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

Zhaosong Lu* Zirui Zhou†

March 26, 2018

Abstract

In this paper we consider a class of convex conic programming. In particular, we propose an inexact augmented Lagrangian (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 computing an ϵ-KKT solution is at most $\mathcal{O}(\epsilon^{-7/4})$. We also propose a modified I-AL method and show that it has an improved iteration-complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$, which is so far the lowest complexity bound among all first-order I-AL type of methods for computing an ϵ-KKT solution. Our complexity analysis of the I-AL methods is mainly based on an analysis on inexact proximal point algorithm (PPA) and the link between the I-AL methods and inexact PPA. It is substantially different from the 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. Compared to the mostly related I-AL methods [11], our modified I-AL method is more practically efficient and also applicable to a broader class of problems.

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

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

*Department of Mathematics, Simon Fraser University, Canada. (email: zhaosong@sfu.ca). This author was supported in part by NSERC Discovery Grant.

†Department of Mathematics, Simon Fraser University, Canada. (email: ziruiz@sfu.ca). This author was supported by NSERC Discovery Grant and the SFU Alan Mekler postdoctoral fellowship.

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

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

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 evaluated.*[1] *The domain of P, denoted by* dom(P)*, is compact.*

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

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

(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

$$\begin{array}{ll} \min\limits_{x \in X} & f(x) \\ \text{s.t.} & g_i(x) \le 0, \ i = 1, \ldots, m_1, \\ & h_j(x) = 0, \ j = 1, \ldots, m_2. \end{array}$$

In this paper we are interested in augmented Lagrangian (AL) methods for solving problem (1). AL methods have been widely regarded as effective methods for solving constrained nonlinear programming (e.g., see [3, 25, 18]). The classical AL method was initially proposed by Hestenes [7] and Powell [20], and has been extensively studied in the literature (e.g., see [21, 2]). Recently, AL methods have been applied to solve some special instances of problem (1) arising in various applications such as compressed sensing [30], image processing [5], and optimal control [8]. They have also been used to solve conic programming (e.g., see [4, 9, 31]).

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

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

[1] The proximal operator associated with P is defined as $\text{prox}_P(x) = \arg\min_y \left\{\frac{1}{2}\|y - x\|^2 + P(y)\right\}$.

where $\mathcal{L}(x, \lambda; \rho)$ is the AL function of (1) defined as (e.g., see [24, Section 11.K] and [26])

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

and $\mathrm{dist}(x, S) := \min\{\|x - z\| : z \in S\}$ for any nonempty closed set $S \subseteq \Re^m$. 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 for updating $\{\lambda^k\}$ 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, first-order methods are often used to approximately solve the AL subproblem (4). For example, Aybat and Iyengar [1] proposed a first-order inexact augmented Lagrangian (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 [16, 27]) to find an approximate solution x^{k+1} of the AL subproblem (4) such that

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

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

$$|F(x) - F^*| \le \epsilon, \quad \mathrm{dist}(g(x), -\mathcal{K}) \le \epsilon \tag{5}$$

for some $\epsilon > 0$ in at most $\mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})$ first-order inner iterations. In addition, Necoara et al. [15] 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 [15] that this method with a suitable choice of $\{\rho_k\}$ and $\{\eta_k\}$ can find an approximate solution x of (1) satisfying (5) in at most $\mathcal{O}(\epsilon^{-1})$ first-order inner iterations. More recently, Xu [29] proposed an I-AL method for solving a special case of (1) with $\mathcal{K}$ being the nonnegative orthant, which can also find an approximate solution x of (1) satisfying (5) in at most $\mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})$ first-order inner iterations. Some other related works on I-AL type of methods can be found, for example, in [12, 19, 28].

Since F^* is typically unknown, (5) generally cannot be used as a termination criterion for AL methods. A common practical termination criterion for AL methods is as follows:

$$\mathrm{dist}(0, \nabla f(x) + \partial P(x) + \nabla g(x)\lambda) \le \epsilon, \quad \mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \epsilon, \quad (x, \lambda) \in \mathrm{dom}(P) \times \mathcal{K}^*. \tag{6}$$

Such x is often referred to as an ϵ-approximate Karush-Kuhn-Tucker (KKT) solution of problem (1). Though the first-order iteration complexity with respect to (5) is established for the I-AL methods [1, 15], it is not clear what first-order iteration-complexity they have in terms of the practical termination criterion (6). In addition, for the I-AL methods [1, 15], $\{\rho_k\}$ and $\{\eta_k\}$ are specifically chosen to achieve low first-order iteration-complexity with respect to (5). Such a choice, however, may not lead to a low first-order iteration-complexity in terms of (6). In fact, there is no theoretical guarantee on the performance of these methods with respect to the practical termination criterion (6).

Lan and Monterio [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 an indicator function of a simple compact convex set X, that is,

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

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 ordinary I-AL steps similar to those in [1] but with $\rho_k \equiv \mathcal{O}(D_\Lambda^{3/4}\epsilon^{-3/4})$ and $\eta_k \equiv \mathcal{O}(D_\Lambda^{1/4}\epsilon^{7/4})$ until a certain approximate $(\tilde{x}, \tilde{\lambda})$ is found, 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 to mainly execute a single I-AL step with $\rho = \rho_k$ and $\eta = \mathcal{O}(\min(D_\Lambda^{3/4}\epsilon^{5/4}, D_\Lambda^{-3/4}\epsilon^{11/4}))$, starting with $(\tilde{x}, \tilde{\lambda})$. They showed that this method can find an ϵ-KKT solution of (7) in at most $\mathcal{O}(\epsilon^{-7/4})$ first-order inner iterations totally. Notice that this I-AL method uses the fixed ρ_k and η_k through all outer iterations and they may be respectively overly large and small, which 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. In addition, the choice of ρ_k and η_k in this method requires some knowledge of D_Λ, which is not known a priori. A "guess-and-check" procedure is thus proposed in [11] to remedy it, which consists of guessing a sequence of estimates $\{t_l\}$ for D_Λ and applying the above I-AL method with D_Λ replaced by t_l until an ϵ-KKT solution of (7) is found. These likely make this method practically inefficient, which is indeed observed in our numerical experiment.

In addition, Lan and Monterio [11] proposed a modified I-AL method by applying the above first-order I-AL method with D_Λ replaced by D_Λ^ϵ to the perturbed problem

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

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

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

first-order inner iterations totally. Since the dependence of D_Λ^ϵ on ϵ is generally unknown, it is not clear how complexity (9) depends on ϵ and also whether or not it improves the first-order iteration-complexity $\mathcal{O}(\epsilon^{-7/4})$ of the above I-AL method [11].

Motivated by the above points, we propose in this paper a practical first-order I-AL method for computing an ϵ-KKT solution of problem (1) and study its iteration-complexity. Our I-AL method, analogous to the one [11], consists of two stages, particularly, primary stage and postprocessing stage. The primary stage is to execute the ordinary I-AL steps with $\{\rho_k\}$ and $\{\eta_k\}$ changing dynamically until either an ϵ-KKT solution of (7) is obtained or a certain approximate $(\tilde{x}, \tilde{\lambda})$ is found. The postprocessing stage is to mainly execute a single I-AL step with ρ being the latest ρ_k obtained from the primary stage and $\eta = \mathcal{O}(\epsilon^2 \min(\rho, 1/\rho))$, starting with $(\tilde{x}, \tilde{\lambda})$. Our I-AL method distinguishes the one in [11] mainly in two aspects: (i) the parameters $\{\rho_k\}$ and $\{\eta_k\}$ of our method dynamically change with the iterations, but those of the latter one are static for all iterations; and (ii) our method does not use any information of D_Λ, but the latter one needs to apply a "guess-and-check" procedure to approximate D_Λ. We show that our I-AL method terminates in a finite number of iterations when $\{\rho_k\}$ and $\{\eta_k\}$ are suitably chosen. Moreover, this method attains its optimal worst-case iteration-complexity $\mathcal{O}(\epsilon^{-7/4})$ for $\rho_k = \mathcal{O}(k^{3/2})$ and $\eta_k = \mathcal{O}(k^{-5/2}\sqrt{\epsilon})$. Though our method shares the same order of worst-case iteration-complexity as the one in [11], it is deemed to be more practically efficient as it uses the dynamic $\{\rho_k\}$ and $\{\eta_k\}$ and also does not need a "guess-and-check" procedure, which is indeed corroborated in our numerical experiment.

Besides, we propose a modified I-AL method with improved worst-case iteration-complexity than our above I-AL method for computing an ϵ-KKT solution of problem (1). It modifies the latter method by adding a regularization term $\|x - x^k\|^2/(2\rho_k)$ to the AL function $\mathcal{L}(x, \lambda^k; \rho_k)$ at each kth outer iteration and also solving the AL subproblems to a higher accuracy. Moreover, it uses a weaker termination criterion and does not need a postprocessing stage. Since this regularization term changes dynamically, it is substantially different from those in [15, 11]. We show that this modified I-AL method terminates in a finite number of iterations when $\{\rho_k\}$ and $\{\eta_k\}$ are suitably chosen. Moreover, this method attains its optimal worst-case iteration-complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$ for $\rho_k = \mathcal{O}(\alpha^k)$ and $\eta_k = \mathcal{O}(\beta^k)$ for any $\alpha > 1$ and $0 < \beta < 1/\alpha$. To the best of our knowledge, this method so far has the lowest iteration-complexity bound among all first-order I-AL type of methods for computing an ϵ-KKT solution of problem (1).

Our iteration-complexity analysis of the first-order I-AL methods is mainly based on an analysis on inexact proximal point algorithm (PPA) and a result that these methods are equivalent to an inexact PPA applied to a monotone inclusion problem. The iteration-complexity of the I-AL methods [1, 15, 11] is, however, obtained by regarding the I-AL methods as an inexact dual gradient method. Therefore, our analyis is substantially different from those in [1, 15, 11]. In addition, as the operator associated with the monotone inlcusion problem linked to the I-AL methods is closely related to the KKT conditions, our approach appears to be more appropriate than the one in [11].

The rest of this paper is organized as follows. In Section 2, we introduce the concept of an ϵ-KKT solution of (1), and study inexact proximal point algorithm for solving monotone inclusion problems and also some optimal first-order methods for solving a class of structured convex optimization. In Section 3, we propose a first-order I-AL method and study its iteration-complexity. In Section 4, we propose a modified first-order I-AL method and derive its iteration-complexity. In Section 5, we present some numercial results for the proposed algorithms. Finally, we make some concluding remarks in Section 6.

