of all signatures to define a tree whereby each node contains a subset of all realizations of the algorithm. To construct the tree, we denote the root node by $N(q_0, w_0)$, where q_0 is the indicator of the event $\tau_0^{\text{trial, true}} < \tau_{-1}$, which is determined based on the initial conditions of the algorithm, and $w_0 = 0$. All realizations of the algorithm follow the same initialization, so q_0 and w_0 are in the signature of every realization. Next, we define a node $N(q_{[k]}, w_{[k]})$ 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 the algorithm for which the signature of the realization up to iteration k is $(q_0, \ldots, q_k, w_0, \ldots, w_k)$. We define the edges in the tree by connecting nodes at neighboring levels, where node $N(q_{[k]}, w_{[k]})$ is connected to node $N(q_{[k]}, q_{k+1}, w_{[k]}, w_{k+1})$ for any $q_{k+1} \in \{0, 1\}$ and $w_{k+1} \in \{w_k, w_k + 1\}$.

Because the behavior of a realization of the algorithm up to iteration $k \in \mathbb{N}$ is 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 $N(q_{[k]}, w_{[k]})$ by writing that $G_{[k-1]} \in N(q_{[k]}, w_{[k]})$. The initial condition, denoted for consistency as $G_{[-1]} \in N(q_0, w_0)$, occurs surely. Based on the description above, the nodes of our tree satisfy the property that for any $k \geq 2$, the event $G_{[k-1]} \in N(q_{[k]}, w_{[k]})$ occurs if and only if

$$Q_k = q_k, \; W_k = w_k, \text{ and } G_{[k-2]} \in N(q_{[k-1]}, w_{[k-1]}). \tag{C.1}$$

Under event $E_\tau(\tau_{\min}^{\text{trial,true}})$ for $\tau_{\min}^{\text{trial,true}} \in \mathbb{R}_{>0}$, one has that $W_{\bar{k}} = \sum_{i=0}^{\bar{k}-1} \mathcal{I}[\mathcal{T}_i < \mathcal{T}_{i-1}] \geq \bar{s}(\tau_{\min}^{\text{trial,true}})$ (see (36)) implies that for all subsequent k one finds that $\mathcal{T}_{k-1} \leq \tau_{\min}^{\text{trial,true}} \leq \mathcal{T}_k^{\text{trial, true}}$.

We are now prepared to state the assumption under which Theorem C.1 is proved. It is similar to Assumption 7 combined with the assumption that (37a) holds for some fixed probability conditioned on the event that $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}$.

Assumption C.1. *For some $\tau_{\min}^{\text{trial,true}} \in \mathbb{R}_{>0}$, the event $E_\tau(\tau_{\min}^{\text{trial,true}})$ occurs. In addition, there exists $p_\tau \in (0,1]$ such that, for all $k \in [k_{\max}]$, one finds*

$$\mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} \mid E_\tau(\tau_{\min}^{\text{trial,true}}), G_{[k-1]} \in N(q_{[k]}, w_{[k]}), \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \geq p_\tau. \tag{C.2}$$

Intuitively, equation (C.2) states that conditioned on $E_\tau(\tau_{\min}^{\text{trial,true}})$, the behavior of the algorithm up to the beginning of iteration k, and $Q_k = 1$, the probability that the merit parameter is decreased in iteration k, is at least p_τ. For simplicity of notation, henceforth we define E as the event that Assumption C.1 holds, and $s_{\max} := \bar{s}(\tau_{\min}^{\text{trial,true}})$. We remark in passing that Lemma 13 is relevant here. In particular, one could apply the same arguments as in the proof of Lemma 13 to show that if

$$\mathbb{P}[G_k^T D_k + U_k^T H_k U_k \geq \nabla f(X_k)^T D_k^{\text{true}} + (U_k^{\text{true}})^T H_k U_k^{\text{true}} \mid E_\tau(\tau_{\min}^{\text{trial,true}}), G_{[k-1]} \in N(q_{[k]}, w_{[k]}), \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \geq p_\tau, \tag{C.3}$$

then (C.2) holds. Hence, (8) could have been written with respect to (C.3), which is an inequality that holds when the distribution of the stochastic gradients satisfies a mild form of symmetry; see example 3.17 in Berahas et al. [1] for a simple example. However, because (C.2) is the inequality that is used in our analysis below, we include this inequality directly in the assumption.

