LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-848453

AN ADAPTIVE SAMPLING AUGMENTED LAGRANGIAN METHOD FOR STOCHASTIC OPTIMIZATION WITH DETERMINISTIC CONSTRAINTS

R. Bollapragada, C. Karamanli, B. Keith, B. Lazarov, S. Petrides, J. Wang

May 1, 2023

Computers & Mathematics with Applications

Disclaimer

LLNL-JRNL-848453

AN ADAPTIVE SAMPLING AUGMENTED LAGRANGIAN METHOD FOR STOCHASTIC OPTIMIZATION WITH DETERMINISTIC CONSTRAINTS *

RAGHU BOLLAPRAGADA†, CEM KARAMANLI†, BRENDAN KEITH‡, BOYAN LAZAROV§, SOCRATIS PETRIDES¶, AND JINGYI WANG¶

Dedicated with respect and admiration to Leszek Demkowicz on the occasion of his 70th birthday anniversary.

Abstract. The primary goal of this paper is to provide an efficient solution algorithm based on the augmented Lagrangian framework for optimization problems with a stochastic objective function and deterministic constraints. Our main contribution is combining the augmented Lagrangian framework with adaptive sampling, resulting in an efficient optimization methodology validated with practical examples. To achieve the presented efficiency, we consider inexact solutions for the augmented Lagrangian subproblems, and through an adaptive sampling mechanism, we control the variance in the gradient estimates. Furthermore, we analyze the theoretical performance of the proposed scheme by showing equivalence to a gradient descent algorithm on a Moreau envelope function, and we prove sublinear convergence for convex objectives and linear convergence for strongly convex objectives with affine equality constraints. The worst-case sample complexity of the resulting algorithm, for an arbitrary choice of penalty parameter in the augmented Lagrangian function, is $\mathcal{O}(\epsilon^{-3-\delta})$, where $\epsilon > 0$ is the expected error of the solution and $\delta > 0$ is a user-defined parameter. If the penalty parameter is chosen to be $\mathcal{O}(\epsilon^{-1})$, we demonstrate that the result can be improved to $\mathcal{O}(\epsilon^{-2})$, which is competitive with the other methods employed in the literature. Moreover, if the objective function is strongly convex with affine equality constraints, we obtain $\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$ complexity. Finally, we empirically verify the performance of our adaptive sampling augmented Lagrangian framework in machine learning optimization and engineering design problems, including topology optimization of a heat sink with environmental uncertainty.

1. Introduction.

We consider constrained stochastic optimization problems of the form

$$\min_{x\in\mathcal{X}} f(x) \quad \text{subject to } c(x) = 0\,, \tag{1.1}$$

where the objective function $f\colon \mathbb{R}^n \to \mathbb{R}$ is the expected value $f(x) = \mathbb{E}_\zeta[F(x,\zeta)]$ of smooth random functions $F(\cdot,\zeta)\colon \mathbb{R}^n \to \mathbb{R}$, the constraint set $\mathcal{X} \subset \mathbb{R}^n$ is compact and convex, and the constraint function $c\colon \mathbb{R}^n \to \mathbb{R}^m$,

$$c(x) \stackrel{\text{def}}{=} Ax - b\,, \tag{1.2}$$

is an affine map with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Our primary motivation is to develop feasible strategies for solving optimal design problems with manufacturing and operational uncertainties [3, 27, 41, 73] (cf. Subsections 6.2 and 6.3) by efficiently solving

*Submitted to the editors 29th March 2024.

Funding: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and the LLNL-LDRD Program under Project tracking No. 22-ERD-009. Release number LLNL-JRNL-848453.

†Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX 78712 (raghu.bollapragada@utexas.edu, cem.karamanli@utexas.edu)

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (brendan_keith@brown.edu).

§Center for Design Optimization, Lawrence Livermore National Laboratory, Livermore, CA 94550 (lazarov2@llnl.gov)

¶Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550 (petrides1@llnl.gov, wang125@llnl.gov)

optimization problems of the form (1.1). Due to the inherently high computational cost, current design problems are often limited to low-dimensional sources of uncertainty or involve smoothly-varying random fields, which can be parameterized by a truncated series expansion with a small number of discrete random variables [46]. The above limitations restrict the practical applicability of some optimization approaches and lead to simplified heuristic procedures requiring subsequent manual intervention and suboptimal design performance [53]. Therefore, designing an efficient and robust optimization framework to address these challenges is crucial. Moreover, constrained stochastic optimization problems like (1.1) also commonly arise in other applications such as machine learning and finance; see, e.g., [4, 28, 75, 78] and references therein.

One of the well-known techniques to solve constrained optimization problems is the augmented Lagrangian method [13, 29, 30, 39, 70]. This method transforms the original constrained optimization problem (1.1) into a sequence of subproblems where the constraint violation is penalized in the objective. The main advantage of this transformation is that it enables using efficient algorithms for solving the subproblems. On the other hand, the major drawback is that multiple subproblems must be solved sequentially. To mitigate the cost of solving the subproblems, inexact solution mechanisms are widely used [43, 50, 54, 56, 72, 77, 86]. Although these mechanisms are well-understood for deterministic problems, the literature on their usage in stochastic settings is limited [55, 65]. Indeed, from our perspective, the main challenge in extending the augmented Lagrangian framework to stochastic approximation techniques lies in defining inexactness criteria for the stochastic methods used to solve the subproblems. In this work, we propose stochastic inexactness termination conditions that address this gap and guarantee convergence in expectation.

Adaptive sampling is a powerful technique that is used in stochastic optimization to control the accuracy of gradient estimates in a computationally efficient manner. The idea comes from the following observation, which is made mathematically precise later in the text: There is little need for an accurate gradient estimate in a stochastic solver when the iterates are far from the optimal solution. However, stochastic algorithms require increasingly accurate gradient estimates as the iterates get closer to the solution. To maintain accuracy, adaptive sampling methods dynamically increase the batch/sample size in response to an *a posteriori* estimate of the variance of the sampled gradients. Theoretical results from the adaptive sampling literature are promising. Indeed, in [23], the authors show that this methodology matches the best achievable complexity bound for *unconstrained* stochastic programs. Adaptive sampling is also known to be efficient in practice [18]. Recently, adaptive sampling methods have been used to develop efficient proximal/projected gradient algorithms for constrained optimization problems [7, 83]. Nevertheless, projecting gradients at every iteration can be challenging or inefficient, depending on the structure of the constraint set. Therefore, we go beyond the work in [7, 83] and consider augmented Lagrangian techniques. In turn, we address a more general class of algorithms and provide greater flexibility for treating the constraint set.

1.1. Contributions. In this paper, we propose an adaptive sampling augmented Lagrangian (ASAL) method by combining the augmented Lagrangian framework with adaptive sampling techniques to solve constrained stochastic optimization problems. We use adaptive sampling to control the accuracy of the gradient estimates when solving the subproblems obtained by penalizing the linear equality constraints. Moreover, we employ inexact solution mechanisms by imposing stochastic inexactness conditions to terminate the inner (i.e., subproblem) iterations. In this way,

we maximize the overall computational efficiency of our approach without sacrificing accuracy. Another important aspect of the methodology is that it relies on proximal/projected gradients to achieve feasibility with respect to the constraint $x \in \mathcal{X}$. Since the method relies only on gradient information, we establish sublinear convergence in the outer iterations for convex objective functions. Furthermore, given a user-defined algorithm parameter $\delta > 0$ and an arbitrary penalty parameter $\alpha > 0$, we find the total expected number of gradient evaluations to achieve an ϵ-accurate solution to be $\mathcal{O}(\epsilon^{-3-\delta})$. Moreover, if the penalty parameter is chosen to be sufficiently large, i.e., $\mathcal{O}(\epsilon^{-1})$, then our result improves to $\mathcal{O}(\epsilon^{-2})$. Finally, the worst-case complexity becomes $\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$ for strongly convex objective functions and when $\mathcal{X} = \mathbb{R}^n$. Table 1.1 compares our setting and theoretical results with the relevant literature. To evaluate the efficacy of our framework, we compare its performance to baseline algorithms in a collection of model problems from machine learning (Subsection 6.1) and engineering (Subsections 6.2 and 6.3).

TABLE 1.1

Summary of the theoretical convergence rate and sample complexity results in the relevant literature under different problem settings. In all the works mentioned here, the constraints are deterministic. Here, K denotes the (outer) iteration number, and ϵ denotes the required accuracy. Convergence rates are deterministic for deterministic problems and are in expectation for stochastic problems. Finally, sample complexity for stochastic solvers denotes the total number of expected stochastic gradient evaluations required to get $\epsilon-$accurate solutions.

paper	objective	set ($\mathcal{X}$)	constraints	rate (outer iter)	sample complexity
[50]	convex deterministic	convex compact	linear	$\mathcal{O}(1/K)$	-
[86]	convex deterministic	convex closed	convex	$\mathcal{O}(1/K)$	-
[84]	strongly convex stochastic	$\mathbb{R}^n$	linear	linear	$\mathcal{O}(\epsilon^{-1})$
[85]	convex stochastic nonsmooth	convex	convex	$\mathcal{O}(1/\sqrt{K})$	$\mathcal{O}(\epsilon^{-2})$
[85]	strongly convex stochastic nonsmooth	convex	convex	$\mathcal{O}(\log(K)/K)$	$\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$
Theorem 4.5 (arbitrary penalty parameter)	convex stochastic	convex compact	linear	$\mathcal{O}(1/K)$	$\mathcal{O}(\epsilon^{-3-\delta})$
Corollary 4.6 ($\mathcal{O}(\epsilon^{-1})$ penalty parameter)	convex stochastic	convex compact	linear	$\mathcal{O}(1/K)$	$\mathcal{O}(\epsilon^{-2})$
Theorem 4.12	strongly convex stochastic	$\mathbb{R}^n$	linear	linear	$\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$

1.2. Literature Review. The augmented Lagrangian method, also known as the method of multipliers, was first proposed by Hestenes [39] and Powell [70]. In [13], its performance is analyzed and compared to other common approaches, such as penalty and Lagrangian methods; see also [11, 33, 71, 72]. Although there have been extensive research efforts to enhance the performance of the basic augmented Lagrangian method to solve deterministic optimization problems (see, e.g., [14,29,30,50,54,56,86]), the current literature on stochastic optimization problems is limited [42,55,85]. In [42], the authors apply a stochastic augmented Lagrangian method to the domain adaptation problem. In [85], Xu developed stochastic primal-dual methods using the augmented Lagrangian function for solving nonsmooth optimization problems with a large number of constraints. In the aforestated approach, a projected stochastic gradient method is employed for the primal updates, while a randomized coordinate method is used for the dual updates.

For structured optimization problems with linear constraints, the alternating dir-

ection method of multipliers (ADMM) framework is often preferred [21]. There has been significant work on stochastic versions of the ADMM method [65, 81, 84, 88, 89]. In [65], the authors consider stochastic ADMM and show a $\mathcal{O}(\log(K)/K)$ convergence rate for strongly convex and $\mathcal{O}(1/\sqrt{K})$ for general convex objective functions. In [84], the authors design an inexact solution mechanism for the subproblems in stochastic ADMM when $\mathcal{X} = \mathbb{R}^n$. There, the authors employ the stochastic gradient method to solve the subproblems and show a linear convergence rate for strongly convex functions. Although our approach also involves inexact solutions, we consider adaptive sampling techniques to solve the subproblems and analyze both general convex and strongly convex functions. Moreover, our formulation allows us to consider implicit constraint sets (i.e., $\mathcal{X} \subset \mathbb{R}^n$) and utilizes only projected (or proximal) stochastic gradients. Other works achieve improved convergence rates by introducing stochastic variance reduction techniques (see, e.g., [59, 81, 88, 89]).

There are many articles on stochastic optimization methods with dynamic sample sizes [7, 16–19, 23, 25, 34–36, 47, 67, 74, 83]. Most of these works focus on unconstrained problems. Of note is the work by Friedlander and Schmidt [35], which shows linear convergence for finite-sum problems when the sample size increases at a geometric rate. Our work relates to the approach taken in Byrd et al. [23], which shows linear convergence of the expected risk minimization problem and calculates the worst-case complexity bounds for the number of gradient evaluations required to get ϵ-accurate solutions. Byrd et al. [23] also study the theoretical and practical aspects of the so-called *norm test*, which controls the sample sizes. Finally, in [7, 83], the authors consider adaptive sampling mechanisms for constrained stochastic programs. In both works, the constraints are represented by an abstract convex set, and the authors propose generalizations of the norm test that utilize projected (reduced) gradients.

Another common methodology to approach (1.1) is using sample average approximation (SAA) techniques [48, 49, 68, 76, 78] which replace the expected value in the objective function with a fixed sample average or other empirical approximation. When it comes to alternative techniques to solve constrained stochastic programs, the sequential quadratic programming (SQP) framework [9, 10, 31, 32, 61, 62] is also often utilized.

1.3. Notation. We denote the set of natural numbers by $\mathbb{N} \stackrel{\text{def}}{=} \{0, 1, 2, \dots\}$, and the set of positive natural numbers as $\mathbb{N}_+ \stackrel{\text{def}}{=} \{1, 2, \dots\}$. Throughout this work, $\|\cdot\|$ denotes the ℓ_2 vector norm or matrix norm and $\langle \cdot, \cdot \rangle$ denotes the ℓ_2-inner product. Finally, a matrix $A \in \mathbb{R}^{m \times n}$ is indicated to be positive definite by writing $A \succ 0$ and positive semi-definite by writing $A \succeq 0$. $A^T \in \mathbb{R}^{n \times m}$ denotes the transpose of a matrix A.

1.4. Organization. This paper is organized as follows. In Section 2, we introduce the preliminary material and assumptions used throughout the paper. The algorithmic framework and its components are given in Section 3. In Section 4, we analyze the convergence and complexity properties of our approach. Practical implementation of the algorithmic components is discussed in Section 5. We demonstrate the numerical performance of our methodology in Section 6. Finally, in Section 7, we provide concluding remarks and discuss avenues for future research.

2. Preliminaries and Assumptions. We provide preliminaries regarding the deterministic augmented Lagrangian method and its interpretation as a gradient descent method applied to the Moreau envelope of the dual function. We also state preliminary assumptions and recall results from the literature that are relied on later

in the paper.

2.1. Deterministic Augmented Lagrangian Method. The Lagrangian function for the problem (1.1) is

$$\ell(x, \lambda) = f(x) - \langle \lambda, c(x) \rangle, \tag{2.1}$$

where $\lambda \in \mathbb{R}^m$ is the Lagrangian (dual) parameter associated to the constraint function $c(x)$. Using (2.1), we can define the saddle-point problem,

$$\min_{x \in \mathcal{X}} \sup_{\lambda \in \mathbb{R}^m} \ell(x, \lambda), \tag{2.2}$$

and note that

$$\sup_{\lambda \in \mathbb{R}^m} \ell(x, \lambda) = \begin{cases} f(x) & \text{for} \quad c(x) = 0, \\ \infty & \text{for} \quad c(x) \neq 0. \end{cases}$$

Hence, if there exists $x \in \mathcal{X} \cap \{x \in \mathbb{R}^n \mid c(x) = 0\}$, then (2.2) is equivalent to (1.1) in the sense that

$$\min_{x \in \mathcal{X}} \sup_{\lambda \in \mathbb{R}^m} \ell(x, \lambda) = \min_{\{x \in \mathcal{X} | c(x) = 0\}} f(x)$$

and

$$\arg\min_{x \in \mathcal{X}} \sup_{\lambda \in \mathbb{R}^m} \ell(x, \lambda) = \arg\min_{\{x \in \mathcal{X} | c(x) = 0\}} f(x).$$

A primal-dual iterate pair $(\hat{x}, \hat{\lambda})$ is said to be a stationary point of (2.2) if

$$(\hat{x}, \hat{\lambda}) \in \left\{ (x, \lambda) \middle| \frac{\text{proj}_{\mathcal{X}}(x - \eta \nabla \ell_x(x, \lambda)) - x}{\eta} = 0 \text{ and } c(x) = 0 \right\}, \tag{2.3}$$

where $\eta > 0$ and

$$\text{proj}_{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{X}} \|x - y\|^2, \tag{2.4}$$

is the projection of $y \in \mathbb{R}^n$ onto the set $\mathcal{X}$ (see [29,50]). We also refer to the conditions in (2.3) as the

$$\textit{feasibility error}: \quad \|c(x)\|, \tag{2.5a}$$

and the

$$\textit{stationarity error}: \quad \left\| \frac{\text{proj}_{\mathcal{X}}(x - \eta \nabla \ell_x(x, \lambda)) - x}{\eta} \right\|. \tag{2.5b}$$

The augmented Lagrangian method is a class of iterative methods that produce stationary points satisfying (2.3) by solving a sequence of subproblems where the objective function is the sum of the Lagrangian function $\ell(x, \lambda)$ and a quadratic penalty term that penalizes violation of the equality constraint $c(x) = 0$. Specifically,

at any given iteration $k \in \mathbb{N}$, the basic primal and dual update rules are given as follows:

$$x_k^* \in \arg\min_{x \in \mathcal{X}} \ \mathcal{L}(x, \lambda_k; \alpha_k), \tag{2.6a}$$

$$\lambda_{k+1} = \lambda_k - \alpha_k c(x_k^*), \tag{2.6b}$$

where $\alpha_k > 0$ is the penalty parameter and

$$\mathcal{L}(x, \lambda; \alpha) = f(x) - \langle \lambda, c(x) \rangle + \frac{\alpha}{2} \|c(x)\|^2, \tag{2.7}$$

is the augmented Lagrangian function. Without restrictions on the objective function $f(x)$, the subproblem in (2.6a) may be unbounded. In this paper, we invoke assumptions that ensure this is not the case (cf. Assumption 2.2 or Assumption 2.3) as well as some other basic assumptions of additional utility.

2.2. Assumptions. We make the following assumptions about the objective function, the constraint function, and the existence of the solution.

ASSUMPTION 2.1. *The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is a convex continuously differentiable function on $\mathcal{X}$. That is, $\nabla^2 f(x) \succeq 0$, for all $x \in \mathcal{X}$. In addition, the gradient of the objective function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant $L < \infty$. That is,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathcal{X}.$$

Assumption 2.1 implies that the augmented Lagrangian function is also a convex function with respect to x on $\mathcal{X}$. That is, for any $\alpha > 0$,

$$\nabla^2_{xx} \mathcal{L}(x, \lambda; \alpha) = \nabla^2 f(x) + \alpha A^T A \succeq 0 \quad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}^m.$$

Note that the affine constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant $\|A\|$. That is, for all $x, y \in \mathcal{X}$,

$$\|c(x) - c(y)\| = \|A(x - y)\| \leq \|A\|\|x - y\|. \tag{2.8}$$

Moreover, as a consequence of Assumption 2.1 and (2.8), we can show that the gradient of the augmented Lagrangian function is Lipschitz continuous with respect to x on $\mathcal{X}$ with Lipschitz constant $L + \alpha\|A\|^2$. That is, due to (2.7),

$$\nabla_x \mathcal{L}(x, \lambda; \alpha) - \nabla_y \mathcal{L}(y, \lambda; \alpha) = \nabla f(x) - \nabla f(y) + \alpha \langle A, Ax - Ay \rangle,$$

and so

$$\|\nabla_x \mathcal{L}(x, \lambda; \alpha) - \nabla_y \mathcal{L}(y, \lambda; \alpha)\| \leq \left(L + \alpha\|A\|^2\right) \|x - y\|, \tag{2.9}$$

for all $x, y \in \mathcal{X}$.

