# Iteration-complexity of first-order augmented Lagrangian methods for convex programming *

Guanghui Lan † Renato D.C. Monteiro ‡

April 30, 2009

**Abstract**

This paper considers a special class of convex programming (CP) problems whose feasible regions consist of a simple compact convex set intersected with an affine manifold. We present first-order methods for this class of problems based on an inexact version of the classical augmented Lagrangian (AL) approach, where the subproblems are approximately solved by means of Nesterov's optimal method. We then establish a bound on the total number of Nesterov's optimal iterations, i.e., the inner iterations, performed throughout the entire inexact AL method to obtain a near primal-dual optimal solution. We also present variants with better iteration-complexity bounds than the original inexact AL method, which consist of applying the original approach directly to a perturbed problem obtained by adding a strongly convex component to the objective function of the CP problem.

**Keywords:** penalty, first-order, augmented Lagrangian method, convex programming, Lagrange multiplier

## 1 Introduction

The basic problem of interest in this paper is the convex programming (CP) problem

$$f^* := \inf\{f(x) : \mathcal{A}(x) = 0,\, x \in X\}, \tag{1}$$

where $f : X \to \mathbf{R}$ is a convex function with Lipschitz continuous gradient, $X \subseteq \Re^n$ is a sufficiently simple compact convex set and $\mathcal{A} : \Re^n \to \Re^m$ is an affine function.

For the case where the feasible region consists only of the set $X$, or equivalently $\mathcal{A} \equiv 0$, Nesterov ([12, 14]) developed a method which finds a point $x \in X$ such that $f(x) - f^* \leq \epsilon$ in at most $\mathcal{O}(\epsilon^{-1/2})$ iterations. Moreover, each iteration of his method requires one gradient evaluation of $f$ and computation of two projections onto $X$. It is shown that his method achieves, uniformly in the dimension, the lower bound on the number of iterations for minimizing convex functions with

*The work of both authors were partially supported by NSF Grant CCF-0808863 and ONR Grant N00014-08-1-0033.

†School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: glan@isye.gatech.edu).

‡School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: monteiro@isye.gatech.edu).

Lipschitz continuous gradient over a closed convex set. When $\mathcal{A}$ is not identically 0, Nesterov's optimal method can still be applied directly to problem (1) but this approach would require the computation of projections onto the feasible region $X \cap \{x : \mathcal{A}(x) = 0\}$, which for most practical problems is as expensive as solving the original problem itself. An alternative approach for solving (1) when $\mathcal{A}$ is not identically 0 is to use first-order methods whose iterations require only computation of projections onto the simple set $X$.

Following this line of investigation, we studied in [9] two first-order methods for solving (1) based on two well-known penalization approaches, namely: the quadratic and the exact penalization approaches. Iteration-complexity bounds for these methods are then derived to obtain two types of near optimal solutions of (1), namely: near primal and near primal-dual optimal solutions. Variants with possibly better iteration-complexity bounds than the aforementioned methods are also discussed. In this paper, we still consider another first-order approach for solving (1) based on the classical augmented Lagrangian approach, where the subproblems are approximately solved by means of Nesterov's optimal method. As a by-product, alternative first-order methods for solving (1) involving only computation of projections onto the simple set $X$ are obtained.

The augmented Lagrangian method, initially proposed by Hestenes [6] and Powell [16] in 1969, is currently regarded as an effective optimization method for solving large-scale nonlinear programming problems (see textbooks or monographs: [1], [2], [5], [15], [17]). More recently, it has been used by the convex programming (CP) community to develop specialized first-order methods for solving large-scale semidefinite programming problems (see, for example, Burer and Monteiro [3, 4], Jarre and Rendl [8], Zhao et al. [18]), due to its lower iteration-cost compared to that of Newton-based interior-point methods. The augmented Lagrangian method applied to problem (1) consists of solving a sequence of subproblems of the form

$$d_\rho(\lambda_k) := \min_{x \in X} \left\{ \mathcal{L}_\rho(x, \lambda_k) := f(x) + \langle \lambda_k, \mathcal{A}(x) \rangle + \frac{\rho}{2} \|\mathcal{A}(x)\|^2 \right\}, \tag{2}$$

where $\rho > 0$ is a given penalty parameter and $\|\cdot\|$ is the norm associated with a given inner product $\langle \cdot, \cdot \rangle$ in $\Re^m$. The multiplier sequence $\{\lambda_k\}$ is generated according to the iterations

$$\lambda_{k+1} = \lambda_k + \rho \mathcal{A}(x_k^*), \tag{3}$$

where $x_k^*$ is a solution of problem (2). Since in most cases (2) can only be solved approximately, $x_k^*$ in (3) is replaced by an $\eta_k$-approximate solution of (2), i.e., a point $x_k \in X$ such that $\mathcal{L}_\rho(x, \lambda_k) - d_\rho(\lambda_k) \leq \eta_k$. The inexact augmented Lagrangian method obtained in this manner, where the subproblems (2) are solved by Nesterov's method, is the main focus of our investigation in this paper. More specifically, we are interested in establishing a bound on the total number of Nesterov's optimal iterations, i.e., the inner iterations, performed throughout the entire inexact AL method.

Several technical issues are addressed in the aforementioned iteration-complexity analysis of the inexact AL method. First, the notion of a near primal-dual optimal solution is introduced and used as a termination criterion by the methods proposed in this paper. Second, it is well-known that $\mathcal{A}(x_k^*)$ is exactly the gradient of the function $d_\rho$ defined in (2) at $\lambda_k$, and hence that (3) can be viewed as a steepest ascent iteration with stepsize $\rho$ applied to the function $d_\rho$. Since, in the inexact AL method, we approximate $d_\rho(\lambda_k) = \mathcal{A}(x_k^*)$ by $\mathcal{A}(x_k)$, where $x_k$ is an approximate solution of (2), we bound the error of the gradient approximation $\mathcal{A}(x_k)$, namely $\|\mathcal{A}(x_k) - \mathcal{A}(x_k^*)\|$, in terms of the accuracy $\eta_k$ of the approximate solution $x_k$, and use this result to derive sufficient conditions on the sequence $\{\eta_k\}$ which guarantee that the corresponding inexact steepest ascent method $\lambda_{k+1} = \lambda_k + \rho \mathcal{A}(x_k)$

has the same rate of convergence as the exact one. Third, as $\rho$ increases, it is well-known that the iteration-complexity of approximately solving each subproblem (2) increases, while the number of dual iterations (3), i.e., the outer iterations, decreases. Ways of choosing the parameter $\rho$ so as to balance these two opposing criterions are then proposed. More specifically, $\rho$ is chosen so as to minimize the overall number of inner iterations performed by the inexact AL method.

It turns out that proper selection of the tolerances $\eta_k$ and the optimal penalty parameter $\rho$ requires knowledge of an upper bound $t$ on $D_\Lambda := \inf_{\lambda \in \Lambda^*} \|\lambda_0 - \lambda^*\|$, where $\Lambda^*$ is the set of Lagrange multipliers associated with the constraint $\mathcal{A}(x) = 0$. Theoretically, choosing the upper bound $t$ so that $t = \mathcal{O}(D_\Lambda)$ yields the lowest provably iteration-complexity bounds obtained by our analysis. However, since $D_\Lambda$ is not known a priori, we present a "guess-and-check" procedure which consists of guessing a sequence of estimates for $D_\Lambda$ and applying the corresponding sequence of inexact AL methods (with pre-specified number of outer-iterations) to (1) until a near primal-dual solution is eventually obtained. It is shown that the above guess-and-check procedure has the same iteration-complexity as the (ideal) inexact AL method for which the exact value of $D_\Lambda$ is known in advance. Finally, we present variants with better iteration-complexity bounds than the original inexact AL method and guess-and-check procedure, which consist of directly applying the original approaches to a perturbed problem obtained by adding a strongly convex component to the objective function of (1).

Our paper is organized as follows. In Section 2, we review Nesterov's smooth first-order method for solving a certain class of smooth CP problems. In Section 3, we describe two inexact AL methods and corresponding guess-and-check procedures for solving (1) and state without proof their iteration-complexity results. More specifically, we discuss the primal-dual termination criterion used in the complexity analysis of the aforementioned methods in Subsection 3.1. Results about the augmented dual function, including a key result about how to approximate its gradient, are discussed in Subsection 3.2. In Subsection 3.3, we describe the first inexact AL method and its corresponding guess-and-check procedure, and present their iteration-complexity results. The second inexact AL method and its corresponding guess-and-check procedure based on applying the above methods to a perturbed problem, obtained by adding a strongly convex component to the objective function of the CP problem (1), are discussed in Subsection 3.4. All technical results of this paper, which can be skipped by readers interested in the main results only, are presented in Sections 4 and 5. More specifically, we present some technical results about the projected gradient in Subsection 4.1 and about the convergence behavior of the sequence $\{\lambda_k\}$ in Subsection 4.2. Subsections 5.1 and 5.2 give the proofs of the main results in Subsection 3.3 and 3.4, respectively. Finally, we give some concluding remarks in Section 6.

### 1.1 Notation and terminology

We denote the set of real numbers by $\mathbf{R}$. Also, $\mathbf{R}_+$ and $\mathbb{R}_{++}$ denote the set of nonnegative and positive real numbers, respectively. In this paper, we use the notation $\Re^p$ to denote a $p$-dimensional vector space inherited with a inner product space $\langle \cdot, \cdot \rangle$ and use $\| \cdot \|$ to denote the inner product norm in $\Re^p$, i.e., $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$. Moreover, we define the projection map onto a given closed convex set $\mathcal{C} \in \Re^p$ by

$$\Pi_{\mathcal{C}}(u) := \operatorname{argmin}\{\|u - c\| : c \in \mathcal{C}\}, \quad \forall u \in \Re^p.$$

A function $f : \mathcal{C} \subseteq \Re^p \to \mathbb{R}$ is said to have $L$-Lipschitz-continuous gradient with respect to $\| \cdot \|$

if it is differentiable and

$$\|\nabla f(\tilde{u}) - \nabla f(u)\| \leq L\|\tilde{u} - u\|, \quad \forall u, \tilde{u} \in \mathcal{C}. \tag{4}$$

It is well-known (see Theorem 2.1.5 of [13]) that, for every $u, \tilde{u} \in \mathcal{C}$, we have:

$$\frac{1}{2L}\|\nabla f(\tilde{u}) - \nabla f(u)\|^2 \;\leq\; f(\tilde{u}) - f(u) - \langle \nabla f(u), (\tilde{u} - u)\rangle \;\leq\; \frac{L}{2}\|\tilde{u} - u\|^2, \tag{5}$$

$$\frac{1}{L}\|\nabla f(\tilde{u}) - \nabla f(u)\|^2 \;\leq\; \langle \nabla f(\tilde{u}) - \nabla f(u), \tilde{u} - u\rangle \;\leq\; L\|\tilde{u} - u\|^2. \tag{6}$$

## 2 Nesterov's Optimal Method

In this section, we review Nesterov's smooth first-order method for solving a certain class of smooth CP problems. Since the variant of the AL method we consider in this paper uses Nesterov's method to solve the augmented Lagrangian subproblems (2), the results of this section will play an important role in the derivation of iteration-complexity bounds for the above AL variant.

The problem of interest in this section is

$$\phi^* := \min_{x \in X} \phi(x), \tag{7}$$

where $X \subset \Re^n$ is a closed convex set and $\phi : X \to \mathbb{R}$ is a convex function that has $L_\phi$-Lipschitz-continuous gradient over $X$ with respect to a given arbitrary norm $\|\cdot\|$ in $\Re^n$. Moreover, we assume that the optimal value $\phi^*$ of problem (7) is finite and that its set of optimal solutions is nonempty.

Let $h : X \to \mathbb{R}$ be a differentiable strongly convex function with modulus $\sigma > 0$ with respect to $\|\cdot\|$, i.e.,

$$h(x) \geq h(\tilde{x}) + \langle \nabla h(\tilde{x}), x - \tilde{x}\rangle + \frac{\sigma}{2}\|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{8}$$

The Bregman distance $d_h : X \times X \to \mathbf{R}$ associated with $h$ is defined as

$$d_h(x; \tilde{x}) \equiv h(x) - l_h(x; \tilde{x}), \quad \forall x, \tilde{x} \in X, \tag{9}$$

where $l_h : \Re^n \times X \to \mathbf{R}$ is the "linear approximation" of $h$ defined as

$$l_h(x; \tilde{x}) = h(\tilde{x}) + \langle \nabla h(\tilde{x}), x - \tilde{x}\rangle, \quad \forall (x, \tilde{x}) \in \Re^n \times X.$$

We are now ready to state Nesterov's smooth first-order method for solving (7). We use the superscript '$sd$' in the sequence obtained by taking a steepest descent step and the superscript '$ag$'

(which stands for 'aggregated gradient') in the sequence obtained by using all past gradients.

**Nesterov's Algorithm:**

0) Let $x_0^{sd} = x_0^{ag} \in X$ be given and set $k = 0$

1) Set $x_k = \frac{2}{k+2} x_k^{ag} + \frac{k}{k+2} x_k^{sd}$ and compute $\phi(x_k)$ and $\phi'(x_k)$.

2) Compute $(x_{k+1}^{sd}, x_{k+1}^{ag}) \in X \times X$ as

$$x_{k+1}^{sd} \in \operatorname{Argmin} \left\{ l_\phi(x; x_k) + \frac{L_\phi}{2} \|x - x_k\|^2 : x \in X \right\}, \tag{10}$$

$$x_{k+1}^{ag} \equiv \operatorname{argmin} \left\{ \frac{L_\phi}{\sigma} d_h(x; x_0) + \sum_{i=0}^{k} \frac{i+1}{2} [l_\phi(x; x_i)] : x \in X \right\}. \tag{11}$$

3) Set $k \leftarrow k + 1$ and go to step 1.

**end**

The main convergence result established by Nesterov [14] regarding the above algorithm is summarized in the following theorem.

**Theorem 1** *The sequence $\{x_k^{sd}\}$ generated by Nesterov's optimal method satisfies*

$$\phi(x_k^{sd}) - \phi^* \leq \frac{4 L_\phi \, d_h(\bar{x}; x_0^{sd})}{\sigma \, k(k+1)}, \quad \forall k \geq 1,$$

*where $\bar{x}$ is an optimal solution of (7). As a consequence, given any $\epsilon > 0$, an iterate $x_k^{sd} \in X$ satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ can be found in no more than*

$$\left\lceil 2 \sqrt{\frac{d_h(\bar{x}; x_0^{sd}) L_\phi}{\sigma \epsilon}} \right\rceil \tag{12}$$

*iterations.*

The following result is as an immediate special case of Theorem 1.

**Corollary 2** *Suppose that $\|\cdot\|$ is a inner product norm and $h : X \to \Re$ is chosen as $h(\cdot) = \|\cdot\|^2/2$ in Nesterov's optimal method. Then, for any $\epsilon > 0$, an iterate $x_k^{sd} \in X$ satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ can be found in no more than*

$$\left\lceil \|x_0^{sd} - \bar{x}\| \sqrt{\frac{2 L_\phi}{\epsilon}} \right\rceil \tag{13}$$

*iterations, where $\bar{x}$ is an optimal solution of (7).*

*Proof.* If $h(x) = \|x\|^2/2$, then (9) implies that $d_h(\bar{x}; x_0^{sd}) = \|x_0^{sd} - \bar{x}\|^2/2$. The corollary clearly follows from this fact and Theorem 1. ∎

Now assume that the objective function $\phi$ is strongly convex over $X$, i.e., for some $\mu > 0$,

$$\langle \nabla\phi(x) - \nabla\phi(\tilde{x}), x - \tilde{x} \rangle \geq \mu \|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{14}$$

Nesterov shows in Theorem 2.2.2 of [13] that, under the assumptions of Corollary 2, a variant of his optimal method finds a solution $x_k \in X$ satisfying $\phi(x_k) - \phi^* \leq \epsilon$ in no more than

$$\left\lceil \sqrt{\frac{L_\phi}{\mu}} \log \frac{L_\phi \|x_0^{sd} - \bar{x}\|^2}{\epsilon} \right\rceil \tag{15}$$

iterations. The following result gives a slightly sharper iteration-complexity bound for Nesterov's optimal method that replaces the term $\log(L_\phi \|x_0^{sd} - \bar{x}\|^2/\epsilon)$ in (15) with $\log(\mu \|x_0^{sd} - \bar{x}\|^2/\epsilon)$. The proof of this result is given in Theorem 8 of [9].

