COMPLEXITY OF AN INEXACT STOCHASTIC SQP ALGORITHM FOR EQUALITY CONSTRAINED OPTIMIZATION *

MICHAEL J. O'NEILL† AND AOJI TANG‡

Abstract. In this paper, we consider nonlinear optimization problems with a stochastic objective function and deterministic equality constraints. We propose an inexact two-stepsize stochastic sequential quadratic programming (SQP) algorithm and analyze its worst-case complexity under mild assumptions. The method utilizes a step decomposition strategy and handles stochastic gradient estimates by assigning different stepsizes to different components of the search direction. We establish the first known $\mathcal{O}(\epsilon_c^{-2})$ worst-case complexity with respect to the infeasibility measure when no constraint qualification is assumed and a worst-case complexity of $\mathcal{O}(\epsilon_c^{-1})$ when LICQ holds, matching the best known result in the literature. In addition, under mild conditions, our method achieves the optimal $\mathcal{O}(\epsilon_L^{-4})$ complexity with respect to the gradient of the Lagrangian regardless of constraint qualifications. Our results provide the first complexity guarantees for the popular Byrd-Omojukun step decomposition strategy and verify its theoretical efficacy. Numerical experiments show that our algorithm has a superior infeasibility convergence performance and a competitive KKT convergence rate compared to the state-of-the-art stochastic SQP method.

Key words. constrained optimization, sequential quadratic programming, stochastic optimization, rank deficient constraint Jacobians, worst case complexity, inexact subproblem solver

MSC codes. 49M37, 65K05, 65K10, 90C15, 90C30, 90C55

1. Introduction.

In this paper, we consider the equality-constrained optimization problem as follows:

$$\min_{x\in\mathbb{R}^n} f(x) = \mathbb{E}[F(x,\omega)] \quad \text{s.t.} \quad c(x) = 0, \tag{1.1}$$

where f is the expectation of $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, under the random variable ω with the associated probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is the constraint function.

Such problems emerge naturally in scenarios where system dynamics are strictly governed by physical laws. For instance, in stochastic optimal power flow problems, the objective is to minimize expected generation costs under highly stochastic renewable energy inputs, while the power flow balance must strictly hold as deterministic equality constraints dictated by Kirchhoff's circuit laws. Similarly, in Physics-Informed Deep Learning (PINN) problems [12, 21], boundary and initial conditions are often enforced as hard deterministic constraints, while only stochastic gradient information is available for computational efficiency. Such formulations are also prevalent in science and engineering, with prominent applications in optimal control [8, 28] and PDE-constrained optimization [19].

There are two general types of methods to solve our desired problem. Stochastic augmented Lagrangian methods [9, 17, 20, 26, 30] penalize the constraints in the objective and solve the corresponding unconstrained subproblem to get a primal update. These methods can benefit from simpler forms of constraints (e.g. linear constraints) and naturally generalize to inequality constrained problems. Meanwhile, stochastic

* **Funding:** This work was partially funded by the Office of Naval Research under award N00014-24-1-2638.

†Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC (mikeoneill@unc.edu).

‡Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC (ajtang@unc.edu).

Sequential Quadratic Programming (SQP) methods are better at coping with more complicated nonlinear constraints. SQP methods compute the iterates via a sequence of quadratic programming subproblems which preserve linearized feasibility.

Existing stochastic SQP frameworks [2, 4, 5, 6, 7, 13, 22, 25] rely heavily on the Linear Independence Constraint Qualification (LICQ) at every iteration. However, this assumption often fails in high-dimensional applications. For example, in PINN problems, enforcing boundary conditions over dense collocation points often leads to redundant constraints, inherently causing rank-deficient constraint Jacobians. When LICQ fails, the KKT system becomes singular. This singularity disrupts standard Newton-type step computations and can cause the merit parameter to vanish prematurely, preventing the algorithm from adequately optimizing the objective function.

To safeguard against these degenerate constraint manifolds, one popular approach is to employ the step decomposition strategy of Byrd-Omojokun [24]. The Byrd-Omojokun method computes two orthogonal components: the normal component (oriented toward feasibility), via a trust-region subproblem, and the tangential component (optimizing the objective within the null space), through the solution of a linear system of equations. These two components are then combined to form the full step direction. This strategy has proven effective for stochastic SQP methods in recent work [3].

1.1. Contributions. Given the critical absence of complexity results under mild assumptions, we propose and analyze an Inexact Two-Stepsize SQP (ITSQP) algorithm. By employing an l_2 merit function, we comprehensively analyze the worst-case complexity results under three specific behaviors of the merit parameter sequence. We prove that our algorithm converges to stationary points (in expectation) in both infeasibility and first-order measures.

Our specific contributions in terms of worst-case complexity results are as follows:

- **Without Constraint Qualifications:** To achieve an ϵ_c-feasible point, our algorithm requires at most $\mathcal{O}(\epsilon_c^{-2})$ iterations regardless of whether the merit parameter sequence has a positive lower bound. To our knowledge, this is the *first known* complexity result for stochastic SQP algorithm under such mild, rank-deficient assumptions.
- **With LICQ:** We prove that the Byrd-Omojokun safeguard does not impede performance. When everywhere LICQ holds, the algorithm perfectly recovers an improved $\mathcal{O}(\epsilon_c^{-1})$ worst-case complexity for the infeasibility measure. This result matches the best known complexity in [25], where LICQ is assumed.
- **First-order Optimality:** When the merit parameter sequence is bounded, our algorithm requires at most $\mathcal{O}(\epsilon_L^{-4})$ iterations to ensure that the first-order optimality condition falls below the tolerance ϵ_L, which matches the optimal complexity bounds for stochastic gradient methods [1].

Furthermore, we exploit the use of inexact iterative solvers in the tangential subproblems and prove equivalent convergence and complexity results under mild assumptions on these solvers. Numerical experiments demonstrate that our ITSQP algorithm yields substantial enhancements in both feasibility convergence speed and accuracy compared to state-of-the-art exact stochastic SQP methods.

We conclude this section with Table 1, which contains the complexity results of different algorithms. Note that the approximate convergence metrics vary under different assumptions.

1.2. Organization. The assumptions and our algorithm are presented in Section 2. Section 3 details the worst-case complexity proofs across the three aforemen-

Algorithm	Assumption	1st-order Comp.	Infeas. Comp.
SPD[18]	x_0 feasible, CQ	$\mathcal{O}(\epsilon_L^{-5})$	$\mathcal{O}(\epsilon_c^{-5})^\dagger$
MLALM[26]	mean-squared smoothness, CQ	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-4})^\dagger$
SSQP[4]	LICQ	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-2})$
SQP-AL[22]	LICQ	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-4})$
TSSQP[25]	LICQ	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-1})$
ITSQP (this paper)	**No CQ assumed**, $\exists\ \tau_{\min}$	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-2})^\ddagger$
	No CQ assumed, $\nexists\ \tau_{\min}$	N/A	$\mathcal{O}(\epsilon_c^{-2})^\ddagger$
	LICQ	$\mathcal{O}(\epsilon_L^{-4})$	$\mathcal{O}(\epsilon_c^{-1})$

TABLE 1

Complexity results of different algorithms solving (1.1). *The default first-order measure is the norm of gradient of the Lagrangian function. The default infeasibility measure is the norm of constraint violation.* †: *The stochastic augmented Lagrangian methods use a combined feasibility and optimality measure.* ‡: *We use a different infeasibility measure when no CQ is assumed, see Corollary* 3.12. *Here,* $\tau_{\min}$ *represents a lower bound on the merit parameter seqeunce.*

tioned cases. Section 4 discusses the implementation of inexact subproblem solutions and proves complexity results for the inexact method. Preliminary experiments are presented in Section 5 and we provide a brief discussion of our work in Section 6.

1.3. Notation. The set of real numbers is denoted as $\mathbb{R}$, the set of positive real numbers is denoted as $\mathbb{R}_{>0}$, the set of natural numbers is denoted as $\mathbb{N}$, 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 symmetric matrices is denoted as $\mathbb{S}^n$. We use $\|\cdot\|$ to denote the Euclidean norm. As in (1.1), f and c denote the objective and constraint functions, respectively. c_i denotes the i-th component of constraint c. Given $A \in \mathbb{R}^{m\times n}$, the null space of A is denoted as $\text{Null}(A)$, and the range space of A^T is denoted as $\text{Range}(A^T)$. We denote the Moore-Penrose psuedoinverse of a matrix A as A^+.

Our algorithm is iterative in the sense that, given a starting point $x_0 \in \mathbb{R}^n$, it generates a sequence of iterations $\{x_k\}_{k\geq 0}$ with $x_k \in \mathbb{R}^n$. At iterate k, we denote $c_k = c(x_k)$ and $J_k = \nabla c(x_k)^T$. Given $J_k \in \mathbb{R}^{m\times n}$, Z_k denotes a matrix whose columns form an orthonormal basis for $\text{Null}(J_k)$.

2. Problem Statement and Algorithm Description. We make the following assumption about the problem (1.1) throughout the paper.

ASSUMPTION 2.1. *Let* $\chi \subset \mathbb{R}^n$ *be an open convex set containing the sequence of iterations* $\{x_k\}$ *generated by our algorithm. The objective function* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable and bounded over* χ *and its gradient* $\nabla f : \mathbb{R}^n \to \mathbb{R}$ *is Lipschitz continuous with constant* $L > 0$ *and bounded over* χ. *The constraint function* $c : \mathbb{R}^n \to \mathbb{R}^m$ *is continuously differentiable and bounded over* χ *and its Jacobian function* $J := \nabla c^T : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ *is bounded over* χ *and Lipschitz continuous with constant* $\Gamma > 0$ *(with respect to* $\|\cdot\|$*) over* χ.

From Assumption 2.1, it follows that there exist positive real numbers $(f_{\text{inf}}, \kappa_g, \kappa_c, \kappa_J) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$, such that

$$f_{\text{inf}} \leq f(x_k),\ \|\nabla f(x_k)\| \leq \kappa_g,\ \|c_k\| \leq \kappa_c,\ \text{and}\ \|J_k\| \leq \kappa_J \text{ for all } k \in \mathbb{N}.$$

Assumption 2.1 are common assumptions on the smoothness of the functions. As for the boundedness of the function and gradient, it is not ideal to assume these remain

bounded in a stochastic setting, since we can only provide convergence in expectation. However, such assumptions are essential in *deterministic* constrained optimization and are common in the stochastic constrained literature [3, 4, 25, 26].

Now we define the Lagrangian function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ corresponding to the problem (1.1) by $L(x, y) = f(x) + c(x)^T y$. The standard KKT condition under the linear independence constraint qualification is

$$\begin{bmatrix} \nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + J(x)^T y \\ c(x) \end{bmatrix} = 0. \tag{2.1}$$

However, without any kind of constraint qualification, it is possible for the KKT system to be degenerate due to the possible rank deficiency of the constraint Jacobian matrix $J(x)$. Then (2.1) might not necessarily be satisfied at a solution to the original problem (1.1), or the problem (1.1) might be infeasible itself. In the latter case, we can only hope that some other measure of constraint violation converges to zero.

To handle this, we employ the l_2-norm of constraints, denoted by $\varphi(x) = \|c(x)\|$, as our infeasibility measure. We say that a point $x \in \mathbb{R}^n$ is stationary with respect to φ, if and only if either:

1. $\varphi(x) = 0$, or
2. $\varphi(x) \neq 0$ and $\nabla\varphi(x) = \frac{J(x)^T c(x)}{\|c(x)\|} = 0$.

We acknowledge that this approach is a compromise, but it is common practice when no constraint qualification is assumed, as in [3]. In our analysis, we mainly work with $\|J(x)^T c(x)\|$ as our infeasibility measure. This is because $\varphi(x) = 0$ will also lead to $\|J(x)^T c(x)\| = 0$, and φ can be recovered from $\|J(x)^T c(x)\|$ when the LICQ holds at every iteration. When considering first-order stationarity, we use the gradient of the Lagrangian, $\|\nabla f(x) + J(x)^T y\|$, with a properly chosen dual variable y.