Our main result, Theorem C.1, essentially shows that the probability that $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k$ occurs infinitely often is zero. Toward proving this result, we first prove a bound on the probability that $\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k$ occurs at least J times for any $J \in \mathbb{N}$ such that $J > \frac{s_{\max}}{p_\tau} + 1$. Given such a J, we can define a number of important sets of nodes in the tree. First, let

$$\mathcal{L}_{\text{good}} := \left\{ N(q_{[k]}, w_{[k]}) : \left(\sum_{i=0}^{k} q_i < J \right) \wedge (w_k = s_{\max} \vee k = k_{\max}) \right\}$$

be the set of nodes at which the sum of the elements of $q_{[k]}$ is sufficiently small (less than J), and either w_k has reached $s_{\max}$ or k has reached $k_{\max}$. Second, let

$$\mathcal{L}_{\text{bad}} := \left\{ N(p_{[k]}, w_{[k]}) : \sum_{i=0}^{k} q_i \geq J \right\}$$

be the nodes in the complement of $\mathcal{L}_{\text{good}}$ at which the sum of the elements of $q_{[k]}$ is at least J. 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 $\mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$. From this restriction and the definitions of $\mathcal{L}_{\text{good}}$ and $\mathcal{L}_{\text{bad}}$, the tree has finite depth, 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

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

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

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

Under these definitions, the paths down the tree terminate at nodes in $\mathcal{L}_{\text{good}} \cup \mathcal{L}_{\text{bad}}$, reaffirming that these nodes are the leaves of the tree. For convenience in the rest of our discussions, let $C(\emptyset) = \emptyset$.

We define the height of node $N(q_{[k]}, w_{[k]})$ as the length of the longest path from $N(q_{[k]}, w_{[k]})$ to a leaf node; that is, the height is denoted as

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

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

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

$$\sum_{i=0}^{j} Q_i = \sum_{i=0}^{j} \mathcal{I}[\mathcal{T}_i^{\text{trial, true}} < \mathcal{T}_{i-1}] \geq J. \tag{C.4}$$

Our first goal in this section is to find a bound on the probability of this event occurring. We will then utilize this bound to prove Theorem C.1. As a first step toward bounding the probability of $E_{\text{bad},k_{\max},J}$, we prove the following result about the leaf nodes of the tree.

Lemma C.1. *For any $k \in [k_{\max}]$, $J \in \mathbb{N}$, and $(q_{[k]}, w_{[k]})$ with $N(q_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{good}}$, one finds*

$$\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] = 0.$$

On the other hand, for any $k \in [k_{\max}]$, $J \in \mathbb{N}$, and $(q_{[k]}, w_{[k]})$ with $N(q_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{bad}}$, one finds

$$\begin{aligned} &\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\ &\leq \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]). \end{aligned}$$

Proof. Consider arbitrary $k \in [k_{\max}]$ and $J \in \mathbb{N}$ as well as an arbitrary pair $(q_{[k]}, w_{[k]})$ such that $N(q_{[k]}, w_{[k]}) \in \mathcal{L}_{\text{good}}$. By the definition of $\mathcal{L}_{\text{good}}$, it follows that $\sum_{i=0}^{k} q_i < J$. Then, by (C.1),

$$\mathbb{P}\left[\sum_{i=0}^{k} Q_i \geq J | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \right] = \mathbb{P}\left[\sum_{i=0}^{k} q_i \geq J \middle| E, G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \right] = 0.$$

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

$$\begin{aligned} \mathbb{P}[G_{[j-1]} \in N(q_{[j]}, w_{[j]}) \wedge \text{(C.4) holds} | E] &= \mathbb{P}[\text{(C.4) holds} | E, G_{[j-1]} \in N(q_{[j]}, w_{[j]})] \\ &\quad \cdot \mathbb{P}[G_{[j-1]} \in N(q_{[j]}, w_{[j]}) | E] = 0. \end{aligned}$$