**Theorem 3** *Let $\epsilon > 0$ be given and suppose that the assumptions of Corollary 2 hold and that the function $\phi$ is strongly convex with modulus $\mu$. Then, the variant where we restart Nesterov's optimal method, with proximal function $h(\cdot) = \|\cdot\|^2/2$, every*

$$K := \left\lceil \sqrt{\frac{8L_\phi}{\mu}} \right\rceil \tag{16}$$

*iterations finds a solution $\tilde{x} \in X$ satisfying $\phi(\tilde{x}) - \phi^* \leq \epsilon$ in no more than $K \max\{1, \lceil \log Q \rceil\}$ iterations, where*

$$Q := \frac{\mu \|x_0^{sd} - \bar{x}\|^2}{2\epsilon} \tag{17}$$

*and $\bar{x} := \operatorname{argmin}_{x \in X} \phi(x)$.*

# 3 The algorithms and main results

In this section, we present the augmented Lagrangian method applied to (1) and discuss its computational complexity. Specifically, we discuss the termination criterion for this method in Subsection 3.1. We review the augmented dual function and discuss some of its properties in Subsection 3.2. In Subsection 3.3, we describe a version of the augmented Lagrangian method and discuss its computational complexity. A variant of this method, for which a perturbation term is added into the objective function of (1), is discussed and analyzed in Subsection 3.4.

## 3.1 Termination criterion

The problem of interest in this paper is the CP problem (1) where $f : X \to \mathbf{R}$ is a convex function with $L_f$-Lipschitz-continuous gradient. The Lagrangian dual function and value function associated with (1) are defined as

$$d(\lambda) := \inf\{f(x) + \langle \lambda, \mathcal{A}(x) \rangle : x \in X\}, \quad \forall \lambda \in \Re^m, \tag{18}$$

$$v(u) := \inf\{f(x) : \mathcal{A}(x) = u, \, x \in X\}, \quad \forall u \in \Re^m. \tag{19}$$

It is well-known that $d$ is always a concave function. Moreover, the assumption we made earlier that $f$ is convex, $\mathcal{A}$ is affine, and $X$ is convex, implies that the function $v$ is convex.

The Lagrangian dual of (1) is the problem

$$d^* := \sup_\lambda d(\lambda). \tag{20}$$

In addition to the convexity assumptions we made about the data of (1), we also assume the following conditions throughout the paper:

**A.1** The function $v(\cdot)$ is closed and $f^* = v(0)$ is finite.

**A.2** The set $\Lambda^*$ of optimal solutions of the dual problem (20) is nonempty.

It is well-known that $d^* = \overline{\text{co}}\, v(0)$, where $\overline{\text{co}}\, v$ is the closed convex hull of $v$. Hence, Assumption A.1 implies that $f^* = v(0) = \overline{\text{co}}\, v(0) = d^*$, i.e., there is no duality gap for the pair of dual problems (1) and (20). Clearly, this implies that $\Lambda^* := \{\lambda^* : d(\lambda^*) = f^*\}$, i.e., $\Lambda^*$ is the set of Lagrange multipliers. Moreover, it is well-known that latter set is also equal to $-\partial v(0)$. It then follows from Assumption A.2 that $v$ is subdifferentiable at 0 and hence that $v$ is proper.

The following result gives a sufficient condition for Assumption A.1 and its proof can be found in the Appendix.

**Proposition 4** *If the set of optimal solutions for problem (1) is nonempty and bounded then Assumption A.1 holds.*

As a consequence of Proposition 4, if $X$ is nonempty and compact, then Assumption A.1 holds.

In this paper, we are interested in obtaining the near-optimal solutions of (1) defined as follows. Note that $x^* \in X$ is an optimal solution of (1) and $\lambda^* \in \Re^m$ is a Lagrange multiplier for (1) if, and only if, $(\tilde{x}, \tilde{\lambda}) = (x^*, \lambda^*)$ satisfies

$$\begin{aligned} &\mathcal{A}(\tilde{x}) = 0, \\ &\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}), \end{aligned} \tag{21}$$

where $\mathcal{N}_X(\tilde{x}) := \{s \in \Re^n : \langle s, x - \tilde{x} \rangle \leq 0, \forall x \in X\}$ denotes the normal cone of $X$ at $\tilde{x}$, and $\mathcal{A}_0$ denotes the linear part of $\mathcal{A}$ defined by $\mathcal{A}_0 := \mathcal{A} - \mathcal{A}(0)$. Based on this observation, we introduce the following notion.

**Definition 1** *For a given pair $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbb{R}_{++}$, $(\tilde{x}, \tilde{\lambda}) \in X \times \Re^m$ is called an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) if*

$$\|\mathcal{A}(x)\|_* \leq \epsilon_p, \tag{22}$$

$$\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}) + \mathcal{B}(\epsilon_d), \tag{23}$$

*where $\mathcal{B}(\eta) := \{x \in \Re^n : \|x\| \leq \eta\}$ for every $\eta \geq 0$.*

The main goal of this paper is to study the iteration-complexity of the augmented Lagrangian method for computing an $(\epsilon_p, \epsilon_d)$-*primal-dual* solution of (1) defined above.

### 3.2 The augmented dual function

In this subsection, we review the definition of the augmented dual function associated with (1) and discuss some of its properties.

Given a penalty parameter $\rho > 0$, the augmented dual function $d_\rho : \Re^m \to \mathbb{R}$ associated with (1) is given by

$$d_\rho(\lambda) := \inf_{x \in X} \left\{ \mathcal{L}_\rho(x, \lambda) := f(x) + \langle \lambda, \mathcal{A}(x) \rangle + \frac{\rho}{2} \|\mathcal{A}(x)\|^2 \right\}, \tag{24}$$

and the augmented dual with parameter $\rho$ is defined as

$$\sup_{\lambda \in \Re^m} d_\rho(\lambda). \tag{25}$$

An alternative characterization for the augmented dual function is given by

$$d_\rho(\lambda) = \inf_u \left\{ v_\rho(u, \lambda) := v(u) + \langle \lambda, u \rangle + \frac{\rho}{2} \|u\|^2 \right\}, \tag{26}$$

where $v(\cdot)$ is the value function given by (19).

**Lemma 5** *The following statements hold:*

*a) problem (26) has an unique optimal solution $u^*_\lambda$;*

*b) the (possibly empty) set of optimal solutions of (24) $X^*_\lambda$ is given by*

$$X^*_\lambda = \{x \in X : \mathcal{A}(x) = u^*_\lambda \; and \; f(x) = v(u^*_\lambda)\}; \tag{27}$$

*c) for any $\lambda \in \Re^m$ and $\rho > 0$, we have*

$$v_\rho(u, \lambda) - d_\rho(\lambda) \geq \frac{\rho}{2} \|u - u^*_\lambda\|^2, \;\; \forall \, u \in \Re^m; \tag{28}$$

*d) problem (25) has the same optimal value and set of optimal solutions as those of (20).*

*Proof.* We first show a). Observe that convexity of $v$ and Assumption A.1 imply that the function $v_\rho(\cdot, \lambda)$ in (26) is a proper lower-semicontinuous convex function for every $\lambda \in \Re^m$ and $\rho > 0$. Moreover, $v_\rho(\cdot, \lambda)$ is strongly convex with modulus $\rho$, that is,

$$v_\rho(\alpha u_1 + (1-\alpha) u_2, \lambda) \leq \alpha v_\rho(u_1, \lambda) + (1-\alpha) v_\rho(u_2, \lambda) - \frac{\rho}{2} \alpha (1-\alpha) \|u_1 - u_2\|^2, \tag{29}$$

for all $(u_1, u_2) \in \Re^m \times \Re^m$ and $\alpha \in (0, 1)$. The above two observations clearly imply a). Statement b) follows directly from a), definition (19), and the equivalence of problems (24) and (26). To show c), we let $u_1 = u$ and $u_2 = u^*_\lambda$ in (29) to obtain

$$\begin{aligned} \frac{\rho}{2} \|u - u^*_\lambda\|^2 &\leq \frac{v_\rho(u, \lambda) - v_\rho(\alpha u + (1-\alpha) u^*_\lambda, \lambda)}{1-\alpha} + \frac{v_\rho(u^*_\lambda, \lambda) - v_\rho(\alpha u + (1-\alpha) u^*_\lambda, \lambda)}{\alpha} \\ &\leq \frac{v_\rho(u, \lambda) - v_\rho(\alpha u + (1-\alpha) u^*_\lambda, \lambda)}{1-\alpha}, \;\; \forall \, \alpha \in (0, 1) \end{aligned}$$

where the last inequality follows from the fact that $u^*_\lambda$ is the optimal solution for problem (26). Letting $\alpha$ go to zero in the above inequality, and using the lower-semicontinuity of $v_\rho$ and the fact that $d_\rho(\lambda) = v_\rho(u^*_\lambda, \lambda)$, we obtain (28). Statement d) is a well-known. ∎

The following proposition summarizes some important properties of $d_\rho$.

**Proposition 6** *For any $\rho > 0$, the function $d_\rho$ is concave, differentiable, and*

$$\nabla d_\rho(\lambda) = u^*_\lambda, \;\; \forall \, \lambda \in \Re^m, \tag{30}$$

*where $u^*_\lambda$ is the unique optimal solution of problem (26). Moreover, $d_\rho$ has $1/\rho$-Lipschitz-continuous gradient with respect to the inner product norm on $\Re^m$.*

*Proof.* Under Assumption A.1, the claim follows immediately from Theorem 1 of [14] applied to the maximization version of (26), i.e., the problem $\max_u\{-v_\rho(u,\lambda)\}$. ∎

In view of Proposition 6 and Lemma 5(b), the exact version of the augmented Lagrangian method stated in Section 1 can be viewed as a version of the steepest ascent method applied to (25). Note that one possible drawback of the exact augmented Lagrangian method is that each iteration of this method requires the solution of problem (2) for computing the gradient $\nabla d_\rho(\lambda_k)$. Since in most applications, problem (2) can only be solved approximately, in this paper we are interested in analyzing the inexact version of the augmented Lagrangian method where the gradient $\nabla d_\rho(\lambda_k)$ is approximated by $\mathcal{A}(x_k)$, where $x_k$ an approximate solution of problem (2).

The following simple but crucial result gives a bound on the error between $\nabla d_\rho(\lambda_k)$ and its aforementioned approximation.

**Proposition 7** *Assume that $(x,\lambda) \in X \times \Re^m$ is such that $\mathcal{L}_\rho(x,\lambda) - d_\rho(\lambda) \leq \eta$. Then, we have*

$$\|\mathcal{A}(x) - \nabla d_\rho(\lambda)\| = \|\mathcal{A}(\tilde{x}) - u_\lambda^*\| \leq \sqrt{\frac{2\eta}{\rho}}, \tag{31}$$

*where $u_\lambda^*$ is the unique optimal solution of (26).*

*Proof.* Letting $u := \mathcal{A}(x)$ and observing that $f(x) \geq v(u)$ due to definition (19), we conclude that

$$\mathcal{L}_\rho(x,\lambda) = f(x) + \langle \lambda, u\rangle + \frac{\rho}{2}\|u\|^2 \geq v(u) + \langle \lambda, u\rangle + \frac{\rho}{2}\|u\|^2 = v_\rho(u,\lambda). \tag{32}$$

This inequality, relation (28), and the assumption that $\mathcal{L}_\rho(x,\lambda) - d_\rho(\lambda) \leq \eta$ then imply that

$$\mathcal{L}_\rho(x,\lambda) - d_\rho(\lambda) \geq v_\rho(u,\lambda) - d_\rho(\lambda) \geq \frac{\rho}{2}\|u - u_\lambda^*\|^2, \tag{33}$$

and hence that (31) holds. ∎

### 3.3 The augmented Lagrangian method

In this subsection, we present the augmented Lagrangian method applied to problem (1) and discuss its convergence behavior.

We start by stating the first inexact AL method that will be studied in this paper.

**The I-AL method:**

**Input:** Initial points $\lambda_0 \in \Re^m$ and $x_{-1} \in X$, penalty parameter $\rho \in \Re_{++}$, outer tolerances $(\epsilon_p, \epsilon_d) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$, iteration limit $\bar{N} \in \mathbb{N} \cup \{+\infty\}$, and inner tolerances $\eta_0, \ldots, \eta_{\bar{N}}$ satisfying

$$0 < \eta_k \leq \frac{\rho \epsilon_p^2}{128}, \quad \forall\, k = 0, \ldots, \bar{K}. \tag{34}$$

0) Set $k = 0$;

1) Using $x_{k-1}$ as starting point, apply Nesterov's optimal method to find an $\eta_k$-approximate solution of problem (2), i.e., a point $x_k \in X$ such that
$$\mathcal{L}_\rho(x_k, \lambda_k) - d_\rho(\lambda_k) \leq \eta_k; \tag{35}$$

2) If $\|\mathcal{A}(x_k)\| \leq 3\epsilon_p/4$, then call subroutine Postprocessing with input $(x, \tilde{\lambda}) = (x_k, \lambda_k)$, report **success**, and terminate the algorithm;

3) Otherwise, if $\|\mathcal{A}(x_k)\| > 3\epsilon_p/4$, set $\lambda_{k+1} = \lambda_k + \rho \mathcal{A}(x_k)$ and increment $k$ by 1;

4) If $k = \bar{N}$, report **failure**, and terminate the algorithm; otherwise, go to step 1.

**end**

We now describe subroutine Postprocessing.

**Postprocessing**$(x, \tilde{\lambda})$:

Set
$$\zeta = \zeta(\rho) := \min\left\{ \frac{\rho \epsilon_p^2}{128}, \frac{\epsilon_d^2}{8 M_\rho} \right\}. \tag{36}$$

P.1) Using $x \in X$ as starting point, apply Nesterov's optimal method to find a $\zeta$-approximate solution $\tilde{x}$ of problem (2);

P.2) Output a pair $(\tilde{x}^+, \tilde{\lambda}^+)$ given by

$$\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla \mathcal{L}_\rho(\tilde{x}, \tilde{\lambda})/M_\rho) \tag{37}$$
$$\tilde{\lambda}^+ := \tilde{\lambda} + \rho \mathcal{A}(\tilde{x}^+). \tag{38}$$

**end**

We will say that an outer iteration of the I-AL method occurs whenever $k$ is incremented by 1 in Step 3. We will refer to an iteration of Nesterov's optimal method to compute $x_k$ in step 1 or $\tilde{x}$

inside subroutine Postprocessing as an inner iteration of the I-AL method.

We now make a few comments about the I-AL method. First, note that the I-AL method is a generic algorithm in the sense that the parameters $\rho$ and $\{\eta_k\}$ have not been specified. Concrete choices of these parameters will be discussed within the context of the convergence results which will be presented in the remaining part of this subsection. Second, in view of Proposition 7, an outer iteration of the I-AL method can be viewed as an iteration of a version of the steepest ascent method with inexact gradient with respect to problem (25). Third, Step 4 ensures that the method terminates in at most $\bar{N}$ outer iterations possibly reporting failure. Fourth, at the beginning of Step 2, the pair $(x_k, \lambda_k)$ satisfies the primal termination condition (22), but not necessarily the dual termination criterion (23). By calling subroutine Postprocessing, the next result, whose proof will be given in Section 5.1, guarantees that the output pair $(\tilde{x}^+, \tilde{\lambda}^+)$ of this subroutine satisfies both (22) and (23).

**Proposition 8** *Let $\rho > 0$, $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, and $\tilde{\lambda} \in \Re^m$ be given and assume that there exists an $x \in X$ satisfying*

$$\|\mathcal{A}(x)\| \leq \frac{3\epsilon_p}{4} \quad \text{and} \quad \mathcal{L}_\rho(x, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \frac{\rho\epsilon_p^2}{128}.$$

*If $\tilde{x} \in X$ is a point satisfying $\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \zeta$, where $\zeta$ is given by (36), then the pair $(\tilde{x}^+, \tilde{\lambda}^+)$ defined by (37) and (38) is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

The following result follows as an immediate consequence of Proposition 8.

**Corollary 9** *If the I-AL method successfully terminates (i.e., at Step 2), then the output pair of subroutine Postprocessing is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

*Proof.* The result follows from Proposition 8, (34), and the fact that at Step 4, conditions (35) and $\|\mathcal{A}(x_k)\| \leq 3\epsilon_p/4$ hold. ∎

Our next result below describes conditions on the parameters $\rho$ and $\{\eta_k\}$ which guarantee the successful termination of the I-AL method.

