Iteration-Complexity of First-Order Augmented Lagrangian Methods for Convex Conic Programming

Zhaosong Lu* Zirui Zhou†

March 8, 2021 (Revised: February 11, 2022; July 2, 2022)

Abstract

In this paper we consider a class of convex conic programming. In particular, we first propose an inexact augmented Lagrangian (I-AL) method that resembles the classical I-AL method for solving this problem, in which the augmented Lagrangian subproblems are solved approximately by a variant of Nesterov's optimal first-order method. We show that the total number of first-order iterations of the proposed I-AL method for finding an ϵ-KKT solution is at most $\mathcal{O}(\epsilon^{-7/4})$. We then propose an adaptively regularized I-AL method and show that it achieves a first-order iteration complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$, which significantly improves existing complexity bounds achieved by first-order I-AL methods for finding an ϵ-KKT solution. Our complexity analysis of the I-AL methods is based on a sharp analysis of inexact proximal point algorithm (PPA) and the connection between the I-AL methods and inexact PPA. It is vastly different from existing complexity analyses of the first-order I-AL methods in the literature, which typically regard the I-AL methods as an inexact dual gradient method.

Keywords: Convex conic programming, augmented Lagrangian method, first-order method, iteration complexity

Mathematics Subject Classification: 90C25, 90C30, 90C46, 49M37

1 Introduction

In this paper we consider convex conic programming in the form of

$$\begin{array}{rl} F^* = \min & \{F(x) := f(x) + P(x)\} \\ \text{s.t.} & g(x) \preceq_{\mathcal{K}} 0, \end{array} \tag{1}$$

where $f, P : \Re^n \to (-\infty, \infty]$ are proper closed convex functions, $\mathcal{K}$ is a closed convex cone in $\Re^m$, the symbol $\preceq_{\mathcal{K}}$ denotes the partial order induced by $\mathcal{K}$, that is, $y \preceq_{\mathcal{K}} z$ if and only if $z - y \in \mathcal{K}$, and the mapping $g : \Re^n \to \Re^m$ is convex with respect to $\mathcal{K}$, that is,

$$g(\alpha x + (1-\alpha)y) \preceq_{\mathcal{K}} \alpha g(x) + (1-\alpha)g(y), \quad \forall x, y \in \Re^n,\ \alpha \in [0,1]. \tag{2}$$

The associated Lagrangian dual problem of (1) is given by

$$d^* = \sup_{\lambda \in \mathcal{K}^*} \inf_x \{f(x) + P(x) + \langle \lambda, g(x) \rangle\}, \tag{3}$$

where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$ (see Section 2). We make the following additional assumptions on problems (1) and (3) throughout this paper.

Assumption 1. *(a) The proximal operator associated with P can be exactly evaluated, and the domain of P, denoted by* $\mathrm{dom}(P)$*, is compact.*[1]

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu). The work of this author was partially supported by NSF Award IIS-2211491.

†Huawei Technologies Canada, Burnaby, BC, Canada (email: zirui.zhou@huawei.com).

[1] Some problems with unbounded $\mathrm{dom}(P)$ can be reformulated as the ones satisfying this assumption. For example, when f is bounded below and P is coercive, the problem generally can be reformulated as the one with the objective $f + \tilde{P}$ for some $\tilde{P}$ with a compact domain. Such a problem often arises in sparse or low-rank learning, in which f is typically a nonnegative loss function and P is the ℓ_1- or nuclear-norm.

(b) *The projection onto $\mathcal{K}$ can be exactly evaluated.*

(c) *The functions f and g are continuously differentiable on an open set Ω containing* $\mathrm{dom}(P)$, *and ∇f and ∇g are Lipschitz continuous on Ω with Lipschitz constants $L_{\nabla f}$ and $L_{\nabla g}$, respectively.*[2]

(d) *The strong duality holds for problems* (1) *and* (3), *that is, both problems have optimal solutions and moreover their optimal values F^* and d^* are equal.*

Problem (1) includes a rich class of problems as special cases. For example, when $\mathcal{K} = \Re^{m_1}_+ \times \{0\}^{m_2}$ for some m_1 and m_2, $g(x) = (g_1(x), \ldots, g_{m_1}(x), h_1(x), \ldots, h_{m_2}(x))^T$ with convex g_i's and affine h_j's, and $P(x)$ is the indicator function of a simple convex set $X \subseteq \Re^n$, problem (1) reduces to an ordinary convex programming problem

$$\min_{x \in X} \{f(x) : g_i(x) \le 0,\ i = 1, \ldots, m_1; h_j(x) = 0,\ j = 1, \ldots, m_2\}.$$

Augmented Lagrangian (AL) methods have been widely regarded as a highly effective method for solving constrained nonlinear programming (e.g., see [3, 24, 17]). The classical AL method was initially proposed by Hestenes [7] and Powell [19], and has been extensively studied in the literature (e.g., see [20, 2]). Recently, AL methods have been applied to solve some instances of problem (1) arising in various applications such as image processing [5] and optimal control [8]. They have also been used to solve large-scale conic programming problems (e.g., see [4, 9, 29]).

When applied to problem (1), AL methods proceed in the following manner. Let $\{\rho_k\}$ be a sequence of nondecreasing positive scalars and $\lambda^0 \in \mathcal{K}^*$ an initial guess of the Lagrangian multiplier of (1). At the kth iteration, x^{k+1} is obtained by approximately solving the AL subproblem

$$\min_x \mathcal{L}(x, \lambda^k; \rho_k), \tag{4}$$

where $\mathcal{L}(x, \lambda; \rho)$ is the AL function of (1) defined as (e.g., see [23, Section 11.K] and [25])

$$\mathcal{L}(x, \lambda; \rho) := f(x) + P(x) + \frac{1}{2\rho}\left[\mathrm{dist}^2(\lambda + \rho g(x), -\mathcal{K}) - \|\lambda\|^2\right],$$

and $\mathrm{dist}(z, -\mathcal{K}) = \min\{\|z + x\| : x \in \mathcal{K}\}$ for any $z \in \Re^m$ Then λ^{k+1} is updated by

$$\lambda^{k+1} = \Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^{k+1})),$$

where $\Pi_{\mathcal{K}^*}(\cdot)$ is the projection operator onto $\mathcal{K}^*$. The iterations that update $\{\lambda^k\}$ are commonly called the *outer iterations* of AL methods, and the iterations of an iterative scheme for solving AL subproblem (4) are referred to as the *inner iterations* of AL methods. In the context of large-scale optimization, a first-order method is often used to approximately solve the AL subproblem (4) and the resulting entire method is usually called a *first-order inexact AL (I-AL) method*.

In this paper we focus on developing first-order I-AL methods and studying their iteration complexity, which is an upper bound on the total number of first-order inner iterations for finding an ϵ-Karush-Kuhn-Tucker (ϵ-KKT) solution of (1), that is, a primal-dual solution (x, λ) satisfying

$$\mathrm{dist}(0, \nabla f(x) + \partial P(x) + \nabla g(x)\lambda) \le \epsilon, \quad \mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \epsilon, \quad (x, \lambda) \in \mathrm{dom}(P) \times \mathcal{K}^* \tag{5}$$

for some prescribed tolerance $\epsilon > 0$. The condition (5) is often checkable in practice and has been broadly used as a termination criterion for I-AL type of methods (e.g., see [29]). It has been shown that under some mild error bound condition any point x satisfying (5) with a small ϵ is close to an optimal solution of problem (1) (e.g., see [20]). As problem (1) arising in various applications is of large scale, a first-order I-AL method with a low complexity bound for finding an ϵ-KKT solution of (1) is highly desirable. The main contributions of this paper consist of: (i) proposing a first-order I-AL method that resembles the classical AL method and establishing its iteration complexity, which reveals how good iteration complexity of the classical I-AL method can be; (ii) proposing an adaptively regularized first-order I-AL method that achieves a significantly improved iteration complexity over existing first-order I-AL methods for finding an ϵ-KKT solution of problem (1); and (iii) a technically new complexity analysis of the I-AL methods based on a sharp analysis of inexact proximal point algorithm (PPA) and the connection between the I-AL methods and inexact PPA, which provides more insights than existing complexity analyses of the first-order I-AL methods in the literature that typically regard the I-AL methods as an inexact dual gradient method.

[2]The symbol ∇g denotes the transpose of the Jacobian of g.

1.1 Related works

Aybat and Iyengar [1] proposed a first-order I-AL method for solving a special case of (1) with affine mapping g. In particular, they applied an optimal first-order method (e.g., see [15, 26]) to find an approximate solution x^{k+1} of the AL subproblem (4) such that

$$\mathcal{L}(x^{k+1}, \lambda^k; \rho_k) - \min_x \mathcal{L}(x, \lambda^k; \rho_k) \leq \eta_k$$

for some $\eta_k > 0$. It is shown in [1] that this method with some suitable choice of $\{\rho_k\}$ and $\{\eta_k\}$ can find an approximate solution x of (1) satisfying

$$|F(x) - F^*| \leq \epsilon, \quad \mathrm{dist}(g(x), -\mathcal{K}) \leq \epsilon \tag{6}$$

for some $\epsilon > 0$ in at most $\mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})$ first-order inner iterations. In addition, Necoara et al. [14] proposed an accelerated first-order I-AL method for solving the same problem as considered in [1], in which an acceleration scheme [6] is applied to $\{\lambda^k\}$ for possibly better convergence. It is claimed in [14] that this method with a suitable choice of $\{\rho_k\}$ and $\{\eta_k\}$ can find an approximate solution x of (1) satisfying (6) in at most $\mathcal{O}(\epsilon^{-1})$ first-order inner iterations. More recently, Xu [28] proposed an I-AL method for solving a special case of (1) with $\mathcal{K}$ being the nonnegative orthant, which can find an approximate solution x of (1) satisfying (6) in at most $\mathcal{O}(\epsilon^{-1})$ first-order inner iterations. Some other related works on I-AL type of methods can be found, for example, in [12, 18, 27].

Since F^* is typically unknown, the criterion (6) is not checkable and cannot be used as a practical termination criterion for the I-AL methods [1, 14, 18, 27, 28] in general. Thus, for the practical use of these methods, one has to terminate them by a checkable criterion, which may result in a substantially different solution from an ϵ-optimal solution defined in (6). Due to this, it may be challenging for them to find such an ϵ-optimal solution *in practice*. Moreover, for these I-AL methods, $\{\rho_k\}$ and $\{\eta_k\}$ are specifically chosen to achieve a low first-order iteration complexity with respect to (6). Such a choice may however not lead to a low first-order iteration complexity with respect to a checkable termination criterion. Therefore, the iteration-complexity results obtained in [1, 14, 18, 27, 28] with respect to the criterion (6) do not seem to have much practical merits in general.

In addition to the aforementioned I-AL methods, Lan and Monteiro [11] proposed a first-order I-AL method for finding an *ϵ-KKT solution* of a special case of (1) with $g = \mathcal{A}(\cdot)$, $\mathcal{K} = \{0\}^m$ and P being the indicator function of a simple compact convex set X, that is,

$$\min \{f(x) : \mathcal{A}(x) = 0, x \in X\}, \tag{7}$$

where $\mathcal{A} : \Re^n \to \Re^m$ is an affine mapping. Roughly speaking, their I-AL method consists of two stages, particularly, primary stage and postprocessing stage. The primary stage is to execute the usual I-AL steps similar to those in [1] but with static $\rho_k \equiv \mathcal{O}(D_\Lambda^{3/4} \epsilon^{-3/4})$ and $\eta_k \equiv \mathcal{O}(D_\Lambda^{1/4} \epsilon^{7/4})$ until a certain approximate solution $(\tilde{x}, \tilde{\lambda})$ is found,[3] where $D_\Lambda = \min\{\|\lambda^0 - \lambda\| : \lambda \in \Lambda^*\}$ and Λ^* is the set of optimal solutions of the Lagrangian dual problem associated with problem (7). The postprocessing stage is mainly to execute a single I-AL step with a penalty parameter $\rho = \mathcal{O}(D_\Lambda^{3/4} \epsilon^{-3/4})$ and an AL subproblem tolerance parameter $\eta = \mathcal{O}(\min(D_\Lambda^{3/4} \epsilon^{5/4}, D_\Lambda^{-3/4} \epsilon^{11/4}))$, starting with $(\tilde{x}, \tilde{\lambda})$. It is shown in [11] that this I-AL method can find an *ϵ-KKT solution* of (7) in at most $\mathcal{O}(\epsilon^{-7/4})$ first-order inner iterations in theory. However, this method is much less practical than the classical I-AL method. Indeed, in the primary stage, this I-AL method uses ρ_k and η_k of same value through all outer iterations, which may be respectively overly large and small. Such a choice of ρ_k and η_k is clearly against the common practical choice that ρ_0 and η_0 are relatively small and large, respectively, and ρ_k gradually increases and η_k gradually decreases as iteration progresses. Moreover, ρ_k and η_k in this method require some knowledge of D_Λ, which is not known a priori and needs to be estimated by a sophisticated and expensive "guess-and-check" procedure proposed in [11]. These two aspects evidently make this I-AL method much more sophisticated and less practical than the classical I-AL method.

Besides, Lan and Monteiro [11] proposed a modified I-AL method by applying their aforementioned first-order I-AL method with D_Λ replaced by D_Λ^ϵ to the perturbed problem

$$\min \left\{ f(x) + \frac{\epsilon}{4D_X} \|x - x^0\|^2 : \mathcal{A}(x) = 0, x \in X \right\}, \tag{8}$$

[3] It means that $\rho_k = \rho$ and $\eta_k = \eta$ for all k for some $\rho = \mathcal{O}(D_\Lambda^{3/4} \epsilon^{-3/4})$ and $\eta = \mathcal{O}(D_\Lambda^{1/4} \epsilon^{7/4})$.

starting with some (x^0, λ^0), where $D_X = \max\{\|x - y\| : x, y \in X\}$, $D_\Lambda^\epsilon = \min\{\|\lambda^0 - \lambda\| : \lambda \in \Lambda_\epsilon^*\}$, and Λ_ϵ^* is the set of Lagrangian dual optimal solutions associated with problem (8). They showed that their modified I-AL method can find an ϵ-*KKT solution* of (7) in at most

$$\mathcal{O}\left\{\left(\frac{\sqrt{D_\Lambda^\epsilon}}{\epsilon}\left[\log\frac{\sqrt{D_\Lambda^\epsilon}}{\epsilon}\right]^{\frac{3}{4}} + \frac{1}{\sqrt{\epsilon}}\log\frac{\sqrt{D_\Lambda^\epsilon}}{\epsilon}\right)\max\left(1, \log\log\frac{\sqrt{D_\Lambda^\epsilon}}{\epsilon}\right)\right\} \tag{9}$$

first-order inner iterations. Since D_Λ^ϵ depends on ϵ (see an example in Appendix A) and its order dependence on ϵ is generally unknown, it is not clear about the order dependence of the iteration complexity (9) on ϵ.

1.2 Main contribution

The goal of this paper is to propose first-order I-AL methods with significantly improved iteration complexity over existing first-order I-AL methods for finding an ϵ-*KKT solution* of problem (1). Our main contribution is listed below.

- We propose a first-order I-AL method that resembles the classical I-AL method establish its first-order iteration complexity $\mathcal{O}(\epsilon^{-7/4})$ for finding an ϵ-*KKT solution* of problem (1). Due to the similarity between this I-AL method and the classical I-AL method, our complexity result reveals that the iteration complexity of the classical I-AL method for finding an ϵ-KKT solution of (1) appears to be $\mathcal{O}(\epsilon^{-7/4})$.
- We propose an *adaptively regularized* first-order I-AL method that establish its first-order iteration complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$ for finding an ϵ-*KKT solution* of problem (1), which significantly improves the previously best-known iteration-complexity $\mathcal{O}(\epsilon^{-7/4})$ achieved by first-order I-AL methods for finding an ϵ-KKT solution of (1). This complexity result implies that the adaptively regularized first-order I-AL method is generally superior to the classical I-AL method for finding an ϵ-KKT solution of (1).
- Our complexity analysis of the I-AL methods is technically new, which is based on a sharp analysis of inexact PPA and the connection between the I-AL methods and inexact PPA. It is vastly different from existing complexity analyses of the first-order I-AL methods, which typically regard the I-AL methods as an inexact dual gradient method. Since the operator associated with the monotone inclusion problem linked to the I-AL methods is closely related to the KKT conditions, our analysis is more appropriate and provides more insights than existing ones in the literature.

