Math. Program., Ser. A (2016) 155:511–547
DOI 10.1007/s10107-015-0861-x

FULL LENGTH PAPER

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

Guanghui Lan · Renato D. C. Monteiro

Received: 19 December 2014 / Accepted: 9 January 2015 / Published online: 24 January 2015

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 possibly 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

Mathematics Subject Classification 90C25 · 90C06 · 90C22 · 49M37

The work has been partially supported by NSF Grant CCF-0808863, CMMI-1000347, CMMI-1254446, and DMS-1319050, and ONR Grant N00014-08-1-0033 and N00014-13-1-0036.

G. Lan (✉)
Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
e-mail: glan@ise.ufl.edu

R. D. C. Monteiro
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
e-mail: monteiro@isye.gatech.edu

1 Introduction

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

$$f^* := \min\{f(x) : \mathcal{A}(x) = 0,\ x \in X\}, \tag{1.1}$$

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 ([15,17]) 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 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 (1.1) 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 (1.1) 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 [11] two first-order methods for solving (1.1) 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 (1.1), 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 (1.1) 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 (1.1) involving only computation of projections onto the simple set X are obtained.

The augmented Lagrangian method, initially proposed by Hestenes [7] and Powell [21] in 1969, is currently regarded as an effective optimization method for solving large-scale nonlinear programming problems (see textbooks or monographs: [1,2,6,19,22]). More recently, it has been used by the convex programming (CP) community to develop specialized first-order methods for solving large-scale semi-definite programming problems (see, for example, Burer and Monteiro [3,4], Jarre and Rendl [8], Zhao et al. [23]), due to its lower iteration-cost compared to that of Newton-based interior-point methods. The augmented Lagrangian method applied to problem (1.1) consists of solving a sequence of subproblems of the form

$$d_\rho(\lambda_k) := \min_{x \in X} \left\{ \mathcal{L}_\rho(x, \lambda_k) := f(x) + \langle \lambda_k, \mathcal{A}(x) \rangle + \frac{\rho}{2} \|\mathcal{A}(x)\|^2 \right\}, \tag{1.2}$$

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

$$\lambda_{k+1} = \lambda_k + \rho \mathcal{A}(x_k^*), \tag{1.3}$$

where x_k^* is a solution of problem (1.2). Since in most cases (1.2) can only be solved approximately, x_k^* in (1.3) is replaced by an η_k-approximate solution of (1.2), 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 (1.2) 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_ρ defined in (1.2) at λ_k, and hence that (1.3) can be viewed as a steepest ascent iteration with stepsize ρ applied to the function d_ρ. 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 (1.2), 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 η_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 ρ increases, it is well-known that the iteration-complexity of approximately solving each subproblem (1.2) increases, while the number of dual iterations (1.3), i.e., the outer iterations, decreases. Ways of choosing the parameter ρ so as to balance these two opposing criterions are then proposed. More specifically, ρ 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 η_k and the optimal penalty parameter ρ requires knowledge of an upper bound t on $D_\Lambda := \min_{\lambda \in \Lambda^*} \|\lambda_0 - \lambda^*\|$, where Λ^* 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_Λ is not known a priori, we present a "guess-and-check" procedure which consists of guessing a sequence of estimates for D_Λ and applying the corresponding sequence of inexact AL methods (with pre-specified number of outer-iterations) to (1.1) 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_Λ 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 (1.1).

Our paper is organized as follows. Section 2 describes the assumptions imposed on (1.1), the definition of an approximate primal-dual solution of (1.1), and the properties of the augmented dual function of (1.1), including a key result about how to approximate its gradient. In Sect. 3, we discuss some basic technical results that will be used in our analysis. In particular, we review Nesterov's smooth first-order method

for solving a certain class of smooth CP problems in Sect. 3.1, present some technical results about the projected gradient in Sect. 3.2 and about the convergence behavior of the steepest descent method with inexact gradient in Sect. 3.3. In Sect. 4, we describe two inexact AL methods and corresponding guess-and-check procedures for solving (1.1) and establish their iteration-complexity results. More specifically, in Sect. 4.1, 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 (1.1), are discussed in Sect. 4.2. Finally, we give some concluding remarks in Sect. 5.

1.1 Notation and terminology

We denote the set of real numbers by $\mathbf{R}$. Also, $\mathbf{R}_+$ and $\mathbf{R}_{++}$ 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 $f : \mathcal{C} \subseteq \Re^p \to \mathbf{R}$ is said to have L-Lipschitz-continuous gradient with respect to $\|\cdot\|$ if it is differentiable and

$$\|\nabla f(\tilde{u}) - \nabla f(u)\| \leq L\|\tilde{u} - u\|, \quad \forall u, \tilde{u} \in \mathcal{C}. \tag{1.4}$$

It is well-known (see Theorem 2.1.5 of [16]) that, for every $u, \tilde{u} \in \mathcal{C}$, we have:

$$\frac{1}{2L}\|\nabla f(\tilde{u}) - \nabla f(u)\|^2 \leq f(\tilde{u}) - f(u) - \langle \nabla f(u), (\tilde{u} - u)\rangle \leq \frac{L}{2}\|\tilde{u} - u\|^2, \tag{1.5}$$

$$\frac{1}{L}\|\nabla f(\tilde{u}) - \nabla f(u)\|^2 \leq \langle \nabla f(\tilde{u}) - \nabla f(u), \tilde{u} - u\rangle \leq L\|\tilde{u} - u\|^2. \tag{1.6}$$

2 The problem of interest

In this section, we describe the CP problem and the basic approach we will study in this paper. More specifically, we discuss the assumptions and termination criterion about problem (1.1) in Sect. 2.1. We review the augmented dual function and discuss some of its properties in Sect. 2.2.

2.1 Assumptions and termination criterion

The problem of interest in this paper is the CP problem (1.1) 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 (1.1) are defined as

$$d(\lambda) := \min\{f(x) + \langle \lambda, \mathcal{A}(x)\rangle : x \in X\}, \quad \forall \lambda \in \Re^m, \tag{2.1}$$

$$v(u) := \min\{f(x) : \mathcal{A}(x) = u,\ x \in X\}, \quad \forall u \in \Re^m. \tag{2.2}$$

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 (1.1) is the problem

$$d^* := \max_{\lambda} d(\lambda). \tag{2.3}$$

In addition to the convexity assumptions we made about (1.1), we also make the following assumptions throughout this paper:

Assumption 1 (1) the set X is nonempty and bounded (and hence $f^* \in \mathbf{R}$);
(2) there exists a Lagrange multiplier for (1.1), i.e., a vector λ^* such that $f^* = d(\lambda^*)$.

Note that $x^* \in X$ is an optimal solution of (1.1) and $\lambda^* \in \Re^m$ is a Lagrange multiplier for (1.1) if, and only if, (x^*, λ^*) satisfies

$$\begin{aligned} &\mathcal{A}(x^*) = 0, \\ &\nabla f(x^*) + (\mathcal{A}_0)^* \lambda^* \in -\mathcal{N}_X(x^*), \end{aligned} \tag{2.4}$$

where $\mathcal{N}_X(x^*) := \{s \in \Re^n : \langle s, x - x^* \rangle \le 0,\ \forall x \in X\}$ denotes the normal cone of X at 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 $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, $(\tilde{x}, \tilde{\lambda}) \in X \times \Re^m$ is called an (ϵ_p, ϵ_d)-*primal-dual* solution of (1.1) if

$$\|\mathcal{A}(\tilde{x})\| \le \epsilon_p, \tag{2.5}$$

$$\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}) + \mathcal{B}(\epsilon_d), \tag{2.6}$$

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

The main goal of this paper is to study the iteration-complexity of the augmented Lagrangian method for computing an (ϵ_p, ϵ_d)-*primal-dual* solution of (1.1) defined above.

2.2 The augmented dual function

In this subsection, we review the definition of the augmented dual function associated with (1.1) and discuss some of its properties.

Given a penalty parameter $\rho > 0$, the augmented dual function $d_\rho : \Re^m \to \mathbf{R}$ associated with (1.1) is given by

$$d_\rho(\lambda) := \min_{x \in X} \left\{ \mathcal{L}_\rho(x, \lambda) := f(x) + \langle \lambda, \mathcal{A}(x)\rangle + \frac{\rho}{2} \|\mathcal{A}(x)\|^2 \right\}, \tag{2.7}$$

and the augmented dual with parameter ρ is defined as

$$\max_{\lambda \in \Re^m} d_\rho(\lambda). \tag{2.8}$$

An alternative characterization for the augmented dual function is given by

$$d_\rho(\lambda) = \min_u \left\{ v_\rho(u, \lambda) := v(u) + \langle \lambda, u \rangle + \frac{\rho}{2} \|u\|^2 \right\}, \tag{2.9}$$

where $v(\cdot)$ is the value function given by (2.2).

