A TWO STEPSIZE SQP METHOD FOR NONLINEAR EQUALITY CONSTRAINED STOCHASTIC OPTIMIZATION

MICHAEL J. O'NEILL

Abstract. We develop a Sequential Quadratic Optimization (SQP) algorithm for minimizing a stochastic objective function subject to deterministic equality constraints. The method utilizes two different stepsizes, one which exclusively scales the component of the step corrupted by the variance of the stochastic gradient estimates and a second which scales the entire step. We prove that this stepsize splitting scheme has a worst-case complexity result which improves over the best known result for this class of problems. In terms of approximately satisfying the constraint violation, this complexity result matches that of deterministic SQP methods, up to constant factors, while matching the known optimal rate for stochastic SQP methods to approximately minimize the norm of the gradient of the Lagrangian. We also propose and analyze multiple variants of our algorithm. One of these variants is based upon popular adaptive gradient methods for unconstrained stochastic optimization while another incorporates a safeguarded line search along the constraint violation. Preliminary numerical experiments show competitive performance against a state of the art stochastic SQP method. In addition, in these experiments, we observe an improved rate of convergence in terms of the constraint violation, as predicted by the theoretical results.

1. Introduction.

We propose a new algorithm for solving equality constrained optimization problems in which the objective function is the expectation of a stochastic funciton. Formally, we consider the optimization problem

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

where $f:\mathbb{R}^n\to\mathbb{R}$, $c:\mathbb{R}^n\to\mathbb{R}^m$, ω is a random variable with associated probability space $(\Omega,\mathcal{F},P)$, $F:\mathbb{R}^n\times\Omega\to\mathbb{R}$, and $\mathbb{E}[\cdot]$ denotes the expectation taken with respect to P. Problems of this form arise in numerous applications, including optimal control [7], PDE-constrained optimization [16, 22], and resource allocation [6] as well as modern machine learning applications, such as physics informed neural networks [11, 21], constraining the output labels of deep neural networks [20] and neural network compression via constraints [10].

The method we design is based on Sequential Quadratic Optimization (SQP) methods, a popular class of algorithms that has seen significant interest in recent years for solving stochastic equality constrained optimization problems, beginning with the influential work of [4]. Numerous extensions of this work have been proposed, such as stochastic SQP methods for problems with rank-deficient Jacobians [3], algorithms for problems with nonlinear inequality constraints [13], worst-case complexity analysis for stochastic SQP methods [12], algorithms which incorporate variance reduction [5] or adaptive sampling [2], as well as stochastic SQP methods which utilize an exact augmented Lagrangian as a merit function [18, 19]. At each iteration, these algorithms generate a search direction by solving a quadratic optimization problem defined in terms of a stochastic gradient estimate subject to a linearization of the constraints and then produce a new iterate by moving along this search direction. For stochastic SQP methods, the chosen step length is generally scaled in such a way as to control the variance of the stochastic gradient estimates, in a manner similar to stepsizes for stochastic gradient methods in unconstrained optimization. Our algorithm takes a different approach and directly utilizes the orthogonal step decomposition of SQP methods[1]. It is well known in the stochastic SQP literature

[1]Computation of the orthogonal decomposition may be unnecessary in certain cases, see Remark 1 for details.

that the normal component of the step decomposition is independent of the current stochastic gradient estimate. Therefore, it is unnecessary to rescale this component by the stepsize which controls the variance in the stochastic gradient estimates in order to ensure convergence. Using this observation, we propose a method which employs two different stepsizes: one which controls the variance of the stochastic gradient estimates and scales only the tangential component and a second stepsize which scales the entire search direction.

We demonstrate the effectiveness of this stepsize splitting approach by developing a worst-case complexity result for our proposed algorithm. We consider the worst-case complexity in terms of finding a point x which satisfies,

$$\mathbb{E}[\|\nabla f(x) + \nabla c(x)y\|] \leq \epsilon_\ell, \qquad \mathbb{E}[\|c(x)\|_1] \leq \epsilon_c, \tag{1.2}$$

where $y \in \mathbb{R}^m$ is some Lagrange multiplier and ϵ_ℓ and ϵ_c are some small tolerances. Few complexity results exist for SQP methods in the literature. The only complexity result for a deterministic SQP method is given in [12], which proved a worst-case complexity result of $\mathcal{O}(\epsilon_\ell^{-2})$ and $\mathcal{O}(\epsilon_c^{-1})$ (this result holds deterministically, not just in expectation). This work also proved a result for the stochastic SQP method of [4], which was shown to have a worst-case complexity of $\mathcal{O}(\epsilon_\ell^{-4})$ and $\mathcal{O}(\epsilon_c^{-2})$ in an idealized setting and $\tilde{\mathcal{O}}(\epsilon_\ell^{-4})$ and $\tilde{\mathcal{O}}(\epsilon_c^{-2})$ otherwise, where $\tilde{\mathcal{O}}$ ignores logarthmic factors. In terms of ϵ_ℓ, this result is optimal, due to information theoretic lower bounds for stochastic gradient methods [1]. However, with respect to the constraint violation, it turns out that this result can be improved. We show that the worst-case complexity of the two stepsize stochastic SQP method proposed in this work has a worst-case complexity of $\mathcal{O}(\epsilon_\ell^{-4})$ and $\mathcal{O}(\epsilon_c^{-1})$. That is, in terms of convergence in the constraint violation, this result matches that of a **deterministic** SQP method, modulo the expectation and constant factors. Furthermore, we avoid unnecessary assumptions which were required to derive a complexity result in [12] by not estimating a merit parameter during the course of the algorithm. Previously this parameter was estimated using stochastic gradient information, which may be highly inaccurate on any given iteration and thus required additional assumptions in order to ensure convergence. In addition to these results, a number of other works have also proposed methods with known worst-case complexity results for solving (1.1), including augmented Lagrangian [15, 23] and stochastic SQP methods [18]. A summary of these worst-case complexity results is given in Table 1.1.

Unfortunately, the complexity result we prove for our initial algorithm requires certain choices of the stepsizes based on potentially difficult to estimate parameters of the problem (such as Lipschitz constants and a reasonable setting of the merit parameter). To remedy this, we propose a variant of our method which incorporates stepsizes inspired by adaptive gradient methods for unconstrained stochastic optimization [14, 17, 25]. Specifically, we build upon the methodolgy commonly known as Adagrad-Norm, which estimates a stepsize using the prior stochastic gradient estimates. We show that we can generate both of the stepsizes used by our algorithm under this framework and derive a worst-case complexity result for this variant of our method of the order $\tilde{\mathcal{O}}(\epsilon_\ell^{-4})$ and $\tilde{\mathcal{O}}(\epsilon_c^{-1})$, without requiring any knowledge of problem specific constants. In addition, both versions of our algorithm guarantee convergence when the stepsizes to be relaxed to lie in a certain set, from which the actual stepsize can be chosen, as was originally proposed in [4]. In order to choose a stepsize from this set, we propose a safeguarded linesearch in terms of the constraint violation and show how this can be implemented when the safeguarding is done in terms of

Algorithm	Conditions	Stationarity	Feasibility
SPD [15]	N/A	$\mathcal{O}\left(\epsilon_\ell^{-6}\right)$	$\mathcal{O}\left(\epsilon_c^{-6}\right)$
SPD [15]	x_0 feasible	$\mathcal{O}\left(\epsilon_\ell^{-5}\right)$	$\mathcal{O}\left(\epsilon_c^{-5}\right)$
MLALM [23]	N/A	$\mathcal{O}\left(\epsilon_\ell^{-5}\right)$	$\mathcal{O}\left(\epsilon_c^{-5}\right)$
MLALM [23]	x_0 near feasible	$\mathcal{O}\left(\epsilon_\ell^{-4}\right)$	$\mathcal{O}\left(\epsilon_c^{-4}\right)$
SSQP-AL [18]	N/A	$\mathcal{O}\left(\epsilon_\ell^{-4}\right)$	$\mathcal{O}\left(\epsilon_c^{-4}\right)$
SSQP [3]	$\tau_{\min}$ known	$\mathcal{O}\left(\epsilon_\ell^{-4}\right)$	$\mathcal{O}\left(\epsilon_c^{-2}\right)$
SSQP [3]	$\tau_{\min}$ unknown	$\tilde{\mathcal{O}}\left(\epsilon_\ell^{-4}\right)$	$\tilde{\mathcal{O}}\left(\epsilon_c^{-2}\right)$
SSQP-AS [2]	N/A	$\mathcal{O}\left(\epsilon_\ell^{-4}\right)$	$\mathcal{O}\left(\epsilon_c^{-2}\right)$
Algorithm 2.1	non-adaptive	$\mathcal{O}\left(\epsilon_\ell^{-4}\right)$	$\mathcal{O}\left(\epsilon_c^{-1}\right)$
Algorithm 2.1	adaptive	$\tilde{\mathcal{O}}\left(\epsilon_\ell^{-4}\right)$	$\tilde{\mathcal{O}}\left(\epsilon_c^{-1}\right)$

Table 1.1: Sample complexity of algorithms for solving (1.1). Convergence of each algorithm is proven underneath slightly different conditions. All methods except MLALM assume that the Jacobian has full rank at each iteration, while MLALM assumes a certain constraint qualification as well as mean-squared smoothness of the stochastic gradients. SSQP and SSQP-AS also make additional assumptions on the behavior of the merit parameter.

the adaptive stepsize rule based on Adagrad-Norm. Finally, we provide preliminary numerical experiments for our algorithm and show that it compares favorably with a state of the art stochstic SQP method. These numerical experiments also demonstrate faster convergence in constraint violation when compared with previously proposed stochastic SQP methods, providing confirmation of our theoretical results.

The rest of this work is organized as follows. In Section 2, we formally define and discuss our proposed algorithm and prove some basic properties. We provide a worst-case complexity analysis in Section 3 for two variants of our algorithm. A safeguarded linesearch procedure is developed in Section 4 and numerical experiments are presented in Section 5. We provide concluding remarks in Section 6.

1.1. Notation. We adopt the notation that $\|\cdot\|$ denotes the ℓ_2-norm for vectors and the vector-induced ℓ_2-norm for matrices. The set of nonnegative integers is denoted as $\mathbb{N} := \{0, 1, 2, \ldots, \}$ and we denote the positive real numbers by $\mathbb{R}_{>0}$.

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

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

1.2. Assumptions and Background. Throughout, we require the following assumptions on f and c:

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

Under this assumption both the gradient of f and the constraint violation are bounded in norm by constants. We denote these constants as $\|\nabla f(x)\| \leq \kappa_g$ and $\|c_k\|_1 \leq \kappa_c$.

Defining the Lagrangian $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ corresponding to (1.1) by $\ell(x, y) := f(x) + c(x)^\top y$, first-order primal-dual stationarity conditions for (1.1), which are necessary for optimality under Assumption 1, are given by

$$0 = \begin{bmatrix} \nabla_x \ell(x, y) \\ \nabla_y \ell(x, y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + \nabla c(x) y \\ c(x) \end{bmatrix}. \tag{1.3}$$

We note that the complexity measure (1.2) is simply an approximate version of these optimality conditions.

As stated above, our algorithm generates a search direction at iteration k by solving the following quadratic optimization problem:

$$\min_{p \in \mathbb{R}^n} f_k + g_k^T p + \frac{1}{2} p^T H_k p \;\text{ subject to }\; c_k + J_k p = 0, \tag{1.4}$$

where g_k is the current stochastic gradient estimate. It is well known that this is equivalent to solving the "Newton SQP system":

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} p_k \\ y_k \end{bmatrix} = - \begin{bmatrix} g_k \\ c_k \end{bmatrix}. \tag{1.5}$$

In order to ensure the solution of this sub-problem is unique, we require the following assumption on H_k.