ASSUMPTION 2.2. *The set $\mathcal{X} \subset \mathbb{R}^n$ is nonempty, convex, and compact. Also, there exists an optimal primal-dual pair (x^*, λ^*) that satisfies the optimality conditions (2.3).*

The compactness of set $\mathcal{X}$ implies that there exists a positive $D < \infty$ such that

$$\|x - y\| \leq D \quad \forall x, y \in \mathcal{X}. \tag{2.10}$$

Also, the existence of an optimal solution x^* implies that the problem in (2.6a) is bounded below. That is, for any $x \in \mathcal{X}$, $\lambda \in \mathbb{R}^m$, and $\alpha \geq 0$,

$$\begin{aligned}\mathcal{L}(x,\lambda;\alpha) \geq f(x) - \langle \lambda, c(x)\rangle &= f(x) - \langle \lambda, c(x) - c(x^*)\rangle \\ &\geq f(x) - \|\lambda\|\|c(x) - c(x^*)\| \\ &\geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle - \|\lambda\|\|c(x) - c(x^*)\| \\ &\geq f(x^*) - \|\nabla f(x^*)\|D - \|\lambda\|\|A\|D\,,\end{aligned}$$

where the first inequality is due to $\|c(x)\|^2 \geq 0$, the equality is due to $c(x^*) = 0$, the third inequality is due to convexity of function f (Assumption 2.1) and (2.10), and the last inequality is due to (2.8) and (2.10). Therefore, (2.6a) is well-defined.

We also develop results for the special case where the augmented Lagrangian function is strongly convex with respect to $x \in \mathcal{X}$.

ASSUMPTION 2.3. *The augmented Lagrangian is μ-strongly convex with respect to $x \in \mathcal{X}$. That is,*

$$\nabla^2_{xx}\mathcal{L}(x,\lambda) \succeq \mu I \quad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}^m,$$

where $I \in \mathbb{R}^{n\times n}$ is an identity matrix.

Note that if the objective function $f(x)$ is μ-strongly convex or A has full column rank, then Assumption 2.3 is trivially satisfied. Moreover, if Assumption 2.3 holds, then (2.6a) is well-defined for any $\lambda_k \in \mathbb{R}^m$.

We also make a standard assumption about the stochastic gradient of $f(x) = \mathbb{E}_\zeta[F(x,\zeta)]$.

ASSUMPTION 2.4. *The variance in the stochastic gradient of $f(x)$ is bounded. That is, there exist constants $\omega_1, \omega_2 \geq 0$ such that*

$$\mathbb{E}_\zeta[\|\nabla F(x,\zeta) - \nabla f(x)\|^2] \leq \omega_1\|\nabla f(x)\|^2 + \omega_2, \quad \forall x \in \mathcal{X}.$$

Using Assumptions 2.1, 2.2 and 2.4, it follows that

$$\begin{aligned}\mathbb{E}_\zeta[\|\nabla F(x,\zeta) - \nabla f(x)\|^2] &\leq 2\omega_1\|\nabla f(x) - \nabla f(x^*)\|^2 + 2\omega_1\|\nabla f(x^*)\|^2 + \omega_2 \\ &\leq 2\omega_1 L^2 D^2 + 2\omega_1\|\nabla f(x^*)\|^2 + \omega_2 \stackrel{\text{def}}{=} \omega\,,\end{aligned}$$

where the first inequality is due to the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^n$. In turn, we note that combining the assumptions above implies the existence of $\omega \geq 0$ such that

$$\mathbb{E}_\zeta[\|\nabla F(x,\zeta) - \nabla f(x)\|^2] \leq \omega, \quad \forall x \in \mathcal{X}. \tag{2.11}$$

2.3. Gradient Descent and the Moreau envelope. The convergence properties of the augmented Lagrangian method are often analyzed by showing its equivalence to a method (e.g., proximal point method) applied to dual problem (cf. [82]). We follow a similar approach in our analysis and show the equivalence of the augmented Lagrangian method and gradient descent method applied to the Moreau envelope [60] of the (negative) dual function. The negative of the dual function of (1.1) is denoted

$$q(\lambda) = -\min_{x\in\mathcal{X}} \ell(x,\lambda)\,, \tag{2.12}$$

and is known to be a convex, proper and continuous function from $\mathbb{R}^m$ to $\mathbb{R}$ [20]. For any given $\alpha > 0$, the Moreau envelope of $q(\lambda)$ is defined as follows [60]:

$$q_\alpha(u) \stackrel{\text{def}}{=} \min_\lambda \left[q(\lambda) + \frac{1}{2\alpha}\|\lambda - u\|^2 \right]. \tag{2.13}$$

In the following lemma, we summarize the important properties of Moreau envelopes.

LEMMA 2.1. *The function $q_\alpha(u)$ given in (2.13) is called the Moreau envelope of $q(\lambda)$ and satisfies the following properties.*

(i) [66, Equation 3.2] The gradient of the Moreau envelope is

$$\nabla q_\alpha(u) = \frac{1}{\alpha}(u - \text{prox}_{\alpha q}(u)), \tag{2.14}$$

where

$$\text{prox}_{\alpha q}(u) = \arg\min_\lambda \left[q(\lambda) + \frac{1}{2\alpha}\|\lambda - u\|^2 \right].$$

(ii) [6, Corollary 18.19] The gradients $\nabla q_\alpha(u)$ are Lipschitz continuous with Lipschitz constant $L_\alpha = \alpha^{-1}$. That is,

$$\|\nabla q_\alpha(u) - \nabla q_\alpha(v)\| \le \alpha^{-1}\|u - v\|, \quad \forall u, v \in \mathbb{R}^m. \tag{2.15}$$

(iii) [66, Page 136] The Moreau envelope retains the optimal value and the set of minimizers. That is,

$$\min_\lambda q(\lambda) = \min_\lambda q_\alpha(\lambda) \quad \textit{and} \quad \arg\min_\lambda q(\lambda) = \arg\min_u q_\alpha(u), \tag{2.16}$$

where the unique common minimizer $\lambda^* \in \mathbb{R}^m$ satisfies the fixed point equation $\lambda^* = \text{prox}_{\alpha q}(\lambda^*)$.

(iv) [69, Lemma 2.23] $q(\lambda)$ is strongly convex with parameter $\mu_q > 0$ if and only if $q_\alpha(u)$ is strongly convex with parameter $\mu_\alpha = \frac{\mu_q}{\mu_q \alpha + 1} > 0$.

Due to Assumption 2.2 and weak duality [20], we have that $q(\lambda)$ is bounded below. That is, the optimal value q^* is finite. Indeed,

$$q^* = \min_\lambda q_\alpha(\lambda) = \min_\lambda q(\lambda) = -\max_\lambda [-q(\lambda)] \ge -\min_{x \in \mathcal{X}, c(x)=0} f(x) = -f(x^*). \tag{2.17}$$

Owing to this fact and the properties of $q_\alpha(\lambda)$ in Lemma 2.1, the dual variable $\lambda \to \lambda^*$ will converge by iteratively minimizing $q_\alpha(\lambda)$ as in the gradient descent method. More explicitly, we may form a convergent sequence of dual variables as follows:

$$\begin{aligned} \lambda_{k+1} &= \lambda_k - \alpha \nabla q_\alpha(\lambda_k) \\ &= \arg\min_\lambda \left[q(\lambda) + \frac{1}{2\alpha}\|\lambda - \lambda_k\|^2 \right] \\ &= \arg\min_\lambda \left[-\min_{x \in \mathcal{X}} [\ell(x, \lambda)] + \frac{1}{2\alpha}\|\lambda - \lambda_k\|^2 \right] \\ &= \arg\max_\lambda \left[\min_{x \in \mathcal{X}} [\ell(x, \lambda) - \frac{1}{2\alpha}\|\lambda - \lambda_k\|^2] \right], \end{aligned} \tag{2.18}$$

where the second equality is due to (2.14) and third equality is due to (2.12). The function $\ell(x,\lambda) - \frac{1}{2\alpha}\|\lambda - \lambda_k\|^2$ is convex with respect to x on $\mathcal{X}$ and strongly concave with respect to λ. By Sion's Minimax Theorem [80], we can interchange the min and max operations (cf. [82, Section 10.5.2]) and obtain an equivalent characterization. That is,

$$\max_{\lambda}\left[\min_{x\in\mathcal{X}}[\ell(x,\lambda) - \frac{1}{2\alpha}\|\lambda-\lambda_k\|^2]\right] = \min_{x\in\mathcal{X}}\left[\max_{\lambda}[\ell(x,\lambda) - \frac{1}{2\alpha}\|\lambda-\lambda_k\|^2]\right] \tag{2.19}$$

$$= \min_{x\in\mathcal{X}}\left[\max_{\lambda}[f(x) - \langle\lambda, c(x)\rangle - \frac{1}{2\alpha}\|\lambda-\lambda_k\|^2]\right]. \tag{2.20}$$

Note that the optimal solution to the max problem (strongly concave in λ) in the second equality is $\lambda = \lambda_k - \alpha c(x)$. Substituting this expression into (2.20), we find

$$\max_{\lambda}\left[\min_{x\in\mathcal{X}}[\ell(x,\lambda) - \frac{1}{2\alpha}\|\lambda-\lambda_k\|^2]\right] = \min_{x\in\mathcal{X}}\mathcal{L}(x,\lambda_k;\alpha).$$

Hence, the dual update λ_{k+1} is given as follows:

$$x_k^* \in \arg\min_{x\in\mathcal{X}}\ \mathcal{L}(x,\lambda_k;\alpha) \tag{2.21a}$$

$$\lambda_{k+1} = \lambda_k - \alpha c(x_k^*). \tag{2.21b}$$

We now observe that the primal updates in (2.6a) and (2.21a) are both minimizers of the augmented Lagrangian function within the set $\mathcal{X}$. This optimization problem can have multiple optimal solutions when the augmented Lagrangian function $\mathcal{L}(x,\lambda_k;\alpha)$ is only a general convex function (not strongly convex). Hence, the updates (2.6a) and (2.21a) may not be the same. However, the dual updates are equivalent due to the following inequality [50, Equation 2.16]: For any $x \in \mathcal{X}$ and $x_k^* \in \arg\min_{x\in\mathcal{X}}\mathcal{L}(x,\lambda_k;\alpha)$,

$$\|c(x_k^*) - c(x)\|^2 \le \frac{2}{\alpha}\left(\mathcal{L}(x,\lambda_k;\alpha) - \mathcal{L}(x_k^*,\lambda_k;\alpha)\right). \tag{2.22}$$

Therefore, all solutions of $\min_{x\in\mathcal{X}}\ \mathcal{L}(x,\lambda_k;\alpha)$ have the same constraint function value $c(x)$ and the augmented Lagrangian method is equivalent to the gradient descent method applied to the Moreau envelope (2.13). Finally, we conclude this section on preliminary material by noting that

$$\nabla q_\alpha(\lambda_k) = c(x_k^*), \tag{2.23}$$

due to (2.18) and (2.21b).

3. Algorithmic Framework. This section begins with a description of a generic inexact augmented Lagrangian framework for solving (1.1). We then provide a complete description of our algorithm, which employs the adaptive sampling proximal gradient method [7,83] to minimize the augmented Lagrangian function (2.7) defined at each iteration.

Each primal variable update (2.6a) in the augmented Lagrangian method involves solving a computationally expensive optimization problem, namely,

$$\min_{x\in\mathcal{X}}\ \mathcal{L}(x,\lambda_k;\alpha), \tag{3.1}$$

where $\alpha > 0$ is the penalty parameter and λ_k is the dual variable at iteration $k \in \mathbb{N}$. Owing to the stochastic nature of sampling the objective function $f(x) = \mathbb{E}_\zeta[F(x,\zeta)]$, the exact solutions to these subproblems cannot be obtained efficiently. Therefore, we work with the inexact augmented Lagrangian framework outlined in Algorithm 3.1. At each iteration of this meta-algorithm, the subproblem (3.1) is solved (inexactly) by a given subproblem solver $\mathcal{S}$ until certain as yet unspecified inexactness conditions hold (cf. Subsection 3.2). Of course, the dual variable update incurs errors attributed to the inexact primal solves. However, if appropriate inexactness conditions are used to terminate the subproblem solver, then Algorithm 3.1 will still converge at the same rate as the exact algorithm (2.6), albeit in expectation.

Algorithm 3.1 Inexact Augmented Lagrangian Framework

Require: $x_{-1} \in \mathbb{R}^n$, $\lambda_0 \in \mathbb{R}^m$, $\alpha > 0$, inexactness conditions, solver $\mathcal{S}$.

1: **for** $k = 0, 1, \ldots$ **do**
2: Set starting point $x_{k,0} \leftarrow x_{k-1}$
3: Find an approximate minimizer x_k of (3.1) using solver $\mathcal{S}$, starting with $x_{k,0}$ such that some inexactness conditions are satisfied
4: Update $\lambda_{k+1} \leftarrow \lambda_k - \alpha c(x_k)$
5: **end for**

REMARK 3.1. We make the following remarks about Algorithm 3.1.

- **Solver and inexactness conditions:** For the sake of generality, we leave the description of the solver and inexactness conditions arbitrary and specify them in Subsection 3.1 and Subsection 3.2 respectively. We assume that the solver $\mathcal{S}$ can compute an approximate minimizer x_k that satisfies the inexactness conditions. The sequences of primal and dual iterates obtained in the algorithm are random due to the stochastic nature of the objective function $f(x)$. Therefore, this assumption is reasonable when the inexactness conditions are also stochastic.
- **Penalty parameter ($\alpha > 0$):** The algorithm employs a constant penalty parameter $\alpha > 0$. In Section 4, we show that the algorithm converges for any choice of this parameter and does not depend on problem characteristics or other algorithmic parameters.
- **Starting points ($x_{k,0}$):** At each iteration, the algorithm uses the previous primal iterate as starting point in the solver $\mathcal{S}$ to solve (3.1). This is meant to reduce the computational effort to solve (3.1). Since the successive augmented Lagrangian functions differ only in the dual variable λ_k, the approximate minimizer of the previous subproblem is an intuitive estimate of the solution to the current problem. In Section 4, we quantify the efficiency of this starting point rule in terms of total computational work.

We now describe the unspecified components of this algorithm: the solver $\mathcal{S}$ and the tolerance conditions.

3.1. Adaptive Sampling Proximal Gradient Method. Projected or proximal stochastic gradient methods are a popular class of methods for solving (3.1) when the projection or proximal operators are easy to compute [66]. The iterate update of a projected stochastic gradient method is given as follows:

$$x_{k,t+1} = x_{k,t} + \eta R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)\,, \tag{3.2}$$

where $\eta > 0$ is the step size parameter, $k \in \mathbb{N}$ denotes the *outer* augmented Lagrangian iteration counter, $t \in \mathbb{N}$ denotes the *inner* projected stochastic gradient iteration counter, $S_{k,t}$ is a set consisting of i.i.d. samples of ζ,

$$R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta) \stackrel{\text{def}}{=} \frac{\text{proj}_{\mathcal{X}}(x_{k,t} - \eta \nabla_x \mathcal{L}_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha)) - x_{k,t}}{\eta}, \tag{3.3}$$

$$\nabla_x \mathcal{L}_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha) \stackrel{\text{def}}{=} \frac{1}{|S_{k,t}|} \sum_{\zeta_i \in S_{k,t}} \nabla_x \mathcal{L}(x_{k,t}, \lambda_k, \zeta_i; \alpha), \tag{3.4}$$

and

$$\nabla_x \mathcal{L}(x_{k,t}, \lambda_k, \zeta_i; \alpha) = \nabla_x F(x_{k,t}, \zeta_i) - \langle \lambda_k, \nabla c(x_{k,t}) \rangle + \alpha \langle c(x_{k,t}), \nabla c(x_{k,t}) \rangle.$$

In what follows, it is helpful to note that $R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)$ denotes a stochastic approximation of the true projected (reduced) gradient

$$R(x_{k,t}, \lambda_k; \alpha, \eta) \stackrel{\text{def}}{=} \frac{\text{proj}_{\mathcal{X}}(x_{k,t} - \eta \nabla_x \mathcal{L}(x_{k,t}, \lambda_k; \alpha)) - x_{k,t}}{\eta}. \tag{3.5}$$

Two adaptive sampling strategies have recently been proposed for the projected stochastic gradient method [7, 83]. Both strategies employ a mechanism for improving the quality of the stochastic gradient approximation by updating the sample size $|S_{k,t}|$ on the fly at each (subproblem) iteration t. In turn, they overcome a significant limitation of fixed sample size strategies without compromising efficiency, while also maintaining the fast convergence of their deterministic counterparts. Indeed, fixed sample size strategies can only guarantee convergence to a neighborhood of the solution or must compromise on the convergence rate.

Adaptive sampling strategies aim to ensure that the variance in the stochastic gradient is controlled by the squared norm of the projected gradient. In [83], this is written as follows:

$$\mathbb{E}_{k,t}\left[\|\nabla_x \mathcal{L}_{S_{k,t}}(x_{k,t}, \lambda_k) - \nabla_x \mathcal{L}(x_{k,t}, \lambda_k)\|^2\right] \leq \theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)]\|^2, \tag{3.6}$$

where $\theta_g > 0$ is a given parameter, and

$$\mathbb{E}_{k,t}[\,\cdot\,] \stackrel{\text{def}}{=} \mathbb{E}[\,\cdot\,|\lambda_k, x_{k,t}], \tag{3.7}$$

denotes the expectation conditioned on the past iterates until $\lambda_k, x_{k,t}$. Specifically, $\mathbb{E}_{k,t}$ is the conditional expectation conditioned on the filtration

$$\mathbb{T}_{k,t} = \sigma(\lambda_0, x_{-1,0}, S_{0,0}, \ldots, S_{0,T_0}, \ldots, S_{k-1,0}, \ldots, S_{k-1,T_{k-1}}, S_{k,T_0}, \ldots, S_{k,t}),$$

where T_i denotes the number of inner iterations performed at the outer iteration i.

Using the definition of the gradient of the augmented Lagrangian function in (3.6) results in the following equivalent condition:

$$\mathbb{E}_{k,t}\left[\|\nabla F_{S_{k,t}}(x_{k,t}) - \nabla f(x_{k,t})\|^2\right] \leq \theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)]\|^2, \tag{3.8}$$

where

$$\nabla F_{S_{k,t}}(x_{k,t}) \stackrel{\text{def}}{=} \frac{1}{|S_{k,t}|} \sum_{\zeta_i \in S_{k,t}} \nabla F(x_{k,t}, \zeta_i). \tag{3.9}$$

Since the samples of ζ are i.i.d., Bienaymé's identity may be used to simplify the left-hand side of (3.8). This results in the following equivalent condition:

CONDITION 3.1 (Theoretical Sampling Condition). *For any given $\theta_g > 0$, the variance in the stochastic gradient of the objective function f is controlled by the squared norm of the expected projected gradient $R_{S_{k,t}}$. That is,*

$$\frac{\mathbb{E}_\zeta[\|\nabla F(x_{k,t},\zeta) - \nabla f(x_{k,t})\|^2]}{|S_{k,t}|} \leq \theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2. \tag{3.10}$$

This condition involves computing true variances and exact projected gradients that are unavailable in practice. Therefore, in Section 5, we also propose a practical version of this condition to control the sample sizes.

We conclude this subsection with remark and the following well-known result (adapted to our setting) [64, Corollary 2.3.2] that is used in the coming analysis.