Lemma 1 *The following statements hold:*

(a) *problem* (2.9) *has an unique optimal solution* u_λ^*;
(b) *the set of optimal solutions of* (2.7) X_λ^* *is given by*

$$X_\lambda^* = \{x \in X : \mathcal{A}(x) = u_\lambda^* \text{ and } f(x) = v(u_\lambda^*)\}; \tag{2.10}$$

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

$$v_\rho(u, \lambda) - d_\rho(\lambda) \geq \frac{\rho}{2} \|u - u_\lambda^*\|^2, \quad \forall u \in \Re^m; \tag{2.11}$$

(d) *problem* (2.8) *has the same optimal value and set of optimal solutions as those of* (2.3).

Proof We first show (a). Observe that convexity of v and Assumption A.1 imply that the function $v_\rho(\cdot, \lambda)$ in (2.9) 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 ρ, that is,

$$v_\rho(\alpha u_1 + (1-\alpha)u_2, \lambda) \leq \alpha v_\rho(u_1, \lambda) + (1-\alpha) v_\rho(u_2, \lambda) - \frac{\rho}{2} \alpha(1-\alpha) \|u_1 - u_2\|^2, \tag{2.12}$$

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 (2.2), and the equivalence of problems (2.7) and (2.9). To show (c), we let $u_1 = u$ and $u_2 = u_\lambda^*$ in (2.12) to obtain

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

where the last inequality follows from the fact that u_λ^* is the optimal solution for problem (2.9). Letting α go to zero in the above inequality, and using the lower-semicontinuity of v_ρ and the fact that $d_\rho(\lambda) = v_\rho(u_\lambda^*, \lambda)$, we obtain (2.11). Statement (d) is a well-known. □

The following proposition summarizes some important properties of d_ρ.

Proposition 1 *For any $\rho > 0$, the function d_ρ is concave, differentiable, and*

$$\nabla d_\rho(\lambda) = u_\lambda^*, \quad \forall\, \lambda \in \Re^m, \tag{2.13}$$

where u_λ^* is the unique optimal solution of problem (2.9). Moreover, d_ρ 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 [17] applied to the maximization version of (2.9), i.e., the problem $\max_u\{-v_\rho(u, \lambda)\}$.

In view of Proposition 1 and Lemma 1(b), the exact version of the augmented Lagrangian method stated in Sect. 1 can be viewed as a version of the steepest ascent method applied to (2.8). Note that one possible drawback of the exact augmented Lagrangian method is that each iteration of this method requires the solution of problem (1.2) for computing the gradient $\nabla d_\rho(\lambda_k)$. Since in most applications, problem (1.2) 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 (1.2).

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

Proposition 2 *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*

$$\|\mathcal{A}(x) - \nabla d_\rho(\lambda)\| = \|\mathcal{A}(x) - u_\lambda^*\| \leq \sqrt{\frac{2\eta}{\rho}}, \tag{2.14}$$

where u_λ^* is the unique optimal solution of (2.9).

Proof Letting $u := \mathcal{A}(x)$ and observing that $f(x) \geq v(u)$ due to definition (2.2), we conclude that

$$\mathcal{L}_\rho(x, \lambda) = f(x) + \langle \lambda, u\rangle + \frac{\rho}{2}\|u\|^2 \geq v(u) + \langle \lambda, u\rangle + \frac{\rho}{2}\|u\|^2 = v_\rho(u, \lambda). \tag{2.15}$$

This inequality, relation (2.11), and the assumption that $\mathcal{L}_\rho(x, \lambda) - d_\rho(\lambda) \leq \eta$ then imply that

$$\mathcal{L}_\rho(x, \lambda) - d_\rho(\lambda) \geq v_\rho(u, \lambda) - d_\rho(\lambda) \geq \frac{\rho}{2}\|u - u_\lambda^*\|^2, \tag{2.16}$$

and hence that (2.14) holds. □

3 Nesterov's optimal method and basic tools

This section discusses some basic technical results that will be used in our analysis. It consists of three subsections. The first one reviews Nesterov's smooth first-order method for solving a certain class of smooth CP problems. The second one develops several technical results involving gradient mapping. The third 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.

3.1 Nesterov's optimal method

In this subsection, we review Nesterov's smooth first-order method. Since the variant of the AL method we consider in this paper uses Nesterov's method to solve the augmented Lagrangian subproblems (1.2), 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

$$\phi^* := \min_{x \in X} \phi(x), \tag{3.1}$$

where $X \subset \Re^n$ is a closed convex set and $\phi : X \to \mathbf{R}$ is a convex function that has L_ϕ-Lipschitz-continuous gradient over X with respect to an inner product norm $\|\cdot\|$ in $\Re^n$. Moreover, we assume that the optimal value ϕ^* of problem (3.1) is finite and that its set of optimal solutions is nonempty.

We are now ready to state Nesterov's smooth first-order method for solving (3.1). 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. We also use $l_\phi : \Re^n \times X \to \mathbf{R}$ to denote the "linear approximation" of ϕ defined as

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

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

$$x_{k+1}^{sd} = \operatorname{argmin} \left\{ l_\phi(x; x_k) + \frac{L_\phi}{2} \|x - x_k\|^2 : x \in X \right\}, \tag{3.2}$$

$$x_{k+1}^{ag} = \operatorname{argmin} \left\{ \frac{L_\phi}{2} \|x - x_0\|^2 + \sum_{i=0}^{k} \frac{i+1}{2} [l_\phi(x; x_i)] : x \in X \right\}. \tag{3.3}$$

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

end

The main convergence result established by Nesterov [17] regarding the above algorithm is summarized in the following theorem.

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

$$\phi(x_k^{sd}) - \phi^* \le \frac{2L_\phi \|x_0^{sd} - \bar{x}\|^2}{k(k+1)}, \quad \forall k \ge 1,$$

where $\bar{x}$ is an optimal solution of (3.1). As a consequence, given any $\epsilon > 0$, the number of iterations to find a point $x_k^{sd} \in X$ satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ can be bounded by

$$\left\lceil \|x_0^{sd} - \bar{x}\| \sqrt{\frac{2L_\phi}{\epsilon}} \right\rceil. \tag{3.4}$$

Moreover, the following result states the iteration complexity of Nesterov's method for solving strongly convex problems satisfying

$$\langle \nabla\phi(x) - \nabla\phi(\tilde{x}), x - \tilde{x} \rangle \geq \mu \|x - \tilde{x}\|^2 \quad \forall x, \tilde{x} \in X \tag{3.5}$$

for some $\mu > 0$ (see Theorem 2.2.2 of [16] and Theorem 8 of [11]).

Theorem 2 *Let $\epsilon > 0$ be given and suppose that (3.5) holds. Then, the variant where we restart Nesterov's optimal method every $K := \lceil \sqrt{8L_\phi/\mu} \rceil$ iterations finds a solution $\tilde{x} \in X$ satisfying $\phi(\tilde{x}) - \phi^* \leq \epsilon$ in no more than*

$$K \max\left\{1, \left\lceil \log \frac{\mu \|x_0^{sd} - \bar{x}\|^2}{2\epsilon} \right\rceil \right\}$$

iterations, where $\bar{x} := argmin_{x \in X} \phi(x)$.

3.2 Gradient mapping

In this subsection, we still consider the CP problem (3.1). It is well-known that $x^* \in X$ is an optimal solution of (3.1) if and only if $\nabla\phi(x^*) \in -\mathcal{N}_X(x^*)$. Moreover, this optimality condition is in turn related to the gradient mapping (or projected gradient) of the function ϕ over X defined as follows.

Definition 2 Given a fixed constant $\tau > 0$, we define the gradient mapping of ϕ at $\tilde{x} \in X$ with respect to X as (see, for example, [16])

$$\nabla\phi(\tilde{x})_X^\tau := \frac{1}{\tau}\left[\tilde{x} - \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))\right], \tag{3.6}$$

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

The following proposition (see Proposition 4 in [11] for the proof) relates the gradient mapping to the aforementioned optimality condition.

Proposition 3 *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 states some properties of the gradient mapping (see [14,16] and Lemma 5 of [11]).

Lemma 2 *Assume that* $x^* \in Argmin_{x\in X}\phi(x)$. *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) $\phi(\tilde{x}^+) - \phi(\tilde{x}) \le -\tau\|\nabla\phi(\tilde{x})]_X^\tau\|^2/2$ *for any* $\tau \le 1/L_\phi$;
(b) *for any* $x \in X$, *we have*

$$\phi(x) - \phi(x^*) \ge \frac{1}{2L_\phi}\|\nabla\phi(x)]_X^{1/L_\phi}\|^2. \tag{3.7}$$

3.3 Steepest descent method with inexact gradient

In this subsection, we consider the unconstrained problem

$$p^* := \min\{p(\lambda) : \lambda \in \Re^m\}, \tag{3.8}$$

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 Γ^* of (3.8) 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. The aim is to apply the results obtained here to $p = -d_\rho$ in (2.7) in order to prove the main convergence results of the augmented Lagrangian methods.