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

In order to analyze our algorithm, we utilize the ℓ-1 merit function $\phi : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$:

$$\phi(x, \tau) = \tau f(x) + \|c(x)\|_1. \tag{1.6}$$

In the above equation, τ is the merit parameter which balances between the function value and constraint violation. For the analysis, we also use the following local model of the merit function $l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \to \mathbb{R}$ defined as

$$l(x, \tau, d) = \tau(f(x) + \nabla f(x)^T d) + \|c(x) + \nabla c(x)^T d\|_1. \tag{1.7}$$

In addition, we consider the reduction in the model for a direction $d \in \mathbb{R}^n$ with $c(x) + \nabla c(x)^T d = 0$ which is $\Delta l : \mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{R}^n \to \mathbb{R}$ defined as

$$\begin{aligned} \Delta l(x, \tau, d) &:= l(x, \tau, 0) - l(x, \tau, d) \\ &= -\tau \nabla f(x)^T d + \|c(x)\|_1. \end{aligned} \tag{1.8}$$

We wish to stress here that unlike previous work, we do not attempt to estimate a good value of τ. We choose to avoid this as previous work relied upon strong assumptions (such as uniformly bounded stochastic gradients [4] or sub-Gaussian stochastic gradients [12]) in order to prove their results. By choosing to relegate the merit function and parameter exclusively to the analysis, we are able to avoid overcomplicating the analysis and adding unnecessary assumptions.

2. Algorithm and Basic Properties. Recall that at each iteration, a search direction p_k is computed as the solution of (1.5). We assume that this step computation is performed in such a way that the orthogonal decomposition

$$p_k = u_k + v_k \text{ where } u_k \in \text{Null}(J_k) \text{ and } v_k \in \text{Range}(J_k^T), \tag{2.1}$$

is known[2]. One important consequence of this decomposition is that the normal component, v_k, does not depend on the current stochastic gradient estimate g_k. Unlike prior work, we do not directly use p_k as our search direction. Instead, we rescale the tangential component, u_k, in order to generate our search direction d_k as follows,

$$d_k = \beta_k u_k + v_k, \tag{2.2}$$

where $\beta_k \in \mathbb{R}_{>0}$. Then, we find the next iterate x_{k+1} by setting $x_{k+1} = x_k + \alpha_k d_k$ for some $\alpha_k \in \mathbb{R}_{>0}$. This procedure is formalized in Algorithm 2.1.

Algorithm 2.1 Generic Two Stepsize Stochastic SQP Algorithm

Require: $x_0 \in \mathbb{R}^n$;
for $k = 0, 1, \ldots$ **do**
  Compute stochastic gradient g_k.
  Compute (p_k, y_k) as the solution of (1.5).
  Choose $\beta_k \in \mathbb{R}_{>0}$.
  Set $d_k \leftarrow v_k + \beta_k u_k$, where $v_k \in \text{Range}(J_k^T)$ and $u_k \in \text{Null}(J_k)$ are the orthogonal decomposition of p_k.
  Choose $\alpha_k \in \mathbb{R}_{>0}$.
  Set $x_{k+1} \leftarrow x_k + \alpha_k d_k$.
end for

The choice of β_k is crucial to ensure convergence of our algorithm, as it controls the variance of the stochastic gradient estimates and plays a similar role as the stepsize in stochastic gradient methods. As such, it is natural to consider β_k to be quite small. Indeed, to ensure our convergence result, we set $\beta_k = O(1/\sqrt{K})$, where K is the total number of iterations we intend to perform. On the other hand, α_k does **not** need to control the error in the stochastic gradients and thus may be set independent of K. Thus, v_k, which is the component of d_k that drives the algorithm towards constraint satisfaction, is only scaled by a stepsize which is independent of K. This is the key insight that leads to our improved complexity.

Algorithm 2.1 is written generically, without specifying how to choose the stepsizes α_k and β_k. We consider two variants for choosing these stepsizes in Section 3 and analyze their behavior. First, in Section 3.1, we consider the case where β_k is defined by a pre-specified sequence and

$$\alpha_k \in [\nu, \nu + \theta\beta_k], \tag{2.3}$$

where $\nu \in \mathbb{R}_{>0}$ and $\theta \in \mathbb{R}_{>0}$. This case is essentially equivalent to the standard stochastic gradient regime with a pre-specified stepsize sequence (modulo the relaxtion of α_k into a range, which was originally suggested for stochastic SQP methods in [4]). For this method, we prove the complexity result foreshadowed in Section 1. However, this result only holds under certain conditions on ν which depend on the Lipschitz

[2]This is not necessary in certain circumstances, see Remark 1 for details.

constants of the gradient of f and Jacobian of c as well as a good estimate of the merit parameter τ. Unfortunately, it may not be reasonable to estimate these parameters a-priori.

To remedy this, in Section 3.2 we analyze a version of Algorithm 2.1 which utilizes adaptive stepsizes based on Adagrad-Norm, which is a popular approach for choosing stepsizes in the stochastic gradient literature [14, 17, 25]. In this case, some additional logarithmic factors appear in the final complexity result, but this approach does not require any knowledge of the Lipschitz constants or the merit parameter.

In addition, in both of the cases we analyze in Section 3, α_k may be chosen from a specific range. In Section 4, we describe a safeguarded line search procedure which can be used to determine α_k. We take adavantage of the assumption that the constraints are deterministic and design a line search which only relies on the constraint violation and does not include stochastic gradient information when computing an α_k. In addition, we provide a fully specified algorithm in Algorithm 4.1 that combines this linesearch procedure with an adaptive lower bound based on the Adagrad-Norm stepsizes developed in Section 3.2.

2.1. Properties of Algorithm 2.1. First, we restate a basic result from [4].

LEMMA 2.1. *([4, Lemma 2.9]) There exists $\kappa_v \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, the normal component v_k satisfies* $\max\{\|v_k\|, \|v_k\|^2\} \le \kappa_v \|c_k\|$.

During the analysis of our algorithm, it is often useful to consider the "true" step computation that would occur if the step was computed using the true gradient, $\nabla f(x_k)$, in place of the stochastic gradient estimate, g_k. Specifically, let $(p_k^{\text{true}}, y_k^{\text{true}})$ be the solution of the linear system:

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} p_k^{\text{true}} \\ y_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ c_k \end{bmatrix}. \tag{2.4}$$

In addition, we define

$$d_k^{\text{true}} = \beta_k u_k^{\text{true}} + v_k, \tag{2.5}$$

where $p_k^{\text{true}} = u_k^{\text{true}} + v_k$ with $u_k^{\text{true}} \in \text{Null}(J_k)$ (we recall here that v_k is independent of g_k and $\nabla f(x_k)$ and thus is the same v_k as in (2.1)).

LEMMA 2.2. *Let Assumptions 1 and 2 hold. Then,*

$$\|u_k^{\text{true}}\| \le \zeta^{-1}\|\nabla f(x_k)\| + \zeta^{-1}\kappa_H\kappa_v\|c_k\| \le \zeta^{-1}\kappa_g + \zeta^{-1}\kappa_H\kappa_v\kappa_c =: \kappa_u.$$

Proof. By the first equation of (2.4) and the definition of u_k^{true}, we have $(u_k^{\text{true}})^T H_k (u_k^{\text{true}} + v_k) = -\nabla f(x_k)^T u_k^{\text{true}}$. Then, by Assumption 2 and Lemma 2.1,

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

Dividing this inequality through by $\|u_k^{\text{true}}\|$ proves the first result. The final result follows by Assumption 1 and Lemma 2.1.

Now we state an important property about the merit parameter τ.

LEMMA 2.3. *Let Assumptions 1 and 2 hold and let $\sigma \in (0,1)$. Let $\beta_k \le \kappa_\beta$ hold for all k and define*

$$\tau_{\min} := \frac{1-\sigma}{\kappa_v(\kappa_\beta \kappa_H \kappa_u + \kappa_g)}. \tag{2.6}$$

Then,

$$\tau_{\min}\left(\nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}}\right) \le (1-\sigma)\|c_k\|_1. \tag{2.7}$$

Proof. By (2.4) and the definition of u_k^{true},

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

Thus, by Assumptions 1 and 2, Lemma 2.1 and Lemma 2.2,

$$\begin{aligned}\nabla f(x_k)^T d_k^{\text{true}} + \beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} &= -\beta_k v_k^T H_k u_k^{\text{true}} + \nabla f(x_k)^T v_k\\ &\le (\beta_k \kappa_H \|u_k^{\text{true}}\| + \|\nabla f(x_k)\|)\|v_k\|\\ &\le \kappa_v(\kappa_\beta \kappa_H \kappa_u + \kappa_g)\|c_k\|_1.\end{aligned}$$

Combining this with (2.6), proves (2.7).

REMARK 1. *Under the condition that H_k preserves the null space of J_k (i.e. for any $u \in Null(J_k), H_k u \in Null(J_k)$), we can sidestep the requirement to compute the orthogonal decomposition of p_k by simply rescaling the matrix H_k by β_k^{-1} and directly use the computed direction as d_k. This additional requirement is necessary when using rescaling in order to prove a result similar to Lemma 2.3, as otherwise the crossing term $v_k^T H_k u_k^{\text{true}}$ picks up a factor of β_k^{-1}. This, in turn, means that it is not possible to provide a bound on $\tau_{\min}$ that is independent of a* ***lower*** *bound on β_k, thus causing serious issues in the final complexity result. For the sake of generality, we don't consider this rescaling approach, though when H_k preserves the nullspace of J_k our results still hold, albeit with potentially different constant factors.*

A direct consequence of the previous lemma is

$$\Delta l(x_k, \tau_{\min}, d_k^{\text{true}}) \ge \tau_{\min}\beta_k (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \sigma\|c_k\|_1, \tag{2.8}$$

which will be used to prove the final convergence result. Given this inequality, it should be clear that with an upper bound on Δl, we would expect convergence in the constraint violation. To see the connection between the quantities in (2.8) and first order stationarity, we prove the following lemma, which shows that the quadratic term can be lower bounded in terms of the gradient of the Lagrangian at x_k for a specific Lagrange multipler.

LEMMA 2.4. *Let Assumptions 1 and 2 hold. Then,*

$$(u_k^{\text{true}})^T H_k u_k^{\text{true}} \ge \zeta \kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - (1+2\kappa_u)\zeta\kappa_v\|c_k\|_1.$$

Proof. By Assumption 2,

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

Then, by (2.4) and Lemmas 2.1 and 2.2,

$$
\begin{aligned}
&\|H_k u_k^{\text{true}} + H_k v_k - H_k v_k\|^2 \\
&\quad = \|H_k u_k^{\text{true}} + H_k v_k\|^2 - 2v_k^T H_k H_k (u_k^{\text{true}} + v_k) + \|H_k v_k\|^2 \\
&\quad \geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - 2v_k^T H_k H_k u_k^{\text{true}} - \|H_k v_k\|^2 \\
&\quad \geq \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 - (1 + 2\kappa_u)\kappa_H^2 \kappa_v \|c_k\|_1,
\end{aligned}
$$

which proves the result.

Thus, given these results, we can see that the convergence rate in terms of the gradient of the Lagrangian should be directly related to the choice of β_k while convergence in the constraint violation will be largely independent of this stepsize (provided it is chosen to sufficiently control the noise in g_k). This is in contrast to the results in [12], where the norm of the constraint violation is multiplied by β_k and is the root cause of the improvement in the complexity result for the constraint violation that we prove in the sequel.

We finish this subsection with the following generic descent lemma.

LEMMA 2.5. *Let Assumptions 1 and 2 hold. Then, with $\tau_{\min}$ defined as in (2.6),*

$$
\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\quad \leq -\alpha_k \Delta l(x_k, \tau_{\min}, d_k^{\text{true}}) + \frac{\alpha_k^2 \beta_k^2}{2}(\tau_{\min} L + \Gamma)\|u_k\|^2 \\
&\qquad + \frac{\alpha_k^2}{2}(\kappa_v(\tau_{\min} L + \Gamma) + 4)\|c_k\|_1 + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}).
\end{aligned}
\tag{2.9}
$$

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

$$
\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\leq \alpha_k \tau_{\min} \nabla f(x_k)^T d_k + \|c_k + \alpha_k J_k d_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&= \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} + |1 - \alpha_k|\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}),
\end{aligned}
$$

where the equality follows from $J_k d_k = -c_k$.