1.3 Outline

The rest of this paper is organized as follows. In Section 2 we introduce some notation and the concept of ϵ-KKT solution. In Section 3 we propose a first-order I-AL method and present its iteration complexity. Also, in Section 4 we propose an adaptively regularized first-order I-AL method and present its iteration complexity. In Section 5 we provide a proof for the technical results stated in Sections 3 and 4. In Section 6 we present some numerical results for the proposed algorithms. Finally, we make some concluding remarks in Section 7.

2 Notation and preliminaries

The following notations will be used throughout this paper. Let $\Re^n$ denote the Euclidean space of dimension n, $\langle\cdot,\cdot\rangle$ denote the standard inner product, and $\|\cdot\|$ stand for the Euclidean norm. The symbols $\Re_+$ and $\Re_{++}$ stand for the set of nonnegative and positive real numbers, respectively.

Given a closed convex function $h : \Re^n \to (-\infty, \infty]$, ∂h and $\mathrm{dom}(h)$ denote the subdifferential and domain of h, respectively. The proximal operator associated with h is denoted by prox_h, that is,

$$\mathrm{prox}_h(z) = \arg\min_{x\in\Re^n}\left\{\frac{1}{2}\|x - z\|^2 + h(x)\right\}, \quad \forall z \in \Re^n.$$

Given a non-empty closed convex set $C \subseteq \Re^n$, $\mathrm{dist}(z, C)$ stands for the Euclidean distance from z to C, and $\Pi_C(z)$ denotes the Euclidean projection of z onto C, namely,

$$\Pi_C(z) = \arg\min\{\|z - x\| : x \in C\}, \quad \mathrm{dist}(z, C) = \|z - \Pi_C(z)\|, \quad \forall z \in \Re^n.$$

The normal cone of C at any $z \in C$ is denoted by $\mathcal{N}_C(z)$. For the closed convex cone $\mathcal{K}$, we use $\mathcal{K}^*$ to denote the dual cone of $\mathcal{K}$, that is, $\mathcal{K}^* = \{y \in \Re^m : \langle y, x\rangle \geq 0, \ \forall x \in \mathcal{K}\}$.

The Lagrangian function $l(x, \lambda)$ of problem (1) is defined as

$$l(x,\lambda) = \begin{cases} f(x) + P(x) + \langle \lambda, g(x)\rangle & \text{if } x \in \mathrm{dom}(P) \text{ and } \lambda \in \mathcal{K}^*, \\ -\infty & \text{if } x \in \mathrm{dom}(P) \text{ and } \lambda \notin \mathcal{K}^*, \\ \infty & \text{if } x \notin \mathrm{dom}(P), \end{cases}$$

which is a closed convex-concave function. The Lagrangian dual function $d : \Re^m \to [-\infty, \infty)$ is defined as

$$d(\lambda) = \inf_x l(x,\lambda) = \begin{cases} \inf\limits_x \{f(x) + P(x) + \langle \lambda, g(x)\rangle\} & \text{if } \lambda \in \mathcal{K}^*, \\ -\infty & \text{if } \lambda \notin \mathcal{K}^*, \end{cases} \tag{10}$$

which is a closed concave function. Moreover, the augmented Lagrangian function for problem (1) is defined as (e.g., see [25])

$$\mathcal{L}(x, \lambda; \rho) = f(x) + P(x) + \frac{1}{2\rho}\Big[\mathrm{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big) - \|\lambda\|^2\Big], \tag{11}$$

where $\rho > 0$ is a penalty parameter. For convenience, we let

$$\mathcal{S}(x, \lambda; \rho) := f(x) + \frac{1}{2\rho}\mathrm{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big). \tag{12}$$

It is clear to see that

$$\mathcal{L}(x, \lambda; \rho) = \mathcal{S}(x, \lambda; \rho) + P(x) - \frac{\|\lambda\|^2}{2\rho}.$$

The augmented Lagrangian dual function of (1) is given by

$$d(\lambda; \rho) = \min_{x \in \Re^n} \mathcal{L}(x, \lambda; \rho). \tag{13}$$

The Lagrangian dual problem (3) can thus be rewritten as

$$d^* = \max_\lambda \ d(\lambda). \tag{14}$$

Let $\partial l : \Re^n \times \Re^m \rightrightarrows \Re^n \times \Re^m$ and $\partial d : \Re^m \rightrightarrows \Re^m$ be respectively the subdifferential mappings associated with l and d (e.g., see [22]). We define two set-valued operators associated with problems (1) and (3) as follows:

$$\mathcal{T}_d : \lambda \to \{u \in \Re^m : -u \in \partial d(\lambda)\}, \quad \forall \lambda \in \Re^m, \tag{15}$$

$$\mathcal{T}_l : (x, \lambda) \to \{(v, u) \in \Re^n \times \Re^m : (v, -u) \in \partial l(x, \lambda)\}, \quad \forall (x, \lambda) \in \Re^n \times \Re^m. \tag{16}$$

It is well known that λ^* is an optimal solution of the Lagrangian dual problem (14) if and only if $0 \in \partial d(\lambda^*)$, and (x^*, λ^*) is a saddle point[4] of l if and only if $(0, 0) \in \partial l(x^*, \lambda^*)$. In addition, it can be verified that

$$\partial l(x, \lambda) = \begin{cases} \begin{pmatrix} \nabla f(x) + \partial P(x) + \nabla g(x)\lambda \\ g(x) - \mathcal{N}_{\mathcal{K}^*}(\lambda) \end{pmatrix}, & \text{if } x \in \mathrm{dom}(P) \text{ and } \lambda \in \mathcal{K}^*, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{17}$$

This enables us to write the KKT optimality condition for problem (1) as follows.

Proposition 1. *Under Assumption 1, $x^* \in \Re^n$ is an optimal solution of (1) if and only if there exists $\lambda^* \in \Re^m$ such that*

$$(0, 0) \in \partial l(x^*, \lambda^*), \tag{18}$$

or equivalently, (x^*, λ^*) satisfies the KKT conditions for (1), that is,

$$0 \in \nabla f(x^*) + \partial P(x^*) + \nabla g(x^*)\lambda^*, \quad \lambda^* \in \mathcal{K}^*, \quad g(x^*) \preceq_{\mathcal{K}} 0, \quad \langle \lambda^*, g(x^*)\rangle = 0.$$

[4] (x^*, λ^*) is called a saddle point of l if it satisfies $\sup_\lambda l(x^*, \lambda) = l(x^*, \lambda^*) = \inf_x l(x, \lambda^*)$.

Proof. The result (18) follows from [22, Theorem 36.6]. By (17), it is not hard to see that (18) holds if and only if $0 \in \nabla f(x^*) + \partial P(x^*) + \nabla g(x^*)\lambda^*$, $\lambda^* \in \mathcal{K}^*$, and $g(x^*) \in \mathcal{N}_{\mathcal{K}^*}(\lambda^*)$. By the definition of $\mathcal{K}^*$ and $\mathcal{N}_{\mathcal{K}^*}$, one can verify that $g(x^*) \in \mathcal{N}_{\mathcal{K}^*}(\lambda^*)$ is equivalent to $g(x^*) \preceq_{\mathcal{K}} 0$ and $\langle \lambda^*, g(x^*)\rangle = 0$. The proof is then completed. □

In practice it is generally impossible to find an exact KKT solution (x^*, λ^*) satisfying (18). We are instead interested in seeking an approximate KKT solution of (1) that is defined as follows.

Definition 1. *Given any $\epsilon > 0$, we say $(x, \lambda) \in \Re^n \times \Re^m$ is an ϵ-KKT solution of* (1) *if there exists $(u, v) \in \partial l(x, \lambda)$ such that $\|u\| \le \epsilon$ and $\|v\| \le \epsilon$.*

Remark 1. *(a) By* (17) *and Definition 1, one can see that (x, λ) is an ϵ-KKT solution of* (1) *if and only if $x \in \mathrm{dom}(P)$, $\lambda \in \mathcal{K}^*$, $\mathrm{dist}(0, \nabla f(x) + \partial P(x) + \nabla g(x)\lambda) \le \epsilon$, and $\mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \epsilon$. It reduces to an ϵ-KKT solution introduced in [11] when g is affine and $\mathcal{K} = \{0\}$.*

(b) For a given (x, λ), it is generally not hard to verify whether it is an ϵ-KKT solution of (1)*. Therefore, Definition 1 gives rise to a checkable termination criterion* (5) *that will be used in this paper.*

3 A first-order I-AL method and its iteration complexity

In this section we propose a first-order I-AL method that resembles the classical I-AL method, and study its first-order iteration complexity for finding an ϵ-KKT solution of problem (1).

Recall Remark 1(a) that (x, λ) is an ϵ-KKT solution of (1) if and only if it satisfies that $x \in \mathrm{dom}(P)$, $\lambda \in \mathcal{K}^*$, $\mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \epsilon$, and $\mathrm{dist}(0, \nabla f(x) + \partial P(x) + \nabla g(x)\lambda) \le \epsilon$. In what follows, we propose an I-AL method to find a pair (x, λ) satisfying these conditions. Given that the proximal operator associated with P and the projection onto $\mathcal{K}$ can be exactly evaluated (see Assumption 1), the first two conditions can be easily satisfied by the iterates of our I-AL method. Observe that the last condition is generally harder to satisfy than the third one since it involves ∇f, ∇g and ∂P. Due to this, our I-AL method consists of two stages, particularly, primary stage and postprocessing stage. In the primary stage, the AL subproblems are solved roughly and the usual I-AL steps are executed until either an ϵ-KKT solution of (1) is obtained or a pair (x^k, λ^k) satisfying nearly the third condition but roughly the last condition is found. In the postprocessing stage, the last AL subproblem arising in the primary stage is resolved to a higher accuracy for obtaining some point $\tilde{x}$, starting with x^k, and a proximal step is then applied to $\mathcal{L}(\cdot, \lambda^k, \rho_k)$ at $\tilde{x}$ and to $l(\tilde{x}, \cdot)$ at λ^k respectively, to generate the output (x^+, λ^+).

Our first-order I-AL method for solving problem (1) is presented as follows.

Algorithm 1 (A first-order I-AL method)**.**

0. Input $\epsilon > 0$, $\lambda^0 \in \mathcal{K}^*$, nondecreasing $\{\rho_k\} \subset \Re_{++}$, and $0 < \eta_k \downarrow 0$. Set $k = 0$.

1. Apply Algorithm 3 (in Appendix B) to the problem $\min_x \mathcal{L}(x, \lambda^k; \rho_k)$ to find $x^{k+1} \in \mathrm{dom}(P)$ satisfying

$$\mathcal{L}(x^{k+1}, \lambda^k; \rho_k) - \min_x \mathcal{L}(x, \lambda^k; \rho_k) \le \eta_k.^5 \tag{19}$$

2. Set $\lambda^{k+1} = \Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^{k+1}))$.

3. If (x^{k+1}, λ^{k+1}) satisfies (5), output $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$ and terminate.

4. If the following inequalities are satisfied

$$\frac{1}{\rho_k}\|\lambda^{k+1} - \lambda^k\| \le \frac{3}{4}\epsilon, \qquad \frac{\eta_k}{\rho_k} \le \frac{\epsilon^2}{128}, \tag{20}$$

   call the subroutine $(x^+, \lambda^+) = \mathrm{Postprocessing}(\lambda^k, \rho_k, x^{k+1}, \epsilon)$, output (x^+, λ^+) and terminate.

5. Set $k \leftarrow k + 1$ and go to Step 1.

[5] In view of Proposition 3, one can terminate Algorithm 3 once a point x^{k+1} satisfying $\Psi(x^{k+1}) - \underline{\Psi}_{k+1} \le \eta_k$ is found, where $\Psi(\cdot) = \mathcal{L}(\cdot, \lambda^k; \rho_k)$, and $\underline{\Psi}_{k+1}$ is defined according to (88) that can be cheaply computed (see Remark 5). Such x^{k+1} clearly satisfies (19).

End.

We next present the subroutine Postprocessing that is used in Step 4 of Algorithm 1. Before proceeding, let L_g be the Lipschitz constant of g on $\mathrm{dom}(P)$ and $M_g = \max_{x\in\mathrm{dom}(P)} \|g(x)\|$.

Subroutine $(x^+, \lambda^+) = \textbf{Postprocessing}(\tilde{\lambda}, \tilde{\rho}, \tilde{x}, \epsilon)$

0. Input $\tilde{\lambda} \in \mathcal{K}^*$, $\tilde{\rho} > 0$, $\tilde{x} \in \mathrm{dom}(P)$, and $\epsilon > 0$.

1. Set
$$\tilde{L} = L_{\nabla f} + L_{\nabla g}\big(\|\tilde{\lambda}\| + \tilde{\rho} M_g\big) + \tilde{\rho} L_g^2, \qquad \tilde{\eta} = \epsilon^2 \cdot \min\left\{\frac{\tilde{\rho}}{128}, \frac{1}{8\tilde{L}}\right\}. \tag{21}$$

2. Apply Algorithm 3 to the problem $\min_x \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho})$ starting with $\tilde{x}$ to find $\hat{x}$ such that
$$\mathcal{L}(\hat{x}, \tilde{\lambda}; \tilde{\rho}) - \min_x \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho}) \le \tilde{\eta}. \tag{22}$$

3. Output the pair (x^+, λ^+), which is computed by
$$x^+ = \mathrm{prox}_{P/\tilde{L}}\big(\hat{x} - \nabla_x \mathcal{S}(\hat{x}, \tilde{\lambda}; \tilde{\rho})/\tilde{L}\big), \quad \lambda^+ = \Pi_{\mathcal{K}^*}\big(\tilde{\lambda} + \tilde{\rho} g(x^+)\big), \tag{23}$$
where $\mathcal{S}$ is defined in (12).

End.

For ease of later reference, we refer to the first-order iterations of Algorithm 3 for solving the AL subproblems as the *inner iterations* of Algorithm 1, and call, the update from (x^k, λ^k) to (x^{k+1}, λ^{k+1}) or the postprocessing step, an *outer iteration* of Algorithm 1. We now make some remarks on Algorithm 1 as follows.

Remark 2. *(a) The subroutine Postprocessing is inspired by [11], in which a similar procedure is proposed for solving a special case of problem* (1) *with affine g and $\mathcal{K} = \{0\}$. The main purpose of this subroutine is to obtain a better iteration complexity.*

(b) Compared to the I-AL method in [11], our I-AL method is much simpler and more closely resembles the classical I-AL method. Indeed, the I-AL method [11] uses the static $\rho_k \equiv \mathcal{O}(D_\Lambda^{3/4}\epsilon^{-3/4})$ and $\eta_k \equiv \mathcal{O}(D_\Lambda^{1/4}\epsilon^{7/4})$ through all outer iterations in the primary stage,[6] where $D_\Lambda = \min\{\|\lambda^0 - \lambda\| : \lambda \in \Lambda^*\}$ and Λ^* is the set of optimal solutions of the Lagrangian dual problem associated with problem (7). *Such $\{\rho_k\}$ and $\{\eta_k\}$ may be overly large and small, respectively. This is clearly against the common practical choice that ρ_0 and η_0 are relatively small and large, respectively, and $\{\rho_k\}$ gradually increases and $\{\eta_k\}$ progressively decreases. Moreover, the above choice of ρ_k and η_k requires some knowledge of D_Λ, which is not known a priori and needs to be estimated by a sophisticated and expensive "guess-and-check" procedure proposed in [11]. In contrast, our I-AL method uses a practical choice of $\{\rho_k\}$ and $\{\eta_k\}$, which dynamically change throughout the iterations. Also, it does not use any knowledge of D_Λ and thus a "guess-and-check" procedure is not required.*

In the rest of this section we present our main results for Algorithm 1, whose proof is deferred to Subsection 5.1. To proceed, we introduce some further notation below. Let Λ^* be the set of optimal solutions of problem (3) and $\hat{\lambda}^* \in \Lambda^*$ such that $\|\lambda^0 - \hat{\lambda}^*\| = \mathrm{dist}(\lambda^0, \Lambda^*)$. In addition, we define
$$D_X = \max_{x,y\in\mathrm{dom}(P)} \|x - y\|, \; D_\Lambda = \|\lambda^0 - \hat{\lambda}^*\|, \; B = L_{\nabla f} + L_{\nabla g}(\|\hat{\lambda}^*\| + D_\Lambda), \tag{24}$$
$$C = L_{\nabla g} M_g + L_g^2, \; \bar{D}_\Lambda = \max\{D_\Lambda, 1\}, \; \bar{B} = \max\{B, 1\}, \; \bar{C} = \max\{C, 1\}, \tag{25}$$
where $L_{\nabla f}$, $L_{\nabla g}$, L_g and M_g are defined above. We start with the following theorem which shows that Algorithm 1 with a suitable choice of $\{\rho_k\}$ and $\{\eta_k\}$ is guaranteed to find an ϵ-KKT solution of problem (1) in a finite number of outer iterations.

