An inexact stochastic variance-reduced gradient method for constrained convex finite-sum optimization

Zhaosong Lu * Yifeng Xiao *

August 5, 2025 (Revised: October 7, 2026)

arXiv:2611.23221v1 [math.OC] 17 Nov 2026

Abstract

In this paper, we study a class of constrained convex finite-sum optimization problems. In many practical applications, all or some of the constraints are hard constraints that represent imperative requirements. Consequently, any desirable approximate solution must (nearly) satisfy these constraints. Motivated by this, we focus on computing an ϵ-*surely feasible stochastic optimal* (ϵ-SFSO) solution, in which the constraint violation is deterministically bounded by ϵ, and the expected optimality gap is at most ϵ. While a recently proposed stochastic variance-reduced first-order penalty method can compute an ϵ-SFSO solution, there remains a significant gap between its first-order oracle complexity for computing an ϵ-SFSO solution and that for computing an ϵ-*expected feasible stochastic optimal* solution, where both the expected constraint violation and the expected optimality gap are bounded by ϵ. To narrow this gap, we propose an inexact stochastic variance-reduced gradient (SVRG) method that significantly improves the previously best-known first-order oracle complexity for computing an ϵ-SFSO solution. Furthermore, when the objective consists of a single function, the proposed method reduces to a novel deterministic first-order method for solving standard convex conic optimization. It achieves a significantly lower gradient evaluation complexity with respect to the objective function compared to existing methods, and can outperform them when computing constraint gradients and projections onto the associated cone is cheaper than evaluating the objective gradient. Preliminary numerical results are presented to illustrate the performance of the proposed method.

Keywords: convex conic optimization, finite-sum optimization, accelerated gradient method, variance reduction

Mathematics Subject Classification: 90C25, 90C35, 90C30, 65K05

1 Introduction

In this paper, we consider constrained convex finite-sum optimization problems of the form

$$\begin{aligned} F^* = \min_x \ & \Big\{ F(x) := \frac{1}{s}\sum_{i=1}^{s} f_i(x) + \psi(x) \Big\} \\ \text{s.t. } & c(x) \in -\mathcal{K}, \end{aligned} \tag{1}$$

where $\mathcal{K} \subseteq \mathbb{R}^m$ is a closed convex cone, f_i is a continuously differentiable convex function on $\mathbb{R}^n$ for all $i = 1, 2, \ldots, s$, $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function with an exactly evaluable proximal operator and domain denoted by X, and $c = (c_1, \ldots, c_m)$ is a Lipschitz smooth and $\mathcal{K}$-convex mapping on X, that is,

$$\alpha c(x) + (1-\alpha)c(y) - c(\alpha x + (1-\alpha)y) \in \mathcal{K} \qquad \forall x, y \in X, \alpha \in [0,1]. \tag{2}$$

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, xiao0414@umn.edu). This work was partially supported by the Office of Naval Research under Award N00014-24-1-2702, the Air Force Office of Scientific Research under Award FA9550-24-1-0343, and the National Science Foundation under Awards 2211491 and 2435911.

Problem (1) includes a broad class of constrained optimization problems as special cases. For instance, when $\mathcal{K} = \mathbb{R}^{m_1}_+ \times \{0\}^{m_2}$ for some $m_1, m_2 \in \mathbb{N}$, and $c = (c_1, \ldots, c_{m_1+m_2})$ with c_j convex for $1 \le j \le m_1$ and affine for $m_1 < j \le m_1 + m_2$, problem (1) reduces to the standard constrained convex finite-sum optimization problem:

$$\min_x \left\{ \frac{1}{s}\sum_{i=1}^{s} f_i(x) + \psi(x) : \ c_j(x) \le 0, \ 1 \le j \le m_1; \ c_j(x) = 0, \ m_1 < j \le m_1 + m_2 \right\}.$$

Moreover, when $c = 0$, problem (1) reduces to the unconstrained finite-sum optimization problem:

$$\min_x \Big\{ \frac{1}{s}\sum_{i=1}^{s} f_i(x) + \psi(x) \Big\}. \tag{3}$$

This class of problems has been extensively studied in the literature (see, e.g., [1, 3, 5, 8, 17, 18, 19, 22, 24]). Notably, several stochastic gradient methods (e.g., [3, 17, 18, 22]) employ variance reduction techniques and achieve $\widetilde{\mathcal{O}}(s\log s + s^{1/2}\epsilon^{-1})$ complexity in terms of the number of ∇f_i evaluations required to compute an ϵ-*stochastic optimal solution*, that is, a point x satisfying $\mathbb{E}[F(x) - F^*] \le \epsilon$.[1] More recently, optimal incremental gradient methods [9] and accelerated stochastic variance-reduced gradient (SVRG) methods [1, 8] have been developed, achieving $\widetilde{\mathcal{O}}(s\log s + s^{1/2}\epsilon^{-1/2})$ or $\mathcal{O}(s\log s + s^{1/2}\epsilon^{-1/2})$ complexity for computing such solutions.

In addition, when $c \neq 0$, $\mathcal{K}$ is the nonnegative orthant, and ψ is the indicator function of a simple convex compact set, a level-set method was proposed in [12] for solving problem (1). This method requires $\mathcal{O}\big(sn + (s+n)\epsilon^{-2}\big)$ evaluations of ∇f_i and ∇c to compute an ϵ-*expected feasible stochastic optimal* (ϵ-EFSO) solution of (1), that is, a stochastic point x satisfying

$$\mathbb{E}[\mathrm{dist}(c(x), -\mathcal{K})] \le \epsilon, \quad \mathbb{E}[F(x) - F^*] \le \epsilon,$$

where $\mathrm{dist}(\cdot,\cdot)$ denotes the Euclidean distance between a point and a set.

In many applications, all or some of the constraints in problem (1) are hard constraints that represent imperative requirements. Consequently, any desirable approximate solution must (nearly) satisfy these constraints. Recently, for the case where $\mathcal{K}$ is the nonnegative orthant, a stochastic variance-reduced first-order penalty (SVRFP) method was proposed in [13] for computing an ϵ-*surely feasible stochastic optimal* (ϵ-SFSO) solution to problem (1), that is, a stochastic point x satisfying

$$\mathrm{dist}\big(c(x), -\mathcal{K}\big) \le \epsilon, \quad \mathbb{E}[F(x) - F^*] \le \epsilon,$$

which is generally a stronger notion than an ϵ-EFSO solution. The SVRFP method applies *only one iteration* of a modified variance-reduced accelerated gradient scheme to a sequence of associated quadratic penalty subproblems:

$$\min_x \Big\{ \frac{1}{s}\sum_{i=1}^{s} f_i(x) + \frac{\rho_k}{2}\mathrm{dist}^2\big(c(x), -\mathcal{K}\big) + \psi(x) \Big\} \tag{4}$$

for a suitably chosen sequence of penalty parameters $\{\rho_k\}$. With an appropriate choice of $\{\rho_k\}$, this method can compute an ϵ-SFSO solution using $\mathcal{O}(s\log s + s^{1/2}\epsilon^{-3/2})$ evaluations of ∇f_i and ∇c, respectively. Moreover, with a suitable choice of $\{\rho_k\}$, it can compute an ϵ-EFSO solution of (1) using $\widetilde{\mathcal{O}}(s\log s + s^{1/2}\epsilon^{-1})$ evaluations of ∇f_i and ∇c, respectively. These are the best-known first-order oracle complexities in the literature for computing an ϵ-SFSO solution and an ϵ-EFSO solution of problem (1), and they are achieved by the SVRFP method. However, there is a significant gap between these two complexity bounds, mainly due to the following two facts established in [13]:

- The solution sequence $\{x_k\}$ generated by the SVRFP method satisfies

$$\mathrm{dist}\big(c(x_k), -\mathcal{K}\big) = \mathcal{O}(\rho_k^{-1/2}), \qquad \mathbb{E}[\mathrm{dist}(c(x_k), -\mathcal{K})] = \mathcal{O}\Big(\rho_k^{-1/2}\sqrt{\mathbb{E}[F_{\rho_k}(x_k) - F^*_{\rho_k}]}\Big),$$

  where $F_{\rho_k}(\cdot)$ and $F^*_{\rho_k}$ denote the objective function and the optimal value of problem (4), respectively.

[1] $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors of the tolerance parameter ϵ in $\mathcal{O}(\cdot)$.

- The sequence $\{\mathbb{E}[F_{\rho_k}(x_k) - F^*_{\rho_k}]\}$ converges to zero, at a rate depending on the choice of $\{\rho_k\}$.

It follows that $\{\mathrm{dist}\big(c(x_k), -\mathcal{K}\big)\}$ converges to zero at a slower rate than $\{\mathbb{E}[\mathrm{dist}(c(x_k), -\mathcal{K})]\}$. As a result, the SVRFP method has a worse complexity bound for finding an ϵ-SFSO solution than for finding an ϵ-EFSO solution of problem (1).

As discussed above, the SVRFP method is essentially an "infeasible" (i.e., exterior) optimization method. It drives the iterates to gradually satisfy the constraints at a relatively slow rate. This behavior leads to the aforementioned significant gap in first-order oracle complexity between finding an ϵ-SFSO solution and an ϵ-EFSO solution of (1). Motivated by this observation, we aim to narrow this gap by proposing a "feasible" method—an inexact SVRG method—in which all iterates nearly satisfy the constraints.

Specifically, observe that problem (1) can be reformulated as

$$\min_x \big\{ f(x) + \psi(x) + \delta_S(x) \big\}, \tag{5}$$

where $S = \{x \in \mathbb{R}^n : c(x) \in -\mathcal{K}\}$, and $\delta_S(\cdot)$ denotes the indicator function of the set S. While problem (5) is structurally similar to (3), the accelerated (exact) SVRG method in [8, Algorithm 1] is not directly applicable, as its development and analysis require exact evaluation of the proximal operator of $\gamma[\psi(\cdot) + \delta_S(\cdot)]$ for $\gamma > 0$. However, this proximal operator cannot be computed exactly in general. Indeed, computing it amounts to solving the constrained optimization problem

$$\begin{aligned} \min_x \ & \Big\{ \psi(x) + \frac{1}{2\gamma}\|x - u\|^2 \Big\} \\ \text{s.t. } & c(x) \in -\mathcal{K} \end{aligned} \tag{6}$$

for any given $u \in \mathbb{R}^n$. Since the constraint set in (6) is typically non-simple, computing its exact optimal solution is challenging. To address this issue, we propose an inexact SVRG method (Algorithm 2) for solving (5) (or equivalently, (1)). Our approach follows a similar framework to [8, Algorithm 1], but requires only an *inexact* solution to problem (6)—particularly an ϵ-Karush-Kuhn-Tucker (ϵ-KKT) solution (x, λ) satisfying

$$\mathrm{dist}\big(0, (x-u)/\gamma + \nabla c(x)\lambda + \partial\psi(x)\big) \le \epsilon, \qquad \mathrm{dist}\big(c(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)\big) \le \epsilon \tag{7}$$

for some $\epsilon > 0$. Such an ϵ-KKT solution can be computed efficiently using a first-order proximal augmented Lagrangian (AL) method (see Algorithm 1). We show that the proposed inexact SVRG method can find an ϵ-SFSO solution of (1) using

$$\mathcal{O}\big(s\log s + s^{1/2}\epsilon^{-1/2}\big) \quad \text{and} \quad \widetilde{\mathcal{O}}\big(s(\log s)^{3/2}\epsilon^{-1/2} + s^{1/4}\epsilon^{-5/4}\big)$$

evaluations of ∇f_i and ∇c, respectively. These results significantly improve upon the best-known first-order oracle complexity for computing an ϵ-SFSO solution achieved by [13, Algorithm 2], and substantially reduce the gap between the complexities for obtaining ϵ-SFSO and ϵ-EFSO solutions of (1).

The main contributions of this paper are summarized as follows.

- We propose an inexact SVRG method (Algorithm 2) for solving problem (1). This method computes an ϵ-SFSO solution—a notion stronger than the commonly studied ϵ-EFSO solution—using $\mathcal{O}\big(s\log s + s^{1/2}\epsilon^{-1/2}\big)$ and $\widetilde{\mathcal{O}}\big(s(\log s)^{3/2}\epsilon^{-1/2} + s^{1/4}\epsilon^{-5/4}\big)$ evaluations of ∇f_i and ∇c, respectively. These complexity bounds significantly improve the previously best-known complexity bound $\mathcal{O}\big(s\log s + s^{1/2}\epsilon^{-3/2}\big)$ established in [13].
- In the special case where $s = 1$, the proposed method reduces to a novel deterministic first-order method and can compute an ϵ-optimal solution of (1) using $\mathcal{O}(\epsilon^{-1/2})$ evaluations of ∇f_1 and $\widetilde{\mathcal{O}}(\epsilon^{-5/4})$ evaluations of ∇c and projections onto $-\mathcal{K}$. This yields a significant improvement in the gradient evaluation complexity with respect to f_1 over the best-known first-order methods (e.g., [6, 14]), which require $\mathcal{O}(\epsilon^{-1})$ or $\widetilde{\mathcal{O}}(\epsilon^{-1})$ evaluations of ∇f_1, ∇c, and projections onto $-\mathcal{K}$. Consequently, when evaluating ∇c and projecting onto $-\mathcal{K}$ are cheaper than evaluating ∇f_1, the proposed method can achieve superior practical performance.

The rest of the paper is organized as follows. Subsection 1.1 introduces the notation and assumptions used throughout the paper. Section 2 presents a first-order proximal AL method for solving constrained strongly convex optimization problems, along with its convergence results. In Section 3, we propose an inexact SVRG method for solving problem (1) and establish its convergence guarantees. Preliminary numerical results are presented in Section 4 to illustrate the performance of the proposed method. The proofs of the main theoretical results are provided in Section 5. Finally, concluding remarks are given in Section 6.

1.1 Notation and assumptions

The following notation will be used throughout this paper. Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space. We use $\langle\cdot,\cdot\rangle$ to denote the standard inner product and $\|\cdot\|$ to denote the Euclidean norm. For any $w\in\mathbb{R}$, let $\lceil w\rceil$ denote the smallest integer greater than or equal to w. The notations $[\cdot]_+$ and $[\cdot]_-$ denote the projection operator onto the nonnegative orthant and nonpositive orthant, respectively.

A mapping ϕ is said to be *L_ϕ-Lipschitz continuous* on a set Ω if $\|\phi(x)-\phi(y)\|\le L_\phi\|x-y\|$ for all $x,y\in\Omega$. It is said to be *$L_{\nabla\phi}$-smooth* on Ω if $\|\nabla\phi(x)-\nabla\phi(y)\|\le L_{\nabla\phi}\|x-y\|$ for all $x,y\in\Omega$, where $\nabla\phi$ denotes the transpose of the Jacobian of ϕ.

Given a nonempty closed convex set $\Omega\subseteq\mathbb{R}^n$, let $\mathrm{int}(\Omega)$ denote its interior, $\mathrm{dist}(x,\Omega)$ the Euclidean distance from x to Ω, and $\Pi_\Omega(x)$ the Euclidean projection of x onto Ω. The indicator function of Ω is denoted by δ_Ω, and the normal cone to Ω at $x\in\Omega$ is denoted by $\mathcal{N}_\Omega(x)$. For a closed convex cone $\mathcal{K}\subseteq\mathbb{R}^n$, its dual cone is denoted by $\mathcal{K}^*$, that is, $\mathcal{K}^*=\{y\in\mathbb{R}^n:\langle y,x\rangle\ge 0\ \forall x\in\mathcal{K}\}$. Finally, let $\mathscr{B}$ denote the unit Euclidean ball, i.e., $\mathscr{B}=\{x:\|x\|\le 1\}$.

Throughout this paper, we impose the following additional assumptions on problem (1).

Assumption 1. *(i) The set X (i.e., the domain of ψ) is compact, and ψ is continuous on X.*

(ii) Each function f_i is L_{f_i}-Lipschitz continuous and $L_{\nabla f_i}$-smooth on X for all $i=1,2,\dots,s$.

(iii) The mapping c is L_c-Lipschitz continuous and $L_{\nabla c}$-smooth on X, and satisfies $\|c(x)\|\le C_c$ for all $x\in X$.

(iv) There exists $x_{\rm int}\in X$ such that $c(x_{\rm int})\in\mathrm{int}(-\mathcal{K})$; that is, there exists $D_c>0$ such that $c(x_{\rm int})+D_c\mathscr{B}\subset-\mathcal{K}$.