Then, when $\alpha_k \leq 1$,

$$
\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\leq \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} + (1 - \alpha_k)\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\
&= \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})
\end{aligned}
$$

On the other hand, when $\alpha_k > 1$,

$$
\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min})
$$

$$
\begin{aligned}
&\le \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} + (\alpha_k - 1)\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\
&= \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k \|c_k\|_1 + 2(\alpha_k - 1)\|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\
&\le \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k \|c_k\|_1 + 2\alpha_k^2 \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})
\end{aligned}
$$

where the second inequality follows by $\alpha_k > 1$.

Therefore, in either case, we have

$$
\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\le \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k \|c_k\|_1 + 2\alpha_k^2 \|c_k\|_1 + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)\|d_k\|^2 \\
&\quad + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}).
\end{aligned}
$$

Then, using the orthogonal decomposition $d_k = \beta_k u_k + v_k$, Lemma 2.1, and (1.8), we have

$$
\begin{aligned}
&\phi(x_k + \alpha_k d_k, \tau_{\min}) - \phi(x_k, \tau_{\min}) \\
&\le \alpha_k \tau_{\min} \nabla f(x_k)^T d_k^{\text{true}} - \alpha_k \|c_k\|_1 + 2\alpha_k^2 \|c_k\|_1 \\
&\quad + \frac{\alpha_k^2}{2}(\tau_{\min} L + \Gamma)(\beta_k^2 \|u_k\|^2 + \|v_k\|^2) + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}) \\
&\le -\alpha_k \Delta l(x_k, \tau_{\min}, d_k^{\text{true}}) + \frac{\alpha_k^2 \beta_k^2}{2}(\tau_{\min} L + \Gamma)\|u_k\|^2 \\
&\quad + \frac{\alpha_k^2}{2}(\kappa_v(\tau_{\min} L + \Gamma) + 4)\|c_k\|_1 + \alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}}),
\end{aligned}
$$

proving the result.

2.2. Stochastic Assumptions and Properties. In order to analyze the convergence of our algorithm, let $\mathcal{F}_k$ denote the natural filtration adapted to Algorithm 2.1 and let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$. Under these definitions, we have the following assumption on our stochastic gradient estimates, g_k.

ASSUMPTION 3. *There exists $M \in \mathbb{R}_{>0}$ such that, for all k, one finds*

$$
\mathbb{E}_k[g_k] = \nabla f(x_k) \quad \textit{and} \quad \mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2^2] \le M. \tag{2.10}
$$

This assumption is largely standard in the stochastic gradient literature. We note that relaxing the uniformly bounded variance assumption to an assumption which allows the variance to grow with the norm of the gradient of f (such as in [9]) is no more general under Assumption 1 since $\|\nabla f(x)\| \le \kappa_g$.

Under Assumption 3, we have the following properties.

LEMMA 2.6. *Let Assumptions 1, 2, and 3 hold. Then, $\mathbb{E}_k[u_k] = u_k^{\text{true}}$, $\mathbb{E}_k[y_k] = y_k^{\text{true}}$,*

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

In addition, when β_k is measurable to $\mathcal{F}_k$, $\mathbb{E}_k[d_k] = d_k^{\text{true}}$, and

$$\mathbb{E}_k[\|d_k - d_k^{\text{true}}\|] \le \beta_k \zeta^{-1}\sqrt{M}.$$

Proof. The first two claims follow directly by the statement of [4, Lemma 3.8] and the third follows directly by the proof of [4, Lemma 3.9] combined with Assumption 2. To prove the fourth result, let Z_k be an orthogonal basis for the null space of J_k (which, by Assumption 1 is a matrix in $\mathbb{R}^{n\times(n-m)}$) and let $u_k = Z_k w_k$ and $u_k^{\text{true}} = Z_k w_k^{\text{true}}$. Then, by (1.5), it follows that

$$Z_k w_k = -Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T(g_k + H_k v_k).$$

Similarly,

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

so that

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

and thus, by Assumptions 2 and 3,

$$\mathbb{E}_k[\|u_k - u_k^{\text{true}}\|^2] \le \mathbb{E}_k[\|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\|^2 \|\nabla f(x_k) - g_k\|^2] \le \zeta^{-2} M.$$

When β_k is measurable to $\mathcal{F}_k$, it follows that

$$\mathbb{E}_k[d_k] = \beta_k \mathbb{E}_k[u_k] + v_k = \beta_k u_k^{\text{true}} + v_k = d_k^{\text{true}}.$$

For the final result, we have that

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

Thus, by Assumptions 2 and 3, as well as Jensen's inequality,

$$\mathbb{E}_k[\|d_k - d_k^{\text{true}}\|] \le \beta_k \mathbb{E}_k[\|Z_k(Z_k^T H_k Z_k)^{-1} Z_k^T\| \|\nabla f(x_k) - g_k\|] \le \beta_k \zeta^{-1}\sqrt{M}.$$

3. Convergence Analysis. In this section, we derive our main convergence results for two variants of Algorithm 2.1, which differ on how α_k and β_k are chosen at each iteration.

3.1. Covergence with Pre-specified Stepsize Sequences. Throughout this subsection, we analyze Algorithm 2.1 when $\{\beta_k\}$ is a pre-specified sequence and α_k lies a pre-specified range, i.e.,

$$\{\beta_k\} \subset \mathbb{R}_{>0}, \qquad \alpha_k \in [\nu, \nu + \theta\beta_k], \quad \forall k, \tag{3.1}$$

for some $\nu \in \mathbb{R}_{>0}$ and $\theta \in \mathbb{R}_{>0}$.

Under this stepsize scheme, we prove a preliminary result about the final term that appears in Lemma 2.5.

Lemma 3.1. *Let Assumptions 1, 2, and 3 hold. Then,*

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

Proof. Let $\xi_k \in [0,1]$ be the random variable such that $\alpha_k = \nu + \xi_k \theta \beta_k$. Then, by Lemma 2.6 and the fact that ν and β_k are measurable to $\mathcal{F}_k$,

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

Now, we are ready to derive our first main result.

THEOREM 3.2. *Let Assumptions 1, 2, and 3 hold. Let $\sigma \in (0,1)$, let $\{\beta_k\} \subset \mathbb{R}_{>0}$ be a pre-specified sequence such that $\beta_k \leq \kappa_\beta$ holds for all k, let $\alpha_k \in [\nu, \nu + \theta\beta_k]$, for some $\theta \in \mathbb{R}_{>0}$, $\nu \in (0, \sigma/(2\kappa_v(\tau_{\min} L + \Gamma) + 4)]$, and let $\tau_{\min}$ be defined as in Lemma 2.3. Let*

$$\begin{aligned}\kappa_1 := &\frac{(\nu + \theta\kappa_\beta)^2 (\tau_{\min} L + \Gamma)(\zeta^{-1} \kappa_H \kappa_u^2 + \zeta^{-2} M)}{2} \\ &+ \theta(\theta\kappa_c(\kappa_v(\tau_{\min} L + \Gamma) + 4) + \tau_{\min} \kappa_g \zeta^{-1} \sqrt{M}).\end{aligned} \tag{3.2}$$

Then, for any $K \in \mathbb{N}$,

$$\begin{aligned}&\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2} \|c_k\|_1] \\ &\quad \leq \tau_{\min}(f(x_0) - f_{\text{low}}) + \|c_0\|_1 + \kappa_1 \sum_{k=0}^{K-1} \beta_k^2.\end{aligned} \tag{3.3}$$

Proof. Taking the conditional expectation on both sides of (2.9) and applying the results of Lemma 2.2, Lemma 2.6, and Lemma 3.1 (noting that β_k is measurable to $\mathcal{F}_k$),

$$\begin{aligned}&\mathbb{E}_k[\phi(x_k + \alpha_k d_k, \tau_{\min})] - \phi(x_k, \tau_{\min}) \\ &\quad \leq -\mathbb{E}_k[\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}})] + \mathbb{E}_k\left[\frac{\alpha_k^2 \beta_k^2}{2} (\tau_{\min} L + \Gamma) \|u_k\|^2\right] \\ &\qquad + \mathbb{E}_k\left[\frac{\alpha_k^2}{2} (\kappa_v(\tau_{\min} L + \Gamma) + 4) \|c_k\|_1\right] + \mathbb{E}_k[\alpha_k \tau_{\min} \nabla f(x_k)^T (d_k - d_k^{\text{true}})] \\ &\quad \leq -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) + (\nu^2 + \theta^2 \beta_k^2)(\kappa_v(\tau_{\min} L + \Gamma) + 4)\|c_k\|_1 \\ &\qquad + \frac{\alpha_k^2 \beta_k^2}{2} (\tau_{\min} L + \Gamma)(\zeta^{-1} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \zeta^{-2} M) + \theta \beta_k^2 \tau_{\min} \kappa_g \zeta^{-1} \sqrt{M} \\ &\quad \leq -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) + \frac{\alpha_k^2 \beta_k^2}{2} (\tau_{\min} L + \Gamma)(\zeta^{-1} \kappa_H \kappa_u^2 + \zeta^{-2} M) \\ &\qquad + \frac{\alpha_k \sigma}{2} \|c_k\|_1 + \beta_k^2 \theta(\theta\kappa_c(\kappa_v(\tau_{\min} L + \Gamma) + 4) + \tau_{\min} \kappa_g \zeta^{-1} \sqrt{M}) \\ &\quad = -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) + \frac{\alpha_k \sigma}{2} \|c_k\|_1 + \beta_k^2 \kappa_1,\end{aligned}$$

where the final inequality follows by $\nu \leq \alpha_k$.

Now, by (2.8),

$$\mathbb{E}_k[\phi(x_k + \alpha_k d_k, \tau_{\min})] - \phi(x_k, \tau_{\min})$$

$$
\begin{aligned}
&\le -\alpha_k \Delta l(x_k, \tau_{\min}, \nabla f(x_k), d_k^{\text{true}}) + \frac{\alpha_k \sigma}{2}\|c_k\|_1 + \beta_k^2 \kappa_1 \\
&\le -\alpha_k (\beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \sigma \|c_k\|_1) + \frac{\alpha_k \sigma}{2}\|c_k\|_1 + \beta_k^2 \kappa_1 \\
&= -\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} - \frac{\alpha_k \sigma}{2}\|c_k\|_1 + \beta_k^2 \kappa_1.
\end{aligned}
$$

Taking the total expectation of this inequality, rearranging and summing from $k = 0, \dots, K-1$,

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2}\|c_k\|_1] \\
&\qquad \le \phi(x_0, \tau_{\min}) - \mathbb{E}[\phi(x_K, \tau_{\min})] + \kappa_1 \sum_{k=0}^{K-1} \beta_k^2.
\end{aligned}
$$

Due to Assumption 1, we have,

$$
-\mathbb{E}[\phi(x_K, \tau_{\min})] = -\mathbb{E}[\tau_{\min} f(x_K) + \|c_K\|_1] \le -\tau_{\min} f_{\text{low}},
$$

so that

$$
\sum_{k=0}^{K-1} \mathbb{E}[\alpha_k \beta_k \tau_{\min} (u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2}\|c_k\|_1] \le \phi(x_0, \tau_{\min}) - \tau_{\min} f_{\text{low}} + \kappa_1 \sum_{k=0}^{K-1} \beta_k^2,
$$

which proves the result.