In addition, (C.4) cannot hold for $j=0$ because $\mathcal{I}[\tau_0^{\text{trial, true}} < \tau_{-1}] = q_0 < J$ by the definition of $\mathcal{L}_{\text{good}}$. Hence, along with the conclusion above, it follows that $E_{\text{bad},k_{\max},J}$ 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 to $N(q_{[k]},w_{[k]})$. 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

$$\mathbb{P}\left[\sum_{i=0}^{k_{\max}} Q_i < J \middle| E, G_{[k-1]} \in N(p_{[k]},w_{[k]})\right] = 1.$$

Otherwise, if $w_k = s_{\max}$, then it follows by the definition of $s_{\max}$ that $\mathcal{T}_{k-1} \le \tau_{\min}^{\text{trial, true}}$ so that $Q_i = \mathcal{I}[\mathcal{T}_i^{\text{trial, true}} < \mathcal{T}_{i-1}] = 0$ holds for all $i \in \{k,\ldots,k_{\max}\}$, and therefore, the equation above again follows. Overall, it follows that $\mathbb{P}[G_{[k-1]} \in N(q_{[k]},w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] = 0$, as desired.

Now consider arbitrary $k \in [k_{\max}]$ and $J \in \mathbb{N}$ as well as an arbitrary pair $(q_{[k]},w_{[k]})$ with $N(q_{[k]},w_{[k]}) \in \mathcal{L}_{\text{bad}}$. One finds that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]},w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
\le\,&\mathbb{P}[G_{[k-1]} \in N(q_{[k]},w_{[k]}) | E] = \mathbb{P}[(\text{C.1}) \text{ holds} | E] \\
=\,&\mathbb{P}[Q_k = q_k | E, W_k = w_k, G_{[k-2]} \in N(q_{[k-1]},w_{[k-1]})] \cdot \mathbb{P}[W_k = w_k \wedge G_{[k-2]} \in N(p_{[k-1]},w_{[k-1]}) | E] \\
=\,&\mathbb{P}[Q_k = q_k | E, W_k = w_k, G_{[k-2]} \in N(q_{[k-1]},w_{[k-1]})] \\
&\cdot \mathbb{P}[W_k = w_k | E, G_{[k-2]} \in N(p_{[k-1]},w_{[k-1]})] \cdot \mathbb{P}[G_{[k-2]} \in N(p_{[k-1]},w_{[k-1]}) | E] \\
=\,&\mathbb{P}[G_{-1} \in N(q_0,w_0)] \\
&\cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]},w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]},w_{[i-1]})]),
\end{aligned}$$

which, because $\mathbb{P}[G_{[-1]} \in N(q_0,w_0)] = 1$, proves the remainder of the result. □

Next, we show that the probability of the occurrence of $E_{\text{bad},k_{\max},J}$ at any node in the tree can be bounded in terms of the probability of the sum of a set of independent Bernoulli random variables being less than a threshold defined by $s_{\max}$.

Lemma C.2. *For any $k \in [k_{\max}]$, $J \in \mathbb{N}$, and $(q_{[k]},w_{[k]})$ with $N(q_{[k]},w_{[k]}) \notin \mathcal{L}_{\text{good}}$, let*

$$\psi_J(q_{[k]}) = J - 1 - \sum_{i=0}^{k} q_i. \tag{C.5}$$

One finds that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]},w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
&\le \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]},w_{[i-1]})] \\
&\quad \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]},w_{[i-1]})]) \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k-1]})} Z_j \le s_{\max} - w_k\right],
\end{aligned} \tag{C.6}$$

where $\{Z_j\}$ are independent Bernoulli random variables with $\mathbb{P}[Z_j = 1] = p_\tau$ for all $j \in \mathbb{N}$.