PROPOSITION 3.1. *Suppose Assumptions 2.1 and 2.2 hold. Then, for any $0 < \eta < \frac{1}{L+\alpha\|A\|^2}$ and for all $x \in \mathcal{X}, \lambda \in \mathbb{R}^m, \alpha > 0$,*

$$\|R(x,\lambda;\alpha,\eta)\|^2 \leq \frac{2}{\eta}\left(\mathcal{L}(x,\lambda,\alpha) - \mathcal{L}(x^*_{\mathcal{L}},\lambda,\alpha)\right), \tag{3.11}$$

where $x^*_{\mathcal{L}} \in \arg\min_{x\in\mathcal{X}} \mathcal{L}(x,\lambda,\alpha)$. Moreover, if Assumption 2.3 also holds, then,

$$\frac{\mu}{2}\|x - x^*_{\mathcal{L}}\|^2 + \frac{\eta}{2}\|R(x,\lambda;\alpha,\eta)\|^2 \leq \langle R(x,\lambda;\alpha,\eta), x - x^*_{\mathcal{L}}\rangle. \tag{3.12}$$

REMARK 3.2 (Alternative Sampling Condition). An alternative sampling condition is proposed in [7] that would replace the right-hand side of (3.6) by a constant factor times the squared norm of the projected gradient (3.5). Following the procedure above, we would then arrive at a somewhat simpler inequality taking the place of (3.10), namely,

$$\frac{\mathbb{E}_\zeta[\|\nabla F(x_{k,t},\zeta) - \nabla f(x_{k,t})\|^2]}{|S_{k,t}|} \leq \tilde{\theta}_g^2 \|R(x_{k,t},\lambda_k;\alpha,\eta)]\|^2, \tag{3.13}$$

for some $\tilde{\theta}_g^2 > 0$. It turns out that the two conditions (3.10) and (3.13) are equivalent in the sense that their right-hand sides bound each other from above and below:

$$\frac{\|R(x_{k,t},\lambda_k;\alpha,\eta)\|}{1+\theta_g} \leq \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)\| \leq \frac{\|R(x_{k,t},\lambda_k;\alpha,\eta)\|}{1-\theta_g}. \tag{3.14}$$

Indeed, note that

$$\begin{aligned}
&\|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta) - R(x_{k,t},\lambda_k;\alpha,\eta)]\|^2 \\
&\qquad \leq \mathbb{E}_{k,t}[\|R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta) - R(x_{k,t},\lambda_k;\alpha,\eta)\|^2] \\
&\qquad = \eta^{-2}\mathbb{E}_{k,t}\Big[\|\operatorname{proj}_{\mathcal{X}}(x_{k,t} - \eta\nabla_x \mathcal{L}_{S_{k,t}}(x_{k,t},\lambda_k;\alpha)) \\
&\qquad\qquad\qquad - \operatorname{proj}_{\mathcal{X}}(x_{k,t} - \eta\nabla_x \mathcal{L}(x_{k,t},\lambda_k;\alpha))\|^2\Big] \\
&\qquad \leq \mathbb{E}_{k,t}\left[\|\nabla F_{S_{k,t}}(x_{k,t}) - \nabla f(x_{k,t})\|^2\right] \\
&\qquad \leq \theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2,
\end{aligned}$$

where the first line follows from Jensen's inequality, the proceeding equality is due to (3.3) and (3.5), the second inequality follows from the non-expansiveness property of projections [64], and the last inequality is due to (3.8). Rearranging terms and using the reverse triangle inequality, $\|a\| - \|b\| \leq \|a - b\|$, for all $a, b \in \mathbb{R}^n$, we arrive at (3.14). Moreover, both conditions (3.10) and (3.13) lead to identical practical algorithms; cf. Section 5. We choose to work with (3.10) instead of (3.13) because it leads to a simpler presentation of the complexity theory in Section 4.

3.2. Inexactness Conditions. The efficiency of an inexact augmented Lagrangian framework depends on its inexactness conditions. These conditions must balance the accuracy of the solution computed at each (outer) iteration and the overall computational efficiency. Due to the stochastic nature of the iterates obtained by our choice of the subproblem solver (cf. Subsection 3.1), these inexactness conditions must also be stochastic. We now propose inexactness conditions that meet these requirements based on the Moreau envelope perspective developed in Subsection 2.3.

Recall from Subsection 2.3 that the exact augmented Lagrangian method can be interpreted as a gradient descent method applied to the Moreau envelope of the (negative) dual function. Therefore, the inexact augmented Lagrangian method leads to inexact dual variable updates. That is, from the dual update (`line 3` in Algorithm 3.1), (2.6b), and (2.23), we have,

$$\begin{aligned} \lambda_{k+1} &= \lambda_k - \alpha c(x_k) \\ &= \lambda_k - \alpha c(x_k^*) + \alpha c(x_k^*) - \alpha c(x_k) \\ &= \lambda_k - \alpha \nabla q_\alpha(\lambda_k) + \alpha \epsilon_k, \end{aligned} \tag{3.15}$$

where x_k^* is an exact minimizer of (3.1) and $\epsilon_k \overset{\text{def}}{=} c(x_k^*) - c(x_k)$. We note that we have not imposed any structure on the subproblems (3.1) that would give the update error ϵ_k zero mean; i.e., $\mathbb{E}_k[\epsilon_k] \neq 0$, where

$$\mathbb{E}_k[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\lambda_k] \tag{3.16}$$

is the expected value operator conditioned on the iterates up until λ_k. Specifically, $\mathbb{E}_k$ is the conditional expectation conditioned on the filtration

$$\mathbb{T}_k = \sigma(\lambda_0, x_{-1,0}, S_{0,0}, \ldots, S_{0,T_0}, \ldots, S_{k-1,0}, \ldots, S_{k-1,T_{k-1}}),$$

where T_i denotes the number of inner iterations performed at the outer iteration i. In turn, we choose to view the additive update rule (3.15) as a *biased* stochastic gradient estimator update.

It is natural to consider an additional sampling condition when the sampling error can control the bias; cf. [7, Condition 2]. Such additional conditions are also common in trust-region methods [1, 5, 15]. In the present setting, however, the error is due to the subproblem solver. To address this, we aim to design a tolerance condition for terminating the inner loop. The following condition allows us to control the inexactness of the Moreau envelope gradient estimates in the dual update:

CONDITION 3.2 (Tolerance Condition I). *For any given $\theta_e \in [0, 1)$ and $\tau_k \geq 0$ with $\lim_{k\to\infty} \tau_k = 0$,*

$$\mathbb{E}_k\left[\|c(x_k^*) - c(x_k)\|^2\right] \leq \theta_e^2 \|c(x_k^*)\|^2 + \tau_k, \tag{3.17}$$

where x_k^* is a minimizer of (3.1).

From the Moreau envelope perspective (3.15), this condition ensures that the expected squared norm of the error, $\mathbb{E}_k[\|\epsilon_k\|^2]$, is controlled by the squared norm of the gradient of $q_\alpha(\lambda_k)$ and a vanishing positive constant τ_k. That is, from (2.23), (3.15), and (3.17), it follows that

$$\mathbb{E}_k[\|\epsilon_k\|^2] = \mathbb{E}_k\left[\|c(x_k^*) - c(x_k)\|^2\right] \leq \theta_e^2 \|c(x_k^*)\|^2 + \tau_k = \theta_e^2 \|\nabla q_\alpha(\lambda_k)\|^2 + \tau_k.$$

Likewise, this condition ensures that inexact gradient information can be employed far away from the solution (i.e., when the gradient's norm is large). Meanwhile, it also ensures accurate gradient information near the solution (i.e., when the gradient norm is small). Although this condition is derived from controlling the inexactness in the dual update, it directly relates to inexactness in the minimization of (3.1) (cf. (2.22)). Therefore, we can replace Condition 3.2 by the following alternative condition:

CONDITION 3.3 (Tolerance Condition II). *For any given $\theta_e \in [0, 1)$ and $\tau_k \geq 0$ with $\lim_{k\to\infty} \tau_k = 0$,*

$$\mathbb{E}_k \left[\mathcal{L}(x_k, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha)\right] \leq \frac{\alpha\theta_e^2 \|c(x_k^*)\|^2}{2} + \frac{\alpha\tau_k}{2}, \tag{3.18}$$

where x_k^* is a minimizer of (3.1).

Condition 3.3 controls the error in the minimization of (3.1) and directly implies Condition 3.2. Indeed, using (2.22) and (3.18), we find

$$\mathbb{E}_k \left[\|c(x_k^*) - c(x_k)\|^2\right] \leq \frac{2}{\alpha}\mathbb{E}_k \left[\mathcal{L}(x_k, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha)\right] \leq \theta_e^2 \|c(x_k^*)\|^2 + \tau_k. \tag{3.19}$$

When the augmented Lagrangian functions are strongly convex, we can also control the norm of the projected gradient (3.5). That is, by Assumption 2.3, (2.8), and (3.12), we have

$$\|c(x_k^*) - c(x_k)\|^2 \leq \|A\|^2 \|x_k^* - x_k\|^2 \leq \frac{4\|A\|^2}{\mu^2} \|R(x_k, \lambda_k; \alpha, \eta)\|^2. \tag{3.20}$$

Therefore, we can impose the following alternate condition when $f(x)$ is strongly convex.

CONDITION 3.4 (Tolerance Condition III). *For any given $\tilde{\theta}_e \in [0, 1)$ and $\tilde{\tau}_k \geq 0$ with $\lim_{k\to\infty} \tilde{\tau}_k = 0$,*

$$\mathbb{E}_k \left[\|R(x_k, \lambda_k; \alpha, \eta)\|^2\right] \leq \tilde{\theta}_e^2 \|c(x_k^*)\|^2 + \tilde{\tau}_k, \tag{3.21}$$

where x_k^* is a minimizer of (3.1).

Condition 3.4 also controls the error in the subproblem (3.1) and implies Condition 3.2. Indeed, set $\tilde{\theta}_e \leq \frac{\mu\theta_e}{2\|A\|}$ and $\tilde{\tau}_k \leq \frac{\mu^2\tau_k}{4\|A\|^2}$. Then, using (3.20) and (3.21), it holds that

$$\mathbb{E}_k \left[\|c(x_k^*) - c(x_k)\|^2\right] \leq \frac{4\|A\|^2}{\mu^2}\mathbb{E}_k \left[\|R(x_k, \lambda_k; \alpha, \eta)\|^2\right] \leq \theta_e^2 \|c(x_k^*)\|^2 + \tau_k. \tag{3.22}$$

REMARK 3.3. We observe that conditions similar to Conditions 3.2 through 3.4 have been proposed in the literature (cf. [50, 56, 72, 86]). The primary advantage of employing our conditions lies in their adaptive control over the subproblem error. Although verifying Conditions 3.2 through 3.4 for a stochastic subproblem solver can be challenging because they each require evaluating deterministic quantities, these conditions can still help us gain insight into the errors permitted in the algorithm while retaining desirable convergence properties. Furthermore, these conditions can guide the development of practical algorithms.

4. Theory. We now establish theoretical convergence guarantees and total sample complexity results for the proposed inexact augmented Lagrangian algorithmic framework when the inexactness conditions proposed in Subsection 3.2 are satisfied. We use the following notation for the full expectation:

$$\mathbb{E}[\,\cdot\,] = \mathbb{E}_0[\,\mathbb{E}_1[\,\cdots \mathbb{E}_k[\,\cdot\,]]]. \tag{4.1}$$

4.1. Convergence Results. We start by establishing a technical lemma.

LEMMA 4.1. *Suppose Assumptions 2.1 and 2.2 hold. For any x_{-1}, λ_0 and $\alpha > 0$, let $\{x_k, \lambda_k\}$ be the sequence of primal-dual iterates generated by Algorithm 3.1. Then, for all $k \in \mathbb{N}$,*

$$q_\alpha(\lambda_{k+1}) \leq q_\alpha(\lambda_k) - \frac{\alpha}{2}\|\nabla q_\alpha(\lambda_k)\|^2 + \frac{\alpha}{2}\|\epsilon_k\|^2, \tag{4.2}$$

where $\epsilon_k = c(x_k^*) - c(x_k)$ and x_k^* is a minimizer of (3.1).

Proof. From the dual update rule (`line 3` in Algorithm 3.1), (2.6b), and (2.23), it follows that

$$\lambda_{k+1} = \lambda_k - \alpha \nabla q_\alpha(\lambda_k) + \alpha \epsilon_k\,.$$

Using the Lipschitz continuity of $\nabla q_\alpha(\lambda)$ with Lipschitz constant $L_\alpha = \alpha^{-1}$ (cf. (2.15)) and the descent lemma [12], we have,

$$\begin{aligned} q_\alpha(\lambda_{k+1}) &\leq q_\alpha(\lambda_k) - \alpha\langle \nabla q_\alpha(\lambda_k) - \epsilon_k, \nabla q_\alpha(\lambda_k)\rangle + \frac{\alpha^2 L_\alpha}{2}\|\nabla q_\alpha(\lambda_k) - \epsilon_k\|^2 \\ &= q_\alpha(\lambda_k) - \frac{\alpha}{2}\|\nabla q_\alpha(\lambda_k)\|^2 + \frac{\alpha}{2}\|\epsilon_k\|^2\,, \end{aligned}$$

as necessary. ∎

We are now ready to establish convergence results for the inexactness conditions developed in Subsection 3.2.

THEOREM 4.2. *Suppose Assumptions 2.1, 2.2 and 2.4 hold. For any x_{-1}, λ_0 and $\alpha > 0$, let $\{(x_k, \lambda_k)\}$ be the sequence of primal-dual iterates generated by Algorithm 3.1. Furthermore, let $\theta_e \in [0, 1)$ and $\tau_k \geq 0$ such that $\tau_0^{-1}\sum_{k=0}^{\infty}\tau_k = a_\infty < \infty$. If any of the following three statements hold at each iteration $k \in N$:*

(a) the primal iterates x_k satisfy Condition 3.2;

(b) the primal iterates x_k satisfy Condition 3.3; or

(c) Assumption 2.3 also holds and the primal iterates x_k satisfy Condition 3.4 with $\tilde{\theta}_e \leq \frac{\mu\theta_e}{2\|A\|}$ and $\tilde{\tau}_k \leq \frac{\mu^2\tau_k}{4\|A\|^2}$;

then

$$\lim_{k\to\infty} \mathbb{E}[\|c(x_k)\|^2] = 0.$$

Moreover, for any $K \in \mathbb{N}$, we have that,

$$\min_{0\leq k\leq K-1} \mathbb{E}[\|c(x_k)\|^2] \leq \frac{4(1+\theta_e^2)}{\alpha(1-\theta_e^2)K}[q_\alpha(\lambda_0) - q^*] + \frac{4}{1-\theta_e^2}\frac{\tau_0 a_\infty}{K}, \tag{4.3}$$

where $q^* > -\infty$ is defined in (2.17). In addition, if either (b) or (c) is satisfied, then

$$\lim_{k\to\infty} \mathbb{E}\left[\left\|\frac{\mathrm{proj}_{\mathcal{X}}(x_k - \eta\nabla\ell_x(x_k, \lambda_{k+1})) - x_k}{\eta}\right\|^2\right] = 0\,, \tag{4.4}$$

for every $0 < \eta < \frac{1}{L+\alpha\|A\|^2}$.

Proof. If (a), (b), or (c) holds, then (3.17) holds as well due to (3.19) and (3.22). By Lemma 4.1, we have,

$$q_\alpha(\lambda_{k+1}) \leq q_\alpha(\lambda_k) - \frac{\alpha}{2}\|\nabla q_\alpha(\lambda_k)\|^2 + \frac{\alpha}{2}\|\epsilon_k\|^2.$$

Taking the conditional expectation (3.16) of both sides and invoking (2.23) and (3.17), we arrive at

$$\begin{aligned}\mathbb{E}_k\left[q_\alpha(\lambda_{k+1})\right] &\leq q_\alpha(\lambda_k) - \frac{\alpha}{2}\|\nabla q_\alpha(\lambda_k)\|^2 + \frac{\alpha}{2}\mathbb{E}_k\left[\|\epsilon_k\|^2\right] \\ &\leq q_\alpha(\lambda_k) - \frac{\alpha(1-\theta_e^2)}{2}\|c(x_k^*)\|^2 + \frac{\alpha}{2}\tau_k. \end{aligned} \tag{4.5}$$

Rearranging terms, we find

$$\|c(x_k^*)\|^2 \leq \frac{2}{\alpha(1-\theta_e^2)}(q_\alpha(\lambda_k) - \mathbb{E}_k[q_\alpha(\lambda_{k+1})]) + \frac{1}{1-\theta_e^2}\tau_k. \tag{4.6}$$

Therefore,

$$\begin{aligned}\mathbb{E}_k\left[\|c(x_k)\|^2\right] &\leq 2\mathbb{E}_k\left[\|c(x_k) - c(x_k^*)\|^2\right] + 2\|c(x_k^*)\|^2 \\ &\leq 2(1+\theta_e^2)\|c(x_k^*)\|^2 + 2\tau_k \\ &\leq \frac{4(1+\theta_e^2)}{\alpha(1-\theta_e^2)}(q_\alpha(\lambda_k) - \mathbb{E}_k[q_\alpha(\lambda_{k+1})]) + \frac{4}{1-\theta_e^2}\tau_k\,,\end{aligned}$$

where the first inequality is due to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^n$, the second inequality is due to (3.17) and the last inequality follows from (4.6). Taking the full expectation (4.1), and summing the above inequality from $k = 0$ to $K-1$, delivers

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\|c(x_k)\|^2\right] &\leq \frac{4(1+\theta_e^2)}{\alpha(1-\theta_e^2)}\mathbb{E}[q_\alpha(\lambda_0) - q_\alpha(\lambda_K)] + \frac{4}{1-\theta_e^2}\sum_{k=0}^{K-1}\tau_k \\ &\leq \frac{4(1+\theta_e^2)}{\alpha(1-\theta_e^2)}[q_\alpha(\lambda_0) - q^*] + \frac{4}{1-\theta_e^2}\tau_0 a_\infty\,,\end{aligned}$$

where the second inequality follows from (2.16) and the assumption $\sum_{k=0}^{\infty}\tau_k = \tau_0 a_\infty < \infty$. Observe that $q^* > -\infty$ due to (2.17), which follows from Assumption 2.2. Therefore,

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|c(x_k)\|^2\right] < \infty\,,$$

which implies that

$$\lim_{k\to\infty}\mathbb{E}\left[\|c(x_k)\|^2\right] = 0.$$

Moreover,

$$\begin{aligned}\min_{0\leq k\leq K-1}\mathbb{E}[\|c(x_k)\|^2] &\leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|c(x_k)\|^2] \\ &\leq \frac{4(1+\theta_e^2)}{\alpha(1-\theta_e^2)K}[q_\alpha(\lambda_0) - q^*] + \frac{4}{1-\theta_e^2}\frac{\tau_0 a_\infty}{K}.\end{aligned}$$

We will now analyze the stationarity error (4.4). Using (2.1), $\lambda_{k+1} = \lambda_k - \alpha c(x_k)$, and (2.7), it follows that,

$$\begin{aligned} x_k - \eta \nabla \ell_x(x_k, \lambda_{k+1}) &= x_k - \eta(\nabla f(x_k) - \langle \lambda_{k+1}, \nabla c(x_k) \rangle) \\ &= x_k - \eta(\nabla f(x_k) - \langle \lambda_k - \alpha c(x_k), \nabla c(x_k) \rangle) \\ &= x_k - \eta \nabla_x \mathcal{L}(x_k, \lambda_k; \alpha). \end{aligned}$$

Therefore,

$$\left\| \frac{\operatorname{proj}_{\mathcal{X}}(x_k - \eta \nabla \ell_x(x_k, \lambda_{k+1})) - x_k}{\eta} \right\|^2 = \|R(x_k, \lambda_k; \alpha, \eta)\|^2 . \tag{4.7}$$

If statement (b) holds, then it follows from (3.11) that

$$\begin{aligned} \mathbb{E}\left[\|R(x_k, \lambda_k; \alpha, \eta)\|^2 \right] &\leq \frac{2}{\eta} \left(\mathbb{E}\left[\mathcal{L}(x_k, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha) \right] \right) \\ &\leq \frac{\alpha \theta_e^2}{\eta} \mathbb{E}[\|c(x_k^*)\|^2] + \frac{\alpha}{\eta} \tau_k. \end{aligned}$$

If statement (c) holds, then

$$\mathbb{E}\left[\|R(x_k, \lambda_k; \alpha, \eta)\|^2 \right] \leq \tilde{\theta}_e^2 \mathbb{E}[\|c(x_k^*)\|^2] + \tilde{\tau}_k.$$

In turn, if either statements (b) or (c) holds, it follows that

$$\mathbb{E}\left[\|R(x_k, \lambda_k; \alpha, \eta)\|^2 \right] \leq \max\left\{ \frac{\alpha \theta_e^2}{\eta}, \tilde{\theta}_e^2 \right\} \mathbb{E}\left[\|c(x_k^*)\|^2 \right] + \max\left\{ \frac{\alpha}{\eta} \tau_k, \tilde{\tau}_k \right\}. \tag{4.8}$$

Taking the full expectation and summing the inequality (4.6) from $k = 0$ to $K - 1$, and observing that $q^* > -\infty$, we arrive at

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\|c(x_k^*)\|^2 \right] \leq \frac{2}{\alpha(1 - \theta_e^2)} [q_\alpha(\lambda_0) - q^*] + \frac{1}{1 - \theta_e^2} \tau_0 a_\infty < \infty, \tag{4.9}$$

which implies that

$$\lim_{k \to \infty} \mathbb{E}\left[\|c(x_k^*)\|^2 \right] = 0. \tag{4.10}$$

Taking limits on both sides of (4.8) and using (4.10) completes the proof. ∎

Theorem 4.2 establishes that the *expected* feasibility error vanishes as $k \to 0$, and meanwhile, the *smallest* feasibility error converges to zero at a sublinear rate. Moreover, the stationarity error also converges to zero *in expectation* when either Condition 3.3 or Condition 3.4 holds. However, the theorem does not guarantee any rate of convergence of the stationarity error. To establish such a result, we can perform one additional update at the iterate at which the expected feasibility error attains a minimum.