[6] It means that $\rho_k = \rho$ and $\eta_k = \eta$ for all k for some $\rho = \mathcal{O}(D_\Lambda^{3/4}\epsilon^{-3/4})$ and $\eta = \mathcal{O}(D_\Lambda^{1/4}\epsilon^{7/4})$.

Theorem 1. *(i) If Algorithm 1 successfully terminates, then the output* (x^+, λ^+) *is an* ϵ*-KKT solution of problem* (1).

(ii) Suppose that $\{\rho_k\}$ *and* $\{\eta_k\}$ *satisfy that*

$$\rho_k > 0 \textit{ is nondecreasing}, \quad 0 < \frac{\eta_k}{\rho_k} \to 0, \quad \frac{\sum_{i=0}^{2k} \sqrt{\rho_i \eta_i}}{\rho_k \sqrt{k+1}} \to 0.^7 \tag{26}$$

Then Algorithm 1 terminates in a finite number of outer iterations.

(iii) Furthermore, if N *is a nonnegative integer such that*

$$\frac{D_\Lambda + 2\sum_{k=0}^{2N} \sqrt{2\rho_k \eta_k}}{\rho_N \sqrt{N+1}} \le \frac{\epsilon}{2}, \quad \frac{\eta_N}{\rho_N} \le \frac{\epsilon^2}{128}, \tag{27}$$

then Algorithm 1 terminates in at most $2N+1$ *outer iterations.*

The next theorem provides an upper bound on the total number of the inner iterations Algorithm 1, that is, the total iterations of Algorithm 3 applied to solve all AL subproblems of Algorithm 1.

Theorem 2. *Let* $\epsilon > 0$ *be given, and* $\bar{C}$, D_X, *and* $\bar{D}_\Lambda$ *be defined in* (24) *and* (25)*. Suppose that* $\{\rho_k\}$ *and* $\{\eta_k\}$ *are chosen as*

$$\rho_k = \rho_0 (k+1)^{\frac{3}{2}}, \quad \eta_k = \eta_0 (k+1)^{-\frac{5}{2}} \cdot \min\{1, \sqrt{\epsilon}\} \tag{28}$$

for some $\rho_0 \ge 1$ *and* $0 < \eta_0 \le 1$*. Then, the total number of inner iterations of Algorithm 1 for finding an* ϵ*-KKT solution of problem* (1) *is at most* $\mathcal{O}\big(\mathcal{T}\big(\min\{1, \epsilon\}\big)\big)$*, where*

$$\mathcal{T}(t) = \frac{D_X \bar{D}_\Lambda^{\frac{3}{2}} \bar{C}}{t^{\frac{7}{4}}} + \frac{D_X \bar{D}_\Lambda^{\frac{5}{4}} \bar{B}^{\frac{1}{2}} (1 + L_{\nabla g}^{\frac{1}{2}})}{t^{\frac{11}{8}}} + \frac{D_X \bar{D}_\Lambda^{\frac{1}{4}} (L_{\nabla g} + L_{\nabla g}^{\frac{1}{2}})}{t^{\frac{9}{8}}} + \frac{D_X \bar{B}}{t} + \frac{\bar{D}_\Lambda^{\frac{1}{2}}}{t^{\frac{1}{2}}}.$$

Remark 3. *(i) One can observe from Lemma 6 and the proof of Theorem 2 that the worst-case upper bound of the total number of inner iterations of Algorithm 1 is given by*

$$\sum_{k=0}^{2N+1} \left\lceil D_X \sqrt{\frac{2(C\rho_k + B + L_{\nabla g} \sum_{i=0}^{k-1} \sqrt{2\rho_i \eta_j})}{\eta_k}} \right\rceil$$

It can be shown that $\rho_k = \mathcal{O}((k+1)^{3/2})$ *and* $\eta_k = \mathcal{O}((k+1)^{-5/2} \min\{1, \sqrt{\epsilon}\})$ *minimize this quantity subject to the constraints given in* (27).

(ii) From Theorem 2, one can see that for any $\epsilon \in (0,1)$*, the first-order iteration complexity of Algorithm 1 for finding an* ϵ*-KKT solution of problem* (1) *is* $\mathcal{O}(\epsilon^{-7/4})$*, which is in the same order as the one for the I-AL method [11]. Nevertheless, Algorithm 1 is much more efficient than the latter method as observed in our numerical experiment. The main reason for this is perhaps that Algorithm 1 uses the dynamic* $\{\rho_k\}$ *and* $\{\eta_k\}$*, while I-AL method [11] uses the static ones through all iterations and also needs a "guess-and-check" procedure to approximate the unknown parameter* D_Λ.

Finally, the following theorem shows that for the ϵ-KKT solution (x^+, λ^+) outputted by Algorithm 1, its primal objective gap, dual objective gap, and constraint violation are at most $\mathcal{O}(\epsilon)$.

Theorem 3. *Consider the same setting as in Theorem 2. Then, the output* (x^+, λ^+) *of Algorithm 1 satisfies*

$$\mathrm{dist}(g(x^+), -\mathcal{K}) \le \epsilon, \qquad -C_1\epsilon \le F(x^+) - F^* \le (D_X + C_2 \cdot \max\{1, \epsilon\})\epsilon, \tag{29}$$

$$0 \le F^* - d(\lambda^+) \le (D_X + C_1 + C_2 \cdot \max\{1, \epsilon\})\epsilon, \tag{30}$$

where $C_1 = \mathrm{dist}(0, \Lambda^*)$ *and* $C_2 = 84\rho_0 \bar{D}_\Lambda + \|\hat{\lambda}^*\|$.

[7] For example, $\rho_k = \hat{C}(k+1)^{3/2}$ and $\eta_k = \tilde{C}(k+1)^{-5/2}$ satisfy (26) for any $\hat{C}$, $\tilde{C} > 0$.

4 An adaptively regularized I-AL method with improved iteration complexity

In this section, we propose an adaptively regularized first-order I-AL method and show that it achieves a significantly improved first-order iteration complexity than Algorithm 1 and existing first-order I-AL methods in the literature for finding an ϵ-KKT solution of (1). In particular, at each kth outer iteration it modifies Algorithm 1 by adding a regularization term $\|x - x^k\|^2/(2\rho_k)$ to the AL function $\mathcal{L}(x, \lambda^k; \rho_k)$ and also solving the AL subproblem to a higher accuracy. Moreover, it does not need a postprocessing stage. Since the regularization terms change dynamically, it is substantially different from those in [14, 11, 28].

Our adaptively regularized first-order I-AL method for problem (1) is presented as follows.

Algorithm 2 (An adaptively regularized I-AL method)**.**

0. Input $\epsilon > 0$, $(x^0, \lambda^0) \in \mathrm{dom}(P) \times \mathcal{K}^*$, nondecreasing $\{\rho_k\} \subset \Re_{++}$, and $0 < \eta_k \downarrow 0$. Set $k = 0$.
1. Apply Algorithm 4 (in Appendix B) to the problem $\min_x \varphi_k(x)$ to find $x^{k+1} \in \mathrm{dom}(P)$ satisfying
$$\mathrm{dist}(0, \partial\varphi_k(x^{k+1})) \le \eta_k,\text{[8]} \tag{31}$$
where
$$\varphi_k(x) = \mathcal{L}(x, \lambda^k; \rho_k) + \frac{1}{2\rho_k}\|x - x^k\|^2. \tag{32}$$
2. Set $\lambda^{k+1} = \Pi_{\mathcal{K}^*}\left(\lambda^k + \rho_k g(x^{k+1})\right)$.
3. If (x^{k+1}, λ^{k+1}) satisfies (5) or the following two inequalities are satisfied
$$\frac{1}{\rho_k}\|(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k)\| \le \frac{\epsilon}{2}, \quad \eta_k \le \frac{\epsilon}{2}, \tag{33}$$
output $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$ and terminate the algorithm.
4. Set $k \leftarrow k + 1$ and go to Step 1.

End.

For ease of later reference, we refer to the iterations of Algorithm 4 for solving the AL subproblems as the *inner iterations* of Algorithm 2, and call the update from (x^k, λ^k) to (x^{k+1}, λ^{k+1}) an *outer iteration* of Algorithm 2. Notice from (32) that φ_k is strongly convex with modulus $1/\rho_k$. The AL subproblem $\min_x \varphi_k(x)$ arising in Algorithm 2 can thus be suitably solved by Algorithm 4.

In the rest of this section, we present our main results for Algorithm 2, whose proof is deferred to Subsection 5.2. Before proceeding, we introduce some further notation that will be used subsequently. Let X^* be the set of optimal solutions of problem (1) and $\hat{x}^* \in X^*$ such that $\|x^0 - \hat{x}^*\| = \mathrm{dist}(x^0, X^*)$. In addition, we define

$$\bar{D}_X = \max\{D_X, 1\},\ D = \mathrm{dist}(x^0, X^*) + D_\Lambda,\ \bar{D} = \max\{D, 1\},\ \hat{B} = L_{\nabla f} + L_{\nabla g}\|\hat{\lambda}^*\| + L_{\nabla g} D, \tag{34}$$

where D_X, D_Λ and $\hat{\lambda}^*$ are defined in (24), and $L_{\nabla f}$ and $L_{\nabla g}$ are the Lipschitz constants of ∇f and ∇g on $\mathrm{dom}(P)$, respectively. The following theorem shows that Algorithm 2 with a suitable choice of $\{\rho_k\}$ and $\{\eta_k\}$ is guaranteed to find an ϵ-KKT solution of problem (1) in a finite number of outer iterations.

Theorem 4. *(i) If Algorithm 2 successfully terminates, then the output (x^+, λ^+) is an ϵ-KKT solution of problem* (1)*.*

(ii) Suppose that $\{\rho_k\}$ and $\{\eta_k\}$ satisfy that
$$\rho_k > 0 \textit{ is nondecreasing}, \quad 0 < \eta_k \downarrow 0, \quad \frac{\sum_{i=0}^{2k} \rho_i \eta_i}{\rho_k \sqrt{k+1}} \to 0.\text{[9]} \tag{35}$$
Then Algorithm 2 terminates in a finite number of outer iterations.

---

[8] In view of Proposition 4, one can terminate Algorithm 4 once a point x^{k+1} satisfying $2L_{\nabla\phi}\|\tilde{x}^{k+1} - x^{k+1}\| \le \eta_k$ is found, where $\phi(\cdot) = \varphi_k(\cdot) - P(\cdot)$, $\tilde{x}^{k+1} = \mathrm{prox}_{P/L_{\nabla\phi}}\left(x^{k+1} - \nabla\phi(x^{k+1})/L_{\nabla\phi}\right)$, and $L_{\nabla\phi}$ is the Lipschitz constant of $\nabla\phi$. Such x^{k+1} clearly satisfies (31).

[9] For example, $\rho_k = \rho_0\alpha^k$ and $\eta_k = \eta_0\beta^k$ satisfy (35) for any $\rho_0 > 0$, $\eta_0 > 0$, $\alpha > 1$ and $0 < \beta < 1/\alpha$.

(iii) Furthermore, if N is a nonnegative integer such that

$$\frac{D+\sum_{k=0}^{N}\rho_k\eta_k}{\rho_N}\le\frac{\epsilon}{2},\quad \eta_N\le\frac{\epsilon}{2}, \tag{36}$$

then Algorithm 2 terminates in at most $N+1$ outer iterations.

The next theorem provides an upper bound on the total number of the inner iterations Algorithm 2, that is, the total iterations of Algorithm 4 applied to solve all AL subproblems of Algorithm 2.

Theorem 5. *Let $\epsilon>0$ be given, and $\bar{D}_X$ and $\bar{D}$ be defined in* (34)*. Suppose that $\{\rho_k\}$ and $\{\eta_k\}$ are chosen as*

$$\rho_k=\rho_0\alpha^k,\quad \eta_k=\eta_0\beta^k \tag{37}$$

for some $\rho_0\ge 1$, $0<\eta_0\le 1$, $\alpha>1$, $0<\beta<1$ such that $\gamma=\alpha\beta<1$. Then, the total number of inner iterations of Algorithm 2 for finding an ϵ-KKT solution of problem (1) *is at most*

$$\mathcal{T}(\epsilon)=\left\lceil\frac{8\alpha^2\sqrt{\hat{C}\rho_0}}{\alpha-1}\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}\right\rceil\max\left\{1,\left\lceil\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}\log_\alpha\frac{2\alpha(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}\right\rceil\right\}, \tag{38}$$

where $\hat{C}=C\rho_0+\hat{B}+L_{\nabla g}\rho_0\eta_0/(1-\gamma)+1$, and C and $\hat{B}$ are defined in (25) *and* (34)*, respectively.*

Remark 4. *One can see from Theorem 5 that the first-order iteration complexity of Algorithm 2 for finding an ϵ-KKT solution of problem* (1) *is $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$, which significantly improves the previously best-known iteration-complexity $\mathcal{O}(\epsilon^{-7/4})$ achieved by first-order I-AL methods for finding an ϵ-KKT solution of* (1)*.*

Finally, the following theorem shows that for the ϵ-KKT solution (x^+,λ^+) outputted by Algorithm 2, its primal objective gap, dual objective gap, and constraint violation are at most $\mathcal{O}(\epsilon)$.

Theorem 6. *Consider the same setting as in Theorem 5. Then, the output (x^+,λ^+) of Algorithm 2 satisfies*

$$\mathrm{dist}(g(x^+),-\mathcal{K})\le\epsilon,\qquad -C_1\epsilon\le F(x^+)-F^*\le C_3\epsilon,\qquad 0\le F^*-d(\lambda^+)\le(C_1+C_3)\epsilon,$$

where $C_1=\mathrm{dist}(0,\Lambda^*)$ and $C_3=D_X+\|\hat{\lambda}^*\|+D+\rho_0\eta_0/(1-\gamma)$.

5 Proof of the main results

In this section we prove our main results presented in Sections 3 and 4, that is, Theorems 1-6. To proceed, we present some technical results that will be used subsequently in our proofs.

Recall that the AL function given in (11) can be written as $\mathcal{L}(x,\lambda;\rho)=\mathcal{S}(x,\lambda;\rho)+P(x)-\|\lambda^2\|/(2\rho)$, where $\mathcal{S}$ is defined in (12). The following lemma states some properties of the function $\mathcal{S}$, whose proof can be found in Appendix C.

Lemma 1. *Let $\mathcal{S}$ be defined in* (12)*. For any $(\lambda,\rho)\in\Re^m\times\Re_{++}$, the following statements hold.*

(i) $\mathcal{S}(x,\lambda;\rho)$ is convex and continuously differentiable in x and

$$\nabla_x\mathcal{S}(x,\lambda;\rho)=\nabla f(x)+\nabla g(x)\Pi_{\mathcal{K}^*}\big(\lambda+\rho g(x)\big). \tag{39}$$

(ii) $\nabla_x\mathcal{S}(x,\lambda;\rho)$ is Lipschitz continuous on $\mathrm{dom}(P)$ with a Lipschitz constant L given by

$$L=L_{\nabla f}+L_{\nabla g}\big(\|\lambda\|+\rho M_g\big)+\rho L_g^2.$$

The lemma below presents some key properties of the AL function, whose proof is a direct extension of the results in [20] and thus omitted.