Proof. Consider any $(q_{[k]},w_{[k]})$ with $h(N(q_{[k]},w_{[k]})) = 0$. Because $N(q_{[k]},w_{[k]}) \notin \mathcal{L}_{\text{good}}$, it follows that $N(q_{[k]},w_{[k]}) \in \mathcal{L}_{\text{bad}}$. Then, by the definition of $\mathcal{L}_{\text{bad}}$, it follows that $\sum_{i=0}^{k} q_i \ge J$. In addition, because $C(N(q_{[k]},w_{[k]})) = \emptyset$ for any node in $\mathcal{L}_{\text{bad}}$, it follows that $P(N(q_{[k]},w_{[k]})) \notin \mathcal{L}_{\text{bad}}$, which implies that $\sum_{i=0}^{k-1} q_i < J$. Thus, $\sum_{i=0}^{k} q_i = J$ and $\sum_{i=0}^{k-1} q_i = J-1$, which implies that $\psi_J(q_{[k-1]}) = 0$. Therefore, overall, the result holds for any $(q_{[k]},w_{[k]})$ with $h(N(q_{[k]},w_{[k]})) = 0$ by Lemma C.1.

We prove the rest of the result by induction on the height of the node. We note that the base case, that is, when $h(N(q_{[k]},w_{[k]})) = 0$, holds by the above argument. Now, assume that (C.6) holds for any $(q_{[k]},w_{[k]})$ with $N(q_{[k]},w_{[k]}) \notin \mathcal{L}_{\text{good}}$ such that $h(N(q_{[k]},w_{[k]})) \le \hat{h}$. Consider arbitrary $(q_{[k]},w_{[k]})$ such that $N(q_{[k]},w_{[k]}) \notin \mathcal{L}_{\text{good}}$ and $h(N(q_{[k]},w_{[k]})) = \hat{h}+1$. By the definition of C, one finds

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]},w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
&= \sum_{\{(q_{k+1},w_{k+1}) : N(q_{[k+1]},w_{[k+1]}) \in C(N(q_{[k]},w_{[k]}))\}} \mathbb{P}[G_{[k]} \in N(q_{[k+1]},w_{[k+1]}) \wedge E_{\text{bad},k_{\max},J} | E].
\end{aligned}$$

Then, by the definition of $q_{[k]}$ and $w_{[k]}$, we can enumerate the children of $N(q_{[k]}, w_{[k]})$ as

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
&= \mathbb{P}[G_{[k]} \in N(q_{[k]}, 0, w_{[k]}, w_k) \wedge E_{\text{bad},k_{\max},J} | E] + \mathbb{P}[G_{[k]} \in N(q_{[k]}, 0, w_{[k]}, w_k + 1) \wedge E_{\text{bad},k_{\max},J} | E] \\
&\quad + \mathbb{P}[G_{[k]} \in N(q_{[k]}, 1, w_{[k]}, w_k) \wedge E_{\text{bad},k_{\max},J} | E] + \mathbb{P}[G_{[k]} \in N(q_{[k]}, 1, w_{[k]}, w_k + 1) \wedge E_{\text{bad},k_{\max},J} | E].
\end{aligned}$$

Now, noting that all children of $N(q_{[k]}, w_{[k]})$ have a height that is at most $\hat{h}$, we apply the induction hypothesis four times to obtain

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
&\leq \Bigg(\mathbb{P}[Q_{k+1} = 0 | E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&\cdot \mathbb{P}[W_{k+1} = w_k | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \leq s_{\max} - w_k\right] \\
&+ \mathbb{P}[Q_{k+1} = 0 | E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&\cdot \mathbb{P}[W_{k+1} = w_k + 1 | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \leq s_{\max} - w_k - 1\right] \\
&+ \mathbb{P}[Q_{k+1} = 1 | E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&\cdot \mathbb{P}[W_{k+1} = w_k | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,3} \leq s_{\max} - w_k\right] \\
&+ \mathbb{P}[Q_{k+1} = 1 | E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&\cdot \mathbb{P}[W_{k+1} = w_k + 1 | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,4} \leq s_{\max} - w_k - 1\right] \Bigg) \\
&\cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})])
\end{aligned}$$

where $Z_{j,p}$ for all $p \in \{1, \ldots, 4\}$ and $j \in \{1, \ldots, \psi(q_{[k]})\}$ are four sets of independent Bernoulli random variables with $\mathbb{P}[Z_{j,p} = 1] = p_\tau$. Now, by the definitions of $Z_{j,1}$, $Z_{j,2}$, $Z_{j,3}$, and $Z_{j,4}$, it follows that

$$\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \leq s_{\max} - w_k\right] = \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,3} \leq s_{\max} - w_k\right],$$

and

$$\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \leq s_{\max} - w_k - 1\right] = \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,4} \leq s_{\max} - w_k - 1\right].$$