(v) The projection onto $\mathcal{K}^*$ can be exactly computed.

In addition, for notational convenience, we define

$$C:=L_{\nabla c}C_c+L_c^2,\qquad D_X:=\max\Big\{\max_{x,y\in X}\|x-y\|,\,1\Big\},\tag{8}$$

$$D_F:=\max_{x,y\in X}\{F(x)-F(y)\},\qquad D_\psi:=\max_{x,y\in X}\{\psi(x)-\psi(y)\},\tag{9}$$

$$f(x):=\frac{1}{s}\sum_{i=1}^s f_i(x),\qquad L_f:=\frac{1}{s}\sum_{i=1}^s L_{f_i},\qquad L_{\nabla f}:=\frac{1}{s}\sum_{i=1}^s L_{\nabla f_i},\tag{10}$$

$$\Lambda:=\frac{5L_{\nabla f}D_X^2/2+L_fD_X+D_\psi}{D_c}.\tag{11}$$

It then follows from Assumption 1 that C, D_X, D_F, and D_ψ are finite. Moreover, f is $L_{\nabla f}$-smooth on X, and

$$\|\nabla f(x)\|\le L_f\qquad\forall x\in X.\tag{12}$$

2 A first-order proximal augmented Lagrangian method for constrained strongly convex optimization

In this section, we present a first-order proximal augmented Lagrangian (FPAL) method for solving a constrained strongly convex optimization problem, which will serve as a subproblem solver for problem (1). Specifically, we consider the problem

$$\begin{aligned} \bar{F}^* = \min_x \ & \{\bar{F}(x) := \bar{f}(x) + \psi(x)\} \\ \text{s.t. } & c(x) \in -\mathcal{K}, \end{aligned} \tag{13}$$

where ψ, c, and $\mathcal{K}$ are defined in Section 1 and satisfy Assumption 1. In addition, we impose the following assumption on problem (13).

Assumption 2. *The function $\bar{f}$ is $L_{\nabla\bar{f}}$-smooth and μ-strongly convex on X, where $0 < \mu \le L_{\nabla\bar{f}}$ and $X = \operatorname{dom}\psi$ is a compact convex set.*

Under Assumptions 1 and 2, strong duality holds for problem (13) and its Lagrangian dual problem, which is given by

$$\max_{\lambda\in\mathcal{K}^*} \min_x \left\{\bar{f}(x) + \langle\lambda, c(x)\rangle + \psi(x)\right\}. \tag{14}$$

As a result, there exists a pair of optimal solutions $(x^*, \lambda^*) \in X \times \mathcal{K}^*$ to problems (13) and (14) that satisfy the following Karush-Kuhn-Tucker (KKT) conditions:

$$0 \in \nabla\bar{f}(x^*) + \nabla c(x^*)\lambda^* + \partial\psi(x^*), \qquad c(x^*) \in \mathcal{N}_{\mathcal{K}^*}(\lambda^*).$$

Since computing a pair of *exact* optimal solutions (x^*, λ^*) is generally intractable, we instead aim to compute an approximate solution, particularly an ϵ-KKT solution as defined below.

Definition 1. *For any $\epsilon > 0$, a pair $(x, \lambda) \in X \times \mathcal{K}^*$ is called an ϵ-KKT solution of problem* (13) *and its dual if it satisfies*

$$\operatorname{dist}\big(0, \nabla\bar{f}(x) + \nabla c(x)\lambda + \partial\psi(x)\big) \le \epsilon, \quad \operatorname{dist}\big(c(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)\big) \le \epsilon.$$

We next adopt the FPAL method proposed in [14, Algorithm 2] to compute an ϵ-KKT solution of problem (13) and its dual. While the original method was developed for the case where $\bar{f}$ is convex (but not necessarily strongly convex), we apply it to the strongly convex setting and derive improved complexity results by leveraging the strong convexity of $\bar{f}$.

Specifically, let $(x_k, \lambda_k) \in X \times \mathcal{K}^*$ denote a pair of approximate solutions generated at the $(k-1)$th iteration. At the kth iteration, the FPAL method applies an accelerated proximal gradient (APG) method to approximately solve the proximal augmented Lagrangian subproblem:

$$\min_x \left\{\overline{\mathcal{L}}(x, \lambda_k; \rho_k) := \mathcal{L}(x, \lambda_k; \rho_k) + \frac{1}{2\rho_k}\|x - x_k\|^2\right\} \tag{15}$$

for some penalty parameter $\rho_k > 0$, where $\mathcal{L}(x, \lambda; \rho)$ is the augmented Lagrangian (AL) function defined as

$$\mathcal{L}(x, \lambda; \rho) = \bar{f}(x) + \frac{1}{2\rho}\Big(\operatorname{dist}^2\big(\lambda + \rho c(x), -\mathcal{K}\big) - \|\lambda\|^2\Big) + \psi(x).$$

The Lagrange multiplier λ_{k+1} is then updated via the standard dual update scheme (see line 3 of Algorithm 1).

We now present the FPAL method in Algorithm 1 for solving problem (13).

Algorithm 1 A first-order proximal augmented Lagrangian method for problem (13)

Input: $\bar{\epsilon} > 0$, $(x_0, \lambda_0) \in X \times \mathcal{K}^*$, $\rho_0 \geq 1$, $0 < \eta_0 \leq 1$, $\zeta > 1$, $0 < \sigma < 1/\zeta$, $\rho_k = \rho_0 \zeta^k$, $\eta_k = \eta_0 \sigma^k$ for all $k \geq 0$.

1: **for** $k = 0, 1, \ldots$ **do**
2: Apply Algorithm 4 in Appendix A with $\tilde{f}(\cdot) \leftarrow \overline{\mathcal{L}}(\cdot, \lambda_k; \rho_k) - \psi(\cdot)$, $\varepsilon \leftarrow \eta_k$, $\mu_{\tilde{f}} \leftarrow \mu + \rho_k^{-1}$, $\tilde{x}_0 \leftarrow x_k$ to the proximal AL subproblem (15), and denote its output by x_{k+1}.
3: Set $\lambda_{k+1} = \Pi_{\mathcal{K}^*}(\lambda_k + \rho_k c(x_{k+1}))$.
4: **if** $\rho_k^{-1} \|(x_{k+1}, \lambda_{k+1}) - (x_k, \lambda_k)\| + \eta_k \leq \bar{\epsilon}$ **then**
5: Terminate the algorithm and output (x_{k+1}, λ_{k+1}).
6: **end if**
7: **end for**

To analyze the complexity of Algorithm 1 for computing an $\bar{\epsilon}$-KKT solution of problem (13) and its dual, we introduce some additional notation and state a property of Algorithm 1 that will be used in the subsequent analysis.

Let (x^*, λ^*) be an arbitrary pair of optimal solutions of (13) and (14). Define

$$\begin{aligned} &D = \max\big\{\|(x_0, \lambda_0) - (x^*, \lambda^*)\|, 1\big\}, \quad B = L_{\nabla \bar{f}} + L_{\nabla c}\|\lambda^*\| + L_{\nabla c} D, \\ &\bar{C} = C\rho_0 + B + L_{\nabla c}\rho_0\eta_0/(1 - \zeta\sigma) + 1, \end{aligned} \tag{16}$$

where the constants C, $L_{\nabla c}$, and $L_{\nabla \bar{f}}$ are defined in Section 1 and Assumption 2, respectively. In addition, it follows from Lemmas 3 and 5 in [14] that

$$\|(x_{k+1}, \lambda_{k+1}) - (x_k, \lambda_k)\| \leq \|(x_0, \lambda_0) - (x^*, \lambda^*)\| + \sum_{i=0}^{k} \rho_i \eta_i \qquad \forall k \geq 0, \tag{17}$$

$$\|(x_k, \lambda_k) - (x^*, \lambda^*)\| \leq \|(x_0, \lambda_0) - (x^*, \lambda^*)\| + \sum_{i=0}^{k-1} \rho_i \eta_i \qquad \forall k \geq 0. \tag{18}$$

We are now ready to present the complexity result of Algorithm 1 for computing an $\bar{\epsilon}$-KKT solution to problem (13) and its dual. The proof is deferred to Subsection 5.1.

Theorem 1. *Let D_X, D, $\bar{C}$, $\bar{\epsilon}$, ρ_0, η_0, ζ, σ, μ be given in* (8), (16), *Algorithm 1, and Assumption 2, respectively, and let*

$$\bar{K} = \left\lceil \max\left\{1, \log_\zeta \frac{2(D + \rho_0\eta_0)}{(1 - \zeta\sigma)\bar{\epsilon}}\right\}\right\rceil, \tag{19}$$

$$\bar{N} = \frac{16\zeta\sqrt{\bar{C}}}{\sqrt{\mu}(\sqrt{\zeta} - 1)} \log \frac{2\zeta\bar{C}D_X}{\eta_0\sigma} \left\lceil \max\left\{1, \log_\zeta \frac{2(D + \rho_0\eta_0)}{(1 - \zeta\sigma)\bar{\epsilon}}\right\}\right\rceil \max\left\{1, \sqrt{\frac{2(D + \rho_0\eta_0)}{(1 - \zeta\sigma)\bar{\epsilon}}}\right\}. \tag{20}$$

Then, Algorithm 1 terminates and outputs an $\bar{\epsilon}$-KKT solution to problem (13) *and its dual in at most $\bar{K} + 1$ iterations. Moreover, the number of evaluations of $\nabla \bar{f}$ and ∇c performed by Algorithm 1 is at most $\bar{N}$, respectively.*

3 An inexact stochastic variance-reduced gradient method for problem (1)

In this section, we propose an inexact stochastic variance-reduced gradient (SVRG) method for solving problem (1).

Observe that when $c = 0$, problem (1) reduces to the unconstrained finite-sum optimization problem (3), which has been extensively studied in the literature (see, e.g., [1, 3, 5, 8, 17, 18, 19, 22, 24]). Notably,

an accelerated SVRG method was proposed in [8, Algorithm 1] for solving (3). This method can be viewed as a restarted accelerated stochastic gradient scheme with two nested loops: the outer loop updates a reference point $\tilde{x}$ and the initialization points x_0, z_0 for the inner loop, while the inner loop performs a sequence of accelerated stochastic gradient steps. At each inner iteration, an SVRG estimator of f is computed using the full gradient $\nabla f(\tilde{x})$ at the reference point $\tilde{x}$, along with the sampled gradients $\nabla f_i(\tilde{x})$ and $\nabla f_i(y)$, where y is updated at each inner iteration, and $i \in \{1, 2, \ldots, s\}$ is drawn according to a prescribed probability distribution.

Note that problem (1) can be reformulated as (5). While problem (5) shares a similar structure with (3), the accelerated SVRG method in [8, Algorithm 1] is not directly applicable to (5) because the proximal operator of $\gamma[\psi(\cdot) + \delta_S(\cdot)]$ is not exactly computable for general $\gamma > 0$. In fact, evaluating this proximal operator requires solving problem (6), which typically involves a non-simple constraint set. As a result, computing an exact solution to (6) is generally challenging. To address this issue, we propose an inexact SVRG method for solving problem (5) (equivalently, problem (1)), which follows a similar framework to [8, Algorithm 1] but only requires an inexact solution to problem (6)—particularly an ϵ-KKT solution (x, λ) satisfying (7) for some suitable $\epsilon > 0$. As discussed in Section 2, such an ϵ-KKT solution can be efficiently computed by Algorithm 1.

We now present an inexact SVRG method for solving problem (1) in Algorithm 2.

Algorithm 2 An inexact SVRG method for problem (1)

Input: $\epsilon > 0$, $R \geq 0$, $\{\alpha_k\} \subset (0, 1]$, $\{p_k\} \subset (0, 1]$, $\{\gamma_k\} \subset (0, \infty)$, $\{T_k\}$, $\{\theta_t\}$, a probability distribution $\mathcal{Q} = \{q_1, q_2, \ldots, q_s\}$.

Call Algorithm 3 in Appendix A to solve the problem $\min_{x \in X} \operatorname{dist}^2\big(c(x), -\mathcal{K}\big)$ to find $x^0 \in X$ such that $\operatorname{dist}\big(c(x^0), -\mathcal{K}\big) \leq \epsilon$.
Set $\tilde{x}_1 = x^0$, $\tilde{z}_1 = x^0$, and $\tilde{\lambda}_1 = 0$.
for $k = 1, 2, \ldots$ **do**
  Set $x_0 = \tilde{x}_k$, $z_0 = \tilde{z}_k$, $\lambda_0 = \tilde{\lambda}_k$, and $\tilde{g}_k = \nabla f(\tilde{x}_k)$.
  for $t = 1, 2, \ldots, T_k$ **do**
    Sample $i_t \in \{1, 2, \ldots, s\}$ according to $\mathcal{Q}$.
    $y_t = (1 - \alpha_k - p_k)x_{t-1} + \alpha_k z_{t-1} + p_k \tilde{x}_k$.
    $g_t = (\nabla f_{i_t}(y_t) - \nabla f_{i_t}(\tilde{x}_k))/(q_{i_t} s) + \tilde{g}_k$.
    Call Algorithm 1 with $\bar{\epsilon} \leftarrow \epsilon$, $\mu \leftarrow 1/\gamma_k$, $(x_0, \lambda_0) \leftarrow (z_{t-1}, \lambda_{t-1} \min\{R/\|\lambda_{t-1}\|, 1\})$ to solve the subproblem

$$\begin{aligned} \min_x \ & \left\{ F_{k,t}(x) := \frac{1}{2\gamma_k} \|x - z_{t-1} + \gamma_k g_t\|^2 + \psi(x) \right\} \\ \text{s.t. } & c(x) \in -\mathcal{K}, \end{aligned} \tag{21}$$

    and denote its output as (z_t, λ_t).[2]
    $x_t = (1 - \alpha_k - p_k)x_{t-1} + \alpha_k z_t + p_k \tilde{x}_k$.
  end for
  Set $\tilde{z}_{k+1} = z_{T_k}$, $\tilde{x}_{k+1} = \sum_{t=1}^{T_k} (\theta_t x_t) / \sum_{t=1}^{T_k} \theta_t$, and $\tilde{\lambda}_{k+1} = \lambda_{T_k}$.
end for

Remark 1. *(i) Algorithm 2 shares some similarities with the gradient sliding (GS) method [7]. In particular, the latter is a double-loop first-order method for solving problems of the form*

$$\min_x \{g(x) + h(x) + \psi(x)\}, \tag{22}$$

where g is a Lipschitz-smooth convex function, ψ is the same as in (1), and h is a nonsmooth convex function such that $\operatorname{ri}(\operatorname{dom} h) \supseteq \operatorname{dom} \psi$.[3] *The outer loop of GS performs an inexact accelerated proximal*

[2] By convention, we set $0/0 = 0$ and $1/0 = \infty$.

[3] Here, $\operatorname{ri}(\operatorname{dom} h)$ denotes the relative interior of $\operatorname{dom} h$, and the latter assumption ensures that $\partial h(x) \neq \emptyset$ for all $x \in \operatorname{dom} \psi$, which is required for applying a proximal subgradient method to solve subproblems of the form (23).

gradient method, whose main computational effort lies in solving subproblems of the form

$$\min_x \Big\{ \frac{1}{2\gamma}\|x-u\|^2 + h(x) + \psi(x) \Big\} \tag{23}$$

for $\gamma > 0$ and a vector u obtained using the gradient of g at some point. The inner loop of GS applies a proximal subgradient method to solve these subproblems, where the main computational cost lies in evaluating the proximal operator of ψ at points obtained using subgradients of h.