2.1. Step computation and algorithm. Standard stochastic SQP algorithms compute the search direction via a sequence of quadratic programming subproblems which satisfy linearized feasibility. At iteration k, the subproblem is formed as

$$\begin{aligned} d_k = \underset{d \in \mathbb{R}^n}{\operatorname{argmin}} \quad & f_k + g_k^T d + \frac{1}{2} d^T H_k d \\ \text{s.t.} \quad & c_k + J_k d = 0, \end{aligned} \tag{2.2}$$

where H_k is a chosen matrix and g_k is a stochastic estimate of $\nabla f(x_k)$.

For a general quadratic programming problem in the form (2.2), there are numerous ways to compute the solution when J_k has full row rank, such as factorization methods and conjugate gradient methods [23]. However, as we do not assume full rank of the Jacobians, rather than solving (2.2) directly for d_k, we employ a step decomposition strategy known as the Byrd-Omojokun method [24].

At each iteration, we first compute the "normal" component $v_k \in \text{Range}(J_k^T)$ of the search direction to minimize linearized constraint violation over a trust-region:

$$\min_{v \in \mathbb{R}^n} \frac{1}{2}\|c_k + J_k v\|^2 \quad \text{s.t.} \quad \|v\| \leq \omega \|J_k^T c_k\|, \tag{2.3}$$

where $\omega > 0$ is a deterministic parameter to control the size of the normal component. Rather than using an adaptive trust-region radius sequence $\{\Delta_k\}$ (like in [14]), we intentionally link v_k with our infeasibility measure $\|J_k^T c_k\|$.

Solving (2.3) exactly may be expensive, but fortunately we only require that the normal step v_k satisfies the following Cauchy decrease condition:

$$\|c_k\| - \|c_k + J_k v_k\| \geq \epsilon_v (\|c_k\| - \|c_k + \alpha_k^C J_k v_k^C\|), \tag{2.4}$$

for some $\epsilon_v \in (0, 1]$. Here, $v_k^C = -J_k^T c_k$, and α_k^C is the solution to the problem $\min_\alpha \frac{1}{2}\|c_k + \alpha J_k v_k^C\|^2$, subject to $\alpha \leq \omega$. Since this condition can be satisfied simply by choosing the Cauchy step $v_k \leftarrow v_k^C$, the normal component can be computed at a relatively low cost. To obtain a more accurate solution, one can apply the linear conjugate gradient method with Steihaug stopping conditions [27] to find a solution that satisfies the Cauchy decrease condition. An important property of (2.3) is that the normal component v_k is independent of the stochasticity introduced by g_k. This property is fundamental to our two-stepsize scheme and lays the theoretical foundation for our superior infeasibility complexity compared to that of [3].

After computing the normal component v_k, our algorithm computes the tangential component u_k through an additional constrained subproblem, involving the stochastic gradient estimate g_k. We start by introducing the tangential subproblem with an assumption on H_k, which assumes that H_k is positive-definite in the null space of J_k and is chosen independently of g_k.

ASSUMPTION 2.2. *For all $k \in \mathbb{N}$, the matrix $H_k \in \mathbb{S}^n$ is chosen independently from g_k, the sequence $\{H_k\}$ is bounded in norm by κ_H, and there exists $\zeta \in \mathbb{R}_{>0}$, such that $u^T H_k u \geq \zeta \|u\|^2$ for all $u \in Null(J_k)$.*

The subproblem to compute the tangential component u_k is formulated as

$$\min_{u \in \mathbb{R}^n} (g_k + H_k v_k)^T u + \frac{1}{2} u^T H_k u \quad \text{s.t.} \quad J_k u = 0. \tag{2.5}$$

Under Assumption 2.2, u_k is unique and can be obtained by solving the corresponding Newton system

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k \\ y_k \end{bmatrix} = - \begin{bmatrix} g_k + H_k v_k \\ 0 \end{bmatrix}, \tag{2.6}$$

although the Lagrange multipliers y_k may have multiple solutions due to rank deficiency. An exact solution to this subproblem can be obtained through factorization methods, while inexact solutions can be obtained with iterative solvers. We will later discuss the possibility of inexactly computing u_k in Section 4.

Throughout, we assume the following standard assumptions about the stochastic gradient estimate g_k.

ASSUMPTION 2.3. *For all $k \in \mathbb{N}$, the stochastic gradient estimate $g_k \in \mathbb{R}^n$ is an unbiased estimator of $\nabla f(x_k)$,* i.e., *$\mathbb{E}_k[g_k] = \nabla f(x_k)$, where $\mathbb{E}_k[\cdot]$ denotes the conditional expectation up to iteration x_k. In addition, there exists $M \in \mathbb{R}_{>0}$ such that $\mathbb{E}_k[\|g_k - \nabla f(x_k)\|^2] \leq M$.*

After computing the two orthogonal components, a natural way to compute the step direction d_k is to simply add them together, as in [3]. However, our algorithm first rescales the tangential part u_k by some pre-defined parameter sequence $\{\beta_k\} > 0$ and then combines the components to form d_k instead. Formally, we set

$$d_k = \beta_k u_k + v_k. \tag{2.7}$$

The next iteration x_{k+1} is then produced by $x_{k+1} = x_k + \alpha_k d_k$, where α_k is the second stepsize. The role of β_k is crucial in our analysis. Simply put, $\{\beta_k\}$ controls the variance introduced by the stochastic gradient estimates g_k, but does not impact convergence in the constraints, which is driven by v_k. Now we can state a general algorithm framework, where the subproblem (2.5) can be solved exactly or inexactly.

Algorithm 2.1 Two-Stepsize Stochastic SQP Algorithm Framework

Initialize $x_0 \in \mathbb{R}^n$.
for $k = 0, 1, 2, ..., K-1$ **do**
  Compute the normal component v_k with (2.3).
  Compute a stochastic gradient estimate g_k.
  Compute the tangential component u_k exactly by (2.5) or inexactly by (4.1).
  Choose $\beta_k > 0$.
  Set $d_k \leftarrow \beta_k u_k + v_k$.
  Choose $\alpha_k > 0$.
  Set $x_{k+1} \leftarrow x_k + \alpha_k d_k$.
end for

There are various ways to choose the stepsizes $\{\beta_k\}$ and $\{\alpha_k\}$. Since the purpose of $\{\beta_k\}$ is to control the variance of stochastic gradient estimates, they are essentially equivalent to the stepsize employed in stochastic gradient methods. Thus, in order to obtain our desired complexity result, we set $\beta_k = O(1/\sqrt{K})$ where K is the total number of iterations that we plan to perform, which is a standard choice in the stochastic gradient literature [15]. For the sake of brevity, we assign $\kappa_\beta \in \mathbb{R}_{>0}$ as an upper bound such that $\beta_k \leq \kappa_\beta$ for all $k \in \mathbb{N}$.

On the other hand, α_k can be chosen to be independent of K because we do not require it to control the stochastic error and it can be viewed as an essentially deterministic step size, which should be inversely proportional to smoothness constants of the problem. Specifically, we choose α_k from an interval involving β_k as follows:

$$\alpha_k \in [\nu, \nu + \theta\beta_k], \tag{2.8}$$

where $\nu \in \mathbb{R}_{>0}$ satisfies properties with respect to problem-specific parameters (see, (3.19)) and $\theta \in \mathbb{R}_{>0}$ is a user-defined constant. This interval type of α_k is common in stochastic SQP methods, as in [4] and [25].

3. Convergence Analysis. To start our analysis, we first define some "true" variables, which are quantities that would be computed if we used the full gradient $\nabla f(x_k)$ rather than stochastic gradient estimates g_k. For instance, let

$$u_k^{\text{true}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \, (\nabla f(x_k) + H_k v_k)^T u + \frac{1}{2} u^T H_k u \quad \text{s.t.} \quad J_k u = 0. \tag{3.1}$$

Equivalently, u_k^{true} can be obtained by solving the "true" Newton system

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k^{\text{true}} \\ y_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) + H_k v_k \\ 0 \end{bmatrix}. \tag{3.2}$$

As mentioned previously, y_k^{true} may not be uniquely defined by (3.2). Throughout our analysis, we impose the choice of the least-squares solution, i.e.,

$$y_k^{\text{true}} := -(J_k^T)^+ (\nabla f(x_k) + H_k v_k + H_k u_k^{\text{true}}). \tag{3.3}$$

Since v_k is independent from the stochastic gradient estimate g_k, v_k^{true} is identical to v_k. Therefore, $d_k^{\text{true}} = \beta_k u_k^{\text{true}} + v_k$, and we have the useful property:

$$J_k d_k^{\text{true}} = J_k(\beta_k u_k^{\text{true}} + v_k) = J_k v_k = J_k d_k. \tag{3.4}$$

3.1. Merit function, local model and preliminary results. To study the convergence of SQP methods, a useful tool is the merit function. We employ the l_2 merit function $\phi : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$ defined by

$$\phi(x, \tau) = \tau f(x) + \|c(x)\| \tag{3.5}$$

in our analysis, where τ is the merit parameter.

We also introduce a local model of the merit function, $l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$:

$$l(x, \tau, g, d) = \tau(f(x) + g^T d) + \|c(x) + J(x)d\|. \tag{3.6}$$

The local model $l(x, \tau, g, d)$ can be seen as an approximation of ϕ near the current point x. As we will see, l serves as a bridge between the merit function ϕ and the infeasibility measure $\|J(x)^T c(x)\|$. We also define the reduction in the local model for a direction d and a gradient estimate g as follows:

$$\begin{aligned} \Delta l(x, \tau, g, d) &:= l(x, \tau, g, 0) - l(x, \tau, g, d) \\ &= -\tau g^T d + \|c(x)\| - \|c(x) + J(x)d\|. \end{aligned} \tag{3.7}$$

Although Algorithm 2.1 does not involve any computation of the merit parameter τ, it still plays an important role in complexity analysis. Similarly to [3], we assume that the sequence of merit parameters $\{\tau_k\}$ can be generated by the following scheme. The only difference is that we substitute values involving stochastic gradients with "true" values and the rescaling parameter β_k.

For some fixed $\sigma \in (0, 1)$ and $\varepsilon_\tau \in [0, 1)$, let

$$\tau_k^{trial} \leftarrow \begin{cases} \infty & \text{if } \nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} \le 0 \\ \frac{(1-\sigma)(\|c_k\| - \|c_k + J_k v_k\|)}{\nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}}} & \text{otherwise,} \end{cases} \tag{3.8}$$

and

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \le \tau_k^{\text{trial}} \\ \min\{(1 - \varepsilon_\tau)\tau_{k-1}, \tau_k^{\text{trial}}\} & \text{otherwise.} \end{cases} \tag{3.9}$$

Hence, the merit parameter sequence $\{\tau_k\}$ is monotonically non-increasing and satisfies

$$\tau_k(\nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}}) \le (1 - \sigma)(\|c_k\| - \|c_k + J_k v_k\|). \tag{3.10}$$

From the fact that $\Delta l(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}}) = -\tau_k \nabla f_k^T d_k^{\text{true}} + (\|c_k\| - \|c_k + J_k v_k\|)$ and $\sigma < 1$, we have the following model reduction inequality:

$$\Delta l(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}}) \ge \tau_k \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \sigma(\|c_k\| - \|c_k + J_k v_k\|). \tag{3.11}$$

Now we state some useful bounds derived from the step computation subproblems (2.3) and (2.5). They link the local model reduction with our infeasibility measure. Notice that these lemmas *do not* depend on the behavior of the merit parameter sequence and can be used for all cases we later discuss. The first lemma provides a tight bound on the infeasibility measure reduction.