**Theorem 10** *Let $\rho \in \mathbf{R}_{++}$ and $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. Assume that the iteration limit $\bar{N}$ of the I-AL method satisfies*

$$\bar{N} \geq N := \left\lceil \frac{16 D_\Lambda^2}{\rho^2 \epsilon_p^2} \right\rceil, \tag{39}$$

*where $D_\Lambda := \inf_{\lambda^* \in \Lambda^*} \|\lambda_0 - \lambda^*\|$, and the sequence $\{\eta_k\}_{k=0}^{\bar{N}-1} \subseteq \mathbb{R}_{++}$ satisfies*

$$\sum_{k=0}^{\bar{N}-1} \eta_k \leq \frac{\rho\epsilon_p^2}{128}. \tag{40}$$

*Then, the I-AL method successfully terminates in at most $N$ outer iterations.*

We now make a few observations about Theorem 10. First, we observe that Theorem 10 holds regardless of the method used to find the approximate solution $x_k$ in step 1 or $\tilde{x}$ in subroutine

Postprocessing. Second, although the number of outer iterations of the I-AL method does not depend on $\epsilon_d$, the number of inner iterations will depend on it, since the number of inner iteration inside subroutine Postprocessing clearly depends on $\epsilon_d$ in view of (36). Third, observe that equation (39) implies that the larger $\rho$ is, the smaller the bound $N$ on the number of outer iterations will be. On the other hand, since the Lipschitz constant of the objective function of subproblem (2) is given by

$$M_\rho := L_f + \rho\|\mathcal{A}\|^2, \tag{41}$$

increasing $\rho$ will increase $M_\rho$, and as a consequence, will increase the iteration-complexity bound of Nesterov's optimal method for finding an approximate solution of (2).

The following result provides a bound on the total number of inner iterations, i.e., the iterations performed by Nesterov's optimal method, in the I-AL algorithm.

**Proposition 11** *Let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, $\rho > 0$, $\bar{N} \in \mathbb{N} \cup \{+\infty\}$ and $\{\eta_k\}_{k=0}^{\bar{N}-1} \subseteq \mathbb{R}_{++}$ be given such that conditions (39) and (40) are satisfied. Then, the I-AL method applied to (1) successfully terminates in $N$ outer iterations, and computes an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) in at most $\mathcal{I}_p + \mathcal{I}_d$ inner iterations, where $N$ is defined in Theorem 10,*

$$\mathcal{I}_p := \left\lfloor \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{N-1} \eta_k^{-\frac{1}{2}} + N \right\rfloor, \quad \mathcal{I}_d := \left\lceil 4 D_X \max\left\{ \frac{4 M_\rho^{\frac{1}{2}}}{\rho^{\frac{1}{2}} \epsilon_p}, \frac{M_\rho}{\epsilon_d} \right\} \right\rceil \tag{42}$$

*and*

$$D_X := \max_{x_1, x_2 \in X} \|x_1 - x_2\|. \tag{43}$$

*Proof.* Clearly, in view of Corollary 2 and Theorem 10, the number of inner iterations performed at step 1 of the I-AL method is bounded by

$$\sum_{k=0}^{N-1} \left\lceil D_X \sqrt{\frac{2M_\rho}{\eta_k}} \right\rceil \leq \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{N-1} \eta_k^{-\frac{1}{2}} + N,$$

and hence by $\mathcal{I}_p$. Moreover, by Corollary 2, the number of inner iterations performed at step 2 (inside subroutine PostProcessing) is bounded by $\lceil D_X \sqrt{2M_\rho/\zeta} \rceil$. Using the definition of $\zeta$ in (36), it follows that the number of inner iterations performed at step 3 is bounded by $\mathcal{I}_d$. The claim then easily follows by combining the previous two observations. ∎

We now present a few consequences of the results obtained in Proposition 11. The first one stated below bounds the total number of inner iterations of the I-AL method when a summable sequence $\{\eta_k\}$ satisfying condition (40) is chosen.

**Theorem 12** *Let $\rho > 0$ be an arbitrary penalty parameter and $(\epsilon_p, \epsilon_d) \in \Re_{++} \times \Re_{++}$ be given. If, for some $\xi > 0$, the I-AL method is applied to problem (1) with input $\bar{N} = +\infty$ and*

$$\eta_k = \frac{\xi \rho \epsilon_p^2}{128(1+\xi)(k+1)^{1+\xi}}, \quad \forall\, k \geq 0, \tag{44}$$

*then the I-AL method successfully terminates in $N$ outer iterations and computes an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) in at most*

$$\mathcal{O}\left(\frac{D_X M_\rho^{\frac{1}{2}}}{\rho^{\frac{1}{2}}\epsilon_p}\left[\left(\frac{D_\Lambda}{\rho\epsilon_p}\right)^{3+\xi}+1\right]+\frac{D_X M_\rho}{\epsilon_d}+\frac{D_\Lambda^2}{\rho^2\epsilon_p^2}+1\right) \tag{45}$$

*inner iterations, where $N$ is given by (39). In particular, if*

$$\rho=\frac{4}{\epsilon_p}\left(\frac{(D_\Lambda)^{3+\xi}\epsilon_d}{\|\mathcal{A}\|}\right)^{\frac{1}{4+\xi}}+\frac{L_f}{\|\mathcal{A}\|^2}, \tag{46}$$

*then the I-AL method successfully terminates in*

$$\left\lceil \min\left\{\left(\frac{D_\Lambda\|\mathcal{A}\|}{\epsilon_d}\right)^{\frac{2}{4+\xi}}, \frac{16D_\Lambda^2\|\mathcal{A}\|^4}{L_f^2\epsilon_p^2}\right\}\right\rceil \tag{47}$$

*outer iterations and computes an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) in at most*

$$\mathcal{O}\left(D_X\left(\frac{\|A\|^{\frac{7+2\xi}{4+\xi}}D_\Lambda^{\frac{3+\xi}{4+\xi}}}{\epsilon_p\epsilon_d^{\frac{3+\xi}{4+\xi}}}+\frac{\|\mathcal{A}\|}{\epsilon_p}+\frac{L_f}{\epsilon_d}\right)+\left(\frac{\|\mathcal{A}\|D_\Lambda}{\epsilon_d}\right)^{\frac{2}{4+\xi}}+1\right) \tag{48}$$

*inner iterations.*

We now make a few observations about Theorem 12. First, in contrast to the quadratic penalty method where the penalty parameter should be chosen larger than a certain threshold value in order to derive provable iteration-complexity results (see Lan and Monteiro [9]), the I-AL method has an iteration-complexity bound, namely (45), which holds regardless of the value of the penalty parameter $\rho$. Second, it is not difficult to see that the choice of $\rho$ in (46) gives the best iteration-complexity bound based on (45) up to a constant factor. Third, a drawback of the above result is that the formula for $\rho$ in (46) depends on the unknown value $D_\Lambda$. This drawback will be remedied by the next two results of this subsection.

Instead of choosing a summable sequence $\{\eta_k\}$, the next result assumes $\bar{N}$ is finite and chooses $\eta_0, \ldots, \eta_{\bar{N}-1}$uniformly, and instead of assuming the exact knowledge of $D_\Lambda$, it assumes that an upper bound $t \geq D_\Lambda$ is given. The motivation for choosing $\eta_0, \ldots, \eta_{\bar{N}-1}$ uniformly is that the minimum of the summation term in the definition of $\mathcal{I}_p$ in (42) subject to a condition like (40) occurs exactly when $\eta_0, \ldots, \eta_{N-1}$ is uniformly chosen.

**Theorem 13** *Let $(\epsilon_p, \epsilon_d) \in \Re_{++}\times\Re_{++}$ be given. If, for some $t \geq D_\Lambda$, the I-AL is applied to problem (1) with input*

$$\rho=\rho(t):=\frac{4t^{\frac{3}{4}}\epsilon_d^{\frac{1}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}}\epsilon_p}+\frac{L_f}{\|\mathcal{A}\|^2},\quad \bar{N}=\bar{N}(t):=\left\lceil\frac{16t^2}{\rho(t)^2\epsilon_p^2}\right\rceil, \tag{49}$$

$$\eta_k=\eta(t):=\frac{\rho(t)\epsilon_p^2}{128\bar{N}(t)},\ \forall\, k\geq 0, \tag{50}$$

*then the method successfully terminates in*

$$\left\lceil \min\left\{ \frac{D_\Lambda^2 \|\mathcal{A}\|^{\frac{1}{2}}}{t^{\frac{3}{2}} \epsilon_d^{\frac{1}{2}}}, \frac{16 D_\Lambda^2 \|\mathcal{A}\|^4}{L_f^2 \epsilon_p^2} \right\} \right\rceil \leq \left\lceil \min\left\{ \frac{D_\Lambda^{\frac{1}{2}} \|\mathcal{A}\|^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}}, \frac{16 D_\Lambda^2 \|\mathcal{A}\|^4}{L_f^2 \epsilon_p^2} \right\} \right\rceil \tag{51}$$

*outer iterations and computes an $(\epsilon_p, \epsilon_d)$-primal-dual solution in at most $\mathcal{O}(\mathcal{I}_{pd}(t))$ inner iterations, where*

$$\mathcal{I}_{pd}(t) := \left\lceil D_X \left( \frac{\|\mathcal{A}\|^{\frac{7}{4}} t^{\frac{3}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}} + \frac{\|\mathcal{A}\|}{\epsilon_p} + \frac{L_f}{\epsilon_d} \right) + \left( \frac{t\|\mathcal{A}\|}{\epsilon_d} \right)^{\frac{1}{2}} \right\rceil, \tag{52}$$

*and $D_X$ and $D_\Lambda$ are defined in Theorem 10 and Proposition 11, respectively.*

Observe that the choice of $\rho$, $\bar{N}$, and $\{\eta_k\}$ given by (49) and (50) requires $t \geq D_\Lambda$ so as to guarantee conditions (39) and (40), and hence that the conclusions of Theorem (10) hold. We now develop a guess-and-check procedure that attempts to find such a constant $t$ while at the same time checks for potentially early termination of the procedure.

**I-AL guess-and-check procedure:**

**Input:** Initial points $\lambda_0 \in \Re^m$ and $x_{-1} \in X$, and tolerances $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$.

0) Set $t_0 = \min\{(\beta_0/\beta_1)^{\frac{4}{3}}, (\beta_0/\beta_2)^2\}$ and $j = 0$, where

$$\beta_0 := 1 + \frac{32 D_X \|\mathcal{A}\|}{\epsilon_p}, \; \beta_1 := \frac{32 D_X \|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}}, \; \beta_2 := \frac{\|\mathcal{A}\|^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}}; \tag{53}$$

1) Run the I-AL method with the above input and with $\rho = \rho(t_j)$, $\bar{N} = \bar{N}(t_j)$ and

$$\eta_k = \eta(t_j), \quad k = 0, \ldots, \bar{N}(t_j);$$

2) If the I-AL method successfully terminates, **stop**; Otherwise, if the I-AL method reports failure, set $t_{j+1} = 2t_j$, $j = j + 1$, and go to step 1.

**end**

The following result gives the iteration-complexity of the above procedure for obtaining an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).

**Theorem 14** *Let $(\epsilon_p, \epsilon_d) \in \mathbb{R}_{++} \times \mathbf{R}_{++}$ be given. The I-AL guess-and-check procedure finds an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) in at most $\mathcal{O}(\mathcal{I}_{pd}(D_\Lambda))$ inner iterations, where $\mathcal{I}_{pd}(t)$ is defined by (52).*

It is interesting to compare the iteration-complexity bound obtained in Theorem 14 with the corresponding one obtained for the quadratic penalty method in [9] to compute an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1), namely,

$$\mathcal{O}\left( D_X \left( \frac{\|\mathcal{A}\|^2 D_\Lambda}{\epsilon_p \epsilon_d} + \frac{\|\mathcal{A}\|}{\epsilon_p} + \frac{L_f}{\epsilon_d} \right) + 1 \right).$$

Clearly, the latter one is worse than $\mathcal{O}(\mathcal{I}_{pd}(D_\Lambda))$ by a factor of $\mathcal{O}((\|\mathcal{A}\|D_\Lambda/\epsilon_d)^{\frac{1}{4}})$.

Finally, we make some observations about the possibility of exploiting the warm-start strategy for solving the augmented Lagrangian subproblems (2). Even though we already stated the I-AL method with the warm-start strategy included, i.e., the one in which the approximate solution of the previous subproblem is used as a starting point for the solution of next subproblem, the proofs of the results stated in this subsection make no use of this feature. The difficulty in exploiting this feature here is due to the fact that the objective functions of the augmented Lagrangian subproblems are convex, but not necessarily strongly convex. But in next subsection, by adding a small strongly convex perturbation to the objective function of problem (1), we will be able to guarantee that the objective functions of the corresponding augmented Lagrangian subproblems will be strongly convex, and thereby exploit the warm start strategy for solving the augmented Lagrangian subproblems, and consequently, the original problem (1).

### 3.4 The I-AL method applied to a perturbation problem

In this subsection, we will exploit the possibility of solving problem (1) by applying a slightly modified version of the I-AL algorithm to a perturbed problem obtained by adding a small strongly convex perturbation to the objective function of (1).

We start by introducing the perturbed problem, namely:

$$f_\gamma^* := \min\{f_\gamma(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 : \mathcal{A}(x) = 0, x \in X\}, \tag{54}$$

where $x_0$ is a fixed point in $X$ and $\gamma > 0$ is a prespecified perturbation parameter. It is well-known that if $\gamma$ is sufficiently small, then an approximate solution of (54) will also be an approximate solution of (1).

The following simple lemma relates the optimal values of the perturbation problem (54) and the original problem (1).

**Lemma 15** *Let $f^*$ and $f_\gamma^*$ be the optimal values defined in (1) and (54), respectively. Then,*

$$0 \leq f_\gamma^* - f^* \leq \gamma D_X^2/2, \tag{55}$$

*where $D_X$ is defined in Proposition 11.*

*Proof.* The first inequality in (55) follows immediately from the fact that $f_\gamma \geq f$. Now, let $x^*$ and $x_\gamma^*$ be optimal solutions of (1) and (54), respectively. Then,

$$f_\gamma^* = f(x_\gamma^*) + \frac{\gamma}{2}\|x_\gamma^* - x_0\|^2 \leq f(x^*) + \frac{\gamma}{2}\|x^* - x_0\|^2 \leq f^* + \frac{\gamma D_X^2}{2},$$

from which the second inequality in (55) follows. ∎

In this section, we will derive an iteration-complexity bound for obtaining an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) by applying the I-AL method directly to the perturbed problem (54) for a conveniently chosen perturbation parameter $\gamma > 0$.

The augmented dual function associated with (54) is given by

$$d_{\rho,\gamma}(\lambda) := \min_{x \in X} \left\{ \mathcal{L}_{\rho,\gamma}(x, \lambda) := f_\gamma(x) + \lambda^T \mathcal{A}(x) + \frac{\rho}{2}\|\mathcal{A}(x)\|^2 \right\}, \tag{56}$$

or alternatively, by

$$d_{\rho,\gamma}(\lambda) = \inf_u \left\{ v_{\rho,\gamma}(u,\lambda) := v_\gamma(u) + \langle \lambda, u \rangle + \frac{\rho}{2} \|u\|^2 \right\}, \tag{57}$$

where $v_\gamma(\cdot)$ is the value function associated with the perturbed problem (54) (see definition (19)). We denote the optimal solution of (57) by $u^*_{\lambda,\gamma}$.

It can be easily seen that the function $\mathcal{L}_{\rho,\gamma}(\cdot,\lambda)$ has $M_{\rho,\gamma}$-Lipschitz continuous gradient where

$$M_{\rho,\gamma} := L_f + \rho\|A\|^2 + \gamma, \tag{58}$$

and that it is strongly convex with modulus $\gamma$ with respect to $\|\cdot\|$.

We now describe a modification of the I-AL method.

**The Modified I-AL method:** This method is the same as I-AL method applied to the perturbed problem (54) (and hence with $M_\rho$, $\mathcal{L}_\rho$, and $d_\rho$ replaced by $M_{\rho,\gamma}$, $\mathcal{L}_{\rho,\gamma}$, and $d_{\rho,\gamma}$ ) except that instead of Nesterov's method, its variant described in Theorem 3 is used to compute the approximate solutions $x_k$ in step 1 and $\tilde{x}$ in subroutine Postprocessing, and the tolerance $\zeta$ in (36) is replaced by

$$\tilde{\zeta} = \tilde{\zeta}(\rho,\gamma) := \min\left\{ \frac{\rho\epsilon_p^2}{128}, \frac{\epsilon_d^2}{32 M_{\rho,\gamma}} \right\}. \tag{59}$$

The next results is a corresponding version of Proposition 8, which guarantees that the output pair $(\tilde{x}^+, \tilde{\lambda}^+)$ of subroutine Postprocessing is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).