However, although problem (5) *shares the same form as* (22)*, it does not belong to this problem class, since* $\mathrm{ri}(\mathrm{dom}\,\delta_S) = \mathrm{ri}(S)$ *typically does not contain* $\mathrm{dom}\,\psi$*. Consequently, the GS method [7] is not applicable to problem* (5) *(or equivalently,* (1)*).*

(ii) The parameter R ensures the boundedness of λ_t, where λ_t is an approximate optimal Lagrange multiplier of subproblem (21) *computed in Line 9 of Algorithm 2. Indeed, as will be shown in Lemmas 2 and 4(iii), it holds that*

$$\|\lambda_t\| \le D_X + R + 2\Lambda + \rho_0\eta_0/(1-\zeta\sigma), \tag{24}$$

where D_X and Λ are defined in (8) *and* (11)*, respectively, and ρ_0, ζ, and σ are input parameters of Algorithm 1.*

In addition, note from Line 9 of Algorithm 2 that $\lambda_0 = \alpha\lambda_{t-1}$, where $\alpha = \min\{R/\|\lambda_{t-1}\|, 1\}$, and λ_0 serves as the initial guess for the optimal Lagrange multiplier of (21)*. Clearly, if $R = 0$, then $\lambda_0 = 0$, which corresponds to a static initialization for Algorithm 1. On the other hand, by* (24) *and the definition of α, we have*

$$\frac{R}{D_X + R + 2\Lambda + \rho_0\eta_0/(1-\zeta\sigma)} \le \alpha \le 1.$$

Consequently, when R is large, $\alpha \approx 1$ and hence $\lambda_0 \approx \lambda_{t-1}$, which provides a dynamic (warm-start) initialization for Algorithm 1.

As will be shown in Theorem 2, setting $R = 0$ (static initialization) yields the lowest first-order oracle complexity bounds among all possible choices of R. However, numerical experiments indicate that larger values of R (warm-start) often lead to faster convergence in practice.

The following theorem presents the convergence results and provides an estimate of the computational cost of Algorithm 2 in terms of the number of evaluations of ∇f_i and ∇c. The proof is deferred to Subsection 5.2.

Theorem 2. *Suppose that Assumption 1 holds. Let D_X, D_F, $L_{\nabla f}$, Λ, C_c, L_c, $L_{\nabla c}$, $L_{\nabla f_i}$, ϵ, ρ_0, η_0, ζ, σ, and R be given in* (8), (9), (10), (11)*, Assumption 1, and Algorithms 1 and 2, respectively, and let*

$$\begin{aligned}
&C_1 = 2D_X^{3/2}\sqrt{L_c^2 + L_{\nabla c}C_c},\\
&C_2 = (L_{\nabla c}C_c + L_c^2)\rho_0 + 3L_{\nabla f}/2 + L_{\nabla c}(D_X + R + 2\Lambda) + L_{\nabla c}\rho_0\eta_0/(1-\zeta\sigma) + 1, &(25)\\
&C_3 = L_{\nabla f}^{-\frac{1}{2}}\frac{16\zeta\sqrt{C_2}}{\sqrt{\zeta}-1}\log\frac{2\zeta C_2 D_X}{\eta_0\sigma}, \qquad C_4 = \frac{2(D_X + R + \Lambda + \rho_0\eta_0)}{1-\zeta\sigma}, &(26)\\
&C_5 = 3D_X + 2R + 4\Lambda + 2\rho_0\eta_0/(1-\zeta\sigma), \qquad C_6 = 2D_F + 3L_{\nabla f}D_X^2/2. &(27)
\end{aligned}$$

Assume that $\{\tilde{x}_k\}$ is generated by Algorithm 2 with the parameters chosen as

$$\begin{aligned}
&k_0 = \lfloor \log_2 s \rfloor + 1, \quad \alpha_k = \begin{cases} \frac{1}{2} & \text{if } 1 \le k \le k_0,\\ \frac{2}{k-k_0+4} & \text{if } k > k_0, \end{cases} \quad p_k = \frac{1}{2}, \quad \gamma_k = \frac{1}{3L_{\nabla f}\alpha_k},\\
&q_i = \frac{L_{\nabla f_i}}{\sum_{j=1}^s L_{\nabla f_j}}, \quad T_k = \begin{cases} 2^{k-1} & \text{if } 1 \le k \le k_0,\\ 2^{k_0-1} & \text{if } k > k_0, \end{cases} \quad \theta_t = \begin{cases} \frac{\gamma_k}{\alpha_k}(\alpha_k + p_k) & \text{if } 1 \le t \le T_k - 1,\\ \frac{\gamma_k}{\alpha_k} & \text{if } t = T_k. \end{cases}
\end{aligned} \tag{28}$$

Let N_f and N_c denote the number of evaluations of ∇f_i and ∇c during the k outer iterations of Algorithm 2 for all $k \geq 1$, respectively. Then, for all $k > k_0$, it holds that

$$\operatorname{dist}\big(c(\tilde{x}_k), -\mathcal{K}\big) \leq \epsilon, \qquad F(\tilde{x}_k) - F^* \geq -D_\lambda \epsilon, \qquad \mathbb{E}[F(\tilde{x}_k) - F^*] \leq \frac{16C_6}{s(k-k_0+3)^2} + 8C_5\epsilon, \tag{29}$$

$$N_f \leq 2sk, \qquad N_c \leq C_1\epsilon^{-1} + 2sC_3\left\lceil \max\left\{1, \log_\zeta \frac{C_4}{\epsilon}\right\}\right\rceil \max\left\{1, \sqrt{\frac{C_4}{\epsilon}}\right\} k^{\frac{3}{2}}, \tag{30}$$

where D_λ denotes the distance between the origin and the set of optimal Lagrange multipliers of (1).

Remark 2. *(i) Letting $k = \lceil k_0 + s^{-1/2}\epsilon^{-1/2}\rceil$, one can observe from* (29) *that*

$$\operatorname{dist}(c(\tilde{x}_k), -\mathcal{K}) \leq \epsilon, \qquad F(\tilde{x}_k) - F^* \geq -D_\lambda\epsilon, \qquad \mathbb{E}[F(\tilde{x}_k) - F^*] \leq 8(C_5 + 2C_6)\epsilon,$$
$$N_f = \mathcal{O}\big(s\log s + s^{1/2}\epsilon^{-1/2}\big), \qquad N_c = \widetilde{\mathcal{O}}\big(s(\log s)^{3/2}\epsilon^{-1/2} + s^{1/4}\epsilon^{-5/4}\big).$$

As a result, Algorithm 2 can find an ϵ-SFSO solution of problem (1) *using $\mathcal{O}\big(s\log s + s^{1/2}\epsilon^{-1/2}\big)$ and $\widetilde{\mathcal{O}}\big(s(\log s)^{3/2}\epsilon^{-1/2} + s^{1/4}\epsilon^{-5/4}\big)$ evaluations of ∇f_i and ∇c, respectively. In contrast, when $\mathcal{K}$ is the nonnegative orthant, the stochastic variance-reduced first-order penalty method proposed in [13, Algorithm 2] requires $\mathcal{O}(s\log s + s^{1/2}\epsilon^{-3/2})$ evaluations of both ∇f_i and ∇c to compute an ϵ-SFSO solution of* (1). *Therefore, Algorithm 2 significantly improves the first-order oracle complexity achieved by [13, Algorithm 2].*

(ii) We now consider the special case of problem (1) *with $s = 1$. In this case, problem* (1) *reduces to a standard convex conic optimization problem, and Algorithm 2 becomes a novel deterministic first-order method. As discussed above, Algorithm 2 can compute an ϵ-optimal solution of* (1) *using $\mathcal{O}(\epsilon^{-1/2})$ and $\widetilde{\mathcal{O}}(\epsilon^{-5/4})$ evaluations of ∇f_1 and ∇c, respectively. In contrast, the best-known first-order methods—such as the extra anchored gradient method [6, Algorithm 3] and the first-order proximal augmented Lagrangian method [14, Algorithm 2]—require $\mathcal{O}(\epsilon^{-1})$ or $\widetilde{\mathcal{O}}(\epsilon^{-1})$ evaluations of ∇f_1, ∇c, and projection onto $-\mathcal{K}$ to obtain an ϵ-optimal solution of* (1). *Compared to these methods, Algorithm 2 significantly reduces the gradient evaluation complexity with respect to f_1, at the cost of a higher complexity with respect to c. Note that the number of projections onto $-\mathcal{K}$ is on the same order as the number of evaluations of ∇c in Algorithm 2. Consequently, when computing ∇c and projecting onto $-\mathcal{K}$ are cheaper than evaluating ∇f_1, Algorithm 2 can achieve superior practical performance.*

4 Numerical results

In this section, we conduct preliminary experiments to test the performance of our proposed method (Algorithm 2), and compare it with the stochastic variance-reduced first-order penalty (SVRFP) method [13, Algorithm 2] and the extra anchored gradient (EAG) method [6, Algorithm 3]. EAG is an optimal first-order method designed for monotone inclusion problems. Since problem (1) can be equivalently reformulated as a monotone inclusion problem, EAG is well-suited for solving it. The experiments are coded in MATLAB and all the computations are conducted on a desktop with a 2 GHz Intel i9-14900F 24-core processor and 32 GB of RAM.

4.1 Fairness-constrained logistic regression

We consider the dataset `a1a` from LIBSVM [2], which consists of approximately 67% male samples and 33% female samples. Due to this imbalance, a classical unconstrained logistic regression model trained

[4]We choose $p_k = 1/2$ for simplicity in the analysis of Algorithm 2. However, dynamic choices are also possible, provided they satisfy the conditions specified in Lemma 7.

on the full dataset typically performs well on the male subgroup but poorly on the female subgroup, indicating potential unfairness. Therefore, incorporating a fairness constraint [4, 23] is necessary.

To address this issue, we minimize the empirical logistic loss subject to a group fairness constraint, defined as a performance guarantee on a validation set V randomly drawn from the female subgroup of the dataset. The resulting fairness-constrained logistic regression model is given by

$$\min_{w,b} \frac{1}{s}\sum_{i=1}^{s} \ell(w,b;\, x_i, y_i) + \lambda\|w\|_1 \quad \text{s.t.} \quad \frac{1}{|V|}\sum_{j\in V} \ell(w,b;\, x_j, y_j) \le \delta \tag{31}$$

for some $\lambda, \delta > 0$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, $i = 1, \ldots, s$, are the sample data from `a1a`, and ℓ is the logistic loss function given by $\ell(w, b; x, y) = \log\left(1 + e^{-y(w^T x + b)}\right)$.

We set $\lambda = 0.01$, $|V| = 100$, and $\delta = (\widehat{L}_V + L_V^*)/2$ for (31), where $\widehat{L}_V = |V|^{-1}\sum_{j\in V} \ell(\hat{w}, \hat{b};\, x_j, y_j)$ is the validation loss evaluated at $(\hat{w}, \hat{b})$, an optimal solution of the unconstrained counterpart of (31) (i.e., problem (31) with $\delta = \infty$), and

$$L_V^* = \min_{w,b} |V|^{-1}\sum_{j\in V} \ell(w, b, x_j, y_j). \tag{32}$$

Clearly, this choice of δ ensures that problem (31) is feasible. Moreover, since $L_V^* < \widehat{L}_V$ typically holds, we have $\delta < \widehat{L}_V$, implying that $(\hat{w}, \hat{b})$ is infeasible for (31). Consequently, the constraint in (31) is generally active at any optimal solution.

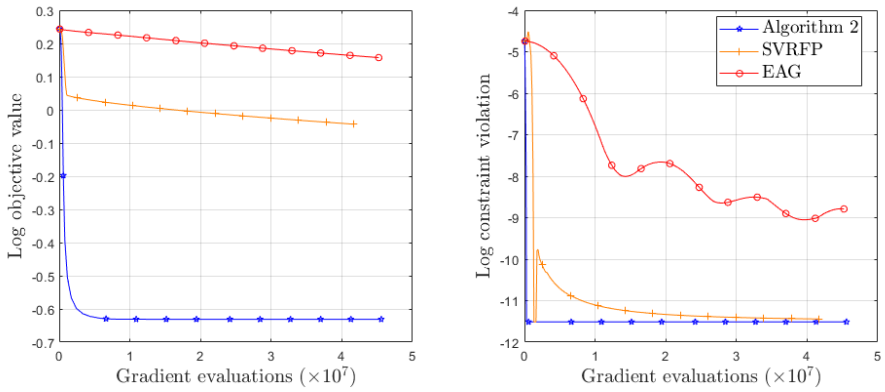


Figure 1: Convergence behavior of objective value and constraint violation for fairness-constrained logistic regression.

To apply Algorithm 2, SVRFP [13], and EAG [6] to solve (31), we specify an initial point and the corresponding parameters. Specifically, we solve problem (32) to obtain an 10^{-2}-optimal solution (w_0, b_0), and use it as the initial point for all three methods. The parameters for these methods are chosen as follows:

- For Algorithm 2, we set $(\epsilon, \rho_0, \eta_0, \zeta, \sigma, R) = (0.01, 10, 0.1, 3, 0.3, 10)$, and choose $(\alpha_k, p_k, \gamma_k, q_i, T_k, \theta_t)$ as specified in Theorem 2.
- For SVRFP, the parameters $(\alpha_k, p_k, \gamma_k, q_i, T_k, \theta_t, \rho_k)$ are selected as in [13, Theorem 4(i)].
- For EAG, we select (β_t, λ) according to [6, Theorem 2].

We now apply these three methods to solve (31) and compare their performance. In particular, we evaluate the objective value and constraint violation as the number of gradient evaluations increases for all three methods. The results are presented in Figure 1, which shows that the stochastic methods—Algorithm 2 and SVRFP—converge significantly faster than the deterministic method EAG, with

Algorithm 2 converging faster than SVRFP in terms of objective value. In addition, Algorithm 2 maintains a consistently low level of constraint violation throughout the iterations, as it is designed to nearly achieve feasibility at each iteration. In contrast, both SVRFP and EAG exhibit an initial increase in constraint violation followed by a gradual decrease.

4.2 Quadratically constrained least squares regression

We consider the following quadratically constrained least-squares problem [10]:

$$\min_x \ \frac{1}{s}\sum_{i=1}^{s} \|A_i x - b_i\|^2 + \lambda\|x\|^2 \quad \text{s.t.} \quad \|Cx - d\|^2 \le r^2 \tag{33}$$

for some $\lambda, r > 0$, where $A_i \in \mathbb{R}^{l\times n}$, $b_i \in \mathbb{R}^l$, $C \in \mathbb{R}^{p\times n}$, and $d \in \mathbb{R}^p$. Quadratic constraints of this form often arise in signal processing (e.g., power budgets) and control (e.g., ellipsoidal uncertainty sets), where feasibility is critical.

We generate data for (33) with $n = 100$, $s = 10{,}000$, $l = 50$, and $p = 20$. Specifically, the data matrices A_i and C have entries drawn independently from $\mathcal{N}(0, l^{-1})$ and $\mathcal{N}(0, p^{-1})$, respectively. We then generate a sparse vector $\hat{x} \in \mathbb{R}^n$ with 10% nonzero entries, whose nonzero components are drawn from the standard normal distribution $\mathcal{N}(0,1)$, and set $b_i = A_i\hat{x} + \xi_i$, where $\xi_i \sim \mathcal{N}(0, 10^{-4}I)$.

We set $\lambda = 0.01$ and $r = 0.5\,\|C\hat{x}^* - d\|$ for (33), where $\hat{x}^* = \left(\sum_{i=1}^s A_i^T A_i + s\lambda I\right)^{-1} \sum_{i=1}^s A_i^T b_i$ is an optimal solution of the unconstrained counterpart of (33) (i.e., problem (33) with $r = \infty$), and $d \sim \mathcal{N}(0, I)$ is generated such that $C\hat{x}^* \neq d$. This condition, together with the choice of r, excludes $\hat{x}^*$ from being an optimal solution of (33). Consequently, the constraint in (33) must be active at any optimal solution.

To apply Algorithm 2, SVRFP [13], and EAG [6] to solve (33), we specify an initial point and the corresponding parameters. Specifically, we solve $\min_x \|Cx - d\|^2$ to obtain a 10^{-2}-feasible solution x_0 of (33), i.e., $\|Cx_0 - d\|^2 \le r^2 + 10^{-2}$. We then use x_0 as the initial point for all three methods. The parameters for these methods are chosen to be the same as those specified in Subsection 4.1.