Lemma 2. *For any $(x,\lambda,\rho)\in\Re^n\times\Re^m\times\Re_{++}$, the following identity holds*

$$\mathcal{L}(x,\lambda;\rho)=\max_{\eta\in\Re^m}\left\{l(x,\eta)-\frac{1}{2\rho}\|\eta-\lambda\|^2\right\}.$$

In addition, if $x\in\mathrm{dom}(P)$, the maximum is attained uniquely at $\bar{\lambda}=\Pi_{\mathcal{K}^*}(\lambda+\rho g(x))$. Consequently, the following statements hold.

(i) For any $(\lambda, \rho) \in \Re^m \times \Re_{++}$, $d(\lambda; \rho)$ satisfies that

$$d(\lambda; \rho) = \max_{\eta \in \Re^m} \left\{ d(\eta) - \frac{1}{2\rho} \|\eta - \lambda\|^2 \right\}. \tag{40}$$

(ii) $\mathcal{L}(x, \lambda; \rho)$ is a convex function in x, and for any $x \in \mathrm{dom}(P)$, we have

$$\partial_x \mathcal{L}(x, \lambda; \rho) = \partial_x l(x, \bar{\lambda}).$$

(iii) $\mathcal{L}(x, \lambda; \rho)$ is a concave function in λ, and for any $x \in \mathrm{dom}(P)$, it is differentiable in λ and

$$\frac{1}{\rho}(\bar{\lambda} - \lambda) = \nabla_\lambda \mathcal{L}(x, \lambda; \rho) \in \partial_\lambda l(x, \bar{\lambda}).$$

We next state some results for inexact proximal point algorithm (PPA), whose proof can be found in Appendix D.

Lemma 3. *Let $\mathcal{T} : \Re^n \rightrightarrows \Re^n$ be a maximally monotone operator and $z^* \in \Re^n$ such that $0 \in \mathcal{T}(z^*)$. Let $\{z^k\}$ be a sequence generated by an inexact PPA, starting with some z^0 and obtaining z^{k+1} by approximately evaluating $\mathcal{J}_{\rho_k}(z^k)$ such that*

$$\|z^{k+1} - \mathcal{J}_{\rho_k}(z^k)\| \le e_k \tag{41}$$

for some $\rho_k > 0$ and $e_k \ge 0$, where $\mathcal{J}_\rho = (\mathcal{I} + \rho \mathcal{T})^{-1}$ and $\mathcal{I}$ is the identify operator. Then we have

$$\|z^s - z^*\| \le \|z^t - z^*\| + \sum_{i=t}^{s-1} e_i, \quad \forall s \ge t \ge 0, \tag{42}$$

$$\|z^{k+1} - z^k\| \le \|z^0 - z^*\| + \sum_{i=0}^{k} e_i, \quad \forall k \ge 0. \tag{43}$$

Moreover, for any $K \ge 1$, we have

$$\min_{K \le k \le 2K} \|z^{k+1} - z^k\| \le \frac{\sqrt{2}\left(\|z^0 - z^*\| + 2\sum_{k=0}^{2K} e_k\right)}{\sqrt{K+1}}. \tag{44}$$

The following two lemmas show that Algorithms 1 and 2 can be viewed as an inexact PPA method applied to solve the monotone inclusion problems $0 \in \mathcal{T}_d(\lambda)$ and $0 \in \mathcal{T}_l(x, \lambda)$, respectively.

Lemma 4. *Let $\{\lambda^k\}$ be the sequence generated by Algorithm 1. Then for any $k \ge 0$, one has*

$$\|\lambda^{k+1} - \mathcal{J}_{\rho_k}(\lambda^k)\| \le \sqrt{2\rho_k \eta_k},$$

where $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k \mathcal{T}_d)^{-1}$ and $\mathcal{T}_d$ is defined in (15).

Proof. It follows from the definition of $\mathrm{dist}(\cdot, -\mathcal{K})$ that for any $\rho > 0$, $\lambda \in \Re^m$ and $x \in \mathrm{dom}(P)$,

$$\mathrm{dist}(\lambda + \rho g(x), -\mathcal{K}) = \min_u \{\|\lambda - u\| : \rho g(x) + u \preceq_{\mathcal{K}} 0\}. \tag{45}$$

By [24, Exercise 2.8], the minimum of (45) is attained uniquely at $\bar{u} = \lambda - \Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big)$. These together with (11) yield

$$\mathcal{L}(x^{k+1}, \lambda^k; \rho_k) = f(x^{k+1}) + P(x^{k+1}) + \frac{1}{2\rho_k}\left[\|\lambda^k - u^k\|^2 - \|\lambda^k\|^2\right], \tag{46}$$

where $u^k = \lambda^k - \Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^{k+1}))$. By this and Step 2 of Algorithm 1, we have $u^k = \lambda^k - \lambda^{k+1}$. Moreover, it follows from (13) and (45) that

$$d(\lambda^k; \rho_k) = \min_u \left\{ v(u) + \frac{1}{2\rho_k}\left[\|\lambda^k - u\|^2 - \|\lambda^k\|^2\right] \right\}, \tag{47}$$

where

$$v(u) = \min_x \{f(x) + P(x) : \rho_k g(x) + u \preceq_{\mathcal{K}} 0\}. \tag{48}$$

Since $f + P$ is convex and g is convex with respect to $\mathcal{K}$, it is not hard to see that v is also convex. Hence, the objective function in (47) is strongly convex in u and it has a unique minimizer $\bar{u}^k$. Claim that $\bar{u}^k = \lambda^k - \mathcal{J}_{\rho_k}(\lambda^k)$. Indeed, it follows from (47) and Danskin's theorem that $\nabla_\lambda d(\lambda^k; \rho_k) = -\bar{u}^k/\rho_k$. In addition, it follows from (40) and the definition of $\mathcal{J}_{\rho_k}(\lambda^k)$ that

$$d(\lambda^k; \rho_k) = \max_{\eta \in \Re^m} \left\{ d(\eta) - \frac{1}{2\rho_k}\|\eta - \lambda^k\|^2 \right\},$$

and the maximum is attained uniquely at $\mathcal{J}_{\rho_k}(\lambda^k)$. By these and Danskin's theorem, we obtain that $\nabla_\lambda d(\lambda^k; \rho_k) = (\mathcal{J}_{\rho_k}(\lambda^k) - \lambda^k)/\rho_k$, which together with $\nabla_\lambda d(\lambda^k; \rho_k) = -\bar{u}^k/\rho_k$ yields $\bar{u}^k = \lambda^k - \mathcal{J}_{\rho_k}(\lambda^k)$ as desired. By this, (46), (47) and (48), we obtain that

$$\begin{aligned} \mathcal{L}(x^{k+1}, \lambda^k; \rho_k) - d(\lambda^k; \rho_k) &= f(x^{k+1}) + P(x^{k+1}) + \frac{1}{2\rho_k}\|\lambda^k - u^k\|^2 - \min_u \left\{ v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2 \right\} \\ &\geq v(u^k) + \frac{1}{2\rho_k}\|\lambda^k - u^k\|^2 - \min_u \left\{ v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2 \right\} \qquad (49) \\ &\geq \frac{1}{2\rho_k}\|u^k - \bar{u}^k\|^2 = \frac{1}{2\rho_k}\|\mathcal{J}_{\rho_k}(\lambda^k) - \lambda^{k+1}\|^2, \qquad (50) \end{aligned}$$

where (49) follows from (48) and the fact that

$$\rho_k g(x^{k+1}) + u^k = \lambda^k + \rho_k g(x^{k+1}) - \Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^{k+1})) = \Pi_{-\mathcal{K}}(\lambda^k + \rho_k g(x^{k+1})) \preceq_{\mathcal{K}} 0,$$

and (50) follows from $\bar{u}^k = \arg\min_u \{v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2\}$, the fact that $v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2$ is strongly convex with modulus $1/\rho_k$, $u^k = \lambda^k - \lambda^{k+1}$, and $\bar{u}^k = \lambda^k - \mathcal{J}_{\rho_k}(\lambda^k)$. The conclusion then follows from (19) and (50). □

Lemma 5. *Let $\{(x^k, \lambda^k)\}$ be generated by Algorithm 2. For any $k \geq 0$, one has*

$$\|(x^{k+1}, \lambda^{k+1}) - \mathcal{J}_{\rho_k}(x^k, \lambda^k)\| \leq \rho_k \eta_k, \qquad (51)$$

where $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k \mathcal{T}_l)^{-1}$ and $\mathcal{T}_l$ is defined in (16).

Proof. By Lemma 2 and $\lambda^{k+1} = \Pi_{\mathcal{K}^*}\big(\lambda^k + \rho_k g(x^{k+1})\big)$, one has

$$\partial_x \mathcal{L}(x^{k+1}, \lambda^k; \rho_k) = \partial_x l(x^{k+1}, \lambda^{k+1}), \quad \frac{1}{\rho_k}(\lambda^{k+1} - \lambda^k) \in \partial_\lambda l(x^{k+1}, \lambda^{k+1}). \qquad (52)$$

By (31), there exists $\|v\| \leq \eta_k$ such that

$$v \in \partial_x \mathcal{L}(x^{k+1}, \lambda^k; \rho_k) + \frac{1}{\rho_k}(x^{k+1} - x^k).$$

This together with (52) implies that

$$x^k + \rho_k v \in \rho_k \partial_x l(x^{k+1}, \lambda^{k+1}) + x^{k+1}, \quad \lambda^k \in -\rho_k \partial_\lambda l(x^{k+1}, \lambda^{k+1}) + \lambda^{k+1}, \qquad (53)$$

which, by the definition of $\mathcal{T}_l$, are equivalent to

$$(x^k + \rho_k v, \lambda^k) \in (\mathcal{I} + \rho_k \mathcal{T}_l)(x^{k+1}, \lambda^{k+1}).$$

It follows from this and $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k \mathcal{T}_l)^{-1}$ that $(x^{k+1}, \lambda^{k+1}) = \mathcal{J}_{\rho_k}(x^k + \rho_k v, \lambda^k)$. By this and the non-expansion of $\mathcal{J}_{\rho_k}$, we obtain

$$\|(x^{k+1}, \lambda^{k+1}) - \mathcal{J}_{\rho_k}(x^k, \lambda^k)\| = \|\mathcal{J}_{\rho_k}(x^k + \rho_k v, \lambda^k) - \mathcal{J}_{\rho_k}(x^k, \lambda^k)\| \leq \|\rho_k v\| \leq \rho_k \eta_k,$$

which yields (51) as desired. □

5.1 Proof of the main results for Algorithm 1

In this subsection we provide a proof for Theorems 1, 2 and 3.

Proof of Theorem 1. (i) One can easily see that (x^+, λ^+) is an ϵ-KKT solution of (1) if Algorithm 1 terminates in Step 3. We now show that it is also an ϵ-KKT solution of (1) if Algorithm 1 terminates in Step 4. To this end, suppose that Algorithm 1 terminates in Step 4 at some iteration k, that is, the inequalities (20) hold for some k. For convenience, let $(\tilde{\lambda}, \tilde{\rho}, \tilde{x}) = (\lambda^k, \rho_k, x^{k+1})$. It then follows that $(x^+, \lambda^+) = \text{Postprocessing}(\tilde{\lambda}, \tilde{\rho}, \tilde{x}, \epsilon)$, and (22) and (23) hold for such $\tilde{\lambda}$ and $\tilde{\rho}$. By Definition 1, it suffices to show that $\text{dist}\big(0, \partial_x l(x^+, \lambda^+)\big) \le \epsilon$ and $\text{dist}\big(0, \partial_\lambda l(x^+, \lambda^+)\big) \le \epsilon$.

We start by showing $\text{dist}\big(0, \partial_x l(x^+, \lambda^+)\big) \le \epsilon$. For convenience, let $\varphi(x) = \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho})$. Notice from Lemma 1 that $\nabla_x \mathcal{S}(x, \tilde{\lambda}; \tilde{\rho})$ is Lipschitz continuous on $\text{dom}(P)$ with Lipschitz constant $\tilde{L}$, where $\tilde{L}$ is given in (21). Then, by (21), (22), (23) and Proposition 2 in Appendix B, one has $\varphi(x^+) \le \varphi(\hat{x})$ and

$$\text{dist}\big(0, \partial\varphi(x^+)\big) \le \sqrt{8\tilde{L}\big(\varphi(\hat{x}) - \min_{x\in\Re^n} \varphi(x)\big)} \le \sqrt{8\tilde{L}\tilde{\eta}} \le \epsilon. \tag{54}$$

In addition, it follows from (23) and Lemma 2 that

$$\partial\varphi(x^+) = \partial_x \mathcal{L}(x^+, \tilde{\lambda}; \tilde{\rho}) = \partial_x l\big(x^+, \Pi_{\mathcal{K}^*}\big(\tilde{\lambda} + \tilde{\rho} g(x^+)\big)\big) = \partial_x l(x^+, \lambda^+).$$

This together with (54) yields $\text{dist}\big(0, \partial_x l(x^+, \lambda^+)\big) \le \epsilon$. It remains to show $\text{dist}\big(0, \partial_\lambda l(x^+, \lambda^+)\big) \le \epsilon$. Let $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k \mathcal{T}_d)^{-1}$, where $\mathcal{T}_d$ is defined in (15). By (20) and Lemma 4, one has

$$\|\lambda^{k+1} - \mathcal{J}_{\rho_k}(\lambda^k)\| \le \sqrt{2\rho_k \eta_k} \le \frac{\rho_k \epsilon}{8}.$$

Using this and the first inequality in (20), we have

$$\|\lambda^k - \mathcal{J}_{\rho_k}(\lambda^k)\| \le \|\lambda^{k+1} - \lambda^k\| + \|\lambda^{k+1} - \mathcal{J}_{\rho_k}(\lambda^k)\| \le \frac{3\rho_k\epsilon}{4} + \frac{\rho_k\epsilon}{8} = \frac{7\rho_k\epsilon}{8},$$

which, together with $\tilde{\lambda} = \lambda^k$ and $\tilde{\rho} = \rho_k$, leads to $\|\tilde{\lambda} - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \le 7\tilde{\rho}\epsilon/8$. In addition, by $\varphi = \mathcal{L}(\cdot, \tilde{\lambda}; \tilde{\rho})$, the second relation in (23), and the same arguments as those for (50), one has

$$\|\lambda^+ - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \le \sqrt{2\tilde{\rho}\big(\mathcal{L}(x^+, \tilde{\lambda}; \tilde{\rho}) - \min_{x\in\Re^n} \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho})\big)} = \sqrt{2\tilde{\rho}\big(\varphi(x^+) - \min_{x\in\Re^n} \varphi(x)\big)}.$$

This, together with $\varphi(x^+) \le \varphi(\hat{x})$, (21) and (22), yields that

$$\|\lambda^+ - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \le \sqrt{2\tilde{\rho}\big(\varphi(\hat{x}) - \min_{x\in\Re^n} \varphi(x)\big)} \le \sqrt{2\tilde{\rho}\tilde{\eta}} \le \frac{\tilde{\rho}\epsilon}{8}.$$

Using this and $\|\tilde{\lambda} - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \le 7\tilde{\rho}\epsilon/8$, we obtain that

$$\|\lambda^+ - \tilde{\lambda}\| \le \|\tilde{\lambda} - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| + \|\lambda^+ - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \le \frac{7\tilde{\rho}\epsilon}{8} + \frac{\tilde{\rho}\epsilon}{8} = \tilde{\rho}\epsilon. \tag{55}$$

Moreover, by Lemma 2 and the second relation in (23), one has $(\lambda^+ - \tilde{\lambda})/\tilde{\rho} \in \partial_\lambda l(x^+, \lambda^+)$. This along with (55) implies $\text{dist}\big(0, \partial_\lambda l(x^+, \lambda^+)\big) \le \epsilon$.