**Proposition 16** *Let $\rho > 0$, $(\epsilon_p, \epsilon_d) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$, and $\tilde{\lambda} \in \Re^m$ be given, and define*

$$\gamma := \frac{\epsilon_d}{2D_X}. \tag{60}$$

*Assume that there exists an $x \in X$ satisfying*

$$\|\mathcal{A}(x)\| \le \frac{3\epsilon_p}{4} \quad \text{and} \quad \mathcal{L}_{\rho,\gamma}(x,\tilde{\lambda}) - d_{\rho,\gamma}(\tilde{\lambda}) \le \frac{\rho\epsilon_p^2}{128}.$$

*If $\tilde{x} \in X$ is a point satisfying $\mathcal{L}_{\rho,\gamma}(\tilde{x},\tilde{\lambda}) - d_{\rho,\gamma}(\tilde{\lambda}) \le \tilde{\zeta}$, where $\tilde{\zeta}$ is given by (59), then the pair $(\tilde{x}^+, \tilde{\lambda}^+)$ defined by (37) and (38) with $\mathcal{L}_\rho$ replaced by $\mathcal{L}_{\rho,\gamma}$ is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

The following result follows as an immediate consequence Proposition 8.

**Corollary 17** *If the modified I-AL method successfully terminates (i.e., at Step 2), then the output pair of subroutine Postprocessing is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

*Proof.* The result follows from Proposition 16, (34), and the fact that at Step 4, conditions (35) and $\|\mathcal{A}(x_k)\| \le 3\epsilon_p/4$ hold. ∎

We now state the corresponding versions of Theorems 13 and 14 with respect to the modified I-AL method.

**Theorem 18** *Let $(\epsilon_p, \epsilon_d) \in \Re_{++} \times \Re_{++}$ be given, and let $\gamma$ be given by (60). For some $t > 0$, consider the modified I-AL method applied to the perturbed problem (54) with input*

$$\rho = \rho_\gamma(t) \;:=\; \frac{4t}{\epsilon_p(\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f + \gamma}{\|\mathcal{A}\|^2}, \tag{61}$$

$$\bar{N} = \bar{N}_\gamma(t) \;:=\; \left\lceil \frac{16t^2}{\rho_\gamma(t)^2 \epsilon_p^2} \right\rceil, \quad \eta_k = \eta_\gamma(t) := \frac{\rho_\gamma(t)\epsilon_p^2}{128\bar{N}_\gamma(t)}, \quad \forall\ k \geq 0, \tag{62}$$

*where*

$$\mathcal{T}(t) := \mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3, \tag{63}$$

$$\mathcal{S}_1 := \sqrt{\frac{D_X \|\mathcal{A}\|^2}{\epsilon_p \epsilon_d}}, \quad \mathcal{S}_2 := \sqrt{\frac{D_X L_f}{\epsilon_d}} + 1 \quad \textit{and } \mathcal{S}_3 := \sqrt{\frac{D_X \|\mathcal{A}\|}{\epsilon_p}} + 3. \tag{64}$$

*Then the following statements hold:*

*a) the total number of inner iterations performed by the above method is bounded by*

$$\mathcal{O}\left\{ \left( \mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 \left[\log \mathcal{T}(t)\right]^{\frac{1}{4}} \right) \left[\log \mathcal{T}(t)\right]^{\frac{3}{4}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t)}{t}\right) \right\}; \tag{65}$$

*b) if $t \geq D_\Lambda^\gamma$, where $D_\Lambda^\gamma := \inf_{\lambda_\gamma \in \Lambda_\gamma^*} \|\lambda_0 - \lambda^*\|$ and $\Lambda_\gamma^*$ denotes the set of Lagrange multipliers associated with (54), then the above method successfully terminates in $\mathcal{O}(\log \mathcal{T}(t))$ outer iterations with an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

Observe that the choice of $\rho$, $\bar{N}$, and $\{\eta_k\}$ given by (61) and (62) requires $t \geq D_\Lambda$ to guarantee the successful termination of the modified I-AL method. We now develop a guess-and-check procedure that attempts to find such a constant $t$ while at the same time checks for potentially early termination of the procedure.

**The modified I-AL guess-and-check procedure:**

**Input:** Initial points $\lambda_0 \in \Re^m$ and $x_{-1} \in X$, and tolerances $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$.

0) Let scalar $\hat{t}$ and function $\psi : \Re^+ \to \Re$ be defined as

$$\hat{t} := \left[ \frac{\mathcal{S}_2^2 + \mathcal{S}_2\sqrt{\mathcal{S}_2^2 + 4(\mathcal{S}_2 + \mathcal{S}_3)}}{2\mathcal{S}_1} \right]^2, \quad \psi(t) := \mathcal{S}_1 t^{\frac{1}{2}} - \mathcal{S}_2 \left[ \log(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \right]^{\frac{1}{4}}, \tag{66}$$

where $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ are given by (64). Find a point $t_0 \in [0, \hat{t}]$ such that $0 \leq \psi(t_0) \leq 1$.

1) Run the modified I-AL method with the above input and with $\rho = \rho_\gamma(t_j)$, $\bar{N} = \bar{N}_\gamma(t_j)$, $\eta_k = \eta_\gamma(t_j)$ for $k \geq 0$, where $\gamma$ is given by (60), and $\rho_\gamma(\cdot)$, $\bar{N}_\gamma(\cdot)$ and $\eta_\gamma(\cdot)$ are defined in (61) and (62).

2) If the modified I-AL method successfully terminates, **stop**; otherwise, set $t_{j+1} = 2t_j$, $j = j + 1$, and go to step 1.

**end**

We now discuss the issue about the existence of $t_0$ satisfying $0 \leq \psi(t_0) \leq 1$. It will be shown in Lemma 31 that $\psi(0) \leq 0$, $\psi(\hat{t}) \geq 0$, and function $\psi$ is non-decreasing. This clearly implies the existence of the required $t_0$. Moreover, $t_0$ can be computed as follows. If $\psi(\hat{t}) \leq 1$, we can take $t_0 = \hat{t}$. Otherwise, a binary search procedure starting with the interval $[0, \hat{t}]$, which must contain the desired scalar $t_0$, determines such a scalar in $\log \hat{t}$ iterations.

The following result gives the iteration-complexity of the above procedure for obtaining an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).

**Theorem 19** *Let $(\epsilon_p, \epsilon_d) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$ be given. The modified I-AL guess-and-check procedure described above finds an $(\epsilon_p, \epsilon_d)$ -primal-dual solution of (1) in at most*

$$\mathcal{O}\left\{\mathcal{S}_1[D_\Lambda^\gamma]^{\frac{1}{2}}[\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma) + \mathcal{S}_2 \log \mathcal{T}(0) \log\log \mathcal{T}(0)\right\}, \tag{67}$$

*inner iterations, where $\mathcal{S}_1, \mathcal{S}_2$, $\mathcal{T}(\cdot)$ and $D_\Lambda^\gamma$ are defined in Theorem 18.*

It is interesting to compare the iteration-complexity bound obtained in Theorem 19 with the corresponding one obtained for the quadratic penalty method in [9] to compute an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1), namely, $\mathcal{O}\left(\mathcal{T}(\|\lambda_\gamma^*\|) \log \mathcal{T}(\|\lambda_\gamma^*\|)\right)$, where $\lambda_\gamma^*$ is the minimum-norm Lagrange multiplier for the perturbed problem (54). Clearly, if the initial multiplier $\lambda_0 = 0$, then $\|\lambda_\gamma^*\| = D_\lambda^\gamma$ and the latter complexity bound reduces to $\mathcal{O}\left(\mathcal{T}(D_\Lambda^\gamma) \log \mathcal{T}(D_\Lambda^\gamma)\right)$. Note that for the situation where

$$\mathcal{S}_2 \log \mathcal{T}(0) \log\log \mathcal{T}(0) = \mathcal{O}\left\{\mathcal{S}_1[D_\Lambda^\gamma]^{\frac{1}{2}}[\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma)\right\}, \tag{68}$$

bound (67) is majorized by $\mathcal{O}\left(\mathcal{T}(D_\Lambda^\gamma)[\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma)\right)$. Clearly, inequality (68) holds if $L_f = 0$. Hence, when $\lambda_0 = 0$ and (68) holds, the first complexity bound is worse than the latter one in Theorem 19 by a factor of $(\log \mathcal{T}(D_\Lambda^\gamma))^{\frac{1}{4}} / \log\log \mathcal{T}(D_\Lambda^\gamma)$. It should be mentioned that if a good warm-start $\lambda_0$ for problem (54) is known, i.e., the ratio $D_\Lambda^\gamma / \|\lambda_\gamma^*\|$ is small, then the complexity bound in Theorem 19 is substantially smaller than the above one.

# 4 Basic Tools

This section discusses some technical results that will be used in our analysis. It consists of two subsections. The first one develops several technical results involving projected gradients. The second subsection develops the convergence results for the steepest descent method with inexact gradient, which will play a crucial role in our analysis for the augmented Lagrangian methods.

## 4.1 Projected gradient and the optimality conditions

In this subsection, we assume that the inner product space $\Re^n$ is endowed with the norm $\|\cdot\|$ associated with its inner product and consider the CP problem (7).

It is well-known that $x^* \in X$ is an optimal solution of (7) if and only if $\nabla\phi(x^*) \in -\mathcal{N}_X(x^*)$. Moreover, this optimality condition is in turn related to the projected gradient of the function $\phi$ over $X$ defined as follows.

**Definition 2** *Given a fixed constant $\tau > 0$, we define the projected gradient of $\phi$ at $\tilde{x} \in X$ with respect to $X$ as (see, for example, [13])*

$$\nabla\phi(\tilde{x})]_X^\tau := \frac{1}{\tau}\left[\tilde{x} - \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))\right], \tag{69}$$

*where $\Pi_X(\cdot)$ is the projection map onto $X$ defined in terms of the inner product norm $\|\cdot\|$ (see Subsection 1.1).*

The following proposition (see Proposition 4 in [9] for the proof) relates the projected gradient to the aforementioned optimality condition.

**Proposition 20** *Let $\tilde{x} \in X$ be given and define $\tilde{x}^+ := \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))$. Then, for any given $\epsilon \geq 0$, the following statements hold:*

*a) $\|\nabla\phi(\tilde{x})]_X^\tau\| \leq \epsilon$ if, and only if, $\nabla\phi(\tilde{x}) \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon)$;*

*b) $\|\nabla\phi(\tilde{x})]_X^\tau\| \leq \epsilon$ implies that $\nabla\phi(\tilde{x}^+) \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}\left((1 + \tau L_\phi)\epsilon\right)$.*

The following result, whose proof is given in Lemma 5 of [9], states some properties of the projected gradient.

**Lemma 21** *Assume that $x^* \in \mathrm{Argmin}_{x\in X}\phi(x)$. Let $\tilde{x} \in X$ be given and define*

$$\tilde{x}^+ := \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x})).$$

*Then, the following statements hold:*

*a) $\phi(\tilde{x}^+) - \phi(\tilde{x}) \leq -\tau\|\nabla\phi(\tilde{x})]_X^\tau\|^2/2$ for any $\tau \leq 1/L_\phi$;*

*b) for any $x \in X$, we have*

$$\phi(x) - \phi(x^*) \geq \frac{1}{2L_\phi}\|\nabla\phi(x)]_X^{1/L_\phi}\|^2. \tag{70}$$

## 4.2 Steepest descent method with inexact gradient

In this subsection, we consider the unconstrained problem

$$p^* := \inf\{p(\lambda) : \lambda \in \Re^m\}, \tag{71}$$

where $p : \Re^m \to \mathbf{R}$ is convex and has $L_p$-Lipschitz-continuous gradient. We assume throughout this subsection that $p^*$ is finite and that the set of optimal solutions $\Gamma^*$ of (71) is nonempty. We are interested in the situation where the gradient $\nabla p(\lambda)$ at any given $\lambda \in \Re^m$ can only be evaluated approximately. This situation arises for example in the case where $p = -d_\rho$, where the computation of the exact gradient requires finding the exact optimal solution of the nonlinear optimization problem (26) (see Proposition 6). The aim is to apply the results obtained here to the function $p = -d_\rho$ in order to prove the main convergence results of the augmented Lagrangian method discussed in Sections 3.3 and 3.4.

An iterate of the steepest descent method with inexact gradient for solving problem (71) consists of:

$$\lambda_{k+1} = \lambda_k - \frac{\alpha_k}{L_p} p'_k \tag{72}$$

where $\alpha_k > 0$ is the stepsize and $p'_k$ is an approximation of the gradient $\nabla p(\lambda_k)$. Define the deviation and the relative deviation between $p'_k$ and $\nabla p(\lambda_k)$ respectively by

$$\delta_k := p'_k - \nabla p(\lambda_k), \qquad e_k := \frac{\|\delta_k\|}{\|p'_k\|}. \tag{73}$$

Before stating the main result of this subsection about the convergence of the inexact steepest descent method, we first present a few technical results.

**Lemma 22** *If $e_k \leq 1 - \alpha_k/2$, then $p(\lambda_{k+1}) \leq p(\lambda_k)$.*

*Proof.* Using the second inequality of (5) with $\lambda = \lambda_k$ and $\tilde{\lambda} = \lambda_{k+1}$, relations (72) and (73), and the Cauchy-Schwartz inequality, we conclude that

$$\begin{aligned} p(\lambda_{k+1}) - p(\lambda_k) &\leq \langle \nabla p(\lambda_k), \lambda_{k+1} - \lambda_k \rangle + \frac{L_p}{2}\|\lambda_{k+1} - \lambda_k\|^2 \\ &= -\frac{\alpha_k}{L_p} \langle p'_k - \delta_k, p'_k \rangle + \frac{\alpha_k^2}{2L_p} \|p'_k\|^2 \;=\; -\frac{\alpha_k}{L_p}\|p'_k\|^2 \left(1 - \frac{\alpha_k}{2} - \frac{\|\delta_k\|}{p'_k}\right) \\ &= -\frac{\alpha_k}{L_p} \|p'_k\|^2 \left(1 - \frac{\alpha_k}{2} - e_k\right) \leq 0, \end{aligned}$$

where the last inequality is due to the assumption that $e_k \leq 1 - \alpha_k/2$. ∎

**Lemma 23** *Assume that $e_k < 1$. Then, for every $\lambda^* \in \Lambda^*$, we have*

$$\alpha_k \beta_k \langle \nabla p(\lambda_k), \lambda_k - \lambda^* \rangle \leq \frac{L_p}{2} \left(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right) + \alpha_k \langle \delta_k, \lambda^* - \lambda_k \rangle, \tag{74}$$

*where*

$$\beta_k := 1 - \alpha_k/[2(1 - e_k)^2]. \tag{75}$$

*Proof.* First note that, by (73), we have

$$\|\nabla p(\lambda_k)\| = \|p'_k - \delta_k\| \geq \|p'_k\| - \|\delta_k\| = (1 - e_k)\|p'_k\|. \tag{76}$$

This inequality, the assumption that $e_k < 1$ and relations (72) and (73) then imply

$$\begin{aligned} \|\lambda_{k+1} - \lambda^*\|^2 &= \left\|\lambda_k - \frac{\alpha_k}{L_p} p'_k - \lambda^*\right\|^2 = \|\lambda_k - \lambda^*\|^2 - \frac{2\alpha_k}{L_p}\langle p'_k, \lambda_k - \lambda^* \rangle + \frac{\alpha_k^2}{L_p^2}\|p'_k\|^2 \\ &\leq \|\lambda_k - \lambda^*\|^2 - \frac{2\alpha_k}{L_p}\langle \nabla p(\lambda_k) + \delta_k, \lambda_k - \lambda^* \rangle + \frac{\alpha_k^2}{L_p^2(1 - e_k)^2}\|\nabla p(\lambda_k)\|^2 \\ &\leq \|\lambda_k - \lambda^*\|^2 + \frac{2\alpha_k}{L_p}\langle \delta_k, \lambda^* - \lambda_k \rangle - \frac{2\alpha_k}{L_p}\left(1 - \frac{\alpha_k}{2(1 - e_k)^2}\right) \langle \nabla p(\lambda_k), \lambda_k - \lambda^* \rangle, \end{aligned}$$

where the last inequality follows from the first inequality in (6) and the fact that $\nabla p(\lambda^*) = 0$. Rearranging the later inequality and using the definition of $\beta_k$, we obtain (74). ∎