Corollary 3.3. *For any $K \in \mathbb{N}_{>0}$, let $\beta_k := \beta = \eta/\sqrt{K}$ for all $k \in [0, K-1]$ where $\eta \in \mathbb{R}_{>0}$, let κ_1 be defined in (3.2) and let*

$$
\kappa_2 := \tau_{\min}(f(x_0) - f_{\text{low}}) + \|c_0\|_1 + \eta^2 \kappa_1. \tag{3.4}
$$

Then, under the conditions of Theorem 3.2, we have

$$
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|c_k\|_1] \le \frac{2\kappa_2}{\nu \sigma K}, \tag{3.5}
$$

and

$$
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2] \le \frac{\kappa_H^2 \kappa_2}{\tau_{\min} \zeta \nu \eta \sqrt{K}} + \frac{2\zeta(1+2\kappa_u)\kappa_v \kappa_H^2 \kappa_2}{\nu \sigma K}. \tag{3.6}
$$

Finally, with probability at least $1 - \delta$,

$$
\begin{aligned}
&\min_{k \in [0, K-1]} \tau_{\min} \zeta \kappa_H^{-2} \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma \sqrt{K}}{2\eta}\|c_k\|_1 \\
&\qquad \le \frac{\kappa_2}{\nu \eta \delta \sqrt{K}} + \frac{2(1+2\kappa_u)\zeta \tau_{\min} \kappa_v \kappa_2}{\sigma \delta K}.
\end{aligned} \tag{3.7}
$$

Proof. By Theorem 3.2, the definition of β_k, and $\nu \le \alpha_k$, it follows that

$$
\sum_{k=0}^{K-1} \mathbb{E}[\|c_k\|_1] \le \frac{2(\tau_{\min}(f(x_0) - f_{\text{low}}) + \|c_0\|_1 + \kappa_1 \sum_{k=0}^{K-1} \beta_k^2)}{\nu \sigma} \le \frac{2\kappa_2}{\nu \sigma}. \tag{3.8}
$$

Dividing both sides of this inequality by K yields the first result.

Now, by Theorem 3.2 and Lemma 2.4 as well as $\alpha_k \le \nu$, we have

$$\sum_{k=0}^{K-1} \mathbb{E}[\nu\beta_k\tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\nu\sigma}{2}\|c_k\|_1 - \nu\tau_{\min}\beta_k(1+2\kappa_u)\zeta\kappa_v\|c_k\|_1]$$
$$\le \sum_{k=0}^{K-1} \mathbb{E}[\alpha_k\beta_k(u_k^{\text{true}})^T H_k(u_k^{\text{true}}) + \frac{\alpha_k\sigma}{2}\|c_k\|_1]$$
$$\le \tau_{\min}(f(x_0) - f_{\text{low}}) + \|c_0\|_1 + \kappa_1 \sum_{k=0}^{K-1}\beta_k^2. \tag{3.9}$$

Rearranging this inequality and using $\beta_k = \eta/\sqrt{K}$,

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2] \le \frac{\kappa_H^2\kappa_2\sqrt{K}}{\nu\eta\tau_{\min}\zeta} + (1+2\kappa_u)\kappa_v\kappa_H^2 \sum_{k=0}^{K-1}\mathbb{E}[\|c_k\|_1].$$

Dividing through by K and applying (3.5), it follows that

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

which proves the second result.

To prove the final result, by (3.9),

$$\sum_{k=0}^{K-1} \mathbb{E}[\nu\beta_k\tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\nu\sigma}{2}\|c_k\|_1]$$
$$\le \tau_{\min}(f(x_0) - f_{\text{low}}) + \|c_0\|_1 + \kappa_1\sum_{k=0}^{K-1}\beta_k^2$$
$$+ \sum_{k=0}^{K-1}\mathbb{E}[\nu\beta_k\tau_{\min}(1+2\kappa_u)\zeta\kappa_v\|c_k\|_1].$$

Applying the definition of β, multiplying through by $\frac{1}{\nu\beta K}$, and using (3.5),

$$\frac{1}{K}\mathbb{E}[\tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma\sqrt{K}}{2\eta}\|c_k\|_1]$$
$$\le \frac{\kappa_2}{\nu\eta\sqrt{K}} + \frac{2(1+2\kappa_u)\zeta\tau_{\min}\kappa_v\kappa_2}{\sigma K}$$

and thus

$$\min_{k\in[0,K-1]} \mathbb{E}[\tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma}{2\beta}\|c_k\|_1]$$
$$\le \frac{\kappa_2}{\nu\sqrt{K}} + \frac{2(1+2\kappa_u)\zeta\tau_{\min}\kappa_v\kappa_2}{\nu\sigma K}.$$

Applying Markov's inequality, it follows that with probability at least $1-\delta$ that

$$\min_{k\in[0,K-1]} \tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k)+J_k^T y_k^{\text{true}}\|^2+\frac{\sigma\sqrt{K}}{2\eta}\|c_k\|_1 \leq \frac{\kappa_2}{\nu\delta\sqrt{K}}+\frac{2(1+2\kappa_u)\zeta\tau_{\min}\kappa_v\kappa_2}{\sigma\delta K},$$

which proves the final result.

From the result of Corollary 3.3, we can easily derive our worst-case complexity results, as promised in Section 1. It should be clear that in terms of the constraint violation, by (3.5), the maximum number of iterations until $\mathbb{E}[\|c_k\|_1]$ falls below ϵ_c is at most $\mathcal{O}(\epsilon_c^{-1})$. Similarly, by Jensen's inequality and (3.6), the maximum number of iterations until $\mathbb{E}[\|\nabla f(x_k)+J_k^T y_k^{\text{true}}\|]\leq\epsilon_\ell$ is $\mathcal{O}(\epsilon_\ell^{-4})$. Finally, if one is interested in a combined result, we obtain the same $\mathcal{O}(K^{-1/2})$ convergence rate as [12], however, our convergence is in terms of a much stronger measure with respect to the constraint violation $\|c_k\|_1$, which is scaled by an additional factor of $\sqrt{K}$. Thus, we expect much faster convergence with respect to the constraint violation than the algorithm in [12] without harming the convergence rate in terms of the gradient of the Lagrangian.

3.2. Convergence with Adaptive Stepsizes. Now, we analyze the case where β_k and α_k are set adaptively, in a manner inspired by Adagrad-Norm [25]. Specifically, at each iteration k, let

$$b_k^2=b_{k-1}^2+\|u_k\|^2,\quad q_k^2=q_{k-1}^2+\|c_k\|_1, \tag{3.10}$$

and

$$\beta_k=\frac{\eta}{b_k},\qquad \alpha_k\in\left[\frac{\nu}{q_k},\frac{\nu}{q_k}+\min\left\{\frac{\theta}{b_k},\frac{\theta}{q_k}\right\}\right], \tag{3.11}$$

for some constants $\eta>0$ and $\nu>0$. We note here that the additional term at the upper end of the range for α_k is due to our adaptive setting of β_k using b_k, which is sufficient to control the stochasticity in g_k, but may be insufficient to control second order terms involving the constraint violation. We remedy this situation via the inclusion of the θ/q_k term. In addition, we remark that q_k can be set in many different ways, such as using $\|v_k\|^2$ or $\|c_k\|_2$ in place of $\|c_k\|_1$. These other strategies may lead to longer stepsizes, which could have important practical implications, however, we choose to use $\|c_k\|_1$ as it obtains the best constant factors in the convergence analysis among the relevant choices.

Throughout this section, since β_k is dependent on g_k, we redefine d_k^{true} as

$$d_k^{\text{true}}:=v_k+\beta_{k-1}u_k^{\text{true}}, \tag{3.12}$$

so that it remains measurable to $\mathcal{F}_k$. We note that under this re-definition, the results of Lemmas 2.3 and 2.5 still hold.

Our subsequent analysis relies on the following lemma, which we give without proof as it is a well-known result in the adaptive gradient literature (see for example, [24, Lemma 10]).

LEMMA 3.4. *Let $\{a_i\}_{i=0}^\infty$ be a series of non-negative real numbers with $a_0\in\mathbb{R}_{>0}$. Then,*

$$\sum_{k=1}^{T}\frac{a_k}{\sum_{i=0}^k a_i}\leq\log\left(\sum_{k=0}^T a_k\right)-\log(a_0) \tag{3.13}$$

In order to prove convergence of our algorithm, the key issue posed by the adaptive stepsizes is the final term in (2.9), which requires a more detailed analysis than in Lemma 3.1 as β_k is no longer measurable to $\mathcal{F}_k$ and d_k^{true} has been redefined in (3.12). We give a bound on this term in the following lemma.

LEMMA 3.5. *Let Assumptions 1, 2, and 3 hold and let*

$$\kappa_3 := \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M) \tag{3.14}$$

and

$$\kappa_4 := \max\left\{\zeta^{-1}\kappa_H^2, \beta_{-1}(1+2\kappa_u)\kappa_H^2\kappa_v\tau_{\min}/\sigma\right\} \tag{3.15}$$

Then,

$$\begin{aligned}\mathbb{E}_k\left[\alpha_k \nabla f(x_k)^T(d_k - d_k^{\text{true}})\right] \le \mathbb{E}_k\Bigg[&\frac{\alpha_k\beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k\sigma}{2\tau_{\min}}\|c_k\|_1 \\ &+ \left(\frac{3\eta^2\kappa_3\kappa_4(\nu+\theta)^2}{2q_{-1}b_{-1}} + \frac{3\kappa_4\theta^2(\eta^2+\beta_{-1}^2\kappa_3)}{2\eta\nu} + \frac{3\zeta^{-1}M\kappa_4\theta^2\beta_{-1}}{2q_{-1}b_{-1}^2}\right)\frac{\|u_k^2\|}{b_k^2}\Bigg].\end{aligned} \tag{3.16}$$

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

$$\mathbb{E}_k\left[\alpha_k \nabla f(x_k)^T(d_k - d_k^{\text{true}})\right] = \mathbb{E}_k\left[\alpha_k \nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\right].$$

Let $\xi_k \in [0,1]$ be the random variable such that $\alpha_k = \frac{\nu}{q_k} + \xi_k \min\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\}$. Then, by Lemma 2.6 and the fact that β_{k-1}, ν, and q_k are measurable to $\mathcal{F}_k$,

$$\begin{aligned}
&\mathbb{E}_k\left[\alpha_k \nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\right] \\
&= \mathbb{E}_k\left[\left(\frac{\nu}{q_k} + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\right)\nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\right] \\
&= \mathbb{E}_k\Big[\frac{\nu}{q_k}(\beta_k - \beta_{k-1})\nabla f(x_k)^T u_k \\
&\quad + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\Big] \\
&= \mathbb{E}_k\Big[\frac{\nu}{q_k}(\beta_k - \beta_{k-1})(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T u_k \\
&\quad + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\Big] \\
&= \mathbb{E}_k\Big[\frac{\nu}{q_k}(\beta_k - \beta_{k-1})(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T u_k \\
&\quad + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\beta_k(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T(u_k - u_k^{\text{true}}) \\
&\quad + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}(\beta_k - \beta_{k-1})(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T u_k^{\text{true}}\Big] \\
&= \mathbb{E}_k\Big[\left(\frac{\nu}{q_k} + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\right)(\beta_k - \beta_{k-1})(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T u_k \\
&\quad + \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\beta_k(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T(u_k - u_k^{\text{true}})
\end{aligned}$$

$$
\begin{aligned}
&+ \xi_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}(\beta_k - \beta_{k-1})(\nabla f(x_k) + J_k^T y_k^{\text{true}})^T (u_k^{\text{true}} - u_k)\Big] \\
\le\ & \mathbb{E}_k\Big[\frac{\nu+\theta}{q_k}|\beta_k - \beta_{k-1}|\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k\| \\
&+ \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\beta_k\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k - u_k^{\text{true}}\| \\
&+ \frac{\theta}{q_k}|\beta_k - \beta_{k-1}|\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k^{\text{true}} - u_k\|\Big],
\end{aligned} \tag{3.17}
$$

where the third equality follows by $u_k, u_k^{\text{true}} \in \text{Null}(J_k)$ and the inequality by the Cauchy-Schwarz inequality and $\xi_k \le 1$.