(ii) Suppose for contradiction that Algorithm 1 does not terminate in a finite number of iterations. It then follows that (20) does not hold for any k. By Lemma 4 and (44), one has that for any $k \ge 1$,

$$\min_{k\le i\le 2k} \|\lambda^{i+1} - \lambda^i\| \le \frac{\sqrt{2}\left(D_\Lambda + 2\sum_{i=0}^{2k} \sqrt{2\rho_i\eta_i}\right)}{\sqrt{k+1}},$$

where D_Λ is defined in (24). Since $\{\rho_k\}$ is assumed to be nondecreasing, we further have

$$\min_{k\le i\le 2k} \frac{1}{\rho_i}\|\lambda^{i+1} - \lambda^i\| \le \frac{\sqrt{2}\left(D_\Lambda + 2\sum_{i=0}^{2k} \sqrt{2\rho_i\eta_i}\right)}{\rho_k\sqrt{k+1}}. \tag{56}$$

By this and (26), one has that $\min_{k\le i\le 2k}\|\lambda^{i+1}-\lambda^i\|/\rho_i\to 0$ and $\eta_k/\rho_k\to 0$ as $k\to\infty$, which imply that (20) is satisfied for some k and thus leads to a contradiction.

(iii) Suppose for contradiction that Algorithm 1 does not terminate within $2N+1$ outer iterations. It then follows that (20) does not hold for all $0\le k\le 2N$. Let $\tilde{k}\in\text{Argmin}_{N\le k\le 2N}\|\lambda^{k+1}-\lambda^k\|/\rho_k$. By this, (56), and the assumption that $\{\rho_k\}$ is nondecreasing and $\{\eta_k\}$ is decreasing, one has

$$\frac{1}{\rho_{\tilde{k}}}\|\lambda^{\tilde{k}+1}-\lambda^{\tilde{k}}\|\le\frac{\sqrt{2}\left(D_\Lambda+2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}\right)}{\rho_N\sqrt{N+1}}<\frac{3}{4}\epsilon,\qquad \frac{\eta_{\tilde{k}}}{\rho_{\tilde{k}}}\le\frac{\eta_N}{\rho_N}\le\frac{\epsilon^2}{128}.$$

Hence, (20) holds for $k=\tilde{k}\le 2N$, which leads to a contradiction. □

To prove Theorems 2 and 3, we need the following result.

Lemma 6. *For any $k\ge 0$, the Lipschitz constant of $\nabla_x\mathcal{S}(x,\lambda^k;\rho_k)$, denoted as L_k, satisfies*

$$L_k\le C\rho_k+B+L_{\nabla g}\sum_{i=0}^{k-1}\sqrt{2\rho_i\eta_i},\tag{57}$$

where B and C are given in (24) *and* (25)*, respectively.*

Proof. By Lemma 1, one has $L_k\le L_{\nabla f}+L_{\nabla g}\big(\|\lambda^k\|+\rho_k M_g\big)+\rho_k L_g^2$. Combining (42) with Lemma 4, and using (24), we have

$$\|\lambda^k\|\le\|\hat{\lambda}^*\|+\|\lambda^k-\hat{\lambda}^*\|\le\|\hat{\lambda}^*\|+D_\Lambda+\sum_{i=0}^{k-1}\sqrt{2\rho_i\eta_i},\tag{58}$$

where $\hat{\lambda}^*$ is defined right above (24). By these and the definitions of B and C, one obtains (57). □

We are now ready to prove Theorems 2 and 3.

Proof of Theorem 2. For convenience, let $\epsilon_0=\min\{1,\epsilon\}$. Let $\bar{N}$ be the number of outer iterations of Algorithm 1. Also, let $\mathcal{I}_k$ and $\mathcal{I}_p$ be the number of iterations executed by Algorithm 3 at the kth outer iteration of Algorithm 1 and in the subroutine Postprocessing, respectively. In addition, let T be the total number of inner iterations of Algorithm 1. Clearly, we have $T=\sum_{k=0}^{\bar{N}-1}\mathcal{I}_k+\mathcal{I}_p$. In what follows, we first derive upper bounds on $\bar{N}$, $\mathcal{I}_k$ and $\mathcal{I}_p$, and then use this formula to obtain an upper bound on T.

First, we derive an upper bound on $\bar{N}$. By (28), we have that $\eta_k=\eta_0(k+1)^{-5/2}\sqrt{\epsilon_0}$ for any $k\ge 0$. Hence, for any $K\ge 0$, it holds that

$$\sum_{k=0}^{K}\sqrt{2\rho_k\eta_k}=\sqrt{2\rho_0\eta_0}\;\epsilon_0^{\frac{1}{4}}\sum_{k=0}^{K}(k+1)^{-\frac{1}{2}}\le 2\sqrt{2\rho_0\eta_0}\;\epsilon_0^{\frac{1}{4}}\sqrt{K+1},\tag{59}$$

where the inequality follows from $\sum_{k=0}^{K}(k+1)^{-1/2}\le 2\sqrt{K+1}$. Let $\gamma=7\bar{D}_\Lambda^{1/2}\epsilon_0^{-1/2}$ and $N=\lceil\gamma\rceil$. It follows from (28), (59), and $\gamma\le N\le\gamma+1$ that

$$\frac{D_\Lambda+2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}}{\rho_N\sqrt{N+1}}\le\frac{\bar{D}_\Lambda+4\sqrt{2\rho_0\eta_0}\;\epsilon_0^{\frac{1}{4}}\sqrt{2N+1}}{\rho_0(N+1)^2}\le\frac{\bar{D}_\Lambda}{\rho_0(N+1)^2}+\frac{8\eta_0^{\frac{1}{2}}\epsilon_0^{\frac{1}{4}}}{\rho_0^{\frac{1}{2}}(N+1)^{\frac{3}{2}}}.\tag{60}$$

Notice that

$$\frac{\bar{D}_\Lambda}{\rho_0(N+1)^2}\le\frac{\bar{D}_\Lambda}{\rho_0\gamma^2}=\frac{\bar{D}_\Lambda}{\rho_0(49\bar{D}_\Lambda\epsilon_0^{-1})}=\frac{\epsilon_0}{49\rho_0}\le\frac{\epsilon}{49},$$

where the first inequality is by $\gamma\le N+1$ and the last inequality follows from $\rho_0\ge 1$ and $\epsilon_0\le\epsilon$. Also, by $\bar{D}_\Lambda\ge 1$, we have $\gamma\ge 7\epsilon_0^{-1/2}$. This together with $\gamma\le N+1$, $\rho_0\ge 1$, and $\eta_0\le 1$ yields

$$\frac{8\eta_0^{\frac{1}{2}}\epsilon_0^{\frac{1}{4}}}{\rho_0^{\frac{1}{2}}(N+1)^{\frac{3}{2}}}\le\frac{8\epsilon_0^{\frac{1}{4}}}{\gamma^{\frac{3}{2}}}\le\frac{8\epsilon_0^{\frac{1}{4}}}{7^{\frac{3}{2}}\epsilon_0^{-\frac{3}{4}}}=\frac{8\epsilon_0}{7^{\frac{3}{2}}}<\frac{4\epsilon_0}{9}\le\frac{4\epsilon}{9},$$

Substituting the above two inequalities into (60), one has

$$\frac{D_\Lambda + 2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}}{\rho_N\sqrt{N+1}} < \frac{\epsilon}{2}. \tag{61}$$

In addition, using $N+1 \geq \gamma \geq 7\epsilon_0^{-1/2}$, (28), $\epsilon_0 \leq 1$, $\rho_0 \geq 1$ and $\eta_0 \leq 1$, we obtain that

$$\frac{\eta_N}{\rho_N} = \frac{\eta_0\epsilon_0^{\frac{1}{2}}}{\rho_0(N+1)^4} \leq \frac{1}{7^4\epsilon_0^{-2}} = \frac{\epsilon_0^2}{7^4} < \frac{\epsilon^2}{128}. \tag{62}$$

By (61), (62) and Theorem 1 (iii), we obtain

$$\bar{N} \leq 2N+1 = 2\left\lceil 7\bar{D}_\Lambda^{\frac{1}{2}}\epsilon_0^{-\frac{1}{2}}\right\rceil + 1. \tag{63}$$

Second, we derive an upper bound on $\mathcal{I}_k$. Let L_k be the Lipschitz constant of $\nabla_x\mathcal{S}(x,\lambda^k;\rho_k)$. It follows from (57) and (28) that for any $k \geq 0$,

$$L_k \leq \bar{C}\rho_0(k+1)^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}L_{\nabla g}(k+1)^{\frac{1}{2}}. \tag{64}$$

This, together with Proposition 3, (19) and (28), yields that

$$\begin{aligned}
\mathcal{I}_k &\leq \left\lceil D_X\sqrt{\frac{2L_k}{\eta_k}}\right\rceil \leq 1 + \sqrt{2}D_X\sqrt{\frac{\bar{C}\rho_0(k+1)^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}L_{\nabla g}(k+1)^{\frac{1}{2}}}{\eta_0(k+1)^{-\frac{5}{2}}\epsilon_0^{\frac{1}{2}}}} \\
&\leq 1 + \sqrt{2}D_X\sqrt{\frac{\bar{C}\rho_0(k+1)^{\frac{3}{2}} + \bar{B}\rho_0 + 2\sqrt{2}\rho_0\epsilon_0^{\frac{1}{4}}L_{\nabla g}(k+1)^{\frac{1}{2}}}{\eta_0(k+1)^{-\frac{5}{2}}\epsilon_0^{\frac{1}{2}}}} \\
&\leq 1 + D_X\sqrt{\frac{2\rho_0}{\eta_0}}\left(\bar{C}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}(k+1)^2 + \bar{B}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}(k+1)^{\frac{5}{4}} + 2L_{\nabla g}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{8}}(k+1)^{\frac{3}{2}}\right),
\end{aligned} \tag{65}$$

where the third inequality is due to $\rho_0 \geq 1$ and $\eta_0 \leq 1$, and the last inequality follows by $\sqrt{a+b+c} \leq \sqrt{a}+\sqrt{b}+\sqrt{c}$ for any $a,b,c \geq 0$.

Third, we derive an upper bound on $\mathcal{I}_p$. Recall that $\bar{N}$ is the number of outer iterations, that is, (20) is satisfied when $k = \bar{N}-1$. It then follows from Algorithm 1 that $(\tilde{\lambda},\tilde{\rho}) = (\lambda^{\bar{N}-1},\rho_{\bar{N}-1})$ and $\tilde{L} = L_{\bar{N}-1}$. By these, Proposition 3, (21), (22) and $\epsilon_0 \leq \epsilon$, we have

$$\mathcal{I}_p \leq \left\lceil D_X\sqrt{\frac{2L_{\bar{N}-1}}{\tilde{\eta}}}\right\rceil \leq \left\lceil\frac{16D_X}{\epsilon_0}\cdot\max\left\{\sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}}, \frac{L_{\bar{N}-1}}{4}\right\}\right\rceil \tag{66}$$

In addition, it follows from (64) that

$$L_{\bar{N}-1} \leq \bar{C}\rho_0\bar{N}^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}L_{\nabla g}\bar{N}^{\frac{1}{2}}. \tag{67}$$

By this and (28), we obtain that for any $\bar{N} \geq 1$,

$$\begin{aligned}
\sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}} &\leq \sqrt{\frac{\bar{C}\rho_0\bar{N}^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}L_{\nabla g}\bar{N}^{\frac{1}{2}}}{\rho_0\bar{N}^{\frac{3}{2}}}} \leq \sqrt{\bar{C} + \bar{B} + 2\sqrt{2}\epsilon_0^{\frac{1}{4}}L_{\nabla g}} \\
&\leq \bar{C}^{\frac{1}{2}} + \bar{B}^{\frac{1}{2}} + 2\epsilon_0^{\frac{1}{8}}L_{\nabla g}^{\frac{1}{2}},
\end{aligned} \tag{68}$$

where the second inequality uses $\bar{N} \geq 1$, $\rho_0 \geq 1$ and $\eta_0 \leq 1$, and the last inequality follows by $\sqrt{a+b+c} \leq \sqrt{a}+\sqrt{b}+\sqrt{c}$ for any $a,b,c \geq 0$. By (67), (68), $\epsilon_0 \leq 1$, $\eta_0 \leq 1$, $\bar{C} \geq 1$ and $\bar{B} \geq 1$, it is not hard to verify that for all $\bar{N} \geq 1$,

$$\max\left\{\sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}}, \frac{L_{\bar{N}-1}}{4}\right\} \leq \bar{C}\rho_0\bar{N}^{\frac{3}{2}} + \bar{B} + 2\rho_0^{\frac{1}{2}}\epsilon_0^{\frac{1}{8}}\bar{N}^{\frac{1}{2}}(L_{\nabla g} + L_{\nabla g}^{\frac{1}{2}}). \tag{69}$$

Substituting (69) into (66), we arrive at

$$\mathcal{I}_p \leq 1 + \frac{16 D_X}{\epsilon_0}\left(\bar{C}\rho_0 \bar{N}^{\frac{3}{2}} + \bar{B} + 2\rho_0^{\frac{1}{2}}\epsilon_0^{\frac{1}{8}}\bar{N}^{\frac{1}{2}}(L_{\nabla g} + L_{\nabla g}^{\frac{1}{2}}).\right). \tag{70}$$

Finally, we use (63), (65) and (70) to derive an upper bound on the overall complexity T. By (63), $N = \lceil \gamma \rceil$ and $\gamma \geq 7$, one has $\bar{N} - 1 \leq 2N \leq 2\gamma + 2 \leq 3\gamma - 1$. This together with (65) yields that

$$\begin{aligned}
\sum_{k=0}^{\bar{N}-1} \mathcal{I}_k &\leq 3\gamma + D_X\sqrt{\frac{2\rho_0}{\eta_0}}\left(\bar{C}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}\sum_{k=0}^{\lfloor 3\gamma\rfloor - 1}(k+1)^2 + \bar{B}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}\sum_{k=0}^{\lfloor 3\gamma\rfloor - 1}(k+1)^{\frac{5}{4}} + 2L_{\nabla g}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{8}}\sum_{k=0}^{\lfloor 3\gamma\rfloor - 1}(k+1)^{\frac{3}{2}}\right) \\
&\leq 3\gamma + D_X\sqrt{\frac{2\rho_0}{\eta_0}}\left(\frac{8}{3}\bar{C}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}(3\gamma)^3 + \frac{2^{\frac{17}{4}}}{9}\bar{B}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}(3\gamma)^{\frac{9}{4}} + \frac{2^{\frac{9}{2}}}{5}L_{\nabla g}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{8}}(3\gamma)^{\frac{5}{2}}\right) \\
&\leq 3\gamma + 72 D_X\sqrt{\frac{2\rho_0}{\eta_0}}\left(\bar{C}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}\gamma^3 + \bar{B}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{4}}\gamma^{\frac{9}{4}} + L_{\nabla g}^{\frac{1}{2}}\epsilon_0^{-\frac{1}{8}}\gamma^{\frac{5}{2}}\right),
\end{aligned}$$

where the second inequality is due to

$$\sum_{k=0}^{K-1}(k+1)^\alpha \leq \frac{1}{1+\alpha}(K+1)^{1+\alpha} \leq \frac{2^{1+\alpha}}{1+\alpha}K^{1+\alpha}, \quad \forall \alpha > 0, K \geq 1.$$

Recall that $\gamma = 7\bar{D}_\Lambda^{1/2}\epsilon_0^{-1/2}$. Substituting this into the above inequality, we obtain

$$\sum_{k=0}^{\bar{N}-1}\mathcal{I}_k = \mathcal{O}\left(\frac{D_X \bar{D}_\Lambda^{3/2}\bar{C}^{1/2}}{\epsilon_0^{7/4}} + \frac{D_X \bar{D}_\Lambda^{9/8}\bar{B}^{\frac{1}{2}} + D_X \bar{D}_\Lambda^{5/4}L_{\nabla g}^{1/2}}{\epsilon_0^{11/8}} + \frac{\bar{D}_\Lambda^{1/2}}{\epsilon_0^{1/2}}\right).$$

In addition, by $\bar{N} \leq 3\gamma$, $\gamma = 7\bar{D}_\Lambda^{1/2}\epsilon_0^{-1/2}$ and (70), we obtain that

$$\mathcal{I}_p = \mathcal{O}\left(\frac{D_X \bar{D}_\Lambda^{3/4}\bar{C}}{\epsilon_0^{7/4}} + \frac{D_X \bar{B}}{\epsilon_0} + \frac{D_X \bar{D}_\Lambda^{1/4}(L_{\nabla g} + L_{\nabla g}^{1/2})}{\epsilon_0^{9/8}}\right).$$

Recall that $T = \sum_{k=0}^{\bar{N}-1}\mathcal{I}_k + \mathcal{I}_p$. By these, $\bar{D}_\Lambda \geq 1$, $\bar{C} \geq 1$, and $\bar{B} \geq 1$, we have

$$T = \mathcal{O}\left(\frac{D_X \bar{D}_\Lambda^{3/2}\bar{C}}{\epsilon_0^{7/4}} + \frac{D_X \bar{D}_\Lambda^{5/4}\bar{B}^{1/2}(1 + L_{\nabla g}^{1/2})}{\epsilon_0^{11/8}} + \frac{D_X \bar{D}_\Lambda^{1/4}(L_{\nabla g} + L_{\nabla g}^{1/2})}{\epsilon_0^{9/8}} + \frac{D_X \bar{B}}{\epsilon_0} + \frac{\bar{D}_\Lambda^{1/2}}{\epsilon_0^{1/2}}\right).$$

This together with $\epsilon_0 = \min\{1, \epsilon\}$ yields the complexity bound in Theorem 2. □

Proof of Theorem 3. Recall from Theorem 1 that the output (x^+, λ^+) of Algorithm 1 is an ϵ-KKT solution of (1). It then follows from Remark 1 that $\mathrm{dist}(g(x^+), \mathcal{N}_{\mathcal{K}^*}(\lambda^+)) \leq \epsilon$. It also follows from Definition 1 that there exist $u \in \partial_x l(x^+, \lambda^+)$ and $v \in \partial_\lambda l(x^+, \lambda^+)$ with $\|u\| \leq \epsilon$ and $\|v\| \leq \epsilon$.