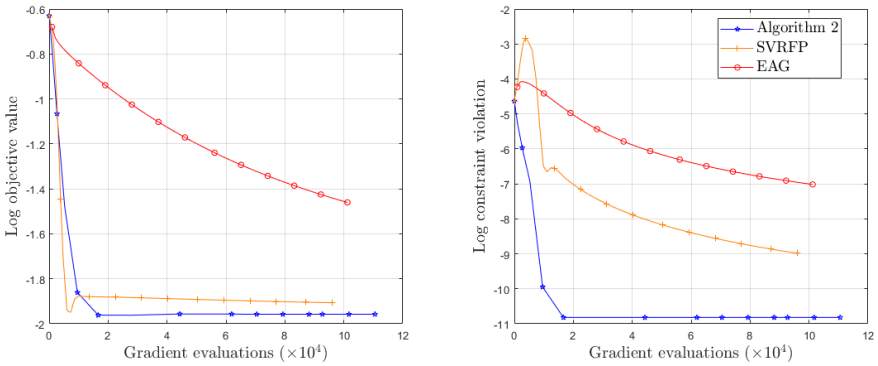


Figure 2: Convergence behavior of objective value and constraint violation for quadratically constrained least squares regression.

We now apply these three methods to solve (33) and compare their performance. In particular, we evaluate the objective value and constraint violation as the number of gradient evaluations increases for all three methods. The results are presented in Figure 2, which shows that the stochastic methods—Algorithm 2 and SVRFP—converge significantly faster than the deterministic method EAG, with Algorithm 2 converging faster than SVRFP in terms of objective value. In addition, Algorithm 2 achieves a rapid reduction in constraint violation and maintains it at a low level, whereas SVRFP exhibits an initial increase in constraint violation followed by a gradual decrease, and EAG maintains a substantially higher level of violation throughout the process.

4.3 Linearly constrained LASSO regression

In this subsection, we consider a special case of problem (1) with $s = 1$. In this case, as discussed in Remark 2(ii), Algorithm 2 reduces to a deterministic first-order method. Moreover, when computing ∇c and projecting onto $-\mathcal{K}$ are cheaper than evaluating ∇f_1, Algorithm 2 can achieve superior practical performance compared to SVRFP [13] and EAG [6]. We now provide a numerical example to illustrate this observation.

Specifically, we consider the following linearly constrained LASSO problem [20]:

$$\min_x \ \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \quad \text{s.t.} \quad Cx \le d \tag{34}$$

for some $\lambda > 0$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{p\times n}$, and $d \in \mathbb{R}^p$. Note that (34) is a special case of problem (1) with $s = 1$, where $f_1(x) = \|Ax - b\|^2/2$, $\psi(x) = \|x\|_1$, $c(x) = Cx - d$, and $\mathcal{K} = \mathbb{R}^p_+$. Moreover, computing ∇c and projecting onto $-\mathcal{K}$ are typically cheaper than evaluating ∇f_1.

We set $\lambda = 0.01$ and generate data for (34) with $m = 2{,}000$, $n = 2{,}000$, and $p = 20$. Specifically, we generate A, b, and C in the same manner as in Subsection 4.2 such that $C\hat{x}^* \neq 0$, where $\hat{x}^*$ is an optimal solution of the unconstrained counterpart of (34), i.e., $\hat{x}^* \in \arg\min_x\{\|Ax - b\|^2/2 + \lambda\|x\|_1\}$. We then set $d = 0.5[C\hat{x}^*]_+ + 2[C\hat{x}^*]_-$. This choice of d, together with $C\hat{x}^* \neq 0$, ensures that $\hat{x}^*$ is infeasible for (34). Consequently, at least one of the constraints in (34) must be active at any optimal solution.

To apply Algorithm 2, SVRFP [13], and EAG [6] to solve (34), we specify an initial point and the corresponding parameters. Specifically, we solve the feasibility problem $\min_x \|[Cx - d]_+\|^2$ to obtain a 10^{-2}-feasible solution x_0 of (34), i.e., $\|[Cx_0 - d]_+\| \le 10^{-2}$. We then use x_0 as the initial point for all three methods. The parameters for Algorithm 2 and EAG are chosen to be the same as those specified in Subsection 4.1, while the parameters $(\alpha_k, p_k, \gamma_k, q_i, T_k, \theta_t, \rho_k)$ for SVRFP are selected as in [13, Theorem 4(ii)] since it reduces to a deterministic method when $s = 1$.

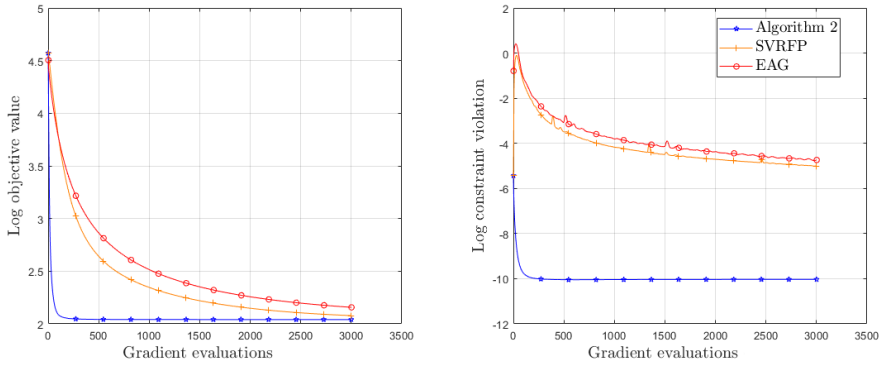


Figure 3: Convergence behavior of objective value and constraint violation for linearly constrained LASSO regression.

We now apply these three methods to solve (34) and compare their performance. Note that all methods reduce to deterministic ones since $s = 1$. In particular, we evaluate the objective value and constraint violation as the number of gradient evaluations increases for all three methods. The results are presented in Figure 3, which shows that Algorithm 2 converges significantly faster than SVRFP and EAG, while SVRFP and EAG exhibit similar convergence behavior in terms of objective value. In addition, Algorithm 2 achieves a rapid reduction in constraint violation and maintains it at a low level, whereas SVRFP and EAG exhibit similar but much slower convergence in constraint violation. This phenomenon is consistent with the observation in Remark 2(ii). Furthermore, the similar performance of SVRFP and EAG is expected, as they have comparable theoretical convergence rates.

5 Proof of the main results

In this section, we present a proof of the main results stated in Sections 2 and 3, which are particularly Theorems 1 and 2.

5.1 Proof of the main results in Section 2

In this subsection, we present a proof of Theorem 1. To this end, we begin by stating a lemma concerning the Lipschitz smoothness of the function $\overline{\mathcal{L}}(\cdot, \lambda_k; \rho_k)$, which was established in [14, Lemma 7].

Lemma 1 ([**14, Lemma 7**]). *Let $\{\rho_k\}$, $\{\eta_k\}$, and $\{\lambda_k\}$ be generated in Algorithm 1, and B, C, and $L_{\nabla c}$ be given in* (8), (16)*, and Assumption 1, respectively. Suppose that Assumptions 1 and 2 hold. Then the function $\overline{\mathcal{L}}(\cdot, \lambda_k; \rho_k)$ is L_k-smooth on X for each $k \geq 0$, where $\overline{\mathcal{L}}(\cdot, \lambda_k; \rho_k)$ is defined in* (15)*, and*

$$L_k = C\rho_k + B + L_{\nabla c} \sum_{i=0}^{k-1} \rho_i \eta_i + \rho_k^{-1}. \tag{35}$$

We are now ready to prove Theorem 1.

Proof of Theorem 1. We first show that Algorithm 1 terminates in at most $\bar{K}+1$ outer iterations, where $\bar{K}$ is defined in (19). Indeed, suppose for contradiction that it runs for more than $\bar{K}+1$ iterations. It then follows from Algorithm 1 that

$$\rho_{\bar{K}}^{-1} \big\| (x_{\bar{K}+1}, \lambda_{\bar{K}+1}) - (x_{\bar{K}}, \lambda_{\bar{K}}) \big\| + \eta_{\bar{K}} > \bar{\epsilon}. \tag{36}$$

Notice from Algorithm 1 that $\zeta > 1$ and $\rho_{\bar{K}} = \rho_0 \zeta^{\bar{K}}$. By these facts and (19), one has

$$\rho_{\bar{K}} \geq \rho_0 \zeta^{\bar{K}} \overset{(19)}{\geq} \frac{2\rho_0(D + \rho_0\eta_0)}{(1-\zeta\sigma)\bar{\epsilon}}. \tag{37}$$

Using this, (16), (17), and the fact that $\rho_0 \geq 1$, $0 < \zeta\sigma < 1$, $\rho_k = \rho_0\zeta^k$, and $\eta_k = \eta_0\sigma^k$, we obtain that

$$\begin{aligned} \frac{\|(x_{\bar{K}+1}, \lambda_{\bar{K}+1}) - (x_{\bar{K}}, \lambda_{\bar{K}})\|}{\rho_{\bar{K}}} &\overset{(16)(17)}{\leq} \frac{D + \sum_{k=0}^{\bar{K}} \rho_k \eta_k}{\rho_{\bar{K}}} \overset{(37)}{\leq} \frac{D + \rho_0\eta_0 \sum_{k=0}^{\bar{K}} (\zeta\sigma)^k}{2\rho_0(D+\rho_0\eta_0)/((1-\zeta\sigma)\bar{\epsilon})} \\ &\leq \frac{D + \rho_0\eta_0/(1-\zeta\sigma)}{2\rho_0(D+\rho_0\eta_0)/((1-\zeta\sigma)\bar{\epsilon})} = \frac{\bar{\epsilon}}{2} \cdot \frac{D(1-\zeta\sigma) + \rho_0\eta_0}{\rho_0(D+\rho_0\eta_0)} \leq \frac{\bar{\epsilon}}{2}. \end{aligned} \tag{38}$$

In addition, notice that $1 < \zeta < \sigma^{-1}$ and $D \geq 1$. By these facts and (19), one has

$$\bar{K} \geq \log_\zeta \frac{2(D+\rho_0\eta_0)}{(1-\zeta\sigma)\bar{\epsilon}} \geq \log_\zeta(2\bar{\epsilon}^{-1}) \geq \log_{\sigma^{-1}}(2\bar{\epsilon}^{-1}) = \log_\sigma(\bar{\epsilon}/2),$$

which together with $0 < \sigma < 1$ implies that $\sigma^{\bar{K}} \leq \bar{\epsilon}/2$. It follows from this, $0 < \eta_0 \leq 1$, and $\eta_{\bar{K}} = \eta_0\sigma^{\bar{K}}$ that $\eta_{\bar{K}} \leq \bar{\epsilon}/2$. This relation and (38) lead to a contradiction to (36). Hence, Algorithm 1 terminates in at most $\bar{K}+1$ outer iterations.

Suppose now that the termination condition $\rho_k^{-1} \|(x_{k+1}, \lambda_{k+1}) - (x_k, \lambda_k)\| + \eta_k \leq \bar{\epsilon}$ holds at some iterate k for Algorithm 1. Then, by an argument similar to that in the proof of [14, Theorem 4(i)], it follows that the output (x_{k+1}, λ_{k+1}) is an $\bar{\epsilon}$-KKT solution of problem (13) and its dual.

In addition, note that Algorithm 4 is called at the kth outer iteration of Algorithm 1 to solve the proximal AL subproblem (15). Moreover, each call of Algorithm 4 performs the same number of evaluations of $\nabla \bar{f}$ and ∇c. Let n_k denote the number of evaluations of either $\nabla \bar{f}$ or ∇c performed by Algorithm 4 during the kth outer iteration of Algorithm 1. Note that when applied to (15), Algorithm 4

performs at most two evaluations of $\nabla \bar{f}$ and ∇c per iteration. By this fact, Lemma 1, and Theorem 4 with $\tilde{f}(\cdot) \leftarrow \overline{\mathcal{L}}(\cdot, \lambda_k; \rho_k) - \psi(\cdot)$, $L_{\nabla \tilde{f}} \leftarrow L_k$, $\mu_{\tilde{f}} \leftarrow \mu + \rho_k^{-1}$, and $\varepsilon \leftarrow \eta_k$, one has

$$n_k \leq 2 \left\lceil \sqrt{\frac{L_k}{\mu + \rho_k^{-1}}} \right\rceil \max \left\{ 1, \left\lceil 2 \log \frac{2 L_k D_X}{\eta_k} \right\rceil \right\}, \tag{39}$$

where L_k is given in (35).

Further, using $\rho_0 \geq 1$, $\zeta > 1$, $0 < \zeta\sigma < 1$, $\rho_i = \rho_0 \zeta^i$, $\eta_i = \eta_0 \sigma^i$, (16), and Lemma 1, we obtain that

$$\begin{aligned} L_k &\leq C\rho_k + B + L_{\nabla c} \sum_{i=0}^{k-1} \rho_i \eta_i + 1 = C\rho_0 \zeta^k + B + L_{\nabla c} \rho_0 \eta_0 \sum_{i=0}^{k-1} (\zeta\sigma)^i + 1 \\ &\leq C\rho_0 \zeta^k + B + \frac{L_{\nabla c} \rho_0 \eta_0}{1 - \zeta\sigma} + 1 \overset{(16)}{\leq} \bar{C} \zeta^k. \end{aligned} \tag{40}$$

Notice that $L_{\nabla \bar{f}} \geq \mu$. By this fact and (16), one can observe that $\bar{C} \geq \mu$, which together with $0 < \eta_0, \sigma \leq 1$ and $\bar{C}, \zeta, D_X \geq 1$ implies that

$$\sqrt{\bar{C}/\mu}\, \zeta^{k/2} \geq 1, \qquad 2 \log \left(2 \bar{C} D_X \zeta^k / (\eta_0 \sigma^k) \right) \geq 2 \log 2 \geq 1. \tag{41}$$

Using these relations, (39), (40), and $\rho_k \geq 1$, we obtain that

$$\begin{aligned} n_k &\overset{(39)}{\leq} 2 \left\lceil \sqrt{\frac{L_k}{\mu}} \right\rceil \max \left\{ 1, \left\lceil 2 \log \frac{2 L_k D_X}{\eta_k} \right\rceil \right\} \overset{(40)}{\leq} 2 \left\lceil \sqrt{\frac{\bar{C}}{\mu}} \zeta^{\frac{k}{2}} \right\rceil \max \left\{ 1, \left\lceil 2 \log \frac{2 \bar{C} D_X \zeta^k}{\eta_0 \sigma^k} \right\rceil \right\} \\ &\leq 2 \left\lceil \sqrt{\frac{\bar{C}}{\mu}} \zeta^{\frac{k}{2}} \right\rceil \left\lceil 2 \log \frac{2 \bar{C} D_X \zeta^k}{\eta_0 \sigma^k} \right\rceil \overset{(41)}{\leq} 16 \sqrt{\frac{\bar{C}}{\mu}} \zeta^{\frac{k}{2}} \log \frac{2 \bar{C} D_X \zeta^k}{\eta_0 \sigma^k} \\ &\leq 16 \sqrt{\frac{\bar{C}}{\mu}} \log \left(\frac{2 \bar{C} D_X \zeta}{\eta_0 \sigma} \right) \max \left\{ 1, k \zeta^{k/2} \right\}, \end{aligned} \tag{42}$$

where the last inequality follows from $2\bar{C} D_X \zeta / \eta_0 \geq 1$ due to $\bar{C}, D_X, \zeta \geq 1$, and $0 < \eta_0 \leq 1$. Observe that

$$1 + \sum_{k=1}^{\bar{K}} k \zeta^{k/2} \leq \bar{K} \sum_{k=0}^{\bar{K}} \zeta^{k/2} \leq \bar{K} \zeta^{(\bar{K}+1)/2} / (\sqrt{\zeta} - 1), \tag{43}$$

$$\zeta^{(\bar{K}+1)/2} \overset{(19)}{\leq} \zeta \max \left\{ 1, \sqrt{\frac{2(D + \rho_0 \eta_0)}{(1 - \zeta\sigma)\epsilon}} \right\}. \tag{44}$$

By these relations, (19), (20), and (42), one has

$$\begin{aligned} \sum_{k=0}^{\bar{K}} n_k = n_0 + \sum_{k=1}^{\bar{K}} n_k &\overset{(42)}{\leq} 16 \sqrt{\frac{\bar{C}}{\mu}} \log \left(\frac{2 \bar{C} D_X \zeta}{\eta_0 \sigma} \right) \left(1 + \sum_{k=1}^{\bar{K}} k \zeta^{\frac{k}{2}} \right) \\ &\overset{(43)}{\leq} \frac{16 \sqrt{\bar{C}}}{\sqrt{\mu}(\sqrt{\zeta} - 1)} \log \left(\frac{2 \bar{C} D_X \zeta}{\eta_0 \sigma} \right) \bar{K} \zeta^{(\bar{K}+1)/2} \overset{(19)(20)(44)}{\leq} \bar{N}, \end{aligned}$$

and hence the conclusion of Theorem 1 holds. □

5.2 Proof of the main results in Section 3

In this subsection, we present a proof of Theorem 2, which establishes the convergence properties of Algorithm 2—an inexact accelerated SVRG method for problem (1).