Now, we focus on the first term in (3.17),

$$
|\beta_k - \beta_{k-1}| = \frac{\eta}{b_{k-1}} - \frac{\eta}{b_k} = \frac{\eta\|u_k\|^2}{b_{k-1}b_k(b_k + b_{k-1})} \le \frac{\eta\|u_k\|}{b_{k-1}b_k}, \tag{3.18}
$$

where the inequality follows by $\|u_k\| \le b_k$. Therefore, applying Young's inequality, we have, for any $\lambda_1 > 0$ measurable to $\mathcal{F}_k$,

$$
\begin{aligned}
&\mathbb{E}_k\left[\frac{\nu+\theta}{q_k}|\beta_{k-1} - \beta_k|\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k\|\right] \\
&\le \eta\mathbb{E}_k\left[\frac{(\nu+\theta)\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k\|^2}{q_k b_{k-1} b_k}\right] \\
&\le \frac{\eta\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2}{2b_{k-1}q_k\lambda_1}\mathbb{E}_k[\|u_k\|^2] + \mathbb{E}_k\left[\frac{\eta\,(\nu+\theta)^2\,\lambda_1\|u_k\|^2}{2q_k b_{k-1} b_k^2}\right] \\
&\le \frac{\eta\kappa_3\beta_{k-1}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2}{2q_k\lambda_1} + \mathbb{E}_k\left[\frac{\eta(\nu+\theta)^2\lambda_1\|u_k\|^2}{2q_k b_{k-1} b_k^2}\right]
\end{aligned} \tag{3.19}
$$

where the final inequality follows by Assumption 2 as well as the results of Lemma 2.2 and Lemma 2.6.

Now, for the second term in (3.17), by Young's inequality, for any $\lambda_2 > 0$ measurable to $\mathcal{F}_k$,

$$
\begin{aligned}
&\mathbb{E}_k\left[\beta_k \min\left\{\frac{\theta}{b_k}, \frac{\theta}{q_k}\right\}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k - u_k^{\text{true}}\|\right] \\
&\le \frac{1}{2q_k b_k \lambda_2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \mathbb{E}_k\left[\frac{\lambda_2\theta^2\beta_k^2}{2}\|u_k - u_k^{\text{true}}\|^2\right].
\end{aligned} \tag{3.20}
$$

Working with the last term in this inequality, since q_k and β_{k-1} are measurable to $\mathcal{F}_k$, by Lemma 2.6,

$$
\begin{aligned}
&\mathbb{E}_k\left[\frac{\lambda_2\theta^2\beta_k^2}{2}\|u_k - u_k^{\text{true}}\|^2\right] \\
&= \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_k^2(\|u_k\|^2 + \|u_k^{\text{true}}\|^2 - 2u_k^T u_k^{\text{true}})\right] \\
&= \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_k^2(\|u_k\|^2 + \|u_k^{\text{true}}\|^2 - 2u_k^T u_k^{\text{true}}) + \beta_{k-1}^2(2u_k^T u_k^{\text{true}} - 2\|u_k^{\text{true}}\|^2)\right] \\
&\le \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_k^2\|u_k\|^2 + 2|\beta_k^2 - \beta_{k-1}^2|\|u_k\|\|u_k^{\text{true}}\| - \beta_{k-1}^2\|u_k^{\text{true}}\|^2\right]
\end{aligned}
$$

$$= \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_k^2\|u_k\|^2 + 2\eta^2\frac{\|u_k\|^2}{b_{k-1}^2 b_k^2}\|u_k\|\|u_k^{\text{true}}\| - \beta_{k-1}^2\|u_k^{\text{true}}\|^2\right]$$

$$\leq \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_k^2\|u_k\|^2 + 2\eta^2\frac{\|u_k\|^2}{b_{k-1}^2 b_k}\|u_k^{\text{true}}\| - \beta_{k-1}^2\|u_k^{\text{true}}\|^2\right] \tag{3.21}$$

where the first inequality follows by $\beta_k \leq \beta_{k-1}$ and the second inequality follows by $\|u_k\| \leq b_k$. Dealing with the second term in this inequality, again applying Young's inequality and using $\beta_{k-1} = \eta/b_{k-1}$, by Lemmas 2.2 and 2.6 as well as Assumption 2,

$$\begin{aligned}
&\frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\frac{2\beta_{k-1}^2\|u_k\|^2}{b_k}\|u_k^{\text{true}}\|\right] \\
&\leq \frac{\lambda_2\theta^2\eta^2}{2}\mathbb{E}_k\left[\frac{\beta_{k-1}^2\|u_k\|^2\|u_k^{\text{true}}\|^2}{\kappa_3} + \frac{\beta_{k-1}^2\kappa_3\|u_k\|^2}{b_k^2}\right] \\
&\leq \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\frac{\beta_{k-1}^2\zeta^{-1}(\kappa_H\kappa_u^2 + \zeta^{-1}M)\|u_k^{\text{true}}\|^2}{\kappa_3} + \frac{\beta_{k-1}^2\kappa_3\|u_k\|^2}{b_k^2}\right] \\
&= \frac{\lambda_2\theta^2}{2}\mathbb{E}_k\left[\beta_{k-1}^2\|u_k^{\text{true}}\|^2 + \frac{\beta_{k-1}^2\kappa_3\|u_k\|^2}{b_k^2}\right],
\end{aligned}$$

so that the first term cancels with the last in (3.21).

Now, for the final term in (3.17), by (3.18), applying Young's inequality for some $\lambda_3 > 0$ that is measurable to $\mathcal{F}_k$, by Lemma 2.6,

$$\begin{aligned}
&\mathbb{E}_k\left[\frac{\theta}{q_k}|\beta_k - \beta_{k-1}|\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k^{\text{true}} - u_k\|\right] \\
&\leq \mathbb{E}_k\left[\frac{\theta}{q_k}\frac{\beta_{k-1}\|u_k\|}{b_{k-1}b_k}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|\|u_k^{\text{true}} - u_k\|\right] \\
&\leq \mathbb{E}_k\left[\frac{\beta_{k-1}}{2\lambda_3 q_k}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2\|u_k^{\text{true}} - u_k\|^2 + \frac{\theta^2\lambda_3\beta_{k-1}\|u_k\|^2}{2q_k b_{k-1}^2 b_k^2}\right] \\
&\leq \mathbb{E}_k\left[\frac{\zeta^{-1}M\beta_{k-1}}{2\lambda_3 q_k}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\theta^2\lambda_3\beta_{k-1}\|u_k\|^2}{2q_k b_{k-1}^2 b_k^2}\right]
\end{aligned}$$

Therefore, combining (3.17), (3.19), (3.20), and (3.21) we have

$$\begin{aligned}
&\mathbb{E}_k\left[\alpha_k\nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\right] \\
&\leq \mathbb{E}_k\left[\left(\frac{\eta\kappa_3\beta_{k-1}}{2q_k\lambda_1} + \frac{1}{2q_k b_k\lambda_2} + \frac{\zeta^{-1}M\beta_{k-1}}{2q_k\lambda_3}\right)\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2\right. \\
&\quad\left. + \left(\frac{\lambda_1\eta(\nu+\theta)^2}{2q_k b_{k-1}} + \frac{\lambda_2\theta^2(\eta^2 + \beta_{k-1}^2\kappa_3)}{2} + \frac{\lambda_3\theta^2\beta_{k-1}}{2q_k b_{k-1}^2}\right)\frac{\|u_k^2\|}{b_k^2}\right]
\end{aligned}$$

Applying Lemma 2.4,

$$\begin{aligned}
&\mathbb{E}_k\left[\alpha_k\nabla f(x_k)^T(\beta_k u_k - \beta_{k-1}u_k^{\text{true}})\right] \\
&\leq \mathbb{E}_k\left[\left(\frac{\eta\kappa_3\beta_{k-1}}{2q_k\lambda_1} + \frac{1}{2q_k b_k\lambda_2} + \frac{\zeta^{-1}M\beta_{k-1}}{2q_k\lambda_3}\right)(\zeta^{-1}\kappa_H^2(u_k^{\text{true}})^T H_k u_k^{\text{true}}\right. \\
&\left. + (1+2\kappa_u)\kappa_H^2\kappa_v\|c_k\|_1) + \left(\frac{\lambda_1\eta(\nu+\theta)^2}{2q_k b_{k-1}} + \frac{\lambda_2\theta^2(\eta^2 + \beta_{k-1}^2\kappa_3)}{2} + \frac{\lambda_3\theta^2\beta_{k-1}}{2q_k b_{k-1}^2}\right)\frac{\|u_k^2\|}{b_k^2}\right].
\end{aligned}$$

Choosing $\lambda_1 = \frac{3\eta\kappa_3\kappa_4}{\nu}$, $\lambda_2 = \frac{3\kappa_4}{\nu\eta}$, and $\lambda_3 = \frac{3\zeta^{-1}M\kappa_4}{\nu}$ and using $\nu/q_k \le \alpha_k$, $q_k \ge q_{-1}$, and $b_{k-1} \ge b_{-1}$ proves the result.

Now, we are prepared to present the first main result of this subsection.

THEOREM 3.6. *Let Assumptions 1, 2, and 3 hold. Let*

$$\kappa_5 := \frac{(\nu+\theta)^2(\kappa_v(\tau_{\min}L+\Gamma)+4)}{2}. \tag{3.22}$$

$$\begin{aligned}\kappa_6 := {} & \frac{\eta^2(\nu+\theta)^2(\tau_{\min}L+\Gamma)}{2q_{-1}^2} + \frac{3\eta^2\kappa_3\kappa_4(\nu+\theta)^2}{2q_{-1}b_{-1}} \\ & + \frac{3\kappa_4\theta^2(\eta^2+\beta_{-1}^2\kappa_3)}{2\eta\nu} + \frac{3\zeta^{-1}M\kappa_4\theta^2\beta_{-1}}{2q_{-1}b_{-1}^2}.\end{aligned} \tag{3.23}$$

Then,

$$\begin{aligned}& \mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\alpha_k\tau_{\min}\beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k\sigma}{2}\|c_k\|_1\right] \\ & \le \tau_{\min}(f_{-1}-f_{\min}) + \|c_{-1}\|_1 + \kappa_5\log(1+\kappa_c K/q_{-1}^2) \\ & \quad + \kappa_6(1+\log(\zeta^{-1}(\kappa_H\kappa_u^2+\zeta^{-1}M)K/b_{-1}^2)).\end{aligned} \tag{3.24}$$

In addition,

$$\begin{aligned}\mathbb{E}[q_{K-1}] \le {} & q_{-1} + \frac{2}{\nu\sigma}(\tau_{\min}(f_{-1}-f_{\min}) + \|c_{-1}\|_1 + \kappa_5\log(1+\kappa_c K/q_{-1}^2) \\ & + \kappa_6\log(1+\zeta^{-1}(\kappa_H\kappa_u^2+\zeta^{-1}M)K/b_{-1}^2).\end{aligned}$$

Proof. By Lemma 2.5, we have

$$\begin{aligned}& \mathbb{E}_k[\phi(x_k+\alpha_k d_k,\tau_{\min})] - \phi(x_k,\tau_{\min}) \\ & \quad \le -\mathbb{E}_k[\alpha_k\Delta l(x_k,\tau_{\min},d_k^{\text{true}})] + \mathbb{E}_k\left[\frac{\alpha_k^2\beta_k^2}{2}(\tau_{\min}L+\Gamma)\|u_k\|^2\right] \\ & \quad + \mathbb{E}_k\left[\frac{\alpha_k^2}{2}(\kappa_v(\tau_{\min}L+\Gamma)+4)\|c_k\|_1\right] + \mathbb{E}_k[\alpha_k\tau_{\min}\nabla f(x_k)^T(d_k-d_k^{\text{true}})].\end{aligned}$$