An iterate of the steepest descent method with inexact gradient for solving problem (3.8) consists of:

$$\lambda_{k+1} = \lambda_k - \frac{\alpha_k}{L_p}p'_k \tag{3.9}$$

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

$$\delta_k := p'_k - \nabla p(\lambda_k), \quad e_k := \frac{\|\delta_k\|}{\|p'_k\|}. \tag{3.10}$$

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

Proof Using the second inequality of (1.5) with $\lambda = \lambda_k$ and $\tilde{\lambda} = \lambda_{k+1}$, relations (3.9) and (3.10), and the Cauchy-Schwartz inequality, we conclude that

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

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

Lemma 4 *Assume that $e_k < 1$. Then, for every $\lambda^* \in \Lambda^*$, we have*

$$\alpha_k\beta_k\langle \nabla p(\lambda_k), \lambda_k - \lambda^* \rangle \leq \frac{L_p}{2}\left(\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right) + \alpha_k\langle \delta_k, \lambda^* - \lambda_k \rangle, \tag{3.11}$$

where

$$\beta_k := 1 - \alpha_k/[2(1 - e_k)^2]. \tag{3.12}$$

Proof First note that, by (3.10), we have

$$\|\nabla p(\lambda_k)\| = \|p'_k - \delta_k\| \geq \|p'_k\| - \|\delta_k\| = (1 - e_k)\|p'_k\|. \tag{3.13}$$

This inequality, the assumption that $e_k < 1$ and relations (3.9) and (3.10) 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 (1.6) and the fact that $\nabla p(\lambda^*) = 0$. Rearranging the later inequality and using the definition of β_k, we obtain (3.11). □

Lemma 5 *Assume that, for some constant $c_1 > 0$, we have*

$$e_k \leq 1 - \sqrt{\frac{\alpha_k + c_1}{2}}. \tag{3.14}$$

Then, for any $\lambda^* \in \Lambda^*$ and for all $k \geq 0$, we have

$$\alpha_k[p(\lambda_k) - p^*] \leq \frac{L_p}{c_1}\left[\left(1 + \frac{2\alpha_k e_k^2}{c_1}\right)\|\lambda_k - \lambda^*\|^2 - \|\lambda_{k+1} - \lambda^*\|^2\right]. \quad (3.15)$$

Proof By the Cauchy-Schwartz inequality and relations (3.10), (1.5), (3.11) and (3.13), we have

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

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

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

Relation (3.15) now follows from the above inequality by noting that (3.12) and (3.14) imply that

$$\beta_k \geq (1 - e_k)^2\beta_k = (1 - e_k)^2 - \frac{\alpha_k}{2} \geq \frac{c_1}{2} > 0 \quad (3.16)$$

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 3 *Assume that for some positive constants c_1, we have*

$$e_k \leq 1 - \sqrt{\frac{\alpha_k + c_1}{2}} \quad (3.17)$$

for every $k \geq 0$. Then, the sequence $\{\lambda_k\}$ generated by the inexact steepest descent method (3.9) satisfies

$$p(\lambda_k) - p^* \leq \frac{L_p}{c_1 \sum_{i=0}^{k} \alpha_i} \left[\|\lambda_0 - \lambda^*\|^2 \exp\left(\sum_{i=0}^{k} \frac{2\alpha_i e_i^2}{c_1} \right) - \|\lambda_{k+1} - \lambda^*\|^2 \right] \tag{3.18}$$

for every $\lambda^* \in \Lambda^*$, where p^* is defined in (3.9).

Proof By Lemma 5 we have $\|\lambda_{k+1} - \lambda^*\|^2 \leq \|\lambda_0 - \lambda^*\|^2 \prod_{i=0}^{k} (1 + 2\alpha_i e_i^2 / c_1)$. Using this observation, (3.15), and an inductive argument, we can show that

$$\sum_{i=0}^{k} \alpha_i [p(\lambda_i) - p^*] \leq \frac{L_p}{c_1} \left[\|\lambda_0 - \lambda^*\|^2 \prod_{i=0}^{k} \left(1 + \frac{2\alpha_i e_i^2}{c_1} \right) - \|\lambda_{k+1} - \lambda^*\|^2 \right] \tag{3.19}$$

for every $k \geq 0$. The above inequality, Lemmas 3 and 5, the inequality $\log(1 + x) \leq x$ for any $x > -1$, and assumption (3.17) 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 3, we obtain the following result which gives an upper bound on the quantities $\|\nabla p(\lambda_k)\|$ and $\|p'(\lambda_k)\|$.

Corollary 1 *Assume that, for some positive constant c_1, relation (3.17) holds for every $k \geq 0$. Then, the sequence $\{\lambda_k\}$ generated by the inexact steepest descent method (3.9) satisfies*

$$\frac{\alpha_k + c_1}{2} \|p_k'\|^2 \leq \|\nabla p(\lambda_k)\|^2 \leq \frac{2L_p^2 \|\lambda_0 - \lambda^*\|^2}{c_1 \sum_{i=0}^{k} \alpha_i} \exp\left(\sum_{i=0}^{k} \frac{2\alpha_i e_i^2}{c_1} \right) \tag{3.20}$$

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 (3.17), imply that $\|\nabla p(\lambda_k)\|^2 \geq (\alpha_k+c_1)\|p'_k\|^2/2$. Moreover, using (1.5), (3.18), 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. □

4 The augmented Lagrangian methods

In this section, we present the augmented Lagrangian methods applied to (1.1) and discuss their computational complexity. Specifically, in Sect. 4.1, 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 (1.1), is discussed and analyzed in Sect. 4.2.

4.1 The inexact augmented Lagrangian (I-AL) method

In this subsection, we present the I-AL method applied to problem (1.1) and discuss its convergence behavior. We start by stating this algorithm as follows.

The I-AL method:

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

$$0 < \eta_k \leq \frac{\rho\epsilon_p^2}{128}, \quad \forall\, k = 0, \ldots, \bar{N}. \tag{4.1}$$

(0) Set $k = 0$;

(1) Using x_{k-1} as starting point, apply Nesterov's optimal method to find an η_k-approximate solution of problem (1.2), i.e., a point $x_k \in X$ such that

$$\mathcal{L}_\rho(x_k, \lambda_k) - d_\rho(\lambda_k) \leq \eta_k; \tag{4.2}$$

(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

$$\zeta = \zeta(\rho) := \min\left\{\frac{\rho\epsilon_p^2}{128}, \frac{\epsilon_d^2}{8M_\rho}\right\}, \tag{4.3}$$

where

$$M_\rho := L_f + \rho\|\mathcal{A}\|^2, \tag{4.4}$$

(P.1) Using $x \in X$ as starting point, apply Nesterov's optimal method to find an approximate solution $\tilde{x}$ of problem (1.2) such that $\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda}) - d_\rho(\tilde{\lambda}) \leq \zeta$;

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

$$\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\mathcal{L}_\rho(\tilde{x}, \tilde{\lambda})/M_\rho) \tag{4.5}$$

$$\tilde{\lambda}^+ := \tilde{\lambda} + \rho\mathcal{A}(\tilde{x}^+). \tag{4.6}$$

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 ρ 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 2, 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 (2.8). Third, Step 4 ensures that the method terminates in at most $\bar{N}$ outer iterations possibly reporting failure. Fourth, in order to check (4.2), it is necessary to generate the lower bounds on $d_\rho(\lambda_k)$ by using Nesterov's method (see, e.g., Theorem 2 of [17] or Theorem 10 of [5]). Finally, at the beginning of Step 2, the pair (x_k, λ_k) satisfies the primal termination condition (2.5), but not necessarily the dual termination criterion (2.6). By calling subroutine Postprocessing, the next result guarantees that the output pair $(\tilde{x}^+, \tilde{\lambda}^+)$ of this subroutine satisfies both (2.5) and (2.6).

Proposition 4 *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 ζ is given by (4.3), then the pair $(\tilde{x}^+, \tilde{\lambda}^+)$ defined by (4.5) and (4.6) is an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).

Proof Clearly, by Lemma 2(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 (4.3), respectively. The above inequality together with (2.7), (4.6), and Proposition 3(b) with $\phi(\cdot) = \mathcal{L}_\rho(\cdot, \tilde{\lambda})$, $L_\phi = M_\rho$, and $\tau = 1/M_\rho$ then imply that

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

where $\tilde{x}^+$ is defined in (4.5). Moreover, it follows from Lemma 2(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 (4.3) 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 2, 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

$$\|\mathcal{A}(\tilde{x}^+)\| \leq \|\mathcal{A}(\tilde{x}^+) - u_\lambda^*\| + \|\mathcal{A}(x) - u_\lambda^*\| + \|\mathcal{A}(x)\| \leq \frac{\epsilon_p}{8} + \frac{\epsilon_p}{8} + \frac{3\epsilon_p}{4} = \epsilon_p. \tag{4.7}$$

We have thus shown that $(\tilde{x}^+, \tilde{\lambda}^+)$ is an (ϵ_p, ϵ_d)-primal-dual solution of (1.1). □

The following result follows as an immediate consequence of Proposition 4.

Corollary 2 *If the I-AL method successfully terminates (i.e., at Step 2), then the output pair of subroutine Postprocessing is an (ϵ_p, ϵ_d)-primal-dual solution of* (1.1).