**Lemma 24** *Assume that, for some constant $c_1 > 0$, we have*

$$e_k \leq 1 - \sqrt{\frac{\alpha_k + c_1}{2}}. \tag{77}$$

*Then, for any $\lambda^* \in \Lambda^*$, we have*

$$\alpha_k[p(\lambda_k) - p^*] \leq \frac{L_p}{c_1}\left[\left(1 + \frac{2\alpha_k e_k^2}{c_1}\right)\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right]. \tag{78}$$

*Proof.* By the Cauchy-Schwartz inequality and relations (73), (5), (74) and (76), we have

$$\begin{aligned}
&\frac{L_p}{2}\left(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right) + \alpha_k e_k \|p_k'\|\,\|\lambda_k - \lambda^*\| \\
&\quad \geq \frac{L_p}{2}\left(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right) + \alpha_k \langle\, \delta_k, \lambda^* - \lambda_k \,\rangle \\
&\quad \geq \alpha_k\beta_k \langle\, \nabla p(\lambda_k), \lambda_k - \lambda^* \,\rangle \;\geq\; \alpha_k\beta_k\left([p(\lambda_k) - p(\lambda^*)] + \frac{1}{2L_p}\|\nabla p(\lambda_k)\|^2\right) \\
&\quad \geq \alpha_k\beta_k\left([p(\lambda_k) - p(\lambda^*)] + \frac{1}{2L_p}(1 - e_k)^2\|p_k'\|^2\right).
\end{aligned}$$

Letting $x = \|p_k'\|/(L_p\|\lambda_k - \lambda^*\|)$ and rearranging the above inequality, we conclude that

$$\alpha\beta_k[p(\lambda_k) - p(\lambda^*)] \leq \frac{L_p}{2}\left[\left(1 + 2\alpha_k e_k x - \alpha_k\beta_k(1 - e_k)^2 x^2\right)\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right].$$

Relation (78) now follows from the above inequality by noting that (75) and (77) imply that

$$\beta_k \geq (1 - e_k)^2\beta_k = (1 - e_k)^2 - \frac{\alpha_k}{2} \geq \frac{c_1}{2} \tag{79}$$

and that the quadratic function $1 + 2\alpha_k e_k x - \alpha_k\beta_k(1 - e_k)^2 x^2$ is bounded above by

$$1 + \frac{\alpha_k e_k^2}{\beta_k(1 - e_k)^2} \leq 1 + \frac{2\alpha_k e_k^2}{c_1}.$$

■

The following theorem states the convergence properties of the inexact steepest descent method described above.

**Theorem 25** *Assume that for some positive constants $c_1$, we have*

$$e_k \leq 1 - \sqrt{\frac{\alpha_k + c_1}{2}} \tag{80}$$

*for every $k \geq 0$. Then, the sequence $\{\lambda_k\}$ generated by the inexact steepest descent method (72) satisfies*

$$p(\lambda_k) - p^* \leq \frac{L_p}{c_1 \sum_{i=0}^k \alpha_i}\left[\|\lambda_0 - \lambda^*\|^2 \exp\left(\sum_{i=0}^k \frac{2\alpha_i e_i^2}{c_1}\right) - \|\lambda_{k+1} - \lambda^*\|^2\right] \tag{81}$$

*for every $\lambda^* \in \Lambda^*$, where $p^*$ is defined in (72).*

*Proof.* Using Lemma 24, it is easy to see by induction that

$$\sum_{i=0}^{k} \alpha_i [p(\lambda_i) - p^*] \leq \frac{L_p}{c_1} \left[ \|\lambda_0 - \lambda^*\|^2 \prod_{i=0}^{k} \left( 1 + \frac{2\alpha_k e_k^2}{c_1} \right) - \|\lambda_{k+1} - \lambda^*\|^2 \right] \tag{82}$$

for every $k \geq 0$. The above inequality, Lemmas 22 and 24, the inequality $\log(1+x) \leq x$ for any $x > -1$ and assumption (80) then imply that

$$\begin{aligned} \left( \sum_{i=0}^{k} \alpha_i \right) [p(\lambda_k) - p^*] &\leq \sum_{i=0}^{k} \alpha_i [p(\lambda_i) - p^*] \\ &\leq \frac{L_p}{c_1} \left[ \|\lambda_0 - \lambda^*\|^2 \exp\left( \sum_{i=0}^{k} \log(1 + 2\alpha_i e_i^2 / c_1) \right) - \|\lambda_{k+1} - \lambda^*\| \right] \\ &\leq \frac{L_p}{c_1} \left[ \|\lambda_0 - \lambda^*\|^2 \exp\left( \sum_{i=0}^{k} \frac{2\alpha_i e_i^2}{c_1} \right) - \|\lambda_{k+1} - \lambda^*\| \right] \end{aligned}$$

for every $k \geq 0$. ∎

As a consequence of Theorem 25, we obtain the following result which gives an upper bound on the quantities $\|\nabla p(\lambda_k)\|$ and $\|p'(\lambda_k)\|$.

**Corollary 26** *Assume that, for some positive constant $c_1$, relation (80) holds for every $k \geq 0$. Then, the sequence $\{\lambda_k\}$ generated by the inexact steepest descent method (72) satisfies*

$$\frac{\alpha_k + c_1}{2} \|p'_k\|^2 \leq \|\nabla p(\lambda_k)\|^2 \leq \frac{2L_p^2 \|\lambda_0 - \lambda^*\|^2}{c_1 \sum_{i=0}^{k} \alpha_i} \exp\left( \sum_{i=0}^{k} \frac{2\alpha_i e_i^2}{c_1} \right) \tag{83}$$

*for every $\lambda^* \in \Lambda^*$.*

*Proof.* Clearly by definition of $e_k$, we have $\|\nabla p(\lambda_k)\| \geq (1 - e_k)\|p'_k\|$, which together with (80), imply that $\|\nabla p(\lambda_k)\|^2 \geq (\alpha_k + c_1)\|p'_k\|^2/2$. Moreover, using (5), (81), and the fact that $\nabla p(\lambda^*) = 0$, we conclude that

$$\|\nabla p(\lambda_k)\|^2 \leq 2L_p(p(\lambda_k) - p^*) \leq \frac{2L_p^2 \|\lambda_0 - \lambda^*\|^2}{c_1 \sum_{i=0}^{k} \alpha_i} \exp\left( \sum_{i=0}^{k} \frac{2\alpha_i e_i^2}{c_1} \right).$$

Our claim clearly follows from the above two observations. ∎

# 5 Convergence Analysis

In this section, we prove the main results presented in Subsections 3.3 and 3.4.

### 5.1 Convergence analysis for the I-AL method

The goal of this subsection is to prove the convergence results for the I-AL method stated in Subsection 3.3, namely: Proposition 8, Proposition 11 and Theorems 10, 12, 13 and 14.

We first give the proof of Proposition 8 which guarantees that subroutine PostProcessing of the I-AL method outputs an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).

**Proof of Proposition 8:** Clearly, by Lemma 21(b) with $\phi(\cdot) = \mathcal{L}_\rho(\cdot, \tilde{\lambda})$ and $L_\phi = M_\rho$, we have

$$\|\nabla \mathcal{L}_\rho(\tilde{x}, \tilde{\lambda})]_X^{1/M_\rho}\| \leq \left\{2M_\rho \left[\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda})\right]\right\}^{\frac{1}{2}} \leq \sqrt{2M_\rho\,\zeta} \leq \frac{\epsilon_d}{2},$$

where the second and last inequalities follow from the assumption that $\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \zeta$ and relation (36), respectively. The above inequality together with (24), (38) and Proposition 20(b) with $\phi(\cdot) = \mathcal{L}_\rho(\cdot, \tilde{\lambda})$, $L_\phi = M_\rho$ and $\tau = 1/M_\rho$ then imply that

$$\nabla f(\tilde{x}^+) + (\mathcal{A}_0)^* \tilde{\lambda}^+ = \nabla f(\tilde{x}^+) + (\mathcal{A}_0)^*(\tilde{\lambda} + \rho \mathcal{A}(\tilde{x}^+)) = \nabla \mathcal{L}_\rho(\tilde{x}^+, \tilde{\lambda}) \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d),$$

where $\tilde{x}^+$ is defined in (37). Moreover, it follows from Lemma 21(a) with $\phi(\cdot) = \mathcal{L}_\rho(\cdot, \tilde{\lambda})$, $L_\phi = M_\rho$ and $\tau = 1/M_\rho$ that $\mathcal{L}_\rho(\tilde{x}^+, \tilde{\lambda}) \leq \mathcal{L}_\rho(\tilde{x}, \tilde{\lambda})$. This observation, the assumption that $\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \zeta$ and (36) then imply that

$$\mathcal{L}_\rho(\tilde{x}^+, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \zeta \leq \frac{\rho \epsilon_p^2}{128}.$$

Using this conclusion, the assumption that $\mathcal{L}_\rho(x, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \rho\epsilon_p^2/128$ and Proposition 7, we then obtain

$$\max\{\|\mathcal{A}(\tilde{x}^+) - u_\lambda^*\|, \|\mathcal{A}(x) - u_\lambda^*\|\} \leq \frac{\epsilon_p}{8},$$

which together with the assumption that $\|\mathcal{A}(x)\| \leq 3\epsilon_p/4$ imply

$$\|\mathcal{A}(\tilde{x}^+)\| \leq \|\mathcal{A}(\tilde{x}^+) - u_\lambda^*\| + \|\mathcal{A}(x) - u_\lambda^*\| + \|\mathcal{A}(x)\| \leq \frac{\epsilon_p}{8} + \frac{\epsilon_p}{8} + \frac{3\epsilon_p}{4} = \epsilon_p. \tag{84}$$

We have thus shown that $(\tilde{x}^+, \tilde{\lambda}^+)$ is an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1). ∎

Theorem 10 states certain conditions on the parameters $\rho$ and $\eta_k$ which guarantee that the I-AL method will successfully terminate in at most $N$ outer iterations. We now give a proof of this result.

**Proof of Theorem 10:** Since $\bar{N} \geq N$ by assumption, the I-AL method does not terminate with failure within the first $N$ outer iterations. Assume for contradiction that the I-AL method does not successfully terminate within the first $N$ outer iterations. This implies that $\|\mathcal{A}(x_k)\| > 3\epsilon_p/4$ for all $0 \leq k \leq N-1$. Letting $\delta_k := \|\mathcal{A}(x_k) - u_{\lambda_k}^*\|$ and $e_k := \delta_k / \|\mathcal{A}(x_k)\|$ for all $k \geq 0$, we conclude from the previous observation, (35), Proposition 7 and assumptions (39) and (40) that

$$\sum_{k=0}^{N-1} e_k^2 = \sum_{k=0}^{N-1} \frac{\delta_k^2}{\|\mathcal{A}(x_k)\|^2} \leq \frac{16}{9\epsilon_p^2} \sum_{k=0}^{N-1} \|\mathcal{A}(x_k) - u_{\lambda_k}^*\|^2 \leq \frac{32}{9\rho\epsilon_p^2} \sum_{k=0}^{N-1} \eta_k \leq \frac{32}{9\rho\epsilon_p^2} \sum_{k=0}^{\bar{N}-1} \eta_k \leq \frac{1}{36}. \tag{85}$$

Noting that (85) implies $e_k \leq 1/6$, and hence that condition (80) holds with $\alpha_k = 1$ and $c_1 = 7/18$, it follows from (85) and Corollary 26 with $p(\cdot) = -d_\rho(\cdot)$, $L_p = 1/\rho$, $p'_k = \mathcal{A}(x_k)$, $c_1 = 7/18$ and $\alpha_k = 1$ that

$$\|\mathcal{A}(x_k)\|^2 \leq \frac{4D_\Lambda^2}{c_1(1+c_1)\,\rho^2(k+1)}\exp\left(\frac{2}{c_1}\sum_{j=0}^{k} e_j^2\right) \leq \frac{1296D_\Lambda^2}{175\rho^2(k+1)}\exp\left(\frac{1}{7}\right) \leq \frac{9D_\Lambda^2}{\rho^2(k+1)}, \tag{86}$$

for every $0 \leq k \leq N-1$. The above inequality with $k = N-1$ together with (39) then imply that

$$\|\mathcal{A}(x_{N-1})\|^2 \leq \frac{9D_\Lambda^2}{\rho^2 N} \leq \frac{9\epsilon_p^2}{16},$$

which clearly contradicts the fact $\|\mathcal{A}(x_{N-1})\| > 3\epsilon_p/4$. ∎

Theorem 12 bounds the total number of inner iterations of the I-AL method when a summable sequence $\{\eta_k\}$ satisfying (40) is used. Before proving this theorem, we first state the following two technical results.

**Proposition 27** *For some positive integer $L$, let positive scalars $p_1, p_2, \cdots, p_L$ be given. Then, there exists a constant $C = C(p_1, \cdots, p_L)$ such that for any nonnegative scalars $\beta_0, \beta_1, \cdots, \beta_L$, $\nu$, and $\bar{t}$, we have*

$$\sum_{k=0}^{K}\left\lceil \beta_0 + \sum_{l=1}^{L}\left(\beta_l\, t_k^{p_l}\right)\right\rceil \max\left\{1, \left\lceil \log\frac{\nu}{t_k}\right\rceil\right\} \leq C\left\lceil \beta_0 + \sum_{l=1}^{L}\left(\beta_l\, \bar{t}^{p_l}\right)\right\rceil \max\left\{1, \left\lceil \log\frac{\nu}{\bar{t}}\right\rceil\right\}, \tag{87}$$

*where*

$$K := \max\left\{0, \left\lceil \log\left(\frac{\bar{t}}{t_0}\right)\right\rceil\right\},\ t_0 := \min_{1\leq l\leq L}\left(\frac{\max(\beta_0,1)}{\beta_l}\right)^{1/p_l},\ t_k = t_0 2^k,\ \forall k = 1,\ldots,K. \tag{88}$$

*In particular, if $\nu = \bar{t}$, then (87) implies that*

$$\sum_{k=0}^{K}\left\lceil \beta_0 + \sum_{l=1}^{L}\left(\beta_l\, t_k^{p_l}\right)\right\rceil \leq C\left\lceil \beta_0 + \sum_{l=1}^{L}\left(\beta_l\, \bar{t}^{p_l}\right)\right\rceil. \tag{89}$$

*Proof.* See Proposition 24 in Lan and Monteiro [9]. ∎

**Lemma 28** *The following statements hold:*

*a) for every $t \geq 1$ and $a, b \geq 0$, we have $(a+b)^t \leq [(2a)^t + (2b)^t]/2$;*

*b) for any $K \geq 1$ and $\xi > 0$, we have*

$$\sum_{k=0}^{+\infty}(k+1)^{-(1+\xi)} \leq 1 + \int_0^{+\infty}(t+1)^{-(1+\xi)}dt \leq (\xi+1)/\xi, \tag{90}$$

$$\sum_{k=0}^{K-1}(k+1)^{\xi} \leq \int_0^{K}(t+1)^{\xi}dt \leq \frac{1}{1+\xi}(K+1)^{1+\xi}. \tag{91}$$

*Proof.* Statement a) follows directly from the convexity of $x^t$ for any $x \geq 0$ and b) is obvious. ■

We are now ready to prove Theorem 12.