LEMMA 3.1. *For any one dimensional optimization problem*

$$\min_{z \in \mathbb{R}^n} \Phi(z) = \frac{1}{2} a z^2 - bz \quad s.t. \quad z \le \omega,$$

where $b > 0$ and $\omega > 0$, its optimal value Φ^* satisfies

$$\Phi^* \le -\frac{b}{2} \min\Big\{ \frac{b}{|a|}, \omega \Big\}.$$

Proof. See [11, Lemma 2.1]. ∎

We know that α_k^C is the solution to the problem $\min_{\alpha>0} \frac{1}{2}\|c_k + \alpha J_k v_k^C\|^2$ s.t. $\alpha \leq \omega$. Using Lemma 3.1, setting $a = \|J_k J_k^T c_k\|^2$ and $b = \|J_k^T c_k\|^2$, we have

$$\Phi^*(\alpha) = \frac{1}{2}\|J_k J_k^T c_k\|^2(\alpha_k^C)^2 - \|J_k^T c_k\|^2(\alpha_k^C) \leq -\frac{\|J_k^T c_k\|^2}{2}\min\Big\{\frac{\|J_k^T c_k\|^2}{\|J_k J_k^T c_k\|^2}, \omega\Big\}. \tag{3.12}$$

LEMMA 3.2. *Let Assumption 2.1 hold. Then, there exists $\kappa_v \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$ with $\|c_k\| > 0$,*

$$\|c_k\|(\|c_k\| - \|c_k + J_k v_k\|) \geq \kappa_v \|J_k^T c_k\|^2.$$

Proof. We first consider the Cauchy step $v_k^C = -J_k^T c_k$ of the subproblem (2.3). Using the previous conclusion (3.12), we have

$$\begin{aligned}
\frac{1}{2}(\|c_k + \alpha_k^C J_k v_k^C\|^2 - \|c_k\|^2) &= \frac{1}{2}(2\alpha_k^C c_k^T J_k v_k^C + (\alpha_k^C)^2\|J_k v_k^C\|^2) \\
&= -\alpha_k^C\|J_k^T c_k\|^2 + \frac{1}{2}(\alpha_k^C)^2\|J_k J_k^T c_k\|^2 \\
&\leq -\frac{1}{2}\|J_k^T c_k\|^2 \min\Big\{\frac{\|J_k^T c_k\|^2}{\|J_k J_k^T c_k\|^2}, \omega\Big\} \\
&\leq -\frac{1}{2}\|J_k^T c_k\|^2 \min\Big\{\frac{1}{\|J_k^T J_k\|}, \omega\Big\}.
\end{aligned} \tag{3.13}$$

Now, since v_k satisfies the Cauchy decreasing condition (2.4), we have the following inequality with the fact $x(x-y) \geq \frac{1}{2}(x^2 - y^2)$ for any $x, y \in \mathbb{R}$ that

$$\begin{aligned}
\|c_k\|(\|c_k\| - \|c_k + J_k v_k\|) &\geq \epsilon_v\|c_k\|(\|c_k\| - \|c_k + \alpha_k^C J_k v_k^C\|) \\
&\geq \frac{1}{2}\epsilon_v(\|c_k\|^2 - \|c_k + \alpha_k^C J_k v_k^C\|^2) \\
&\geq \frac{\epsilon_v}{2}\|J_k^T c_k\|^2 \min\{\omega, \frac{\omega^2}{\|J_k^T J_k\|}\}.
\end{aligned} \tag{3.14}$$

Since $\|J_k^T J_k\|$ is bounded from above by Assumption 2.1, there exists a constant $\kappa_v \in (0, \min\{\omega, \frac{\omega^2}{\|J_k^T J_k\|}\})$ such that

$$\|c_k\|(\|c_k\| - \|c_k + J_k v_k\|) \geq \kappa_v \|J_k^T c_k\|^2,$$

which proves the result. ∎

LEMMA 3.3. *There exists $\underline{\omega} \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$ with $\|c_k\| > 0$,*

$$\underline{\omega}\|J_k^T c_k\|^2 \leq \|v_k\| \leq \omega\|J_k^T c_k\|.$$

Proof. The right inequality comes from the trust-region constraint of (2.3). From the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\|c_k\| - \|c_k + J_k v_k\| \leq \|J_k v_k\| \leq \|J_k\|\|v_k\|.$$

If $\|J_k\| = 0$ or $\|c_k\| = 0$, the desired inequality is trivial. Otherwise, by Lemma 3.2,

$$\|v_k\| \geq \frac{\|c_k\| - \|c_k + J_k v_k\|}{\|J_k\|} \geq \frac{\kappa_v\|J_k^T c_k\|^2}{\|J_k\|\|c_k\|}.$$

Since J_k and c_k are assumed to be bounded in norm, there exists some $\underline{\omega} > 0$ that satisfy $\|v_k\| \geq \underline{\omega}\|J_k^T c_k\|^2$ for all k. ∎

Lemma 3.2 and Lemma 3.3 are are very useful in the Byrd-Omojokun type of methods [24] when constraint qualifications do not hold. We additionally prove stronger versions of them when LICQ holds in Subsection 3.3. Next, we consider a bound on the "true" tangential component, u_k^{true}.

LEMMA 3.4. *Let Assumptions 2.1 and 2.2 hold. Let y_k^{true} be defined by* (3.3) *for all $k \in \mathbb{N}$. Then, we have*

$$\|u_k^{true}\| \leq \zeta^{-1}(\|\nabla f(x_k) + J_k^T y_k^{true}\| + \kappa_H \omega \|J_k^T c_k\|),$$

and

$$\|u_k^{true}\| \leq \zeta^{-1}(\kappa_g + \kappa_H \omega \kappa_J \kappa_c) =: \kappa_u.$$

Proof. From (3.2), we know $H_k u_k^{\text{true}} + J_k^T y_k^{\text{true}} = -\nabla f(x_k) + H_k v_k$. By Assumption 2.2, since u_k^{true} lies strictly in the null space of J_k, we have

$$\begin{aligned}
\zeta \|u_k^{\text{true}}\|^2 &\leq (u_k^{\text{true}})^T H_k u_k^{\text{true}} \\
&= -\nabla f(x_k)^T u_k^{\text{true}} - v_k^T H_k u_k^{\text{true}} \\
&= -(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T u_k^{\text{true}} - v_k^T H_k u_k^{\text{true}} \\
&\leq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| \|u_k^{\text{true}}\| + \kappa_H \|v_k\| \|u_k^{\text{true}}\| \\
&\leq \|u_k^{\text{true}}\| (\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| + \kappa_H \omega \|J_k^T c_k\|).
\end{aligned}$$

Dividing both sides by $\zeta \|u_k^{\text{true}}\|$ yields the first result. Similarly,

$$\begin{aligned}
\zeta \|u_k^{\text{true}}\|^2 &\leq -\nabla f(x_k)^T u_k^{\text{true}} - v_k^T H_k u_k^{\text{true}} \\
&\leq \|\nabla f(x_k)\| \|u_k^{\text{true}}\| + \kappa_H \|v_k\| \|u_k^{\text{true}}\| \\
&\leq \|u_k^{\text{true}}\| (\kappa_g + \kappa_H \omega \|J_k^T c_k\|) \\
&\leq \|u_k^{\text{true}}\| (\kappa_g + \kappa_H \omega \kappa_J \kappa_c).
\end{aligned}$$

As before, dividing through yields the second result. ∎

The following lemma bounds the distance in expectation between the computed direction d_k and the true direction d_k^{true}, which reflects the variance of the stochastic estimates.

LEMMA 3.5. *Let Assumptions 2.2 and 2.3 hold. Then, for all $k \in \mathbb{N}$, $\mathbb{E}_k[u_k] = u_k^{true}$, $\mathbb{E}_k[d_k] = d_k^{true}$, and $\mathbb{E}_k[\|d_k - d_k^{true}\|] \leq \beta_k \zeta^{-1} \sqrt{M}$.*

Proof. Under Assumption 2.2, there exists a matrix Z_k whose columns form an orthogonal basis for $\text{Null}(J_k)$, and vectors w_k, w_k^{true} such that $u_k = Z_k w_k$ and $u_k^{\text{true}} = Z_k w_k^{\text{true}}$. From subproblem (2.5), explicit forms of w_k and w_k^{true} are given as:

$$w_k = -(Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k); \quad w_k^{\text{true}} = -(Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(x_k) + H_k v_k).$$

Since $(Z_k^T H_k Z_k)^{-1} Z_k^T$ and Z_k are both linear operators, from Assumption 2.3 and the property of expectation, we have $\mathbb{E}_k[u_k] = u_k^{\text{true}}$. In addition, since β_k is independent of g_k, $\mathbb{E}_k[d_k] = \mathbb{E}_k[\beta_k u_k + v_k] = \beta_k \mathbb{E}_k[u_k] + v_k = \beta_k u_k^{\text{true}} + v_k = d_k^{\text{true}}$.

The rest of the proof follows from Jensen's inequality and Assumptions 2.2 and 2.3.

$$\begin{aligned}
\mathbb{E}_k[\|d_k - d_k^{\text{true}}\|] &= \mathbb{E}_k[\|\beta_k (u_k - u_k^{\text{true}})\|] \\
&= \beta_k \mathbb{E}_k[\|Z_k (w_k - w_k^{\text{true}})\|] \\
&= \beta_k \mathbb{E}_k[\|Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k - \nabla f(x_k)\|] \\
&\leq \beta_k \mathbb{E}_k[\|Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T \| \|g_k - \nabla f(x_k)\|] \\
&= \beta_k \|(Z_k^T H_k Z_k)^{-1}\| \mathbb{E}_k[\|g_k - \nabla f(x_k)\|] \\
&\leq \beta_k \zeta^{-1} \sqrt{\mathbb{E}_k[\|g_k - \nabla f(x_k)\|^2]} \\
&\leq \beta_k \zeta^{-1} \sqrt{M}.
\end{aligned}$$

∎

Next, we provide an important upper bound on the expectation of $\|u_k\|^2$.

LEMMA 3.6. *Let Assumptions 2.2 and 2.3 hold. Then, for all* $k \in \mathbb{N}$,

$$\mathbb{E}_k[\|u_k\|^2] \leq \zeta^{-2} M + \zeta^{-1} (u_k^{true})^T H_k u_k^{true}.$$

Proof. Since $u_k, u_k^{\text{true}} \in \text{Null}(J_k)$, by Assumption 2.2 and (2.6), we have

$$\begin{aligned}
&\mathbb{E}_k[\|u_k\|^2] - \zeta^{-1}(u_k^{\text{true}})^T H_k u_k^{\text{true}} \\
&\leq \zeta^{-1}(\mathbb{E}_k[u_k^T H_k u_k] - (u_k^{\text{true}})^T H_k u_k^{\text{true}}) \\
&= \zeta^{-1}(\mathbb{E}_k[u_k^T(-u_k^T J_k^T y_k - g_k - H_k v_k)] - (u_k^{\text{true}})^T(-(u_k^{\text{true}})^T J_k^T y_k^{\text{true}} - \nabla f(x_k) - H_k v_k) \\
&= \zeta^{-1}(\nabla f(x_k)^T u_k - \mathbb{E}_k[g_k^T u_k]).
\end{aligned}$$

Since $u_k = Z_k w_k = Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k)$, we know

$$\begin{aligned}
\mathbb{E}_k[g_k^T u_k] &= \mathbb{E}_k[-g_k^T Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k)] \\
&\leq -\zeta^{-1}\mathbb{E}_k[\|g_k^T Z_k\|^2] - \zeta^{-1}\mathbb{E}_k[g_k^T H_k v_k] \\
&= -\zeta^{-1}(\mathbb{E}_k[\|g_k^T Z_k\|^2] + \nabla f(x_k)^T H_k v_k).
\end{aligned}$$

Similarly, $\nabla f(x_k)^T u_k \leq -\zeta^{-1}(\|g_k^T Z_k\|^2 + \nabla f(x_k)^T H_k v_k)$. Combining the inequalities and Assumption 2.3, we have

$$\mathbb{E}_k[\|u_k\|^2] - \zeta^{-1}(u_k^{\text{true}})^T H_k u_k^{\text{true}}) \leq \zeta^{-2}(-\|Z_k^T g_k\|^2 + \mathbb{E}_k[\|Z_k^T g_k\|^2]) \leq \zeta^{-2} M,$$

which proves the result. ∎

The following lemma allows us to connect (3.11) with the gradient of the Lagrangian.