Proof The result follows from Proposition 4, (4.1), and the fact that at Step 4, conditions (4.2) and $\|\mathcal{A}(x_k)\| \leq 3\epsilon_p/4$ hold. □

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

Theorem 4 *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*

$$\bar{N} \geq N := \left\lceil \frac{16 D_\Lambda^2}{\rho^2 \epsilon_p^2} \right\rceil, \tag{4.8}$$

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

$$\sum_{k=0}^{\bar{N}-1} \eta_k \leq \frac{\rho \epsilon_p^2}{128}. \tag{4.9}$$

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

Proof 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, (4.2), Proposition 2, and assumptions (4.8) and (4.9) that

$$\begin{aligned} \sum_{k=0}^{N-1} e_k^2 = \sum_{k=0}^{N-1} \frac{\delta_k^2}{\|\mathcal{A}(x_k)\|^2} &\leq \frac{16}{9\epsilon_p^2} \sum_{k=0}^{N-1} \|\mathcal{A}(x_k) - u_{\lambda_k}^*\|^2 \\ &\leq \frac{32}{9\rho\epsilon_p^2} \sum_{k=0}^{N-1} \eta_k \leq \frac{32}{9\rho\epsilon_p^2} \sum_{k=0}^{\bar{N}-1} \eta_k \leq \frac{1}{36}. \end{aligned} \tag{4.10}$$

Noting that (4.10) implies $e_k \leq 1/6$, and hence that condition (3.17) holds with $\alpha_k = 1$ and $c_1 = 7/18$, it follows from (4.10) and Corollary 1 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

$$\begin{aligned} \|\mathcal{A}(x_k)\|^2 &\leq \frac{4 D_\Lambda^2}{c_1(1+c_1)\,\rho^2(k+1)} \exp\left(\frac{2}{c_1} \sum_{j=0}^{k} e_j^2\right) \\ &\leq \frac{1296 D_\Lambda^2}{175 \rho^2 (k+1)} \exp\left(\frac{1}{7}\right) \leq \frac{9 D_\Lambda^2}{\rho^2 (k+1)}, \end{aligned} \tag{4.11}$$

for every $0 \leq k \leq N-1$. The above inequality with $k = N-1$ together with (4.8) then imply that

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

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

We now make a few observations about Theorem 4. First, we observe that Theorem 4 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 ϵ_d, the number of inner iterations will depend on it, since the number of inner iteration inside subroutine Postprocessing clearly depends on ϵ_d in view of (4.3). Third, observe that Eq. (4.8) implies that the larger ρ 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 (1.2) is given by M_ρ (see (4.4)), increasing ρ will increase M_ρ, and as a consequence, will increase the iteration-complexity bound of Nesterov's optimal method for finding an approximate solution of (1.2).

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 5 *Let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, $\rho > 0$, $\bar{N} \in \mathbb{N} \cup \{+\infty\}$, and $\{\eta_k\}_{k=0}^{\bar{N}-1} \subseteq \mathbf{R}_{++}$ be given such that conditions (4.8) and (4.9) are satisfied. Then, the I-AL method applied to (1.1) successfully terminates in N outer iterations, and computes an (ϵ_p, ϵ_d)-primal-dual solution of (1.1) in at most $\mathcal{I}_p + \mathcal{I}_d$ inner iterations, where N is defined in Theorem 4,*

$$\mathcal{I}_p := \left\lfloor \sqrt{2} D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{N-1} \eta_k^{-\frac{1}{2}} + N \right\rfloor, \quad \mathcal{I}_d := \left\lceil 4 D_X \max\left\{ \frac{4 M_\rho^{\frac{1}{2}}}{\rho^{\frac{1}{2}} \epsilon_p}, \frac{M_\rho}{\epsilon_d} \right\} \right\rceil \tag{4.12}$$

and

$$D_X := \max_{x_1, x_2 \in X} \|x_1 - x_2\|. \tag{4.13}$$

Proof Clearly, in view of Theorem 1 and Theorem 4, 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 Theorem 1, 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 ζ in (4.3), it follows that the number of inner iterations performed at step 2 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 5. The first one stated below bounds the total number of inner iterations of the I-AL method by assuming that $\bar{N}$ is finite and $\eta_0, \ldots, \eta_{\bar{N}-1}$ is uniform. In addition, instead of assuming the exact knowledge of D_Λ, 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 (4.12) subject to a condition like (4.9) occurs exactly when $\eta_0, \ldots, \eta_{N-1}$ is uniformly chosen. It should be noted, however, that it is possible to choose non-uniform η_k's in order to guarantee the convergence of the I-AL method.

Theorem 5 *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 (1.1) with input*

$$\rho = \rho(t) := \frac{4\, t^{\frac{3}{4}} \epsilon_d^{\frac{1}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}} \epsilon_p} + \frac{L_f}{\|\mathcal{A}\|^2}, \quad \bar{N} = \bar{N}(t) := \left\lceil \frac{16 t^2}{\rho(t)^2 \epsilon_p^2} \right\rceil, \tag{4.14}$$

$$\eta_k = \eta(t) := \frac{\rho(t) \epsilon_p^2}{128 \bar{N}(t)}, \quad \forall\, k \geq 0, \tag{4.15}$$

then the method successfully terminates in

$$\left\lceil \min \left\{ \frac{D_\Lambda^2 \|\mathcal{A}\|^{\frac{1}{2}}}{t^{\frac{3}{2}} \epsilon_d^{\frac{1}{2}}}, \frac{16 D_\Lambda^2 \|\mathcal{A}\|^4}{L_f^2 \epsilon_p^2} \right\} \right\rceil \leq \left\lceil \min \left\{ \frac{D_\Lambda^{\frac{1}{2}} \|\mathcal{A}\|^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}}, \frac{16 D_\Lambda^2 \|\mathcal{A}\|^4}{L_f^2 \epsilon_p^2} \right\} \right\rceil \tag{4.16}$$

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

$$\mathcal{I}_{pd}(t) := \left\lceil D_X \left(\frac{\|\mathcal{A}\|^{\frac{7}{4}} t^{\frac{3}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}} + \frac{\|\mathcal{A}\|}{\epsilon_p} + \frac{L_f}{\epsilon_d} \right) + \left(\frac{t \|\mathcal{A}\|}{\epsilon_d} \right)^{\frac{1}{2}} \right\rceil, \tag{4.17}$$

and D_X and D_Λ are defined in Theorem 4 and Proposition 5, respectively.

Proof Using (4.14) and the assumption that $t \geq D_\Lambda$, we obtain

$$\bar{N}(t) \geq \left\lceil \frac{16 D_\Lambda^2}{\rho^2 \epsilon_p^2} \right\rceil = N. \tag{4.18}$$

Also note that (4.14) and (4.15) imply that $\sum_{k=0}^{\bar{N}-1} \eta_k = \bar{N} \eta(t) = \bar{N}(t) \eta(t) = \rho \epsilon_p^2 / 128$. We have thus shown that conditions (4.8) and (4.9) hold. It then follows from Proposition 5 that the total number of outer iterations is bounded by N, where N is defined by (4.8). Bound (4.16) now follows by combining the definition of N in (4.8) with the fact

$$\rho = \rho(t) \geq \max \left\{ \frac{4 t^{\frac{3}{4}} \epsilon_d^{\frac{1}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}} \epsilon_p}, \frac{L_f}{\|\mathcal{A}\|^2} \right\} \geq \max \left\{ \frac{4 D_\Lambda^{\frac{3}{4}} \epsilon_d^{\frac{1}{4}}}{\|\mathcal{A}\|^{\frac{1}{4}} \epsilon_p}, \frac{L_f}{\|\mathcal{A}\|^2} \right\}. \tag{4.19}$$

It also follows from Proposition 5 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 (4.12). Noting that by (4.14), (4.15), and the convexity of $q^{\frac{3}{2}}$ for $q \in \mathbf{R}$,

we have

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

we then conclude from (4.12), (4.14), and (4.18) that

$$\mathcal{I}_p \le \sqrt{2}D_X M_\rho^{\frac{1}{2}} \sum_{k=0}^{\bar{N}(t)-1} \eta_k^{-\frac{1}{2}} + \bar{N}(t)$$
$$\le \frac{16\sqrt{2}D_X M_\rho^{\frac{1}{2}}}{\rho(t)^{\frac{1}{2}}\epsilon_p}\left(\frac{64t^3}{\rho(t)^3\epsilon_p^3}+1\right) + \frac{16t^2}{\rho(t)^2\epsilon_p^2}+1. \tag{4.20}$$

Now, by using the first relation in (4.14), we have that $\rho(t) \ge L_f/\|\mathcal{A}\|^2$, and hence that

$$M_\rho = L_f + \rho(t)\|\mathcal{A}\|^2 \le 2\rho(t)\|\mathcal{A}\|^2. \tag{4.21}$$

This conclusion together with (4.19) and (4.20) then imply that

$$\mathcal{I}_p \le \frac{32D_X\|\mathcal{A}\|}{\epsilon_p}\left(\frac{64t^3}{\rho(t)^3\epsilon_p^3}+1\right) + \frac{16t^2}{\rho(t)^2\epsilon_p^2}+1$$
$$\le \frac{32D_X\|\mathcal{A}\|}{\epsilon_p}\left(\frac{\|\mathcal{A}\|^{\frac{3}{4}}t^{\frac{3}{4}}}{\epsilon_d^{\frac{3}{4}}}+1\right) + \frac{\|\mathcal{A}\|^{\frac{1}{2}}t^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}}+1. \tag{4.22}$$