1.1 Notations

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 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,

$$\text{prox}_h(z) = \arg\min_{x\in\Re^n} \left\{ \frac{1}{2}\|x-z\|^2 + h(x) \right\}, \quad \forall z \in \Re^n. \tag{10}$$

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

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

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 (1) is defined as

$$l(x,\lambda) = \begin{cases} f(x) + P(x) + \langle \lambda, g(x)\rangle & \text{if } x \in \text{dom}(P) \text{ and } \lambda \in \mathcal{K}^*, \\ -\infty & \text{if } x \in \text{dom}(P) \text{ and } \lambda \notin \mathcal{K}^*, \\ +\infty & \text{if } x \notin \text{dom}(P), \end{cases} \tag{11}$$

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

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

which is a closed concave function. The Lagrangian dual problem (3) can thus be rewritten as

$$d^* = \max_\lambda \ d(\lambda). \tag{12}$$

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 [23]). It can be verified that

$$\partial l(x,\lambda) = \begin{cases} \begin{pmatrix} \nabla f(x) + \partial P(x) + \nabla g(x)\lambda \\ g(x) - \mathcal{N}_{\mathcal{K}^*}(\lambda) \end{pmatrix}, & \text{if } x \in \text{dom}(P) \text{ and } \lambda \in \mathcal{K}^*, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{13}$$

It is well known that λ^* is an optimal solution of the Lagrangian dual problem (12) if and only if $0 \in \partial d(\lambda^*)$, and (x^*, λ^*) is a saddle point[2] of l if and only if $(0,0) \in \partial l(x^*, \lambda^*)$.

Finally, we define two set-valued operators associated with problems (1) and (3) as follows:

$$\mathcal{T}_d : \lambda \to \{u \in \Re^m : -u \in \partial d(\lambda)\}, \quad \forall \lambda \in \Re^m, \tag{14}$$

$$\mathcal{T}_l : (x,\lambda) \to \{(v,u) \in \Re^n \times \Re^m : (v, -u) \in \partial l(x,\lambda)\}, \quad \forall (x,\lambda) \in \Re^n \times \Re^m. \tag{15}$$

2 Technical preliminaries

In this section we introduce ϵ-KKT solutions for problem (1). Also, we study an inexact proximal point algorithm for solving the monotone inclusion problem. Finally, we discuss some variants of Nesterov's optimal first-order method for solving a class of structured convex optimization.

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

2.1 ϵ-KKT solutions

The following result provides a characterization of an optimal solution of (1).

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*

$$(0,0) \in \partial l(x^*, \lambda^*), \tag{16}$$

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

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

Proof. The result (16) follows from [23, Theorem 36.6]. By (13), it is not hard to see that (16) 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 (16). Instead, we are 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* (13) *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* (6) *that will be used in this paper.*

2.2 Inexact proximal point algorithm

In this subsection, we review the inexact proximal point algorithm (PPA) for solving the monotone inclusion problem and study some of its properties.

A set-valued operator $\mathcal{T} : \Re^n \rightrightarrows \Re^n$ is called *monotone* if

$$\langle z - z', w - w' \rangle \ge 0 \quad \text{whenever} \quad w \in \mathcal{T}(z), w' \in \mathcal{T}(z').$$

Further, $\mathcal{T}$ is called *maximally monotone* if its graph is not properly contained in the graph of any other monotone operators. For example, the two operators $\mathcal{T}_d$ and $\mathcal{T}_l$ defined in (14) are maximally monotone (e.g., see [23, Corollaries 31.5.2 and 37.5.2]).

In what follows, we assume that the operator $\mathcal{T} : \Re^n \rightrightarrows \Re^n$ is maximally monotone and moreover $\{z : 0 \in \mathcal{T}(z)\} \ne \emptyset$. Let

$$\mathcal{J}_\rho = (\mathcal{I} + \rho\mathcal{T})^{-1}, \quad \forall \rho > 0.$$

Then $\mathcal{J}_\rho$ is a single-valued mapping from $\Re^n$ into $\Re^n$ (see [14]). Moreover, $\mathcal{J}_\rho$ is non-expansive, that is,

$$\|\mathcal{J}_\rho(z) - \mathcal{J}_\rho(z')\| \le \|z - z'\|, \quad \forall z, z' \in \Re^n,$$

and $\mathcal{J}_\rho(z) = z$ if and only if $0 \in \mathcal{T}(z)$. Furthermore, for any z^* such that $0 \in \mathcal{T}(z^*)$, one has (e.g., see [22, Proposition 1])

$$\|\mathcal{J}_\rho(z) - z^*\|^2 + \|\mathcal{J}_\rho(z) - z\|^2 \le \|z - z^*\|^2, \tag{17}$$

which implies that

$$\|\mathcal{J}_\rho(z) - z^*\| \le \|z - z^*\|, \quad \|\mathcal{J}_\rho(z) - z\| \le \|z - z^*\|, \quad \forall z \in \Re^n. \tag{18}$$

Analagous to the classical fixed-point method, the following inexact PPA was proposed for solving the monotone inclusion problem $0 \in \mathcal{T}(z)$ (e.g., see [22]).

Algorithm 1 (Inexact proximal point algorithm)

0. Input $z^0 \in \Re^n$, $\{e_k\} \subset \Re_+$ and $\{\rho_k\} \subset \Re_{++}$. Set $k = 0$.
1. Find z^{k+1} by approximately evaluating $\mathcal{J}_{\rho_k}(z^k)$ such that

$$\|z^{k+1} - \mathcal{J}_{\rho_k}(z^k)\| \le e_k. \tag{19}$$

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

End.

The following convergence result is established in [22, Theorem 1].

Theorem 1 *Let $\{z^k\}$ be generated by Algorithm 1. Suppose that $\inf_k \rho_k > 0$ and $\sum_{k=0}^\infty e_k < \infty$. Then $\{z^k\}$ converges to a point z^∞ satisfying $0 \in \mathcal{T}(z^\infty)$.*

We next study some properties of Algorithm 1, which will be used to analyze the first-order I-AL methods in later sections.

Theorem 2 *Let z^* be a vector such that $0 \in \mathcal{T}(z^*)$, and $\{z^k\}$ be the sequence generated by Algorithm 1. Then it holds that*

$$\|z^s - z^*\| \le \|z^t - z^*\| + \sum_{i=t}^{s-1} e_i, \quad \forall s \ge t \ge 0. \tag{20}$$

Moreover, for any $K \ge 0$, we have

$$\sum_{k=K}^{2K} \|z^{k+1} - z^k\|^2 \le 2\left(\|z^0 - z^*\| + 2\sum_{k=0}^{2K} e_k\right)^2. \tag{21}$$

Proof. Let $\xi^k = z^{k+1} - \mathcal{J}_{\rho_k}(z^k)$ for all $k \ge 0$. By this, (18) and (19) with $\rho = \rho_k$ and $z = z^k$, one has

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

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

$$\|z^s - z^*\| \le \|z^t - z^*\| + \sum_{i=t}^{s-1} \|\xi^i\|, \quad \forall s \ge t \ge 0. \tag{22}$$

Notice from (19) that $\|\xi^k\| \le e_k$ for all $k \ge 0$, which along with (22) leads to (20). In addition, by the definition of ξ^k, and (17) 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 (22), we obtain that

$$\begin{aligned}
\sum_{k=K}^{2K} \|\mathcal{J}_{\rho_k}(z^k) - z^k\|^2 &\le \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\sum_{k=K}^{2K} \|\xi^k\| \left(\|z^K - z^*\| + \sum_{j=K}^{k} \|\xi^j\| \right) \\
&= \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + 2\sum_{k=K}^{2K}\sum_{j=K}^{k} \|\xi^k\|\|\xi^j\| \\
&= \|z^K - z^*\|^2 - \sum_{k=K}^{2K} \|\xi^k\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + \sum_{k=K}^{2K} \|\xi^k\|^2 + \left(\sum_{k=K}^{2K} \|\xi^k\|\right)^2 \\
&= \|z^K - z^*\|^2 + 2\|z^K - z^*\| \cdot \sum_{k=K}^{2K} \|\xi^k\| + \left(\sum_{k=K}^{2K} \|\xi^k\|\right)^2 \\
&= \left(\|z^K - z^*\| + \sum_{k=K}^{2K} \|\xi^k\| \right)^2 \le \left(\|z^0 - z^*\| + \sum_{k=0}^{2K} \|\xi^k\| \right)^2,
\end{aligned} \tag{23}$$

where (23) follows from (22) 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 (23) 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 (21). The proof is then completed. $\square$

Corollary 1 *Let z^* be a vector such that $0 \in \mathcal{T}(z^*)$, and $\{z^k\}$ be the sequence generated by Algorithm 1. Then, it follows that*

$$\|z^{k+1} - z^k\| \leq \|z^0 - z^*\| + \sum_{i=0}^{k} e_i. \tag{24}$$

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

$$\min_{K \leq k \leq 2K} \|z^{k+1} - z^k\| \leq \frac{\sqrt{2}\left(\|z^0 - z^*\| + 2\sum_{k=0}^{2K} e_k\right)}{\sqrt{K+1}}. \tag{25}$$

Proof. By (18) with $\rho = \rho_k$ and $z = z^k$, one has $\|\mathcal{J}_{\rho_k}(z^k) - z^k\| \leq \|z^k - z^*\|$. This together with (19) and (20) yields that

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

In addition, (25) follows directly from (21). □

2.3 Optimal first-order methods for structured convex optimization

In this subsection we consider a class of structured convex optimization in the form of

$$\phi_h^* = \min_{x \in \Re^n} \left\{ \phi_h(x) := \phi(x) + h(x) \right\}, \tag{26}$$

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, we assume that $\mathrm{dom}(h)$ is compact. Therefore, the optimal value ϕ_h^* of (26) is finite.

We first study a property of problem (26), which will be used subsequently.