LEMMA 3.7. *Let Assumptions 2.1 and 2.2 hold and let* y_k^{true} *be defined by* (3.3). *Then,*

$$(u_k^{true})^T H_k (u_k^{true}) \geq \frac{\zeta}{2\kappa_H^2} \|\nabla f(x_k) + J_k^T y_k^{true}\|^2 - \kappa_0 \|J_k^T c_k\|^2,$$

where $\kappa_0 := \omega^2(2\kappa_H + 2\zeta^{-1}\kappa_H^2 + \zeta)$.

Proof. From Assumption 2.2, we know

$$(u_k^{\text{true}})^T H_k (u_k^{\text{true}}) \geq \zeta \|u_k^{\text{true}}\|^2 \geq \zeta \kappa_H^{-2} \|H_k u_k^{\text{true}}\|^2.$$

Then, from (3.2), Lemma 3.3, and Lemma 3.4, we can conclude that

$$\begin{aligned}
\|H_k u_k^{\text{true}}\|^2 &= \|H_k u_k^{\text{true}} + H_k v_k - H_k v_k\|^2 \\
&= \|H_k u_k^{\text{true}} + H_k v_k\|^2 - 2 v_k^T H_k H_k (u_k^{\text{true}} + v_k) + \|H_k v_k\|^2 \\
&\geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - 2 v_k^T H_k H_k u_k^{\text{true}} - \|H_k v_k\|^2 \\
&\geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - \kappa_H^2 (2\|u_k^{\text{true}}\| + \|v_k\|)\|v_k\| \\
&\geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 \\
&\quad - \kappa_H^2 (2\zeta^{-1}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| + 2\zeta^{-1}\kappa_H \omega \|J_k^T c_k\| + \|v_k\|)\|v_k\| \\
&\geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - 2\omega\zeta^{-1}\kappa_H^2 \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| \|J_k^T c_k\| \\
&\quad - \kappa_H^2 \omega^2 (2\zeta^{-1}\kappa_H + 1)\|J_k^T c_k\|^2.
\end{aligned}$$

Next, applying Young's inequality to the 2nd term in the above inequality,

$$2\omega\zeta^{-1}\kappa_H^2 \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| \|J_k^T c_k\| \leq \frac{1}{2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + 2\omega^2\zeta^{-2}\kappa_H^4 \|J_k^T c_k\|^2,$$

and thus

$$\|H_k u_k^{\text{true}}\|^2 \geq \frac{1}{2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - \kappa_H^2 \omega^2 (2\zeta^{-1}\kappa_H + 2\zeta^{-2}\kappa_H^2 + 1)\|J_k^T c_k\|^2,$$

which proves the result. ∎

Recall that the merit parameter sequence $\{\tau_k\}$ is monotonically non-increasing and non-negative. Therefore, by the monotone convergence theorem, there are two possible tail behaviors:

(i) There exists some $\tau_{\min} \in \mathbb{R}_{>0}$ such that $\lim_{k\to\infty} \tau_k = \tau_{\min}$,
(ii) $\{\tau_k\}$ vanishes, *i.e.,* $\lim_{k\to\infty} \tau_k = 0$.

We formalize event (i) in the following assumption.

ASSUMPTION 3.8. *There exists* $\tau_{\min} > 0$ *such that*

$$\tau_{\min}(\nabla f(x_k)^T d_k^{true} + \beta_k (u_k^{true})^T H_k u_k^{true}) \leq (1-\sigma)(\|c_k\| - \|c_k + J_k v_k\|), \tag{3.15}$$

holds for all $k \in \mathbb{N}$.

As an immediate consequence of this assumption, by (3.11) we know that for all $k \in \mathbb{N}$,

$$\Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) \geq \tau_{\min}\beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \sigma(\|c_k\| - \|c_k + J_k v_k\|). \tag{3.16}$$

This generic bound on the model reduction frees us from computing τ_k values, thus avoiding the over-complication of the analysis and extra computation in practice, unlike in [3] which relies on stochastic estimates of τ_k at each iteration.

Event (ii) leads the merit function ϕ to ignore the objective f in the limit and to minimize the constraint violation $\|c_k\|$ only. Such behavior can occur even in *deterministic* SQP methods when $\{J_k\}$ does not have full rank, in which case the best one can hope for is convergence to an infeasible stationary point, which we prove below.

We consider the convergence of Algorithm 2.1 and associated complexity results under these different events. In fact, Assumption 3.8 could hold regardless of whether LICQ holds, which leads to different complexity results in terms of the infeasibility measure. Therefore, we will mainly discuss *three* cases in the following subsections.

3.2. Case I: Assumption 3.8 holds without constraint qualifications. In this subsection, we assume that $\tau_k \searrow \tau_{\min} > 0$ without any constraint qualifications. We first state a generic merit function descent lemma that relates the improvement of the merit function to the local model.

LEMMA 3.9. *Let Assumptions 2.1, 2.2, 2.3, and 3.8 hold, and* α_k *is chosen as in* (2.8). *Then if* $\alpha_k \leq 1$, *we have*

$$\begin{aligned} &\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\ &\leq -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{true}) + \frac{\alpha_k^2 \beta_k^2}{2}(\tau_{\min} L + \Gamma)\|u_k\|^2 \\ &\quad + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|v_k\|^2 + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{true}). \end{aligned} \tag{3.17}$$

Proof. By the L-Lipschitz continuity of $\nabla f(x)$ and Γ-Lipschitz continuity of J_k, we have

$$\begin{aligned} &\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\ &= \tau_{\min}[f(x + \alpha_k d_k) - f(x_k)] + \|c(x_k + \alpha_k d_k)\| - \|c(x_k)\| \\ &\leq \alpha_k \tau_{\min} \nabla f(x_k)^T d_k + \alpha_k \|c_k + J_k d_k\| + |1 - \alpha_k| \|c_k\| \\ &\quad - \|c_k\| + \frac{1}{2}(\tau_{\min} L + \Gamma)\alpha_k^2 \|d_k\|^2 \\ &= \alpha_k \tau_{\min} \nabla f(x_k)^T d_k - \alpha_k(\|c_k\| - \|c_k + J_k d_k\|) + \frac{1}{2}(\tau_{\min} L + \Gamma)\alpha_k^2 \|d_k\|^2. \end{aligned} \tag{3.18}$$

Since $d_k = \beta_k u_k + v_k$ and u_k, v_k are orthogonal, we have $\|d_k\|^2 = \beta_k^2\|u_k\|^2 + \|v_k\|^2$. Combining with the fact $J_k d_k = J_k d_k^{\text{true}}$, we have

$$\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\leq \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k(\|c_k\| - \|c_k + J_k d_k^{\text{true}}\|) \\
&\quad + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)(\beta_k^2\|u_k\|^2 + \|v_k\|^2) + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\
&\leq -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) + \frac{\alpha_k^2 \beta_k^2}{2}(\tau_{\min} L + \Gamma)\|u_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|v_k\|^2. \qquad \square
\end{aligned}$$

Lemma 3.9 relates the reduction in the merit function ϕ to the local model l, the norms of v_k and u_k, and the stochastic error in d_k. It also reveals how β_k^2 acts on $\|u_k\|^2$ to control the stochastic variance.

Recall that α_k is chosen from the interval $[\nu, \nu + \theta\beta_k]$. Now we choose

$$\nu \in \left(0, \min\left\{\frac{\sigma \kappa_v \kappa_c^{-1}}{2(\tau_{\min} L + \Gamma)\omega^2}, 1 - \theta\kappa_\beta\right\}\right], \tag{3.19}$$

which ensures $\alpha_k \leq 1$. A natural drawback of this step size choosing scheme is that we must have knowledge of the Lipschitz constants Γ, L, and the bound of the merit parameter $\tau_{\min}$, which may be unreasonable in practice. Practically, one could consider adaptive parameter choosing schemes such as AdaGrad or Adam style stepsizes, which will yield similar complexity results with additional logarithmic terms [25, 29].

LEMMA 3.10. *Let Assumptions 2.1, 2.2, 2.3, and 3.8 hold, then*

$$\mathbb{E}_k[\alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{true})] \leq \beta_k^2 \theta \tau_{\min} \kappa_g \zeta^{-1} \sqrt{M}.$$

Proof. Let $\xi_k \in [0, 1]$ be a random variable such that $\alpha_k = \nu + \theta\xi_k\beta_k$. Then from Lemma 3.5,

$$\begin{aligned}
&\mathbb{E}_k[\alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})] \\
&= \mathbb{E}_k[(\nu + \theta\xi_k\beta_k)\tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})] \\
&= \nu \tau_{\min} \nabla f(x_k)^T \mathbb{E}[d_k - d_k^{\text{true}}] + \mathbb{E}_k[\theta\xi_k\beta_k\tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})] \\
&\leq \theta \tau_{\min} \mathbb{E}_k[\|\nabla f(x_k)\| \|d_k - d_k^{\text{true}}\|] \\
&\leq \beta_k^2 \theta \tau_{\min} \kappa_g \zeta^{-1} \sqrt{M}. \qquad \square
\end{aligned}$$

With all of these results in hand, we can prove our first main theorem, which is stated generically for any choice of β_k.

THEOREM 3.11. *Let Assumptions 2.1, 2.2, 2.3, and 3.8 hold. Choose ν as in* (3.19)*, α_k as in* (2.8) *and let*

$$\kappa_1 := \frac{1}{2}(\tau_{\min} L + \Gamma)(\zeta^{-1}\kappa_H \kappa_u^2 + \zeta^{-2} M) + \theta\kappa_g \tau_{\min} \zeta^{-1}\sqrt{M} + \theta^2(\tau_{\min} L + \Gamma)\omega\kappa_J\kappa_c.$$

Then we have, for all $K \in \mathbb{N}$,

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{true})^T H_k u_k^{true} + \frac{1}{2}\alpha_k \sigma \kappa_v \kappa_c^{-1} \|J_k^T c_k\|^2] \\
&\leq \tau_{\min}(f_0 - f_{\inf}) + \tau_{\min}\|c_0\| + \kappa_1 \sum_{k=0}^{K-1} \beta_k^2.
\end{aligned}$$

Proof. Taking the conditional expectation of the k-th iterate on both sides of (3.17) and applying the results of Lemmas 3.6 and 3.10, we have

$$\begin{aligned}
&\mathbb{E}_k[\phi(x_k+\alpha_k d_k,\tau_{\min})]-\phi(x_k,\tau_{\min})\\
&\le \mathbb{E}_k[-\alpha_k\Delta l(x_k,\tau_{\min},\nabla f(x_k),d_k^{\text{true}})]+\frac{\alpha_k^2\beta_k^2}{2}(\tau_{\min}L+\Gamma)\mathbb{E}_k[\|u_k\|^2]\\
&\quad+\frac{\alpha_k^2}{2}(\tau_{\min}L+\Gamma)\|v_k\|^2+\mathbb{E}_k[\alpha_k\tau_{\min}\nabla f(x_k)^T(d_k-d_k^{\text{true}})]\\
&\le -\alpha_k\Delta l(x_k,\tau_{\min},\nabla f(x_k),d_k^{\text{true}})+(\theta\kappa_g\tau_{\min}\zeta^{-1}\sqrt{M})\beta_k^2+\frac{\alpha_k^2}{2}(\tau_{\min}L+\Gamma)\|v_k\|^2\\
&\quad+\frac{\alpha_k^2\beta_k^2}{2}(\tau_{\min}L+\Gamma)(\zeta^{-1}(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\zeta^{-2}M)\\
&\le -\alpha_k\Delta l(x_k,\tau_{\min},\nabla f(x_k),d_k^{\text{true}})+(\nu^2+\theta^2\beta_k^2)(\tau_{\min}L+\Gamma)\|v_k\|^2\\
&\quad+\beta_k^2[\frac{\alpha_k^2}{2}(\tau_{\min}L+\Gamma)(\zeta^{-1}(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\zeta^{-2}M)+\theta\kappa_g\tau_{\min}\zeta^{-1}\sqrt{M}]\\
&\le -\alpha_k\Delta l(x_k,\tau_{\min},\nabla f(x_k),d_k^{\text{true}})+\beta_k^2\kappa_1+\nu^2(\tau_{\min}L+\Gamma)\|v_k\|^2.
\end{aligned}$$