We start by proving the first inequality in (29). It is not hard to verify that $\mathcal{N}_{\mathcal{K}^*}(\lambda^+) \subseteq -\mathcal{K}$, which together with $\mathrm{dist}(g(x^+), \mathcal{N}_{\mathcal{K}^*}(\lambda^+)) \leq \epsilon$ implies that $\mathrm{dist}(g(x^+), -\mathcal{K}) \leq \epsilon$ holds as desired.

We next prove the second inequality in (29). Let $\lambda^* \in \Lambda^*$ be such that $\|\lambda^*\| = \mathrm{dist}(0, \Lambda^*)$. Also, let x^* be any optimal solution of problem (1). Then, (x^*, λ^*) is a saddle point of l, which implies that $l(x^+, \lambda^*) - l(x^*, \lambda^*) \geq 0$ and $F^* = l(x^*, \lambda^*)$. By these and $l(x^+, \lambda^*) = F(x^+) + \langle \lambda^*, g(x^+)\rangle$, we obtain

$$F(x^+) - F^* \geq -\langle \lambda^*, g(x^+)\rangle. \tag{71}$$

In addition, by $\lambda^* \in \mathcal{K}^*$ and $\mathrm{dist}(g(x^+), -\mathcal{K}) \leq \epsilon$, one has that

$$\langle \lambda^*, g(x^+)\rangle = \langle \lambda^*, g(x^+) - \Pi_{-\mathcal{K}}(g(x^+))\rangle + \underbrace{\langle \lambda^*, \Pi_{-\mathcal{K}}(g(x^+))\rangle}_{\leq 0} \leq \|\lambda^*\| \cdot \mathrm{dist}(g(x^+), -\mathcal{K}) \leq \|\lambda^*\|\epsilon,$$

which, along with (71) and $C_1 = \mathrm{dist}(0, \Lambda^*) = \|\lambda^*\|$, implies that $F(x^+) - F^* \geq -C_1\|\epsilon$.

We now prove the last inequality in (29). Recall that $u \in \partial_x l(x^+, \lambda^+)$ and $v \in \partial_\lambda l(x^+, \lambda^+)$ with $\|u\| \le \epsilon$ and $\|v\| \le \epsilon$. By (10), $\lambda^+ \in \mathcal{K}^*$ and Assumption 1, there exists an $\tilde{x} \in \mathrm{dom}(P)$ satisfying $d(\lambda^+) = l(\tilde{x}, \lambda^+)$. These, together with the convexity of $l(\cdot, \lambda^+)$, $x^+ \in \mathrm{dom}(P)$ and (24), imply that

$$l(x^+, \lambda^+) - d(\lambda^+) = l(x^+, \lambda^+) - l(\tilde{x}, \lambda^+) \le \langle x^+ - \tilde{x}, u\rangle \le \|x^+ - \tilde{x}\|\|u\| \le D_X \epsilon. \tag{72}$$

Also, using $v \in \partial_\lambda l(x^+, \lambda^+)$ and (17), we obtain that $g(x^+) - v \in \mathcal{N}_{\mathcal{K}^*}(\lambda^+)$. Then, by the definition of $\mathcal{N}_{\mathcal{K}^*}(\lambda^+)$ and the fact that $\mathcal{K}^*$ is a closed convex cone, it is not hard to see that $\langle g(x^+) - v, \lambda^+\rangle = 0$. Thus, we obtain

$$\langle \lambda^+, g(x^+)\rangle = \langle \lambda^+, v\rangle \ge -\|\lambda^+\|\|v\| \ge -\epsilon\|\lambda^+\|. \tag{73}$$

Let $\bar{N}$ be the number of outer iterations of Algorithm 1. Claim that

$$\|\lambda^+\| \le \rho_{\bar{N}-1}\epsilon + \|\hat{\lambda}^*\| + D_\Lambda + \sum_{i=0}^{\bar{N}-1} \sqrt{2\rho_i \eta_i}. \tag{74}$$

Indeed, due to (58), (74) clearly holds if Algorithm 1 terminates in Step 3. We now assume that it terminates in Step 4. Then $(x^+, \lambda^+) = \mathrm{Postprocessing}(\lambda^{\bar{N}-1}, \rho_{\bar{N}-1}, x^{\bar{N}}, \epsilon)$. This together with (55) and (58) yields that

$$\|\lambda^+\| \le \|\lambda^+ - \lambda^{\bar{N}-1}\| + \|\lambda^{\bar{N}-1}\| \le \rho_{\bar{N}-1}\epsilon + \|\hat{\lambda}^*\| + D_\Lambda + \sum_{i=0}^{\bar{N}-2} \sqrt{2\rho_i \eta_i},$$

and hence (74) holds. By (28), (63), $\bar{D}_\Lambda \ge 1$, and $\epsilon_0 = \min\{1, \epsilon\}$, we obtain

$$\rho_{\bar{N}-1}\epsilon = \rho_0 \bar{N}^{\frac{3}{2}}\epsilon \le \rho_0 \left(17\bar{D}_\Lambda^{\frac{1}{2}}\epsilon_0^{-\frac{1}{2}}\right)^{\frac{3}{2}} \epsilon \le 71\rho_0 \bar{D}_\Lambda \cdot \max\{1, \epsilon\}.$$

Also, by (59), (63), $\bar{D}_\Lambda \ge 1$, $\rho_0 \ge 1$ and $\eta_0 \le 1$, we obtain

$$\sum_{i=0}^{\bar{N}-1} \sqrt{2\rho_i\eta_i} \le 2\sqrt{2\rho_0\eta_0}\epsilon_0^{\frac{1}{4}}\sqrt{\bar{N}} \le 2\sqrt{2\rho_0\eta_0}\epsilon_0^{\frac{1}{4}}\sqrt{17\bar{D}_\Lambda^{\frac{1}{2}}\epsilon_0^{-\frac{1}{2}}} \le 12\sqrt{\rho_0\eta_0}\bar{D}_\Lambda^{\frac{1}{4}} \le 12\rho_0\bar{D}_\Lambda.$$

These, together with (74), $\rho_0 \ge 1$ and $D_\Lambda \le \bar{D}_\Lambda$, yield that $\|\lambda^+\| \le C_2 \cdot \max\{1, \epsilon\}$, where $C_2 = 84\rho_0\bar{D}_\Lambda + \|\hat{\lambda}^*\|$. By this, (72), (73), and $l(x^+, \lambda^+) = F(x^+) + \langle \lambda^+, g(x^+)\rangle$, we obtain

$$F(x^+) - d(\lambda^+) \le D_X\epsilon - \langle \lambda^+, g(x^+)\rangle \le (D_X + \|\lambda^+\|)\epsilon \le (D_X + C_2 \cdot \max\{1, \epsilon\})\epsilon, \tag{75}$$

which along with $d(\lambda^+) \le F^*$ results in the last inequality in (29).

Finally, the first inequality in (30) trivially holds, while the second inequality in (30) follows from (75) and the second inequality in (29). □

5.2 Proof of the main results for Algorithm 2

In this subsection we provide a proof for Theorems 4, 5, and 6.

Proof of Theorem 4. (i) Suppose that Algorithm 2 terminates in Step 3 at some iteration k. It then follows that (x^+, λ^+) is already an ϵ-KKT solution of problem (1) or the inequalities (33) hold for such k. We next show that for the latter case, $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$ is also an ϵ-KKT solution of (1). To this end, suppose that (33) holds at the kth iteration. Notice that (53) holds for some $\|v\| \le \eta_k$, which yields

$$\frac{1}{\rho_k}(x^k - x^{k+1}) + v \in \partial_x l(x^{k+1}, \lambda^{k+1}), \quad \frac{1}{\rho_k}(\lambda^{k+1} - \lambda^k) \in \partial_\lambda l(x^{k+1}, \lambda^{k+1}).$$

By these, (33) and $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$, we obtain

$$\begin{aligned}
\mathrm{dist}(0, \partial_x l(x^+, \lambda^+)) &\le \frac{1}{\rho_k}\|x^{k+1} - x^k - \rho_k v\| \le \frac{1}{\rho_k}\|x^{k+1} - x^k\| + \|v\| \le \epsilon,\\
\mathrm{dist}(0, \partial_\lambda l(x^+, \lambda^+)) &\le \frac{1}{\rho_k}\|\lambda^{k+1} - \lambda^k\| \le \epsilon,
\end{aligned}$$

which along with Definition 1 imply that (x^+, λ^+) is an ϵ-KKT solution of problem (1).

(ii) Suppose for contradiction that Algorithm 2 does not terminate in a finite number of iterations. It then follows that (33) does not hold for any k. By (44) and Lemma 5, one has that

$$\min_{k\le i\le 2k} \|(x^{i+1}, \lambda^{i+1}) - (x^i, \lambda^i)\| \le \frac{\sqrt{2}\left(D + 2\sum_{i=0}^{2k} \rho_i\eta_i\right)}{\sqrt{k+1}},$$

where D is defined in (34). Since $\{\rho_k\}$ is assumed to be nondecreasing, we further have

$$\min_{k\le i\le 2k} \frac{1}{\rho_i}\|(x^{i+1}, \lambda^{i+1}) - (x^i, \lambda^i)\| \le \frac{\sqrt{2}\left(D + 2\sum_{i=0}^{2k} \rho_i\eta_i\right)}{\rho_k\sqrt{k+1}}.$$

By this and (35), one has that $\min_{k\le i\le 2k} \|(x^{i+1}, \lambda^{i+1}) - (x^i, \lambda^i)\|/\rho_i \to 0$ and $\eta_k \to 0$ as $k \to \infty$, which implies that (33) is satisfied for some k and thus leads to a contradiction.

(iii) Let $\hat{x}^*$ and $\hat{\lambda}^*$ be defined right above (24) and (34), respectively. Suppose for contradiction that Algorithm 2 does not terminate within $N+1$ outer iterations. It then follows that (33) does not hold for all $0 \le k \le N$. Then, by (34), (43), and (51), we have that for all $0 \le k \le N$,

$$\|(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k)\| \le \|(x^0, \lambda^0) - (\hat{x}^*, \hat{\lambda}^*)\| + \sum_{i=0}^{k} \rho_i\eta_i \le D + \sum_{i=0}^{k} \rho_i\eta_i,$$

where D is given in (34). By this and (36), one can see that (33) is satisfied when $k = N$, which leads to a contradiction. □

To prove Theorems 5 and 6, we need the following result.

Lemma 7. *Let $s_k(x) = \mathcal{S}(x, \lambda^k; \rho_k) + \|x - x^k\|^2/(2\rho_k)$. Then s_k is continuously differentiable, and moreover, ∇s_k is Lipschitz continuous on* $\mathrm{dom}(P)$ *with a Lipschitz constant L_k given by*

$$L_k = C\rho_k + \hat{B} + L_{\nabla g}\sum_{i=0}^{k-1} \rho_i\eta_i + \rho_k^{-1}, \tag{76}$$

where C and $\hat{B}$ are defined in (25) *and* (34)*, respectively.*

Proof. By the definition of $s_k(x)$ and Lemma 1, one has

$$\|\nabla s_k(x) - \nabla s_k(y)\| \le \left(L_{\nabla f} + L_{\nabla g}(\|\lambda^k\| + \rho_k M_g) + \rho_k L_g^2 + \rho_k^{-1}\right)\|x - y\|, \quad \forall x, y \in \mathrm{dom}(P). \tag{77}$$

In addition, it follows from Lemmas 3 and 5 that

$$\|(x^k, \lambda^k) - (\hat{x}^*, \hat{\lambda}^*)\| \le \|(x^0, \lambda^0) - (\hat{x}^*, \hat{\lambda}^*)\| + \sum_{i=0}^{k-1} \rho_i\eta_i,$$

where $\hat{x}^*$ and $\hat{\lambda}^*$ are defined right above (24) and (34), respectively. Hence, we have that

$$\|\lambda^k\| \le \|\hat{\lambda}^*\| + \|\lambda^k - \hat{\lambda}^*\| \le \|\hat{\lambda}^*\| + D + \sum_{i=0}^{k-1} \rho_i\eta_i, \tag{78}$$

where D is given in (34). Substituting this into (77), and using the definitions of $\hat{B}$ and C, we obtain that $\|\nabla s_k(x) - \nabla s_k(y)\| \le L_k\|x - y\|$ for all $x, y \in \mathrm{dom}(P)$. Hence, the conclusion holds. □

We are now ready to provide a proof for Theorems 5 and 6.

Proof of Theorem 5. Let $\bar{N}$ be the number of outer iterations of Algorithm 2, and let $\mathcal{I}_k$ be the number of first-order iterations executed by Algorithm 4 at the kth outer iteration of Algorithm 2. In addition, let T be the total number of first-order inner iterations of Algorithm 2. Clearly, we have $T = \sum_{k=0}^{\bar{N}-1} \mathcal{I}_k$. In what follows, we first derive upper bounds on $\bar{N}$ and $\mathcal{I}_k$, and then use this formula to obtain an upper bound on T.