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

$$\mathrm{dist}\big(0, \partial\phi_h(x^+)\big) \leq \sqrt{8L_{\nabla\phi}\big(\phi_h(x) - \phi_h^*\big)}, \tag{27}$$

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

Proof. Since $\nabla\phi$ is Lipschitz continuous on $\mathrm{dom}(h)$ with Lipschitz constant $L_{\nabla\phi}$, we have that (e.g., see [17, Lemma 1.2.3])

$$\phi(y) \leq \phi(x) + \langle \nabla\phi(x), y - x \rangle + \frac{L_{\nabla\phi}}{2}\|y - x\|^2, \quad \forall x, y \in \mathrm{dom}(h). \tag{28}$$

Let $x \in \mathrm{dom}(h)$ be arbitrarily chosen. By the definition of x^+ and (10), we have that $x^+ \in \mathrm{dom}(h)$ and

$$x^+ = \arg\min_{z \in \Re^n} \left\{ \phi(x) + \langle \nabla\phi(x), z - x \rangle + \frac{L_{\nabla\phi}}{2}\|z - x\|^2 + h(z) \right\}. \tag{29}$$

Notice that the objective function in (29) is strongly convex with modulus $L_{\nabla\phi}$. Hence, we have

$$\begin{aligned} &\phi(x) + \langle \nabla\phi(x), x^+ - x \rangle + \frac{L_{\nabla\phi}}{2}\|x^+ - x\|^2 + h(x^+) \\ &\quad \leq \phi(x) + \langle \nabla\phi(x), z - x \rangle + \frac{L_{\nabla\phi}}{2}\|z - x\|^2 + h(z) - \frac{L_{\nabla\phi}}{2}\|z - x^+\|^2, \quad \forall z \in \mathrm{dom}(h). \end{aligned} \tag{30}$$

This together with (28) yields that

$$\phi_h(x^+) \le \phi(x) + \langle \nabla\phi(x), x^+ - x\rangle + \frac{L_{\nabla\phi}}{2}\|x^+ - x\|^2 + h(x^+) \le \phi_h(x) - \frac{L_{\nabla\phi}}{2}\|x^+ - x\|^2, \tag{31}$$

where the first inequality is due to (28) with $y = x^+$, and the second one is by (30) with $z = x$. It then follows that $\phi_h(x^+) \le \phi_h(x)$. Moreover, the optimality condition of (29) yields that $0 \in \nabla\phi(x) + \partial h(x^+) + L_{\nabla\phi}(x^+ - x)$. This gives

$$\nabla\phi(x^+) - \nabla\phi(x) - L_{\nabla\phi}(x^+ - x) \in \nabla\phi(x^+) + \partial h(x^+) = \partial\phi_h(x^+).$$

Hence, we have

$$\mathrm{dist}\big(0, \partial\phi_h(x^+)\big) \le \|\nabla\phi(x^+) - \nabla\phi(x) - L_{\nabla\phi}(x^+ - x)\| \le 2L_{\nabla\phi}\|x^+ - x\|, \tag{32}$$

where the second inequality is due to the Lipschitz continuity of $\nabla\phi$. Combining (31) and (32) gives

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

which is the desired inequality (27). The proof is then completed. □

In the rest of this subsection we study some optimal first-order methods for solving problem (26). We start by considering the case of problem (26) in which ϕ is convex but not necessarily strongly convex. In particular, we review a method presented in [27, Section 3] for solving (26) with a general convex ϕ, which is a variant of Nesterov's optimal first-order methods [16, 17].

Algorithm 2 (An optimal first-order method for (26) **with general convex** ϕ**)**

0. Input $x^0 = z^0 \in \mathrm{dom}(h)$. Set $k = 0$.

1. Set $y^k = \frac{k}{k+2}x^k + \frac{2}{k+2}z^k$.

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

$$\ell(x; y) := \phi(y) + \langle \nabla\phi(y), x - y\rangle + h(x). \tag{33}$$

3. Set $x^{k+1} = \frac{k}{k+2}x^k + \frac{2}{k+2}z^{k+1}$.

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

End.

The main convergence result of Algorithm 2 is summarized below, whose proof can be found in [27, Corollary 1].

Proposition 3 *Let* $\{x^k\}$ *and* $\{y^k\}$ *be the sequences generated by Algorithm 2. Then we have*

$$\phi_h(x^{k+1}) - \phi_h^* \leq \phi_h(x^{k+1}) - \phi_h^k \leq \frac{2L_{\nabla\phi}D_h^2}{(k+1)(k+3)}, \quad \forall k \geq 0, \tag{34}$$

where D_h *and* ϕ_h^k *are defined as*

$$D_h = \max_{x,y\in\mathrm{dom}(h)} \|x-y\|, \qquad \phi_h^k = \frac{4}{(k+1)(k+3)} \min_x \left\{ \sum_{i=0}^{k} \frac{i+2}{2}\ell(x;y^i) \right\}, \quad \forall k \geq 0. \tag{35}$$

Remark 2 *Since* h *is proper and* $\mathrm{dom}(h)$ *is compact, it is not hard to see that* D_h *and* ϕ_h^k *are finite for all* $k \geq 0$*. From Proposition 3, one can see that Algorithm 2 finds an* ϵ*-optimal solution* x^{k+1} *satisfying* $\phi_h(x^{k+1}) - \phi_h^* \leq \epsilon$ *once* $2L_{\nabla\phi}D_h^2/((k+1)(k+3)) \leq \epsilon$ *or* $\phi_h(x^{k+1}) - \phi_h^k \leq \epsilon$ *holds. Therefore, these two inequalities can be used as a termination criterion for Algorithm 2. The latter one is, however, a better termination criterion due to* (34).

The following result is an immediate consequence of Proposition 3, which provides an iteration-complexity of Algorithm 2 for finding an ϵ-optimal solution of problem (26).

Corollary 2 *For any given* $\epsilon > 0$*, Algorithm 2 finds an approximate solution* x^k *of problem* (26) *such that* $\phi_h(x^k) - \phi_h^* \leq \epsilon$ *in no more than* $K(\epsilon)$ *iterations, where*

$$K(\epsilon) = \left\lceil D_h \sqrt{\frac{2L_{\nabla\phi}}{\epsilon}} \right\rceil.$$

We next consider the case of problem (26) in which ϕ is strongly convex, that is, there exists a constant $\mu \in (0, L_{\nabla\phi})$ such that

$$\langle \nabla\phi(x) - \nabla\phi(y), x-y \rangle \geq \mu\|x-y\|^2, \quad \forall x,y \in \mathrm{dom}(h). \tag{36}$$

We now propose a slight variant of Nesterov's optimal method [17, 13] for solving problem (26) with a strongly convex ϕ.

Algorithm 3 (An optimal first-order method for (26) with strongly convex ϕ)

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

$$x^0 = \mathrm{prox}_{h/L_{\nabla\phi}} \left(x^{-1} - \frac{1}{L_{\nabla\phi}} \nabla\phi(x^{-1}) \right). \tag{37}$$

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

1. Set $y^k = \dfrac{x^k + \alpha z^k}{1+\alpha}$.

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 (33).

3. Set $x^{k+1} = (1-\alpha)x^k + \alpha z^k$.

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

End.

Remark 3 *Algorithm 3 differs from Nesterov's optimal method [17, 13] in that it executes a proximal step* (37) *to generate* x^0 *while the latter method simply sets* $x^0 = x^{-1}$.

The main convergence result of Algorithm 3 is presented as follows.

Proposition 4 *Suppose that* (36) *holds. Let* $\{x^k\}$ *be generated by Algorithm 3. Then we have*

$$\phi_h(x^k) - \phi_h^* \leq \frac{L_{\nabla\phi} D_h^2}{2}\left(1 - \sqrt{\frac{\mu}{L_{\nabla\phi}}}\right)^k, \quad \forall k \geq 0, \tag{38}$$

where D_h *is defined in* (35).

Proof. Observe that $\{x^k\}_{k\geq 0}$ is identical to the sequence generated by the Nesterov's optimal method [17, 13] starting with x^0. Hence, it follows from [13, Theorem 1] that

$$\phi_h(x^k) - \phi_h^* \leq \left(1 - \sqrt{\frac{\mu}{L_{\nabla\phi}}}\right)^k \left(\phi_h(x^0) - \phi_h^* + \frac{\mu}{2}\|x^0 - x^*\|^2\right), \quad \forall k \geq 0. \tag{39}$$

Notice that x^0 is computed by (37). It follows from (10) and (37) that

$$x^0 = \arg\min_x \left\{\langle \nabla\phi(x^{-1}), x - x^{-1}\rangle + \frac{L_{\nabla\phi}}{2}\|x - x^{-1}\|^2 + h(x)\right\}. \tag{40}$$

Let x^* is be the optimal solution of (26). By (40) and the Lipschitz continuity of $\nabla\phi(x)$, one has that

$$\phi(x^0) + h(x^0) \leq \phi(x^{-1}) + \langle \nabla\phi(x^{-1}), x^0 - x^{-1}\rangle + \frac{L_{\nabla\phi}}{2}\|x^0 - x^{-1}\|^2 + h(x^0) \tag{41}$$

$$\leq \phi(x^{-1}) + \langle \nabla\phi(x^{-1}), x^* - x^{-1}\rangle + \frac{L_{\nabla\phi}}{2}\|x^* - x^{-1}\|^2 + h(x^*) - \frac{L_{\nabla\phi}}{2}\|x^0 - x^*\|^2 \tag{42}$$

$$\leq \phi(x^*) + h(x^*) - \frac{L_{\nabla\phi}}{2}\|x^0 - x^*\|^2 + \frac{L_{\nabla\phi} D_h^2}{2} \tag{43}$$

where (41) follows from (28), (42) is due to (40) and the fact that the objective function in (40) is strongly convex with modulus $L_{\nabla\phi}$, and (43) follows from the convexity of ϕ and $\|x^* - x^{-1}\| \leq D_h$. Using (43), $\mu < L_{\nabla\phi}$ and $\phi_h = \phi + h$, we obtain that

$$\phi_h(x^0) - \phi_h^* + \frac{\mu}{2}\|x^0 - x^*\|^2 \leq \phi_h(x^0) - \phi_h^* + \frac{L_{\nabla\phi}}{2}\|x^0 - x^*\|^2 \leq \frac{L_{\nabla\phi} D_h^2}{2}.$$

This together with (39) leads to (38) as desired. □

The following result is a consequence of Propositions 2 and 4, regarding the iteration-complexity of Algorithm 3 for finding a certain approximate solution of problem (26) with strongly convex ϕ.

Corollary 3 *Suppose that* (36) *holds. Let* $\{x^k\}$ *be the sequence generated by Algorithm 3 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 an approximate solution* $\tilde{x}^k$ *of problem* (26) *satisfying* $\mathrm{dist}(0, \partial\phi_h(\tilde{x}^k)) \leq \epsilon$ *is generated by running Algorithm 3 for at most* $\widetilde{K}(\epsilon)$ *iterations, where*

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

Proof. It follows from Proposition 2 and $\tilde{x}^k = \text{prox}_{h/L_{\nabla\phi}}\big(x^k - \nabla\phi(x^k)/L_{\nabla\phi}\big)$ that

$$\text{dist}(0, \partial\phi_h(\tilde{x}^k)) \leq \sqrt{8L_{\nabla\phi}(\phi_h(x^k) - \phi_h^*)}. \tag{44}$$

By (38), it is not hard to verify that

$$\phi_h(x^k) - \phi_h^* \leq \frac{\epsilon^2}{8L_{\nabla\phi}}, \quad \forall k \geq \widetilde{K}(\epsilon),$$

which together with (44) implies that $\text{dist}(0, \partial\phi_h(\tilde{x}^k)) \leq \epsilon$ for all $k \geq \widetilde{K}(\epsilon)$. Hence, the conclusion of this corollary holds. □

2.4 Augmented Lagrangian function and its properties

In this subsection we introduce the augmented Lagrangian function for problem (1) and study some of its properties.