From (3.16) and Lemma 3.2, we have

$$\begin{aligned}
&\mathbb{E}_k[\phi(x_k+\alpha_k d_k,\tau_{\min})]-\phi(x_k,\tau_{\min})\\
&\le -\alpha_k(\tau_{\min}\beta_k(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\sigma(\|c_k\|-\|c_k+J_kv_k\|))\\
&\quad+\beta_k^2\kappa_1+\nu^2(\tau_{\min}L+\Gamma)\|v_k\|^2\\
&\le -\alpha_k\tau_{\min}\beta_k(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\beta_k^2\kappa_1\\
&\quad+(-\alpha_k\sigma+\nu^2(\tau_{\min}L+\Gamma)\frac{\omega^2\kappa_c}{\kappa_v})(\|c_k\|-\|c_k+J_kv_k\|)\\
&\quad\Big(\text{Since }\nu\le\min\Big\{\frac{\sigma\kappa_v}{2(\tau_{\min}L+\Gamma)\omega^2\kappa_c},1-\theta\kappa_\beta\Big\}\Big)\\
&\le -\alpha_k\tau_{\min}\beta_k(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\beta_k^2\kappa_1-\frac{1}{2}\alpha_k\sigma(\|c_k\|-\|c_k+J_kv_k\|) \qquad (3.20)\\
&\le -\alpha_k\tau_{\min}\beta_k(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\beta_k^2\kappa_1-\frac{1}{2}\alpha_k\sigma\kappa_v\kappa_c^{-1}\|J_k^Tc_k\|^2.
\end{aligned}$$

Taking the total expectation, rearranging and summing the inequalities from $k=0$ to $K-1$,

$$\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}[\alpha_k\beta_k\tau_{\min}(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\frac{1}{2}\alpha_k\sigma\kappa_v\kappa_c^{-1}\|J_k^Tc_k\|^2]\\
&\le\phi(x_0,\tau_{\min})-\mathbb{E}[\phi(x_K,\tau_{\min})]+\kappa_1\sum_{k=0}^{K-1}\beta_k^2.
\end{aligned}$$

Since $-\mathbb{E}[\phi(x_K,\tau_{\min})]=-\mathbb{E}[\tau_{\min}f(x_K)+\|c_K\|]\le-\tau_{\min}f_{\inf}$, we can conclude that

$$\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}[\alpha_k\beta_k\tau_{\min}(u_k^{\text{true}})^TH_ku_k^{\text{true}}+\frac{1}{2}\alpha_k\sigma\kappa_v\kappa_c^{-1}\|J_k^Tc_k\|^2]\\
&\le\phi(x_0,\tau_{\min})-\tau_{\min}f_{\inf}+\kappa_1\sum_{k=0}^{K-1}\beta_k^2.
\end{aligned}$$

∎

Theorem 3.11 shows that under Assumption 3.8 and $\sum\beta_k^2<\infty$, the infeasibility measure $\|J_k^Tc_k\|$ converges in expectation. The convergence rate is determined by problem-specific constants, such as L and Γ, and the starting point x_0. In fact, if the exact order of $\{\beta_k\}$ is known, we can obtain a worst-case complexity for $\mathbb{E}\|J_k^Tc_k\|$ as well as $\mathbb{E}\|\nabla f(x_k)+J_k^Ty_k^{\text{true}}\|$.

COROLLARY 3.12. *Let the assumptions of Theorem* 3.11 *hold. For any* $K \in \mathbb{N}$, *let* $\beta_k := \beta = \eta/\sqrt{K}$ *for all* $k \in [0, K-1]$, *where* $\eta > 0$. *Let* $\kappa_2 := \tau_{\min}(f_0 - f_{\inf}) + \|c_0\| + \eta^2\kappa_1$, *then*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \frac{2\kappa_c\kappa_2}{\nu\sigma\kappa_v K}.$$

Proof. From Theorem 3.11, the definition of β_k, Assumption 2.2, and the fact that $\nu \leq \alpha_k$ for all $k \in \mathbb{N}$, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \frac{2\kappa_c(\tau_{\min}(f_0 - f_{\inf}) + \|c_0\| + \kappa_1\sum\beta_k^2)}{\nu\sigma\kappa_v K} \leq \frac{2\kappa_c\kappa_2}{\nu\sigma\kappa_v K}. \qquad \square$$

COROLLARY 3.13. *Let the assumptions of Theorem* 3.11 *hold. For any* $K \in \mathbb{N}$, *let* $\beta_k := \beta = \eta/\sqrt{K}$ *for all* $k \in [0, K-1]$, *where* $\eta > 0$. *Let* y_k^{true} *be defined in* (3.3). *Then,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{true}\|^2] \leq \frac{2\kappa_H^2\kappa_2}{\nu\eta\tau_{\min}\zeta\sqrt{K}} + \frac{4\kappa_H^2\kappa_c\kappa_0\kappa_2}{\nu\sigma\kappa_v\zeta K}. \tag{3.21}$$

Proof. Similarly to Corollary 3.12, we know

$$\sum_{k=0}^{K-1}\mathbb{E}[\alpha_k\beta_k\tau_{\min}(u_k^{\text{true}})^T H_k u_k^{\text{true}}] \leq \phi(x_0, \tau_{\min}) - \tau_{\min}f_{\inf} + \kappa_1\sum_{k=0}^{K-1}\beta_k^2.$$

From Lemma 3.7, it follows that

$$\begin{aligned}&\sum_{k=0}^{K-1}\mathbb{E}[\frac{\alpha_k\beta_k\tau_{\min}\zeta}{2\kappa_H^2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2]\\ &\leq \phi(x_0, \tau_{\min}) - \tau_{\min}f_{\inf} + \kappa_1\sum_{k=0}^{K-1}\beta_k^2 + \kappa_0\sum_{k=0}^{K-1}\mathbb{E}[\alpha_k\beta_k\tau_{\min}\|J_k^T c_k\|^2].\end{aligned}$$

Substituting β_k with $\eta/\sqrt{K}$, rearranging the inequality, dividing both sides by K, and applying Corollary 3.12, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2] \leq \frac{2\kappa_H^2\kappa_2}{\nu\eta\tau_{\min}\zeta\sqrt{K}} + \frac{4\kappa_H^2\kappa_c\kappa_0\kappa_2}{\nu\sigma\kappa_v\zeta K},$$

which proves the result. $\square$

Corollary 3.12 and Corollary 3.13 prove our first set of complexity results. It is obvious that for any small tolerance $\epsilon_c, \epsilon_L > 0$, to achieve $\mathbb{E}\|J_k^T c_k\| \leq \epsilon_c$, we need at most $\mathcal{O}(\epsilon_c^{-2})$ iterations. To achieve $\mathbb{E}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\| \leq \epsilon_L$, we need at most $\mathcal{O}(\epsilon_L^{-4})$ iterations. To our knowledge, the complexity with respect to feasibility represents the best known result in the absence of a constraint qualification. The latter result, in terms of ϵ_L, is known to be *optimal* [1] for stochastic gradient methods to find first-order stationary points under our assumptions.

3.3. Case II: LICQ holds. In this subsection, we assume that the Linear Independence Constraint Qualification is satisfied for every iteration. The analysis is parallel to Section 3.2, however, under this assumption, we are able to prove tighter bounds with respect to convergence in the infeasibility measure and the gradient of the Lagrangian. It is worth mentioning that most of the existing literature on complexity results in constrained optimization assumes LICQ or other types of constraint qualification in their analysis, such as variants of MFCQ [10, 26]. We formally state LICQ as an assumption here.

ASSUMPTION 3.14. *(LICQ) For all* $k \in \mathbb{N}$, *the constraint Jacobian matrix* $J(x_k)$ *has full row rank. Equivalently, there exists a positive constant* $\sigma_J > 0$ *such that* $\|J_k^T c_k\| \geq \sigma_J\|c_k\|$.

The following lemma is an improved version of Lemma 3.2. Since LICQ holds for all iterations, we can now use the exact infeasibility measure $\varphi = \|c(x)\|$.

LEMMA 3.15. *Let Assumptions 2.1 and 3.14 hold for all $k \in \mathbb{N}$, then with $\|c_k\| > 0$,*

$$\|c_k\| - \|c_k + J_k v_k\| \geq \kappa_v \sigma_J \|J_k^T c_k\| \geq \kappa_v \sigma_J^2 \|c_k\|. \tag{3.22}$$

Proof. From Lemma 3.2,

$$\|c_k\|(\|c_k\| - \|c_k + J_k v_k\|) \geq \kappa_v \|J_k^T c_k\|^2 \geq \kappa_v \sigma_J \|J_k^T c_k\| \|c_k\| \geq \kappa_v \sigma_J^2 \|c_k\|^2.$$

Dividing both sides by $\|c_k\|$ yields both results. ∎

The following lemma shows that LICQ serves as a sufficient condition for the existence of $\tau_{\min} > 0$, thus implying that Assumption 3.8 holds throughout this subsection.

LEMMA 3.16. *Let Assumptions 2.1, 2.2, 2.3 and 3.14 hold and let $\sigma \in (0,1)$ and $\beta_k \leq \kappa_\beta$ hold for all $k \in \mathbb{N}$. Then, for all $k \in \mathbb{N}$, there exists such $\tau_{\min} := \frac{(1-\sigma)\kappa_v \sigma_J}{\omega(\kappa_\beta \kappa_H \kappa_u + \kappa_g)}$ satisfying*

$$\tau_{\min}(\nabla f(x_k)^T d_k^{true} + \beta_k (u_k^{true})^T H_k u_k^{true}) \leq (1-\sigma)(\|c_k\| - \|c_k + J_k v_k\|).$$

Proof. By the definition of d_k^{true},

$$\begin{aligned}\nabla f(x_k)^T d_k^{\text{true}} &= \nabla f(x_k)^T (\beta_k u_k^{\text{true}} + v_k) \\ &= -\beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} - \beta_k v_k^T H_k u_k^{\text{true}} + \nabla f(x_k)^T v_k.\end{aligned}$$

From the trust-region constraint in subproblem (2.3) and Lemma 3.15, we have

$$\begin{aligned}\nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} &= -\beta_k v_k^T H_k u_k^{\text{true}} + \nabla f(x_k)^T v_k \\ &\leq (\beta_k \kappa_H \|u_k^{\text{true}}\| + \|\nabla f(x_k)\|)\|v_k\| \\ &\leq (\kappa_\beta \kappa_H \kappa_u + \kappa_g)\omega \|J_k^T c_k\| \\ &\leq \omega(\kappa_\beta \kappa_H \kappa_u + \kappa_g) \frac{(\|c_k\| - \|c_k + J_k v_k\|)}{\kappa_v \sigma_J}.\end{aligned}$$

Multiplying $\tau_{\min}$ on both sides proves the result. ∎

We can now prove the main theorem for Case II.