**Proof of Theorem 12:** We first show that condition (40) holds. Indeed, by (44), (90) and the assumption that $\bar{N} = +\infty$, we have

$$\sum_{k=0}^{\bar{N}-1} \eta_k = \frac{\xi \rho \epsilon_p^2}{128(\xi+1)} \sum_{k=0}^{\infty} \frac{1}{(k+1)^{1+\xi}} \leq \frac{\rho \epsilon_p^2}{128}.$$

It then follows from Proposition 11 that the method will successfully terminate in $N$ outer iterations and the total number of inner iterations is bounded by $\mathcal{I}_p + \mathcal{I}_d$, where $N$, $\mathcal{I}_p$ and $\mathcal{I}_d$ are defined in (39) and (42). Observe that by (44), (91), (39) and Lemma 28(a) with $a = 16D_\Lambda^2/(\rho^2\epsilon_p^2)$, $b = 2$ and $t = (3+\xi)/2$, we have

$$\begin{aligned}
\sum_{k=0}^{N-1} \eta_k^{-\frac{1}{2}} &= \frac{8\sqrt{2}(1+\xi)^{\frac{1}{2}}}{\xi^{\frac{1}{2}}\rho^{\frac{1}{2}}\epsilon_p} \sum_{k=0}^{N-1} (k+1)^{\frac{1+\xi}{2}} \leq \frac{16\sqrt{2}(1+\xi)^{\frac{1}{2}}}{(3+\xi)\xi^{\frac{1}{2}}\rho^{\frac{1}{2}}\epsilon_p} (N+1)^{\frac{3+\xi}{2}} \\
&\leq \frac{16\sqrt{2}(1+\xi)^{\frac{1}{2}}}{(3+\xi)\xi^{\frac{1}{2}}\rho^{\frac{1}{2}}\epsilon_p} \left(\frac{16D_\Lambda^2}{\rho^2\epsilon_p^2} + 2\right)^{\frac{3+\xi}{2}} = \frac{64C(\xi)}{\rho^{\frac{1}{2}}\epsilon_p} \left(\frac{8D_\Lambda^2}{\rho^2\epsilon_p^2} + 1\right)^{\frac{3+\xi}{2}} \\
&\leq \frac{32C(\xi)}{\rho^{\frac{1}{2}}\epsilon_p} \left[\left(\frac{4D_\Lambda}{\rho\epsilon_p}\right)^{3+\xi} + 2^{\frac{3+\xi}{2}}\right],
\end{aligned}$$

where $C(\xi) := (1+\xi)^{\frac{1}{2}} 2^{\frac{\xi}{2}} / [(3+\xi)\xi^{\frac{1}{2}}]$. This relation together with (39) and (42) then imply that

$$\begin{aligned}
\mathcal{I}_p &\leq \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{N-1} \eta_k^{-\frac{1}{2}} + N \\
&\leq \frac{32\sqrt{2}C(\xi) D_X M_\rho^{\frac{1}{2}}}{\rho^{\frac{1}{2}}\epsilon_p} \left[\left(\frac{4D_\Lambda}{\rho\epsilon_p}\right)^{3+\xi} + 2^{\frac{3+\xi}{2}}\right] + \frac{16D_\Lambda^2}{\rho^2\epsilon_p^2} + 1.
\end{aligned}$$

Moreover, it can be easily seen from (42) that

$$\mathcal{I}_d \leq 4D_X \left\{ \frac{4M_\rho^{\frac{1}{2}}}{\rho^{\frac{1}{2}}\epsilon_p} + \frac{M_\rho}{\epsilon_d} \right\} + 1.$$

Combining the previous two inequalities, we immediately see that the the total number of inner iterations performed by the I-AL method is bounded by (45).

Assume now that $\rho$ is chosen as in (46). Then, bound (47) follows by combining the definition of $N$ in (39) with the fact that by (46),

$$\rho \geq \max\left\{ \frac{1}{\epsilon_p} \left( \frac{(D_\Lambda)^{3+\xi}\epsilon_d}{\|\mathcal{A}\|} \right)^{\frac{1}{4+\xi}}, \frac{L_f}{\|\mathcal{A}\|^2} \right\}. \tag{92}$$

Also, (92) implies that $\rho \geq L_f/\|\mathcal{A}\|^2$, and hence that

$$M_\rho = L_f + \rho\|\mathcal{A}\|^2 \leq 2\rho\|\mathcal{A}\|^2 = \frac{2\|\mathcal{A}\|^2}{\epsilon_p}\left(\frac{(D_\Lambda)^{3+\xi}\epsilon_d}{\|\mathcal{A}\|}\right)^{\frac{1}{4+\xi}} + 2L_f = 2\epsilon_d\left(\frac{\|A\|^{\frac{7+2\xi}{4+\xi}} D_\Lambda^{\frac{3+\xi}{4+\xi}}}{\epsilon_p \epsilon_d^{\frac{3+\xi}{4+\xi}}} + \frac{L_f}{\epsilon_d}\right). \quad (93)$$

Hence, bound (45) is majorized by

$$\mathcal{O}\left(\frac{D_X\|\mathcal{A}\|}{\epsilon_p}\left[\left(\frac{D_\Lambda}{\rho\epsilon_p}\right)^{3+\xi} + 1\right] + D_X\left(\frac{\|A\|^{\frac{7+2\xi}{4+\xi}} D_\Lambda^{\frac{3+\xi}{4+\xi}}}{\epsilon_p \epsilon_d^{\frac{3+\xi}{4+\xi}}} + \frac{L_f}{\epsilon_d}\right) + \frac{D_\Lambda^2}{\rho^2\epsilon_p^2} + 1\right). \quad (94)$$

Also, by (92), we have

$$\frac{D_\Lambda}{\rho\epsilon_p} \leq D_\Lambda\left(\frac{\|\mathcal{A}\|}{D_\Lambda^{3+\xi}\epsilon_d}\right)^{\frac{1}{4+\xi}} = \left(\frac{D_\Lambda\|\mathcal{A}\|}{\epsilon_d}\right)^{\frac{1}{4+\xi}}.$$

Substituting the above inequality into (94), we obtain bound (48). ∎

Theorem 13 provides a bound on the total number inner iterations of the I-AL method when a uniform sequence $\{\eta_k\}$ is used, under the assumption that an upper bound $t$ on $D_\Lambda$, is known. We will now provide a proof of Theorem 13.

**Proof of Theorem 13** Using (49) and the assumption that $t \geq D_\Lambda$, we obtain

$$\bar{N}(t) \geq \left\lceil \frac{16 D_\Lambda^2}{\rho^2\epsilon_p^2} \right\rceil = N. \quad (95)$$

Also note that (49) and (50) imply that

$$\sum_{k=0}^{\bar{N}-1} \eta_k = \bar{N}\eta(t) = \bar{N}(t)\eta(t) = \frac{\rho\epsilon_p^2}{128}.$$

We have thus shown that conditions (39) and (40) hold. It then follows from Proposition 11 that the total number of outer iterations is bounded by $N$, where $N$ is defined by (39). Bound (51) now follows by combining the definition of $N$ in (39) with the fact that

$$\rho = \rho(t) \geq \max\left\{\frac{4t^{\frac{3}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}}\epsilon_p}, \frac{L_f}{\|\mathcal{A}\|^2}\right\} \geq \max\left\{\frac{4D_\Lambda^{\frac{3}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}}\epsilon_p}, \frac{L_f}{\|\mathcal{A}\|^2}\right\}. \quad (96)$$

It also follows from Proposition 11 that the total number of inner iterations is bounded by $\mathcal{I}_p + \mathcal{I}_d$, where $\mathcal{I}_p$ and $\mathcal{I}_d$ are given by (42). Noting that by (49), (50) and Lemma 28(a) with $a = 16t^2/(\rho^2\epsilon_p^2)$, $b = 1$ and $t = 3/2$, we have

$$\sum_{k=0}^{\bar{N}(t)-1} \eta_k^{-\frac{1}{2}} = \frac{8\sqrt{2}}{\rho(t)^{\frac{1}{2}}\epsilon_p}\bar{N}(t)^{\frac{3}{2}} \leq \frac{8\sqrt{2}}{\rho(t)^{\frac{1}{2}}\epsilon_p}\left(\frac{16t^2}{\rho(t)^2\epsilon_p^2} + 1\right)^{\frac{3}{2}} \leq \frac{16}{\rho(t)^{\frac{1}{2}}\epsilon_p}\left(\frac{64t^3}{\rho(t)^3\epsilon_p^3} + 1\right),$$

we then conclude from (42), (49) and (95) that

$$\mathcal{I}_p \leq \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{\bar{N}(t)-1} \eta_k^{-\frac{1}{2}} + \bar{N}(t) \leq \frac{16\sqrt{2} D_X M_\rho^{\frac{1}{2}}}{\rho(t)^{\frac{1}{2}} \epsilon_p} \left( \frac{64 t^3}{\rho(t)^3 \epsilon_p^3} + 1 \right) + \frac{16 t^2}{\rho(t)^2 \epsilon_p^2} + 1. \tag{97}$$

Now, by using the first relation in (49), we have that $\rho(t) \geq L_f / \|\mathcal{A}\|^2$, and hence that

$$M_\rho = L_f + \rho(t) \|\mathcal{A}\|^2 \leq 2 \rho(t) \|\mathcal{A}\|^2. \tag{98}$$

This conclusion together with (96) and (97) then imply that

$$\begin{aligned} \mathcal{I}_p &\leq \frac{32 D_X \|\mathcal{A}\|}{\epsilon_p} \left( \frac{64 t^3}{\rho(t)^3 \epsilon_p^3} + 1 \right) + \frac{16 t^2}{\rho(t)^2 \epsilon_p^2} + 1 \\ &\leq \frac{32 D_X \|\mathcal{A}\|}{\epsilon_p} \left( \frac{\|\mathcal{A}\|^{\frac{3}{4}} t^{\frac{3}{4}}}{\epsilon_d^{\frac{3}{4}}} + 1 \right) + \frac{\|\mathcal{A}\|^{\frac{1}{2}} t^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}} + 1. \end{aligned} \tag{99}$$

Moreover, it easily follows from (42), (98) and (49) that

$$\begin{aligned} \mathcal{I}_d &\leq 4 D_X \left( \frac{4 M_\rho^{\frac{1}{2}}}{\rho(t)^{\frac{1}{2}} \epsilon_p} + \frac{M_\rho}{\epsilon_d} \right) + 1 \leq 4 D_X \left( \frac{4\sqrt{2} \|\mathcal{A}\|}{\epsilon_p} + \frac{2 \rho(t) \|\mathcal{A}\|^2}{\epsilon_d} \right) + 1 \\ &= \frac{16\sqrt{2} D_X \|\mathcal{A}\|}{\epsilon_p} + 8 D_X \left( \frac{4 t^{\frac{3}{4}} \|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}} + \frac{L_f}{\epsilon_d} \right) + 1. \end{aligned} \tag{100}$$

Combining (99) and (100), we easily see that the I-AL method computes an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1) in at most $\mathcal{O}(\mathcal{I}_{pd}(t))$ inner iterations, where $\mathcal{I}_{pd}(t)$ is defined by (52). ∎

We now give the proof of Theorem 14, which establishes the iteration-complexity of the I-AL guess-and-check procedure.

**Proof of Theorem 14** Suppose that the I-AL guess-and-check procedure terminates when the iteration count $j$ is equal to $J$. Letting

$$\bar{J} := \max\{0, \lceil \log(D_\Lambda / t_0) \rceil\} \tag{101}$$

and noting that $t_{\bar{J}} = t_0 2^{\bar{J}} \geq D_\Lambda$, we conclude from Theorem 13 that $J \leq \bar{J}$. Let $\mathcal{I}_{p,j}$, $j = 1, \ldots, J$, denote the number of inner iterations performed at step 1) of the I-AL method during loop $j$ of the I-AL guess-and-check procedure, and let $\mathcal{I}_{d,J}$ denote the number of inner iterations performed by subroutine Postprocessing during loop $J$ of the I-AL guess-and-check procedure. Then, the overall number of inner iterations performed by the I-AL guess-and-check procedure is bounded by

$$\sum_{j=0}^{J} \mathcal{I}_{p,j} + \mathcal{I}_{d,J} \leq \sum_{j=0}^{\bar{J}} \mathcal{I}_{p,j} + \mathcal{I}_{d,J}. \tag{102}$$

Since the total number of outer iterations at the $j$th loop is bounded by $N(t_j)$, it follows from Corollary 2 that

$$\mathcal{I}_{p,j} \leq \sum_{k=0}^{\bar{N}(t_j)-1} \left\lceil D_X \sqrt{\frac{2M_\rho}{\eta_k}} \right\rceil \leq \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{\bar{N}(t_j)-1} \eta_k^{-\frac{1}{2}} + \bar{N}(t_j).$$

Hence, similar to the proof of (97), (98) and (99), we can show that for $j = 0, \ldots, J$, we have

$$\mathcal{I}_{p,j} \leq 32 D_X \left[ \frac{t_j^{\frac{3}{4}} \|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}} + \frac{\|\mathcal{A}\|}{\epsilon_p} \right] + \frac{t_j^{\frac{1}{2}} \|\mathcal{A}\|^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}} + 1 \leq \left\lceil \beta_0 + \beta_1 t_j^{\frac{3}{4}} + \beta_2 t_j^{\frac{1}{2}} \right\rceil,$$

where $\beta_0$, $\beta_1$, and $\beta_2$ are given by (53). Noting that $t_j = t_0 2^j$ for every $j$ and the definition of $t_0$ in step 0) of the I-AL guess-and-check procedure, it follows from the previous inequality and relation (89) with $L = 2$, $p_1 = 3/4$, $p_2 = 1/2$, $\bar{t} = D_\Lambda$, $J = \bar{J}$, and $\beta_0$, $\beta_1$, and $\beta_2$ as above that

$$\sum_{j=0}^{\bar{J}} \mathcal{I}_{p,j} = \mathcal{O}(1) \left\lceil \beta_0 + \beta_1 D_\Lambda^{\frac{3}{4}} + \beta_2 D_\Lambda^{\frac{1}{2}} \right\rceil. \tag{103}$$

Now, using (101), it is easy to see that $t_J \leq t_{\bar{J}} \leq \max\{t_0, 2D_\Lambda\}$ and hence that

$$t_J^{\frac{3}{4}} \leq \max\left\{ t_0^{\frac{3}{4}}, (2D_\Lambda)^{\frac{3}{4}} \right\} \leq \max\left\{ \frac{\beta_0}{\beta_1}, (2D_\Lambda)^{\frac{3}{4}} \right\} \leq \frac{\beta_0}{\beta_1} + (2D_\Lambda)^{\frac{3}{4}}, \tag{104}$$

where the last inequality is due to the definition of $t_0$ in Step 0 of the I-AL guess-and-check procedure. Using this inequality, the definition of $\beta_0$ and $\beta_1$ in (53), and an argument similar to the proof of (100), we have

$$\begin{aligned} \mathcal{I}_{d,J} &\leq \frac{16\sqrt{2} D_X \|\mathcal{A}\|}{\epsilon_p} + 8 D_X \left[ \frac{4 t_J^{\frac{3}{4}} \|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}} + \frac{L_f}{\epsilon_d} \right] + 1 \\ &\leq \beta_0 + \beta_1 t_J^{\frac{3}{4}} + \frac{8 D_X L_f}{\epsilon_d} \leq 2\beta_0 + \beta_1 (2D_\Lambda)^{\frac{3}{4}} + \frac{8 D_X L_f}{\epsilon_d}. \end{aligned} \tag{105}$$

Now, using (103) and (105), it is easy to see that the right-high-side of (102) is bounded by $\mathcal{O}(\mathcal{I}_{pd}(D_\Lambda))$, where $\mathcal{I}_{pd}(\cdot)$ is defined in (52). ∎

## 5.2 Convergence analysis for the I-AL method applied to the perturbed problem

The goal of this subsection is to prove the convergence results stated in Subsection 3.4, namely, Proposition 16 and Theorems 18 and 19.

We first prove Proposition 16 which guarantees that subroutine Postprocessing of the modified I-AL method outputs an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).

**Proof of Proposition 16:** As in the proof of Proposition 8 with $\zeta$ replaced by $\tilde{\zeta}$, we can show that

$$\nabla f_\gamma(\tilde{x}^+) + (\mathcal{A}_0)^* \tilde{\lambda}^+ \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}\left(\frac{\epsilon_d}{2}\right),$$

where $\tilde{x}^+$ is defined in (37) with $\mathcal{L}_\rho$ replaced by $\mathcal{L}_{\rho,\gamma}$. Noting that

$$\nabla f_\gamma(\tilde{x}^+) = \nabla f(\tilde{x}^+) + \gamma(\tilde{x}^+ - x_0)$$

and that (43) and (60) imply that

$$\gamma\|\tilde{x}^+ - x_0\| \leq \gamma D_X = \frac{\epsilon_d}{2},$$

we then conclude that

$$\nabla f(\tilde{x}^+) + (\mathcal{A}_0)^*\tilde{\lambda}^+ \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d).$$

Moreover, similar to the proof of Proposition 8, we can show that $\|\mathcal{A}(\tilde{x}^+)\| \leq \epsilon_p$. Thus, $(\tilde{x}^+, \lambda_k^+)$ is an $(\epsilon_p, \epsilon_d)$-primal-dual solution for (1). ■

Theorems 18 provides a bound on the total number of inner iterations performed by the modified I-AL method. Before proving this result, we first present two technical lemmas. The first one stated below establishes an important technical result that allows us to take the advantage of the "warm-start" strategy described in the end of Subsection 3.3.