COROLLARY 4.3. *Suppose Assumptions 2.1, 2.2 and 2.4 hold. Let k_* be the iteration number at which $\min_{0 \leq k \leq K-1} \mathbb{E}[\|c(x_k)\|^2]$ is attained. That is,*

$$\mathbb{E}[\|c(x_{k_*})\|^2] = \min_{0 \leq k \leq K-1} \mathbb{E}[\|c(x_k)\|^2].$$

For any given $\tilde{\alpha} > 0$, $h \geq 0$, and $0 < \tilde{\eta} < \frac{1}{L + \tilde{\alpha}\|A\|^2}$, let $\tilde{x}$ be an approximate minimizer of $\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda_{k_*}, \tilde{\alpha})$ obtained with the starting point x_{k_*} that satisfies either of the following two statements:

(i) $\tilde{x}$ satisfies

$$\mathbb{E}_{k_*}\left[\mathcal{L}(\tilde{x},\lambda_{k_*};\tilde{\alpha})-\mathcal{L}(x^*_{k_*},\lambda_{k_*};\tilde{\alpha})\right]\le\frac{\tilde{\alpha}h}{2K}\,;\ \textit{or}$$

(ii) Assumption 2.3 also holds and $\tilde{x}$ satisfies

$$\mathbb{E}_{k_*}\left[\|R(\tilde{x},\lambda_{k_*};\tilde{\alpha},\tilde{\eta})\|^2\right]\le\frac{\mu^2 h}{4\|A\|^2K}\,.$$

Then, for $\tilde{\lambda}=\lambda_{k_*}-\tilde{\alpha}c(\tilde{x})$, we have

$$\mathbb{E}[\|c(\tilde{x})\|^2]\le\frac{2h}{K}+\frac{8(1+\theta_e^2)}{\alpha(1-\theta_e^2)K}[q_\alpha(\lambda_0)-q^*]+\frac{8}{1-\theta_e^2}\frac{\tau_0 a_\infty}{K} \tag{4.11a}$$

and

$$\mathbb{E}\left[\left\|\frac{\mathrm{proj}_{\mathcal{X}}(\tilde{x}-\tilde{\eta}\nabla\ell_x(\tilde{x},\tilde{\lambda}))-\tilde{x}}{\tilde{\eta}}\right\|^2\right]\le\frac{b}{K}, \tag{4.11b}$$

where $b=\frac{\tilde{\alpha}h}{\tilde{\eta}}$ if (i) holds and $b=\frac{\mu^2 h}{4\|A\|^2}$ if (ii) holds.

Proof. If either (i) or (ii) is satisfied, then using $\tilde{\alpha}$ as the penalty parameter in (3.19) and (3.22), it follows that

$$\mathbb{E}_{k_*}[\|c(\tilde{x})-c(x_{k_*})\|^2]\le\frac{h}{K}.$$

Therefore, taking the full expectation,

$$\begin{aligned}\mathbb{E}[\|c(\tilde{x})\|^2]&=\mathbb{E}[\|c(\tilde{x})-c(x_{k_*})+c(x_{k_*})\|^2]\\&\le 2\mathbb{E}[\|c(\tilde{x})-c(x_{k_*})\|^2]+2\mathbb{E}[\|c(x_{k_*})\|^2]\\&\le\frac{2h}{K}+\frac{8(1+\theta_e^2)}{\alpha(1-\theta_e^2)K}[q_\alpha(\lambda_0)-q^*]+\frac{8}{1-\theta_e^2}\frac{\tau_0 a_\infty}{K},\end{aligned}$$

where the first inequality is due to $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$ for any $a,b\in\mathbb{R}^n$, and the last inequality follows from (4.3). Now, consider the stationarity error. Similar to (4.7), we can show that

$$\mathbb{E}\left[\left\|\frac{\mathrm{proj}_{\mathcal{X}}(\tilde{x}-\tilde{\eta}\nabla\ell_x(\tilde{x},\tilde{\lambda}))-\tilde{x}}{\tilde{\eta}}\right\|^2\right]=\mathbb{E}\left[\|R(\tilde{x},\lambda_{k_*};\tilde{\alpha},\tilde{\eta})\|^2\right]\,. \tag{4.12}$$

Therefore, if (i) holds, it follows from (3.11) with $\tilde{\alpha}$ as the penalty parameter that

$$\mathbb{E}\left[\|R(\tilde{x},\lambda_{k_*};\tilde{\alpha},\tilde{\eta})\|^2\right]\le\frac{2}{\tilde{\eta}}\left(\mathbb{E}\left[\mathcal{L}(\tilde{x},\lambda_{k_*};\tilde{\alpha})-\mathcal{L}(x^*_{k_*},\lambda_{k_*};\tilde{\alpha}\right]\right)\le\frac{\tilde{\alpha}h}{\tilde{\eta}K}\,.$$

Likewise, if (ii) holds, then

$$\mathbb{E}\left[\|R(\tilde{x},\lambda_{k_*};\tilde{\alpha},\tilde{\eta})\|^2\right]\le\frac{\mu^2 h}{4\|A\|^2K}.$$

Substituting these inequalities into (4.12) completes the proof. ∎

4.2. Sample Complexity. We now establish the sample complexity for our inexact augmented Lagrangian algorithm, i.e., we estimate the worst-case expected total number of stochastic gradient evaluations to reach an $\epsilon-$accurate solution. To define accuracy, we specifically consider the following metric:

$$\max\left\{\mathbb{E}[\|c(\tilde{x})\|^2], \mathbb{E}\left[\left\|\frac{\mathrm{proj}_{\mathcal{X}}(\tilde{x}-\tilde{\eta}\nabla\ell_x(\tilde{x},\tilde{\lambda}))-\tilde{x}}{\tilde{\eta}}\right\|^2\right]\right\} \leq \epsilon, \tag{4.13}$$

for some $\epsilon \in (0,1)$. For the sake of brevity in this analysis, we employ Condition 3.3 as the inexactness condition with $\theta_e = 0$. At any outer iteration k, x_{k-1} is used as the starting point in the adaptive sampling proximal gradient method to solve the inner subproblem (3.1) until Condition 3.3 is satisfied. Recall that we define the index for the inner iterations as t, and the iterates in the inner loop as $x_{k,t}$. Since x_{k-1} is used as the starting iterate, we set $x_{k,0} \overset{\text{def}}{=} x_{k-1}$.

The adaptive sampling projected gradient method used to solve the inner subproblems (see Subsection 3.1) converges at a sublinear rate (cf. [7, Theorem 2.11], [83, Theorem 3.7]). The following theorem reformulates this result for augmented Lagrangian subproblems (3.1).

THEOREM 4.4. *Suppose Assumptions 2.1 and 2.3 hold. If* $\eta = \frac{(1-2\theta_g^2)}{L+\alpha\|A\|^2}$ *with* $\theta_g \in [0, \frac{1}{\sqrt{2}})$ *and Condition 3.1 is satisfied, then for any outer iteration* $k \in \mathbb{N}$ *and inner iteration* $t \in \mathbb{N}_+$*, it holds that*

$$\mathbb{E}_k[\mathcal{L}(x_{k,t},\lambda_k;\alpha) - \mathcal{L}(x_k^*,\lambda_k;\alpha)] \leq \frac{(L+\alpha\|A\|^2)\min_{x_k^*\in\mathcal{X}_k^*}\|x_{k-1}-x_k^*\|^2}{2(1-2\theta_g^2)t}, \tag{4.14}$$

where $\mathcal{X}_k^* = \arg\min_{x\in\mathcal{X}}\mathcal{L}(x,\lambda_k;\alpha)$.

Recall that $\mathbb{E}_k$ denotes expectation conditioned on the filtration $\mathbb{T}_k$, and note that the initial distance to optimality is in this filtration, i.e., $\min_{x_k^*\in\mathcal{X}_k^*}\|x_{k-1}-x_k^*\|^2 \in \mathbb{T}_k$.

Adaptive sampling methods are more efficient and robust in practice than methods that increase the sample sizes at predetermined rates. However, their sample complexity analysis has proven to be difficult, and establishing an upper bound on the sample sizes at each iteration poses significant challenges. Therefore, we make the following assumption based on the sample size growth rate over inner iterations t.

ASSUMPTION 4.1. *At any given outer iteration* $k \in \mathbb{N}$*, the expected sample size required to satisfy Condition 3.1 increases at a polynomial rate over the inner iterations* t*. More specifically, there exists* $c_0 \geq 0$ *and* $\delta_0 > 0$ *arbitrarily close to zero, such that*

$$\mathbb{E}_k[|S_{k,t}|] = \frac{c_0\omega}{\theta_g^2\min_{x_k^*\in\mathcal{X}_k^*}\|x_{k-1}-x_k^*\|^2}(t+1)^{1+\delta_0} \quad \forall k,t\in\mathbb{N}, \tag{4.15}$$

where $\mathcal{X}_k^* = \arg\min_{x\in\mathcal{X}}\mathcal{L}(x,\lambda_k;\alpha)$ *and* ω *is defined in (2.11).*

Predetermined sample growth rates similar to Assumption 4.1 are employed in unconstrained and constrained stochastic optimization settings [9, 67]. We acknowledge that Assumption 4.1 pertains to algorithmic quantities and is, therefore, less than ideal. Nevertheless, while we cannot rigorously prove this statement, we provide the following set of supporting (heuristic) arguments.

Consider rewriting Condition 3.1 in the following way:

$$b_{k,t} \stackrel{\text{def}}{=} \frac{\mathbb{E}_\zeta[\|\nabla f(x_{k,t},\zeta) - \nabla F(x_{k,t})\|^2]}{\theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2} \leq |S_{k,t}|\,. \tag{4.16}$$

This inequality is tight, i.e., $|S_{k,t}| = b_{k,t}$ when Condition 3.1 is satisfied with equality. In this case, due to (2.11), it follows that

$$b_{k,t} \leq \frac{\omega}{\theta_g^2 \|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2}. \tag{4.17}$$

On the other hand, using (4.14) and taking the expected value of both sides of (3.11) yields

$$\mathbb{E}_k[\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2] \leq \frac{(L+\alpha\|A\|^2)\min_{x_k^*\in\mathcal{X}_k^*}\|x_{k-1}-x_k^*\|^2}{\eta(1-2\theta_g^2)t}, \quad \forall t\in\mathbb{N}_+. \tag{4.18}$$

This inequality implies that the expected squared norm of the reduced gradient goes to zero at a sublinear rate.

Now, recall (3.14). In particular,

$$\|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2 \leq \frac{1}{(1-\theta_g)^2}\|R(x_{k,t},\lambda_k;\alpha,\eta)]\|^2\,.$$

Taking the conditional expectation $\mathbb{E}_k$ of both sides and invoking (4.18), it follows that

$$\mathbb{E}_k[\|\mathbb{E}_{k,t}[R_{S_{k,t}}(x_{k,t},\lambda_k;\alpha,\eta)]\|^2] \leq \frac{(L+\alpha\|A\|^2)\min_{x_k^*\in\mathcal{X}_k^*}\|x_{k-1}-x_k^*\|^2}{(1-\theta_g)^2\,\eta(1-2\theta_g^2)\,t}, \quad \forall t\in\mathbb{N}_+.$$

This inequality implies that the expected squared norm of the stochastic reduced gradient goes to zero at a sublinear rate. Therefore, it is possible to replace the right-hand-side of Condition 3.1 with a sublinearly convergent sequence $((t+1)^{-(1+\delta_0)})$ for any $t\in\mathbb{N}$ and achieve a similar sublinear convergence result as in Theorem 4.4. In such a scenario, the sample sizes satisfy Assumption 4.1. For the sake of brevity, in the rest of our analysis, we assume $\delta_0 = 0$ since δ_0 is arbitrarily close to zero.

We now state an equation that is useful to bound finite sum expressions in the complexity analysis. For any $\delta > 0$ and $K\in\mathbb{N}_+$, we have

$$\sum_{k=0}^{K} k^{1+\delta} < \int_{t=0}^{K+1} t^{1+\delta}dt = \frac{(K+1)^{2+\delta}}{2+\delta}. \tag{4.19}$$

We are now ready to prove the main theorem about outer iteration and sample complexity.

THEOREM 4.5. *Suppose Assumptions 2.1, 2.2 and 4.1 hold with $\delta_0 = 0$. For any x_{-1},λ_0 and $\alpha>0$, let $\{(x_k,\lambda_k)\}$ be the sequence of primal-dual iterates generated by Algorithm 3.1 where x_k satisfies Condition 3.3 at each outer iteration k with $\theta_e = 0$, $\tau_k = \frac{\tau_0}{(k+1)^{1+\delta/2}}$, $a_\infty = \sum_{k=0}^{\infty}\frac{1}{(k+1)^{1+\delta/2}}$, with $\tau_0\geq 0$, and $\delta>0$. Suppose the sample sizes $|S_{k,t}|$ satisfy Condition 3.1 with $\theta_g\in[0,\frac{1}{\sqrt{2}})$, $\eta = \frac{(1-2\theta_g^2)}{L+\alpha\|A\|^2}$. Under the conditions of Corollary 4.3 with $\tilde{x}$ satisfying (i) with $h\geq 0$, $\tilde{\eta} = \frac{(1-2\theta_g^2)}{L+\tilde{\alpha}\|A\|^2}$ and $\tilde{\alpha}>0$, the number of outer iterations to get an $\epsilon-$accurate solution $(\tilde{x},\tilde{\lambda})$ satisfying (4.13) is*

$$K_\epsilon = \left\lceil \frac{1}{\epsilon}\max\left\{\frac{8}{\alpha}[q_\alpha(\lambda_0)-q^*] + 2(4\tau_0 a_\infty + h), \frac{\tilde{\alpha}h}{\tilde{\eta}}\right\}\right\rceil. \tag{4.20}$$

Moreover, if the gradients of the Lagrangian function are bounded, i.e., $\|\nabla \mathcal{L}(x_{k-1}, \lambda_k; \alpha)\|^2 \leq D_{\mathcal{L}}$ *for all* $k \in \mathbb{N}$, *then the expected number of stochastic gradient evaluations is*

$$\mathbb{E}[\mathcal{W}] \leq \frac{B(K_\epsilon + 1)^{3+\delta}}{\alpha^2 \tau_0^2 (3+\delta)} + \frac{\tilde{B} K_\epsilon^2}{\tilde{\alpha}^2 h^2}, \tag{4.21}$$

where

$$B = \frac{2c_0\omega}{\theta_g^2}\left(\frac{(L + \alpha\|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}}\right),$$

and

$$\tilde{B} = \frac{2c_0\omega}{\theta_g^2}\left(\frac{(L + \tilde{\alpha}\|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}}\right).$$

Proof. Substituting $\theta_e = 0$ into (4.11), we obtain

$$\mathbb{E}[\|c(\tilde{x})\|^2] \leq \frac{8}{\alpha K}[q_\alpha(\lambda_0) - q^*] + \frac{2(4\tau_0 a_\infty + h)}{K},$$

and

$$\mathbb{E}\left[\left\|\frac{\operatorname{proj}_{\mathcal{X}}(\tilde{x} - \tilde{\eta}\nabla \ell_x(\tilde{x}, \tilde{\lambda})) - \tilde{x}}{\tilde{\eta}}\right\|^2\right] \leq \frac{\tilde{\alpha} h}{\tilde{\eta} K}.$$

Therefore, for any $K \geq K_\epsilon$ defined in (4.20), $(\tilde{x}, \tilde{\lambda})$ satisfies (4.13).