Moreover, it easily follows from (4.12), (4.21) and (4.14) that

$$\mathcal{I}_d \le 4D_X\left(\frac{4M_\rho^{\frac{1}{2}}}{\rho(t)^{\frac{1}{2}}\epsilon_p}+\frac{M_\rho}{\epsilon_d}\right)+1 \le 4D_X\left(\frac{4\sqrt{2}\|\mathcal{A}\|}{\epsilon_p}+\frac{2\rho(t)\|\mathcal{A}\|^2}{\epsilon_d}\right)+1$$
$$= \frac{16\sqrt{2}D_X\|\mathcal{A}\|}{\epsilon_p} + 8D_X\left(\frac{4t^{\frac{3}{4}}\|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p\epsilon_d^{\frac{3}{4}}}+\frac{L_f}{\epsilon_d}\right)+1. \tag{4.23}$$

Combining (4.22) and (4.23), we easily see that the I-AL method computes an (ϵ_p, ϵ_d)-primal-dual solution of (1.1) in at most $\mathcal{O}(\mathcal{I}_{pd}(t))$ inner iterations, where $\mathcal{I}_{pd}(t)$ is defined by (4.17). □

Observe that the choice of ρ, $\bar{N}$, and $\{\eta_k\}$ given by (4.14) and (4.15) requires $t \ge D_\Lambda$ so as to guarantee conditions (4.8) and (4.9), and hence that the conclusions

of Theorem 4 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

$$\beta_0 := 1 + \frac{32 D_X \|\mathcal{A}\|}{\epsilon_p}, \quad \beta_1 := \frac{32 D_X \|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p \epsilon_d^{\frac{3}{4}}}, \quad \beta_2 := \frac{\|\mathcal{A}\|^{\frac{1}{2}}}{\epsilon_d^{\frac{1}{2}}}; \tag{4.24}$$

(1) Run the I-AL method with the above input and with $\rho = \rho(t_j)$, $\bar{N} = \bar{N}(t_j)$ given by (4.14) and $\eta_k = \eta(t_j)$, $k = 0, \ldots, \bar{N}(t_j)$, where $\eta(t_j)$ is given by (4.15);

(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

Before establishing the complexity bound for the above I-AL guess-and-check procedure, we first state the following technical result.

Proposition 6 *For some positive integer L, let positive scalars $p_1, p_2, \ldots, p_L$ be given. Then, there exists a constant $C = C(p_1, \ldots, p_L)$ such that for any nonnegative scalars $\beta_0, \beta_1, \ldots, \beta_L$, ν, and $\bar{t}$, we have*

$$\begin{aligned} &\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, t_k^{p_l}\right) \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil\right\} \\ &\quad \le C \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, \bar{t}^{p_l}\right) \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\}, \end{aligned} \tag{4.25}$$

where

$$\begin{aligned} K &:= \max\left\{0, \left\lceil \log\left(\frac{\bar{t}}{t_0}\right) \right\rceil\right\}, \\ t_0 &:= \min_{1 \le l \le L} \left(\frac{\max(\beta_0, 1)}{\beta_l}\right)^{1/p_l}, \quad t_k = t_0 2^k, \ \forall k = 1, \ldots, K. \end{aligned} \tag{4.26}$$

In particular, if $\nu = \bar{t}$, then (4.25) implies that

$$\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, t_k^{p_l}\right) \right\rceil \le C \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, \bar{t}^{p_l}\right) \right\rceil. \tag{4.27}$$

Proof The inequality (4.27) is shown in Lemma 13 of [11], while the more general result in (4.25) is shown in Proposition 5.4.1 of [9]. □

We are now ready to state the iteration-complexity of the above I-AL guess-and-check procedure.

Theorem 6 *Let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. The I-AL guess-and-check procedure finds an (ϵ_p, ϵ_d)-primal-dual solution of (1.1) in at most $\mathcal{O}(\mathcal{I}_{pd}(D_\Lambda))$ inner iterations, where $\mathcal{I}_{pd}(t)$ is defined by (4.17).*

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

$$\bar{J} := \max\{0, \lceil \log(D_\Lambda / t_0) \rceil\} \tag{4.28}$$

and noting that $t_{\bar{J}} = t_0 2^{\bar{J}} \geq D_\Lambda$, we conclude from Theorem 5 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

$$\sum_{j=0}^{J} \mathcal{I}_{p,j} + \mathcal{I}_{d,J} \leq \sum_{j=0}^{\bar{J}} \mathcal{I}_{p,j} + \mathcal{I}_{d,J}. \tag{4.29}$$

Since the total number of outer iterations at the jth loop is bounded by $N(t_j)$, it follows from Theorem 1 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 (4.20), (4.21) and (4.22), we can show that for $j = 0, \ldots, J$, we have

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

where β_0, β_1, and β_2 are given by (4.24). 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 (4.27) with $L = 2$, $p_1 = 3/4$, $p_2 = 1/2$, $\bar{t} = D_\Lambda$, $J = \bar{J}$, and β_0, β_1, and β_2 as above that

$$\sum_{j=0}^{\bar{J}} \mathcal{I}_{p,j} = \mathcal{O}(1) \left\lceil \beta_0 + \beta_1 D_\Lambda^{\frac{3}{4}} + \beta_2 D_\Lambda^{\frac{1}{2}} \right\rceil. \tag{4.30}$$

Now, using (4.28), it is easy to see that $t_J \leq t_{\bar{J}} \leq \max\{t_0, 2D_\Lambda\}$ and hence that

$$t_J^{\frac{3}{4}} \leq \max\left\{t_0^{\frac{3}{4}}, (2D_\Lambda)^{\frac{3}{4}}\right\} \leq \max\left\{\frac{\beta_0}{\beta_1}, (2D_\Lambda)^{\frac{3}{4}}\right\} \leq \frac{\beta_0}{\beta_1} + (2D_\Lambda)^{\frac{3}{4}}, \quad (4.31)$$

where the second 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 β_0 and β_1 in (4.24), and an argument similar to the proof of (4.23), we have

$$\begin{aligned}\mathcal{I}_{d,J} &\leq \frac{16\sqrt{2}D_X\|\mathcal{A}\|}{\epsilon_p} + 8D_X\left[\frac{4t_J^{\frac{3}{4}}\|\mathcal{A}\|^{\frac{7}{4}}}{\epsilon_p\epsilon_d^{\frac{3}{4}}} + \frac{L_f}{\epsilon_d}\right] + 1 \\ &\leq \beta_0 + \beta_1 t_J^{\frac{3}{4}} + \frac{8D_X L_f}{\epsilon_d} \leq 2\beta_0 + \beta_1(2D_\Lambda)^{\frac{3}{4}} + \frac{8D_X L_f}{\epsilon_d}. \quad (4.32)\end{aligned}$$

Now, using (4.30) and (4.32), it is easy to see that the right-high-side of (4.29) is bounded by $\mathcal{O}(\mathcal{I}_{pd}(D_\Lambda))$, where $\mathcal{I}_{pd}(\cdot)$ is defined in (4.17). □

It is interesting to compare the iteration-complexity bound obtained in Theorem 6 with the corresponding one obtained for the quadratic penalty method in [11] to compute an (ϵ_p, ϵ_d)-primal-dual solution of (1.1), 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 (1.2). 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 (1.1), 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 (1.1).

4.2 The I-AL method applied to a perturbation problem

In this subsection, we will exploit the possibility of solving problem (1.1) by applying a slightly modified version of the I-AL algorithm to a perturbed problem obtained by adding a small strongly convex perturbation to f in (1.1), i.e.,

$$f_\gamma^* := \min\left\{f_\gamma(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 : \mathcal{A}(x) = 0, x \in X\right\}. \quad (4.33)$$

Here x_0 is a fixed point in X and $\gamma > 0$ is a prespecified perturbation parameter. The following result, whose proof can be found in Lemma 15 of [11], shows that if γ is sufficiently small, then an approximate solution of (4.33) will also be an approximate solution of (1.1).

Lemma 6 *Let f^* and f^*_γ be the optimal values defined in (1.1) and (4.33), respectively, and D_X be defined in Proposition 5. Then,*

$$0 \le f^*_\gamma - f^* \le \gamma D_X^2/2. \tag{4.34}$$

Our goal in this section is to derive an iteration-complexity bound for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1.1) by applying the I-AL method directly to problem (4.33) for a conveniently chosen perturbation parameter $\gamma > 0$.