Therefore, it follows that

$$\begin{aligned}
&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E] \\
&\leq \Bigg((\mathbb{P}[Q_{k+1} = 0 | E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&+ \mathbb{P}[Q_{k+1} = 1 | E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})]) \\
&\cdot \mathbb{P}[W_{k+1} = w_k | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \leq s_{\max} - w_k\right] \\
&+ (\mathbb{P}[Q_{k+1} = 0 | E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \\
&+ \mathbb{P}[Q_{k+1} = 1 | E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})]) \\
&\cdot \mathbb{P}[W_{k+1} = w_k + 1 | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \leq s_{\max} - w_k - 1\right] \Bigg) \\
&\cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).
\end{aligned}$$

Now, by the law of total probability, it follows that

$$\begin{aligned}1 = \mathbb{P}[Q_{k+1} = 0 \mid E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})]\\ + \mathbb{P}[Q_{k+1} = 1 \mid E, W_{k+1} = w_k, G_{[k-1]} \in N(q_{[k]}, w_{[k]})],\end{aligned}$$

and

$$\begin{aligned}1 = \mathbb{P}[Q_{k+1} = 0 \mid E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})]\\ + \mathbb{P}[Q_{k+1} = 1 \mid E, W_{k+1} = w_k + 1, G_{[k-1]} \in N(q_{[k]}, w_{[k]})].\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad}, k_{\max}, J} \mid E]\\ &\le \left(\mathbb{P}[W_{k+1} = w_k \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k\right]\right.\\ &\quad \left. + \mathbb{P}[W_{k+1} = w_k + 1 \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] \cdot \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \le s_{\max} - w_k - 1\right]\right)\\ &\quad \cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i \mid E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i \mid E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).\end{aligned} \tag{C.7}$$

We proceed by considering two cases. First, suppose $q_k = 1$. By Assumption C.1, it follows that

$$\begin{aligned}&\mathbb{P}[W_{k+1} = w_k + 1 \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})]\\ &= \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]}), \mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_{k-1}] \ge p_\tau.\end{aligned}$$

Additionally, using the law of total probability, we have

$$1 = \mathbb{P}[W_{k+1} = w_k \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] + \mathbb{P}[W_{k+1} = w_k + 1 \mid E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})].$$

Therefore, it follows that

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad}, k_{\max}, J} \mid E]\\ &\le \max_{p \in [p_\tau, 1]} \left((1-p)\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k\right] + p\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \le s_{\max} - w_k - 1\right]\right)\\ &\quad \cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i \mid E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i \mid E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).\end{aligned} \tag{C.8}$$

In addition, by the definition of $Z_{j,1}$ and $Z_{j,2}$, one finds that

$$\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \le s_{\max} - w_k - 1\right] \le \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k\right].$$

Therefore, it follows that the max in (C.8) is given by $p = p_\tau$. Thus,

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad}, k_{\max}, J} \mid E]\\ &\le \left((1-p_\tau)\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k\right] + p_\tau\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k - 1\right]\right)\\ &\quad \cdot \prod_{i=1}^{k} (\mathbb{P}[Q_i = q_i \mid E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i \mid E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]),\end{aligned}$$

where, by the definitions of $Z_{j,1}$ and $Z_{j,2}$, we have used the fact that

$$\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \le s_{\max} - w_k - 1\right] = \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \le s_{\max} - w_k - 1\right].$$

Now, for all $j \in \{1,\ldots,\psi_J(q_{[k]})\}$, define $Z_j = Z_{j,1}$, and let $Z_{\psi_J(q_{[k]})+1}$ be a Bernoulli random variable with $\mathbb{P}[Z_{\psi_J(q_{[k]})+1} = 1] = p_\tau$. Then, it follows that

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max}} | E]\\ &\leq \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})+1} Z_j \leq s_{\max} - w_k\right]\\ &\quad \cdot \prod_{i=1}^{k}(\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).\end{aligned}$$