Let T_k be the first inner iteration at which Condition 3.3 is satisfied with $\theta_e = 0$. If $T_k = 0$, then we would have a sufficiently accurate starting point for the algorithm to terminate before the first complete iteration. Therefore, without loss of generality, we assume that $T_k > 0$. By Theorem 4.4, the inner subproblem termination condition, Condition 3.3 with $\theta_e = 0$, is satisfied at a given inner iteration $t \in \mathbb{N}_+$ if

$$\frac{(L + \alpha\|A\|^2)\min_{x_k^* \in \mathcal{X}_k^*}\|x_{k-1} - x_k^*\|^2}{2(1-2\theta_g^2)t} \leq \frac{\alpha\tau_k}{2}.$$

Thus, we have a deterministic upper bound $\Omega_k > 0$ on the random variable T_k; namely,

$$T_k \leq \Omega_k \stackrel{\text{def}}{=} \left\lceil \frac{(L + \alpha\|A\|^2)}{(1-2\theta_g^2)\alpha\tau_k} \min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x_k^*\|^2 \right\rceil. \tag{4.22}$$

We now analyze the total number of expected stochastic gradient evaluations. First, consider the expected sample complexity at each outer iteration k:

$$\begin{aligned}\mathbb{E}_k[\mathcal{W}_k] = \mathbb{E}_k\left[\sum_{t=0}^{T_k-1} |S_{k,t}|\right] &\leq \frac{c_0\omega}{\theta_g^2 \min_{x_k^* \in \mathcal{X}_k^*}\|x_{k-1} - x^*\|^2} \sum_{t=0}^{\Omega_k - 1}(t+1) \\ &\leq \frac{c_0\omega\Omega_k^2}{\theta_g^2 \min_{x_k^* \in \mathcal{X}_k^*}\|x_{k-1} - x^*\|^2}, \end{aligned} \tag{4.23}$$

where the first inequality is due to Assumption 4.1 with $\delta_0 = 0$. Substituting (4.22) into (4.23), using $\lceil x \rceil \le x+1$ and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^n$, we have that

$$\mathbb{E}_k[\mathcal{W}_k] \le \frac{2c_0\omega(L+\alpha\|A\|^2)^2 \min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x_k^*\|^2}{\theta_g^2(1-2\theta_g^2)^2\alpha^2\tau_k^2} + \frac{2c_0\omega}{\theta_g^2 \min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x^*\|^2}. \tag{4.24}$$

Now, $T_k > 0$ implies that Condition 3.3 is violated at $x_{k,0} = x_{k-1}$. That is,

$$\mathbb{E}_k\left[\mathcal{L}(x_{k-1}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha)\right] > \frac{\alpha\tau_k}{2}.$$

Recalling that x_{k-1} is in the filtration $\mathbb{T}_k$, and using convexity of $\mathcal{L}$, it follows that

$$\begin{aligned} \frac{\alpha\tau_k}{2} &< \mathcal{L}(x_{k-1}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha) \\ &\le \|\nabla\mathcal{L}(x_{k-1}, \lambda_k; \alpha)\| \|x_{k-1} - x_k^*\| \\ &\le \sqrt{D_{\mathcal{L}}} \|x_{k-1} - x_k^*\|, \end{aligned}$$

for all $x_k^* \in \mathcal{X}_k^*$. Therefore,

$$\min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x_k^*\|^2 > \frac{\alpha^2\tau_k^2}{4D_{\mathcal{L}}} \tag{4.25}$$

Now, summing the inequality (4.24) from $k = 0$ to $K-1$, taking full expectation, using (2.10) and (4.25), $\tau_k = \tau_0(k+1)^{-1-\delta/2}$, and (4.19), it follows that

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{K-1} \mathcal{W}_k\right] &\le \frac{2c_0\omega(L+\alpha\|A\|^2)^2}{\theta_g^2(1-2\theta_g^2)^2\alpha^2} \sum_{k=0}^{K-1} \frac{\mathbb{E}[\min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x_k^*\|^2]}{\tau_k^2} + \frac{8c_0\omega D_{\mathcal{L}}}{\theta_g^2\alpha^2} \sum_{k=0}^{K-1} \frac{1}{\tau_k^2} \\ &\le \frac{B}{\alpha^2\tau_0^2} \sum_{k=0}^{K-1} (k+1)^{2+\delta} \\ &\le \frac{B}{\alpha^2\tau_0^2(3+\delta)} (K+1)^{3+\delta}. \end{aligned} \tag{4.26}$$

We now consider the total number of stochastic gradients evaluated in the final step described in Corollary 4.3 with $\tilde{x}$ satisfying (i). Following a similar approach to the derivation of (4.24) and (4.25), and using (2.10), we have that

$$\mathbb{E}[\tilde{W}] \le \frac{2c_0\omega(L+\tilde{\alpha}\|A\|^2)^2K^2D^2}{\theta_g^2(1-2\theta_g^2)^2\tilde{\alpha}^2\tilde{h}^2} + \frac{8c_0\omega K^2 D_{\mathcal{L}}}{\theta_g^2\tilde{\alpha}^2\tilde{h}^2} = \frac{\tilde{B}K^2}{\tilde{\alpha}^2h^2}. \tag{4.27}$$

Finally, we can define the expected total number of gradient evaluations as

$$\mathbb{E}[\mathcal{W}] = \mathbb{E}\left[\sum_{k=0}^{K-1} \mathcal{W}_k\right] + \mathbb{E}[\tilde{\mathcal{W}}] \tag{4.28}$$

$$\le \frac{B(K+1)^{3+\delta}}{\alpha^2\tau_0^2(3+\delta)} + \frac{\tilde{B}K^2}{\tilde{\alpha}^2h^2}. \tag{4.29}$$

Substituting $K = K_\epsilon$ into (4.29) completes the proof. ∎

REMARK 4.1. In Theorem 4.5, we state an additional assumption related to the boundedness of the gradients of the augmented Lagrangian functions at the iterates computed by the algorithm. We note that this is a mild assumption and can be proven using Assumptions 2.1 and 2.2, and if the dual variables λ_k are bounded. Due to the convergence results established in Subsection 4.1, it is reasonable to assume that the dual variables are bounded.

4.3. Sample Complexity: $\alpha = \mathcal{O}(\epsilon^{-1})$. Theorem 4.5 establishes total outer iteration complexity, K_ϵ, and expected sample complexity, $\mathbb{E}[\mathcal{W}]$, for any choice of the penalty parameter α. If α and the other parameters (e.g., τ_0, h) given in Theorem 4.5 are chosen to be independent of the accuracy ϵ, then $K_\epsilon = \mathcal{O}(\epsilon^{-1})$ and $\mathbb{E}[\mathcal{W}] = \mathcal{O}(\epsilon^{-3-\delta})$. However, this sample complexity bound is not tight as the optimal sample complexity for stochastic convex programs is $\mathcal{O}(\epsilon^{-2})$ [51, 85]. The next corollary establishes that this optimal sample complexity can be achieved when $\alpha = \mathcal{O}(\epsilon^{-1})$.

COROLLARY 4.6. *Under the conditions of Theorem 4.5, if $\alpha = c_\alpha \epsilon^{-1}$, $\tau_0 = c_\tau \epsilon$, and $h = c_h \epsilon$ for some $c_\alpha, c_\tau, c_h \in (0, \infty)$. Then $K_\epsilon = \mathcal{O}(1)$ and*

$$\mathbb{E}[\mathcal{W}] = \mathcal{O}(\epsilon^{-2}). \tag{4.30}$$

Proof. Substituting α, τ_0, and h values into (4.20), it follows that

$$\begin{aligned} K_\epsilon &\leq 1 + \frac{1}{\epsilon} \max\left\{ \frac{8\epsilon}{c_\alpha}[q_\alpha(\lambda_0) - q_\alpha(\lambda^*)] + 2\epsilon(4c_\tau a_\infty + c_h), \frac{\tilde{\alpha} c_h \epsilon}{\tilde{\eta}} \right\} \\ &= 1 + \max\left\{ \frac{8}{c_\alpha}[q_\alpha(\lambda_0) - q_\alpha(\lambda^*)] + 2(4c_\tau a_\infty + c_h), \frac{\tilde{\alpha} c_h}{\tilde{\eta}} \right\} \\ &= \mathcal{O}(1)\,. \end{aligned} \tag{4.31}$$

We now analyze the sample complexity. Using α, τ_0, and h values, and $\epsilon < 1$, it follows that

$$\begin{aligned} \frac{B(K_\epsilon+1)^{3+\delta}}{\alpha^2 \tau_0^2 (3+\delta)} &= \frac{2c_0\omega}{\theta_g^2 c_\alpha^2 c_\tau^2 (3+\delta)} \left(\frac{(L + c_\alpha \epsilon^{-1} \|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}} \right) (K_\epsilon + 1)^{3+\delta} \\ &\leq \frac{2c_0\omega}{\epsilon^2 \theta_g^2 c_\alpha^2 c_\tau^2 (3+\delta)} \left(\frac{(L + c_\alpha \|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}} \right) (K_\epsilon + 1)^{3+\delta}, \end{aligned} \tag{4.32}$$

and

$$\frac{\tilde{B} K_\epsilon^2}{\tilde{\alpha}^2 h^2} = \frac{2c_0\omega}{\epsilon^2 \theta_g^2 \tilde{\alpha}^2 c_h^2} \left(\frac{(L + \tilde{\alpha} \|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}} \right) K_\epsilon^2. \tag{4.33}$$

Substituting (4.32) and (4.33) in (4.21), we have that,

$$\begin{aligned} \mathbb{E}[\mathcal{W}] &\leq \frac{2c_0\omega}{\epsilon^2 \theta_g^2 c_\alpha^2 c_\tau^2 (3+\delta)} \left(\frac{(L + c_\alpha \|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}} \right) (K_\epsilon + 1)^{3+\delta} \\ &\quad + \frac{2c_0\omega}{\epsilon^2 \theta_g^2 \tilde{\alpha}^2 c_h^2} \left(\frac{(L + \tilde{\alpha} \|A\|^2)^2 D^2}{(1-2\theta_g^2)^2} + 4D_{\mathcal{L}} \right) K_\epsilon^2 \\ &= \mathcal{O}(\epsilon^{-2}), \end{aligned}$$

where the last equality is due to the fact that all other constants in the inequality are independent of the choice of ϵ. ∎

REMARK 4.2. We observe that the complexity results given in Theorem 4.5 and Corollary 4.6 do not exploit the benefits of using the previous iterate $x_{k-1} = x_{k,0}$ as the starting point for solving the current subproblem. That is, the bound on $\mathbb{E}\left[\min_{x_k^* \in \mathcal{X}_k^*} \|x_{k-1} - x_k^*\|^2\right]$ is not tight. The difficulty in exploiting the benefits of this procedure is due to the fact that the augmented Lagrangian functions are only convex but not necessarily strongly convex. In Subsection 4.4, we consider strongly convex functions and establish the advantages of this procedure.

4.4. Sample Complexity: $\mathcal{X} = \mathbb{R}^n$. We provide improved convergence and complexity results when $\mathcal{X} = \mathbb{R}^n$ and the objective function f is μ-strongly convex.

ASSUMPTION 4.2. *The objective function f is μ-strongly convex. That is,*

$$\nabla^2 f(x) \succeq \mu I \quad \forall x \in \mathbb{R}^n$$

where $I \in \mathbb{R}^{n \times n}$ is an identity matrix.

We should note that Assumption 4.2 implies Assumption 2.3. In this case, the inner subproblems are unconstrained and have unique optimal solutions. Moreover, the optimality conditions given in (2.3) can be written as

$$\nabla \ell_x(x, \lambda) = 0 \quad \text{and} \quad c(x) = 0.$$

It can also be shown that the negative dual function $q(\lambda)$ is strongly convex in this setting, as stated in the following proposition (cf. [37, Propositions 3.1 and 3.3] and the references therein, [90, Theorem 1], [38, Proposition 2.5]).

PROPOSITION 4.7. *If Assumptions 2.1 and 4.2 hold with $\mathcal{X} = \mathbb{R}^n$, then $q(\lambda)$ defined in (2.12) is strongly convex with the strong convexity parameter $\mu_q = \frac{\sigma}{\mu + L}$ where $\sigma = \lambda_{\min}(AA^T)$.*

For the sake of completeness, we include the proof of this proposition in Appendix A. We also state the following well-known result for strongly convex functions with Lipschitz continuous gradients (cf. [64, Theorem 2.1.5 and Theorem 2.1.10])

PROPOSITION 4.8. *If the function $q_\alpha(\lambda)$ is strongly convex with parameter μ_α and has a Lipschitz continuous gradient with Lipschitz constant L_α, then for any $\lambda \in \mathbb{R}^m$, it holds that*

$$2\mu_\alpha(q_\alpha(\lambda) - q_\alpha(\lambda^*)) \leq \|\nabla q_\alpha(\lambda)\|^2 \leq 2L_\alpha(q_\alpha(\lambda) - q_\alpha(\lambda^*)), \tag{4.34}$$

where $\lambda^* = \arg\min_\lambda q_\alpha(\lambda)$.

Note that $L_\alpha = \alpha^{-1}$ by Lemma 2.1. We now establish a linear rate of convergence of both *feasibility error* and *stationarity error*. For the sake of brevity, we only consider Condition 3.4.

THEOREM 4.9. *Suppose Assumptions 2.1 and 4.2 hold and $\mathcal{X} = \mathbb{R}^n$. For any x_{-1}, λ_0 and $\alpha > 0$, let $\{(x_k, \lambda_k)\}$ be the sequence of primal-dual iterates generated by Algorithm 3.1. If the primal iterates x_k satisfy Condition 3.4 at each iteration $k \in \mathbb{N}$ with $\tilde{\theta}_e \leq \frac{\mu \theta_e}{2\|A\|}$, $\tilde{\tau}_k = \frac{\mu^2 \tau_k}{4\|A\|^2}$, $\theta_e \in [0, 1)$, and $\tau_k = \tau_0(1/a)^k$ for some $\tau_0 > 0$ and $a > 1$, then*

$$\mathbb{E}[\|c(x_k)\|^2] \leq A_1 \rho^k \quad \text{and} \quad \mathbb{E}\left[\|\nabla \ell_x(x_k, \lambda_{k+1}))\|^2\right] \leq A_2 \rho^k, \tag{4.35}$$

where $A_1 = 4(1+\theta_e^2)L_\alpha A_3 + 2\tau_0$, $A_2 = 2L_\alpha\tilde{\theta}_e^2 A_3 + \frac{\mu^2\tau_0}{4\|A\|^2}$, $A_3 = \max\left\{q_\alpha(\lambda_0) - q_\alpha(\lambda^*), \frac{\tau_0}{\mu_\alpha(1-\theta_e^2)}\right\}$, $\rho = \max\left\{1 - \frac{\alpha\mu_\alpha(1-\theta_e^2)}{2}, \frac{1}{a}\right\} < 1$, *and* $\mu_\alpha = \frac{\mu_q}{\mu_q\alpha+1}$.

Proof. Using Proposition 4.7 and Lemma 2.1, it follows that the $q_\alpha(\lambda)$ is a strongly convex function with strong convexity parameter $\frac{\mu_q}{\mu_q\alpha+1}$. Therefore, substituting (4.34) into (4.5), using (2.23), subtracting $q_\alpha(\lambda^*)$ from both sides and taking full expectation we obtain

$$\begin{aligned}\mathbb{E}[q_\alpha(\lambda_{k+1}) - q_\alpha(\lambda^*)] &\leq \left(1 - \alpha\mu_\alpha(1-\theta_e^2)\right)\mathbb{E}[q_\alpha(\lambda_k) - q_\alpha(\lambda^*)] + \frac{\alpha\tau_k}{2} \\ &\leq \left(1 - \alpha\mu_\alpha(1-\theta_e^2)\right)\mathbb{E}[q_\alpha(\lambda_k) - q_\alpha(\lambda^*)] + \frac{\alpha\tau_0}{2a^k},\end{aligned}$$

where the second inequality is due to $\tau_k = \tau_0(1/a)^k$. It is now a straightforward exercise in mathematical induction to show that

$$\mathbb{E}[q_\alpha(\lambda_k) - q_\alpha(\lambda^*)] \leq A_3\rho^k \quad \forall k \in \mathbb{N}. \tag{4.36}$$

The statement is trivially true for $k = 0$. Let's assume it is true for iteration k. For iteration $k+1$, it follows that

$$\begin{aligned}\mathbb{E}[q_\alpha(\lambda_{k+1}) - q_\alpha(\lambda^*)] &\leq \left(1 - \alpha\mu_\alpha(1-\theta_e^2)\right)\mathbb{E}[q_\alpha(\lambda_k) - q_\alpha(\lambda^*)] + \frac{\alpha\tau_0}{2a^k} \\ &\leq A_3\rho^k\left(1 - \alpha\mu_\alpha(1-\theta_e^2) + \frac{\alpha\tau_0}{2A_3(a\rho)^k}\right) \\ &\leq A_3\rho^k\left(1 - \alpha\mu_\alpha(1-\theta_e^2) + \frac{\alpha\mu_\alpha(1-\theta_e^2)}{2}\right) \\ &\leq A_3\rho^{k+1},\end{aligned}$$

where the second inequality is due to the statement of the induction, third inequality is due to $\rho \geq 1/a$ and the definition of A_3, and the last inequality is due to the definition of ρ. Hence, (4.36) is satisfied. Substituting $\lambda = \lambda_k$ in (4.34), by (2.23) and taking expectation of both sides it follows that,

$$\begin{aligned}\mathbb{E}[\|c(x_k^*)\|^2] &\leq 2L_\alpha\mathbb{E}[q_\alpha(\lambda_k) - q_\alpha(\lambda^*)] \\ &\leq 2L_\alpha A_3\rho^k.\end{aligned} \tag{4.37}$$

Therefore, using the definitions of A_1 and ρ, we have that

$$\begin{aligned}\mathbb{E}[\|c(x_k)\|^2] &\leq 2\mathbb{E}[\|c(x_k^*) - c(x_k)\|^2] + 2\mathbb{E}[\|c(x_k^*)\|^2] \\ &\leq 2(1+\theta_e^2)\mathbb{E}[\|c(x_k^*)\|^2] + 2\tau_k \\ &\leq \rho^k\left(4(1+\theta_e^2)L_\alpha A_3 + \frac{2\tau_0}{(\rho a)^k}\right) \\ &\leq A_1\rho^k.\end{aligned}$$

Using (4.37), Condition 3.4, and (4.7), it follows that

$$\begin{aligned}\mathbb{E}\left[\|\nabla\ell_x(x_k, \lambda_{k+1}))\|^2\right] &\leq \tilde{\theta}_e^2\mathbb{E}[\|c(x_k^*)\|^2] + \tilde{\tau}_k \\ &\leq 2L_\alpha\tilde{\theta}_e^2 A_3\rho^k + \frac{\mu^2\tau_0}{4\|A\|^2 a^k} \\ &\leq A_2\rho^k,\end{aligned}$$

where the last inequality is due to definitions of A_2 and $\rho a \geq 1$. ∎

We now derive the sample complexity results. We will use the fact that the metric (4.13) can be simplified to

$$\max\left\{\mathbb{E}[\|c(x_k)\|^2], \mathbb{E}\left[\|\nabla \ell_x(x_k, \lambda_{k+1})\|^2\right]\right\} \leq \epsilon \in (0,1), \tag{4.38}$$

since $\mathcal{X} = \mathbb{R}^n$. Moreover, the adaptive sampling projected gradient method employed to solve the inner subproblems converges at a linear rate as stated below (cf. [7, Theorem 2.10], [83, Theorem 3.7]).

THEOREM 4.10. *Suppose Assumptions 2.1, 2.4 and 4.2 hold. If* $\eta = \frac{(1-2\theta_g^2)}{L+\alpha\|A\|^2}$ *with* $\theta_g \in [0, \frac{1}{2})$*, and Condition 3.1 is satisfied. Then, for any outer iteration* $k \in \mathbb{N}$ *and inner iteration* $t \in \mathbb{N}$*, it holds that*

$$\mathbb{E}_k[\mathcal{L}(x_{k,t}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha)] \leq \rho_{\mathcal{L}}^t(\mathcal{L}(x_{k,0}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda_k; \alpha)), \tag{4.39}$$

where $\rho_{\mathcal{L}} = 1 - \frac{(1-2\theta_g^2)\mu}{L+\alpha\|A\|^2} \in [0,1)$.