To prove the result, we need to bound the final three terms. Starting with the first of these, we have that

$$\mathbb{E}_k\left[\frac{\alpha_k^2\beta_k^2}{2}(\tau_{\min}L+\Gamma)\|u_k\|^2\right] \le \frac{\eta^2(\nu+\theta)^2(\tau_{\min}L+\Gamma)}{2q_{-1}^2}\mathbb{E}_k\left[\frac{\|u_k\|^2}{b_k^2}\right],$$

where the inequality follows due to the definition of α_k and $q_k \ge q_{-1}$. For the next term,

$$\frac{\alpha_k^2(\kappa_v(\tau_{\min}L+\Gamma)+4)}{2}\|c_k\|_1 \le \frac{(\nu+\theta)^2(\kappa_v(\tau_{\min}L+\Gamma)+4)}{2q_k^2}\|c_k\|_1 = \frac{\kappa_5\|c_k\|_1}{q_k^2}.$$

Now, applying the result of Lemma 3.5, we have

$$\mathbb{E}_k[\phi(x_k+\alpha_k d_k,\tau_{\min})] - \phi(x_k,\tau_{\min})$$

$$\leq -\mathbb{E}_k[\alpha_k \Delta l(x_k, \tau_{\min}, d_k^{\text{true}})] + \mathbb{E}_k\left[\frac{\alpha_k \tau_{\min} \beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}}\right]$$
$$+ \mathbb{E}_k\left[\frac{\alpha_k \sigma}{2}\|c_k\|_1\right] + \frac{\kappa_5 \|c_k\|_1}{q_k^2} + \kappa_6 \mathbb{E}_k\left[\frac{\|u_k\|^2}{b_k^2}\right].$$

Then, applying (2.8) (where we note that under the re-definition of d_k^{true} in (3.12), β_k is replaced by β_{k-1}), it follows that

$$\begin{aligned}
&\mathbb{E}_k[\phi(x_k + \alpha_k d_k, \tau_{\min})] - \phi(x_k, \tau_{\min}) \\
&\quad \leq -\mathbb{E}_k\left[\frac{\alpha_k \tau_{\min} \beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}}\right] - \mathbb{E}_k\left[\frac{\alpha_k \sigma}{2}\|c_k\|_1\right] \\
&\quad + \frac{\kappa_5 \|c_k\|_1}{q_k^2} + \kappa_6 \mathbb{E}_k\left[\frac{\|u_k\|^2}{b_k^2}\right].
\end{aligned}$$

Next, taking the total expectation of this inequality and summing for all $k = 0, \ldots, K-1$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\alpha_k \tau_{\min} \beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2}\|c_k\|_1\right] \\
&\leq \phi(x_{-1}, \tau_{\min}) - \mathbb{E}[\phi(x_K, \tau_{\min})] + \mathbb{E}\left[\kappa_5 \sum_{k=0}^{K-1} \frac{\|c_k\|_1}{q_k^2}\right] + \mathbb{E}\left[\kappa_6 \sum_{k=0}^{K-1} \frac{\|u_k\|^2}{b_k^2}\right].
\end{aligned}$$

By the definition of ϕ and Assumption 1, it follows that

$$\begin{aligned}
\phi(x_{-1}, \tau_{\min}) - \mathbb{E}[\phi(x_K, \tau_{\min})] &= \tau_{\min} f_{-1} + \|c_{-1}\|_1 - \mathbb{E}[\tau_{\min} f_K - \|c_K\|_1] \\
&\leq \tau_{\min}(f_{-1} - f_{\min}) + \|c_{-1}\|_1.
\end{aligned}$$

Now, applying Lemma 3.4 twice, by Assumption 1, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\alpha_k \tau_{\min} \beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2}\|c_k\|_1\right] \\
&\leq \tau_{\min}(f_{-1} - f_{\min}) + \|c_{-1}\|_1 + \kappa_5 \log(1 + \kappa_c K / q_{-1}^2) \\
&\quad + \kappa_6 \mathbb{E}\left[\log\left(\frac{b_{-1}^2 + \sum_{k=0}^{K-1} \|u_k\|^2}{b_{-1}^2}\right)\right].
\end{aligned}$$

Using Jensen's inequality, the tower rule, and the results of Lemma 2.2 and Lemma 2.6,

$$\mathbb{E}\left[\log\left(\frac{b_{-1}^2 + \sum_{k=0}^{K-1} \|u_k\|^2}{b_{-1}}\right)\right] \leq \log\left(1 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M) K / b_{-1}^2\right)$$

and thus,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\alpha_k \tau_{\min} \beta_{k-1}}{2}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\alpha_k \sigma}{2}\|c_k\|_1\right] \\
&\leq \tau_{\min}(f_{-1} - f_{\min}) + \|c_{-1}\|_1 + \kappa_5 \log(1 + \kappa_c K / q_{-1}^2)
\end{aligned}$$

$$+ \kappa_6 \log(1 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M)K/b_{-1}^2),$$

proving the first result.

To prove the second result, note that

$$q_{K-1} = \frac{q_{-1}^2 + \sum_{k=0}^{K-1} \|c_k\|_1}{q_K} \le q_{-1} + \sum_{k=0}^{K-1} \frac{\|c_k\|_1}{q_k} \le q_{-1} + \frac{1}{\nu} \sum_{k=0}^{K-1} \alpha_k \|c_k\|_1,$$

and therefore, by (3.24),

$$\mathbb{E}[q_{K-1}] \le q_{-1} + \frac{2}{\nu\sigma}(\tau_{\min}(f_{-1} - f_{\min}) + \|c_{-1}\|_1 + \kappa_5 \log(1 + \kappa_c K/q_{-1}^2) \\ + \kappa_6 \log(1 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M)K/b_{-1}^2).$$

Next, we derive the following corollary, from which our complexity results for this subsection will follow directly.

COROLLARY 3.7. *Let the assumptions of Theorem 3.6 hold. Let*

$$\kappa_7(K) := \tau_{\min}(f_{-1} - f_{\min}) + \|c_{-1}\|_1 + \kappa_5 \log(1 + \kappa_c K/q_{-1}) \\ + \kappa_6 \log(1 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M)K/b_{-1}) \tag{3.25}$$

and

$$\kappa_8(K) := \sqrt{b_{-1}^2 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M)K}. \tag{3.26}$$

Then, with probability at least $1 - \delta_1$*,*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \|c_k\|_1\right] \le \frac{2(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\sigma^2\delta_1 K}, \tag{3.27}$$

with probability at least $1 - \delta_2$*,*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2\right] \\ \le \frac{8\kappa_H^2 \kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\tau_{\min}\nu^2\eta\zeta\sigma\delta_2^2 K} \\ + \frac{4(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_H^2(1 + 2\kappa_u)\kappa_v\kappa_7(K)}{\nu^2\sigma^2\delta_2 K}, \tag{3.28}$$

and with probability at least $1 - \delta_3$*,*

$$\min_{k\in[0,K-1]} \tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma\kappa_8(K)}{\eta}\|c_k\|_1 \\ \le \frac{54\kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\eta\sigma\delta_3^3 K} \\ + \frac{18(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)\zeta\tau_{\min}(1 + 2\kappa_u)\kappa_v}{\nu^2\sigma^2\delta_3^2 K}. \tag{3.29}$$

Proof. By Theorem 3.6 and Markov's inequality, it follows that with probability at least $1 - \delta_1$,

$$\alpha_{K-1} \geq \frac{\nu}{q_K} \geq \frac{\nu^2 \sigma \delta_1}{\nu \sigma q_{-1} + 2\kappa_7(K)}. \tag{3.30}$$

Therefore, by (3.24), Assumption 2, and the fact that α_K is non-increasing, it follows that with probability at least $1 - \delta_1$,

$$\mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\nu^2 \sigma^2 \delta_1}{2(\nu \sigma q_{-1} + 2\kappa_7(K))} \|c_k\|_1\right] \leq \mathbb{E}\left[\sum_{k=0}^{K-1} \frac{\alpha_k \sigma}{2} \|c_k\|_1\right] \leq \kappa_7(K),$$

and thus

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} \|c_k\|_1\right] \leq \frac{2(\nu \sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2 \sigma^2 \delta_1 K},$$

which proves the first result.

Next, by the law of iterated expectation, Jensen's inequality, and the results of Lemmas 2.2 and 2.6,

$$\mathbb{E}[b_{K-1}] = \mathbb{E}\left[\sqrt{b_{-1}^2 + \sum_{k=0}^{K-1} \|u_k\|^2}\right] \leq \sqrt{b_{-1}^2 + \zeta^{-1}(\kappa_H \kappa_u^2 + \zeta^{-1} M)K} = \kappa_8(K). \tag{3.31}$$

Therefore, using (3.30) with $\delta_1 = \delta_2/2$ and Markov's inequality with (3.31), with probability at least $1 - \delta_2$, by Assumption 2, (3.24), and the union bound,

$$\mathbb{E}\left[\sum_{k=0}^{K-1} (u_k^{\text{true}})^T H_k u_k^{\text{true}}\right] \leq \frac{8\kappa_8(K)(\nu \sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\tau_{\min} \nu^2 \eta \sigma \delta_2^2}.$$

Next, applying Lemma 2.4,

$$\mathbb{E}\left[\sum_{k=0}^{K-1} \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2\right] \leq \frac{8\kappa_H^2 \kappa_8(K)(\nu \sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\tau_{\min} \nu^2 \eta \zeta \sigma \delta_2^2} + \kappa_H^2 (1 + 2\kappa_u)\kappa_v \mathbb{E}\left[\sum_{k=0}^{K-1} \|c_k\|_1\right].$$

Noting that this result holds under the same event as in (3.27) (with $\delta_1 = \delta_2/2$), it follows that with probability at least $1 - \delta_2$,

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2\right] \leq \frac{8\kappa_H^2 \kappa_8(K)(\nu \sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\tau_{\min} \nu^2 \eta \zeta \sigma \delta_2^2 K} + \frac{4(\nu \sigma q_{-1} + 2\kappa_7(K))\kappa_H^2 (1 + 2\kappa_u)\kappa_v \kappa_7(K)}{\nu^2 \sigma^2 \delta_2 K}.$$

Finally, using (3.24), (3.30), (3.31), Markov's inequality and the union bound, with probability at least $1 - \frac{2}{3}\delta_3$,

$$\mathbb{E}\left[\sum_{k=0}^{K-1} \tau_{\min}(u_k^{\text{true}})^T H_k u_k^{\text{true}} + \frac{\sigma\kappa_8(K)}{\eta}\|c_k\|_1\right] \leq \frac{18\kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\eta\sigma\delta_3^2}.$$

Thus, by Lemma 2.4

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma\kappa_8(K)}{\eta}\|c_k\|_1\right] \\
&\leq \frac{18\kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\eta\sigma\delta_3^2 K} + \zeta\tau_{\min}(1+2\kappa_u)\kappa_v \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|c_k\|_1\right] \\
&\leq \frac{18\kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\eta\sigma\delta_3^2 K} \\
&\quad + \frac{6(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)\zeta\tau_{\min}(1+2\kappa_u)\kappa_v}{\nu^2\sigma^2\delta_3 K}.
\end{aligned}$$

Therefore, applying Markov's inequality and the union bound, it follows that with probability at least $1 - \delta_3$,

$$\begin{aligned}
&\min_{k\in[0,K-1]} \tau_{\min}\zeta\kappa_H^{-2}\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|^2 + \frac{\sigma\kappa_8(K)}{\eta}\|c_k\|_1 \\
&\leq \frac{54\kappa_8(K)(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)}{\nu^2\eta\sigma\delta_3^3 K} \\
&\quad + \frac{18(\nu\sigma q_{-1} + 2\kappa_7(K))\kappa_7(K)\zeta\tau_{\min}(1+2\kappa_u)\kappa_v}{\nu^2\sigma^2\delta_3^2 K}.
\end{aligned}$$

By the definitions of $\kappa_7(K) = \mathcal{O}(\log(K))$ and $\kappa_8(K) = \mathcal{O}(\sqrt{K})$, it follows that the results of Corollary 3.7 match, up to log factors, those we derived in Section 3.1 for the pre-specified stepsize setting. Thus, in terms of the complexity measures (1.2), this variant of Algorithm 2.1 has a worst-case complexity of $\tilde{\mathcal{O}}(\epsilon_\ell^{-4})$ and $\tilde{\mathcal{O}}(\epsilon_c^{-1})$.