Before proceeding, we highlight the main technical challenges in analyzing Algorithm 2 compared to the classical (exact) accelerated SVRG method [8]. To this end, we introduce the following notation:

$$\ell_f(u,v) := f(v) + \langle \nabla f(v), u - v\rangle, \qquad \delta_t := g_t - \nabla f(y_t), \tag{45}$$

$$\Omega := \{x \in X : c(x) \in -\mathcal{K}\}, \qquad \Omega_\epsilon := \{x \in X : \mathrm{dist}(c(x), -\mathcal{K}) \le \epsilon\}. \tag{46}$$

In the classical accelerated SVRG method, z_t is assumed to be the exact optimal solution of the subproblem (21) in Algorithm 2. Under this assumption, the following key inequality holds:

$$\gamma_k\big(\ell_f(z_t,y_t) - \ell_f(x,y_t) + \psi(z_t) - \psi(x)\big) \le \frac{1}{2}\|x - z_{t-1}\|^2 - \frac{1}{2}\|x - z_t\|^2 - \frac{1}{2}\|z_{t-1} - z_t\|^2 + \gamma_k\langle \delta_t, x - z_t\rangle \tag{47}$$

for all $1 \le t \le T_k$, $k \ge 1$, and $x \in \Omega$ (see, e.g., [8, Lemma 5]). The analysis of the classical accelerated SVRG method relies crucially on this inequality.

However, in Algorithm 2, z_t is only an approximate solution of the subproblem (21), and thus (47) does not generally hold. Instead, by leveraging the fact that (z_t, λ_t) is an ϵ-KKT pair for (21), together with several technical arguments, we establish the following inequality:

$$\gamma_k\big(\ell_f(z_t,y_t) - \ell_f(x,y_t) + \psi(z_t) - \psi(x)\big) \le \frac{1}{2}\|x - z_{t-1}\|^2 - \frac{1}{2}\|x - z_t\|^2 - \frac{1}{2}\|z_{t-1} - z_t\|^2 + \gamma_k\langle \delta_t, x - z_t\rangle + \mathcal{O}(\gamma_k\epsilon)$$

for all $1 \le t \le T_k$, $k \ge 1$, and $x \in \Omega_\epsilon$, where $\{\gamma_k\}$ is the stepsize sequence of Algorithm 2 (see Lemma 5 for details). Using this inequality, together with an appropriate choice of $\{\gamma_k\}$, we show that the outer loop of Algorithm 2 achieves the same convergence rate as the classical accelerated SVRG method, despite solving its subproblems inexactly.

We now present the proof of Theorem 2. To this end, we begin by establishing several technical lemmas. The following lemma provides a uniform upper bound on the optimal Lagrange multipliers of problem (21).

Lemma 2. *Suppose that Assumption 1 holds. Let Λ be defined in* (11)*, and $\lambda^*_{k,t}$ be an optimal Lagrange multiplier of problem* (21) *that is generated in Algorithm 2 with $\{\gamma_k\}$ and $\{q_i\}$ given in* (28).[5] *Then we have*

$$\|\lambda^*_{k,t}\| \le \Lambda \qquad \forall 1 \le t \le T_k, k \ge 1. \tag{48}$$

Proof. Fix any $k \ge 1$ and $1 \le t \le T_k$. Observe from Algorithm 2 that $y_t, \tilde{x}_k \in X$, and

$$g_t = (\nabla f_{i_t}(y_t) - \nabla f_{i_t}(\tilde{x}_k))/(q_{i_t}s) + \tilde{g}_k.$$

Using these observations, (8), (9), (10), (12), (28), and Assumption 1, we obtain that

$$\begin{aligned}\|g_t\| &= \|(\nabla f_{i_t}(y_t) - \nabla f_{i_t}(\tilde{x}_k))/(q_{i_t}s) + \tilde{g}_k\| \le L_{\nabla f_{i_t}}\|y_t - \tilde{x}_k\|/(q_{i_t}s) + \|\tilde{g}_k\| \\ &\overset{(10)(28)}{=} L_{\nabla f}\|y_t - \tilde{x}_k\| + \|\nabla f(\tilde{x}_k)\| \overset{(8)(12)}{\le} L_{\nabla f}D_X + L_f.\end{aligned} \tag{49}$$

Notice from (28) that $\gamma_k^{-1} = 3L_{\nabla f}\alpha_k \le 3L_{\nabla f}/2$. By this relation, (8), (9), (49), and the definition of $F_{k,t}$

[5] The existence of $\lambda^*_{k,t}$ is guaranteed by the strong convexity of problem (21) and Assumption 1(iv).

in (21), one has that for all $x, y \in X$,

$$\begin{aligned} F_{k,t}(x) - F_{k,t}(y) &\overset{(21)}{=} \frac{1}{2\gamma_k}\big(\|x - z_{t-1} + \gamma_k g_t\|^2 - \|y - z_{t-1} + \gamma_k g_t\|^2\big) + \psi(x) - \psi(y) \\ &= \frac{1}{2\gamma_k}\langle x - y, x - z_{t-1} + y - z_{t-1} + 2\gamma_k g_t\rangle + \psi(x) - \psi(y) \\ &\le \frac{1}{2\gamma_k}\|x - y\|\|x - z_{t-1} + y - z_{t-1} + 2\gamma_k g_t\| + \psi(x) - \psi(y) \\ &\le \frac{1}{2\gamma_k}\|x - y\|\big(\|x - z_{t-1}\| + \|y - z_{t-1}\| + 2\gamma_k\|g_t\|\big) + \psi(x) - \psi(y) \\ &\le \frac{1}{2\gamma_k} D_X\big(2D_X + 2\gamma_k(L_{\nabla f} D_X + L_f)\big) + D_\psi \\ &= D_X^2\gamma_k^{-1} + (L_{\nabla f} D_X + L_f)D_X + D_\psi \le 5L_{\nabla f} D_X^2/2 + L_f D_X + D_\psi, \end{aligned} \tag{50}$$

where the third inequality follows from (8), (9), and (49), and the last inequality is due to $\gamma_k^{-1} \le 3L_{\nabla f}/2$.

Let $x_{k,t}^*$ be the optimal solution of problem (21). Since $\lambda_{k,t}^*$ is an optimal Lagrange multiplier of (21), one has that $\lambda_{k,t}^* \in \mathcal{K}^*$, $\langle \lambda_{k,t}^*, c(x_{k,t}^*)\rangle = 0$, and

$$x_{k,t}^* = \arg\min_x \big\{F_{k,t}(x) + \langle \lambda_{k,t}^*, c(x)\rangle\big\}.$$

Notice from Assumption 1(iv) that $x_{\rm int} \in X$ and $c(x_{\rm int}) + D_c\mathscr{B} \subset -\mathcal{K}$. It then follows that

$$F_{k,t}(x_{\rm int}) + \langle \lambda_{k,t}^*, c(x_{\rm int})\rangle \ge F_{k,t}(x_{k,t}^*) + \langle \lambda_{k,t}^*, c(x_{k,t}^*)\rangle = F_{k,t}(x_{k,t}^*),$$

which together with (50) implies that

$$\langle \lambda_{k,t}^*, -c(x_{\rm int})\rangle \le F_{k,t}(x_{\rm int}) - F_{k,t}(x_{k,t}^*) \le 5L_{\nabla f} D_X^2/2 + L_f D_X + D_\psi. \tag{51}$$

If $\lambda_{k,t}^* = 0$, (48) clearly holds. Now suppose $\lambda_{k,t}^* \neq 0$. By $c(x_{\rm int}) + D_c\mathscr{B} \subset -\mathcal{K}$, one has $-c(x_{\rm int}) - D_c\lambda_{k,t}^*/\|\lambda_{k,t}^*\| \in \mathcal{K}$. Using this and $\lambda_{k,t}^* \in \mathcal{K}^*$, we obtain that

$$\langle \lambda_{k,t}^*, -c(x_{\rm int}) - D_c\lambda_{k,t}^*/\|\lambda_{k,t}^*\|\rangle \ge 0,$$

which implies that $\|\lambda_{k,t}^*\| \le D_c^{-1}\langle \lambda_{k,t}^*, -c(x_{\rm int})\rangle$. Combining this relation with (11) and (51), we conclude that (48) holds. □

The following lemma establishes the convexity of the set Ω_ϵ defined in (46).

Lemma 3. *Let Ω_ϵ be defined in* (46)*. Then Ω_ϵ is convex for any $\epsilon > 0$.*

Proof. Fix any $x, y \in \Omega_\epsilon$ and $\alpha \in [0, 1]$. In view of (2) and the fact that $\mathcal{K}$ is a convex cone, one has

$$\alpha\mathcal{K} + (1-\alpha)\mathcal{K} = \mathcal{K}, \qquad -\mathcal{K} + c(\alpha x + (1-\alpha)y) - \alpha c(x) - (1-\alpha)c(y) \subseteq -\mathcal{K}.$$

Using these relations, we obtain that

$$\begin{aligned} &\alpha\,\mathrm{dist}\big(c(x), -\mathcal{K}\big) + (1-\alpha)\mathrm{dist}\big(c(y), -\mathcal{K}\big) \ge \mathrm{dist}\big(\alpha c(x) + (1-\alpha)c(y), -\mathcal{K}\big) \\ &= \mathrm{dist}\big(c(\alpha x + (1-\alpha)y), -\mathcal{K} + c(\alpha x + (1-\alpha)y) - \alpha c(x) - (1-\alpha)c(y)\big) \\ &\ge \mathrm{dist}\big(c(\alpha x + (1-\alpha)y), -\mathcal{K}\big), \end{aligned}$$

which implies that $\mathrm{dist}(c(\cdot), -\mathcal{K})$ is a convex function. The convexity of Ω_ϵ then follows from this fact and the convexity of X. □

The next lemma establishes some properties of the sequences $\{\tilde{x}_k\}$, $\{\tilde{z}_k\}$, $\{x_t\}$, $\{y_t\}$, $\{z_t\}$, and $\{\lambda_t\}$ generated by Algorithm 2, which will be used in the subsequent analysis.

Lemma 4. *Suppose that Assumption 1 holds. Let D_X, Λ, Ω_ϵ, ρ_0, η_0, ζ, σ, R be given in* (8), (11), (46)*, and Algorithms 1 and 2, respectively, and $\{\tilde{x}_k\}$, $\{\tilde{z}_k\}$, $\{x_t\}$, $\{y_t\}$, $\{z_t\}$, and $\{\lambda_t\}$ be generated in Algorithm 2. Then the following statements hold.*

(i) $\tilde{x}_k, \tilde{z}_k \in \Omega_\epsilon$ *for all* $k \geq 1$.

(ii) $x_t, z_t \in \Omega_\epsilon$ *for all* $0 \leq t \leq T_k$ *and* $k \geq 1$, *and* $y_t \in \Omega_\epsilon$ *for all* $1 \leq t \leq T_k$ *and* $k \geq 1$.

(iii) $\|(x_t, \lambda_t) - (x^*_{k,t}, \lambda^*_{k,t})\| \leq D_X + R + \Lambda + \rho_0\eta_0/(1-\zeta\sigma)$ *for all* $1 \leq t \leq T_k$ *and* $k \geq 1$, *where* $(x^*_{k,t}, \lambda^*_{k,t})$ *is an arbitrary pair of optimal solution and Lagrange multiplier of problem* (21).

Proof. For all $1 \leq t \leq T_k$ and $k \geq 1$, it follows from Theorem 1 that (z_t, λ_t) is a pair of ϵ-KKT solution of problem (21). By this fact and the definition of ϵ-KKT solution in (7), one can see that $z_t \in X$, $\lambda_t \in \mathcal{K}^*$, and $\mathrm{dist}(c(z_t), \mathcal{N}_{\mathcal{K}^*}(\lambda_t)) \leq \epsilon$. Hence, there exists $u \in \mathcal{N}_{\mathcal{K}^*}(\lambda_t)$ such that $\|c(z_t) - u\| \leq \epsilon$. By $\lambda_t \in \mathcal{K}^*$, $u \in \mathcal{N}_{\mathcal{K}^*}(\lambda_t)$, and the fact that $\mathcal{K}^*$ is a cone, one can observe that $u^T\lambda_t = 0$ and $u^T(v - \lambda_t) \leq 0$ for all $v \in \mathcal{K}^*$. It follows that $u^T v \leq 0$ for all $v \in \mathcal{K}^*$, which leads to $u \in -\mathcal{K}$. Using this and $\|c(z_t) - u\| \leq \epsilon$, we see that $\mathrm{dist}(c(z_t), -\mathcal{K}) \leq \epsilon$, which together with $z_t \in X$ implies that $z_t \in \Omega_\epsilon$ for all $1 \leq t \leq T_k$ and $k \geq 1$. In addition, notice from Algorithm 2 that $x^0 \in \Omega_\epsilon$ and $\tilde{x}_1 = \tilde{z}_1 = x^0$. Hence, $\tilde{x}_1, \tilde{z}_1 \in \Omega_\epsilon$. Also, for each $k \geq 1$, $x_0 = \tilde{x}_k$, $z_0 = \tilde{z}_k$, $\tilde{z}_{k+1} = z_{T_k}$, $\tilde{x}_{k+1}$ is a convex combination of $x_1, \ldots, x_{T_k}$, y_t is a convex combination of $x_{t-1}, z_{t-1}, \tilde{x}_k$, and x_t is a convex combination of x_{t-1}, z_t, and $\tilde{x}_k$ for all $1 \leq t \leq T_k$. Based on these facts, the convexity of Ω_ϵ, and an inductive argument, one can readily verify that statements (i) and (ii) hold.

We next show that statement (iii) holds. Indeed, since (x_t, λ_t) is generated by Algorithm 1 applied to problem (21), it follows from (18) that

$$\|(x_t, \lambda_t) - (x^*_{k,t}, \lambda^*_{k,t})\| \leq \|(x_0, \lambda_0) - (x^*_{k,t}, \lambda^*_{k,t})\| + \sum_{i=0}^{\infty} \rho_i\eta_i \qquad \forall 0 \leq t \leq T_k, k \geq 1. \tag{52}$$

Notice from Algorithm 2 that $x_0, x^*_{k,t} \in X$, $\|\lambda_0\| \leq R$, $\rho_i = \rho_0\zeta^i$, and $\eta_i = \eta_0\sigma^i$ for all $i \geq 0$. Also, recall from Lemma 2 that $\|\lambda^*_{k,t}\| \leq \Lambda$. In view of these facts, Assumption 1, and $0 < \zeta\sigma < 1$, one has

$$\begin{aligned}\|(x_0, \lambda_0) - (x^*_{k,t}, \lambda^*_{k,t})\| + \sum_{i=0}^{\infty} \rho_i\eta_i &\leq \|x_0 - x^*_{k,t}\| + \|\lambda_0\| + \|\lambda^*_{k,t}\| + \rho_0\eta_0 \sum_{i=0}^{\infty} (\zeta\sigma)^i \\ &\leq D_X + R + \Lambda + \rho_0\eta_0/(1-\zeta\sigma),\end{aligned}$$

which together with (52) implies that statement (iii) holds as desired. □

The following lemma establishes a key property of the inexact solution z_t to problem (21) generated in Algorithm 2.