The augmented Lagrangian function for problem (1) is defined as (e.g., see [26])

$$\mathcal{L}(x, \lambda; \rho) := f(x) + P(x) + \frac{1}{2\rho}\Big[\text{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big) - \|\lambda\|^2\Big], \tag{45}$$

where $\rho > 0$ is a penalty parameter. The augmented Lagrangian dual function of (1) is given by

$$d(\lambda; \rho) := \min_{x\in\Re^n} \mathcal{L}(x, \lambda; \rho). \tag{46}$$

For convenience, we let

$$\mathcal{S}(x, \lambda; \rho) := f(x) + \frac{1}{2\rho}\text{dist}^2\big(\lambda + \rho g(x), -\mathcal{K}\big). \tag{47}$$

It is clear to see that

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

Recall that g is continuously differentiable on an open set containing $\text{dom}(P)$. By this and the compactness of $\text{dom}(P)$, we know that

$$M_g := \max_{x\in\text{dom}(P)} \|g(x)\| \tag{48}$$

is finite. Moreover, there exists some $L_g > 0$ such that g is Lipschitz continuous on $\text{dom}(P)$ with Lipschitz constant L_g and also $\|\nabla g(x)\| \leq L_g$ for any $x \in \text{dom}(P)$. We next study some properties of the functions $\mathcal{S}(x, \lambda; \rho)$ and $\mathcal{L}(x, \lambda; \rho)$.

Proposition 5 *For any $(\lambda, \rho) \in \Re^m \times \Re_{++}$, the following statements hold.*

(i) $\mathcal{S}(x, \lambda; \rho)$ is a convex function in x.

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

$$\nabla_x\mathcal{S}(x, \lambda; \rho) = \nabla f(x) + \nabla g(x)\Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big). \tag{49}$$

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

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

Proof. (i) Let $x, x' \in \Re^n$ and $\alpha \in [0, 1]$ be arbitrarily chosen. By (2), one has

$$\lambda + \rho g\big(\alpha x + (1-\alpha)x'\big) = \lambda + \rho[\alpha g(x) + (1-\alpha)g(x')] + \underbrace{\rho\left(g\big(\alpha x + (1-\alpha)x'\big) - [\alpha g(x) + (1-\alpha)g(x')]\right)}_{\in -\mathcal{K}}.$$

It follows that $\lambda + \rho\big[\alpha g(x) + (1-\alpha)g(x')\big] \preceq_{-\mathcal{K}} \lambda + \rho g\big(\alpha x + (1-\alpha)x'\big)$. Using this and Lemma 3 in Appendix A with $\mathcal{K}$ replaced by $-\mathcal{K}$, we have

$$\mathrm{dist}^2\Big(\lambda + \rho g\big(\alpha x + (1-\alpha)x'\big), -\mathcal{K}\Big) \le \mathrm{dist}^2\Big(\lambda + \rho[\alpha g(x) + (1-\alpha)g(x')], -\mathcal{K}\Big). \tag{50}$$

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

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

which along with (50) leads to

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

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.

(ii) 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 Lemma 4 in Appendix A.

(iii) Recall that ∇f, ∇g and g are Lipschitz continuous on $\mathrm{dom}(P)$. By this and (49), we have that for any $x, x' \in \mathrm{dom}(P)$,

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

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

The following proposition is an extension of the results in [21] to problem (1). For the sake of completeness, we include a proof for it.

Proposition 6 *For any $(x, \lambda, \rho) \in \Re^n \times \Re^m \times \Re_{++}$, the following identity holds*

$$\mathcal{L}(x, \lambda; \rho) = \max_{\eta \in \Re^m} \left\{ l(x, \eta) - \frac{1}{2\rho}\|\eta - \lambda\|^2 \right\}. \tag{51}$$

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

$$d(\lambda; \rho) = \max_{\eta \in \Re^m} \left\{ d(\eta) - \frac{1}{2\rho}\|\eta - \lambda\|^2 \right\}. \tag{52}$$

(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

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

Proof. We first show that (51) holds. Indeed, if $x \notin \mathrm{dom}(P)$, (51) trivially holds since both sides equal ∞. Now suppose that $x \in \mathrm{dom}(P)$. By the definition of l in (11), we have that for any $\eta \in \mathcal{K}^*$,

$$\begin{aligned} l(x, \eta) - \frac{1}{2\rho}\|\eta - \lambda\|^2 &= f(x) + P(x) + \langle \eta, g(x) \rangle - \frac{1}{2\rho}\|\eta - \lambda\|^2 \\ &= f(x) + P(x) - \frac{1}{2\rho}\|\lambda\|^2 + \frac{1}{2\rho}\|\lambda + \rho g(x)\|^2 - \frac{1}{2\rho}\|\eta - (\lambda + \rho g(x))\|^2. \end{aligned} \tag{53}$$

Also, for any $\eta \notin \mathcal{K}^*$, $l(x, \eta) - \frac{1}{2\rho}\|\eta - \lambda\|^2 = -\infty$. Hence, the maximum in (51) is attained at

$$\bar{\lambda} = \arg\min_{\eta \in \mathcal{K}^*} \|\eta - (\lambda + \rho g(x))\|^2,$$

which is unique and equals $\Pi_{\mathcal{K}^*}(\lambda + \rho g(x))$. Substituting this into (53), we obtain that

$$\begin{aligned} \max_{\eta \in \Re^m} \left\{ l(x, \eta) - \frac{1}{2\rho}\|\eta - \lambda\|^2 \right\} &= f(x) + P(x) - \frac{1}{2\rho}\|\lambda\|^2 + \frac{1}{2\rho}\|\lambda + \rho g(x)\|^2 - \frac{1}{2\rho}\|\bar{\lambda} - (\lambda + \rho g(x))\|^2 \\ &= f(x) + P(x) - \frac{1}{2\rho}\|\lambda\|^2 + \frac{1}{2\rho}\left(\|\lambda + \rho g(x)\|^2 - \mathrm{dist}^2(\lambda + \rho g(x), \mathcal{K}^*)\right) \\ &= f(x) + P(x) - \frac{1}{2\rho}\|\lambda\|^2 + \frac{1}{2\rho}\mathrm{dist}^2(\lambda + \rho g(x), -\mathcal{K}) = \mathcal{L}(x, \lambda; \rho), \end{aligned}$$

where the third equality is due to Lemma 4. Therefore, (51) holds as desired.

By (46) and (51), one has

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

where the third equality is due to the fact that the function inside the brace is strongly concave in λ. Therefore, statement (i) holds. Finally, statements (ii) and (iii) follow from (51) and Danskin's theorem. □

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

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

From Remark 1 (a), we know that (x, λ) is an ϵ-KKT solution of (1) if and only if it satisfy $x \in \mathrm{dom}(P)$, $\lambda \in \mathcal{K}^*$, $\mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \epsilon$, 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 generate a pair (x, λ) to satisfy these conditions. Given that the proximal operator associated with P and the projection onto $\mathcal{K}$ can be evaluated (see Assumption 1), the first two conditions can be easily satisfied by the iterates of our proposed 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, the primary stage and the postprocessing stage. In the primary stage, the AL subproblems are solved roughly, and a pair (x^k, λ^k) is found in the end that satisfies nearly the third condition but roughly the last condition. In the postprocessing stage, the latest AL subproblem arising in the primary stage is re-optimized to a higher accuracy to obtain some point $\tilde{x}$, starting at x^k. A proximal step is then applied to $\mathcal{L}(\cdot, \lambda^k, \rho_k)$ at $\tilde{x}$ and $l(\tilde{x}, \cdot)$ 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 4 (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 2 to the problem $\min_x \mathcal{L}(x, \lambda^k; \rho_k)$ to find $x^{k+1} \in \mathrm{dom}(P)$ satisfying
$$\mathcal{L}(x^{k+1}, \lambda^k; \rho_k) - \min_x \mathcal{L}(x, \lambda^k; \rho_k) \le \eta_k. \tag{54}$$

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 (6), set $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$ and output (x^+, λ^+).

4. If the following inequalities are satisfied
$$\frac{1}{\rho_k}\|\lambda^{k+1} - \lambda^k\| \le \frac{3}{4}\epsilon, \qquad \frac{\eta_k}{\rho_k} \le \frac{\epsilon^2}{128}, \tag{55}$$
call the subroutine $(x^+, \lambda^+) = \mathrm{Postprocessing}(\lambda^k, \rho_k, x^{k+1}, \epsilon)$ and output (x^+, λ^+).

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

End.

The subroutine Postprocessing in Step 4 of Algorithm 4 is presented as follows.

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
$$\tilde{L} = L_{\nabla f} + L_{\nabla g}\big(\|\tilde{\lambda}\| + \tilde{\rho} M_g\big) + \tilde{\rho} L_g^2, \qquad \tilde{\eta} = \epsilon^2 \cdot \min\left\{\frac{\tilde{\rho}}{128}, \frac{1}{8\tilde{L}}\right\}. \tag{56}$$

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

$$\mathcal{L}(\hat{x}, \tilde{\lambda}; \tilde{\rho}) - \min_x \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho}) \leq \tilde{\eta}. \tag{57}$$

3. Output the pair (x^+, λ^+), which is computed by

$$x^+ = \mathrm{prox}_{P/\tilde{L}}\big(\hat{x} - \nabla_x \mathcal{S}(\hat{x}, \tilde{\lambda}; \tilde{\rho})/\tilde{L}\big), \quad \lambda^+ = \Pi_{\mathcal{K}^*}\big(\tilde{\lambda} + \tilde{\rho} g(x^+)\big), \tag{58}$$

where $\mathcal{S}$ is defined in (47).

End.

For ease of later reference, we refer to the first-order iterations of Algorithm 2 for solving the AL subproblems as the *inner iterations* of Algorithm 4, and call the update from (x^k, λ^k) to (x^{k+1}, λ^{k+1}) an *outer iteration* of Algorithm 4. We now make some remarks on Algorithm 4 as follows.

Remark 4 *(a) By Proposition 5, $\mathcal{L}(\cdot, \lambda; \rho)$ is in the form of (26) with $\phi = \mathcal{S}(\cdot, \lambda; \rho)$ and $h = P$. Therefore, Algorithm 2 can be suitably applied to solve AL subproblems (54) and (57).*

(b) The subroutine Postprocessing is inspired by [11], where 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.

(c) The I-AL method [11] uses the fixed $\rho_k \equiv \mathcal{O}(D_\Lambda^{3/4}\epsilon^{-3/4})$ and $\eta_k \equiv \mathcal{O}(D_\Lambda^{1/4}\epsilon^{7/4})$ through all outer iterations in the primary stage, 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. In addition, the choice of ρ_k and η_k in the I-AL method [11] requires some knowledge of D_Λ, which is not known a priori. A "guess-and-check" procedure is thus proposed in [11] to remedy it, which consists of guessing a sequence of estimates $\{t_l\}$ for D_Λ and applying their I-AL method with D_Λ replaced by t_l until an ϵ-KKT solution of (7) is found. These likely make this method practically inefficient, which is indeed observed in our numerical experiment. By 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.