Using Proposition 4.8, it can be shown that the gradient of the augmented Lagrangian function also converges to zero. That is, applying Proposition 4.8 to the augmented Lagrangian function, we have that

$$\begin{aligned} 2\mu(\mathcal{L}(x_{k,t}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda; \alpha)) &\leq \|\nabla_x \mathcal{L}(x_{k,t}, \lambda_k; \alpha)\|^2 \\ &\leq 2(L + \alpha\|A\|^2)(\mathcal{L}(x_{k,t}, \lambda_k; \alpha) - \mathcal{L}(x_k^*, \lambda; \alpha)). \end{aligned} \tag{4.40}$$

Combining (4.39) and (4.40), it follows that,

$$\mathbb{E}_k[\|\nabla_x \mathcal{L}(x_{k,t}, \lambda_k; \alpha)\|^2] \leq \frac{L + \alpha\|A\|^2}{\mu} \rho_{\mathcal{L}}^t \|\nabla_x \mathcal{L}(x_{k,0}, \lambda_k; \alpha)\|^2. \tag{4.41}$$

The next theorem establishes pessimistic upper bounds on the sample sizes employed at each outer iteration $k \in \mathbb{N}$ and each inner iteration $t \in \mathbb{N}$, and the number of inner iterations T_k required to satisfy Condition 3.4. For the sake of brevity, in this complexity analysis, we employ Condition 3.4 with $\tilde{\theta}_e = 0$, and also assume that Assumption 2.4 holds with $\omega_1 = 1$ and $\omega_2 = \omega$, i.e.,

$$\mathbb{E}_\zeta[\|\nabla F(x, \zeta) - \nabla f(x)\|^2] \leq \omega. \tag{4.42}$$

THEOREM 4.11. *Suppose the conditions of Theorem 4.10 are satisfied and (4.42) holds. Then the number of inner iterations* T_k *required to satisfy Condition 3.4 with* $\tilde{\theta}_e = 0$ *are bounded from above as follows:*

$$T_k \leq \left\lceil \log_{1/\rho_{\mathcal{L}}} \left(\frac{2(L + \alpha\|A\|^2)\left(\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + \|R(x_{k-1}, \lambda_{k-1}; \alpha, \eta)\|^2\right)}{\mu\tilde{\tau}_k} \right) \right\rceil. \tag{4.43}$$

Moreover, for any inner iteration $t < T_k$*, the sample sizes* $|S_{k,t}|$ *are at most*

$$|S_{k,t}| \leq \frac{\omega}{\theta_g^2 \tilde{\tau}_k}. \tag{4.44}$$

Proof. At any outer iteration $k \in N$, let T_k denote the first inner iteration t at which the following condition holds:

$$\min\left\{\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2, \mathbb{E}_{k,t}\left[\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2\right]\right\} \leq \tilde{\tau}_k. \tag{4.45}$$

Hence, at $t = T_k$, Condition 3.4 is satisfied with $\tilde{\theta}_e = 0$ at $x_k = x_{k,t}$. Therefore, for all $t < T_k$, it follows that

$$\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2 > \tilde{\tau}_k. \tag{4.46}$$

Using (3.14), (4.42), and (4.46), and choosing the smallest sample size $|S_{k,t}|$ satisfying (4.16), it follows that

$$|S_{k,t}| \leq \frac{\omega(1+\theta_g)^2}{\theta_g^2 \tilde{\tau}_k}. \tag{4.47}$$

Now, let us bound the number of inner iterations required to satisfy (4.45). Using (3.5) and (4.41) and $\mathcal{X} = \mathbb{R}^n$ it follows that

$$\mathbb{E}_k\left[\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2\right] \leq \frac{L+\alpha\|A\|^2}{\mu}\rho_{\mathcal{L}}^t \|R(x_{k,0},\lambda_k;\alpha,\eta)\|^2. \tag{4.48}$$

Substituting $x_{k,0} = x_{k-1}$, it follows that

$$\begin{aligned}
&\|R(x_{k-1},\lambda_k;\alpha,\eta)\|^2 \\
&\quad = \|R(x_{k-1},\lambda_k;\alpha,\eta) - R(x_{k-1},\lambda_{k-1};\alpha,\eta) + R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2 \\
&\quad \leq 2\|R(x_{k-1},\lambda_k;\alpha,\eta) - R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2 + 2\|R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2
\end{aligned} \tag{4.49}$$

where the last inequality is due to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^n$. Consider

$$\begin{aligned}
\|R(x_{k-1},\lambda_k;\alpha,\eta) - R(x_{k-1},\lambda_{k-1};\alpha,\eta)\| &= \|\nabla_x \mathcal{L}(x_{k-1},\lambda_k;\alpha) - \nabla_x \mathcal{L}(x_{k-1},\lambda_{k-1};\alpha)\| \\
&= \|\langle \lambda_k - \lambda_{k-1}, \nabla c(x_{k-1})\rangle\| \\
&\leq \alpha\|A\|\|c(x_{k-1})\|
\end{aligned} \tag{4.50}$$

where the first equality is due to (3.5) and the inequality is due to $\lambda_k = \lambda_{k-1} - \alpha c(x_{k-1})$. Using (4.48), (4.49), and (4.50), it follows that

$$\begin{aligned}
&\mathbb{E}_k\left[\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2\right] \\
&\qquad \leq \frac{L+\alpha\|A\|^2}{\mu}\rho_{\mathcal{L}}^t \left(2\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + 2\|R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2\right).
\end{aligned} \tag{4.51}$$

Therefore, for any

$$t \geq \left\lceil \log_{1/\rho_{\mathcal{L}}}\left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + \|R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2\right)}{\mu\tilde{\tau}_k}\right)\right\rceil$$

we have,

$$\mathbb{E}_k\left[\|R(x_{k,t},\lambda_k;\alpha,\eta)\|^2\right] \leq \tilde{\tau}_k.$$

Using (4.45), it follows that,

$$T_k \leq \left\lceil \log_{1/\rho_{\mathcal{L}}} \left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + \|R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2\right)}{\mu\tilde{\tau}_k} \right) \right\rceil \tag{4.52}$$

which completes the proof. ∎

We are now ready to provide the main complexity theorem for this subsection.

THEOREM 4.12. *Suppose the conditions of Theorems 4.9 and 4.10, and (4.42) hold. Then the number of outer iterations to get an $\epsilon-$accurate solution (x_k, λ_{k+1}) satisfying (4.38) is*

$$K_\epsilon = \left\lceil \log_{1/\rho} \left(\frac{\max\{A_1, A_2\}}{\epsilon} \right) \right\rceil = \mathcal{O}\left(\log(1/\epsilon)\right), \tag{4.53}$$

where A_1, A_2 are defined in Theorem 4.9, and the expected number of stochastic gradient evaluations is

$$\mathbb{E}[\mathcal{W}] = \mathcal{O}\left(\epsilon^{-1}\log(1/\epsilon)\right). \tag{4.54}$$

Proof. Equation (4.53) directly follows from (4.35). Now, consider the sample complexity at each outer iteration k

$$\mathcal{W}_k \overset{\text{def}}{=} \sum_{t=0}^{T_k-1} |S_{k,t}| \leq \frac{\omega(1+\theta_g)^2}{\theta_g^2\tilde{\tau}_k} T_k, \tag{4.55}$$

where the inequality is due to (4.44). Therefore, the expected total number of gradient evaluations is found to be

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{K-1}\mathcal{W}_k\right] &\leq \mathbb{E}\left[\sum_{k=0}^{K-1}\frac{\omega(1+\theta_g)^2}{\theta_g^2\tilde{\tau}_k}T_k\right] \\ &\leq \sum_{k=0}^{K-1}\frac{\omega(1+\theta_g)^2}{\theta_g^2\tilde{\tau}_k}\mathbb{E}[T_k] \\ &\leq \sum_{k=0}^{K-1}\frac{4\omega(1+\theta_g)^2\|A\|^2a^k}{\theta_g^2\tau_0\mu^2}\mathbb{E}[T_k], \end{aligned} \tag{4.56}$$

where the last inequality is due to $\tilde{\tau}_k = \frac{\mu^2\tau_k}{4\|A\|^2}$ and $\tau_k = \tau_0(1/a)^k$. Using (4.43) and

taking the full expectation of both sides, it follows that

$$
\begin{aligned}
&\mathbb{E}[T_k - 1] \\
&\leq \mathbb{E}\left[\log_{1/\rho_{\mathcal{L}}}\left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + \|R(x_{k-1},\lambda_{k-1};\alpha,\eta)\|^2\right)}{\mu\tilde{\tau}_k}\right)\right] \\
&= \mathbb{E}\left[\log_{1/\rho_{\mathcal{L}}}\left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2\|c(x_{k-1})\|^2 + \|\nabla_x \ell(x_{k-1},\lambda_k;\alpha,\eta)\|^2\right)}{\mu\tilde{\tau}_k}\right)\right] \\
&\leq \log_{1/\rho_{\mathcal{L}}}\left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2\mathbb{E}[\|c(x_{k-1})\|^2] + \mathbb{E}[\|\nabla_x \ell(x_{k-1},\lambda_k;\alpha,\eta)\|^2]\right)}{\mu\tilde{\tau}_k}\right) \\
&\leq \log_{1/\rho_{\mathcal{L}}}\left(\frac{2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2 A_1 + A_2\right)\rho^{k-1}}{\mu\tilde{\tau}_k}\right) \\
&\qquad (4.57) \\
&= \log_{1/\rho_{\mathcal{L}}}\left(\frac{8\|A\|^2(L+\alpha\|A\|^2)\left(\alpha^2\|A\|^2 A_1 + A_2\right)\rho^{k-1}a^k}{\mu^3\tau_0}\right) = \mathcal{O}(k)\,,
\end{aligned}
$$

where the second line is due to (4.7), third line due to Jensen's inequality, fourth line is due to Theorem 4.9, and the last line follows from $\tilde{\tau}_k = \frac{\mu^2\tau_k}{4\|A\|^2}$ and $\tau_k = \tau_0(1/a)^k$. Therefore, (4.57) shows that there exist $s_1 > 0$ and $s_2 > 0$ such that

$$
\mathbb{E}[T_k] \leq s_1 + s_2 k. \tag{4.58}
$$

Substituting (4.58) into (4.56), we have that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{K-1}\mathcal{W}_k\right] &\leq \sum_{k=0}^{K-1}\frac{4\omega(1+\theta_g)^2\|A\|^2 a^k}{\theta_g^2\tau_0\mu^2}(s_1 + s_2 k) \\
&\leq \sum_{k=0}^{K-1}\frac{4\omega(1+\theta_g)^2\|A\|^2 a^k}{\theta_g^2\tau_0\mu^2}(s_1 + s_2 K) \\
&\leq \frac{4\omega(1+\theta_g)^2\|A\|^2 a^K}{\theta_g^2\tau_0\mu^2(a-1)}(s_1 + s_2 K) \\
&= \mathcal{O}\left(\epsilon^{-1}\log(1/\epsilon)\right),
\end{aligned}
$$

where the last line is due to (4.53). ∎

Remark 4.3. It is important to emphasize that performing sampling complexity analysis for adaptive sampling methods is quite challenging with present optimization techniques. However, these methods fall under a general class of increasing batch size methods where one can establish theoretical sample complexity analysis that shows stochastic gradient and increasing batch size mechanisms have similar total sample complexity results (see, e.g., [23]). We have established pessimistic (i.e., worst-case) complexity bounds where the sample sizes at each inner iteration are bounded above by the largest sample size employed across all inner iterations at any given outer iteration k (cf. (4.44)). Owing to this pessimistic bound on sample sizes, the overall complexity bound $\mathcal{O}\left(\epsilon^{-1}\log(1/\epsilon)\right)$ is slightly worse than the optimal sample complexity $\mathcal{O}\left(\epsilon^{-1}\right)$ for strongly convex stochastic programming problems [19, 84].

5. Practical Algorithm. In this section, we present a complete and practical adaptive sampling augmented Lagrangian (ASAL) algorithm that uses an adaptive sampling proximal gradient method to inexactly solve the augmented Lagrangian subproblems. We describe the mechanism by which the sample size is selected at each inner iteration and the mechanism to terminate the subproblem solver.

The sample size selection and inexactness conditions described in Subsections 3.1 and 3.2 respectively are impractical as they require computing exact variances or deterministic quantities such as $\mathcal{L}(x_k, \lambda_k; \alpha)$ and $R(x_k, \lambda_k; \alpha, \eta)$. That being said, these quantities can be approximated using sample variances and sampled stochastic counterparts of the deterministic quantities following the ideas proposed in [7, 16, 83].

Sample Size Selection. We propose the following practical sampling test to approximate Condition 3.1 where the left-hand-side is the sample variance that approximates the exact variance and the right-hand-side is the stochastic projected (reduced) gradient that approximates the expectation of this quantity.

TEST 5.1 (Practical Sampling Test). *For any given $\theta_g \geq 0$, the sample size $|S_{k,t}|$ satisfies*

$$\frac{\frac{1}{|S_{k,t}|-1}\sum_{\zeta_i \in S_{k,t}} \|\nabla F(x_{k,t}, \zeta_i) - \nabla F_{S_{k,t}}(x_{k,t})\|^2}{|S_{k,t}|} \leq \theta_g^2 \|R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)\|^2. \tag{5.1}$$

In our practical Algorithm 5.1, we aim to satisfy Test 5.1 at each inner iteration using the following procedure. Whenever (5.1) is *not* satisfied at the current inner iteration t, we attempt to ensure (5.1) will be satisfied at the next inner iteration $t+1$ by using the relative variance,

$$\nu_t \overset{\text{def}}{=} \frac{\frac{1}{|S_{k,t}|-1}\sum_{\zeta_i \in S_{k,t}} \|\nabla F(x_{k,t}, \zeta_i) - \nabla F_{S_{k,t}}(x_{k,t})\|^2}{\theta_g^2 |S_{k,t}| \|R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)\|^2}, \tag{5.2}$$

to select the next sample size. More specifically, we set $|S_{k,t+1}| = \lceil \nu_t |S_{k,t}| \rceil$ whenever $\nu_t > 1$.

On the other hand, if (5.1) *is* satisfied at the current inner iteration t (i.e., $\nu_t \leq 1$), then keeping the sample size unchanged, $|S_{k,t+1}| = |S_{k,t}|$, is a simple rule to maintain control over the sample variance. However, if $\nu_t \ll 1$ is sufficiently small and the current sample size $|S_{k,t}| \gg 1$ is sufficiently large, then it may be beneficial to reduce cost by decreasing the sample size. We explore this possibility by providing an opportunity for the sample size to decrease like $|S_{k,t+1}| = \lceil \nu_t |S_{k,t}| \rceil$ until $|S_{k,t}|$ reaches a minimum value.[1] `Lines 8` through `16` in Algorithm 5.1 encapsulate the sample size selection procedure.

Inexactness Conditions. We propose a practical test to terminate the inner subproblem solver. Owing to the difficulty in computing the optimal quantities $c(x_k^*)$ and $\mathcal{L}(x_k^*, \lambda_k; \alpha)$, and the equivalence of Conditions 3.2 through 3.4, we design the practical test based on Condition 3.4. Following a similar procedure employed in approximating the sample size test conditions, we approximate the projected (reduced) gradient with its stochastic counterpart and the optimal constraint violation with the current constraint violation. The resulting practical test is as follows:

[1] Although the sample sizes are allowed to decrease, we do not observe sample size decreases in our numerical experiments; cf. Remark 6.1.

TEST 5.2 (Practical Tolerance Test). *For any given* $\tilde{\theta}_e \in [0,1)$ *and* $\tilde{\tau}_k \geq 0$ *with* $\lim_{k\to\infty} \tilde{\tau}_k = 0$,

$$\|R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)\|^2 \leq \tilde{\theta}_e^2 \|c(x_{k,t})\|^2 + \tilde{\tau}_k. \tag{5.3}$$

We terminate the inner subproblem whenever (5.3) is violated. Algorithm 5.1 provides a complete description of the ASAL algorithm.

Algorithm 5.1 Adaptive Sampling Augmented Lagrangian (ASAL) Method

Input: $x_{-1} \in \mathbb{R}^n$, $\lambda_0 \in \mathbb{R}^m$, step size $\eta > 0$, penalty parameter $\alpha > 0$, initial sample size $|S_{0,0}|$, sample size test parameters $(\theta_g > 0, \nu_l \in (0,1), s_l > 0, s_{min} > 0)$, inexactness tolerance parameters $(\tilde{\theta}_e \in [0,1), \tilde{\tau}_k \geq 0)$

Initialization: Set $k \leftarrow 0$

loop
  Set $t \leftarrow 0$
  Set $x_{k,0} \leftarrow x_{k-1}$
  repeat
    Choose a set $S_{k,t}$ consisting of $|S_{k,t}|$ i.i.d. realizations of ζ
    Compute $R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)$ via (3.3) and (3.4)
    Update $x_{k,t+1} \leftarrow x_{k,t} + \eta R_{S_{k,t}}(x_{k,t}, \lambda_k; \alpha, \eta)$
    if Test 5.1 is not satisfied **then**
      Set $|S_{k,t+1}| \leftarrow \lceil \nu_t |S_{k,t}| \rceil$
    else
      if $\nu_t < \nu_l$ and $|S_{k,t}| > s_l$ **then**
        Set $|S_{k,t+1}| \leftarrow \max\{s_{min}, \lceil \nu_t |S_{k,t}| \rceil\}$
      else
        Set $|S_{k,t+1}| \leftarrow |S_{k,t}|$
      end if
    end if
    Set $t \leftarrow t + 1$
  until Test 5.2 is satisfied
  Set $x_k \leftarrow x_{k,t}$
  Update $\lambda_{k+1} \leftarrow \lambda_k - \alpha c(x_k)$
  Set $|S_{k+1,0}| \leftarrow |S_{k,t}|$
  Set $k \leftarrow k + 1$
end loop

6. Numerical results. In this section, we study the performance of ASAL (Algorithm 5.1) using model problems from machine learning (Subsection 6.1) and engineering (Subsections 6.2 and 6.3). We implement Test 5.2 with $\tilde{\theta}_e = 0$ and $\tilde{\tau}_k = \tau_0/(k+1)$, treating τ_0 as a hyperparameter for this numerical study.

6.1. Logistic regression with multiple disparate impact constraints. We first consider a constrained logistic regression problem. A decision-making system suffers from *disparate impact* if it provides outputs that affect a group of people sharing a value of a sensitive feature more frequently than other groups [4]. In [87, Section 4.4], it is shown that disparate impact can be controlled in binary classification problems by applying deterministic constraints. More explicitly, we consider the optimization

problem

$$\begin{aligned} &\text{minimize} && \frac{1}{N}\sum_{i=1}^{N}\left[\log(1+\exp(-z_i\langle x, y_i\rangle))\right] + \frac{\gamma}{2}\|x\|^2 \\ &\text{subject to} && \langle a_1, x\rangle = b_1, \quad |\langle a_2, x\rangle| \le b_2, \end{aligned} \tag{6.1}$$

where $x \in \mathbb{R}^n$ is the optimization variable and $(y_i, z_i) \in \mathbb{R}^n \times \{-1, 1\}$ are input/output pairs from a classification data set. Here, $\gamma > 0$ is a fixed Tikhonov regularization parameter. Meanwhile, a_1, $a_2 \in \mathbb{R}^n$ and b_1, $b_2 \geq 0$ are constraint parameters. In [4], it is suggested to take, e.g., $a_1 = \mathbb{E}_{y,s}\big[(s - \mathbb{E}_s[s])y\big]$, where s is a secondary observable, in addition to y. However, for the purpose of demonstration, we arbitrarily set a_1 and a_2 from samples drawn for a standard multivariate normal distribution. Likewise, we set $b_1 = 0.1, b_2 = 0.02$. The initial x_{-1} and λ_0 variables are chosen to be zero vectors, and we set $\gamma = 1/N$.

In this experiment, we use the `mushroom` classification data set from the LIBSVM collection [26]. The size of this data set is $N = 8124$, and the dimension of the problem is $n = 112$. In order to evaluate the performance of ASAL, we record the feasibility and stationarity errors (2.5) until 200 training epochs (i.e., $200N$ cumulative gradient evaluations) have elapsed. We then compare ASAL to three separately-tuned fixed-batch-size implementations of ASAL using 10%, 20%, and 50% of the data set size at each iteration, respectively. In this experiment we use $\theta_g = 0.99$, $\nu_1 = 0.5$, $s_1 = 0.1N$, and $s_{\min} = 0.1N$. The value of θ_g is *not tuned* and is, instead, set at an arbitrary value close to the suggestion for unconstrained problems in [24, 34]. The values of the other three fixed hyperparameters are also set arbitrarily. Yet, they appear to have little to no effect on performance; cf. Remark 6.1.

We treat τ_0, α and the step size η as tunable hyperparameters. All of the hyperparameters are tuned using the following procedure: We run each augmented Lagrangian algorithm for all possible combinations of $\tau_0 = 10^4, 10^3, 10^2, 10^1, 10^0, 10^{-1}$, $\eta = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$, and $\alpha = 10^2, 10^1, 10^0, 10^{-1}, 10^{-2}$. Then, for each algorithm, we select the run with the smallest average objective function value in the final 5 inner iterations among all runs whose minimum feasibility error in the final 30 inner iterations is less than feasibility tolerance 10^{-4}.

The stationarity and feasibility errors corresponding to the best hyperparameters for each algorithm are overlaid in Figure 6.1. Because the hyperparameter tuning procedure we have used seeks the best stationarity error among runs reaching a feasibility error threshold, it is no surprise that ASAL and each of the three baseline algorithms achieve a similar *minimal* feasibility error (around feasibility tolerance 10^{-4}). Nevertheless, we observe that ASAL outperforms the three baseline algorithms with respect to stationarity error. We also present similar results for `australian` data set from the LIBSVM collection [26] in Appendix B.