We first derive an upper bound on $\bar{N}$. Due to (37) and $0<\gamma<1$, we have that

$$\sum_{k=0}^{K}\rho_k\eta_k=\rho_0\eta_0\sum_{k=0}^{K}\gamma^k\le\rho_0\eta_0\sum_{k=0}^{\infty}\gamma^k=\frac{\rho_0\eta_0}{1-\gamma},\quad\forall K\ge 0. \tag{79}$$

Let

$$N=\max\left\{1,\left\lceil\log_\alpha\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}\right\rceil\right\}.$$

Since $N\ge\log_\alpha\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}$, we have from (37) that

$$\rho_N\ge\frac{2\rho_0(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}.$$

By this, (79), $D\le\bar{D}$, and $\rho_0\ge 1$, we obtain

$$\frac{D+\sum_{k=0}^{N}\rho_k\eta_k}{\rho_N}\le\frac{\bar{D}+\frac{\rho_0\eta_0}{1-\gamma}}{\frac{2\rho_0(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}}=\frac{\epsilon}{2}\cdot\frac{\bar{D}(1-\gamma)+\rho_0\eta_0}{\rho_0(\bar{D}+\rho_0\eta_0)}\le\frac{\epsilon}{2}\cdot\frac{\bar{D}+\rho_0\eta_0}{\bar{D}+\rho_0\eta_0}=\frac{\epsilon}{2}.$$

In addition, one can observe that $1<\alpha<\beta^{-1}$ and $\bar{D}+\rho_0\eta_0\ge 1-\gamma$. By these, we have

$$N\ge\log_\alpha\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}\ge\log_{\beta^{-1}}\frac{2}{\epsilon},$$

which, together with (37), $\beta<1$ and $\eta_0\le 1$, implies that $\eta_N\le\epsilon/2$. It then follows from these and Theorem 4 (iii) that

$$\bar{N}\le N+1\le\max\left\{1,\left\lceil\log_\alpha\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}\right\rceil\right\}+1. \tag{80}$$

We next derive an upper bound on $\mathcal{I}_k$. By (37), (76), $\alpha>1$ and $\rho_0\ge 1$, one has that for any $k\ge 0$,

$$L_k\le C\rho_0\alpha^k+\hat{B}+\frac{L_{\nabla g}\rho_0\eta_0}{1-\gamma}+\frac{1}{\rho_0\alpha^k}\le\hat{C}\alpha^k,$$

where $\hat{C}=C\rho_0+\hat{B}+L_{\nabla g}\rho_0\eta_0/(1-\gamma)+1$. Notice that $\varphi_k(x)$ is strongly convex with modulus $\mu_k=1/\rho_k$. By this, (31), $\rho_k=\rho_0\alpha^k$, $\hat{C}\ge 1$, $\bar{D}_X\ge 1$, $\alpha>1$, $\beta<1$, $\rho_0\ge 1$, $\eta_0\le 1$, and Proposition 4 (in Appendix B), we obtain that for any $k\ge 0$,

$$\begin{aligned}\mathcal{I}_k&\le\left\lceil\sqrt{\frac{L_k}{\mu_k}}\right\rceil\max\left\{1,\left\lceil 2\log\frac{2L_kD_X}{\eta_k}\right\rceil\right\}\le\left\lceil\sqrt{\hat{C}\rho_0}\ \alpha^k\right\rceil\max\left\{1,\left\lceil 2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k}\right\rceil\right\}\\&\le\left\lceil\sqrt{\hat{C}\rho_0}\ \alpha^k\right\rceil\left\lceil 2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k}\right\rceil\le\left(\sqrt{\hat{C}\rho_0}\ \alpha^k+1\right)\left(2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k}+1\right)\\&\le 8\sqrt{\hat{C}\rho_0}\ \alpha^k\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k}\le 8\sqrt{\hat{C}\rho_0}\ k\alpha^k\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta},\end{aligned} \tag{81}$$

where the third and fifth inequalities follow from $\sqrt{\hat{C}\rho_0}\ \alpha^k\ge 1$ and $2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k}\ge 2\log 2\ge 1$.

Finally, we derive an upper bound on T. By (81), one has

$$T=\sum_{k=0}^{\bar{N}-1}\mathcal{I}_k\le 8\sqrt{\hat{C}\rho_0}\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}\sum_{k=0}^{\bar{N}-1}k\alpha^k\le\frac{8\sqrt{\hat{C}\rho_0}}{\alpha-1}\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}(\bar{N}-1)\alpha^{\bar{N}}, \tag{82}$$

where the last inequality is due to $\sum_{k=0}^{K}k\alpha^k\le K\alpha^{K+1}/(\alpha-1)$ for any $K\ge 0$. We divide the rest of the proof into the following two cases.

Case (a): $\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon} \geq \alpha$. This along with (80) implies that $\bar{N} \leq \log_\alpha \frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon} + 2$. By this and (82), one has

$$T \leq \frac{8\sqrt{\hat{C}\rho_0}}{\alpha - 1} \log \frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta} \log_\alpha \frac{2\alpha(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon} \cdot \frac{2\alpha^2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}.$$

Case (b): $\frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon} < \alpha$. This together with (80) implies that $\bar{N} \leq 2$. By this and (82), one has

$$T \leq \frac{8\alpha^2\sqrt{\hat{C}\rho_0}}{\alpha - 1} \log \frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}.$$

Combining the results in the above two cases, we obtain (38) as desired. □

Proof of Theorem 6. Let $\bar{N}$ be the number of outer iterations of Algorithm 2. By (78) and $\lambda^+ = \lambda^{\bar{N}}$, one has $\|\lambda^+\| \leq \|\hat{\lambda}^*\| + D + \sum_{i=0}^{\bar{N}-1}\rho_i\eta_i$, where $\hat{\lambda}^*$ is defined right above (34) and D is given in (34). This together with (37) and (79) yields that $\|\lambda^+\| \leq \|\hat{\lambda}^*\| + D + \rho_0\eta_0/(1-\gamma)$. The rest of the proof follows the same arguments as those in the proof of Theorem 3. □

6 Numerical results

In this section we conduct some preliminary numerical experiments to test the performance of our proposed algorithms (Algorithms 1 and 2), and compare them with a closely related I-AL method and its modified version proposed in [11], which are named as I-AL$_1$ and I-AL$_2$ respectively for ease of reference. In particular, we apply all these algorithms to the linear programming (LP) problem

$$\min_{x\in\Re^n} \left\{c^T x : Ax = b,\ l \leq x \leq u\right\} \tag{83}$$

for some $A \in \Re^{m\times n}$, $b \in \Re^m$, $c \in \Re^n$, and $l, u \in \Re$. It is clear that (83) is a special case of problem (1) with $f(x) = c^T x$, P being the indicator function of the set $\{x \in \Re^n : l \leq x \leq u\}$, $g(x) = Ax - b$, and $\mathcal{K} = \{0\}^m$. All the algorithms are coded in Matlab and all the computations are performed on a Dell desktop with a 3.40-GHz Intel Core i7-3770 processor and 16 GB of RAM.

In our experiment, we choose $\epsilon = 0.01$ for all the aforementioned algorithms. In addition, the parameters $\{\rho_k\}$ and $\{\eta_k\}$ of these algorithms are set as follows. For Algorithm 1, we set them by (28) with $\rho_0 = 100$ and $\eta_0 = 1$. For Algorithm 2, we choose them by (37) with $\rho_0 = 100$, $\eta_0 = 0.1$, $\alpha = 1.1$ and $\beta = 0.8$. For the algorithms I-AL$_1$ and I-AL$_2$, we choose $\{\rho_k\}$ and $\{\eta_k\}$ as described in [11] and set $t_0 = 1$ as the initial value in their "guess-and-check" procedures.

We randomly generate 20 instances for problem (83), each of which is generated by a similar manner as described in [10]. In particular, given positive integers $m < n$ and a scalar $0 < \zeta \leq 1$, we first randomly generate a matrix $A \in \Re^{m\times n}$ with density ζ, whose entries are randomly chosen from the standard normal distribution.[10] We then generate a vector $x \in \Re^n$ with entries randomly chosen from the uniform distribution on $[-5, 5]$ and set $b = Ax$. Also, we generate a vector $c \in \Re^n$ with entries randomly chosen from the standard normal distribution. Finally, we randomly choose l and u from the uniform distribution on $[-10, -5]$ and $[5, 10]$, respectively.

The computational results of all the algorithms for solving problem (83) with the above 20 instances are presented in Table 1. In detail, the parameters n, m, and ζ of each instance are listed in the first three columns, respectively. For each instance, the total number of first-order iterations and the CPU time (in seconds) for these algorithms are given in the next four columns and the last four columns, respectively. One can observe that Algorithm 2 performs best in terms of both number of iterations and CPU time, which is not surprising as it has the lowest first-order iteration complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$ among these algorithms. In addition, although Algorithm 1 and I-AL$_1$ share the same order of first-order iteration complexity $\mathcal{O}(\epsilon^{-7/4})$, one can observe that the practical performance of Algorithm 1 is substantially better than that of I-AL$_1$. The main reason is perhaps that Algorithm 1 uses the dynamic $\{\rho_k\}$ and $\{\eta_k\}$, while I-AL$_1$ uses the static ones through all iterations and also needs a "guess-and-check" procedure for approximating the unknown parameter D_Λ. Finally, we observe that I-AL$_2$ performs much better than I-AL$_1$ and generally better than Algorithm 1, but it is substantially outperformed by Algorithm 2.

[10] The matrix A is generated via the Matlab command A = sprandn(m,n,ζ).

Table 1: Computational results for solving problem (83)

Parameters			Iterations ($\times 10^3$)				CPU Time (in seconds)			
n	m	ζ	Algorithm 1	Algorithm 2	I-AL$_1$	I-AL$_2$	Algorithm 1	Algorithm 2	I-AL$_1$	I-AL$_2$
1,000	100	0.01	5	13	164	52	0.7	0.9	18.8	6.6
1,000	100	0.05	8	13	200	23	1.2	1.2	31.5	3.8
1,000	100	0.10	8	16	200	25	1.8	2.0	41.7	5.4
1,000	500	0.01	22	16	200	30	3.8	1.7	33.7	5.3
1,000	500	0.05	23	19	300	35	10.8	6.3	136.9	16.5
1,000	500	0.10	22	15	300	22	17.5	8.9	237.2	17.0
1,000	900	0.01	150	20	900	77	35.2	3.0	208.0	18.6
1,000	900	0.05	124	19	1,100	64	94.3	10.7	876.0	51.8
1,000	900	0.10	132	21	600	49	197.2	23.9	903.3	71.0
5,000	500	0.01	19	27	200	78	17.2	13.6	181.0	74.0
5,000	500	0.05	20	31	200	49	46.5	49.9	505.1	126.9
5,000	500	0.10	19	26	200	42	129.9	149.6	1,357.3	288.3
5,000	2,500	0.01	79	20	300	49	225.8	40.5	852.1	140.7
5,000	2,500	0.05	80	27	300	61	1,706.4	505.1	6,406.2	1,309.8
5,000	2,500	0.10	81	31	300	54	3,577.7	1,240.9	13,324.2	2,530.2
5,000	4,500	0.01	400	27	1,400	191	2,953.1	167.9	10,364.8	1,425.8
5,000	4,500	0.05	406	29	1,300	207	17,724.6	1,067.8	55,608.2	8,812.9
5,000	4,500	0.10	300	32	1,200	172	26,489.9	2,449.3	104,523.0	15,002.9
10,000	1,000	0.01	27	30	200	54	76.7	52.2	572.8	157.0
10,000	5,000	0.01	116	29	400	111	1,988.5	406.6	6,895.0	1,931.0

7 Concluding remarks

In this paper our analyses of the I-AL methods rely on the assumption that the domain of the function P is compact. One natural question is whether this assumption can be dropped. In addition, can the first-order iteration complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$ for finding an ϵ-KKT solution of problem (1) be further improved for an I-AL method? These will be left for the future research.

A An example about the dependence of D_Λ^ϵ on ϵ

In this part we present an example to demonstrate $D_\Lambda^\epsilon = \Theta(1/\epsilon)$, where $D_\Lambda^\epsilon = \min\{\|\lambda^0 - \lambda\| : \lambda \in \Lambda_\epsilon^*\}$ and Λ_ϵ^* is the set of optimal Lagrangian multipliers associated with the perturbed problem (8).

Consider the linear programming problem:

$$\begin{aligned} \min_{x\in X} \quad & x_1 + x_2 - \delta x_3 \\ \text{s.t.} \quad & x_1 = 1,\ \delta x_2 = -\delta \end{aligned} \tag{84}$$

for some $\delta \in (0,1)$, where

$$X = \left\{ x \in \Re^3 : -x_2 + \frac{\delta}{1-\delta} x_3 \le \frac{1}{1-\delta},\ -2 \le x_1, x_2, x_3 \le 2 \right\}.$$

We also consider a perturbed problem for (84) given by

$$\begin{aligned} \min_{x\in X} \quad & x_1 + x_2 - \delta x_3 + \frac{\epsilon}{4D_X}\|x - x^0\|^2 \\ \text{s.t.} \quad & x_1 = 1,\ \delta x_2 = -\delta, \end{aligned} \tag{85}$$

where $x^0 = (1,-1,-1)^T \in X$, $\epsilon = 2D_X\delta$, and $D_X = \max\{\|x-y\| : x, y \in X\}$. It is clear that (84) and (85) are a special instance of problems (7) and (8), respectively. In addition, one can verify that for any $\delta \in (0,1)$, $\bar{x}^\epsilon = (1,-1,0)^T$ is the optimal solution of (85) and $\bar{\lambda}^\epsilon = (1, 2D_X/\epsilon)^T$ is the unique optimal

Lagrangian multiplier associated with the constraints $x_1 = 1$ and $\delta x_2 = -\delta$ of (85). Assume without loss of generality that $\lambda^0 = 0$. Then, for any $\delta \in (0,1)$, we have $D_\Lambda^\epsilon = \|\lambda^0 - \bar{\lambda}^\epsilon\| = \epsilon^{-1}\sqrt{\epsilon^2 + 4D_X^2}$, whose dependence on ϵ is roughly $\Theta(\epsilon^{-1})$ when ϵ is small.

B Optimal first-order methods for unconstrained convex optimization problems

In this part we review optimal first-order methods for solving a class of convex optimization problems in the form of

$$\Psi^* = \min_{x\in\Re^n} \left\{\Psi(x) := \phi(x) + h(x)\right\}, \tag{86}$$

where $\phi, h : \Re^n \to (-\infty, \infty]$ are closed convex functions, ϕ is continuously differentiable on an open set containing $\mathrm{dom}(h)$, and $\nabla\phi$ is Lipschitz continuous with Lipschitz constant $L_{\nabla\phi}$ on $\mathrm{dom}(h)$. In addition, assume that $\mathrm{dom}(h)$ is compact and let $D_h := \max_{x,y\in\mathrm{dom}(h)} \|x-y\|$.

We first state a property of problem (86), which is used in the proof of some main results of this paper. Its proof follows from some standard arguments and is omitted due to page limit.

Proposition 2. *For any $x \in \mathrm{dom}(h)$, we have $\Psi(x^+) \le \Psi(x)$ and*

$$\mathrm{dist}\big(0, \partial\Psi(x^+)\big) \le 2L_{\nabla\phi}\|x^+ - x\| \le \sqrt{8L_{\nabla\phi}\big(\Psi(x) - \Psi^*\big)},$$

where $x^+ = \mathrm{prox}_{h/L_{\nabla\phi}}\big(x - \nabla\phi(x)/L_{\nabla\phi}\big)$.

We next present an optimal first-order method for solving (86) in which ϕ is convex but not necessarily strongly convex. It is a variant of Nesterov's optimal first-order methods [15, 16] and has been studied in, for example, [26, Section 3].

Algorithm 3 (An optimal first-order method for (86) with general convex ϕ)**.**

0. Input $x^0 = z^0 \in \mathrm{dom}(h)$. Set $k = 0$.
1. Set $y^k = (kx^k + 2z^k)/(k+2)$.
2. Compute z^{k+1} as
$$z^{k+1} = \arg\min_z \left\{\ell(z; y^k) + \frac{L_{\nabla\phi}}{k+2}\|z - z^k\|^2\right\},$$
where
$$\ell(x;y) := \phi(y) + \langle\nabla\phi(y), x - y\rangle + h(x). \tag{87}$$
3. Set $x^{k+1} = (kx^k + 2z^{k+1})/(k+2)$.
4. Set $k \leftarrow k+1$ and go to Step 1.

End.

The following result provides an iteration-complexity of Algorithm 3 for finding an ϵ-optimal solution of (86). It is an immediate consequence of [26, Corollary 1] and its proof is thus omitted.

Proposition 3. *Let $\{(x^k, y^k)\}$ be generated by Algorithm 3 and $\ell(\cdot;\cdot)$ be defined in (87). Then, $\Psi(x^k) - \Psi^* \le \Psi(x^k) - \underline{\Psi}_k$ for all $k \ge 1$. Moreover, for any given $\epsilon > 0$, Algorithm 3 finds an approximate solution x^k of problem (86) such that $\Psi(x^k) - \Psi^* \le \Psi(x^k) - \underline{\Psi}_k \le \epsilon$ in no more than $K(\epsilon)$ iterations, where*

$$K(\epsilon) = \left\lceil D_h\sqrt{\frac{2L_{\nabla\phi}}{\epsilon}}\right\rceil, \qquad \underline{\Psi}_k = \frac{4}{k(k+2)}\min_x\left\{\sum_{i=0}^{k-1}\frac{i+2}{2}\ell(x;y^i)\right\}. \tag{88}$$

Remark 5. *Observe from (88) that $\underline{\Psi}_k = \frac{2}{k(k+2)}\left[v_k + \min_x\left\{\langle u^k, x\rangle + h(x)\right\}\right]$, where*

$$u^k = \sum_{i=0}^{k-1}(i+2)\nabla\phi(y^i), \quad v_k = \sum_{i=0}^{k-1}(i+2)\left(\phi(y^i) - \langle\nabla\phi(y^i), y^i\rangle\right).$$

Note that (u^k, v_k) can be recursively and thus cheaply computed. Once (u^k, v_k) is available, computing $\underline{\Psi}_k$ only requires solving the problem $\min_x\left\{\langle u^k, x\rangle + h(x)\right\}$, which typically has a closed-form solution.