We next study global convergence of Algorithm 4, and also its first-order iteration-complexity for a special choice of $\{\rho_k\}$ and $\{\eta_k\}$. To proceed, we establish a crucial result as follows, which shows that each outer iteration of Algorithm 4 can be viewed as a step of an inexact PPA applied to solve the monotone inclusion problem $0 \in \mathcal{T}_d(\lambda)$, where $\mathcal{T}_d$ is defined in (14). It generalizes the result of [21, Proposition 6] that is for a special case of problem (1) with $\mathcal{K} = \{0\}^{m_1} \times \Re^{m_2}_+$.

Proposition 7 *Let $\{\lambda^k\}$ be the sequence generated by Algorithm 4 for solving problem (1). Then for any $k \geq 0$, one has*

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

where $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k\mathcal{T}_d)^{-1}$ and $\mathcal{T}_d$ is defined in (14).

Proof. It follows from the definition of $\text{dist}(\cdot, -\mathcal{K})$ and Lemma 4 (a) in Appendix that for any $\rho > 0$, $\lambda \in \Re^m$ and $x \in \text{dom}(P)$,

$$\text{dist}(\lambda + \rho g(x), -\mathcal{K}) = \min_u \{\|\lambda - u\| : \rho g(x) + u \preceq_{\mathcal{K}} 0\}, \tag{59}$$

and the minimum is attained uniquely at $\bar{u} = \lambda - \Pi_{\mathcal{K}^*}\big(\lambda + \rho g(x)\big)$. These together with (45) yield

$$\mathcal{L}(x^{k+1}, \lambda^k; \rho_k) = f(x^{k+1}) + P(x^{k+1}) + \frac{1}{2\rho_k}\left[\|\lambda^k - u^k\|^2 - \|\lambda^k\|^2\right], \tag{60}$$

where $u^k = \lambda^k - \Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^{k+1}))$. By this and Step 2 of Algorithm 4, we have $u^k = \lambda^k - \lambda^{k+1}$. Moreover, it follows from (46) and (59) that

$$\begin{aligned} d(\lambda^k; \rho_k) &= \min_x \left\{ f(x) + P(x) + \frac{1}{2\rho_k}\left[\text{dist}^2(\lambda^k + \rho_k g(x), -\mathcal{K}) - \|\lambda^k\|^2\right] \right\} \\ &= \min_{x,u} \left\{ f(x) + P(x) + \frac{1}{2\rho_k}\left[\|\lambda^k - u\|^2 - \|\lambda^k\|^2\right] : \rho_k g(x) + u \preceq_{\mathcal{K}} 0 \right\} \\ &= \min_u \left\{ v(u) + \frac{1}{2\rho_k}\left[\|\lambda^k - u\|^2 - \|\lambda^k\|^2\right] \right\}, \end{aligned} \tag{61}$$

where

$$v(u) = \min_x \{ f(x) + P(x) : \rho_k g(x) + u \preceq_{\mathcal{K}} 0 \}. \tag{62}$$

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 (61) 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 (61) and Danskin's theorem that $\nabla_\lambda d(\lambda^k; \rho_k) = -\bar{u}^k/\rho_k$. In addition, it follows from (52) 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, (60), (61) and (62), we obtain that

$$\begin{aligned} \mathcal{L}(x^{k+1}, \lambda^k; \rho_k) - d(\lambda^k; \rho_k) &= f(x^{k+1}) + P(x^{k+1}) + \frac{1}{2\rho_k}\|\lambda^k - u^k\|^2 - \min_u \left\{ v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2 \right\} \\ &\geq v(u^k) + \frac{1}{2\rho_k}\|\lambda^k - u^k\|^2 - \min_u \left\{ v(u) + \frac{1}{2\rho_k}\|\lambda^k - u\|^2 \right\} \qquad (63) \\ &\geq \frac{1}{2\rho_k}\|u^k - \bar{u}^k\|^2 = \frac{1}{2\rho_k}\|\mathcal{J}_{\rho_k}(\lambda^k) - \lambda^{k+1}\|^2, \qquad (64) \end{aligned}$$

where (63) follows from (62) 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 (64) 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 (54) and (64). □

We are now ready to establish the global convergence of Algorithm 4.

Theorem 3 *(i) If Algorithm 4 successfully terminates (i.e., at Step 3 or 4), then the output (x^+, λ^+) is an ϵ-KKT solution of problem* (1).

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

$$\rho_k > 0 \text{ is nondecreasing}, \quad 0 < \frac{\eta_k}{\rho_k} \to 0, \quad \frac{\sum_{i=0}^{2k} \sqrt{\rho_i \eta_i}}{\rho_k \sqrt{k+1}} \to 0.^3 \tag{65}$$

Then Algorithm 4 terminates in a finite number of iterations. Moreover, its output (x^+, λ^+) is an ϵ-KKT solution of problem (1).

Proof. (i) One can easily see that (x^+, λ^+) is an ϵ-KKT solution of (1) if Algorithm 4 terminates at Step 3. We now show that it is also an ϵ-KKT solution of (1) if Algorithm 4 terminates at Step 4. To this end, suppose that Algorithm 4 terminates at Step 4 for some iteration k, that is, the inequalities (55) hold at 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 (57) and (58) 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 $\varphi_p(x) = \mathcal{L}(x, \tilde{\lambda}; \tilde{\rho})$. Notice from Proposition 5 that $\nabla_x \mathcal{S}(x, \tilde{\lambda}; \tilde{\rho})$ is Lipschitz continuous on $\text{dom}(P)$ with Lipschitz constant $\tilde{L}$. Hence, φ_p is in the form of (26) with $\phi = \mathcal{S}(\cdot, \tilde{\lambda}; \tilde{\rho})$ and $h = P$. By (56), (57), (58) and Proposition 2, one has $\varphi_p(x^+) \le \varphi_p(\hat{x})$ and

$$\text{dist}\big(0, \partial \varphi_p(x^+)\big) \le \sqrt{8\tilde{L}\big(\varphi_p(\hat{x}) - \min_{x \in \Re^n} \varphi_p(x)\big)} \le \sqrt{8\tilde{L}\tilde{\eta}} \le \epsilon. \tag{66}$$

In addition, it follows from (58) and Proposition 6 that

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

This together with (66) yields $\text{dist}\big(0, \partial_x l(x^+, \lambda^+)\big) \le \epsilon$ as desired.

It remains to show that $\text{dist}\big(0, \partial_\lambda l(x^+, \lambda^+)\big) \le \epsilon$. By (55) and Proposition 7, 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 (55), 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 $\varphi_p = \mathcal{L}(\cdot, \tilde{\lambda}; \tilde{\rho})$, the second relation in (58), and the same arguments as those for (64), one has

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

This together with $\varphi_p(x^+) \le \varphi_p(\hat{x})$, (56) and (57) yields that

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

[3] For example, $\rho_k = \hat{C}(k+1)^{3/2}$ and $\eta_k = C_2(k+1)^{-5/2}$ satisfy (65) for any $\hat{C}$, $C_2 > 0$.

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

$$\|\lambda^+ - \tilde{\lambda}\| \leq \|\tilde{\lambda} - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| + \|\lambda^+ - \mathcal{J}_{\tilde{\rho}}(\tilde{\lambda})\| \leq \frac{7\tilde{\rho}\epsilon}{8} + \frac{\tilde{\rho}\epsilon}{8} = \tilde{\rho}\epsilon. \tag{67}$$

Moreover, by Proposition 6 and the second relation in (58), one has

$$\lambda^+ = \Pi_{\mathcal{K}^*}\big(\tilde{\lambda} + \tilde{\rho} g(x^+)\big) = \arg\max_{\lambda \in \Re^m} \left\{ l(x^+, \lambda) - \frac{1}{2\tilde{\rho}}\|\lambda - \tilde{\lambda}\|^2 \right\}.$$

Its first-order optimality condition yields that $(\lambda^+ - \tilde{\lambda})/\tilde{\rho} \in \partial_\lambda l(x^+, \lambda^+)$. This together with (67) implies $\mathrm{dist}\big(0, \partial_\lambda l(x^+, \lambda^+)\big) \leq \epsilon$.

(ii) Suppose for contradiction that Algorithm 4 does not terminate. Let $\{\lambda^k\}$ be generated by Algorithm 4. By Proposition 7, one can observe that $\{\lambda^k\}$ can be viewed as the one generated by Algorithm 1 applied to the problem $0 \in \mathcal{T}(\lambda)$ with $\mathcal{T} = \mathcal{T}_d$ and $e_k = \sqrt{2\rho_k\eta_k}$. It then follows from Corollary 1 that

$$\min_{k \leq i \leq 2k} \|\lambda^{i+1} - \lambda^i\| \leq \frac{\sqrt{2}\left(\|\lambda^0 - \lambda^*\| + 2\sum_{i=0}^{2k}\sqrt{2\rho_i\eta_i}\right)}{\sqrt{k+1}} \tag{68}$$

for any λ^* satisfying $0 \in \mathcal{T}_d(\lambda^*)$, which, together with the assumption that $\{\rho_k\}$ is nondecreasing, implies that

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

By this and (65), one has that $\min_{k \leq i \leq 2k} \|\lambda^{i+1} - \lambda^i\|/\rho_i \to 0$ and $\eta_k/\rho_k \to 0$ as $k \to \infty$. It follows that the inequalities (55) must hold at some iteration k. This implies that Algorithm 4 terminates at iteration k, which leads to a contradiction. Hence, Algorithm 4 terminates in a finite number of iterations. It then follows from statement (i) that the output (x^+, λ^+) is an ϵ-KKT solution of (1). □

In the remainder of this section, we study the first-order iteration-complexity of Algorithm 4. In particular, we derive an upper bound on the total number of its inner iterations, i.e., all iterations of Algorithm 2 applied to solve the AL subproblems of Algorithm 4. To proceed, we introduce some further notation that will be used subsequently. 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

$$D_X := \max_{x,y \in \mathrm{dom}(P)} \|x - y\|, \; D_\Lambda := \|\lambda^0 - \hat{\lambda}^*\|, \; B := L_{\nabla f} + L_{\nabla g}\|\hat{\lambda}^*\| + L_{\nabla g} D_\Lambda, \tag{69}$$

$$C := L_{\nabla g} M_g + L_g^2, \; \bar{D}_\Lambda := \max\{D_\Lambda, 1\}, \; \bar{B} := \max\{B, 1\}, \; \bar{C} := \max\{C, 1\}, \tag{70}$$

where M_g is defined in (48), and $L_{\nabla f}$, $L_{\nabla g}$ and L_g are the Lipschitz constants of ∇f, ∇g and g on $\mathrm{dom}(P)$, respectively.

We next establish two technical lemmas that will be used subsequently.