**Lemma 29** *Let $(x_k, \lambda_k) \in X \times \Re^m$ be given and let $\lambda_{k+1} = \lambda_k + \rho\mathcal{A}(x_k)$. If $\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_k) \leq \eta_k$, then*

$$\frac{\gamma}{2}\|x_k - x_{k+1}^*\|^2 \leq \mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - d_{\rho,\gamma}(\lambda_{k+1}) \leq \left(\sqrt{\eta_k} + \sqrt{\frac{\rho}{2}}\|\mathcal{A}(x_k)\|\right)^2, \tag{106}$$

*where $x_{k+1}^*$ is the unique solution of $\min_{x\in X}\mathcal{L}_{\rho,\gamma}(x, \lambda_{k+1})$.*

*Proof.* The first inequality in (106) follows immediately from the strong convexity of $\mathcal{L}_{\rho,\gamma}(\cdot, \lambda_{k+1})$. Hence, it suffices to show the second inequality in (106). Clearly, by definition (56) and the fact that $\lambda_{k+1} = \lambda_k + \rho\mathcal{A}(x_k)$, we have

$$\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - \mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) = \rho\|\mathcal{A}(x_k)\|^2.$$

The above observation together with the assumption $\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_k) \leq \eta_k$ then imply that

$$\begin{aligned}
\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - d_{\rho,\gamma}(\lambda_{k+1}) &= [\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - \mathcal{L}_{\rho,\gamma}(x_k, \lambda_k)] + [\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})] \\
&= \rho\|\mathcal{A}(x_k)\|^2 + [\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_k)] + [d_{\rho,\gamma}(\lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})] \\
&\leq \rho\|\mathcal{A}(x_k)\|^2 + \eta_k + [d_{\rho,\gamma}(\lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})].
\end{aligned} \tag{107}$$

Moreover, in view of Proposition 6 applied to the perturbed problem (54), the function $d_{\rho,\gamma}(\cdot)$ is concave and has $1/\rho$-Lipschitz-continuous gradient and $\nabla d_{\rho,\gamma}(\lambda) = u_{\lambda,\gamma}^*$. It then follows from (5) that

$$\begin{aligned}
-d_{\rho,\gamma}(\lambda_{k+1}) + d_{\rho,\gamma}(\lambda_k) &\leq \langle -u_{\lambda_k,\gamma}^*, \lambda_{k+1} - \lambda_k\rangle + \frac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\
&= -\rho\langle u_{\lambda_k,\gamma}^*, \mathcal{A}(x_k)\rangle + \frac{\rho}{2}\|\mathcal{A}(x_k)\|^2,
\end{aligned} \tag{108}$$

where the last equality follows from the fact that $\lambda_{k+1} - \lambda_k = \rho \mathcal{A}(x_k)$. Combining (107) and (108), we obtain

$$\begin{aligned}
\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - d_{\rho,\gamma}(\lambda_{k+1}) &\leq \eta_k + \rho \langle \mathcal{A}(x_k) - u^*_{\lambda_k,\gamma}, \mathcal{A}(x_k) \rangle + \frac{\rho}{2} \|\mathcal{A}(x_k)\|^2 \\
&\leq \eta_k + \rho \|\mathcal{A}(x_k) - u^*_{\lambda_k,\gamma}\| \|\mathcal{A}(x_k)\| + \frac{\rho}{2} \|\mathcal{A}(x_k)\|^2 \\
&\leq \eta_k + \sqrt{2\rho\,\eta_k}\, \|\mathcal{A}(x_k)\| + \frac{\rho}{2} \|\mathcal{A}(x_k)\|^2 = \left( \sqrt{\eta_k} + \sqrt{\frac{\rho}{2}} \|\mathcal{A}(x_k)\| \right)^2,
\end{aligned}$$

where the last inequality follows from Proposition 7 with $\mathcal{L}_\rho = \mathcal{L}_{\rho,\gamma}$, $d_\rho = d_{\rho,\gamma}$, and $u^*_{\lambda_k} = u^*_{\lambda_k,\gamma}$. ■

The following technical result states a bound on the number of inner iterations performed by the modified I-AL method applied to (54) when a constant sequence $\{\eta_k\}$ is applied.

**Lemma 30** *Let $\rho > 0$, $(\epsilon_p, \epsilon_d) \in \Re_{++} \times \Re_{++}$ and $\bar{N} \in \mathbb{N}$ be given, and let $\gamma$ be given by (60). Consider the modified I-AL method applied to the perturbed problem (54) with penalty parameter $\rho$, iteration limit $\bar{N}$ and inner tolerances $\eta_0, \ldots, \eta_{\bar{N}}$ given by*

$$\eta_k = \eta_\gamma := \frac{\rho \epsilon_p^2}{128 \bar{N}}, \quad k = 0, \ldots, \bar{N} - 1. \tag{109}$$

*Then the following statements hold:*

*a) the total number of inner iterations performed by the above method is bounded by*

$$\left\lceil \sqrt{\frac{8 M_{\rho,\gamma}}{\gamma}} \right\rceil \left\{ 2 \max\left(1, \left\lceil \log \frac{64 \gamma \bar{N} D_X^2}{\rho \epsilon_p^2} \right\rceil \right) + \min(\bar{N}, N_\gamma) \left\lceil 2 \log\left(1 + \frac{24 \bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil \right.$$
$$\left. + \max\left(1, \left\lceil \log \frac{16 \gamma M_{\rho,\gamma} D_X^2}{\epsilon_d^2} \right\rceil \right) \right\}, \tag{110}$$

*where*

$$N_\gamma := \left\lceil \frac{16 [D_\Lambda^\gamma]^2}{\rho^2 \epsilon_p^2} \right\rceil; \tag{111}$$

*b) if $\bar{N} \geq N_\gamma$, then the above method successfully terminates in $N_\gamma$ outer iterations with an $(\epsilon_p, \epsilon_d)$-primal-dual solution of (1).*

*Proof.* Statement b) immediately follows from the assumption $\bar{N} \geq N_\gamma$ and Theorem 10 applied to the perturbed problem (54). We now show part a). Note that by Statement b), the number of outer iterations of the above method is bounded by $\min\{\bar{N}, N_\gamma\}$. Assume that the method terminates at the $K$-th outer iteration for some

$$0 \leq K \leq \min\{\bar{N}, N_\gamma\} - 1. \tag{112}$$

Clearly, $\|\mathcal{A}(x_k)\| > 3\epsilon_p/4$ for all $0 \leq k \leq K - 1$. Hence, by using an argument similar to the one preceding (86), we can show that

$$\|\mathcal{A}(x_k)\|^2 \leq \frac{9 [D_\Lambda^\gamma]^2}{\rho^2 (k+1)}, \quad k = 1, \ldots, K - 1. \tag{113}$$

For $k = 0, \ldots, K$, let $x_k^* := \operatorname{argmin}_{x \in X} \mathcal{L}_{\rho,\gamma}(x, \lambda_k)$, and $l_k$ denote the number of inner iterations performed at step 1 of the modified I-AL method. By Theorem 3 with $\phi(\cdot) = \mathcal{L}_{\rho,\gamma}(\cdot, \lambda_0)$, $L_\phi = M_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \eta_\gamma$, (43) and (109), we have

$$\begin{aligned} l_0 &\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma \|x_{-1} - x_0^*\|^2}{2\eta_\gamma} \right\rceil \right\} \leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma D_X^2}{2\eta_\gamma} \right\rceil \right\} \\ &= \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{64\gamma \bar{N} D_X^2}{\rho \epsilon_p^2} \right\rceil \right\}. \end{aligned} \tag{114}$$

It also follows from Theorem 3 that

$$l_k \leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma \|x_{k-1} - x_k^*\|^2}{2\eta_\gamma} \right\rceil \right\}, \quad \forall\, k = 1, \ldots, K.$$

Now by using (106) and (113), we have

$$\frac{\gamma \|x_{k-1} - x_k^*\|^2}{2} \leq \left( \sqrt{\eta_\gamma} + \sqrt{\frac{\rho}{2}} \|\mathcal{A}(x_{k-1})\| \right)^2 \leq \left( \sqrt{\eta_\gamma} + \frac{3D_\Lambda^\gamma}{\sqrt{2\rho k}} \right)^2.$$

We then conclude from the previous two observations and (109) that

$$\begin{aligned} l_k &\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil 2\log\left(1 + \frac{3D_\Lambda^\gamma}{\sqrt{2\rho k \eta_\gamma}}\right) \right\rceil \right\} \\ &= \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil 2\log\left(1 + \frac{3D_\Lambda^\gamma}{\sqrt{2\rho k \eta_\gamma}}\right) \right\rceil \leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil 2\log\left(1 + \frac{3D_\Lambda^\gamma}{\sqrt{2\rho \eta_\gamma}}\right) \right\rceil \\ &= \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil 2\log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil, \quad \forall\, k = 1, \ldots, K. \end{aligned}$$

The above conclusion together with (110) and (114) then clearly imply that the total number of inner iterations performed at step 1) of the modified I-AL method is bounded by

$$\begin{aligned} l_0 + \sum_{k=1}^K l_k &\leq l_0 + K \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil 2\log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil \\ &\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\{ 2\max\left(1, \left\lceil \log \frac{64\gamma \bar{N} D_X^2}{\rho \epsilon_p^2} \right\rceil\right) + K \left\lceil \log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil \right\}. \end{aligned} \tag{115}$$

Moreover, let $\tilde{l}_K$ denote the number of inner iterations performed by subroutine PostProcessing. By using Theorem 3 with $\phi(\cdot) = \mathcal{L}_{\rho,\gamma}(\cdot, \lambda_K)$, $L_\phi = M_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \tilde{\zeta}$ and (59), we have

$$\begin{aligned} \tilde{l}_K &\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma D_X^2}{2\tilde{\zeta}} \right\rceil \right\} \\ &\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left[ \max\left\{1, \left\lceil \log \frac{64\gamma D_X^2}{\rho \epsilon_p^2} \right\rceil \right\} + \max\left\{1, \left\lceil \log \frac{16\gamma M_{\rho,\gamma} D_X^2}{\epsilon_d^2} \right\rceil \right\} \right]. \end{aligned} \tag{116}$$

Combining inequalities (112), (115) and (116), we can easily see that the total number of inner iterations performed by the modified I-AL method is bounded by (110). ∎

We are now ready to prove Theorem 18.

**Proof of Theorem 18:** We first show part a). It immediately follows from Lemma 30(a) that the total number of inner iterations performed by the modified I-AL method is bounded by (110) with $\bar{N} = \bar{N}_\gamma(t)$ and $\rho = \rho_\gamma(t)$. Note that by (61), (62), (123) and the fact that, by (63) and (64), $\log \mathcal{T}(t) \geq 2$, we have

$$\bar{N}_\gamma(t) \leq \frac{16t^2}{\rho_\gamma(t)^2 \epsilon_p^2} + 1 \leq \log \mathcal{T}(t) + 1 \leq 2 \log \mathcal{T}(t). \tag{117}$$

Also, using definitions (58) and (61), we have that

$$\gamma \leq M_{\rho,\gamma} = L_f + \gamma + \rho \|\mathcal{A}\|^2 \leq 2\rho \|\mathcal{A}\|^2. \tag{118}$$

This observation together with (60) and (61) then imply that

$$\begin{aligned}
\left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil &\leq \left\lceil 4\sqrt{\frac{\rho\|\mathcal{A}\|^2}{\gamma}} \right\rceil = \left\lceil 4 \left( \frac{4t\|\mathcal{A}\|^2}{\gamma \epsilon_p (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f}{\gamma} + 1 \right)^{\frac{1}{2}} \right\rceil \\
&\leq 4 \left( \frac{4D_X t \|\mathcal{A}\|^2}{\epsilon_p \epsilon_d (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{D_X L_f}{\epsilon_d} + 1 \right)^{\frac{1}{2}} + 1 \\
&\leq 8\sqrt{\frac{D_X t \|\mathcal{A}\|^2}{\epsilon_p \epsilon_d}} (\log \mathcal{T}(t))^{-\frac{1}{4}} + 4\sqrt{\frac{D_X L_f}{\epsilon_d}} + 5.
\end{aligned} \tag{119}$$

Observe that, by (117), (118), (63) and (64),

$$\begin{aligned}
\log \frac{64\gamma D_X^2 \bar{N}_\gamma(t)}{\rho_\gamma(t)\epsilon_p^2} &\leq \log \frac{128\gamma D_X^2 \log \mathcal{T}(t)}{\rho_\gamma(t)\epsilon_p^2} \leq \log \frac{256\|\mathcal{A}\|^2 D_X^2 \log \mathcal{T}(t)}{\epsilon_p^2} \\
&= 8 + 4\log \left( \frac{\|\mathcal{A}\| D_X}{\epsilon_p} \right)^{\frac{1}{2}} + \log\log \mathcal{T}(t) = \mathcal{O}\left(\log \mathcal{T}(t)\right),
\end{aligned} \tag{120}$$

and that, by (62), the fact that $\log x \leq x$, and (117),

$$\begin{aligned}
\min(\bar{N}_\gamma(t), N_\gamma) \left\lceil 2\log\left(1 + \frac{24 D_\Lambda^\gamma [\bar{N}_\gamma(t)]^{\frac{1}{2}}}{\rho_\gamma(t)\epsilon_p}\right) \right\rceil &\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{6D_\Lambda^\gamma}{t} \left[\log \bar{N}_\gamma(t)\right]^{\frac{1}{2}} [\bar{N}_\gamma(t)]^{\frac{1}{2}}\right) \right\rceil \\
&\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{6D_\Lambda^\gamma}{t} \bar{N}_\gamma(t)\right) \right\rceil \\
&\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{12 D_\Lambda^\gamma}{t} \log \mathcal{T}(t)\right) \right\rceil \\
&= \mathcal{O}\left\{ \log \mathcal{T}(t) \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t)}{t}\right) \right\}.
\end{aligned} \tag{121}$$

It also follows from (118), (60), (61), (62), (63) and (64) that

$$\begin{aligned}
\log \frac{16\gamma M_{\rho,\gamma} D_X^2}{\epsilon_d^2} &\leq \log \frac{16 M_{\rho,\gamma}^2 D_X^2}{\epsilon_d^2} \leq \log \left( \frac{8\rho \|\mathcal{A}\|^2 D_X}{\epsilon_d} \right)^2 \\
&\leq 2 \log \left[ \frac{8\|\mathcal{A}\|^2 D_X}{\epsilon_d} \left( \frac{4t}{\epsilon_p (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f + \gamma}{\|\mathcal{A}\|^2} \right) \right] \\
&= 2 \log \left[ \frac{8\|\mathcal{A}\|^2 D_X}{\epsilon_d} \left( \frac{4t}{\epsilon_p (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f}{\|\mathcal{A}\|^2} + \frac{\epsilon_d}{2 D_X \|\mathcal{A}\|^2} \right) \right] \\
&\leq 2 \log \left[ 8 D_X \left( \frac{4t\|\mathcal{A}\|^2}{\epsilon_p \epsilon_d} + \frac{L_f}{\epsilon_d} \right) + 4 \right] = \mathcal{O}(\log \mathcal{T}(t)). \qquad (122)
\end{aligned}$$

Now substituting bounds (119), (120), (121), and (122) into bound (110), we obtain bound (65). Statement b) follows immediately from Lemma 30(b) and the fact that, by (62), the assumption $t \geq D_\Lambda^\gamma$ and (117),

$$N_\gamma = \left\lceil \frac{16[D_\Lambda^\gamma]^2}{\rho_\gamma(t)^2 \epsilon_p^2} \right\rceil \leq \left\lceil \frac{16 t^2}{\rho_\gamma(t)^2 \epsilon_p^2} \right\rceil = \bar{N}_\gamma(t) \leq 2 \log \mathcal{T}(t). \qquad (123)$$

■

Before proving Theorem 19, we first state two technical results that summarize some properties of the function $\psi$ defined in (66).