THEOREM 3.17. *Let Assumptions 2.1, 2.2, 2.3 and 3.14 hold for all $k \in \mathbb{N}$. Choose $\beta_k = \eta/\sqrt{K}$ and α_k as in (2.8). Then for any $K \in \mathbb{N}$,*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|c_k\|] \leq \frac{2\kappa_2}{\nu \sigma \kappa_v \sigma_J^2 K}, \tag{3.23}$$

and

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{true}\|^2] \leq \frac{2\kappa_H^2 \kappa_2}{\nu \eta \tau_{\min} \zeta \sqrt{K}} + \frac{4\kappa_2 \kappa_H^2 \kappa_0 \kappa_J^2 \kappa_c}{\nu \sigma \kappa_v \sigma_J^2 \zeta K}. \tag{3.24}$$

Proof. By Lemma 3.16, we know that $\tau_{\min} > 0$ exists. Therefore, the conditions for Theorem 3.11 are satisfied and thus, by (3.20) and $\beta_k = \eta/\sqrt{K}$,

$$\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{1}{2}\alpha_k \sigma(\|c_k\| - \|c_k + J_k v_k\|)] \leq \phi(x_0, \tau_{\min}) - \tau_{\min} f_{\inf} + \kappa_1 \eta^2.$$

From Lemma 3.15, we have

$$\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{1}{2}\alpha_k \sigma \kappa_v \sigma_J^2 \|c_k\|] \leq \phi(x_0, \tau_{\min}) - \tau_{\min} f_{\inf} + \kappa_1 \eta^2.$$

Therefore, by Assumption 2.2,

$$\sum_{k=0}^{K-1} \frac{1}{2}\nu\sigma\kappa_v\sigma_J^2\mathbb{E}[\|c_k\|] \leq \kappa_2.$$

Dividing both sides by K yields the first result. Next, from Lemma 3.7, we have

$$\begin{aligned}&\sum_{k=0}^{K-1} \mathbb{E}\left[\frac{\zeta\alpha_k\beta_k\tau_{\min}}{2\kappa_H^2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - \alpha_k\beta_k\tau_{\min}\kappa_0\|J_k^T c_k\|^2\right]\\ &\leq \sum_{k=0}^{K-1} \mathbb{E}[\alpha_k\beta_k\tau_{\min}(u_k^{\text{true}})^T H_k u_k^{\text{true}}] \leq \kappa_2.\end{aligned}$$

Therefore,

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2] \leq \frac{2\kappa_H^2\kappa_2\sqrt{K}}{\nu\eta\tau_{\min}\zeta} + 2\zeta^{-1}\kappa_H^2\kappa_0\kappa_J^2\kappa_c\sum_{k=0}^{K-1}\mathbb{E}[\|c_k\|]. \tag{3.25}$$

Dividing both sides by K and using (3.23), yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2] \leq \frac{2\kappa_H^2\kappa_2}{\nu\eta\tau_{\min}\zeta\sqrt{K}} + \frac{4\kappa_2\kappa_H^2\kappa_0\kappa_J^2\kappa_c}{\nu\sigma\kappa_v\sigma_J^2\zeta K}.$$

∎

Parallel to Corollary 3.12 and 3.13, we can obtain worst-case complexity results from Theorem 3.17. To achieve $\mathbb{E}[\|c_k\|] \leq \epsilon_c$, we only need at most $\mathcal{O}(\epsilon_c^{-1})$ iterations. This result matches the complexity of the algorithm proposed in [25] to reach an ϵ_c-feasible point (although it uses a l_1-norm and assumes the exact d_k is accessible), which is the best known result when LICQ holds. Clearly, the improvement when compared with the results of Section 3.2 is directly as a result of Assumption 3.14. On the other hand, we still require at most $\mathcal{O}(\epsilon_L^{-4})$ iterations to achieve $\mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|] \leq \epsilon_L$, matching the optimal complexity with respect to this measure.

3.4. Case III: Merit parameter sequence vanishes. For the analysis in the previous subsections, we assumed that $\tau_{\min} > 0$ exists, even though our algorithm does not directly compute it. Unfortunately, in certain cases, such $\tau_{\min}$ does not exist, *i.e.,*the merit parameter sequence $\{\tau_k\}$ may decrease to 0 in the limit. This behavior is fundamentally tied to the existence of a (sub)sequence of iterates along which the constraint Jacobians tends toward rank deficiency. In such a case, the best result we can hope for is convergence to an infeasible stationary point, i.e., convergence in $\|J_k^T c_k\|$. In this setting, we can prove our convergence result under the less restrictive choice of ν, given by,

$$\nu \in \left(0, \min\left\{\frac{\kappa_v\kappa_c^{-1}}{2\Gamma\omega^2}, 1 - \theta\kappa_\beta\right\}\right]. \tag{3.26}$$

Armed with this definition, we now derive the fundamental lemma of this section.

LEMMA 3.18. *Under Assumptions 2.1, 2.2 and 2.3, it follows for all $k \in \mathbb{N}$ that*

$$\mathbb{E}_k[\|c_k\| - \|c(x_k + \alpha_k d_k)\|] \geq \frac{1}{2}\nu\kappa_v\kappa_c^{-1}\mathbb{E}_k[\|J_k^T c_k\|^2] - \kappa_3\beta_k^2, \tag{3.27}$$

where we define $\kappa_3 := \Gamma(\theta^2\kappa_J^2\kappa_c^2\omega^2 + \frac{1}{2}(\zeta^{-1}\kappa_H\kappa_u^2 + \zeta^{-2}M))$.

Proof. By Lipschitz continuity of the Jacobian of c and $\alpha_k \leqslant 1$,

$$\begin{aligned}
\|c_k\| - \|c(x_k + \alpha_k d_k)\| &\geq \alpha_k(\|c_k\| - \|c_k + J_k v_k\|) - \frac{\Gamma}{2}\alpha_k^2\|d_k\|^2 \\
&\geq \nu\kappa_v\kappa_c^{-1}\|J_k^T c_k\|^2 - \frac{\Gamma}{2}\alpha_k^2(\|v_k\|^2 + \beta_k^2\|u_k\|^2) \\
&\geq \nu\kappa_v\kappa_c^{-1}\|J_k^T c_k\|^2 - \Gamma(\nu^2 + \theta^2\beta^2)\|v_k\|^2 - \frac{\Gamma\beta_k^2}{2}\|u_k\|^2 \\
&\geq \nu\kappa_v\kappa_c^{-1}\|J_k^T c_k\|^2 - \Gamma\nu^2\omega^2\|J_k^T c_k\|^2 - \beta_k^2\Gamma\theta^2\kappa_J^2\kappa_c^2\omega^2 - \frac{\Gamma\beta_k^2}{2}\|u_k\|^2 \\
&\geq \frac{1}{2}\nu\kappa_v\kappa_c^{-1}\|J_k^T c_k\|^2 - \beta_k^2\Gamma\theta^2\kappa_J^2\kappa_c^2\omega^2 - \frac{\Gamma\beta_k^2}{2}\|u_k\|^2
\end{aligned}$$

Taking the conditional expectation on both sides and applying Lemmas 3.4 and 3.6 yields the result. ∎

Now we present our main result for this section.

THEOREM 3.19. *Let Assumptions 2.1, 2.2 and 2.3 hold for all $k \in \mathbb{N}$. Set $\beta_k = \eta/\sqrt{K}$, α_k as in* (2.8)*, and ν as in* (3.26)*. Then for any $K \in \mathbb{N}$,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \frac{2\kappa_c\left(\|c_0\| + \eta^2\kappa_3\right)}{\nu\kappa_v K}. \tag{3.28}$$

Proof. Summing (3.27) for $k = 0, \ldots, K-1$ and rearranging, we have

$$\frac{1}{2}\nu\kappa_v\kappa_c^{-1}\sum_{k=0}^{K-1}\mathbb{E}_k[\|J_k^T c_k\|^2] \leq \sum_{k=0}^{K-1}\mathbb{E}_k[\|c_k\| - \|c(x_{k+1})\|] + \kappa_3\beta_k^2.$$

Taking the total expectation and applying $\beta_k = \eta/\sqrt{K}$, we have

$$\frac{1}{2}\nu\kappa_v\kappa_c^{-1}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \|c_0\| + \eta^2\kappa_3.$$

Rearranging terms and dividing through by K yields the result. ∎

To our knowledge, this represents the first complexity result for convergence to an infeasible stationary point, and matches the $\mathcal{O}(\epsilon_c^{-2})$ complexity result obtained in Subsection 3.2. While convergence to an infeasible stationary point is obviously undesired, it is the best possible convergence behavior one can hope for under such mild assumptions, even when f is deterministic.

4. The Inexact Version of Stochastic SQP Algorithm. In our previous discussion, we assumed that an *exact* solution of the subproblem (2.5) is obtained so that condition $u_k \in \text{Null}(J_k)$, *i.e.,* $J_k u_k = 0$ is strictly satisfied. However, such an exact u_k is expensive to compute in practice, especially when the problem is large-scale. More importantly, since u_k arises from the stochastic subproblem (2.5) involving g_k, it is natural that we need not expend too much computation solving for u_k, which is fundamentally corrupted by stochastic noise. Therefore, we turn to *inexact* solvers for the subproblem (2.5). We introduce termination tests, which are used as a stopping criterion for iterative linear system solvers such as MINRES and enable our analysis to track the residuals of the linear system. We now introduce and (re)define some variables for the analysis in this section.

In this section, we assume u_k and y_k are computed using an inexact iterative linear system solver. This solver also generates a pair of residuals $(\rho_k, r_k) \in \mathbb{R}^n \times \mathbb{R}^m$ such that, for every iteration $k \in \mathbb{N}$, (ρ_k, r_k) satisfy our termination tests. We denote the exact solution of the subproblem (2.5) by $u_{k,*}$, which lies in the Null space of J_k (*i.e.,*$u_{k,*}$ is the u_k used

in previous sections). Recalling the general KKT system (2.6) of the tangential subproblem, we define the residual pair (ρ_k, r_k) formally as follows:

$$\begin{bmatrix}\rho_k\\ r_k\end{bmatrix} := \begin{bmatrix}H_k & J_k^T\\ J_k & 0\end{bmatrix}\begin{bmatrix}u_k\\ y_k\end{bmatrix} + \begin{bmatrix}g_k + H_k v_k\\ 0\end{bmatrix}. \tag{4.1}$$

We make the following assumptions on the iterative solver we use. Note that Assumption 4.1 is a common property of linear system solvers. In fact, Krylov Subspace solvers, such as MINRES, can produce an exact solution *i.e.,* $(\rho_k, r_k) = (0, 0)$ after $t = n + m$ iterations.

ASSUMPTION 4.1. *For all $k \in \mathbb{N}$, the iterative system solver used to compute u_k and y_k generates a sequence $\{(u_{k,t}, \rho_{k,t}, r_{k,t})\}_{t\geq 0}$ with*

$$\begin{bmatrix}\rho_{k,t}\\ r_{k,t}\end{bmatrix} := \begin{bmatrix}H_k & J_k^T\\ J_k & 0\end{bmatrix}\begin{bmatrix}u_{k,t}\\ y_{k,t}\end{bmatrix} + \begin{bmatrix}g_k + H_k v_k\\ 0\end{bmatrix}, \tag{4.2}$$

such that $\lim_{t\to\infty} \|(u_{k,t}, \rho_{k,t}, r_{k,t}) - (u_{k,*}, 0, 0)\| = 0$.

Note that, since the y_k part of the solution to (2.5) is not unique when J_k is rank deficient and we do not need y_k in our analysis, we do not include it in the assumption. Formally, we ask for the following termination tests of the iterative linear system solver.

TERMINATION TEST 1. *Let $(\gamma_r, \gamma_\rho) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and $\beta_k \in \mathbb{R}_{>0}$ be a predefined rescaling parameter. For all $k \in \mathbb{N}$, we terminate the iterative linear solver as long as the residuals satisfy the following conditions:*

$$\|r_k\| \leq \gamma_r \beta_k \quad \textit{and} \quad \|\rho_k\| \leq \gamma_\rho \beta_k.$$

Under Assumption 4.1, the termination test is always satisfied for all sufficiently large t.

Since our inexact u_k is no longer contained in the Null space of J_k now, we first prove a few parallel lemmas to Lemma 3.6 to continue our convergence analysis.