Lemma 1 *If N is a nonnegative integer such that*

$$\frac{D_\Lambda + 2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}}{\rho_N\sqrt{N+1}} \leq \frac{\epsilon}{2}, \quad \frac{\eta_N}{\rho_N} \leq \frac{\epsilon^2}{128}, \tag{71}$$

then the number of outer iterations of Algorithm 4 is at most $2N+1$.

Proof. It follows from (68) that

$$\min_{N\le k\le 2N}\|\lambda^{k+1}-\lambda^k\| \le \frac{\sqrt{2}\left(D_\Lambda + 2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}\right)}{\sqrt{N+1}}.$$

By this, (71) and the assumption that $\{\rho_k\}$ is nondecreasing, there exists some $N \le \tilde{k} \le 2N$ such that

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

In addition, since $\{\rho_k\}$ is nondecreasing and $\{\eta_k\}$ is decreasing, we obtain from (71) that

$$\frac{\eta_{\tilde{k}}}{\rho_{\tilde{k}}} \le \frac{\eta_N}{\rho_N} \le \frac{\epsilon^2}{128}.$$

Hence, the inequalities (55) hold for $k = \tilde{k}$. Since $\tilde{k} \le 2N$, Algorithm 4 terminates within at most $2N+1$ outer iterations. □

Lemma 2 *For any $k \ge 0$, the Lipschitz constant of $\nabla_x \mathcal{S}(x,\lambda^k;\rho_k)$, denoted as L_k, satisfies*

$$L_k \le C\rho_k + B + L_{\nabla g}\sum_{i=0}^{k-1}\sqrt{2\rho_i\eta_i}, \tag{72}$$

where B and C are given in (69) *and* (70)*.*

Proof. By Proposition 5 (iii), one has $L_k \le L_{\nabla f} + L_{\nabla g}\big(\|\lambda^k\| + \rho_k M_g\big) + \rho_k L_g^2$. In addition, recall that $\{\lambda^k\}$ can be viewed as the one generated by Algorithm 1 applied to the problem $0 \in \mathcal{T}(\lambda)$ with $\mathcal{T} = \mathcal{T}_d$ and $e_k = \sqrt{2\rho_k\eta_k}$. It thus follows from (69) and Theorem 2 that

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

where $\hat{\lambda}^*$ is defined above. By these and the definitions of B and C, we obtain the desired bound (72). □

We are now ready to establish the first-order iteration-complexity of Algorithm 4.

Theorem 4 *Let $\epsilon > 0$ be given, and $\bar{C}$, D_X, and $\bar{D}_\Lambda$ be defined in* (69) *and* (70)*. Suppose that $\{\rho_k\}$ and $\{\eta_k\}$ are chosen as*

$$\rho_k = \rho_0(k+1)^{\frac{3}{2}}, \quad \eta_k = \eta_0(k+1)^{-\frac{5}{2}}\cdot\min\{1,\sqrt{\epsilon}\} \tag{73}$$

for some $\rho_0 \ge 1$ and $0 < \eta_0 \le 1$. Then, the total number of inner iterations of Algorithm 4 for finding an ϵ-KKT solution of problem (1) *is at most $\mathcal{O}\big(\mathcal{T}\big(\min\{1,\epsilon\}\big)\big)$, where*

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

Proof. For convenience, let $\epsilon_0 = \min\{1, \epsilon\}$. Let $\bar{N}$ be the number of outer iterations of Algorithm 4. Also, let $\mathcal{I}_k$ and $\mathcal{I}_p$ be the number of iterations executed by Algorithm 2 at the outer iteration k of Algorithm 4 and in the subroutine Postprocessing, respectively. In addition, let T be the total number of inner iterations of Algorithm 4. 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 (73), 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

$$\sum_{k=0}^{K} \sqrt{2\rho_k\eta_k} = \sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}} \sum_{k=0}^{K} (k+1)^{-\frac{1}{2}} \le 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}\sqrt{K+1}, \tag{74}$$

where the inequality follows by $\sum_{k=0}^{K}(k+1)^{-1/2} \le 2\sqrt{K+1}$. Let $\gamma = 7\bar{D}_\Lambda^{1/2}\epsilon_0^{-1/2}$ and $N = \lceil\gamma\rceil$. It follows from (73), (74), and $\gamma \le N \le \gamma + 1$ that

$$\frac{D_\Lambda + 2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}}{\rho_N\sqrt{N+1}} \le \frac{\bar{D}_\Lambda + 4\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}}\sqrt{2N+1}}{\rho_0(N+1)^2} \le \frac{\bar{D}_\Lambda}{\rho_0(N+1)^2} + \frac{8\eta_0^{\frac{1}{2}}\epsilon_0^{\frac{1}{4}}}{\rho_0^{\frac{1}{2}}(N+1)^{\frac{3}{2}}}. \tag{75}$$

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 (75), one has

$$\frac{D_\Lambda + 2\sum_{k=0}^{2N}\sqrt{2\rho_k\eta_k}}{\rho_N\sqrt{N+1}} < \frac{\epsilon}{2}. \tag{76}$$

In addition, using $N+1 \ge \gamma \ge 7\epsilon_0^{-1/2}$, (73), $\epsilon_0 \le 1$, $\rho_0 \ge 1$ and $\eta_0 \le 1$, we obtain that

$$\frac{\eta_N}{\rho_N} = \frac{\eta_0\epsilon_0^{\frac{1}{2}}}{\rho_0(N+1)^4} \le \frac{1}{7^4\epsilon_0^{-2}} = \frac{\epsilon_0^2}{7^4} < \frac{\epsilon^2}{128}. \tag{77}$$

By (76), (77) and Lemma 1, we obtain

$$\bar{N} \le 2N + 1 = 2\left\lceil 7\bar{D}_\Lambda^{\frac{1}{2}}\epsilon_0^{-\frac{1}{2}}\right\rceil + 1. \tag{78}$$

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 (72) and (73) that for any $k \ge 0$,

$$L_k \le \bar{C}\rho_0(k+1)^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}} L_{\nabla g}(k+1)^{\frac{1}{2}}. \tag{79}$$

This together with Corollary 2, (54) and (73) yields that

$$
\begin{aligned}
\mathcal{I}_k &\leq \left\lceil D_X \sqrt{\frac{2L_k}{\eta_k}} \right\rceil \leq 1 + \sqrt{2} D_X \sqrt{\frac{\bar{C}\rho_0 (k+1)^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}} L_{\nabla g}(k+1)^{\frac{1}{2}}}{\eta_0 (k+1)^{-\frac{5}{2}} \epsilon_0^{\frac{1}{2}}}} \\
&\leq 1 + \sqrt{2} D_X \sqrt{\frac{\bar{C}\rho_0 (k+1)^{\frac{3}{2}} + \bar{B}\rho_0 + 2\sqrt{2}\rho_0 \epsilon_0^{\frac{1}{4}} L_{\nabla g}(k+1)^{\frac{1}{2}}}{\eta_0 (k+1)^{-\frac{5}{2}} \epsilon_0^{\frac{1}{2}}}} \\
&\leq 1 + D_X \sqrt{\frac{2\rho_0}{\eta_0}} \left(\bar{C}^{\frac{1}{2}} \epsilon_0^{-\frac{1}{4}} (k+1)^2 + \bar{B}^{\frac{1}{2}} \epsilon_0^{-\frac{1}{4}} (k+1)^{\frac{5}{4}} + 2L_{\nabla g}^{\frac{1}{2}} \epsilon_0^{-\frac{1}{8}} (k+1)^{\frac{3}{2}} \right),
\end{aligned} \tag{80}
$$

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, (55) is satisfied when $k = \bar{N} - 1$. It then follows that $(\tilde{\lambda}, \tilde{\rho}) = (\lambda^{\bar{N}-1}, \rho_{\bar{N}-1})$ and $\tilde{L} = L_{\bar{N}-1}$. By these, Corollary 2, (56), (57) and $\epsilon_0 \leq \epsilon$, we have

$$
\mathcal{I}_p \leq \left\lceil D_X \sqrt{\frac{2L_{\bar{N}-1}}{\tilde{\eta}}} \right\rceil \leq \left\lceil \frac{16 D_X}{\epsilon_0} \cdot \max \left\{ \sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}}, \frac{L_{\bar{N}-1}}{4} \right\} \right\rceil \tag{81}
$$

In addition, it follows from (79) that

$$
L_{\bar{N}-1} \leq \bar{C}\rho_0 \bar{N}^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}} L_{\nabla g} \bar{N}^{\frac{1}{2}}. \tag{82}
$$

By this and (73), we obtain that for any $\bar{N} \geq 1$,

$$
\begin{aligned}
\sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}} &\leq \sqrt{\frac{\bar{C}\rho_0 \bar{N}^{\frac{3}{2}} + \bar{B} + 2\sqrt{2\rho_0\eta_0}\ \epsilon_0^{\frac{1}{4}} L_{\nabla g} \bar{N}^{\frac{1}{2}}}{\rho_0 \bar{N}^{\frac{3}{2}}}} \leq \sqrt{\bar{C} + \bar{B} + 2\sqrt{2}\epsilon_0^{\frac{1}{4}} L_{\nabla g}} \\
&\leq \bar{C}^{\frac{1}{2}} + \bar{B}^{\frac{1}{2}} + 2\epsilon_0^{\frac{1}{8}} L_{\nabla g}^{\frac{1}{2}},
\end{aligned} \tag{83}
$$

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 (82), (83), $\epsilon_0 \leq 1$, $\eta_0 \leq 1$, $\bar{C} \geq 1$ and $\bar{B} \geq 1$, it is not hard to verify that for all $\bar{N} \geq 1$,

$$
\max \left\{ \sqrt{\frac{L_{\bar{N}-1}}{\rho_{\bar{N}-1}}}, \frac{L_{\bar{N}-1}}{4} \right\} \leq \bar{C}\rho_0 \bar{N}^{\frac{3}{2}} + \bar{B} + 2\rho_0^{\frac{1}{2}} \epsilon_0^{\frac{1}{8}} \bar{N}^{\frac{1}{2}} (L_{\nabla g} + L_{\nabla g}^{\frac{1}{2}}). \tag{84}
$$

Substituting (84) into (81), we arrive at

$$
\mathcal{I}_p \leq 1 + \frac{16 D_X}{\epsilon_0} \left(\bar{C}\rho_0 \bar{N}^{\frac{3}{2}} + \bar{B} + 2\rho_0^{\frac{1}{2}} \epsilon_0^{\frac{1}{8}} \bar{N}^{\frac{1}{2}} (L_{\nabla g} + L_{\nabla g}^{\frac{1}{2}}). \right). \tag{85}
$$

Finally, we use (78), (80) and (85) to derive an upper bound on the overall complexity T. By (78),

$N=\lceil\gamma\rceil$ and $\gamma\geq 7$, one has $\bar{N}-1\leq 2N\leq 2\gamma+2\leq 3\gamma-1$. This together with (80) yields that

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

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

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

In addition, by $\bar{N}\leq 3\gamma$, $\gamma=7\bar{D}_{\Lambda}^{1/2}\epsilon_0^{-1/2}$ and (85), we obtain that

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

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

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

This together with $\epsilon_0=\min\{1,\epsilon\}$ yields the complexity bound in Theorem 4. □

Remark 5 *(i) It can be shown that $\rho_k=\mathcal{O}((k+1)^{3/2})$ and $\eta_k=\mathcal{O}((k+1)^{-5/2})$ minimizes the worst-case upper bound of the total number of inner iterations of Algorithm 4. The derivation is, however, rather tedious and thus omitted.*

(ii) Algorithm 4 shares the same order of worst-case iteration-complexity in terms of ϵ as the I-AL method [11]. It is, however, much more efficient than the latter method as observed in our numerical experiment. The main reason is perhaps that Algorithm 4 uses the dynamic $\{\rho_k\}$ and $\{\eta_k\}$, but 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}.