This proves the result in this case by noting that $q_k = 1$ implies

$$\psi_J(q_{[k]}) + 1 = J - 1 - \sum_{i=0}^{k} q_i + 1 = J - 1 - \sum_{i=0}^{k-1} q_i = \psi_J(q_{[k-1]}).$$

Next, consider the case where $q_k = 0$. Recalling that

$$1 = \mathbb{P}[W_{k+1} = w_k | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})] + \mathbb{P}[W_{k+1} = w_k + 1 | E, G_{[k-1]} \in N(q_{[k]}, w_{[k]})],$$

it follows from (C.7) that

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E]\\ &\leq \max_{p \in [0,1]}\left((1-p)\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \leq s_{\max} - w_k\right] + p\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \leq s_{\max} - w_k - 1\right]\right) \qquad \text{(C.9)}\\ &\quad \cdot \prod_{i=1}^{k}(\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).\end{aligned}$$

Similar to before, noting that

$$\mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,2} \leq s_{\max} - w_k - 1\right] \leq \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_{j,1} \leq s_{\max} - w_k\right],$$

it follows that the max in (C.9) is given by $p = 0$, so with $Z_j = Z_{j,1}$ for all $j \in \{1,\ldots,\psi_J(q_{[k]})\}$,

$$\begin{aligned}&\mathbb{P}[G_{[k-1]} \in N(q_{[k]}, w_{[k]}) \wedge E_{\text{bad},k_{\max},J} | E]\\ &\leq \mathbb{P}\left[\sum_{j=1}^{\psi_J(q_{[k]})} Z_j \leq s_{\max} - w_k\right]\\ &\quad \cdot \prod_{i=1}^{k}(\mathbb{P}[Q_i = q_i | E, W_i = w_i, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})] \cdot \mathbb{P}[W_i = w_i | E, G_{[i-2]} \in N(q_{[i-1]}, w_{[i-1]})]).\end{aligned}$$

The result follows from this inequality and the fact that $\psi_J(q_{[k]}) = \psi_J(q_{[k-1]})$ because $q_k = 0$. □

We now apply Lemma C.2 to obtain a high-probability bound.

Lemma C.3. *For any $J > \frac{s_{\max}}{p_\tau} + 1$, one finds that*

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \geq J\right] \leq e^{-\frac{p_\tau(J-1)}{2}\left(1 - \frac{s_{\max}}{p_\tau(J-1)}\right)^2}. \qquad \text{(C.10)}$$

Proof. Recalling that the initial condition for the tree, $G_{[-1]} \in N(q_0, w_0)$, occurs with probability one, by Lemma C.2, it follows that there exist $J-1$ independent Bernoulli random variables Z_j with $\mathbb{P}[Z_j = 1] = p_\tau$ for all $j \in \{1,\ldots,J-1\}$ such that

$$\mathbb{P}[E_{\text{bad},k_{\max},J} | E] = \mathbb{P}[G_{[-1]} \in N(q_0, w_0) \wedge E_{\text{bad},k_{\max}} | E] \leq \mathbb{P}\left[\sum_{j=1}^{J-1} Z_j \leq s_{\max}\right].$$

Let

$$\mu := \sum_{j=1}^{J-1} \mathbb{P}[Z_j = 1] = p_\tau(J-1) \text{ and } \rho := 1 - s_{\max}/\mu.$$

Noting that $\rho \in (0,1)$ by the definition of J, by the multiplicative form of Chernoff's bound,

$$\mathbb{P}\left[\sum_{j=1}^{J-1} Z_j \le s_{\max}\right] \le e^{-\frac{1}{2}\mu\rho^2} = e^{-\frac{1}{2}\mu(1-s_{\max}/\mu)^2} = e^{-\frac{p_\tau(J-1)}{2}\left(1-\frac{s_{\max}}{p_\tau(J-1)}\right)^2}.$$

For all $k \in [k_{\max}]$, we have $\mathcal{T}_k \le \mathcal{T}_{k-1}$. Thus, by the definition of $E_{\text{bad},k_{\max},J}$, it follows that

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J | E\right] \le \mathbb{P}\left[\sum_{k=0}^{k_{\max}} Q_k \ge J | E\right] \le \mathbb{P}[E_{\text{bad},k_{\max},J} | E] \le e^{-\frac{p_\tau(J-1)}{2}\left(1-\frac{s_{\max}}{p_\tau(J-1)}\right)^2},$$

as desired. □

We now prove the main result of this Appendix.