Lemma 5. *Suppose that Assumption 1 holds. Let f, C_5, ℓ_f, δ_t, Ω_ϵ, and ϵ be given in* (10), (27), (45), (46)*, and Algorithm 2, respectively, and let $\{g_t\}$, $\{y_t\}$, $\{z_t\}$ be generated in the kth outer iteration of Algorithm 2 with $\{\gamma_k\}$ and $\{q_i\}$ given in* (28). *Then, for all $1 \leq t \leq T_k$, $k \geq 1$, and $x \in \Omega_\epsilon$, we have*

$$\gamma_k\big(\ell_f(z_t, y_t) - \ell_f(x, y_t) + \psi(z_t) - \psi(x)\big) \leq \frac{1}{2}\|x - z_{t-1}\|^2 - \frac{1}{2}\|x - z_t\|^2 - \frac{1}{2}\|z_{t-1} - z_t\|^2 + \gamma_k\langle\delta_t, x - z_t\rangle + \gamma_k C_5\epsilon. \tag{53}$$

Proof. Fix any $k \geq 1$. Notice from Algorithm 2 that (z_t, λ_t) is a pair of ϵ-KKT solution of problem (21). By this fact and the definition of ϵ-KKT solution in (7), one has that $z_t \in X$, $\lambda_t \in \mathcal{K}^*$, and

$$\mathrm{dist}\big(0, (z_t - z_{t-1} + \gamma_k g_t)/\gamma_k + \nabla c(z_t)\lambda_t + \partial\psi(z_t)\big) \leq \epsilon, \quad \mathrm{dist}\big(c(z_t), \mathcal{N}_{\mathcal{K}^*}(\lambda_t)\big) \leq \epsilon. \tag{54}$$

Hence, there exists $u \in \mathcal{N}_{\mathcal{K}^*}(\lambda_t)$ such that $\|c(z_t) - u\| \leq \epsilon$. By $\lambda_t \in \mathcal{K}^*$, $u \in \mathcal{N}_{\mathcal{K}^*}(\lambda_t)$, and the fact that $\mathcal{K}^*$ is a cone, one can observe that $\langle\lambda_t, u\rangle = 0$. In addition, let $\lambda^*_{k,t}$ be an optimal Lagrange multiplier of problem (21). Invoking Lemmas 2 and 4(iii), we have

$$\|\lambda^*_{k,t}\| \leq \Lambda, \qquad \|\lambda_t - \lambda^*_{k,t}\| \leq D_X + R + \Lambda + \rho_0\eta_0/(1-\zeta\sigma),$$

which imply that

$$\|\lambda_t\| \leq \|\lambda_{k,t}^*\| + \|\lambda_t - \lambda_{k,t}^*\| \leq D_X + R + 2\Lambda + \rho_0\eta_0/(1-\zeta\sigma). \tag{55}$$

Using this relation, $\langle \lambda_t, u\rangle = 0$, and $\|c(z_t) - u\| \leq \epsilon$, we obtain that

$$\langle \lambda_t, c(z_t)\rangle = \langle \lambda_t, c(z_t) - u\rangle \geq -\epsilon\|\lambda_t\| \geq -\epsilon\big(D_X + R + 2\Lambda + \rho_0\eta_0/(1-\zeta\sigma)\big). \tag{56}$$

Similarly, for any $x \in \Omega_\epsilon$, it follows from the second relation in (46) that there exists $\tilde{u} \in -\mathcal{K}$ such that $\|c(x) - \tilde{u}\| \leq \epsilon$. This together with $\lambda_t \in \mathcal{K}^*$ and (55) implies that $\langle \lambda_t, \tilde{u}\rangle \leq 0$, and hence

$$\langle \lambda_t, c(x)\rangle \leq \langle \lambda_t, c(x) - \tilde{u}\rangle \leq \epsilon\|\lambda_t\| \overset{(55)}{\leq} \epsilon\big(D_X + R + 2\Lambda + \rho_0\eta_0/(1-\zeta\sigma)\big) \qquad \forall x \in \Omega_\epsilon. \tag{57}$$

Also, by the first relation in (54), there exists a vector $\psi'(z_t) \in \partial\psi(z_t)$ such that

$$\|(z_t - z_{t-1} + \gamma_k g_t)/\gamma_k + \nabla c(z_t)\lambda_t + \psi'(z_t)\| \leq \epsilon. \tag{58}$$

In addition, by the $\mathcal{K}$-convexity of $c(\cdot)$ and $\lambda_t \in \mathcal{K}^*$, one has

$$\langle \lambda_t, \alpha c(x) + (1-\alpha)c(y) - c(\alpha x + (1-\alpha)y)\rangle \geq 0 \qquad \forall x, y \in X, \alpha \in [0,1],$$

which implies that the function $\langle \lambda_t, c(\cdot)\rangle$ is convex. It then follows that the function $\|x - z_{t-1} + \gamma_k g_t\|^2/(2\gamma_k) + \langle \lambda_t, c(x)\rangle + \psi(x)$ is strongly convex with modulus γ_k^{-1}. Using this fact, (58), and Assumption 1, we obtain that for all $x \in \Omega_\epsilon$,

$$\begin{aligned}
&\frac{\|x - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \langle \lambda_t, c(x)\rangle + \psi(x) \geq \frac{\|z_t - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \langle \lambda_t, c(z_t)\rangle + \psi(z_t) \\
&\quad + \langle (z_t - z_{t-1} + \gamma_k g_t)/\gamma_k + \nabla c(z_t)\lambda_t + \psi'(z_t), x - z_t\rangle + \frac{1}{2\gamma_k}\|x - z_t\|^2 \\
&\overset{(58)}{\geq} \frac{\|z_t - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \langle \lambda_t, c(z_t)\rangle + \psi(z_t) - \|x - z_t\|\epsilon + \frac{1}{2\gamma_k}\|x - z_t\|^2 \\
&\geq \frac{\|z_t - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \langle \lambda_t, c(z_t)\rangle + \psi(z_t) - D_X\epsilon + \frac{1}{2\gamma_k}\|x - z_t\|^2,
\end{aligned}$$

where the last inequality follows from the fact that $\|x - z_t\| \leq D_X$ for all $x \in \Omega_\epsilon$ due to Assumption 1. Combining the above inequality with (56) and (57) leads to

$$\begin{aligned}
\frac{\|x - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \psi(x) \geq \frac{\|z_t - z_{t-1} + \gamma_k g_t\|^2}{2\gamma_k} + \psi(z_t) + \frac{1}{2\gamma_k}\|x - z_t\|^2 \\
- \big(3D_X + 2R + 4\Lambda + 2\rho_0\eta_0/(1-\zeta\sigma)\big)\epsilon \qquad \forall x \in \Omega_\epsilon.
\end{aligned} \tag{59}$$

Moreover, by (45), one has

$$\begin{aligned}
&\|z_t - z_{t-1} + \gamma_k g_t\|^2 - \|x - z_{t-1} + \gamma_k g_t\|^2 = \|z_t - z_{t-1}\|^2 - \|x - z_{t-1}\|^2 + 2\gamma_k\langle g_t, z_t - x\rangle \\
&\overset{(45)}{=} \|z_t - z_{t-1}\|^2 - \|x - z_{t-1}\|^2 + 2\gamma_k\langle \delta_t, z_t - x\rangle + 2\gamma_k\langle \nabla f(y_t), z_t - x\rangle \\
&\overset{(45)}{=} \|z_t - z_{t-1}\|^2 - \|x - z_{t-1}\|^2 + 2\gamma_k\langle \delta_t, z_t - x\rangle + 2\gamma_k\big(\ell_f(z_t, y_t) - \ell_f(x, y_t)\big).
\end{aligned}$$

Substituting this relation into (59) and using (27), we obtain (53), which completes the proof of the lemma. □

The next lemma, established in [8, Lemma 3], shows that g_t is an unbiased stochastic estimator of $\nabla f(y_t)$ and provides an upper bound on its variance.

Lemma 6 ([**8, Lemma 3**]). *Suppose that Assumption 1 holds. Let f, $L_{\nabla f}$ and δ_t be defined in* (10) *and* (45)*, respectively, and $\tilde{x}_k$ and $\{y_t\}$ be generated in the kth outer iteration of Algorithm 2 with $\{q_i\}$ given in* (28)*. Then, for all $1 \le t \le T_k$ and $k \ge 1$, we have*

$$\mathbb{E}[\delta_t|\Xi_{k,t-1}] = 0, \qquad \mathbb{E}[\|\delta_t\|^2|\Xi_{k,t-1}] \le 2L_{\nabla f}\big(f(\tilde{x}_k) - f(y_t) - \langle \nabla f(y_t), \tilde{x}_k - y_t\rangle\big), \tag{60}$$

where $\Xi_{k,t-1}$ denotes the collection of all samples generated by Algorithm 2 prior to the t-th inner iteration of the kth outer iteration.

The following lemma establishes a recursion for $\{\mathbb{E}[F(x_t) - F(x)]\}$ for all $1 \le t \le T_k$, $k \ge 1$, and $x \in \Omega_\epsilon$.

Lemma 7. *Suppose that Assumption 1 holds. Let $L_{\nabla f}$, C_5, Ω_ϵ, ϵ be given in* (10)*,* (27)*,* (46)*, and Algorithm 2, respectively, and let $\{x_t\}$, $\{y_t\}$, $\{z_t\}$, $\{g_t\}$, and $\tilde{x}_k$ be generated in the kth outer iteration of Algorithm 2 with $\{q_i\}$ and γ_k given in* (28)*, and α_k, γ_k, and p_k satisfying*

$$1 - L_{\nabla f}\alpha_k\gamma_k > 0, \tag{61}$$

$$p_k - \frac{L_{\nabla f}\alpha_k\gamma_k}{1 - L_{\nabla f}\alpha_k\gamma_k} \ge 0 \tag{62}$$

for all $k \ge 1$. Then, for all $1 \le t \le T_k$, $k \ge 1$, and $x \in \Omega_\epsilon$, we have

$$\begin{aligned}\frac{\gamma_k}{\alpha_k}\mathbb{E}[F(x_t) - F(x)] + \frac{1}{2}\mathbb{E}[\|x - z_t\|^2] \le \frac{\gamma_k}{\alpha_k}(1 - \alpha_k - p_k)\mathbb{E}[F(x_{t-1}) - F(x)] + \frac{\gamma_k p_k}{\alpha_k}\mathbb{E}[F(\tilde{x}_k) - F(x)] \\ + \frac{1}{2}\mathbb{E}[\|x - z_{t-1}\|^2] + \gamma_k C_5\epsilon.\end{aligned} \tag{63}$$

Proof. Fix any $k \ge 1$. Let $\{x_t\}$, $\{y_t\}$, and $\{z_t\}$ be generated in the kth outer iteration of Algorithm 2 for all $1 \le t \le T_k$. Then it follows from the expressions of x_t and y_t in Algorithm 2 that $x_t - y_t = \alpha_k(z_t - z_{t-1})$. By this relation, (53), the $L_{\nabla f}$-smoothness and the convexity of f, and the expressions of x_t and $\ell_f(\cdot,\cdot)$, one has

$$\begin{aligned}
f(x_t) &\le \ell_f(x_t, y_t) + \frac{L_{\nabla f}}{2}\|x_t - y_t\|^2 \\
&= (1 - \alpha_k - p_k)\ell_f(x_{t-1}, y_t) + \alpha_k\ell_f(z_t, y_t) + p_k\ell_f(\tilde{x}_k, y_t) + \frac{L_{\nabla f}\alpha_k^2}{2}\|z_t - z_{t-1}\|^2 \\
&\overset{(53)}{\le} (1 - \alpha_k - p_k)\ell_f(x_{t-1}, y_t) + \alpha_k\Big(\ell_f(x, y_t) + \psi(x) - \psi(z_t) + \frac{1}{2\gamma_k}\big(\|x - z_{t-1}\|^2 - \|x - z_t\|^2 \\
&\qquad - \|z_{t-1} - z_t\|^2\big) + \langle \delta_t, x - z_t\rangle + C_5\epsilon\Big) + p_k\ell_f(\tilde{x}_k, y_t) + \frac{L_{\nabla f}\alpha_k^2}{2}\|z_t - z_{t-1}\|^2 \\
&= (1 - \alpha_k - p_k)\ell_f(x_{t-1}, y_t) + \alpha_k\Big(\ell_f(x, y_t) + \psi(x) - \psi(z_t) + \frac{1}{2\gamma_k}\|x - z_{t-1}\|^2 - \frac{1}{2\gamma_k}\|x - z_t\|^2\Big) \\
&\quad + p_k\ell_f(\tilde{x}_k, y_t) - \frac{\alpha_k}{2\gamma_k}(1 - L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1} - z_t\|^2 + \alpha_k\langle \delta_t, x - z_t\rangle + \alpha_k C_5\epsilon \\
&\le (1 - \alpha_k - p_k)f(x_{t-1}) + \alpha_k\Big(F(x) - \psi(z_t) + \frac{1}{2\gamma_k}\|x - z_{t-1}\|^2 - \frac{1}{2\gamma_k}\|x - z_t\|^2\Big) \\
&\quad + p_k\ell_f(\tilde{x}_k, y_t) - \frac{\alpha_k}{2\gamma_k}(1 - L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1} - z_t\|^2 + \alpha_k\langle \delta_t, x - z_t\rangle + \alpha_k C_5\epsilon,
\end{aligned}$$

where the first relation follows from the $L_{\nabla f}$-smoothness of f, the second relation is due to the definition of $\ell_f(\cdot,\cdot)$, and the last relation follows from the convexity of f and the expression of F. Also, by the expression of x_t and the convexity of ψ, one has

$$\psi(x_t) \le (1 - \alpha_k - p_k)\psi(x_{t-1}) + \alpha_k\psi(z_t) + p_k\psi(\tilde{x}_k).$$

Summing up these two inequalities and using the expression of F, we have

$$F(x_t) \leq (1-\alpha_k-p_k)F(x_{t-1}) + \alpha_k\Big(F(x) + \frac{1}{2\gamma_k}\|x-z_{t-1}\|^2 - \frac{1}{2\gamma_k}\|x-z_t\|^2\Big) + p_k\ell_f(\tilde{x}_k, y_t)$$
$$- \frac{\alpha_k}{2\gamma_k}(1-L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1}-z_t\|^2 + \alpha_k\langle\delta_t, x-z_t\rangle + p_k\psi(\tilde{x}_k) + \alpha_k C_5\epsilon. \tag{64}$$

In addition, observe that

$$\begin{aligned} &p_k\ell_f(\tilde{x}_k, y_t) - \frac{\alpha_k}{2\gamma_k}(1-L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1}-z_t\|^2 + \alpha_k\langle\delta_t, x-z_t\rangle \\ &= p_k\ell_f(\tilde{x}_k, y_t) - \frac{\alpha_k}{2\gamma_k}(1-L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1}-z_t\|^2 + \alpha_k\langle\delta_t, z_{t-1}-z_t\rangle + \alpha_k\langle\delta_t, x-z_{t-1}\rangle \\ &\leq p_k\ell_f(\tilde{x}_k, y_t) + \frac{\alpha_k\gamma_k\|\delta_t\|^2}{2(1-L_{\nabla f}\alpha_k\gamma_k)} + \alpha_k\langle\delta_t, x-z_{t-1}\rangle, \end{aligned} \tag{65}$$

where the last relation follows from (61) and the Young's inequality. Using (60), (62), the definition of $\ell_f(\cdot,\cdot)$, and the convexity of f, we obtain that

$$\begin{aligned} &p_k\ell_f(\tilde{x}_k, y_t) + \frac{\alpha_k\gamma_k\mathbb{E}[\|\delta_t\|^2|\Xi_{k,t-1}]}{2(1-L_{\nabla f}\alpha_k\gamma_k)} + \alpha_k\mathbb{E}[\langle\delta_t, x-z_{t-1}\rangle|\Xi_{k,t-1}] \\ &\overset{(60)}{\leq} p_k\ell_f(\tilde{x}_k, y_t) + \frac{L_{\nabla f}\alpha_k\gamma_k}{1-L_{\nabla f}\alpha_k\gamma_k}\left(f(\tilde{x}_k) - \ell_f(\tilde{x}_k, y_t)\right) \\ &= \Big(p_k - \frac{L_{\nabla f}\alpha_k\gamma_k}{1-L_{\nabla f}\alpha_k\gamma_k}\Big)\ell_f(\tilde{x}_k, y_t) + \frac{L_{\nabla f}\alpha_k\gamma_k}{1-L_{\nabla f}\alpha_k\gamma_k}f(\tilde{x}_k) \leq p_k f(\tilde{x}_k), \end{aligned} \tag{66}$$

where the last inequality follows from (62) and the convexity of f. It then follows from (65) and (66) that

$$\begin{aligned} &\mathbb{E}\Big[p_k\ell_f(\tilde{x}_k, y_t) - \frac{\alpha_k}{2\gamma_k}(1-L_{\nabla f}\alpha_k\gamma_k)\|z_{t-1}-z_t\|^2 + \alpha_k\langle\delta_t, x-z_t\rangle\Big] \\ &\overset{(65)}{\leq} \mathbb{E}\Big[p_k\ell_f(\tilde{x}_k, y_t) + \frac{\alpha_k\gamma_k\|\delta_t\|^2}{2(1-L_{\nabla f}\alpha_k\gamma_k)} + \alpha_k\langle\delta_t, x-z_{t-1}\rangle\Big] \overset{(66)}{\leq} p_k\mathbb{E}[f(\tilde{x}_k)]. \end{aligned}$$

Using this inequality and taking expectation on both sides of (64), we obtain that for all $1 \leq t \leq T_k$,

$$\mathbb{E}[F(x_t)] \leq (1-\alpha_k-p_k)\mathbb{E}[F(x_{t-1})] + \alpha_k\Big(F(x) + \frac{1}{2\gamma_k}\mathbb{E}[\|x-z_{t-1}\|^2] - \frac{1}{2\gamma_k}\mathbb{E}[\|x-z_t\|^2]\Big) + p_k\mathbb{E}[F(\tilde{x}_k)] + \alpha_k C_5\epsilon.$$

The relation (63) then follows from this inequality by rearranging the terms. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. We first show that the first two relations in (29) hold. Indeed, it follows from (46) and Lemma 4(i) that $\mathrm{dist}(c(\tilde{x}_k), -\mathcal{K}) \leq \epsilon$. Hence, there exists $u \in -\mathcal{K}$ such that $\|c(\tilde{x}_k) - u\| \leq \epsilon$. Let λ^* be the minimum-norm optimal Lagrange multiplier of problem (1). Then, one has that $\lambda^* \in \mathcal{K}^*$ and

$$F^* = \min_x F(x) + \langle\lambda^*, c(x)\rangle.$$

Using these facts, $u \in -\mathcal{K}$, and $\|c(\tilde{x}_k) - u\| \leq \epsilon$, we obtain that

$$\langle\lambda^*, c(\tilde{x}_k)\rangle \leq \langle\lambda^*, c(\tilde{x}_k) - u\rangle \leq \|\lambda^*\|\epsilon, \qquad F(\tilde{x}_k) + \langle\lambda^*, c(\tilde{x}_k)\rangle \geq F^*.$$

In addition, by the definition of D_λ in this theorem, one has $D_\lambda = \|\lambda^*\|$. Using these relations, we have

$$F(\tilde{x}_k) \geq F^* - \langle\lambda^*, c(\tilde{x}_k)\rangle \geq F^* - \|\lambda^*\|\epsilon = F^* - D_\lambda\epsilon.$$

Hence, the first two relations in (29) hold as desired.