**Lemma 31** *Let $\psi(t)$ and $\hat{t}$ be defined in (66). Then, the following statements hold:*

*a) $\psi(t)$ is continuous and non-decreasing for $t \geq 0$;*

*b) $\psi(0) \leq 0$ and $\psi(\hat{t}) \geq 0$.*

*Proof.* Statement a) immediately following from the fact that, by (66),

$$\begin{aligned}
\psi'(t) &= \mathcal{S}_1 \left\{ 1 - \frac{\mathcal{S}_2}{4(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3)} \left[ \log(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \right]^{-\frac{3}{4}} \right\} \frac{1}{2\sqrt{t}} \\
&\geq \mathcal{S}_1 (1 - 1/4) \frac{1}{2\sqrt{t}} \geq \frac{3\mathcal{S}_1}{8\sqrt{t}} \geq 0, \quad \forall\, t > 0,
\end{aligned}$$

where in the first inequality we use the fact that $\log(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \geq 2$ in view of (64). It can be easily seen from (66) that $\psi(0) \leq 0$. Noting that, by the definition of $\hat{t}$ in (66),

$$\mathcal{S}_1^2 \hat{t} - \mathcal{S}_2^2 (\mathcal{S}_1 \hat{t}^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) = \mathcal{S}_1^2 \hat{t} - \mathcal{S}_1 \mathcal{S}_2^2 \hat{t}^{\frac{1}{2}} - \mathcal{S}_2^2 (\mathcal{S}_2 + \mathcal{S}_3) = 0,$$

we conclude from (66) and the fact that $\log \tau \leq \tau \leq \tau^2$ for $\tau \geq 1$ that

$$\psi(\hat{t}) = \mathcal{S}_1 \hat{t}^{\frac{1}{2}} - \mathcal{S}_2 \left[ \log(\mathcal{S}_1 \hat{t}^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \right]^{\frac{1}{4}} \geq \mathcal{S}_1 \hat{t}^{\frac{1}{2}} - \mathcal{S}_2 (\mathcal{S}_1 \hat{t}^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3)^{\frac{1}{2}} = 0.$$

We have thus shown that b) holds. ■

**Lemma 32** *Let $\psi(t)$ and $\hat{t}$ be defined in (66). Then, there exists $t_0 \in [0, \hat{t}]$ such that $0 \leq \psi(t_0) \leq 1$. Moreover, we have*

$$\begin{aligned} \mathcal{S}_1 t_0^{\frac{1}{2}} &\leq \mathcal{S}_2 \left[\log \mathcal{T}(t_0)\right]^{\frac{1}{4}} + 1, & (124) \\ \mathcal{S}_1 t^{\frac{1}{2}} &\geq \mathcal{S}_2 [\log \mathcal{T}(t)]^{\frac{1}{4}}, \quad \forall\, t \geq t_0, & (125) \\ \log \mathcal{T}(t_0) &= \mathcal{O}\left(\log \mathcal{T}(0)\right), & (126) \end{aligned}$$

*where $\mathcal{T}(\cdot)$, $\mathcal{S}_1$ and $\mathcal{S}_2$ are defined in (63) and (64).*

*Proof.* The existence of $t_0 \in [0, \hat{t}]$ satisfying $0 \leq \psi(t_0) \leq 1$ follows immediately from Lemma 31. Inequality (124) follows from (63), (66) and the fact $\psi(t_0) \leq 1$. Moreover, we conclude from (63), (66), the assumption $\psi(t_0) \geq 0$ and Lemma 31(a) that

$$\mathcal{S}_1 t^{\frac{1}{2}} - \mathcal{S}_2 \left[\log \mathcal{T}(t)\right]^{\frac{1}{4}} = \mathcal{S}_1 t^{\frac{1}{2}} - \mathcal{S}_2 \left[\log(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3)\right]^{\frac{1}{4}} = \psi(t) \geq \psi(t_0) \geq 0$$

for any $t \geq t_0$, and hence that (125) holds. Also note that by (63), (66) and the fact that $t_0 \leq \hat{t}$, we have

$$\log \mathcal{T}(t_0) = \log(\mathcal{S}_1 t_0^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \leq \log(\mathcal{S}_1 \hat{t}^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) = \mathcal{O}\left(\log(\mathcal{S}_2 + \mathcal{S}_3)\right) = \mathcal{O}\left(\log \mathcal{T}(0)\right).$$

■

We are now ready to prove Theorem 19.

**Proof of Theorem 19:** Consider parameter $t_0$ computed in step 0 of the modified I-AL guess-and-check procedure. Assume first that $t_0 \geq D_\Lambda^\gamma$. Using this assumption, Theorem 18, relations (124) and (126), and the fact that, by (63) and (64), $\mathcal{T}(t) \geq 4$ for every $t \geq 0$, we conclude that the modified I-AL guess-and-check procedure will successfully terminate after the first loop and that the total number of inner iterations is bounded by

$$\begin{aligned} & \mathcal{O}\left\{\left(\mathcal{S}_1 t_0^{\frac{1}{2}} + \mathcal{S}_2 \left[\log \mathcal{T}(t_0)\right]^{\frac{1}{4}}\right) \left[\log \mathcal{T}(t_0)\right]^{\frac{3}{4}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t_0)}{t_0}\right)\right\} \\ &= \mathcal{O}\left\{\left(\mathcal{S}_1 t_0^{\frac{1}{2}} + \mathcal{S}_2 \left[\log \mathcal{T}(t_0)\right]^{\frac{1}{4}}\right) \left[\log \mathcal{T}(t_0)\right]^{\frac{3}{4}} \log\log \mathcal{T}(t_0)\right\} \\ &= \mathcal{O}\left\{\mathcal{S}_2 \log \mathcal{T}(t_0) \log\log \mathcal{T}(t_0)\right\} = \mathcal{O}\left\{\mathcal{S}_2 \log \mathcal{T}(0) \log\log \mathcal{T}(0)\right\}, \end{aligned}$$

which is clearly bounded by (67).

Now assume that $t_0 < D_\Lambda^\gamma$. Suppose that the modified I-AL guess-and-check procedure terminates when the iteration count $j$ is equal to $J$. Let

$$\bar{J} := \max\{0, \lceil \log(D_\Lambda^\gamma / t_0) \rceil\} \tag{127}$$

and note that

$$2 D_\Lambda^\gamma \geq t_{\bar{J}} := t_0 2^{\bar{J}} \geq D_\Lambda^\gamma. \tag{128}$$

Theorem 18(b) and the second inequality in (128) then imply that $J \leq \bar{J}$. Also observe that, by relation (87) with $L = 1$, $p_1 = 1/2$, $\bar{t} = D_\Lambda^\gamma$, $K = \bar{J}$, $\nu = D_\Lambda^\gamma \log \mathcal{T}(t_{\bar{J}})$, $\beta_0 = 0$ and $\beta_1 = 1/\sqrt{t_0}$, we

have

$$\begin{aligned}
\sum_{j=0}^{\bar{J}} t_j^{\frac{1}{2}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{t_j}\right) &\leq \sqrt{t_0} \sum_{j=0}^{\bar{J}} \left\lceil \frac{1}{\sqrt{t_0}} t_j^{\frac{1}{2}} \right\rceil \max\left(1, \left\lceil \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{t_j} \right\rceil\right) \\
&= \mathcal{O}\left\{ \sqrt{t_0} \left\lceil \frac{1}{\sqrt{t_0}} [D_\Lambda^\gamma]^{\frac{1}{2}} \right\rceil \max\left(1, \left\lceil \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{D_\Lambda^\gamma} \right\rceil\right) \right\} \\
&= \mathcal{O}\left\{ \left( [D_\Lambda^\gamma]^{\frac{1}{2}} + \sqrt{t_0} \right) \max\left(1, \lceil \log \log \mathcal{T}(2D_\Lambda^\gamma) \rceil\right) \right\} \\
&= \mathcal{O}\left( [D_\Lambda^\gamma]^{\frac{1}{2}} \log \log \mathcal{T}(D_\Lambda^\gamma) \right), \qquad (129)
\end{aligned}$$

where the last identity follows from the facts that $t_0 \leq D_\Lambda^\gamma$ and $\log \mathcal{T}(D_\Lambda^\gamma) \geq 2$. Using the facts that $J \leq \bar{J}$ and the function $\mathcal{T}$ given by (63) is non-decreasing, Theorem 18(a), relations (125) and (129), and the simple observation that by (128), we have $t_0 \leq t_j \leq 2D_\Lambda^\gamma$ for every $j = 1, \ldots, \bar{J}$, we conclude that the total number of inner iterations performed by the modified I-AL guess-and-check procedure is bounded by

$$\begin{aligned}
&\mathcal{O}\left\{ \sum_{j=0}^{\bar{J}} \left[ \left( \mathcal{S}_1 t_j^{\frac{1}{2}} + \mathcal{S}_2 [\log \mathcal{T}(t_j)]^{\frac{1}{4}} \right) [\log \mathcal{T}(t_j)]^{\frac{3}{4}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t_j)}{t_j}\right) \right] \right\} \\
&= \mathcal{O}\left\{ \sum_{j=0}^{\bar{J}} \left[ \mathcal{S}_1 t_j^{\frac{1}{2}} [\log \mathcal{T}(t_j)]^{\frac{3}{4}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t_j)}{t_j}\right) \right] \right\} \\
&= \mathcal{O}\left\{ [\log \mathcal{T}(2D_\Lambda^\gamma)]^{\frac{3}{4}} \mathcal{S}_1 \sum_{j=0}^{\bar{J}} \left[ t_j^{\frac{1}{2}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{t_j}\right) \right] \right\} \\
&= \mathcal{O}\left\{ [\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \mathcal{S}_1 [D_\Lambda^\gamma]^{\frac{1}{2}} \log \log \mathcal{T}(D_\Lambda^\gamma) \right\},
\end{aligned}$$

which is clearly bounded by (67). ∎

# 6 Concluding remarks

In this section, we compare the results obtained in this paper for the inexact AL methods with another possible approach for solving variational inequalities (VI) studied in Nemirovski ([11]) for bounded sets, and Monteiro and Svaiter ([10]) for unbounded sets.

Given a closed convex set $\Omega \in \Re^p$ and a monotone continuous function $F : \Omega \to \Re^p$. The (monotone) VI problem with respect to the pair $(F, X)$, denoted by $VIP(F, \Omega)$, consists of finding $w^*$ such that

$$w^* \in \Omega, \qquad \langle w - w^*, F(w^*) \rangle \geq 0, \quad \forall w \in \Omega. \qquad (130)$$

It is well-known that, under the assumption that $F$ is monotone and continuous, (130) is equivalent to

$$w^* \in \Omega, \qquad \langle w - w^*, F(w) \rangle \geq 0, \quad \forall w \in \Omega.$$

Relaxing the above two conditions, we obtain the following two notions of approximate solutions of $VIP(F, \Omega)$.

**Definition 3** *A point $\bar{w} \in \Omega$ is a $(\varrho, \epsilon)$-strong (resp., $(\varrho, \epsilon)$-weak) solution of $VIP(F, \Omega)$ if there exists $r \in \Re^n$ such that $\|r\| \leq \varrho$ and, for every $w \in \Omega$, $\langle w - \bar{w}, F(\bar{w}) - r\rangle \geq -\epsilon$ (resp., $\langle w - \bar{w}, F(w) - r\rangle \geq -\epsilon$).*

It is well-known that the CP problem (1) is equivalent to solving the $VIP(F, \Omega)$, where $\Omega := X \times \Re^m$ and

$$F(w) = F(x, \lambda) := \begin{pmatrix} \nabla f(x) + \mathcal{A}_0^* \lambda \\ -\mathcal{A}(x) \end{pmatrix}.$$

Moreover, defining the norm on $\Re^n \times \Re^m$ as $\|w\| := (\|x\|^2 + \|\lambda\|^2)^{1/2}$, then it is easy to see that an $(\epsilon_p, \epsilon_d)$-primal-dual solution $(\bar{x}, \bar{\lambda})$ is a $(\varrho, 0)$-strong solution, where $\varrho = \max\{\epsilon_p, \epsilon_d\}$. Disregarding $L_f$, $\|\mathcal{A}\|$, $D_X$, $D_\Lambda$ and $D_\Lambda^\gamma$, it has been shown in Monteiro and Svaiter ([10]) that, given $(\varrho, \epsilon) \in \Re_{++} \times \Re_{++}$, a variant of the Korpelevich's method can find an $(\varrho, \epsilon)$-strong solution for $VIP(F, \Omega)$ in $\mathcal{O}(\varrho^{-2} + \epsilon^{-1})$. On the other hand, we show in this paper that a $(\varrho, 0)$-strong solution, and hence an approximate solution as above, can be found in

$$\mathcal{O}\left(\frac{1}{\varrho}(\log \varrho^{-1})^{3/4} \log\log \varrho^{-1}\right)$$

by applying the modified guess-and-check procedure in Subsection 3.2 with $\epsilon_p = \epsilon_d = \varrho/\sqrt{2}$. Hence, the complexity in this paper is better than the one in [10] by at least a factor of

$$\varrho(\log \varrho^{-1})^{3/4} \log\log \varrho^{-1}.$$

It should be noted that [10] also shows that an $(\varrho, \epsilon)$-weak solution for $VIP(F, \Omega)$ can be found in

$$\mathcal{O}(\varrho^{-1} + \epsilon^{-1}). \tag{131}$$

It would be interesting to see whether our analysis in this paper can be modified to the context of finding a weak solution of $VIP(F, \Omega)$ so as to obtain a better iteration-complexity bound than (131).

# Appendix

**Proof of Proposition 4:** Let $\{(b_k, r_k)\}$ be a sequence of epi$f$ converging to $(b, r)$ for $k \to +\infty$. It suffices to show that $v(b) \leq r$. First notice that the fact that $v(b_k) \leq r_k$ implies that there exists $x_k \in \mathcal{F}(b_k)$ such that $f(x_k) = v(b_k) \leq r_k$. Now we claim that the sequence $\{x_k\}$ is bounded. Hence, by using this claim, there exists an accumulation point $x$ of the sequence $\{x_k\}$ such that $x \in \mathcal{F}(b), f(x) \leq r$ as $k \to +\infty$, which clearly implies that $v(b) \leq r$. Now it remains to show that the sequence$\{x_k\}$ is bounded. Indeed, let $f'_\infty(\cdot)$ denote the recession function of $f$, and $\mathcal{F}(b)_\infty$ denote the recession cone of the set $\mathcal{F}(b)$. Also let $\phi_b(\cdot) := f(\cdot) + \mathrm{I}_{\mathcal{F}(b)}(\cdot)$, using the assumption that the set of optimal solutions for (1) is nonempty and bounded, we have

$$\{\phi_0\}'_\infty(d) = f'_\infty(d) + \mathrm{I}_{\mathcal{F}(0)_\infty} > 0$$

for all $d \neq 0$ (see Definitions 2.2.2 and 3.2.3, Remark 3.2.8 and Proposition 3.2.9 in [7]). It can also be easily seen that the recession cone $\mathcal{F}(b_k)_\infty \equiv \mathcal{F}(0)_\infty$. It then follows from the above two relations that $\{\phi_{b_k}\}'_\infty(d) > 0$ for all $d \neq 0$, which, by Remark 3.2.8 in [7], implies that $x_k \in \mathrm{Argmin}_x \phi_{b_k}(x)$ is bounded. ∎

## References


[1] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York, first edition, 1982.

[2] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, New York, second edition, 1984.

[3] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 95:329–357, 2003.

[4] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103:427–444, 2005.

[5] E.G. Golshtein and N.V. Tretyakov. *Modified Lagrangians and monotone maps in optimization*. Springer-Verlag, New York, USA, 1996.

[6] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization and Application*, 4:303–320, 1969.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization algorithms I*, volume 305 of *Comprehensive Study in Mathematics*. Springer-Verlag, New York, 1993.

[8] F. Jarre and F. Rendl. An augmented primal-dual method for linear conic programs. Manuscript, Institut fur Mathematik, Universit at Dusseldorf, Germany, Austria, April 2007.

[9] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. Manuscript, School of Industrial Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, June 2008.

[10] R.D.C. Monteiro and B.F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, March 2009.

[11] A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2004.

[12] Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. translated as Soviet Math. Docl.

[13] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[14] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[15] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer-Verlag, New York, USA, 1999.

[16] M.M.D. Powell. An efficient method for nonlinear constraints in minimization problems. In ed. R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, 1969.

[17] A. Ruszczynski. *Nonlinear Optimization.* Princeton University Press, first edition, 2006.

[18] X. Zhao, D. Sun, and K. Toh. A newton-cg augmented lagrangian method for semidefinite programming. Manuscript, National University of Singapore, Singapore, March 2008.