REMARK 6.1. Notice from Figure 6.1 that the ASAL sample size never decreases. This is despite the safeguarding mechanism in `line 17` of Algorithm 5.1. We have witnessed this non-decreasing sample size property in all of our experiments with ASAL after tuning the hyperparameters α, η, and τ_0. Thus, we see little justification for allowing sample size decreases in future implementations of ASAL and do not report the hyperparameters ν_1, s_1, and $s_{\min}$ in the remaining experiments.

REMARK 6.2. The starting values for the cumulative gradient evaluations in Figure 6.1 represent the fact that we are recording errors only *after* advancing a single optimization step. Each algorithm began with the same initial guesses x_{-1} and λ_0.

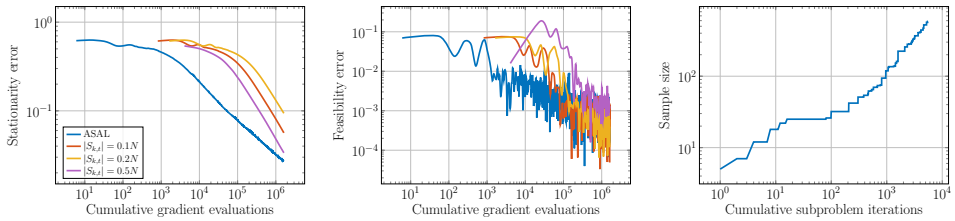

FIGURE 6.1. *Results of running Algorithm 5.1 on the constrained logistic regression problem (6.1) with the* `mushroom` *classification data set. Notice that ASAL achieves the lowest average stationarity errors while matching the minimal feasibility error of the three baseline algorithms. To generate these results, we used the algorithm parameters* $\theta_g = 0.99$ *and individually tuned* α*,* η*, and* τ_0*.*

REMARK 6.3. Observe that the expected feasibility error with ASAL steadily decreases in Figure 6.1. Meanwhile, the feasibility error in each of the other algorithms plateaus after around 5×10^5 cumulative gradient evaluations. This is due to the stable sample size growth provided by our adaptive sampling strategy and the fact that a fixed number of samples are used for each of the baselines; i.e., the baseline algorithms can only converge in expectation to a neighborhood of the solution. As a result, even though the slopes of the stationarity errors for the baseline algorithms are higher than ASAL after 200 epochs, we conclude that ASAL would remain the better practical algorithm even if a larger epoch threshold had been used.

REMARK 6.4. The tuning procedure used in this experiment is expensive and impractical for more expensive problems. Owing to this fact, in the remaining sections, we only compare ASAL to baseline algorithms with a shared set of hyperparameters.

6.2. Optimal truss design. We consider optimizing the simply supported truss structure shown in Figure 6.2 in a problem inspired by an example presented in [73]. The truss elements are numbered as shown in Figure 6.2, and a random force F, pointing downwards, is applied in the middle of the bottom chord. The cross-sections of the truss elements are denoted by x^i, $i = 1, 2, \ldots, 7$, and the yield stress associated with the members with σ_i, $i = 1, 2, \ldots, 7$. The first two yield stress limits σ_i, $i = 1, 2$, are log-normal random variables with mean 100N/mm^2 and standard deviation 20N/mm^2. The yield stresses for all other members are also log-normal, but with mean 200N/mm^2 and standard deviation 40N/mm^2. The correlation coefficient between σ_1 and σ_2 is 0.8, and between σ_i, $i = 1, 2$, and σ_j, $j = 3, 4, 5, 6, 7$, the correlation coefficients are each 0.5. The correlation coefficients between each $\sigma_i \neq \sigma_j$, $i, j \in \{3, 4, 5, 6, 7\}$, are set to 0.8. The applied force f is independent of the yield stresses and is distributed log-normally with mean 1000kN and standard deviation 400kN. The structure will fail if any member exceeds the associated yield stress, i.e., for each member, we can define the following random limit state function:

$$g_i(x; f, \boldsymbol{\sigma}) = \frac{f}{c_i x^i} - \sigma_i, \quad i \in \{1, 2, \ldots, 7\}, \tag{6.2}$$

where the fixed parameters c_i depend on the geometry and the loads. For this structure, $c_{\{i=1,2\}} = 1/\left(2\sqrt{3}\right)$ and $c_{\{i=3,\ldots,7\}} = 1/\sqrt{3}$.

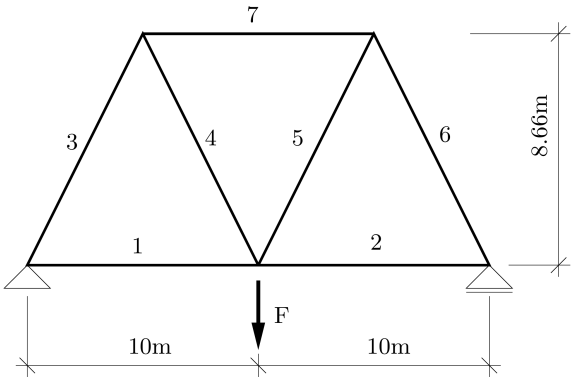

FIGURE 6.2. *Definition of the geometry and the load applied to the truss.*

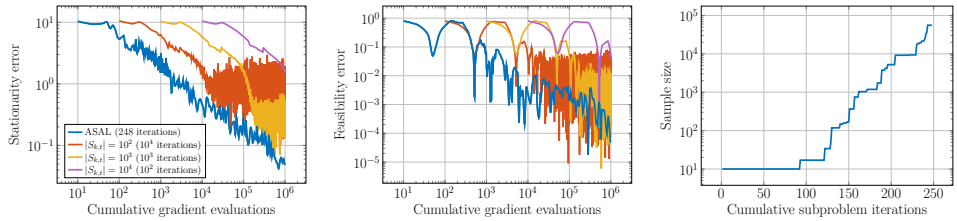

FIGURE 6.3. *Results of running Algorithm 5.1 on the truss optimization problem (6.3). Notice that ASAL achieves the lowest average stationary and feasibility errors while simultaneously requiring the smallest number of iterations. To generate these results, we used the algorithm parameters* $\theta_g = 0.99$, $\alpha = 0.01$, $\eta = 1.0$, *and primal tolerance sequence* $\tau_k = 10/k$.

We pose the following stochastic optimization problem:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{7\alpha}\mathbb{E}\left[\ln\left(\sum_{i=1}^{7}\exp(\alpha g_i(x))\right)\right], \\ &\text{subject to} \quad A \le x \le B, \quad \langle 1, x\rangle \le C, \end{aligned} \tag{6.3}$$

where $\alpha = 1$, $A = 1 \times 10^4\text{mm}^2$, $B = 5 \times 10^4\text{mm}^2$, and $C = 15 \times 10^4\text{mm}^2$ are user-defined parameters. The components of the optimal solution are estimated to be

$$x_{\{i=1,2\}} = 4.342 \times 10^4\text{mm}^2 \quad \text{and} \quad \mathrm{x}_{\{\mathrm{i}=3,\ldots,7\}} = 1.263 \times 10^4\text{mm}^2\,. \tag{6.4}$$

To solve this problem, we use ASAL with $\theta_g = 0.99$ and compare its performance to the stochastic augmented Lagrangian method with fixed sample sizes under a 1 million cumulative sample budget. In each experiment, we use the penalty and step size values $\alpha = 0.01$ and $\eta = 1.0$. Figure 6.3 documents our findings. Notice that, even though it used less than 25% of the total iterations, the stationarity and feasibility errors from ASAL (248 iterations) are significantly lower after the sample budget expires than the best-performing fixed sample size algorithm (1000 iterations).

6.3. Optimal design of a heat sink. We close with a non-convex optimization problem of engineering interest. In this final experiment, we consider the optimal design of a heat sink within a hypothetical square domain $\Omega = (0, 1)^2$ with a stochastic

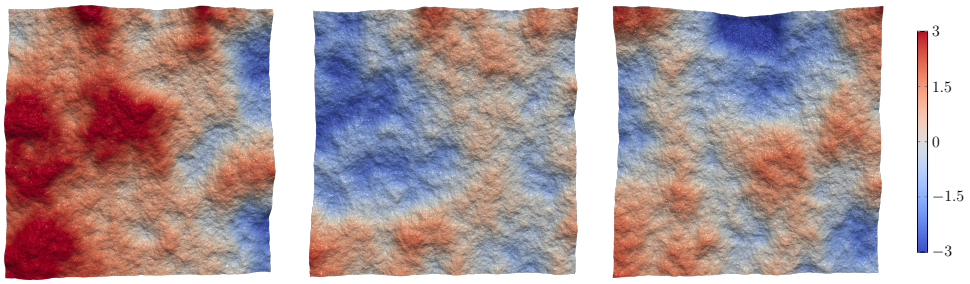

FIGURE 6.4. *Three independent realizations of the Gaussian random field $f(\boldsymbol{x})$ generated by solving (6.5) on the square domain $\Omega = (0,1)^2$.*

heat source $f(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, described by a spatial Gaussian random field with Mátern covariance. More specifically, we follow [57, 58] and define

$$-\kappa^2 \Delta f + f = \mathcal{W} \ \text{ in } \Omega\,, \quad \nabla f \cdot \boldsymbol{n} = 0 \ \text{ on } \partial\Omega\,, \tag{6.5}$$

where $\mathcal{W}$ is spatial additive white Gaussian noise, $\kappa > 0$ is a correlation length parameter, and $\boldsymbol{n}$ denotes the outward-facing unit normal vector field on $\partial\Omega$. Mátern random fields can be used to model various random spatial phenomena [44, 45, 57], which makes them reasonable for modeling the heat source in this example. Figure 6.4 depicts three representative solutions to (6.5) for the reader's interest.

We use the two-field filtered density approach to topology optimization [79, Section 3.1.2] to formulate the optimal heat sink design problem. The goal is to find a material distribution $0 \le \rho \le 1$, where zero indicates no material, and one indicates the complete presence of material, that induces the smallest thermal compliance, $\int_\Omega uf \,\mathrm{d}\boldsymbol{x}$, in expectation. In the aforestated expression, the temperature distribution u is determined by ρ and f through the heat diffusion equation $-\operatorname{div} r(\widetilde{\rho})\nabla u = f$, where $\widetilde{\rho}$ is a regularized (filtered) distribution function [22, 52] and $r(\widetilde{\rho}) > 0$ is a thermal conductivity model. In this work, we use the well-known (modified) solid isotropic material penalization (SIMP) model $r(\widetilde{\rho}) = \rho_- + \widetilde{\rho}^3(1-\rho_-)$, where $0 < \rho_- \ll 1$ is a nominal thermal diffusivity constant assigned to void regions in order to prevent the stiffness matrix from becoming singular [2].

The full problem formulation is written as follows:

$$\min_{\rho \in L^2(\Omega),\, u \in H^1(\Omega)} \ \left\{ \widehat{F}(\rho, u) := \mathbb{E}\left[\int_\Omega uf \,\mathrm{d}\boldsymbol{x} \right] \right\}, \tag{6.6a}$$

subject to the constraints

$$\begin{cases} -\epsilon^2 \Delta \widetilde{\rho} + \widetilde{\rho} = \rho \ \text{ in } \Omega\,, \quad \nabla \widetilde{\rho} \cdot \boldsymbol{n} = 0 \ \text{ on } \partial\Omega\,, \\ -\operatorname{div}\big(r(\widetilde{\rho})\nabla u\big) = f \ \text{ in } \Omega\,, \quad u = 0 \ \text{ on } \Gamma_0\,, \quad \nabla u \cdot \boldsymbol{n} = 0 \ \text{ on } \partial\Omega \setminus \Gamma_0\,, \\ \displaystyle\int_\Omega \rho(\boldsymbol{x}) d\boldsymbol{x} \le \gamma |\Omega|\,, \ \text{ and } \ 0 \le \rho \le 1 \ \text{ in } \Omega\,, \end{cases} \tag{6.6b}$$

where $0 < \gamma < 1$ is the *volume fraction*, which constrains the fraction of the domain occupied by design, and $\epsilon > 0$ is a *length scale* for the final design. The boundary conditions and solution to the optimization problem (6.6) with $\rho_- = 10^{-3}$, $\gamma = 0.5$, $\epsilon = 0.01$, $\kappa = 0.2$ are depicted in Figure 6.5.

To remove the PDE constraints from the optimization problem, we employ a reduced space formulation, often referred to in the literature as a nested formulation [8], which can be written as

$$\min_{\rho \in L^2(\Omega)} \quad \left\{ F(\rho) := \mathbb{E}\left[\int_\Omega u\left(\widetilde{\rho}\left(\rho\right)\right) f \, \mathrm{d}\boldsymbol{x} \right] \right\}, \tag{6.7a}$$

subject to the constraints

$$\int_\Omega \rho(\boldsymbol{x}) d\boldsymbol{x} \leq \gamma |\Omega| \,, \quad \text{and} \quad 0 \leq \rho \leq 1 \quad \text{in } \Omega \,. \tag{6.7b}$$

In this formulation, it is understood that the temperature field $u = u\left(\widetilde{\rho}\left(\rho\right)\right)$ solves the state equation

$$-\operatorname{div}\big(r(\widetilde{\rho})\nabla u\big) = f \quad \text{in } \Omega \,, \quad u = 0 \quad \text{on } \Gamma_0 \,, \quad \nabla u \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega \setminus \Gamma_0 \,, \tag{6.8}$$

and the filtered density $\widetilde{\rho} = \widetilde{\rho}\left(\rho\right)$ solves the screened Poisson equation,

$$-\epsilon^2 \Delta \widetilde{\rho} + \widetilde{\rho} = \rho \quad \text{in } \Omega \,, \quad \nabla \widetilde{\rho} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega \,. \tag{6.9}$$

Since the inequality constraint in (6.7b) is always active, it is replaced by an equality constraint that our ASAL algorithm can handle. The gradients of the reduced objective function in (6.7a) are computed with FEM-discretized representations of the temperature u and filtered density $\widetilde{\rho}$ using standard adjoint analysis techniques [8]. Finally, $L^2(\Omega)$ projections are used to enforce the box constraints found in (6.7b).

For comparison, Figure 6.5 also depicts a reference solution to (6.6) corresponding to the (deterministic) uniform heat field $f \equiv 1$. Close examination reveals significant differences between the designs with deterministic and stochastic inputs. The deterministic case results in an organic tree-like structure that aims to transfer the heat generated at any point in the computational domain using the shortest possible way to the Dirichlet boundary with zero temperature. The design does not depend on the magnitude of the heat source, and any constant input will result in the same material distribution if the initial material distribution is in the vicinity of the local solution. On the other hand, due to the oscillatory nature of the stochastic input, the heat source term can take positive and negative values. Such input distribution allows the optimization process to balance the heat transfer locally without linking the local subdomain directly to the boundary with a fixed temperature. Thus, the role of the closed loops of material appearing in the design with stochastic input is to establish a local heat equilibrium. In this case, the global tree-like structure transfers only the excess heat, which cannot be balanced locally.

In this experiment, we use ASAL with $\theta_g = 2$ and compare its performance to stochastic augmented Lagrangian with fixed sample sizes, $|S_{k,t}| = 10^j$, $j = 1, 2, 3$, under a 10^5 cumulative sample budget. In each execution, we use the step size values $\alpha = 0.1$ and $\eta = 2.0$. Figure 6.6 documents our findings. ASAL achieves the lowest combined average stationary and feasibility errors while requiring less than 20% of the iterations of the best-performing fixed sample size run ($|S_{k,t}| = 10^2$). Although the average feasibility errors with ASAL and this fixed sample size run are similar, the variance of the fixed sample size run is much greater. Finally, the average stationarity error for the best-performing fixed sample size run is significantly larger than the average stationarity error with ASAL.

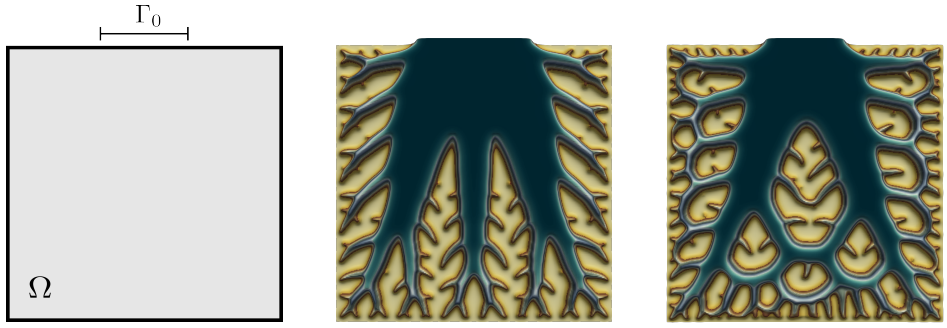

FIGURE 6.5. *Left: Depiction of the subsets of the domain boundary $\partial\Omega$ for the boundary conditions for the heat field u in (6.5). We define $u = 0$ on Γ_0 and $\nabla u \cdot \boldsymbol{n} = 0$ on $\partial\Omega \setminus \Gamma_0$. Middle: Reference density field $\tilde{\rho}$ for the solution of the* deterministic *thermal compliance optimization problem (6.6) with $f \equiv 1$ everywhere in Ω. Right: The filtered density $\widetilde{\rho}$ for the solution of the* expected value *thermal compliance optimization problem (6.6) with f given by (6.5). The presence of closed-loop branches in the optimal solution on the right indicates a preference for balancing the heat locally and transferring only the excess unbalanced heat to the external environment through Γ_0.*

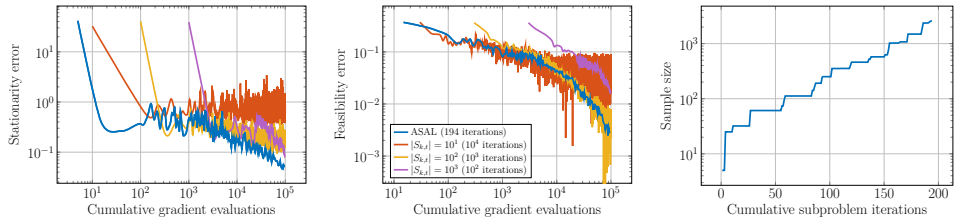

FIGURE 6.6. *Results of running ASAL (Algorithm 5.1) on the topology optimization problem (6.6). Notice that ASAL achieves the lowest combined average stationary and feasibility errors while simultaneously requiring the smallest number of iterations. Indeed, the average stationarity error for the best-performing fixed sample size run, $|S_{k,t}| = 10^j$, $j = 1, 2, 3$, ends up larger than the average stationarity error with ASAL. Moreover, although the average feasibility errors of ASAL and the best-performing fixed sample size run are similar, the variance for the fixed sample size run is much greater. To generate these results, we used the algorithm parameters $\theta_g = 2$, $\alpha = 0.1$, $\eta = 2.0$, and primal tolerance sequence $\tau_k = 1/k$.*

7. Final Remarks. Motivated by a growing interest in developing optimization algorithms for constrained stochastic optimization problems, we introduced a framework that combines augmented Lagrangian methods with adaptive sampling techniques. In our framework, we employed stochastic solvers for the subproblems and imposed stochastic tolerance criteria for the inexact solutions. We analyzed various theoretical tolerance conditions and designed a practical test. To establish convergence results, we first showed that our framework is equivalent to an inexact gradient descent algorithm on the Moreau envelope. Second, we showed sublinear convergence in the outer iterations when f is convex and linear convergence when f is strongly convex with $\mathcal{X} = \mathbb{R}^n$. We also analyzed the worst-case expected work complexity of our approach in terms of the number of gradient evaluations required to obtain an ϵ-accurate solution. For convex f and compact $\mathcal{X}$, we showed $\mathcal{O}(\epsilon^{-3-\delta})$ complexity where $\delta > 0$ is a user-defined parameter. This result improves to $\mathcal{O}(\epsilon^{-2})$ when the penalty parameter $\alpha = \mathcal{O}(\epsilon^{-1})$. If f is strongly convex and $\mathcal{X} = \mathbb{R}^n$, we proved $\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$ complexity.