The augmented dual function associated with (4.33) is given by

$$d_{\rho,\gamma}(\lambda) := \min_{x\in X} \left\{ \mathcal{L}_{\rho,\gamma}(x,\lambda) := f_\gamma(x) + \lambda^T \mathcal{A}(x) + \frac{\rho}{2}\|\mathcal{A}(x)\|^2 \right\}, \tag{4.35}$$

or alternatively, by

$$d_{\rho,\gamma}(\lambda) = \min_u \left\{ v_{\rho,\gamma}(u,\lambda) := v_\gamma(u) + \langle \lambda, u\rangle + \frac{\rho}{2}\|u\|^2 \right\}, \tag{4.36}$$

where $v_\gamma(\cdot)$ is the value function associated with the perturbed problem (4.33) (see definition (2.2)). We denote the optimal solution of (4.36) 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

$$M_{\rho,\gamma} := L_f + \rho\|A\|^2 + \gamma, \tag{4.37}$$

and that it is strongly convex with modulus γ 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 (4.33) (and hence with M_ρ, $\mathcal{L}_\rho$, and d_ρ 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 2 is used to compute the approximate solutions x_k in step 1 and $\tilde{x}$ in subroutine Postprocessing, and the tolerance ζ in (4.3) is replaced by

$$\tilde{\zeta} = \tilde{\zeta}(\rho,\gamma) := \min\left\{ \frac{\rho\epsilon_p^2}{128}, \frac{\epsilon_d^2}{32 M_{\rho,\gamma}} \right\}. \tag{4.38}$$

The next results is a corresponding version of Proposition 4, which guarantees that the output pair $(\tilde{x}^+, \tilde{\lambda}^+)$ of subroutine Postprocessing is an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).

Proposition 7 *Let $\rho > 0$, $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, and $\tilde{\lambda} \in \Re^m$ be given, and define*

$$\gamma := \frac{\epsilon_d}{2D_X}. \tag{4.39}$$

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

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

If $\tilde{x} \in X$ is a point satisfying $\mathcal{L}_{\rho,\gamma}(\tilde{x}, \tilde{\lambda}) - d_{\rho,\gamma}(\tilde{\lambda}) \leq \tilde{\zeta}$, where $\tilde{\zeta}$ is given by (4.38), then the pair $(\tilde{x}^+, \tilde{\lambda}^+)$ defined by (4.5) and (4.6) with $\mathcal{L}_\rho$ replaced by $\mathcal{L}_{\rho,\gamma}$ is an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).

Proof As in the proof of Proposition 4 with ζ 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 (4.5) 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 (4.13) and (4.39) imply that $\gamma\|\tilde{x}^+ - x_0\| \leq \gamma D_X = \epsilon_d/2$, 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 4, we can show that $\|\mathcal{A}(\tilde{x}^+)\| \leq \epsilon_p$. Thus, $(\tilde{x}^+, \lambda_k^+)$ is an (ϵ_p, ϵ_d)-primal-dual solution for (1.1). □

The following result follows as an immediate consequence Proposition 4.

Corollary 3 *If the modified I-AL method successfully terminates (i.e., at Step 2), then the output pair of subroutine Postprocessing is an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).*

Proof The result follows from Proposition 7, (4.1), and the fact that at Step 4, conditions (4.2) and $\|\mathcal{A}(x_k)\| \leq 3\epsilon_p/4$ hold. □

Our goal in the remaining part of this subsection is to establish 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 Sect. 4.1.

Lemma 7 *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*

$$\frac{\gamma}{2}\|x_k - x_{k+1}^*\|^2 \leq \mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - d_{\rho,\gamma}(\lambda_{k+1}) \leq \left(\sqrt{\eta_k} + \sqrt{\frac{\rho}{2}}\|\mathcal{A}(x_k)\|\right)^2, \tag{4.40}$$

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 (4.40) follows immediately from the strong convexity of $\mathcal{L}_{\rho,\gamma}(\cdot, \lambda_{k+1})$. Hence, it suffices to show the second inequality in (4.40). Clearly, by definition (4.35) 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 $\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_k) \le \eta_k$ then imply that

$$\begin{aligned}\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - d_{\rho,\gamma}(\lambda_{k+1}) &= \big[\mathcal{L}_{\rho,\gamma}(x_k, \lambda_{k+1}) - \mathcal{L}_{\rho,\gamma}(x_k, \lambda_k)\big] \\ &\quad + \big[\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})\big] \\ &= \rho\|\mathcal{A}(x_k)\|^2 + \big[\mathcal{L}_{\rho,\gamma}(x_k, \lambda_k) - d_{\rho,\gamma}(\lambda_k)\big] \\ &\quad + \big[d_{\rho,\gamma}(\lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})\big] \\ &\le \rho\|\mathcal{A}(x_k)\|^2 + \eta_k + [d_{\rho,\gamma}(\lambda_k) - d_{\rho,\gamma}(\lambda_{k+1})].\end{aligned} \tag{4.41}$$

Moreover, in view of Proposition 1 applied to the perturbed problem (4.33), 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 (1.5) that applied to $f(\cdot) = -d_{\rho,\gamma}(\cdot)$ that

$$\begin{aligned}-d_{\rho,\gamma}(\lambda_{k+1}) + d_{\rho,\gamma}(\lambda_k) &\le \langle -u^*_{\lambda_k,\gamma}, \lambda_{k+1} - \lambda_k\rangle + \frac{1}{2\rho}\|\lambda_{k+1} - \lambda_k\|^2 \\ &= -\rho\langle u^*_{\lambda_k,\gamma}, \mathcal{A}(x_k)\rangle + \frac{\rho}{2}\|\mathcal{A}(x_k)\|^2\end{aligned} \tag{4.42}$$

where the last equality follows from the fact that $\lambda_{k+1} - \lambda_k = \rho\mathcal{A}(x_k)$. Combining (4.41) and (4.42), we obtain

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

where the last inequality follows from Proposition 2 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 (4.33) when a constant sequence $\{\eta_k\}$ is applied.

Lemma 8 *Let $\rho > 0$, $(\epsilon_p, \epsilon_d) \in \Re_{++} \times \Re_{++}$ and $\bar{N} \in \mathbb{N}$ be given, and let γ be given by (4.39). Consider the modified I-AL method applied to the perturbed problem*

(4.33) *with penalty parameter* ρ*, iteration limit* $\bar{N}$ *and inner tolerances* $\eta_0, \ldots, \eta_{\bar{N}}$ *given by*

$$\eta_k = \eta_\gamma := \frac{\rho\epsilon_p^2}{128\bar{N}}, \quad k = 0, \ldots, \bar{N} - 1. \tag{4.43}$$

Then the following statements hold:

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

$$\left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\{ 2\max\left(1, \left\lceil \log \frac{64\gamma\bar{N}D_X^2}{\rho\epsilon_p^2} \right\rceil\right) + \min(\bar{N}, N_\gamma) \left\lceil 2\log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho\epsilon_p}\right)\right\rceil \right.$$
$$\left. + \max\left(1, \left\lceil \log \frac{16\gamma M_{\rho,\gamma} D_X^2}{\epsilon_d^2} \right\rceil\right)\right\}, \tag{4.44}$$

where

$$N_\gamma := \left\lceil \frac{16[D_\Lambda^\gamma]^2}{\rho^2\epsilon_p^2} \right\rceil; \tag{4.45}$$

(b) *if* $\bar{N} \geq N_\gamma$*, then the above method successfully terminates in* N_γ *outer iterations with an* (ϵ_p, ϵ_d)*-primal-dual solution of* (1.1).

Proof Statement (b) immediately follows from the assumption $\bar{N} \geq N_\gamma$ and Theorem 4 applied to the perturbed problem (4.33). 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

$$0 \leq K \leq \min\{\bar{N}, N_\gamma\} - 1. \tag{4.46}$$

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 (4.11), we can show that

$$\|\mathcal{A}(x_k)\|^2 \leq \frac{9[D_\Lambda^\gamma]^2}{\rho^2(k+1)}, \quad k = 1, \ldots, K - 1. \tag{4.47}$$

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 2 with $\phi(\cdot) = \mathcal{L}_{\rho,\gamma}(\cdot, \lambda_0)$, $L_\phi = M_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \eta_\gamma$ (4.13) and (4.43), we have

$$l_0 \leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma\|x_{-1} - x_0^*\|^2}{2\eta_\gamma} \right\rceil\right\}$$
$$\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma D_X^2}{2\eta_\gamma} \right\rceil\right\}$$

$$= \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{64\gamma \bar{N} D_X^2}{\rho \epsilon_p^2} \right\rceil \right\}. \tag{4.48}$$

It also follows from Theorem 2 that

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

Now by using (4.40) and (4.47), we have

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

We then conclude from the previous two observations and (4.43) that, $\forall k = 1, \ldots, K$,

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

The above conclusion together with (4.44) and (4.48) then clearly imply that the total number of inner iterations performed at step 1) of the modified I-AL method is bounded by

$$\begin{aligned} l_0 + \sum_{k=1}^{K} l_k &\le l_0 + K \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil 2\log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil \\ &\le \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left\{ 2\max\left(1, \left\lceil \log \frac{64\gamma \bar{N} D_X^2}{\rho \epsilon_p^2} \right\rceil\right) \right. \\ &\quad \left. + K \left\lceil \log\left(1 + \frac{24\bar{N}^{\frac{1}{2}} D_\Lambda^\gamma}{\rho \epsilon_p}\right) \right\rceil \right\}. \end{aligned} \tag{4.49}$$

Moreover, let $\tilde{l}_K$ denote the number of inner iterations performed by subroutine Post-Processing. By using Theorem 2 with $\phi(\cdot) = \mathcal{L}_{\rho,\gamma}(\cdot, \lambda_K)$, $L_\phi = M_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \tilde{\zeta}$ and (4.38), we have

$$\tilde{l}_K \leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\gamma D_X^2}{2\tilde{\zeta}} \right\rceil \right\}$$
$$\leq \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \left[\max\left\{1, \left\lceil \log \frac{64\gamma D_X^2}{\rho\epsilon_p^2} \right\rceil \right\} + \max\left\{1, \left\lceil \log \frac{16\gamma M_{\rho,\gamma} D_X^2}{\epsilon_d^2} \right\rceil \right\} \right]. \tag{4.50}$$

Combining inequalities (4.46), (4.49) and (4.50), we can easily see that the total number of inner iterations performed by the modified I-AL method is bounded by (4.44). □

We now state the corresponding version of Theorem 5 with respect to the modified I-AL method.