We now turn to consider the case of problem (86) in which ϕ is strongly convex, that is, there exists a constant $\mu \in (0, L_{\nabla\phi})$ such that

$$\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \geq \mu\|x-y\|^2, \quad \forall x, y \in \mathrm{dom}(h). \tag{89}$$

We next propose a slight variant of Nesterov's optimal method [16, 13] for solving problem (86) with a strongly convex ϕ.

Algorithm 4 (An optimal first-order method for (86) with strongly convex ϕ)**.**

0. Input $x^{-1} \in \mathrm{dom}(h)$, $L_{\nabla\phi} > 0$ and $0 < \mu < L_{\nabla\phi}$. Compute

$$x^0 = \mathrm{prox}_{h/L_{\nabla\phi}}\left(x^{-1} - \frac{1}{L_{\nabla\phi}}\nabla\phi(x^{-1})\right). \tag{90}$$

   Set $z^0 = x^0$, $\alpha = \sqrt{\mu/L_{\nabla\phi}}$ and $k = 0$.

1. Set $y^k = (x^k + \alpha z^k)/(1+\alpha)$.
2. Compute z^{k+1} as

$$z^{k+1} = \arg\min_z \left\{\ell(z; y^k) + \frac{\alpha L_{\nabla\phi}}{2}\|z - \alpha y^k - (1-\alpha)z^k\|^2\right\},$$

   where $\ell(x; y)$ is defined in (87).

3. Set $x^{k+1} = (1-\alpha)x^k + \alpha z^{k+1}$.
4. Set $k \leftarrow k+1$ and go to Step 1.

End.

Remark 6. *Algorithm 4 differs from Nesterov's optimal method [16, 13] in that it executes a proximal step* (90) *to generate x^0 while the latter method simply sets $x^0 = x^{-1}$.*

We next state an iteration-complexity result for Algorithm 4 for finding an approximate optimal solution of problem (86). Its proof follows from [13, Theorem 1], Proposition 2 and some standard arguments, and is omitted due to page limit.

Proposition 4. *Suppose that* (89) *holds. Let $\{x^k\}$ be the sequence generated by Algorithm 4 and $\tilde{x}^k = \mathrm{prox}_{h/L_{\nabla\phi}}\left(x^k - \nabla\phi(x^k)/L_{\nabla\phi}\right)$ for all $k \geq 0$. Then, $\mathrm{dist}(0, \partial\Psi(\tilde{x}^k)) \leq 2L_{\nabla\phi}\|\tilde{x}^k - x^k\|$ for all $k \geq 0$. Moreover, for any given $\epsilon > 0$, an approximate solution $\tilde{x}^k$ of problem* (86) *satisfying $\mathrm{dist}(0, \partial\Psi(\tilde{x}^k)) \leq 2L_{\nabla\phi}\|\tilde{x}^k - x^k\| \leq \epsilon$ is generated by running Algorithm 4 for at most $\widetilde{K}(\epsilon)$ iterations, where*

$$\widetilde{K}(\epsilon) = \left\lceil \sqrt{\frac{L_{\nabla\phi}}{\mu}}\, \right\rceil \max\left\{1, \left\lceil 2\log\frac{2L_{\nabla\phi}D_h}{\epsilon}\right\rceil\right\}.$$

C Proof of Lemma 1

Proof. (i) We first show that $\mathcal{S}(\cdot, \lambda; \rho)$ is convex. Let $x, x' \in \Re^n$ and $\alpha \in [0,1]$ be arbitrarily chosen. Using (2) and the relation

$$\lambda + \rho g\big(\alpha x + (1-\alpha)x'\big) = \lambda + \rho[\alpha g(x) + (1-\alpha)g(x')] + \rho\left(g\big(\alpha x + (1-\alpha)x'\big) - [\alpha g(x) + (1-\alpha)g(x')]\right),$$

we see that $\lambda + \rho\big[\alpha g(x) + (1-\alpha)g(x')\big] \preceq_{-\mathcal{K}} \lambda + \rho g\big(\alpha x + (1-\alpha)x'\big)$. By this and the fact that $\mathcal{K}$ is a convex cone, it is not hard to show that

$$\mathrm{dist}^2\Big(\lambda + \rho g(\alpha x + (1-\alpha)x'), -\mathcal{K}\Big) \leq \mathrm{dist}^2\Big(\lambda + \rho[\alpha g(x) + (1-\alpha)g(x')], -\mathcal{K}\Big). \tag{91}$$

In addition, by the convexity of $\mathrm{dist}^2(\cdot, -\mathcal{K})$, one has

$$\begin{aligned}\mathrm{dist}^2\Big(\lambda + \rho[\alpha g(x) + (1-\alpha)g(x')], -\mathcal{K}\Big) &= \mathrm{dist}^2\Big(\alpha\big(\lambda + \rho g(x)\big) + (1-\alpha)\big(\lambda + \rho g(x')\big), -\mathcal{K}\Big) \\ &\leq \alpha \cdot \mathrm{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big) + (1-\alpha)\mathrm{dist}^2\big(\lambda + \rho g(x'), -\mathcal{K}\big),\end{aligned}$$

which along with (91) leads to

$$\mathrm{dist}^2\Big(\lambda + \rho g\big(\alpha x + (1-\alpha)x'\big), -\mathcal{K}\Big) \leq \alpha \cdot \mathrm{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big) + (1-\alpha)\mathrm{dist}^2\big(\lambda + \rho g(x'), -\mathcal{K}\big).$$

It thus follows that $\mathrm{dist}^2(\lambda+\rho g(\cdot), -\mathcal{K})$ is convex. This together with the convexity of f implies that $\mathcal{S}(\cdot, \lambda; \rho)$ is convex. Next we show that $\mathcal{S}(\cdot, \lambda; \rho)$ is continuously differentiable. By the definition of $\mathrm{dist}(\cdot, -\mathcal{K})$, one has

$$\mathcal{S}(x, \lambda; \rho) = f(x) + \frac{1}{2\rho} \min_{v \in -\mathcal{K}} \|\lambda + \rho g(x) - v\|^2,$$

where the minimum is attained uniquely at $v = \Pi_{-\mathcal{K}}\big(\lambda + \rho g(x)\big)$. Using Danskin's theorem (e.g., see [3]), we conclude that $\mathcal{S}(x, \lambda; \rho)$ is differentiable in x and

$$\nabla_x \mathcal{S}(x, \lambda; \rho) = \nabla f(x) + \nabla g(x)\big[\lambda + \rho g(x) - \Pi_{-\mathcal{K}}\big(\lambda + \rho g(x)\big)\big] = \nabla f(x) + \nabla g(x)\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big),$$

where the second equality follows from [24, Exercise 2.8].

(ii) Recall that ∇f, ∇g and g are Lipschitz continuous on $\mathrm{dom}(P)$. By this and (39), we have that for any $x, x' \in \mathrm{dom}(P)$,

$$\begin{aligned}&\|\nabla_x \mathcal{S}(x, \lambda; \rho) - \nabla_x \mathcal{S}(x', \lambda; \rho)\| = \|\nabla f(x) + \nabla g(x)\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big) - \nabla f(x') - \nabla g(x')\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x')\big)\| \\ &\leq \|\nabla g(x)\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big) - \nabla g(x')\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x')\big)\| + \|\nabla f(x) - \nabla f(x')\| \\ &\leq L_{\nabla g}\|x - x'\|\|\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big)\| + \|\nabla g(x')\|\|\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big) - \Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x')\big)\| + L_{\nabla f}\|x - x'\| \\ &\leq L_{\nabla g}\|x - x'\|\|\lambda + \rho g(x)\| + \rho L_g\|g(x) - g(x')\| + L_{\nabla f}\|x - x'\| \\ &\leq \big(L_{\nabla g}(\|\lambda\| + \rho M_g) + \rho L_g^2 + L_{\nabla f}\big)\|x - x'\|\end{aligned}$$

where the third inequality is due to the non-expansiveness of the projection operator $\Pi_{\mathcal{K}^*}$ and $\|\nabla g(x')\| \leq L_g$, and the last one follows from $\|g(x)\| \leq M_g$ and the Lipschitz continuity of g on $\mathrm{dom}(P)$. □

D Proof of Lemma 3

Proof. Since $0 \in \mathcal{T}(z^*)$ and $\mathcal{T}$ is maximally monotone, it follows from [21, Proposition 1] that

$$\|\mathcal{J}_\rho(z) - z^*\|^2 + \|\mathcal{J}_\rho(z) - z\|^2 \leq \|z - z^*\|^2, \tag{92}$$

which implies that

$$\|\mathcal{J}_\rho(z) - z^*\| \leq \|z - z^*\|, \quad \|\mathcal{J}_\rho(z) - z\| \leq \|z - z^*\|, \quad \forall z \in \Re^n. \tag{93}$$

Let $\xi^k = z^{k+1} - \mathcal{J}_{\rho_k}(z^k)$ for all $k \geq 0$. By this, (93), and (41) with $\rho = \rho_k$ and $z = z^k$, one has

$$\|z^{k+1} - z^*\| \leq \|z^{k+1} - \mathcal{J}_{\rho_k}(z^k)\| + \|\mathcal{J}_{\rho_k}(z^k) - z^*\| \leq \|\xi^k\| + \|z^k - z^*\|, \quad \forall k \geq 0.$$

Summing up the above inequality from $k = t$ to $k = s - 1$ yields

$$\|z^s - z^*\| \leq \|z^t - z^*\| + \sum_{i=t}^{s-1} \|\xi^i\|, \quad \forall s \geq t \geq 0. \tag{94}$$

Notice from (41) that $\|\xi^k\| \leq e_k$ for all $k \geq 0$, which along with (94) leads to (42). Besides, by (93) with $\rho = \rho_k$ and $z = z^k$, one has $\|\mathcal{J}_{\rho_k}(z^k) - z^k\| \leq \|z^k - z^*\|$. These together with (41) and (42) yield

$$\|z^{k+1} - z^k\| \leq \|z^{k+1} - \mathcal{J}_{\rho_k}(z^k)\| + \|\mathcal{J}_{\rho_k}(z^k) - z^k\| \leq e_k + \|z^k - z^*\| \leq \|z^0 - z^*\| + \sum_{i=0}^{k} e_i.$$

In addition, by the definition of ξ^k, and (92) with $\mathcal{J} = \mathcal{J}_{\rho_k}$ and $z = z^k$, one has

$$\begin{aligned}\|\mathcal{J}_{\rho_k}(z^k) - z^k\|^2 &\le \|z^k - z^*\|^2 - \|\mathcal{J}_{\rho_k}(z^k) - z^*\|^2 = \|z^k - z^*\|^2 - \|\mathcal{J}_{\rho_k}(z^k) - z^{k+1} + z^{k+1} - z^*\|^2 \\ &\le \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 - \|\xi^k\|^2 + 2\|\xi^k\|\|z^{k+1} - z^*\|.\end{aligned}$$

Summing up the above inequality from $k = K$ to $k = 2K$ and using (94), we obtain that

$$\begin{aligned}\sum_{k=K}^{2K} \|\mathcal{J}_{\rho_k}(z^k) - z^k\|^2 &\le \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\sum_{k=K}^{2K} \|\xi^k\| \left(\|z^K - z^*\| + \sum_{j=K}^{k} \|\xi^j\|\right) \\ &= \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + 2\sum_{k=K}^{2K}\sum_{j=K}^{k} \|\xi^k\|\|\xi^j\| \\ &= \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + \sum_{k=K}^{2K} \|\xi^k\|^2 + \left(\sum_{k=K}^{2K} \|\xi^k\|\right)^2 \\ &= \|z^K - z^*\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + \left(\sum_{k=K}^{2K} \|\xi^k\|\right)^2 \\ &= \left(\|z^K - z^*\| + \sum_{k=K}^{2K} \|\xi^k\|\right)^2 \le \left(\|z^0 - z^*\| + \sum_{k=0}^{2K} \|\xi^k\|\right)^2, \qquad (95)\end{aligned}$$

where (95) follows from (94) with $t = 0$ and $s = K$. Again, by the definition of ξ^k, one has

$$\|z^{k+1} - z^k\|^2 = \|\mathcal{J}_{\rho_k}(z^k) + \xi^k - z^k\|^2 \le 2\left(\|\mathcal{J}_{\rho_k}(z^k) - z^k\|^2 + \|\xi^k\|^2\right).$$

This together with (95) yields

$$\begin{aligned}\sum_{k=K}^{2K} \|z^{k+1} - z^k\|^2 &\le 2\sum_{k=K}^{2K} \|\mathcal{J}_{\rho_k}(z^k) - z^k\|^2 + 2\sum_{k=K}^{2K} \|\xi^k\|^2 \le 2\left(\|z^0 - z^*\| + \sum_{k=0}^{2K} \|\xi^k\|\right)^2 + 2\sum_{k=0}^{2K} \|\xi^k\|^2 \\ &\le 2\left(\|z^0 - z^*\| + 2\sum_{k=0}^{2K} \|\xi^k\|\right)^2,\end{aligned}$$

which along with $\|\xi^k\| \le e_k$ leads to (44). The proof is then completed. □

References

[1] N. S. Aybat and G. Iyengar. An augmented Lagrangian method for conic convex programming. arXiv:1302.6322, preprint, 2013.

[2] D. P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Academic press, 1982.

[3] D. P. Bertsekas. Nonlinear Programming. Athena Scientific, 1999.

[4] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, 95(2):329–357, 2003.

[5] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen. An augmented Lagrangian method for total variation video restoration. *IEEE Trans. Image Process.*, 20(11):3097–3111, 2011.

[6] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2): 37–75, 2014.

[7] M. R. Hestenes. Multiplier and gradient methods. *J. Optim. Theory Appl.*, 4(5):303–320, 1969.

[8] K. Ito and K. Kunisch. The augmented Lagrangian method for parameter estimation in elliptic systems. *SIAM J. Control Optim.*, 28(1):113–136, 1990.

[9] F. Jarre and F. Rendl. An augmented primal-dual method for linear conic programs. *SIAM J. Optim.*, 19(2):808–823, 2008.

[10] G. Lan, Z. Lu, and R. D. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Math. Program.*, 126(1):1–29, 2011.

[11] G. Lan and R. D. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.*, 155(1-2):511–547, 2016.

[12] Y.-F. Liu, X. Liu, and S. Ma. On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Math. Oper. Res.*, 44(2):632–650, 2019.

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

[14] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optim. Methods Softw.*, 34(2):305–335, 2019.

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

[16] Y. Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Springer, 2004.

[17] J. Nocedal and S. Wright. Numerical Optimization, 2nd Edition. Springer, 2006.

[18] A. Patrascu, I. Necoara, and T.-D. Quoc. Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optim. Lett.*, 11(3):609–626, 2017.

[19] M. J. D. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, R. Fletcher (ed.), Academic Press, page 283–298, 1969.

[20] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.

[21] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[22] R. T. Rockafellar. Convex Analysis. Princeton University Press, 1997.

[23] R. T. Rockafellar and R. J.-B. Wets. Variational Analysis. Springer, 2009.

[24] A. P. Ruszczyński. Nonlinear Optimization. Princeton University Press, 2006.

[25] A. Shapiro and J. Sun. Some properties of the augmented Lagrangian in cone constrained optimization. *Math. Oper. Res.*, 29(3):479–491, 2004.

[26] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Unpublished manuscript, 2008.

[27] Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *INFORMS Journal on Optimization*, 3(1):89–117, 2021.

[28] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Program.*, 185:199–244, 2021.

[29] X.-Y. Zhao, D. Sun, and K.-C. Toh. A Newton-CG augmented Lagrangian method for semidefinite programming. *SIAM J. Optim.*, 20(4):1737–1765, 2010.