To evaluate our framework's practical performance, we tested it on a constrained machine learning problem and in engineering applications. Here, we observed that our

method minimizes the objective function more efficiently and reaches a feasible solution in a more stable manner than benchmark stochastic approximation algorithms.

Although our analysis holds for any penalty parameter $\alpha > 0$, this parameter should be tuned for optimal performance in practice. The other main hyperparameters are the step size $\eta > 0$ for the inner problems, and the subproblem tolerance values $\tau_k > 0$. Since tuning is computationally expensive, it would be helpful to develop methods that adaptively select these hyper-parameters in order to further improve the practical efficacy of our adaptive sampling framework. Two other natural extensions would be generalizing our methods to include nonlinear constraints and chance constraints.

Appendix A. Proof of Proposition 4.7.

Proof. Due to the strong convexity of f, (2.12) has a unique optimal solution, denoted as $x(\lambda)$. Using [40, Corollary 4.5.3], we can show that $q(\lambda)$ is differentiable and

$$\nabla q(\lambda) = Ax(\lambda) - b. \tag{A.1}$$

Moreover, from the optimality conditions of (2.12), we have

$$\nabla f(x(\lambda)) - A^T\lambda = 0. \tag{A.2}$$

Let $\lambda_1, \lambda_2 \in \mathbb{R}^m$. Consider,

$$\begin{aligned}
\langle \nabla q(\lambda_2) - \nabla q(\lambda_1), \lambda_2 - \lambda_1 \rangle &= \langle A(x(\lambda_2) - x(\lambda_1)), \lambda_2 - \lambda_1 \rangle \\
&= \langle x(\lambda_2) - x(\lambda_1), A^T(\lambda_2 - \lambda_1) \rangle \\
&= \langle x(\lambda_2) - x(\lambda_1), \nabla f(x(\lambda_2)) - \nabla f(x(\lambda_1)) \rangle \\
&\geq \frac{1}{\mu + L} \|\nabla f(x(\lambda_2)) - \nabla f(x(\lambda_1))\|^2 \\
&= \frac{1}{\mu + L} \|A^T(\lambda_2 - \lambda_1)\|^2 \\
&\geq \frac{\lambda_{\min}(AA^T)}{\mu + L} \|\lambda_2 - \lambda_1\|^2 ,
\end{aligned}$$

where the first equality is due to (A.1), the second and the third equalities are due to (A.2), and the first inequality is due to [63, Theorem 2.1.11]. Therefore, using [63, Theorem 2.1.9], we can claim that $q(\lambda)$ is strongly convex with parameter $\frac{\sigma}{L+\mu}$. ∎

Appendix B. Logistic regression with multiple disparate impact constraints, `australian` dataset.

We consider problem (6.1) with `australian` classification data set from the LIBSVM collection [26]. The data set has $N = 690$ rows, and the dimension of the problem is $n = 14$. Considering the budget of cumulative gradient evaluations as $200N$, and the fixed hyperparameters as $\theta_g = 0.99$, $\nu_1 = 0.5$, $s_1 = 0.1N, s_{\min} = 0.1N$, we compare three separately-tuned fixed-batch-size implementations of ASAL using 10%, 20%, and 50% of the data set size. We tune τ_0, α and the step size η using the same procedure described in Subsection 6.1 with the sets of $\tau_0 = 10^{i-1}$, $\eta = 10^{j-5}$, and $\alpha = 10^{j-4}$, where $i = 0, 1, 2, 3, 4, 5$ and $j = 0, 1, 2, 3, 4, 5, 6$.

For each algorithm, we select the run with the smallest average objective function value in the final 10 inner iterations among all runs whose minimum feasibility error in the final 50 inner iterations is less than the feasibility tolerance 10^{-3}. These

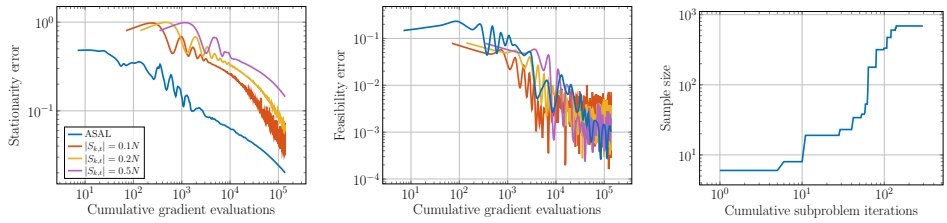

FIGURE B.1. *Results of running Algorithm 5.1 on the constrained logistic regression problem (6.1) with the* `australian` *classification data set. Notice that ASAL achieves the lowest average stationarity errors while matching the minimal feasibility error of the three baseline algorithms. To generate these results, we used the algorithm parameters* $\theta_g = 0.99$ *and individually tuned* α, η, *and* τ_0.

values (i.e., 10, 50, and 10^{-3}, respectively) are slightly different than the values given in Subsection 6.1 to ensure that the best combinations of hyperparameter values correspond to a more stable set of runs. Because of the same reason, we restrict $\alpha = 10^{-1}$ for the ASAL algorithm while tuning, as we observe this value results in choosing runs that show a good balance between stationarity and feasibility errors. The comparison of the algorithms is given in Figure B.1. Similar to Subsection 6.1, we observe that ASAL and each of the three baseline algorithms achieve a similar *minimal* feasibility error (around feasibility tolerance 10^{-3}) and that ASAL performs better than the three baseline algorithms when it comes to stationarity error.

REFERENCES

[1] *Stochastic optimization using a trust-region method and random models*, Mathematical Programming, 169 (2018), pp. 447–487.

[2] E. ANDREASSEN, A. CLAUSEN, M. SCHEVENELS, B. S. LAZAROV, AND O. SIGMUND, *Efficient topology optimization in MATLAB using 88 lines of code*, Structural and Multidisciplinary Optimization, 43 (2011), pp. 1–16.

[3] E. ANDREASSEN, B. S. LAZAROV, AND O. SIGMUND, *Design of manufacturable 3D extremal elastic microstructure*, Mechanics of Materials, 69 (2014), pp. 1 – 10, https://doi.org/http://dx.doi.org/10.1016/j.mechmat.2013.09.018, http://www.sciencedirect.com/science/article/pii/S0167663613002093.

[4] S. BAROCAS AND A. D. SELBST, *Big data's disparate impact*, Calif. L. Rev., 104 (2016), p. 671.

[5] F. BASTIN, C. CIRILLO, AND P. L. TOINT, *An adaptive monte carlo algorithm for computing mixed logit estimators*, Computational Management Science, 3 (2006), pp. 55–79.

[6] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2nd ed., 2017.

[7] F. BEISER, B. KEITH, S. URBAINCZYK, AND B. WOHLMUTH, *Adaptive sampling strategies for risk-averse stochastic optimization with constraints*, IMA Journal of Numerical Analysis, (2023), https://doi.org/10.1093/imanum/drac083. drac083.

[8] M. P. BENDSOE AND O. SIGMUND, *Topology optimization: theory, methods, and applications*, Springer Science & Business Media, 2003.

[9] 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).

[10] 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.

[11] D. BERTSEKAS, *Convex optimization algorithms*, Athena Scientific, 2015.

[12] D. BERTSEKAS, A. NEDIC, AND A. OZDAGLAR, *Convex analysis and optimization*, vol. 1, Athena Scientific, 2003.

[13] D. P. BERTSEKAS, *Constrained optimization and Lagrange multiplier methods*, Academic press, 2014.

[14] E. G. Birgin, *Augmented lagrangian method with nonmonotone penalty parameters for constrained optimization*, Computational Optimization and Applications, 51 (2012), pp. 941–965.
[15] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, *Convergence rate analysis of a stochastic trust-region method via supermartingales*, INFORMS journal on optimization, 1 (2019), pp. 92–119.
[16] R. Bollapragada, R. Byrd, and J. Nocedal, *Adaptive sampling strategies for stochastic optimization*, SIAM Journal on Optimization, 28 (2018), pp. 3312–3343.
[17] R. Bollapragada, R. H. Byrd, and J. Nocedal, *Exact and inexact subsampled newton methods for optimization*, IMA Journal of Numerical Analysis, 39 (2019), pp. 545–578.
[18] R. Bollapragada, J. Nocedal, D. Mudigere, H.-J. Shi, and P. T. P. Tang, *A progressive batching l-bfgs method for machine learning*, in International Conference on Machine Learning, PMLR, 2018, pp. 620–629.
[19] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM review, 60 (2018), pp. 223–311.
[20] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine learning, 3 (2011), pp. 1–122.
[22] T. E. Bruns and D. A. Tortorelli, *Topology optimization of non-linear elastic structures and compliant mechanisms*, Computer methods in applied mechanics and engineering, 190 (2001), pp. 3443–3459.
[23] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, *Sample size selection in optimization methods for machine learning*, Mathematical programming, 134 (2012), pp. 127–155.
[24] A. Carlon, L. Espath, R. Lopez, and R. Tempone, *Multi-iteration stochastic optimizers*, arXiv preprint arXiv:2011.01718, (2020).
[25] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Mathematical Programming, 169 (2018), pp. 337–375.
[26] C.-C. Chang and C.-J. Lin, *Libsvm: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011), https://doi.org/10.1145/1961189.1961199, https://doi.org/10.1145/1961189.1961199.
[27] S. Chen, W. Chen, and S. Lee, *Level set based robust shape and topology optimization under random field uncertainties*, Structural and Multidisciplinary Optimization, 41 (2010), pp. 507–524, https://doi.org/10.1007/s00158-009-0449-2, http://dx.doi.org/10.1007/s00158-009-0449-2.
[28] G. Cornuejols and R. Tütüncü, *Optimization methods in finance*, vol. 5, Cambridge University Press, 2006.
[29] F. E. Curtis, N. I. Gould, H. Jiang, and D. P. Robinson, *Adaptive augmented lagrangian methods: algorithms and practical numerical experience*, Optimization Methods and Software, 31 (2016), pp. 157–186.
[30] F. E. Curtis, H. Jiang, and D. P. Robinson, *An adaptive augmented lagrangian method for large-scale constrained optimization*, Mathematical Programming, 152 (2015), pp. 201–245.
[31] F. E. Curtis, M. J. O'Neill, and D. P. Robinson, *Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization*, arXiv preprint arXiv:2112.14799, (2021).
[32] F. E. Curtis, D. P. Robinson, and B. Zhou, *Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints*, arXiv preprint arXiv:2107.03512, (2021).
[33] J. Eckstein and W. Yao, *Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results*, RUTCOR Research Reports, 32 (2012), p. 44.
[34] L. Espath, S. Krumscheid, R. Tempone, and P. Vilanova, *On the equivalence of different adaptive batch size selection strategies for stochastic gradient descent methods*, 2021, https://arxiv.org/abs/2109.10933.
[35] M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1380–A1405.
[36] S. Ganesh and F. Nobile, *Gradient-based optimisation of the conditional-value-at-risk using the multi-level Monte Carlo method*, arXiv preprint arXiv:2210.03485, (2022).
[37] V. Guigues, *On the strong concavity of the dual function of an optimization problem*, arXiv preprint arXiv:2006.16781, (2020).

[38] V. Guigues, *Inexact stochastic mirror descent for two-stage nonlinear stochastic programs*, Mathematical Programming, 187 (2021), pp. 533–577.
[39] M. R. Hestenes, *Multiplier and gradient methods*, Journal of optimization theory and applications, 4 (1969), pp. 303–320.
[40] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*, vol. 305, Springer science & business media, 2013.
[41] M. Jansen, B. Lazarov, M. Schevenels, and O. Sigmund, *On the similarities between micro/nano lithography and topology optimization projection methods*, Structural and Multidisciplinary Optimization, 48 (2013), pp. 717–730, https://doi.org/10.1007/s00158-013-0941-6, http://dx.doi.org/10.1007/s00158-013-0941-6.
[42] 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.
[43] M. Kang, M. Kang, and M. Jung, *Inexact accelerated augmented lagrangian methods*, Computational Optimization and Applications, 62 (2015), pp. 373–404.
[44] B. Keith, U. Khristenko, and B. Wohlmuth, *A fractional PDE model for turbulent velocity fields near solid walls*, Journal of Fluid Mechanics, 916 (2021), p. A21.
[45] U. Khristenko, A. Constantinescu, P. L. Tallec, J. T. Oden, and B. Wohlmuth, *A statistical framework for generating microstructures of two-phase random materials: application to fatigue analysis*, Multiscale Modeling & Simulation, 18 (2020), pp. 21–43.
[46] U. Khristenko, L. Scarabosio, P. Swierczynski, E. Ullmann, and B. Wohlmuth, *Analysis of boundary effects on PDE-based sampling of Whittle–Matérn random fields*, SIAM/ASA Journal on Uncertainty Quantification, 7 (2019), pp. 948–974.
[47] A. Kodakkal, B. Keith, U. Khristenko, A. Apostolatos, K.-U. Bletzinger, B. Wohlmuth, and R. Wüchner, *Risk-averse design of tall buildings for uncertain wind conditions*, Computer Methods in Applied Mechanics and Engineering, 402 (2022), p. 115371.
[48] D. P. Kouri and A. Shapiro, *Optimization of PDEs with uncertain inputs*, Frontiers in PDE-Constrained Optimization, (2018), pp. 41–81.
[49] D. P. Kouri and T. M. Surowiec, *A primal-dual algorithm for risk minimization*, Mathematical Programming, 193 (2022), pp. 337–363.
[50] G. Lan and R. D. Monteiro, *Iteration-complexity of first-order augmented lagrangian methods for convex programming*, Mathematical Programming, 155 (2016), pp. 511–547.
[51] G. Lan and Z. Zhou, *Algorithms for stochastic optimization with function or expectation constraints*, Computational Optimization and Applications, 76 (2020), pp. 461–498.
[52] B. S. Lazarov and O. Sigmund, *Filters in topology optimization based on Helmholtz-type differential equations*, International Journal for Numerical Methods in Engineering, 86 (2011), pp. 765–781.
[53] B. S. Lazarov, F. Wang, and O. Sigmund, *Length scale and manufacturability in density-based topology optimization*, Archive of Applied Mechanics, 86 (2016), pp. 189–218, https://doi.org/10.1007/s00419-015-1106-4, http://dx.doi.org/10.1007/s00419-015-1106-4.
[54] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, *Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2170–2178.
[55] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, *Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization*, arXiv preprint arXiv:2212.09513, (2022).
[56] Z. Li and Y. Xu, *Augmented lagrangian–based first-order methods for convex-constrained programs with weakly convex objective*, INFORMS Journal on Optimization, 3 (2021), pp. 373–397.
[57] F. Lindgren, D. Bolin, and H. Rue, *The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running*, Spatial Statistics, (2022), p. 100599.
[58] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 423–498.
[59] Y. Liu, F. Shang, H. Liu, L. Kong, L. Jiao, and Z. Lin, *Accelerated variance reduction stochastic admm for large-scale machine learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (2021), pp. 4242–4255, https://doi.org/10.1109/TPAMI.2020.3000512.
[60] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France, 93 (1965), pp. 273–299.
[61] S. Na, M. Anitescu, and M. Kolar, *An adaptive stochastic sequential quadratic program-*

ming with differentiable exact augmented lagrangians, Mathematical Programming, (2022), pp. 1–71.
[62] S. Na, M. Anitescu, and M. Kolar, *Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming*, Mathematical Programming, (2023), pp. 1–75.
[63] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2003.
[64] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018.
[65] H. Ouyang, N. He, L. Tran, and A. Gray, *Stochastic alternating direction method of multipliers*, in International conference on machine learning, PMLR, 2013, pp. 80–88.
[66] N. Parikh, S. Boyd, et al., *Proximal algorithms*, Foundations and trends® in Optimization, 1 (2014), pp. 127–239.
[67] R. Pasupathy, P. Glynn, S. Ghosh, and F. S. Hashemi, *On sampling rates in simulation-based recursions*, SIAM Journal on Optimization, 28 (2018), pp. 45–73.
[68] C. Phelps, J. O. Royset, and Q. Gong, *Optimal control of uncertain systems using sample average approximations*, SIAM Journal on Control and Optimization, 54 (2016), pp. 1–29.
[69] C. Planiden and X. Wang, *Strongly convex functions, moreau envelopes, and the generic nature of convex functions with strong minimizers*, SIAM Journal on Optimization, 26 (2016), pp. 1341–1364.
[70] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, Optimization, (1969), pp. 283–298.
[71] R. T. Rockafellar, *Augmented lagrange multiplier functions and duality in nonconvex programming*, SIAM Journal on Control, 12 (1974), pp. 268–285.
[72] R. T. Rockafellar, *Augmented lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of operations research, 1 (1976), pp. 97–116.
[73] R. T. Rockafellar and J. O. Royset, *On buffered failure probability in design and optimization of structures*, Reliability Engineering & System Safety, 95 (2010), pp. 499 – 510, https://doi.org/http://dx.doi.org/10.1016/j.ress.2010.01.001.
[74] F. Roosta-Khorasani and M. W. Mahoney, *Sub-sampled newton methods*, Mathematical Programming, 174 (2019), pp. 293–326.
[75] J. O. Royset, *Risk-adaptive approaches to learning and decision making: A survey*, 2022, https://doi.org/10.48550/ARXIV.2212.00856, https://arxiv.org/abs/2212.00856.
[76] J. O. Royset and R. Szechtman, *Optimal budget allocation for sample average approximation*, Operations Research, 61 (2013), pp. 762–776.
[77] M. F. Sahin, A. Alacaoglu, F. Latorre, V. Cevher, et al., *An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints*, Advances in Neural Information Processing Systems, 32 (2019).
[78] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: Modeling and theory*, SIAM, 2021.
[79] O. Sigmund and K. Maute, *Topology optimization approaches: A comparative review*, Structural and Multidisciplinary Optimization, 48 (2013), pp. 1031–1055.
[80] M. Sion, *On general minimax theorems.*, Pacific Journal of mathematics, 8 (1958), pp. 171–176.
[81] T. Suzuki, *Stochastic dual coordinate ascent with alternating direction method of multipliers*, in International Conference on Machine Learning, PMLR, 2014, pp. 736–744.
[82] S. Wright and B. Recht, *Optimization for Data Analysis*, Cambridge University Press, 2022.
[83] Y. Xie, R. Bollapragada, R. Byrd, and J. Nocedal, *Constrained and composite optimization via adaptive sampling methods*, IMA Journal of Numerical Analysis, (2023), p. drad020.
[84] Y. Xie and U. V. Shanbhag, *Si-admm: A stochastic inexact admm framework for stochastic convex programs*, IEEE Transactions on Automatic Control, 65 (2019), pp. 2355–2370.
[85] Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM Journal on Optimization, 30 (2020), pp. 1664–1692.
[86] Y. Xu, *Iteration complexity of inexact augmented lagrangian methods for constrained convex programming*, Mathematical Programming, 185 (2021), pp. 199–244.
[87] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, *Fairness constraints: A flexible approach for fair classification*, J. Mach. Learn. Res., 20 (2019), pp. 1–42.
[88] S. Zheng and J. T. Kwok, *Fast-and-light stochastic ADMM*, in International Joint Conference on Artificial Intelligence, 2016.
[89] W. Zhong and J. T. Kwok, *Fast stochastic alternating direction method of multipliers*, in International conference on machine learning, PMLR, 2014, pp. 46–54.
[90] X. Zhou, *On the fenchel duality between strong convexity and lipschitz continuous gradient*, 2018, https://arxiv.org/abs/1803.06573.