We now derive the explicit form of our inexact u_k. With J_k^+, we can write u_k as the decomposition $u_k = u_{k,1} + u_{k,2}$, where $u_{k,1} \in \text{Range}(J_k)$ and $u_{k,2} \in \text{Null}(J_k)$.

From (4.1), we have

$$u_{k,1} = J_k^+ r_k, \quad u_{k,2} = -Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k + H_k J_k^+ r_k - \rho_k). \tag{4.3}$$

Thus the *inexact* solution u_k of (4.1) is

$$u_k = J_k^+ r_k - Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k + H_k J_k^+ r_k - \rho_k). \tag{4.4}$$

Also, recall that the tangential component computed exactly with the true gradient is

$$u_k^{\text{true}} = -Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (\nabla f(x_k) + H_k v_k).$$

To derive a uniform bound on $\|u_k\|^2$, we make the following assumption on the behavior of constraint Jacobian matrices $\{J_k\}$.

ASSUMPTION 4.2. *For all $k \in \mathbb{N}$, the nonzero singular values of J_k is uniformly lower bounded by some positive value $\bar{\sigma}_{\min} \in \mathbb{R}_{>0}$.*

This assumption covers the LICQ case, where all the singular values of the constraint Jacobians are positive and bounded from below. To clarify, since the assumption does not require all singular values to be positive, it includes certain "ideal" cases of rank-deficient Jacobians. We note that this requirement is necessary to provide any meaningful bound on $\|u_{k,1}\|$ and thus is fundamental to our analysis with an inexact u_k. In particular, Assumption 4.2 yields a uniform upper bound that is $\|J_k^+\| \leq 1/\bar{\sigma}_{\min}$ for all $k \in \mathbb{N}$.

LEMMA 4.1. *Let Assumptions 4.1 and 4.2 hold. Then, for all $k \in \mathbb{N}$, $\|\mathbb{E}_k[u_k - u_k^{true}]\| \leq \kappa_4 \beta_k$, where $\kappa_4 := (\bar{\sigma}_{\min}^{-1}\gamma_r + \zeta^{-1}\gamma_\rho)$.*

Proof. From (4.4) we know

$$\begin{aligned}\mathbb{E}_k[u_k - u_k^{\text{true}}] &= \mathbb{E}_k[J_k^+ r_k - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T (g_k - \nabla f(x_k)) \\ &\quad - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k J_k^+ r_k + Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T \rho_k] \\ &= \mathbb{E}_k[(I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^+ r_k] + \mathbb{E}_k[Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T \rho_k].\end{aligned}$$

From [3, Lemma 11], we know $\|I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k\| \leq 1$ and $\|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\| \leq \zeta^{-1}$. Taking the norm of both sides and using Cauchy-Schwarz and the triangle inequality, we have

$$\begin{aligned}\|\mathbb{E}_k(u_k - u_k^{\text{true}})\| &\leq \|\mathbb{E}_k[(I - Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T H_k) J_k^+ r_k]\| \\ &\quad + \|\mathbb{E}_k[Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T \rho_k]\| \\ &\leq \frac{1}{\bar{\sigma}_{\min}} \gamma_r \beta_k + \zeta^{-1} \gamma_\rho \beta_k,\end{aligned}$$

which proves the result. ∎

LEMMA 4.2. *Let Assumptions 2.1, 2.2, 2.3, 4.1, and 4.2 hold. For all $k \in \mathbb{N}$, $\mathbb{E}_k\|u_k\|^2$ is bounded from above by $\tilde{\kappa}_u := 2\kappa_u^2 + 2(\zeta^{-1}\sqrt{M} + \kappa_4\kappa_\beta)^2$.*

Proof. From [3, Lemma 11], we know that $\mathbb{E}_k[\|u_k - u_k^{\text{true}}\|] \leq \zeta^{-1}\sqrt{M} + \kappa_4\beta_k$. Therefore, $\mathbb{E}_k\|u_k\|^2 \leq \mathbb{E}_k[2(\|u_k^{\text{true}}\|^2 + \|u_k - u_k^{\text{true}}\|^2)] \leq 2\kappa_u^2 + 2(\zeta^{-1}\sqrt{M} + \kappa_4\kappa_\beta)^2 = \tilde{\kappa}_u$. ∎

Now, we present our first main result of this section, when Assumption 3.8 holds as well.

THEOREM 4.3. *Let Assumptions 2.1, 2.2, 2.3, 3.8, 4.1, and 4.2 hold and let $\beta_k = \eta/\sqrt{K}$ and $\alpha_k \in [\nu, \nu + \theta\beta_k]$. Let*

$$\nu \in \left(0, \min\left\{\frac{\sigma\kappa_v\kappa_c^{-1}}{4(\tau_{\min}L + \Gamma)\omega^2}, 1 - \theta\kappa_\beta\right\}\right),$$

and define

$$\tilde{\kappa}_1 := \tau_{\min}\kappa_g\kappa_4 + (\tau_{\min}L + \Gamma)\tilde{\kappa}_u + 2\theta^2(\tau_{\min}L + \Gamma)\omega^2\kappa_J^2\kappa_c^2 + \gamma_r,$$

$$\textit{and} \quad \tilde{\kappa}_2 := \tau_{\min}(f_0 - f_{\inf}) + \|c_0\| + \eta^2\tilde{\kappa}_1.$$

Then, we have

$$\sum_{k=0}^{K-1} \mathbb{E}[\nu\beta_k\tau_{\min}(u_k^{true})^T H_k u_k^{true} + \frac{1}{2}\nu\sigma(\|c_k\| - \|c_k + J_k v_k\|)] \leq \tilde{\kappa}_2. \tag{4.5}$$

Proof. Similar to the proof of Lemma 3.9,

$$\begin{aligned}&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\ &\leq -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f_k, d_k^{\text{true}}) + \alpha_k\tau_{\min}\nabla f_k^T(d_k - d_k^{\text{true}}) \\ &\quad - \alpha_k(\|c_k + J_k v_k\| - \|c_k + J_k v_k + \beta_k r_k\|) + \frac{\alpha_k^2}{2}(\tau_{\min}L + \Gamma)\|d_k\|^2.\end{aligned}$$

Using $\|d_k\|^2 \leq 2(\beta_k^2\|u_k\|^2 + \|v_k\|^2)$ and rearranging the order of the inequality, we have

$$\begin{aligned}&\alpha_k \Delta l(x_k, \tau_{\min}\nabla f_k, d_k^{\text{true}}) \\ &\leq (\phi(x_k, \tau_{\min}) - \phi(x_k + \alpha_k d_k, \tau_{\min})) + \alpha_k\tau_{\min}\nabla f_k^T(d_k - d_k^{\text{true}}) \\ &\quad + \alpha_k(\|c_k + J_k v_k + \beta_k r_k\| - \|c_k + J_k v_k\|) + (\tau_{\min}L + \Gamma)\alpha_k^2(\beta_k^2\|u_k\|^2 + \|v_k\|^2).\end{aligned}$$

By the definition of ν and α_k, taking the conditional expectation, by (3.11),

$$\begin{aligned}
&\nu\mathbb{E}_k[\tau_{\min}\beta_k(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \sigma(\|c_k\| - \|c_k + J_k v_k\|)] \\
&\leq \mathbb{E}_k[(\phi(x_k,\tau_{\min}) - \phi(x_k+\alpha_k d_k,\tau_{\min})] + \mathbb{E}_k[\alpha_k\tau_{\min}\nabla f_k^T(d_k - d_k^{\text{true}})] \\
&\quad + \alpha_k\beta_k\mathbb{E}_k[\|r_k\|] + \mathbb{E}_k[\alpha_k^2\beta_k^2(\tau_{\min}L+\Gamma)\|u_k\|^2] \\
&\quad + 2\nu^2(\tau_{\min}L+\Gamma)\omega^2\|J_k^T c_k\|^2 + 2\theta^2\beta_k^2(\tau_{\min}L+\Gamma)\omega^2\kappa_J^2\kappa_c^2.
\end{aligned}$$

Working with the second term on the right-hand side of this inequality, by Lemma 4.2,

$$\mathbb{E}_k[\alpha_k\tau_{\min}\nabla f_k^T(d_k - d_k^{\text{true}})] = \mathbb{E}_k[\alpha_k\tau_{\min}\beta_k\nabla f_k^T(u_k - u_k^{\text{true}})] \leq \beta_k^2\tau_{\min}\kappa_g\kappa_4. \tag{4.6}$$

Since $\alpha_k \leq 1$ and $\|r_k\| \leq \gamma_r\beta_k$, we know

$$\alpha_k\beta_k\mathbb{E}_k\|r_k\| \leq \gamma_r\beta_k^2.$$

By the choice of ν and Lemma 3.2 we have

$$2\nu^2(\tau_{\min}L+\Gamma)\omega^2\|J_k^T c_k\|^2 \leq \frac{1}{2}\nu\sigma(\|c_k\| - \|c_k + J_k v_k\|).$$

Combining the results in Lemma 4.1 and 4.2, we can conclude that

$$\begin{aligned}
&\mathbb{E}_k[\nu\tau_{\min}\beta_k(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{1}{2}\nu\sigma(\|c_k\| - \|c_k + J_k v_k\|)] \\
&\leq \mathbb{E}_k[\phi(x_k,\tau_{\min}) - \phi(x_k+\alpha_k d_k,\tau_{\min})] \\
&\quad + \beta_k^2(\tau_{\min}\kappa_g\kappa_4 + (\tau_{\min}L+\Gamma)\tilde{\kappa}_u + 2\theta^2(\tau_{\min}L+\Gamma)\omega^2\kappa_J^2\kappa_c^2 + \gamma_r) \\
&= \mathbb{E}_k[\phi(x_k,\tau_{\min}) - \phi(x_k+\alpha_k d_k,\tau_{\min})] + \tilde{\kappa}_1\beta_k^2.
\end{aligned}$$

Thus, summing from $k = 0$ to $K-1$, and substituting $\beta_k = \eta/\sqrt{K}$, we have

$$\sum_{k=0}^{K-1}\mathbb{E}[\nu\beta_k\tau_{\min}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{1}{2}\nu\sigma(\|c_k\| - \|c_k + J_k v_k\|)] \leq \tilde{\kappa}_2. \tag{4.7}$$

∎

Theorem 4.3 is an essentially parallel result to Theorem 3.11. The difference lies mainly in the constants, as $\tilde{\kappa}_1$ includes additional terms that are dependent on γ_ρ and γ_r. This is natural in that our Termination Test 1 forces the errors of residuals below a threshold that is controlled by β_k.

For the sake of brevity, we only list the complexity results of the inexact algorithm under two different cases. Their proofs follow essentially in the same manner as those of Corollary 3.12, 3.13, and Theorem 3.17.

COROLLARY 4.4. *Under the same assumptions as in Theorem 4.3, we have the following results:*

1. *When no constraint qualification is assumed, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \frac{2\kappa_c\tilde{\kappa}_2}{\nu\sigma\kappa_v K}, \text{ and}$$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{true}\|^2] \leq \frac{2\kappa_H^2\tilde{\kappa}_2}{\nu\eta\tau_{\min}\zeta\sqrt{K}} + \frac{4\kappa_H^2\kappa_c\kappa_0\tilde{\kappa}_2}{\nu\sigma\kappa_v\zeta K}.$$

2. *When LICQ holds for all iterate k, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|c_k\|] \leq \frac{2\tilde{\kappa}_2}{\nu\sigma\kappa_v\sigma_J^2 K}, \text{ and}$$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{true}\|^2] \leq \frac{2\kappa_H^2\tilde{\kappa}_2}{\nu\eta\tau_{\min}\zeta\sqrt{K}} + \frac{4\tilde{\kappa}_2\kappa_H^2\kappa_0\kappa_J^2\kappa_c}{\nu\sigma\kappa_v\sigma_J^2\zeta K}.$$

Again, we need at most $\mathcal{O}(\epsilon_c^{-2})$ and $\mathcal{O}(\epsilon_c^{-1})$ iterations to reach a ϵ_c-stationary feasible point for case I and case II, respectively. The only difference lies in the constant. For example, the inexact algorithm needs roughly $(\tilde{\kappa}_2 - \kappa_2) \times 2(\kappa_c/\nu\sigma\kappa_v)$ more iterations in the worst case. With respect to the gradient of the Lagrangian, both cases yield the same $\mathcal{O}(\epsilon_L^{-4})$ complexity.