We next show that the third relation in (29) holds. To this end, let $\tilde{x}^* \in \arg\min_{x\in\Omega_\epsilon} F(x)$. One can observe from (1) and (46) that $F^* = \min_{x\in\Omega} F(x)$ and $\Omega \subseteq \Omega_\epsilon$. It then follows that $F(\tilde{x}^*) \leq F^*$. In addition, one can verify that $\{q_i\}$, $\{\alpha_k\}$, $\{\gamma_k\}$, and $\{p_k\}$ that are given in (28) satisfy the assumptions stated in Lemma 7. Hence, Lemma 7 holds for such parameters. It then follows from (63) with $x = \tilde{x}^*$ that for all $k \geq 1$,

$$\begin{aligned}\frac{\gamma_k}{\alpha_k}\mathbb{E}[F(x_t) - F(\tilde{x}^*)] \leq \frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(x_{t-1}) - F(\tilde{x}^*)] + \frac{\gamma_k p_k}{\alpha_k}\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)]\\ + \frac{1}{2}\big(\mathbb{E}[\|z_{t-1}-\tilde{x}^*\|^2] - \mathbb{E}[\|z_t - \tilde{x}^*\|^2]\big) + \gamma_k C_5\epsilon \qquad \forall 1 \leq t \leq T_k. \end{aligned} \tag{67}$$

Summing up (67) from $t = 1$ to T_k, and using the fact that $x_0 = \tilde{x}_k$, $z_0 = \tilde{z}_k$, $z_{T_k} = \tilde{z}_{k+1}$, we have

$$\begin{aligned}\sum_{t=1}^{T_k}\frac{\gamma_k}{\alpha_k}\mathbb{E}[F(x_t) - F(\tilde{x}^*)] &\leq \frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(x_0) - F(\tilde{x}^*)] + \sum_{t=1}^{T_k-1}\frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(x_t) - F(\tilde{x}^*)]\\ &\quad + \frac{T_k\gamma_k p_k}{\alpha_k}\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] + \frac{1}{2}\big(\mathbb{E}[\|z_0-\tilde{x}^*\|^2] - \mathbb{E}[\|z_{T_k}-\tilde{x}^*\|^2]\big) + \gamma_k T_k C_5\epsilon\\ &= \frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] + \sum_{t=1}^{T_k-1}\frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(x_t) - F(\tilde{x}^*)]\\ &\quad + \frac{T_k\gamma_k p_k}{\alpha_k}\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] + \frac{1}{2}\big(\mathbb{E}[\|\tilde{z}_k-\tilde{x}^*\|^2] - \mathbb{E}[\|\tilde{z}_{k+1}-\tilde{x}^*\|^2]\big) + \gamma_k T_k C_5\epsilon. \end{aligned} \tag{68}$$

Let $\mathcal{Q}_k$ and $\mathcal{R}_k$ be defined as

$$\mathcal{Q}_k := \frac{\gamma_k}{\alpha_k} + (T_k-1)\frac{\gamma_k(\alpha_k+p_k)}{\alpha_k}, \qquad \mathcal{R}_k := \frac{\gamma_k}{\alpha_k}(1-\alpha_k) + (T_k-1)\frac{\gamma_k p_k}{\alpha_k} \qquad \forall k \geq 1, \tag{69}$$

where α_k, γ_k, p_k, and T_k are given in (28). Notice from (69), the definition of θ_t in (28), and Algorithm 2 that $\mathcal{Q}_k = \sum_{t=1}^{T_k}\theta_t$ and $\tilde{x}_{k+1} = \sum_{t=1}^{T_k}(\theta_t x_t)/\sum_{t=1}^{T_k}\theta_t$. Using these, (28), (68), and the convexity of F, we have

$$\begin{aligned}\mathcal{Q}_k\mathbb{E}[F(\tilde{x}_{k+1}) - F(\tilde{x}^*)] &= \Big(\sum_{t=1}^{T_k}\theta_t\Big)\mathbb{E}[F(\tilde{x}_{k+1}) - F(\tilde{x}^*)] \leq \sum_{t=1}^{T_k}\theta_t\mathbb{E}[F(x_t) - F(\tilde{x}^*)]\\ &\overset{(28)}{=} \sum_{t=1}^{T_k}\frac{\gamma_k}{\alpha_k}\mathbb{E}[F(x_t) - F(\tilde{x}^*)] - \sum_{t=1}^{T_k-1}\frac{\gamma_k}{\alpha_k}(1-\alpha_k-p_k)\mathbb{E}[F(x_t) - F(\tilde{x}^*)]\\ &\overset{(68)}{\leq} \Big(\frac{\gamma_k}{\alpha_k}(1-\alpha_k) + (T_k-1)\frac{\gamma_k p_k}{\alpha_k}\Big)\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] + \frac{1}{2}\big(\mathbb{E}[\|\tilde{z}_k-\tilde{x}^*\|^2] - \mathbb{E}[\|\tilde{z}_{k+1}-\tilde{x}^*\|^2]\big) + \gamma_k T_k C_5\epsilon\\ &\overset{(69)}{=} \mathcal{R}_k\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] + \frac{1}{2}\big(\mathbb{E}[\|\tilde{z}_k-\tilde{x}^*\|^2] - \mathbb{E}[\|\tilde{z}_{k+1}-\tilde{x}^*\|^2]\big) + \gamma_k T_k C_5\epsilon. \end{aligned} \tag{70}$$

We claim that $\mathcal{Q}_k - \mathcal{R}_{k+1} \geq 0$ for all $k \geq 1$. Indeed, notice from (28) that if $1 \leq k < k_0$, $\alpha_{k+1} = \alpha_k = 1/2$, $p_k = 1/2$, $\gamma_{k+1} = \gamma_k$, and $T_{k+1} = 2T_k$. Using these relations and (69), we have that for all $1 \leq k < k_0$,

$$\begin{aligned}\mathcal{Q}_k - \mathcal{R}_{k+1} &\overset{(69)}{=} \frac{\gamma_k}{\alpha_k} + (T_k-1)\frac{\gamma_k(\alpha_k+p_k)}{\alpha_k} - \frac{\gamma_{k+1}}{\alpha_{k+1}}(1-\alpha_{k+1}) - (T_{k+1}-1)\frac{\gamma_{k+1}p_{k+1}}{\alpha_{k+1}}\\ &= \frac{\gamma_k}{\alpha_k}\Big(1 + (T_k-1)(\alpha_k+p_k) - (1-\alpha_k) - (2T_k-1)p_k\Big) = 0, \end{aligned}$$

and hence the above claim holds for all $1 \le k < k_0$. On the other hand, if $k \ge k_0$, notice from $k_0 \ge 1$ and (28) that $T_k = T_{k+1} = T_{k_0} \ge 1$, $\alpha_k = 2/(k-k_0+4)$, $p_k = 1/2$, and $\gamma_k = 1/(3L_{\nabla f}\alpha_k) = (k-k_0+4)/(6L_{\nabla f})$. Using these relations and (69), we obtain that for all $k \ge k_0$,

$$\begin{aligned}
\mathcal{Q}_k - \mathcal{R}_{k+1} &\overset{(69)}{=} \frac{\gamma_k}{\alpha_k} + (T_k - 1)\frac{\gamma_k(\alpha_k + p_k)}{\alpha_k} - \frac{\gamma_{k+1}}{\alpha_{k+1}}(1-\alpha_{k+1}) - (T_{k+1}-1)\frac{\gamma_{k+1}p_{k+1}}{\alpha_{k+1}} \\
&= \frac{\gamma_k}{\alpha_k} + (T_{k_0} - 1)\frac{\gamma_k(\alpha_k + p_k)}{\alpha_k} - \frac{\gamma_{k+1}}{\alpha_{k+1}}(1-\alpha_{k+1}) - (T_{k_0}-1)\frac{\gamma_{k+1}p_{k+1}}{\alpha_{k+1}} \\
&= \frac{\gamma_k}{\alpha_k} - \frac{\gamma_{k+1}}{\alpha_{k+1}}(1-\alpha_{k+1}) + (T_{k_0}-1)\Big(\frac{\gamma_k(\alpha_k+p_k)}{\alpha_k} - \frac{\gamma_{k+1}p_{k+1}}{\alpha_{k+1}}\Big) \\
&= \frac{(k-k_0+4)^2}{12L_{\nabla f}} - \frac{(k-k_0+5)^2}{12L_{\nabla f}}\Big(1 - \frac{2}{k-k_0+5}\Big) \\
&\quad + (T_{k_0}-1)\Big[\frac{(k-k_0+4)^2}{12L_{\nabla f}}\Big(\frac{2}{k-k_0+4} + \frac{1}{2}\Big) - \frac{(k-k_0+5)^2}{24L_{\nabla f}}\Big] \\
&= \frac{(k-k_0+4)^2 - (k-k_0+5)^2 + 2(k-k_0+5)}{12L_{\nabla f}} \\
&\quad + \frac{(T_{k_0}-1)\big[4(k-k_0+4) + (k-k_0+4)^2 - (k-k_0+5)^2\big]}{24L_{\nabla f}} \\
&= \frac{1}{12L_{\nabla f}} + \frac{(T_{k_0}-1)(2(k-k_0)+7)}{24L_{\nabla f}} \ge 0,
\end{aligned}$$

where the second equality is due to $T_k = T_{k+1} = T_{k_0}$, and the fourth equality follows from $\alpha_k = 2/(k-k_0+4)$, $p_k = 1/2$, and $\gamma_k = (k-k_0+4)/(6L_{\nabla f})$. Hence, the above claim also holds for all $k \ge k_0$.

Recall from Lemma 4(i) that $\tilde{x}_k \in \Omega_\epsilon$ for all $k \ge 1$. It follows from this fact and the definition $\tilde{x}^*$ that $F(\tilde{x}_k) - F(\tilde{x}^*) \ge 0$. Also, note from (28) and (69) that $\mathcal{R}_1 = 2/(3L_{\nabla f})$. Summing up (70) with k replaced by i from $i = 1$ to $k-1$, and using the fact that $F(\tilde{x}_i) - F(\tilde{x}^*) \ge 0$, $\mathcal{Q}_i - \mathcal{R}_{i+1} \ge 0$ for all $i \ge 1$, $\tilde{x}_1 = \tilde{z}_1 = x^0$, and $\mathcal{R}_1 = 2/(3L_{\nabla f})$, we have

$$\begin{aligned}
\mathcal{Q}_{k-1}\mathbb{E}[F(\tilde{x}_k) - F(\tilde{x}^*)] &\le \sum_{i=1}^{k-2}(\mathcal{R}_{i+1} - \mathcal{Q}_i)\mathbb{E}[F(\tilde{x}_{i+1}) - F(\tilde{x}^*)] + \mathcal{R}_1(F(\tilde{x}_1) - F(\tilde{x}^*)) \\
&\quad + \frac{1}{2}\|\tilde{z}_1 - \tilde{x}^*\|^2 + C_5\epsilon\sum_{i=1}^{k-1}\gamma_i T_i \\
&\le \frac{2}{3L_{\nabla f}}(F(x^0) - F(\tilde{x}^*)) + \frac{1}{2}\|x^0 - \tilde{x}^*\|^2 + C_5\epsilon\sum_{i=1}^{k-1}\gamma_i T_i. \qquad (71)
\end{aligned}$$

In addition, recall that $k_0 = \lfloor \log_2 s \rfloor + 1$ and $T_{k_0} = 2^{k_0-1}$, which imply that $s/2 \le T_{k_0} \le s$. By this relation, (28), and (69), one has that for all $k > k_0$,

$$\begin{aligned}
\mathcal{Q}_{k-1} &\overset{(69)}{=} \frac{\gamma_{k-1}}{\alpha_{k-1}}\big(1 + (T_{k-1}-1)(\alpha_{k-1} + p_{k-1})\big) \\
&\overset{(28)}{=} \frac{1}{3L_{\nabla f}\alpha_{k-1}^2}\left(1 + (T_{k-1}-1)\Big(\alpha_{k-1} + \frac{1}{2}\Big)\right) = \frac{T_{k-1}-1}{3L_{\nabla f}\alpha_{k-1}} + \frac{T_{k-1}+1}{6L_{\nabla f}\alpha_{k-1}^2} \\
&\overset{(28)}{=} \frac{(T_{k_0}-1)(k-k_0+3)}{6L_{\nabla f}} + \frac{(T_{k_0}+1)(k-k_0+3)^2}{24L_{\nabla f}} \\
&\ge \frac{(s-2)(k-k_0+3)}{12L_{\nabla f}} + \frac{(s+2)(k-k_0+3)^2}{48L_{\nabla f}} \\
&= \frac{(s-2)(k-k_0+3)}{12L_{\nabla f}} + \frac{2(k-k_0+3)^2}{48L_{\nabla f}} + \frac{s(k-k_0+3)^2}{48L_{\nabla f}}
\end{aligned}$$

$$\geq \frac{(s-2)(k-k_0+3)}{12L_{\nabla f}} + \frac{8(k-k_0+3)}{48L_{\nabla f}} + \frac{s(k-k_0+3)^2}{48L_{\nabla f}}$$
$$= \frac{s(k-k_0+3)}{12L_{\nabla f}} + \frac{s(k-k_0+3)^2}{48L_{\nabla f}} \geq \frac{s(k-k_0+3)^2}{48L_{\nabla f}}, \tag{72}$$

$$\sum_{i=1}^{k-1}\gamma_i T_i = \sum_{i=1}^{k_0-1}\gamma_i T_i + \sum_{i=k_0}^{k-1}\gamma_i T_i \overset{(28)}{=} \frac{2}{3L_{\nabla f}}\sum_{i=1}^{k_0-1}2^{i-1} + \frac{T_{k_0}}{3L_{\nabla f}}\sum_{i=k_0}^{k-1}\frac{i-k_0+4}{2}$$
$$\leq \frac{2^{k_0}}{3L_{\nabla f}} + \frac{T_{k_0}}{3L_{\nabla f}}\frac{(k-k_0+7)(k-k_0)}{4} = \frac{T_{k_0}\big((k-k_0)^2+7(k-k_0)+8\big)}{12L_{\nabla f}}$$
$$\leq \frac{T_{k_0}\big(2(k-k_0)^2+12(k-k_0)+18\big)}{12L_{\nabla f}} \leq \frac{s(k-k_0+3)^2}{6L_{\nabla f}}. \tag{73}$$

Notice that $x^0, \tilde{x}^* \in X$. It then follows from Assumption 1 that $F(x^0)-F(\tilde{x}^*) \leq D_F$ and $\|x^0-\tilde{x}^*\| \leq D_X$. Using these relations, (71), (72), and (73), we obtain that

$$\mathbb{E}[F(\tilde{x}_k)-F(\tilde{x}^*)] \leq \frac{16}{s(k-k_0+3)^2}\Big(2(F(x^0)-F(\tilde{x}^*)) + \frac{3L_{\nabla f}}{2}\|x^0-\tilde{x}^*\|^2\Big) + 8C_5\epsilon$$
$$\leq \frac{16}{s(k-k_0+3)^2}\Big(2D_F + \frac{3L_{\nabla f}D_X^2}{2}\Big) + 8C_5\epsilon,$$

which together with (27) and $F(\tilde{x}^*) \leq F^*$ implies that the third relation in (29) holds.