4 A modified I-AL method with improved iteration-complexity

In this section, we propose a modified first-order I-AL method and show that it has a better iteration-complexity than Algorithm 4 for computing an ϵ-KKT solution of (1). In particular, it modifies the latter method by adding a regularization term $\|x - x^k\|^2/(2\rho_k)$ to the AL function $\mathcal{L}(x, \lambda^k; \rho_k)$ at each kth outer iteration and also solving the AL subproblems to a higher accuracy. Moreover, it uses a weaker termination criterion and does not need a postprocessing stage. Since this regularization term changes dynamically, it is substantially different from those in [15, 11, 29].

Our modified first-order I-AL method is presented as follows.

Algorithm 5 (The modified 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 3 to the problem $\min_x \varphi_k(x)$ to find $x^{k+1} \in \mathrm{dom}(P)$ satisfying
$$\mathrm{dist}(0, \partial\varphi_k(x^{k+1})) \le \eta_k, \tag{86}$$
where
$$\varphi_k(x) = \mathcal{L}(x, \lambda^k; \rho_k) + \frac{1}{2\rho_k}\|x - x^k\|^2. \tag{87}$$
2. Set $\lambda^{k+1} = \Pi_{\mathcal{K}^*}\big(\lambda^k + \rho_k g(x^{k+1})\big)$.
3. If (x^{k+1}, λ^{k+1}) satisfies (6) or the following two inequalities are satisfied
$$\frac{1}{\rho_k}\|(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k)\| \le \frac{\epsilon}{2}, \quad \eta_k \le \frac{\epsilon}{2}, \tag{88}$$
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 3 for solving the AL subproblems as the *inner iterations* of Algorithm 5, and call the update from (x^k, λ^k) to (x^{k+1}, λ^{k+1}) an *outer iteration* of Algorithm 5. Notice from (87) that φ_k is strongly convex with modulus $1/\rho_k$. Therefore, the AL subproblem $\min_x \varphi_k(x)$ arising in Algorithm 5 is in the form of (26) and it can be suitably solved by Algorithm 3

We next study the global convergence of Algorithm 5, and also its first-order iteration-complexity for a special choice of $\{\rho_k\}$ and $\{\eta_k\}$. To proceed, we establish a crucial result as follows, which shows that each outer iteration of Algorithm 4 can be viewed as a step of an inexact PPA applied to solve the monotone inclusion problem $0 \in \mathcal{T}_l(x, \lambda)$, where $\mathcal{T}_l$ is defined in (15). It generalizes the result of [21, Proposition 8] that is for a special case of problem (1) with $\mathcal{K} = \{0\}^{m_1} \times \Re^{m_2}_+$.

Proposition 8 *Let $\{(x^k, \lambda^k)\}$ be generated by Algorithm 5. For any $k \ge 0$, one has*
$$\|(x^{k+1}, \lambda^{k+1}) - \mathcal{J}_{\rho_k}(x^k, \lambda^k)\| \le \rho_k\eta_k, \tag{89}$$
where $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k\mathcal{T}_l)^{-1}$ and $\mathcal{T}_l$ is defined in (15).

Proof. By Proposition 6 and $\lambda^{k+1} = \Pi_{\mathcal{K}^*}\big(\lambda^k + \rho_k g(x^{k+1})\big)$, one has

$$\partial_x \mathcal{L}(x^{k+1}, \lambda^k; \rho_k) = \partial_x l(x^{k+1}, \lambda^{k+1}), \quad \frac{1}{\rho_k}(\lambda^{k+1} - \lambda^k) \in \partial_\lambda l(x^{k+1}, \lambda^{k+1}). \tag{90}$$

By (86), there exists $\|v\| \le \eta_k$ 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 (90) implies that

$$x^k + \rho_k v \in \rho_k \partial_x l(x^{k+1}, \lambda^{k+1}) + x^{k+1}, \quad \lambda^k \in -\rho_k \partial_\lambda l(x^{k+1}, \lambda^{k+1}) + \lambda^{k+1}, \tag{91}$$

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

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

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

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

which yields (89) as desired. □

We are now ready to establish the global convergence of Algorithm 5.

Theorem 5 *(i) If Algorithm 5 successfully terminates (i.e., at Step 3), then the output* (x^+, λ^+) *is an* ϵ*-KKT solution of problem* (1).

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

$$\rho_k > 0 \text{ is nondecreasing}, \quad 0 < \eta_k \to 0, \quad \frac{\sum_{i=0}^{2k} \rho_i \eta_i}{\rho_k \sqrt{k+1}} \to 0.^4 \tag{92}$$

Then Algorithm 5 terminates in a finite number of iterations. Moreover, its output (x^+, λ^+) *is an* ϵ*-KKT solution of problem* (1).

Proof. (i) Suppose that Algorithm 5 terminates at Step 3 for some iteration k. It then follows that (x^+, λ^+) is already an ϵ-KKT solution of problem (1) or the inequalities (88) 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). Notice that (91) holds for some $\|v\| \le \eta_k$. By (91), one has

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

By this, (88) and $(x^+, \lambda^+) = (x^{k+1}, \lambda^{k+1})$, we obtain

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

[4] For example, $\rho_k = \rho_0 \alpha^k$ and $\eta_k = \eta_0 \beta^k$ satisfy (92) for any $\rho_0 > 0$, $\eta_0 > 0$, $\alpha > 1$ and $0 < \beta < 1/\alpha$.

In view of Definition 1, (x^+, λ^+) is an ϵ-KKT solution of problem (1).

(ii) Suppose for contradiction that Algorithm 5 does not terminate. Let $\{(x^k, \lambda^k)\}$ be generated by Algorithm 5. By Proposition 8, one can observe that $\{(x^k, \lambda^k)\}$ can be viewed as the one generated by Algorithm 1 applied to the problem $0 \in \mathcal{T}(x, \lambda)$ with $\mathcal{T} = \mathcal{T}_l$ and $e_k = \rho_k \eta_k$. It then follows from Corollary 1 that

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

for any (x^*, λ^*) satisfying $0 \in \mathcal{T}_l(x^*, \lambda^*)$, which, together with the assumption that $\{\rho_k\}$ is nondecreasing, implies that

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

By this and (92), 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 $k \to \infty$. It follows that the inequalities (88) must hold at some iteration k. This implies that Algorithm 5 terminates at iteration k, which leads to a contradiction. Hence, Algorithm 5 terminates in a finite number of iterations. It then follows from statement (i) that the output (x^+, λ^+) is an ϵ-KKT solution of (1). $\square$

In the rest of this section, we study the first-order iteration-complexity of Algorithm 5. In particular, we derive an upper bound on the total number of its inner iterations, i.e., all iterations of Algorithm 3 applied to solve the AL subproblems of Algorithm 5. Before proceeding, we introduce some further notation that will be used subsequently. Let X^* be the set of optimal solutions of problem (7) and $\hat{x}^* \in X^*$ such that $\|x^0 - \hat{x}^*\| = \mathrm{dist}(x^0, X^*)$. In addition, we define

$$\bar{D}_X := \max\{D_X, 1\}, \ D := \mathrm{dist}(x^0, X^*) + D_\Lambda, \ \bar{D} := \max\{D, 1\}, \ \hat{B} := L_{\nabla f} + L_{\nabla g}\|\hat{\lambda}^*\| + L_{\nabla g} D, \quad (93)$$

where D_X, D_Λ and $\hat{\lambda}^*$ are defined in (69), and $L_{\nabla f}$ and $L_{\nabla g}$ are the Lipschitz constants of ∇f and ∇g on $\mathrm{dom}(P)$, respectively.

We next establish two technical lemmas that will be used subsequently.

Proposition 9 *If N is a nonnegative integer such that*

$$\frac{D + \sum_{k=0}^{N} \rho_k \eta_k}{\rho_N} \le \frac{\epsilon}{2}, \quad \eta_N \le \frac{\epsilon}{2}, \quad (94)$$

then the number of outer iterations of Algorithm 5 is at most $N+1$.

Proof. Recall that $\hat{x}^*$ and $\hat{\lambda}^*$ are the optimal solutions of problems (1) and (3), respectively. It then follows that $0 \in \mathcal{T}_l(\hat{x}^*, \hat{\lambda}^*)$. Also, recall that $\{(x^k, \lambda^k)\}$ can be viewed as the one generated by Algorithm 1 applied to the problem $0 \in \mathcal{T}(x, \lambda)$ with $\mathcal{T} = \mathcal{T}_l$ and $e_k = \rho_k \eta_k$. These together with (24), (89), and (93) yield that for any $k \ge 0$,

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

where the last relation is due to (93). By this and (94), one can see that (88) is satisfied when $k = N$. Hence, Algorithm 5 terminates within $N+1$ outer iterations. $\square$

Proposition 10 *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*

$$L_k = C\rho_k + \hat{B} + L_{\nabla g} \sum_{i=0}^{k-1} \rho_i \eta_i + \rho_k^{-1}, \tag{95}$$

where C *and* $\hat{B}$ *are defined in* (70) *and* (93)*, respectively.*

Proof. By the definition of $s_k(x)$ and Proposition 5 (iii), one has

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

Recall from the proof of Proposition 9 that $0 \in \mathcal{T}_l(\hat{x}^*, \hat{\lambda}^*)$. Since $\{(x^k, \lambda^k)\}$ can be viewed as the one generated by Algorithm 1 applied to the problem $0 \in \mathcal{T}(x, \lambda)$ with $\mathcal{T} = \mathcal{T}_l$ and $e_k = \rho_k \eta_k$, it follows from Theorem 2 that

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

where the last relation is due to (93). Substituting this into the above inequality, and using the definitions of $\hat{B}$ and C, we obtain that $\|\nabla s_k(x) - \nabla s_k(y)\| \leq L_k \|x - y\|$ for all $x, y \in \mathrm{dom}(P)$. Hence, the conclusion holds. □

We are now ready to establish the first-order iteration-complexity of Algorithm 5.