Theorem 7 *Let $(\epsilon_p, \epsilon_d) \in \Re_{++} \times \Re_{++}$ be given, and let γ be given by (4.39). For some $t > 0$, consider the modified I-AL method applied to the perturbed problem (4.33) with input*

$$\rho = \rho_\gamma(t) := \frac{4t}{\epsilon_p (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f + \gamma}{\|\mathcal{A}\|^2}, \tag{4.51}$$

$$\bar{N} = \bar{N}_\gamma(t) := \left\lceil \frac{16t^2}{\rho_\gamma(t)^2 \epsilon_p^2} \right\rceil, \quad \eta_k = \eta_\gamma(t) := \frac{\rho_\gamma(t)\epsilon_p^2}{128\bar{N}_\gamma(t)}, \quad \forall\, k \geq 0, \tag{4.52}$$

where

$$\mathcal{T}(t) := \mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3, \tag{4.53}$$

$$\mathcal{S}_1 := \sqrt{\frac{D_X \|\mathcal{A}\|^2}{\epsilon_p \epsilon_d}}, \quad \mathcal{S}_2 := \sqrt{\frac{D_X L_f}{\epsilon_d}} + 1 \text{ and } \mathcal{S}_3 := \sqrt{\frac{D_X \|\mathcal{A}\|}{\epsilon_p}} + 3. \tag{4.54}$$

Then the following statements hold:

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

$$\mathcal{O}\left\{ \left(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 [\log \mathcal{T}(t)]^{\frac{1}{4}} \right) [\log \mathcal{T}(t)]^{\frac{3}{4}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t)}{t}\right) \right\}; \tag{4.55}$$

(b) *if $t \geq D_\Lambda^\gamma$, where $D_\Lambda^\gamma := \min_{\lambda_\gamma \in \Lambda_\gamma^*} \|\lambda_0 - \lambda^*\|$ and Λ_γ^* denotes the set of Lagrange multipliers associated with (4.33), then the above method successfully terminates in $\mathcal{O}(\log \mathcal{T}(t))$ outer iterations with an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).*

Proof We first show part (a). It immediately follows from Lemma 8a that the total number of inner iterations performed by the modified I-AL method is bounded by

(4.44) with $\bar{N} = \bar{N}_\gamma(t)$ and $\rho = \rho_\gamma(t)$. Note that by (4.51), (4.52), and the fact that, by (4.53) and (4.54), $\log \mathcal{T}(t) \geq 2$, we have

$$\bar{N}_\gamma(t) \leq \frac{16t^2}{\rho_\gamma(t)^2 \epsilon_p^2} + 1 \leq \log \mathcal{T}(t) + 1 \leq 2 \log \mathcal{T}(t). \tag{4.56}$$

Also, using definitions (4.37) and (4.51), we have that

$$\gamma \leq M_{\rho,\gamma} = L_f + \gamma + \rho \|\mathcal{A}\|^2 \leq 2\rho \|\mathcal{A}\|^2. \tag{4.57}$$

This observation together with (4.39) and (4.51) then imply that

$$\begin{aligned}
\left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \leq \left\lceil 4\sqrt{\frac{\rho\|\mathcal{A}\|^2}{\gamma}} \right\rceil &= \left\lceil 4\left(\frac{4t\|\mathcal{A}\|^2}{\gamma \epsilon_p (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{L_f}{\gamma} + 1 \right)^{\frac{1}{2}} \right\rceil \\
&\leq 4\left(\frac{4D_X t\|\mathcal{A}\|^2}{\epsilon_p \epsilon_d (\log \mathcal{T}(t))^{\frac{1}{2}}} + \frac{D_X L_f}{\epsilon_d} + 1 \right)^{\frac{1}{2}} + 1 \\
&\leq 8\sqrt{\frac{D_X t\|\mathcal{A}\|^2}{\epsilon_p \epsilon_d}} (\log \mathcal{T}(t))^{-\frac{1}{4}} + 4\sqrt{\frac{D_X L_f}{\epsilon_d}} + 5.
\end{aligned} \tag{4.58}$$

Observe that, by (4.56), (4.57), (4.53), and (4.54),

$$\begin{aligned}
\log \frac{64\gamma D_X^2 \bar{N}_\gamma(t)}{\rho_\gamma(t)\epsilon_p^2} \leq \log \frac{128\gamma D_X^2 \log \mathcal{T}(t)}{\rho_\gamma(t)\epsilon_p^2} &\leq \log \frac{256\|\mathcal{A}\|^2 D_X^2 \log \mathcal{T}(t)}{\epsilon_p^2} \\
&= 8 + 4\log\left(\frac{\|\mathcal{A}\| D_X}{\epsilon_p} \right)^{\frac{1}{2}} + \log\log \mathcal{T}(t) = \mathcal{O}\left(\log \mathcal{T}(t)\right),
\end{aligned} \tag{4.59}$$

and that, by (4.52), the fact that $\log x \leq x$, and (4.56),

$$\begin{aligned}
&\min(\bar{N}_\gamma(t), N_\gamma) \left\lceil 2\log\left(1 + \frac{24 D_\Lambda^\gamma [\bar{N}_\gamma(t)]^{\frac{1}{2}}}{\rho_\gamma(t)\epsilon_p} \right) \right\rceil \\
&\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{6D_\Lambda^\gamma}{t} \left[\log \bar{N}_\gamma(t)\right]^{\frac{1}{2}} [\bar{N}_\gamma(t)]^{\frac{1}{2}} \right) \right\rceil \\
&\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{6D_\Lambda^\gamma}{t} \bar{N}_\gamma(t) \right) \right\rceil \\
&\leq \bar{N}_\gamma(t) \left\lceil 2\log\left(1 + \frac{12 D_\Lambda^\gamma}{t} \log \mathcal{T}(t) \right) \right\rceil \\
&= \mathcal{O}\left\{ \log \mathcal{T}(t) \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(t)}{t} \right) \right\}.
\end{aligned} \tag{4.60}$$

It also follows from (4.57), (4.39), (4.51), (4.52), (4.53) and (4.54) that

$$\begin{aligned}\log\frac{16\gamma M_{\rho,\gamma}D_X^2}{\epsilon_d^2} &\le \log\frac{16M_{\rho,\gamma}^2D_X^2}{\epsilon_d^2} \le \log\left(\frac{8\rho\|\mathcal{A}\|^2D_X}{\epsilon_d}\right)^2\\ &\le 2\log\left[\frac{8\|\mathcal{A}\|^2D_X}{\epsilon_d}\left(\frac{4t}{\epsilon_p(\log\mathcal{T}(t))^{\frac{1}{2}}}+\frac{L_f+\gamma}{\|\mathcal{A}\|^2}\right)\right]\\ &= 2\log\left[\frac{8\|\mathcal{A}\|^2D_X}{\epsilon_d}\left(\frac{4t}{\epsilon_p(\log\mathcal{T}(t))^{\frac{1}{2}}}+\frac{L_f}{\|\mathcal{A}\|^2}+\frac{\epsilon_d}{2D_X\|\mathcal{A}\|^2}\right)\right]\\ &\le 2\log\left[8D_X\left(\frac{4t\|\mathcal{A}\|^2}{\epsilon_p\epsilon_d}+\frac{L_f}{\epsilon_d}\right)+4\right]=\mathcal{O}(\log\mathcal{T}(t)).\end{aligned}\tag{4.61}$$

Now substituting bounds (4.58), (4.59), (4.60), and (4.61) into bound (4.44), we obtain bound (4.55). Statement (b) follows immediately from Lemma 8b and the fact that, by (4.52), the assumption $t \ge D_\Lambda^\gamma$ and (4.56),

$$N_\gamma = \left\lceil\frac{16[D_\Lambda^\gamma]^2}{\rho_\gamma(t)^2\epsilon_p^2}\right\rceil \le \left\lceil\frac{16t^2}{\rho_\gamma(t)^2\epsilon_p^2}\right\rceil = \bar{N}_\gamma(t) \le 2\log\mathcal{T}(t).\tag{4.62}$$

□

Observe that the choice of ρ, $\bar{N}$, and $\{\eta_k\}$ given by (4.51) and (4.52) requires $t \ge D_\Lambda$ 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

$$\hat{t} := \left[\frac{\mathcal{S}_2^2+\mathcal{S}_2\sqrt{\mathcal{S}_2^2+4(\mathcal{S}_2+\mathcal{S}_3)}}{2\mathcal{S}_1}\right]^2,\tag{4.63}$$

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

where $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ are given by (4.54). Find a point $t_0 \in [0, \hat{t}]$ such that $0 \le \psi(t_0) \le 1$.