4. Safeguarded Line Search.

The convergence analysis in Section 3 specifies proper ranges for α_k in Algorithm 2.1 in order to ensure convergence, but does not provide any recommendations on how to choose α_k in this range. Commonly, in other stochastic SQP methods, the procedure used to set α_k incorporates the merit parameter τ_k, which is adaptively estimated at each iteration. However, the estimation of τ_k may be highly inaccurate and noisy due to only having stochastic access to the gradient of f. For this reason, we do not attempt rely on the stochastic gradient information in order to choose α_k and instead solely utilize the constraints.

Consider first the case where α_k satisfies $\alpha_k \in [\nu, \nu + \theta\beta_k]$ as it does in the analysis in Section 3.1. Then, we can find an α_k in this range through a safeguarded backtracking procedure. Starting from $\hat{\alpha}_k = \nu + \theta\beta_k$, we backtrack until

$$\|c(x_k + \hat{\alpha}_k d_k)\|_1 \leq (1 - \xi\hat{\alpha}_k)\|c_k\|_1, \tag{4.1}$$

holds for some $\xi \in (0,1)$ where, when (4.1) fails to hold for $\hat{\alpha}_k$, we set $\hat{\alpha}_k = \rho\hat{\alpha}_k$ for some $\rho \in (0,1)$. However, as we cannot guarantee termination, we safeguard this linesearch by ceasing the search procedure if $\hat{\alpha}_k$ ever falls below ν. When (4.1) holds for some $\hat{\alpha}_k \geq \nu$, we set $\alpha_k = \hat{\alpha}_k$. On the other hand, if (4.1) fails to hold prior to $\hat{\alpha}_k < \nu$, we instead set $\alpha_k = \nu$. Thus, this procedure is guaranteed to output an α_k in the specified range and therefore the convergence results of Section 3.1 hold. In addition, on any step where (4.1) is satisfied for $\hat{\alpha}_k \geq \nu$, we have confirmation of sufficient decrease in the constraint violation. Finally, we note that the number of backtracking steps at any iteration k is at most $\log(\nu/(\nu + \theta\beta_k))/\log(\rho)$ due to terminating the backtracking as soon as $\hat{\alpha}_k < \nu$.

Unfortunately, the convergence theory only holds for the previous procedure under certain conditions on ν. To relax these conditions, we once again turn to the one of the adaptive stepsize rules of Section 3.2. In particular, we consider the case where β_k is chosen as a pre-specified sequence and the lower bound for α_k is chosen in a manner similar to that of (3.11). As we use a slight modification of this stepsize, we give the full procedure (which is the algorithm used in the computational results of Section 5) in Algorithm 4.1.

The backtracking procedure in Algorithm 4.1 is very similar to the one described above, with a few minor differences. In particular, we set the lower bound adaptively, using the stepsize rule in Section 3.2. In addition, unlike in Section 3.2, we only update the lower bound when the backtracking procedure fails to satisfy the sufficient decrease condition prior to reducing $\hat{\alpha}_k$ below the lower bound. The logic for this is simple; if the lower bound was reached, then it is probably too large and should be reduced. On the other hand, when the sufficient decrease condition is satisfied at iteration k, we keep the lower bound as it was at the start of this iteration, since it is already sufficiently small to find a good steplength in terms of reducing the constraint violation.

Algorithm 4.1 Two Stepsize Stochastic SQP with Adaptive Backtracking

Require: $x_0 \in \mathbb{R}^n$, $\{\beta_k\} \subset \mathbb{R}_{>0}$, $\nu \in \mathbb{R}_{>0}$, $q_{-1} \in \mathbb{R}_{>0}$, $\theta \in \mathbb{R}_{>0}$, $\xi \in (0,1)$, $\rho \in (0,1)$;

for $k = 0, 1, \ldots$ **do**
    Compute stochastic gradient g_k.
    Compute (p_k, y_k) as the solution of (1.5).
    Set $d_k \leftarrow v_k + \beta_k u_k$, where $v_k \in \text{Range}(J_k^T)$ and $u_k \in \text{Null}(J_k)$ are the orthogonal decomposition of p_k.
    Set $\hat{q}_k^2 \leftarrow q_{k-1}^2 + \|c_k\|_1$ and $\hat{\alpha}_k \leftarrow \frac{\nu}{\hat{q}_k} + \theta\beta_k$.
    while $\|c(x_k + \hat{\alpha}_k d_k)\|_1 > (1 - \xi\hat{\alpha}_k)\|c_k\|_1$ and $\hat{\alpha}_k > \frac{\nu}{\hat{q}_k}$ **do**
        Set $\hat{\alpha}_k \leftarrow \rho\hat{\alpha}_k$.
    end while
    if $\hat{\alpha}_k \geq \frac{\nu}{\hat{q}_k}$ **then**
        Set $\alpha_k \leftarrow \hat{\alpha}_k$ and $q_k = q_{k-1}$.
    else
        Set $\alpha_k \leftarrow \frac{\nu}{\hat{q}_k}$ and $q_k = \hat{q}_k$.
    end if
    Set $x_{k+1} \leftarrow x_k + \alpha_k d_k$.
end for

While this is a relatively simple variant of Algorithm 2.1, the analysis in Section 3 does not directly translate. We provide the following lemma which provides a starting point for the analysis that can then easily be combined with the techniques in Section 3 to obtain a worst-case complexity result.

LEMMA 4.1. *Let Assumptions 1, 2, and 3 hold and let x_k be generated by Algorithm 4.1. Let $\beta_k = \eta/\sqrt{K}$ hold for all k. Let*

$$\kappa_9 := \xi^{-1}(\|c_0\|_1 - (2+\Gamma\kappa_v/2)\nu^2\log(q_{-1}^2) + \nu^2\eta^2\Gamma\zeta^{-1}(\kappa_H\kappa_u^2 + \zeta^{-1}M)/(2q_{-1}^2)) \tag{4.2}$$

and

$$\kappa_{10} := 2(q_{-1} + \kappa_9/\nu) + 8(4+\Gamma\kappa_v)\xi^{-1}\nu\log(e + (4+\Gamma\kappa_v)\xi^{-1}\nu). \tag{4.3}$$

Then, $\mathbb{E}[q_{K-1}] \leq \kappa_{10}$ *and*

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\alpha_k\|c_k\|_1\right] \leq \kappa_9 + (4+\Gamma\kappa_v)\xi^{-1}\nu\log(\kappa_{10}) \tag{4.4}$$

Proof. Let $\mathcal{K}_\alpha$ denote the index set of iterations k such that $\alpha_k = \frac{\nu}{\hat{q}_k}$. Then, for any $k \in \mathcal{K}_\alpha$, by Γ-Lipschitz continuity of the Jacobian of c,

$$\begin{aligned}\|c(x_k+\alpha_k d_k)\|_1 - \|c_k\|_1 &\leq \|c_k + \alpha_k J_k d_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2 \\ &= |1-\alpha_k|\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2,\end{aligned}$$

where the equality follows from $J_k d_k = -c_k$. Therefore, when $\alpha_k \leq 1$,

$$\begin{aligned}\|c(x_k+\alpha_k d_k)\|_1 - \|c_k\|_1 &\leq (1-\alpha_k)\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2 \\ &\leq -\alpha_k\|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2.\end{aligned}$$

On the other hand, when $\alpha_k > 1$,

$$\begin{aligned}\|c(x_k+\alpha_k d_k)\|_1 - \|c_k\|_1 &\leq (\alpha_k - 1)\|c_k\|_1 - \|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2 \\ &= -\alpha_k\|c_k\|_1 + 2(\alpha_k-1)\|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2 \\ &\leq -\alpha_k\|c_k\|_1 + 2\alpha_k^2\|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2,\end{aligned}$$

where the final inequality follows by $\alpha_k > 1$.

Therefore, in either case, whenever $k \in \mathcal{K}_\alpha$, we have

$$\|c(x_k+\alpha_k d_k)\|_1 - \|c_k\|_1 \leq -\xi\alpha_k\|c_k\|_1 + 2\alpha_k^2\|c_k\|_1 + \frac{\alpha_k^2\Gamma}{2}\|d_k\|^2,$$

where we used $\xi \in (0,1)$.

Next, for any iteration where $k \in \mathcal{K}_\alpha^c$, it follows that

$$\|c(x_k+\alpha_k d_k)\|_1 - \|c_k\|_1 \leq (1-\xi\alpha_k)\|c_k\|_1 - \|c_k\|_1 = -\xi\alpha_k\|c_k\|_1.$$

Combining these cases and summing this inequality for $k = 0, \ldots, K-1$, it follows that

$$\|c(x_K)\|_1 - \|c_0\|_1 \leq -\sum_{k=0}^{K-1}\xi\alpha_k\|c_k\|_1 + \sum_{j\in\mathcal{K}_\alpha} 2\alpha_j^2\|c_j\|_1 + \frac{\alpha_j^2\Gamma}{2}\|d_j\|^2.$$

Next, by the orthogonal decomposition $d_k = v_k + \beta_k u_k$ and Lemma 2.1, we have

$$\begin{aligned}
&\|c(x_K)\|_1 - \|c_0\|_1 \\
&\le -\sum_{k=0}^{K-1} \xi\alpha_k\|c_k\|_1 + \sum_{j\in\mathcal{K}_\alpha} 2\alpha_j^2\|c_j\|_1 + \frac{\alpha_j^2\Gamma}{2}(\|v_j\|^2 + \beta_j^2\|u_j\|^2) \\
&\le -\sum_{k=0}^{K-1} \xi\alpha_k\|c_k\|_1 + \sum_{j\in\mathcal{K}_\alpha} (2+\Gamma\kappa_v/2)\alpha_j^2\|c_j\|_1 + \frac{\alpha_j^2\beta_j^2\Gamma}{2}\|u_j\|^2 \\
&\le -\sum_{k=0}^{K-1} \xi\alpha_k\|c_k\|_1 + \sum_{j\in\mathcal{K}_\alpha} \frac{(2+\Gamma\kappa_v/2)\nu^2}{q_{-1}^2 + \sum_{\ell\in\mathcal{K}_\alpha,\ell\le j}\|c_\ell\|_1}\|c_j\|_1 + \frac{\nu^2\eta^2\Gamma}{2Kq_{-1}^2}\|u_j\|^2,
\end{aligned}$$

where the final inequality follows by the definitions of β_k and α_k for any $k \in \mathcal{K}_\alpha$. Next, taking the expectation of both sides of this inequality, rearranging and using the law of iterated expectation with the result of Lemma 2.6, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[\alpha_k\|c_k\|_1\right] \\
&\le \xi^{-1}\|c_0\|_1 + \mathbb{E}\left[\sum_{j\in\mathcal{K}_\alpha} \frac{\xi^{-1}(2+\Gamma\kappa_v/2)\nu^2}{q_{-1}^2 + \sum_{\ell\in\mathcal{K}_\alpha,\ell\le j}\|c_\ell\|_1}\|c_j\|_1\right] + \mathbb{E}\left[\frac{\nu^2\eta^2\Gamma}{2Kq_{-1}^2}\|u_j\|^2\right] \\
&\le \xi^{-1}\|c_0\|_1 + \xi^{-1}(2+\Gamma\kappa_v/2)\nu^2\mathbb{E}\left[\log(q_{-1}^2 + \sum_{j\in\mathcal{K}_\alpha}\|c_j\|_1) - \log(q_{-1}^2)\right] \\
&\quad + \frac{\nu^2\eta^2\xi^{-1}\Gamma\zeta^{-1}(\kappa_H\kappa_u^2 + \zeta^{-1}M)}{2q_{-1}^2} \\
&\le \xi^{-1}\|c_0\|_1 + \xi^{-1}(2+\Gamma\kappa_v/2)\nu^2(2\log(\mathbb{E}[q_{K-1}]) - \log(q_{-1}^2)) \\
&\quad + \frac{\nu^2\eta^2\xi^{-1}\Gamma\zeta^{-1}(\kappa_H\kappa_u^2 + \zeta^{-1}M)}{2q_{-1}^2},
\end{aligned} \tag{4.5}$$

where the second inequality follows by Lemma 3.4 and the final inequality follows by Jensen's inequality and the concavity of $\log(x)$.