Theorem 6 *Let* $\epsilon > 0$ *be given, and* $\bar{D}_X$ *and* $\bar{D}$ *be defined in* (93)*. Suppose that* $\{\rho_k\}$ *and* $\{\eta_k\}$ *are chosen as*

$$\rho_k = \rho_0 \alpha^k, \quad \eta_k = \eta_0 \beta^k \tag{96}$$

for some $\rho_0 \geq 1$*,* $0 < \eta_0 \leq 1$*,* $\alpha > 1$*,* $0 < \beta < 1$ *such that* $\gamma = \alpha\beta < 1$*. Then, the total number of inner iterations of Algorithm 5 for finding an* ϵ*-KKT solution of problem* (1) *is at most*

$$\mathcal{T}(\epsilon) = \left\lceil \frac{8\alpha^2 \sqrt{\hat{C}\rho_0}}{\alpha - 1} \log \frac{2\alpha \hat{C} \bar{D}_X}{\eta_0 \beta} \right\rceil \max\left\{1, \left\lceil \frac{2(\bar{D} + \rho_0 \eta_0)}{(1-\gamma)\epsilon} \log_\alpha \frac{2\alpha(\bar{D} + \rho_0 \eta_0)}{(1-\gamma)\epsilon} \right\rceil \right\}, \tag{97}$$

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* (70) *and* (93)*, respectively.*

Proof. Let $\bar{N}$ be the number of outer iterations of Algorithm 5, and let $\mathcal{I}_k$ be the number of first-order iterations executed by Algorithm 3 at the outer iteration k of Algorithm 5. In addition, let T be the total number of first-order inner iterations of Algorithm 5. 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 (96) and $0 < \gamma < 1$, we have that

$$\sum_{k=0}^{K} \rho_k \eta_k = \rho_0 \eta_0 \sum_{k=0}^{K} \gamma^k \leq \rho_0 \eta_0 \sum_{k=0}^{\infty} \gamma^k = \frac{\rho_0 \eta_0}{1-\gamma}, \quad \forall K \geq 0. \tag{98}$$

Let

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

Since $N \geq \log_\alpha \frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon}$, we have from (96) that

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

By this, (98), $D \leq \bar{D}$, and $\rho_0 \geq 1$, we obtain

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

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

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

which together with (96), $\beta < 1$ and $\eta_0 \leq 1$ implies that $\eta_N \leq \epsilon/2$. It then follows from these and Proposition 9 that

$$\bar{N} \leq N + 1 \leq \max\left\{1, \left\lceil \log_\alpha \frac{2(\bar{D} + \rho_0\eta_0)}{(1-\gamma)\epsilon} \right\rceil \right\} + 1. \tag{99}$$

We next derive an upper bound on $\mathcal{I}_k$. By (95), (96), $\alpha > 1$ and $\rho_0 \geq 1$, one has that for any $k \geq 0$,

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

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, (86), $\rho_k = \rho_0\alpha^k$, $\hat{C} \geq 1$, $\bar{D}_X \geq 1$, $\alpha > 1$, $\beta < 1$, $\rho_0 \geq 1$, $\eta_0 \leq 1$, and Corollary 3, we obtain that for any $k \geq 0$,

$$\begin{aligned}
\mathcal{I}_k &\leq \left\lceil \sqrt{\frac{L_k}{\mu_k}} \right\rceil \max\left\{1, \left\lceil 2\log\frac{2L_kD_X}{\eta_k} \right\rceil \right\} \leq \left\lceil \sqrt{\hat{C}\rho_0}\ \alpha^k \right\rceil \max\left\{1, \left\lceil 2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k} \right\rceil \right\} \\
&\leq \left\lceil \sqrt{\hat{C}\rho_0}\ \alpha^k \right\rceil \left\lceil 2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k} \right\rceil \leq \left(\sqrt{\hat{C}\rho_0}\ \alpha^k + 1\right)\left(2\log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k} + 1\right) \\
&\leq 8\sqrt{\hat{C}\rho_0}\ \alpha^k \log\frac{2\alpha^k\hat{C}\bar{D}_X}{\eta_0\beta^k} \leq 8\sqrt{\hat{C}\rho_0}\ k\alpha^k \log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta},
\end{aligned} \tag{100}$$

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

Finally, we derive an upper bound on T. By (100), one has

$$T = \sum_{k=0}^{\bar{N}-1} \mathcal{I}_k \leq 8\sqrt{\hat{C}\rho_0}\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}\sum_{k=0}^{\bar{N}-1} k\alpha^k \leq \frac{8\sqrt{\hat{C}\rho_0}}{\alpha-1}\log\frac{2\alpha\hat{C}\bar{D}_X}{\eta_0\beta}(\bar{N}-1)\alpha^{\bar{N}}, \tag{101}$$

where the last inequality is due to $\sum_{k=0}^{K} k\alpha^k \leq K\alpha^{K+1}/(\alpha-1)$ for any $K \geq 0$. 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 (99) implies that $\bar{N} \leq \log_\alpha \frac{2(\bar{D}+\rho_0\eta_0)}{(1-\gamma)\epsilon} + 2$. By this and (101), 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 (99) implies that $\bar{N} \leq 2$. By this and (101), 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 (97) as desired. □

Remark 6 *One can see from Theorem 6 that the first-order iteration-complexity of Algorithm 5 for finding an ϵ-KKT solution of problem* (1) *is $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$, which clearly improves that of Algorithm 4 in terms of dependence on ϵ.*

5 Numerical results

In this section we conduct some preliminary numerical experiments to test the performance of our proposed algorithms (Algorithms 4 and 5), and compare them with a closely related I-AL method and its modified version proposed in [11], which are denoted by 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

$$\min_{x\in\Re^n} \left\{c^T x : Ax = b,\ l \leq x \leq u\right\} \tag{102}$$

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 (102) 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 4, we set them by (73) with $\rho_0 = 100$ and $\eta_0 = 1$. For Algorithm 5, we choose them by (96) 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 (102), 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.[5] 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.

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

Table 1: Computational results for solving problem (102)

Parameters			Iterations ($\times 10^3$)				CPU Time (in seconds)			
n	m	ζ	Algorithm 4	Algorithm 5	I-AL$_1$	I-AL$_2$	Algorithm 4	Algorithm 5	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

The computational results of all the algorithms for solving problem (102) 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 5 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 4 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 4 is substantially better than that of I-AL$_1$. The main reason is perhaps that Algorithm 4 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 4, but it is substantially outperformed by Algorithm 5.

6 Concluding remarks

In this paper we considered a class of convex conic programming. In particular, we proposed an inexact augmented Lagrangian (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 showed that the total number of first-order iterations of the proposed I-AL method for computing an ϵ-KKT solution is at most $\mathcal{O}(\epsilon^{-7/4})$. We also proposed a modified I-AL method and showed that it has an improved iteration-complexity $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$, which is so far the lowest complexity bound among all first-order I-AL type of methods for computing an ϵ-KKT solution. Our complexity analysis of the I-AL methods is mainly based on an analysis on inexact proximal point algorithm (PPA) and the link between the I-AL methods and inexact PPA, which is substantially different from the existing complexity analyses of the first-order I-AL methods in the literature. The computational results on a set of randomly generated LP problems demonstrated that our modified I-AL method substantially outperforms those in [11] in terms of both total number of first-order iterations and CPU time.

Our current 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 computing an ϵ-KKT solution of problem (1) be further improved for an I-AL method? These will be left for the future research.

A Some properties of closed convex cones

We review some properties of convex cones in this part. Let $\emptyset \neq \mathcal{K} \subseteq \Re^m$ be a closed convex cone and $\mathcal{K}^*$ its dual cone.

Lemma 3 *For any $u, v \in \Re^m$, $\mathrm{dist}(u+v,\mathcal{K}) \le \mathrm{dist}(u,\mathcal{K}) + \mathrm{dist}(v,\mathcal{K})$. Moreover, if $u \preceq_{\mathcal{K}} v$, then $\mathrm{dist}(u,\mathcal{K}) \ge \mathrm{dist}(v,\mathcal{K})$.*

Proof. Notice that $\mathcal{K} = \{w_1 + w_2 : w_1, w_2 \in \mathcal{K}\}$. It follows that

$$\begin{aligned}\mathrm{dist}(u+v,\mathcal{K}) &= \min\{\|u+v-w\| : w \in \mathcal{K}\} = \min\{\|u+v-w_1-w_2\| : w_1, w_2 \in \mathcal{K}\} \\ &\le \min\{\|u-w_1\| : w_1 \in \mathcal{K}\} + \min\{\|v-w_2\| : w_2 \in \mathcal{K}\} = \mathrm{dist}(u,\mathcal{K}) + \mathrm{dist}(v,\mathcal{K}).\end{aligned}$$

Suppose $u \preceq_{\mathcal{K}} v$. Then $v - u \in \mathcal{K}$. Hence, we have

$$\begin{aligned}\mathrm{dist}(v,\mathcal{K}) &= \min\{\|v-w_1-w_2\| : w_1, w_2 \in \mathcal{K}\} \le \min\{\|v-(v-u)-w_2\| : w_2 \in \mathcal{K}\} \\ &= \min\{\|u-w_2\| : w_2 \in \mathcal{K}\} = \mathrm{dist}(u,\mathcal{K}).\end{aligned}$$

□

Lemma 4 *For any $v \in \Re^m$, the following statements hold.*

(a) $v = \Pi_{-\mathcal{K}}(v) + \Pi_{\mathcal{K}^*}(v)$.

(b) $\mathrm{dist}(v,-\mathcal{K}) = \|\Pi_{\mathcal{K}^*}(v)\|$ *and* $\mathrm{dist}(v,\mathcal{K}^*) = \|\Pi_{-\mathcal{K}}(v)\|$.

(c) $\|v\|^2 = \|\Pi_{-\mathcal{K}}(v)\|^2 + \|\Pi_{\mathcal{K}^*}(v)\|^2 = \mathrm{dist}^2(v,-\mathcal{K}) + \mathrm{dist}^2(v,\mathcal{K}^*)$.

Proof. It follows from [25, Exercise 2.8] that for any $v \in \Re^m$, we have

$$v = \Pi_{-\mathcal{K}}(v) + \Pi_{\mathcal{K}^*}(v), \quad \langle \Pi_{-\mathcal{K}}(v), \Pi_{\mathcal{K}^*}(v) \rangle = 0.$$

Using these two equalities, it is not hard to verify that the statements in Lemma 4 hold. □

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. Devolde, 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. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, October, 2017.

[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] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29(3):341–346, 1962.

[15] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. To appear in *Optim. Methods Softw.*, 2017.

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

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

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

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

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

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

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

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

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

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

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

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

[28] Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. arXiv:1711.08020, preprint, 2017.

[29] Y. Xu. Global convergence rates of augmented Lagrangian methods for constrained convex programming. arXiv:1711.05812, preprint, 2017.

[30] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168, 2008.

[31] 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.