Theorem C.1. *Under Assumption C.1, it follows that*

$$\mathbb{P}\left[\sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] < \infty \middle| E\right] = 1. \tag{C.11}$$

Proof. By Lemma C.3, for any $k_{\max} \in \mathbb{N} \setminus \{0\}$ and $J > \frac{s_{\max}}{p_\tau} + 1$, it follows that

$$\mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J \middle| E\right] \le e^{-\frac{p_\tau(J-1)}{2}\left(1-\frac{s_{\max}}{p_\tau(J-1)}\right)^2}.$$

Let $A_{k_{\max}}$ denote the event that

$$\sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J.$$

It follows from this definition that $A_{k_{\max}} \subseteq A_{k_{\max}+1}$ for any $k_{\max} \in \mathbb{N} \setminus \{0\}$. Therefore, by the properties of an increasing sequence of events (see, for example, section 1.5 in Stirzaker [33]), it follows that

$$\begin{aligned}\mathbb{P}\left[\sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J \middle| E\right] &= \mathbb{P}\left[\lim_{k_{\max}\to\infty} \sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J \middle| E\right] \\ &= \lim_{k_{\max}\to\infty} \mathbb{P}\left[\sum_{k=0}^{k_{\max}} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] \ge J \middle| E\right] \le e^{-\frac{p_\tau(J-1)}{2}\left(1-\frac{s_{\max}}{p_\tau(J-1)}\right)^2}.\end{aligned}$$

Next, let A_J denote the event that

$$\sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] < J.$$

From the definition of A_J, it follows that $A_J \subseteq A_{J+1}$ for any $J > \frac{s_{\max}}{p_\tau} + 1$. Thus, as above,

$$\begin{aligned}\mathbb{P}\left[\sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] < \infty \middle| E\right] &= \mathbb{P}\left[\lim_{J\to\infty} \sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] < J \middle| E\right] \\ &= \lim_{J\to\infty} \mathbb{P}\left[\sum_{k=0}^{\infty} \mathcal{I}[\mathcal{T}_k^{\text{trial, true}} < \mathcal{T}_k] < J \middle| E\right] \\ &\ge \lim_{J\to\infty} 1 - e^{-\frac{p_\tau(J-1)}{2}\left(1-\frac{s_{\max}}{p_\tau(J-1)}\right)^2} = 1,\end{aligned}$$

which is the desired conclusion. □

References

[1] Berahas AS, Curtis FE, Robinson DP, Zhou B (2021) Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM J. Optim.* 31(2):1352–1379.

[2] Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

[3] Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2(3):1–27.

[4] Chen C, Tung F, Vedula N, Mori G (2018) Constraint-aware deep neural network compression. *Proceedings of the European Conference on Computer Vision (ECVC)*, 400–415.

[5] Cuomo S, Di Cola VS, Giampaolo F, Rozza G, Raissi M, Piccialli F (2022) Scientific machine learning through physics–informed neural networks: Where we are and what's next. *J. Sci. Comput.* 92(88).

[6] Curtis FE, Nocedal J, Wächter A (2009) A matrix-free algorithm for equality constrained optimization problems with rank deficient Jacobians. *SIAM J. Optim.* 20(3):1224–1249.

[7] Curtis FE, O'Neill MJ, Robinson DP (2023) Worst-case complexity of an SQP method for nonlinear equality constrained stochastic optimization. *Math. Program.*, ePub ahead of print June 7, https://doi.org/10.1007/s10107-023-01981-1.

[8] Curtis FE, Robinson DP, Zhou B (2023) Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. Preprint, submitted February 28, https://arxiv.org/abs/2302.14790.

[9] Curtis FE, Kungurtsev V, Robinson DP, Wang Q (2023) A stochastic-gradient-based interior-point algorithm for solving smooth bound-constrained optimization problems. Preprint, submitted April 28, https://arxiv.org/abs/2304.14907.