Therefore, since $\alpha_k \ge \nu/q_{K-1}$ for all $k \le K-1$, it follows that

$$\mathbb{E}\left[\sum_{k\in\mathcal{K}_\alpha}\frac{\|c_k\|_1}{q_{K-1}}\right] \le \kappa_9/\nu + (4+\Gamma\kappa_v)\xi^{-1}\nu\log(\mathbb{E}[q_{K-1}]).$$

Now, by the definition of q_{K-1} and the fact that $x \le a + b\log(x)$ implies $x \le 2a + 8b\log(e+b)$ for any $a > 0$ and $b > 0$ [24],

$$\begin{aligned}
\mathbb{E}[q_{K-1}] &= \mathbb{E}\left[\frac{q_{-1}^2 + \sum_{k\in\mathcal{K}_\alpha}\|c_k\|_1}{q_{K-1}}\right] \\
&\le q_{-1} + \kappa_9/\nu + (4+\Gamma\kappa_v)\xi^{-1}\nu\log(\mathbb{E}[q_{K-1}]) \\
&\le 2(q_{-1} + \kappa_9/\nu) + 8(4+\Gamma\kappa_v)\xi^{-1}\nu\log(e + (4+\Gamma\kappa_v)\xi^{-1}\nu),
\end{aligned}$$

proving the first result. Thus, by (4.5), it follows that

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\alpha_k \|c_k\|_1\right] \leq \kappa_9 + (4+\Gamma\kappa_v)\xi^{-1}\nu\log(\mathbb{E}[q_{K-1}]) \leq \kappa_9 + (4+\Gamma\kappa_v)\xi^{-1}\nu\log(\kappa_{10}).$$

From this proof, we can see that we still obtain a convergence rate of $\mathcal{O}(1/K)$ in terms of the average constraint violation. In addition, as in Section 3.1, any second order terms in the convergence analysis can be split into either terms involving β_k^2 or $\alpha_k^2\|c_k\|_1$ terms. Since α_k is bounded from above by a constant, it should be clear by the prior lemma that the sum of the $\alpha_k^2\|c_k\|_1$ terms are bounded, in expectation, by a constant factor. In addition, given the bound on $\mathbb{E}[q_{K-1}]$, we can combine the analysis of Sections 3.1 and 3.2 to derive a converegence result with a worst-case complexity of $\mathcal{O}(\epsilon_\ell^{-4})$ and $\mathcal{O}(\epsilon_c^{-1})$, matching the results of Section 3.1. We leave the full complexity analysis as an exercise to the reader.

5. Numerical Experiments. In this section, we numerically validate the performance of our proposed algorithm. We focus our attention on Algorithm 4.1, as it is a fully specified version of the generic Algorithm 2.1. We consider the performance of Algorithm 4.1 on a subset of the equality constrained problems from the CUTE collection [8]. We follow the experimental setup of [4] and select equality constrained optimization problems for which (i) f is not a constant function, (ii) n+$m \leq 1000$ and (iii) the Jacobian of c was non-singular at every iteration performed in our experiments. This selection resulted in a total of 60 problems, each of which has specified initial point, which we used in our experiments. We consider these problems at three different noise levels of $\epsilon_N \in \{10^{-5}, 10^{-3}, 10^{-1}\}$. At iteration k, a stochastic gradient is generated such that $g_k \sim \mathcal{N}(\nabla f(x_k), \epsilon_N I)$. For each problem and noise level, we ran a total of 20 instances for each algorithm. For each instance, all algorithms were given a total budget of 1000 iterations.

We compare Algorithm 4.1 with the Github implementation of Algorithm 3 in [4] and use the parameter settings provided in [4][3]. For Algorithm 4.1, for all problems and noise levels, we use the parameter settings $\beta_k = \beta = 10^{-1}$, $\nu = 1$, $q_{-1} = 10^{-9}$, $\theta = 1$, $\xi = 10^{-3}$ and $\rho = 1/2$.

For every run performed, we computed a resulting feasibility and optimality error. If a trial produced a sufficiently feasible iterate in the sense that $\|c_k\|_\infty \leq 10^{-6}$ for some k, then, we report the feasibility error as $\|c_k\|_\infty$ and the optimality error was reported as $\|\nabla f(x_k) + J_k^T y_k^{\text{true}}\|_\infty$, where y_k^{true} was computed as a least-squares multiplier using the true gradient $\nabla f(x_k)$ and J_k. (This ensures that the reported optimality error is not based on a stochastic gradient and is instead an accurate measure of optimality corresponding to the iterate x_k.) On the other hand, if no sufficiently feasible iterate was produced on a given run, then the feasibility error and optimality error were computed using the same measures at the least infeasible iterate computed. In addition to terminating when the maximum iteration limit is reached, the algorithms were terminated if they ever computed a point x_k which was both sufficiently feasible and the stationarity error was smaller than 10^{-4}. The results of this experiment are presented in Figure 5.1. In this figure, as well as in the following discussion, we refer to Algorithm 3 of [4] as SSQP, while Algorithm 4.1 is referred to as TSSQP.

As we can see from this plot, the computed stationarity error are relatively similar between these algorithms across all noise levels, with SSQP slightly outperforming

[3] https://github.com/frankecurtis/StochasticSQP

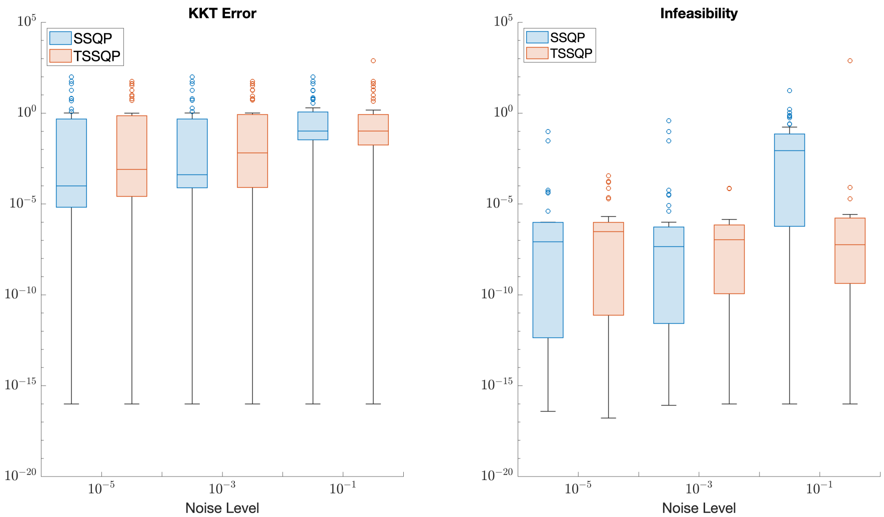

Fig. 5.1: Box plots of optimality error (left) and feasibility error (right) across various noise levels on CUTE problems. SSQP is Algorithm 3 of [4] and TSSQP is Algorithm 4.1.

TSSQP in stationarity error when the noise level is lower. This may be attributed to SSQP using an estimate of the merit parameter τ, which is more likely to be well-behaved in a low noise setting. However, once the noise level increases to $\epsilon_N = 10^{-1}$, the gap between these algorithms vanishes for the stationarity error. On the other hand, when the noise level is low, these algorithms perform similarly in terms of the infeasibility error. However, as the noise level increases, the performance of SSQP severely degrades with respect to infeasibility, while the performance of TSSQP is nearly unchanged. We view this as confirmation of our theoretical results as it demonstrates the superior ability of TSSQP to converge with respect to constraint violation while having minimal to no impact on its ability converge with respect to the KKT error.

6. Conclusion. In this paper, we propose and analyze a new SQP method for equality constrained optimization with a stochastic objective function. The algorithm uses a stepsize splitting scheme in order to improve upon the worst-case complexity of recently proposed stochastic SQP methods. We show that the proposed method matches the rate of convergence of a determinstic SQP method in terms of constraint violation and obtains the optimal rate for a stochastic method in terms of the gradient of the Lagrangian.

There are number of possible directions of future research. Fundamentally, this stepsize splitting scheme can be incorporated into any of the previously proposed stochastic SQP methods in the literature, including those for rank deficient Jacobians, inequality constraints, and inexact subproblem solutions. Designing new algorithms for these cases and deriving a worst-case complexity analysis are potential direction of future work.

REFERENCES

[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and V. 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, (2023).

[4] A. S. Berahas, F. E. Curtis, D. P. Robinson, and B. Zhou, *Sequential quadratic optimization for nonlinear equality constrained stochastic optimization*, SIAM Journal on Optimization, 31 (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] D. P. Bertsekas, *Network optimization: continuous and discrete models*, Athena Scientific Belmont, 1998.

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

[8] I. Bongartz, A. R. Conn, N. Gould, and P. L. Toint, *Cute: Constrained and unconstrained testing environment*, ACM Transactions on Mathematical Software (TOMS), 21 (1995), pp. 123–160.

[9] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM review, 60 (2018), pp. 223–311.

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

[11] S. Cuomo, V. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, *Scientific machine learning through physics–informed neural networks: Where we are and what's next*, Journal of Scientific Computing, 92 (2022), p. 88.

[12] F. E. Curtis, M. J. O'Neill, and D. P. Robinson, *Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization*, Mathematical Programming, 205 (2024), pp. 431–483.

[13] F. E. Curtis, D. P. Robinson, and B. Zhou, *Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints*, arXiv preprint arXiv:2302.14790, (2023).

[14] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization.*, Journal of machine learning research, 12 (2011).

[15] 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.

[16] F. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced sqp method*, Computational Optimization and Applications, 1 (1992), pp. 113–135.

[17] B. H. McMahan and M. Streeter, *Adaptive bound optimization for online convex optimization*, arXiv preprint arXiv:1002.4908, (2010).

[18] 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.

[19] S. Na, M. Anitescu, and M. Kolar, *Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming*, Mathematical Programming, 202 (2023), pp. 279–353.

[20] Y. Nandwani, A. Pathak, and P. Singla, *A primal dual formulation for deep learning with constraints*, Advances in neural information processing systems, 32 (2019).

[21] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, Journal of Computational physics, 378 (2019), pp. 686–707.

[22] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for pde-constrained optimization*, SIAM Journal on Scientific Computing, 32 (2010), pp. 271–298.

[23] Q. Shi, X. Wang, and H. Wang, *A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization*, Optimization Online, (2022).

[24] B. Wang, H. Zhang, Z. Ma, and W. Chen, *Convergence of adagrad for non-convex objectives:*

Simple proofs and relaxed assumptions, in The Thirty Sixth Annual Conference on Learning Theory, PMLR, 2023, pp. 161–190.
[25] R. Ward, X. Wu, and L. Bottou, *Adagrad stepsizes: Sharp convergence over nonconvex landscapes*, Journal of Machine Learning Research, 21 (2020), pp. 1–30.