In the remainder of this proof, we show that (30) holds. To this end, let $N_{k,f}$ denote the number of evaluations of ∇f_i performed in the kth outer iteration of Algorithm 2 for any $k \geq 1$, and let $N_{k,t}$ denote the number of evaluations of ∇c performed in the t-th inner iteration at the kth outer iteration of Algorithm 2 for all $1 \leq t \leq T_k$ and $k \geq 1$.

Observe from Algorithm 2 that $N_{k,f} = s + T_k$ for all $k \geq 1$. We next bound $N_{k,t}$ for all $1 \leq t \leq T_k$ and $k \geq 1$. To this end, let $(x^*_{k,t}, \lambda^*_{k,t})$ be a pair of optimal solutions of problem (21) and its dual, and define

$$D_{k,t} = \max\big\{\big\|(z_{t-1}, \lambda_{t-1}\min\{R/\|\lambda_{t-1}\|, 1\}) - (x^*_{k,t}, \lambda^*_{k,t})\big\|, 1\big\}, \tag{74}$$
$$B_{k,t} = 1/\gamma_k + L_{\nabla c}\|\lambda^*_{k,t}\| + L_{\nabla c}D_{k,t}, \tag{75}$$
$$\bar{C}_{k,t} = C\rho_0 + B_{k,t} + L_{\nabla c}\rho_0\eta_0/(1-\zeta\sigma) + 1. \tag{76}$$

By (8), (48), (74), and $z_{t-1}, x^*_{k,t} \in X$, one has

$$D_{k,t} \leq \max\big\{\|z_{t-1}-x^*_{k,t}\| + \|\lambda_{t-1}\min\{R/\|\lambda_{t-1}\|, 1\}\| + \|\lambda^*_{k,t}\|, 1\big\} \leq D_X + R + \Lambda. \tag{77}$$

Observe from (28) that $1/\gamma_k = 3L_{\nabla f}\alpha_k \leq 3L_{\nabla f}/2$. In view of this, (48), (75), and (77), we obtain that

$$B_{k,t} \leq 3L_{\nabla f}/2 + L_{\nabla c}\Lambda + L_{\nabla c}(D_X + R + \Lambda) = 3L_{\nabla f}/2 + L_{\nabla c}(D_X + R + 2\Lambda).$$

Using this relation, (25), (76), and the definition of C in (8), we have

$$\bar{C}_{k,t} \overset{(8)(76)}{=} (L_{\nabla c}C_c + L_c^2)\rho_0 + B_{k,t} + L_{\nabla c}\rho_0\eta_0/(1-\zeta\sigma) + 1 \overset{(25)}{\leq} C_2. \tag{78}$$

Note from (21) that the smooth component of $F_{k,t}(x)$ is $\|x - z_{t-1} + \gamma_k g_t\|^2/(2\gamma_k)$, which is $1/\gamma_k$-smooth and $1/\gamma_k$-strongly convex in $\mathbb{R}^n$. Moreover, notice from Algorithm 2 that Algorithm 1 with $\bar{\epsilon} = \epsilon$, $\mu = 1/\gamma_k$, $(x_0, \lambda_0) = (z_{t-1}, \lambda_{t-1}\min\{R/\|\lambda_{t-1}\|, 1\})$ is called to solve problem (21). By these observations, (25), (26), (77), (78), and Theorem 1, one has

$$N_{k,t} \overset{\text{Theorem 1}}{\leq} \frac{16\zeta\sqrt{\bar{C}_{k,t}}}{\sqrt{1/\gamma_k}(\sqrt{\zeta}-1)}\log\frac{2\zeta\bar{C}_{k,t}D_X}{\eta_0\sigma}\left\lceil\max\left\{1, \log_\zeta\frac{2(D_{k,t}+\rho_0\eta_0)}{(1-\zeta\sigma)\epsilon}\right\}\right\rceil\max\left\{1, \sqrt{\frac{2(D_{k,t}+\rho_0\eta_0)}{(1-\zeta\sigma)\epsilon}}\right\}$$
$$\leq L_{\nabla f}^{\frac{1}{2}}C_3\sqrt{\gamma_k}\left\lceil\max\left\{1, \log_\zeta\frac{C_4}{\epsilon}\right\}\right\rceil\max\left\{1, \sqrt{\frac{C_4}{\epsilon}}\right\}. \tag{79}$$

In addition, since $k_0 = \lfloor \log_2 s \rfloor + 1$, it follows that $k_0 \geq 1$, and hence $i - k_0 + 4 \leq 4i$ for all $i \geq 1$. Using this and the definitions of α_i and γ_i in (28), we have

$$\sum_{i=1}^{k} \sqrt{\gamma_i} = (3L_{\nabla f})^{-\frac{1}{2}} \sum_{i=1}^{k} \alpha_i^{-\frac{1}{2}} = (3L_{\nabla f})^{-\frac{1}{2}} \Big(\sum_{i=1}^{k_0} \sqrt{2} + \sum_{i=k_0+1}^{k} \sqrt{(i-k_0+4)/2} \Big) \leq (3L_{\nabla f})^{-\frac{1}{2}} \sum_{i=1}^{k} \sqrt{2i}$$
$$\leq \frac{2}{3}(3L_{\nabla f}/2)^{-\frac{1}{2}}(k+1)^{\frac{3}{2}} \leq \frac{2}{3}(3L_{\nabla f}/2)^{-\frac{1}{2}}(2k)^{\frac{3}{2}} \leq 2L_{\nabla f}^{-\frac{1}{2}} k^{\frac{3}{2}}, \tag{80}$$

where the last inequality is due to the fact that $\sum_{i=1}^{k} \sqrt{i} \leq \int_{\tau=0}^{k} \sqrt{\tau+1} d\tau < 2(k+1)^{3/2}/3$.

Let N_0 denote the number of evaluations of ∇c performed by Algorithm 3 applied to the problem $\min_{x \in X} \operatorname{dist}^2(c(x), -\mathcal{K})$ for finding $x^0 \in X$ satisfying $\operatorname{dist}(c(x^0), -\mathcal{K}) \leq \epsilon$. One can verify that the function $\operatorname{dist}^2(c(x), -\mathcal{K})$ is $2(L_c^2 + L_{\nabla c} C_c) D_X$-smooth on X. In addition, notice that one evaluation of ∇c is performed by Algorithm 3 per iteration. It then follows from (25) and Theorem 3 with $L_{\nabla \tilde{f}} = 2(L_c^2 + L_{\nabla c} C_c) D_X$ and $\varepsilon = \epsilon^2$ that

$$N_0 \leq 2 D_X^{3/2} \sqrt{L_c^2 + L_{\nabla c} C_c} \epsilon^{-1} \overset{(25)}{=} C_1 \epsilon^{-1}. \tag{81}$$

Recall that N_f and N_c denote the number of evaluations of ∇f_i and ∇c during the k outer iterations of Algorithm 2 for all $k \geq 1$, respectively. Notice from (28) that $T_i \leq T_{k_0} \leq s$ for all $i \geq 1$. Using this, (79), (80), (81), and Algorithm 2, we can observe that for all $k \geq 1$,

$$N_f = \sum_{i=1}^{k} (s + T_i) \leq 2sk,$$
$$N_c = N_0 + \sum_{i=1}^{k} \sum_{t=1}^{T_i} N_{i,t} \overset{(79)}{\leq} N_0 + L_{\nabla f}^{\frac{1}{2}} C_3 \left\lceil \max\left\{1, \log_\zeta \frac{C_4}{\epsilon}\right\} \right\rceil \max\left\{1, \sqrt{\frac{C_4}{\epsilon}}\right\} \sum_{i=1}^{k} T_i \sqrt{\gamma_i}$$
$$\leq N_0 + s L_{\nabla f}^{\frac{1}{2}} C_3 \left\lceil \max\left\{1, \log_\zeta \frac{C_4}{\epsilon}\right\} \right\rceil \max\left\{1, \sqrt{\frac{C_4}{\epsilon}}\right\} \sum_{i=1}^{k} \sqrt{\gamma_i},$$

which together with (80) and (81) implies that (30) holds. □

6 Concluding remarks

In this paper, we propose an inexact stochastic variance-reduced gradient method for solving a class of constrained convex finite-sum optimization problems. The method computes an *ϵ-surely feasible stochastic optimal* (ϵ-SFSO) solution using $\mathcal{O}\big(s \log s + s^{1/2} \epsilon^{-1/2}\big)$ and $\widetilde{\mathcal{O}}\big(s(\log s)^{3/2} \epsilon^{-1/2} + s^{1/4} \epsilon^{-5/4}\big)$ evaluations of ∇f_i and ∇c, respectively. These bounds significantly improve upon the previously best-known first-order oracle complexity for obtaining an ϵ-SFSO solution.

It is worth noting that the proposed method can be extended to the case where the objective function is additionally strongly convex. In this setting, the extended method can compute an ϵ-SFSO solution using

$$\mathcal{O}\big(s \log s + s^{1/2} \log \epsilon^{-1}\big) \quad \text{and} \quad \mathcal{O}\Big(\epsilon^{-1} + s^{1/2} \epsilon^{-1/2} (\log \epsilon^{-1})^2 + s \log s\, \epsilon^{-1/2} \log \epsilon^{-1}\Big)$$

evaluations of ∇f_i and ∇c, respectively. These bounds further improve upon those obtained under the merely convex setting.

A Accelerated proximal gradient methods for unconstrained convex optimization

In this part, we review accelerated proximal gradient (APG) methods for solving a class of unconstrained convex optimization problems of the form

$$\tilde{F}^* = \min_{x\in\mathbb{R}^n} \left\{ \tilde{F}(x) := \tilde{f}(x) + \psi(x) \right\}, \tag{82}$$

where ψ is a closed convex function with compact domain denoted by X, and $\tilde{f}$ is an $L_{\nabla\tilde{f}}$-smooth convex function on X. Throughout this section, we define $D_X = \max\{1, \max_{x,y\in X} \|x-y\|\}$.

We first present an APG method for solving (82) in which $\tilde{f}$ is convex but not necessarily strongly convex. This method is a variant of Nesterov's optimal first-order methods [15, 16], and has been studied in, for example, [21, Section 3].

Algorithm 3 An APG method for problem (82) with general convex $\tilde{f}$

Input: $x_0 \in X$.

1: Set $z_0 = x_0$.

2: **for** $k = 0, 1, \ldots$ **do**

3: Compute

$$\begin{aligned} y_k &= (kx_k + 2z_k)/(k+2), \\ z_{k+1} &= \arg\min_x \left\{ \langle \nabla\tilde{f}(y_k), x\rangle + \psi(x) + \frac{L_{\nabla\tilde{f}}}{k+2}\|x - z_k\|^2 \right\}, \\ x_{k+1} &= (kx_k + 2z_{k+1})/(k+2). \end{aligned}$$

4: **end for**

The following result provides the iteration complexity of Algorithm 3 for finding an ε-optimal solution to problem (82). It follows directly from [21, Corollary 1], and thus its proof is omitted.

Theorem 3. *For any given $\varepsilon > 0$, Algorithm 3 finds an approximate solution x_k of problem* (82) *such that $\tilde{F}(x_k) - \tilde{F}^* \le \varepsilon$ in no more than $\left\lceil D_X\sqrt{2L_{\nabla\tilde{f}}/\varepsilon} \right\rceil$ iterations.*

We next present an optimal APG method [14, Algorithm 4], which is a slight variant of Nesterov's optimal first-order methods [11, 15] for solving strongly convex problem (82), where $\tilde{f}$ is further assumed to be $\mu_{\tilde{f}}$-strongly convex on X with $0 < \mu_{\tilde{f}} \le L_{\nabla\tilde{f}}$.

Algorithm 4 An APG method for problem (82) with strongly convex $\tilde{f}$

Input: $\varepsilon > 0$ and $\tilde{x}_0 \in X$.

1: Set $x_0 = \tilde{x}_0$ and compute

$$x_1 = \arg\min_x \Big\{ \langle \nabla \tilde{f}(x_0), x \rangle + \psi(x) + \frac{L_{\nabla \tilde{f}}}{2} \|x - x_0\|^2 \Big\}.$$

2: Set $z_1 = x_1$ and $\alpha = \sqrt{\mu_{\tilde{f}} / L_{\nabla \tilde{f}}}$.

3: **for** $k = 1, 2, \ldots$ **do**

4: Compute

$$\begin{aligned} y_k &= (x_k + \alpha z_k)/(1+\alpha), \\ z_{k+1} &= \arg\min_x \Big\{ \langle \nabla \tilde{f}(y_k), x \rangle + \psi(x) + \frac{\alpha L_{\nabla \tilde{f}}}{2} \|x - \alpha y_k - (1-\alpha) z_k\|^2 \Big\}, \\ x_{k+1} &= (1-\alpha) x_k + \alpha z_{k+1}. \end{aligned}$$

5: Compute

$$\hat{x}_{k+1} = \arg\min_x \Big\{ \langle \nabla \tilde{f}(x_{k+1}), x \rangle + \psi(x) + \frac{L_{\nabla \tilde{f}}}{2} \|x - x_{k+1}\|^2 \Big\}.$$

6: **if** $\|x_{k+1} - \hat{x}_{k+1}\| \le \varepsilon / (2 L_{\nabla \tilde{f}})$ **then**

7: Terminate the algorithm and output $\hat{x}_{k+1}$.

8: **end if**

9: **end for**

The following result provides the iteration complexity of Algorithm 4 for computing an approximately optimal solution to problem (82). Its proof follows from [11, Theorem 1], combined with standard arguments, and is therefore omitted.

Theorem 4. *For any given $\varepsilon > 0$, Algorithm 4 terminates in at most $\widetilde{K}$ iterations, and outputs an approximate solution $\hat{x}_{k+1}$ of problem* (82) *satisfying* $\operatorname{dist}(0, \partial \tilde{F}(\hat{x}_{k+1})) \le \varepsilon$, *where*

$$\widetilde{K} = \left\lceil \sqrt{\frac{L_{\nabla \tilde{f}}}{\mu_{\tilde{f}}}} \right\rceil \max \left\{ 1, \left\lceil 2 \log \frac{2 L_{\nabla \tilde{f}} D_X}{\varepsilon} \right\rceil \right\}.$$

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[3] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[4] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

[5] H. Hendrikx, F. Bach, and L. Massoulie. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.

[6] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[7] G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1):201–235, 2016.

[8] G. Lan, Z. Li, and Y. Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171:167–215, 2018.

[10] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. SIAM, 1995.

[11] Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015.

[12] Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3112–3121, 2018.

[13] Z. Lu and Y. Xiao. First-order methods for stochastic and finite-sum convex optimization with deterministic constraints. 2025. Accepted by *Mathematics of Operations Research*.

[14] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *SIAM Journal on Optimization*, 33(2):1159–1190, 2023.

[15] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

[16] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[17] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[18] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1), 2013.

[19] C. Song, Y. Jiang, and Y. Ma. Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33:833–844, 2020.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[21] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2009. Manuscript.

[22] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[23] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970, 2017.

[24] D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pages 7574–7583, 2019.