Next, we turn our attention to the case where $\tau_{\min}$ does not exist and provide a lemma parallel to Lemma 3.18.

LEMMA 4.3. *Let ν be defined by* (3.26)*. Then, under Assumptions 2.1, 2.2, 2.3, 4.1, and 4.2, it follows for all $k \in \mathbb{N}$ that*

$$\mathbb{E}_k[\|c_k\| - \|c(x_k + \alpha_k d_k)\|] \geq \frac{1}{2}\nu\kappa_v\kappa_c^{-1}\mathbb{E}_k\|J_k^T c_k\|^2 - \tilde{\kappa}_3\beta_k^2, \tag{4.8}$$

where we define

$$\tilde{\kappa}_3 := \Big[\Gamma(\frac{1}{2}\tilde{\kappa}_u + \theta^2\kappa_J^2\kappa_c^2\omega^2) + \gamma_r\Big].$$

Proof. The proof follows directly by the proof of Lemma 3.18 combined with the techniques used in the proof of Theorem 4.3. ∎

Given this, we can present our final complexity result.

THEOREM 4.4. *Let Assumptions 2.1, 2.2, 2.3, 4.1, and 4.2 hold for all $k \in \mathbb{N}$. Set $\beta_k = \eta/\sqrt{K}$, α_k as in* (2.8)*, and ν as in* (3.26)*. Then for any $K \in \mathbb{N}$,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|J_k^T c_k\|^2] \leq \frac{2\kappa_c\left(\|c_0\| + \eta^2\tilde{\kappa}_3\right)}{\nu\kappa_v K}. \tag{4.9}$$

Proof. The proof follows via an identical argument as the proof of Theorem 3.19. ∎

Clearly, this result parallels that of Theorem 3.19 and proves a worst-case complexity of $\mathcal{O}(\epsilon_c^{-2})$ with respect to our infeasibility measure. Thus, all of our complexity results translate directly to the setting of inexact solves for u_k, under Assumptions 4.1 and 4.2.

5. Numerical Experiments. In this section, we validate the performance of our inexact two-stepsize SQP (ITSQP) method with numerical experiments. We tested ITSQP together with the original stochastic SQP method (SSQP) [4] and a step-lengthening stochastic SQP method (SSQPL) on a subset of the equality constrained optimization problems from the CUTEst collection [16]. Since the singularity of Jacobians is detected in more than half of the problems, we use the step decomposition strategy for all three algorithms to make the results comparable.

We use an experiment setup similar to [4], with a total of 110 equality constrained problems. Multiple levels of stochasticity (noise that follows multivariate normal distribution $\mathcal{N}(0, \epsilon_N I)$, where $\epsilon_N \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$) are injected into the gradient estimates to simulate the stochastic gradient. To "tune" the algorithms, we tested 5 different fixed scaling stepsizes $\beta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ under 15 different seeds. Therefore, a total of 440(=110×4) problems are tested, and for each problem, every algorithm is run 75(=5 × 15) times with different stepsizes. For the stepsize α_k, we use the default stepsize selection scheme in [4] for SSQP, a step-lengthening stepsize for SSQPL and the adaptive α_k proposed in [25, Algorithm 4.1] for ITSQP.

For each run of the algorithm, we give a budget of 10,000 iterations and report the best iterate with the following scheme: for any iterate k, if $\|c_k\|_\infty \leq 10^{-6}$, we treat it as a feasible point. We then pick the best iterate with the lowest KKT error $\|\nabla f_k + J_k^T y_k^{\text{true}}\|_\infty$ from all feasible points. If $\|\nabla f_k + J_k^T y_k^{\text{true}}\|_\infty \leq 10^{-4}$ on an iterate satisfying $\|c_k\|_\infty \leq 10^{-6}$, we terminate the algorithm. If such feasible points do not exist, we report the first-order measure of the most feasible point (one with the lowest $\|c_k\|_\infty$). This is commonly used in the stochastic SQP literature, such as [13] and [25].

The figures contain the KKT errors and infeasibility errors of three algorithms tested under different noise levels. For each of the 440 problems, we compared the average infeasibility errors and KKT errors over five different stepsizes and chose the one stepsize that

yields the best results to plot the figures. It is $\beta_k = 1$ for SSQP and SSQPL, and $\beta_k = 10^{-3}$ for ITSQP.

We can conclude from the figures that our ITSQP method outperforms the SSQP and SSQPL methods in terms of the infeasibility measure across all noise levels. This advantage becomes increasingly pronounced as the noise level increases, highlighting the superior convergence rate of our two-stepsize algorithm.

The SSQP and SSQPL methods compute the merit parameter sequence τ_k and choose the stepsize based on τ_k, which appears to help them achieve better KKT convergence. The higher first-order errors may also be attributed to the more pessimistic stepsize α_k we use in our algorithm.

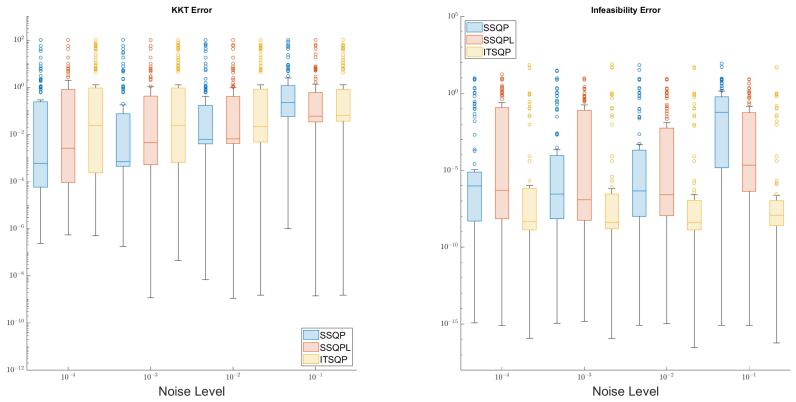

FIG. 1. *Comparison of KKT error and infeasibility error under different noise levels*

6. Conclusion. In this paper, we proposed and analyzed an inexact two-stepsize stochastic SQP method that can handle the possibility of rank deficient Jacobians. We prove the first-known $\mathcal{O}(\epsilon_c^{-2})$ complexity for the infeasibility measure $\|J_k^T c_k\|$ to fall below ϵ_c under mild assumptions, and an improved $\mathcal{O}(\epsilon_c^{-1})$ complexity when LICQ holds, which matches the best known result in the literature. These results also hold in the case where the tangential component is computed inexactly. Numerical experiments also show that our algorithm converges more efficiently in terms of the infeasibility measure than those without the two-stepsize scheme.

Extending similar SQP strategies to the more general, inequality constrained setting, remains an open problem. In addition, extending this method to the case of stochastic equality constraints is a natural future direction.

REFERENCES

[1] Y. ARJEVANI, Y. CARMON, J. C. DUCHI, D. J. FOSTER, N. SREBRO, AND B. WOODWORTH, *Lower bounds for non-convex stochastic optimization*, Mathematical Programming, 199 (2023), pp. 165–214.

[2] A. S. BERAHAS, R. BOLLAPRAGADA, AND B. ZHOU, *An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization*, arXiv preprint arXiv:2206.00712, (2022).

[3] A. S. BERAHAS, F. E. CURTIS, M. J. O'NEILL, AND D. P. ROBINSON, *A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient jacobians*, Mathematics of Operations Research, 49 (2024), pp. 2212–2248.

[4] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou, *Sequential quadratic optimization for nonlinear equality constrained stochastic optimization*, SIAM Journal on Optimization, 31 (2021), pp. 1352–1379.
[5] A. S. Berahas, J. Shi, Z. Yi, and B. Zhou, *Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction*, Computational Optimization and Applications, 86 (2023), pp. 79–116.
[6] A. S. Berahas, J. Shi, and B. Zhou, *Optimistic noise-aware sequential quadratic programming for equality constrained optimization with rank-deficient jacobians*, arXiv preprint arXiv:2503.06702, (2025).
[7] A. S. Berahas, M. Xie, and B. Zhou, *A sequential quadratic programming method with high-probability complexity bounds for nonlinear equality-constrained stochastic optimization*, SIAM Journal on Optimization, 35 (2025), pp. 240–269.
[8] J. T. Betts, *Practical methods for optimal control and estimation using nonlinear programming*, SIAM, 2010.
[9] R. Bollapragada, C. Karamanli, B. Keith, B. Lazarov, S. Petrides, and J. Wang, *An adaptive sampling augmented lagrangian method for stochastic optimization with deterministic constraints*, Computers & Mathematics with Applications, 149 (2023), pp. 239–258.
[10] D. Boob, Q. Deng, and G. Lan, *Level constrained first order methods for function constrained optimization*, Mathematical Programming, 209 (2025), pp. 1–61.
[11] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Mathematical programming, 89 (2000), pp. 149–185.
[12] H. Chen, G. E. C. Flores, and C. Li, *Physics-informed neural networks with hard linear equality constraints*, Computers & Chemical Engineering, 189 (2024), p. 108764.
[13] F. E. Curtis, D. P. Robinson, and B. Zhou, *A stochastic inexact sequential quadratic optimization algorithm for nonlinear equality-constrained optimization*, INFORMS Journal on Optimization, 6 (2024), pp. 173–195.
[14] Y. Fang, S. Na, M. W. Mahoney, and M. Kolar, *Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems*, SIAM Journal on Optimization, 34 (2024), pp. 2007–2037.
[15] S. Ghadimi and G. Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM journal on optimization, 23 (2013), pp. 2341–2368.
[16] S. Gratton and P. L. Toint, *S2mpj and cutest optimization problems for matlab, python and julia*, Optimization Methods and Software, (2025), pp. 1–33.
[17] Z. Jiang, C. Liu, Y. M. Lee, C. Hegde, S. Sarkar, and D. Jiang, *The stochastic augmented lagrangian method for domain adaptation*, Knowledge-Based Systems, 235 (2022), p. 107593.
[18] L. Jin and X. Wang, *A stochastic primal-dual method for a class of nonconvex constrained optimization*, Computational Optimization and Applications, 83 (2022), pp. 143–180.
[19] D. P. Kouri and D. Ridzal, *Inexact trust-region methods for pde-constrained optimization*, in Frontiers in PDE-constrained optimization, Springer, 2018, pp. 83–121.
[20] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, *Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization*, Computational Optimization and Applications, 87 (2024), pp. 117–147.
[21] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson, *Physics-informed neural networks with hard constraints for inverse design*, SIAM Journal on Scientific Computing, 43 (2021), pp. B1105–B1132.
[22] S. Na, M. Anitescu, and M. Kolar, *An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians*, Mathematical Programming, 199 (2023), pp. 721–791.
[23] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 2006.
[24] E. O. Omojokun, *Trust region algorithms for optimization with nonlinear equality and inequality constraints*, University of Colorado at Boulder, 1989.
[25] M. J. O'Neill, *A two stepsize sqp method for nonlinear equality constrained stochastic optimization*, arXiv preprint arXiv:2408.16656, (2024).
[26] Q. Shi, X. Wang, and H. Wang, *A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization*, Mathematics of Operations Research, (2025).
[27] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM Journal on Numerical Analysis, 20 (1983), pp. 626–637, https://doi.org/10.1137/0720042.
[28] K. L. Teo, B. Li, C. Yu, V. Rehbock, et al., *Applied and computational optimal control*,

Optimization and Its Applications, (2021).
[29] Q. Wang, C. Piermarini, Y. Zhu, and F. E. Curtis, *Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning*, arXiv preprint arXiv:2601.11795, (2026).
[30] Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM Journal on Optimization, 30 (2020), pp. 1664–1692.