[10] Davis D, Drusvyatskiy D, Kakade S (2020) Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* 20:119–154.

[11] Fang Y, Na S, Mahoney MW, Kolar M (2022) Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. Preprint, submitted November 29, https://arxiv.org/abs/2211.15943.

[13] Gould NIM, Orban D, Toint PL (2015) CUTEst: A constrained and unconstrained testing environment with safe threads for mathematical optimization. *Comput. Optim. Appl.* 60:545–557.

[12] Gould NIM, Lucidi S, Roma M, Toint PL (1999) Solving the trust-region subproblem using the Lanczos method. *SIAM J. Optim.* 9(2):504–525.

[14] Han SP (1977) A globally convergent method for nonlinear programming. *J. Optim. Theory Appl.* 22(3):297–309.

[15] Han SP, Mangasarian OL (1979) Exact penalty functions in nonlinear programming. *Math. Program.* 17:251–269.

[16] Hazan E, Luo H (2016) Variance-reduced and projection-free stochastic optimization. *Proceedings of International Conference on Machine Learning (ICML)*, 1263–1271.

[17] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L (2021) Physics-informed machine learning. *Nat. Rev. Phys.* 3:422–440.

[18] Kumar Roy S, Mhammedi Z, Harandi M (2018) Geometry aware constrained optimization techniques for deep learning. *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 4460–4469.

[19] Locatello F, Yurtsever A, Fercoq O, Cevher V (2019) Stochastic Frank-Wolfe for composite convex minimization. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 14269–14279.

[20] Lu H, Freund RM (2020) Generalized stochastic Frank-Wolfe algorithm with stochastic "substitute" gradient for structured convex optimization. *Math. Program.* 187:317–349.

[21] Lu L, Pestourie R, Yao W, Wang Z, Verdugo F, Johnson SG (2021) Physics-informed neural networks with hard constraints for inverse design *SIAM J. Sci. Comput.* 43(6):B1105–B1132.

[22] Na S, Anitescu M, Kolar M (2023) An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Math. Program.* 199:721–791.

[23] Nandwani Y, Pathak A, Singla P (2019) A primal-dual formulation for deep learning with constraints. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 12157–12168.

[24] Négiar G, Mahoney MW, Krishnapriyan AS (2023) Learning differentiable solvers for systems with hard constraints. Preprint, submitted July 18, https://arxiv.org/abs/2207.08675.

[25] Nocedal J, Wright S (2006) Numerical optimization. *Springer Series in Operations Research and Financial Engineering* (Springer-Verlag, New York).

[26] Omojokun EO (1989) Trust region algorithms for optimization with nonlinear equality and inequality constraints. PhD thesis, University of Colorado, Boulder, CO.

[27] Powell MJD (1978) A fast algorithm for nonlinearly constrained optimization calculations. *Numerical Analysis*, Lecture Notes in Mathematics, vol. 630 (Springer, Berlin), 144–157.

[28] Ravi SN, Dinh T, Lokhande VS, Singh V (2019) Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. *Proc. Conf. AAAI Artif. Intell.* 33:4772–4779.

[29] Reddi SJ, Sra S, Póczos B, Smola A (2016) Stochastic Frank-Wolfe methods for nonconvex optimization. *2016 54th Annual Allerton Conference*, 1244–1251 (IEEE, Piscataway, NJ).

[30] Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

[31] Robbins H, Siegmund D (1971) A convergence theorem for nonnegative almost supermartingales and some applications. Rustagi JS, ed., *Optimizing Methods in Statistics* (Academic Press, New York).

[32] Steihaug T (1983) The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.* 20(3):626–637.

[33] Stirzaker D (2003) *Elementary Probability* (Cambridge University Press, Cambridge).

[34] Wächter A, Biegler LT (2000) Failure of global convergence for a class of interior point methods for nonlinear programming. *Math. Program.* 88(3):565–574.

[35] Zhang M, Shen Z, Mokhtari A, Hassani H, Karbasi A (2020) One sample stochastic Frank-Wolfe. *AISTATS* 2020:4012–4023.