(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 $k \ge 0$, where γ is given by (4.39), and $\rho_\gamma(\cdot)$, $\bar{N}_\gamma(\cdot)$ and $\eta_\gamma(\cdot)$ are defined in (4.51) and (4.52).

(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 $0 \le \psi(t_0) \le 1$.

Lemma 9 *Let $\psi(t)$ and $\hat{t}$ be defined in (4.63). Then, the following statements hold:*

(a) *$\psi(t)$ is continuous and non-decreasing for $t \ge 0$;*
(b) *$\psi(0) \le 0$ and $\psi(\hat{t}) \ge 0$;*
(c) *there exists $t_0 \in [0, \hat{t}]$ such that $0 \le \psi(t_0) \le 1$. Moreover, we have*

$$\mathcal{S}_1 t_0^{\frac{1}{2}} \le \mathcal{S}_2 [\log \mathcal{T}(t_0)]^{\frac{1}{4}} + 1, \tag{4.64}$$

$$\mathcal{S}_1 t^{\frac{1}{2}} \ge \mathcal{S}_2 [\log \mathcal{T}(t)]^{\frac{1}{4}}, \quad \forall t \ge t_0, \tag{4.65}$$

$$\log \mathcal{T}(t_0) = \mathcal{O}(\log \mathcal{T}(0)), \tag{4.66}$$

where $\mathcal{T}(\cdot)$, $\mathcal{S}_1$ and $\mathcal{S}_2$ are defined in (4.53) and (4.54).

Proof Statement (a) immediately following from the fact that, by (4.63),

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

where in the first inequality we use the fact that $\log(\mathcal{S}_1 t^{\frac{1}{2}} + \mathcal{S}_2 + \mathcal{S}_3) \ge 2$ in view of (4.54). It can be easily seen from (4.63) that $\psi(0) \le 0$. Noting that, by the definition of $\hat{t}$ in (4.63),

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

we conclude from (4.63) and the fact that $\log \tau \le \tau \le \tau^2$ for $\tau \ge 1$ that

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

We have thus shown that (b) holds. We now show that part (c) holds. The existence of $t_0 \in [0, \hat{t}]$ satisfying $0 \le \psi(t_0) \le 1$ follows immediately from Lemma 9. Inequality (4.64) follows from (4.53), (4.63) and the fact $\psi(t_0) \le 1$. Moreover, we conclude from (4.53), (4.63), the assumption $\psi(t_0) \ge 0$ and Lemma 9(a) that

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

for any $t \ge t_0$, and hence that (4.65) holds. Also note that by (4.53), (4.63) and the fact that $t_0 \le \hat{t}$, we have

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

□

Clearly, the existence of the required t_0 follows from Lemma 9.c). 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.

We are now ready to establish the iteration-complexity of the above modified I-AL guess-and-check procedure for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1.1).

Theorem 8 *Let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. The modified I-AL guess-and-check procedure described above finds an (ϵ_p, ϵ_d) -primal-dual solution of (1.1) in at most*

$$\mathcal{O}\left\{\mathcal{S}_1 [D_\Lambda^\gamma]^{\frac{1}{2}} [\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma) + \mathcal{S}_2 \log \mathcal{T}(0) \log\log \mathcal{T}(0)\right\}, \quad (4.67)$$

inner iterations, where $\mathcal{S}_1, \mathcal{S}_2, \mathcal{T}(\cdot)$, and D_Λ^γ are defined in Theorem 7.

Proof 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 7, relations (4.64) and (4.66), and the fact that, by (4.53) and (4.54), $\mathcal{T}(t) \geq 4$ 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

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

which is clearly bounded by (4.67). 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

$$\bar{J} := \max\{0, \lceil \log(D_\Lambda^\gamma / t_0) \rceil\} \quad (4.68)$$

and note that

$$2 D_\Lambda^\gamma \geq t_{\bar{J}} := t_0 2^{\bar{J}} \geq D_\Lambda^\gamma. \quad (4.69)$$

Theorem 7b and the second inequality in (4.69) then imply that $J \leq \bar{J}$. Also observe that, by relation (4.25) 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

$$\begin{aligned}
&\sum_{j=0}^{\bar{J}} t_j^{\frac{1}{2}} \max\left(1, \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{t_j}\right) \\
&\le \sqrt{t_0} \sum_{j=0}^{\bar{J}} \left\lceil \frac{1}{\sqrt{t_0}} t_j^{\frac{1}{2}} \right\rceil \max\left(1, \left\lceil \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{t_j} \right\rceil\right) \\
&= \mathcal{O}\left\{\sqrt{t_0} \left\lceil \frac{1}{\sqrt{t_0}} [D_\Lambda^\gamma]^{\frac{1}{2}} \right\rceil \max\left(1, \left\lceil \log \frac{D_\Lambda^\gamma \log \mathcal{T}(2D_\Lambda^\gamma)}{D_\Lambda^\gamma} \right\rceil\right)\right\} \\
&= \mathcal{O}\left\{\left([D_\Lambda^\gamma]^{\frac{1}{2}} + \sqrt{t_0}\right) \max\left(1, \lceil \log\log \mathcal{T}(2D_\Lambda^\gamma)\rceil\right)\right\} \\
&= \mathcal{O}\left([D_\Lambda^\gamma]^{\frac{1}{2}} \log\log \mathcal{T}(D_\Lambda^\gamma)\right), \qquad (4.70)
\end{aligned}$$

where the last identity follows from the facts that $t_0 \le D_\Lambda^\gamma$ and $\log \mathcal{T}(D_\Lambda^\gamma) \ge 2$. Using the facts that $J \le \bar{J}$ and the function $\mathcal{T}$ given by (4.53) is non-decreasing, Theorem 7(a), relations (4.65) and (4.70), and the simple observation that by (4.69), we have $t_0 \le t_j \le 2D_\Lambda^\gamma$ 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 (4.67). □

It is interesting to compare the iteration-complexity bound obtained in Theorem 8 with the corresponding one obtained for the quadratic penalty method in [11] to compute an (ϵ_p, ϵ_d)-primal-dual solution of (1.1), namely, $\mathcal{O}\left(\mathcal{T}(\|\lambda_\gamma^*\|) \log \mathcal{T}(\|\lambda_\gamma^*\|)\right)$, where λ_γ^* is the minimum-norm Lagrange multiplier for the perturbed problem (4.33). 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

$$\mathcal{S}_2 \log \mathcal{T}(0) \log\log \mathcal{T}(0) = \mathcal{O}\left\{\mathcal{S}_1 [D_\Lambda^\gamma]^{\frac{1}{2}} [\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma)\right\}, \quad (4.71)$$

bound (4.67) 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 (4.71) holds if $L_f = 0$. Hence, when $\lambda_0 = 0$ and (4.71) holds, the

first complexity bound is worse than the latter one in Theorem 8 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 λ_0 for problem (4.33) is known, i.e., the ratio $D_\Lambda^\gamma / \|\lambda_\gamma^*\|$ is small, then the complexity bound in Theorem 8 is substantially smaller than the above one.

Note that we cannot compare the iteration-complexity of Theorem 6 with that obtained in Theorem 8 since the first one is expressed in terms of D_Λ and the latter in terms of D_Λ^γ. However, if $D_\Lambda^\gamma = \mathcal{O}(D_\Lambda)$ and (4.71) holds, then it can be easily seen that

$$\frac{\mathcal{S}_1 [D_\Lambda^\gamma]^{\frac{1}{2}} [\log \mathcal{T}(D_\Lambda^\gamma)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda^\gamma)}{\mathcal{I}_{pd}(D_\Lambda)} \leq \frac{\epsilon_d^{\frac{1}{4}} [\log \mathcal{T}(D_\Lambda)]^{\frac{3}{4}} \log\log \mathcal{T}(D_\Lambda)}{\mathcal{S}_1 (\|\mathcal{A}\| D_X D_\Lambda)^{\frac{1}{4}}}. \tag{4.72}$$

Hence, the second complexity is better than the first one whenever $D_\Lambda^\gamma = \mathcal{O}(D_\Lambda)$ and $\mathcal{S}_1$ is sufficiently large.

5 Concluding Remarks

In this paper, we establish the complexity of an I-AL method to solve a special class of convex programming problems. We also present a variant of this method with possibly better complexity, obtained by applying the original I-AL method to a perturbed problem. We demonstrate that both of these complexities compare favorably with the corresponding ones obtained in [11] for the quadratic penalty methods. More specifically, to compute a pair of (ϵ_p, ϵ_d)-solution of problem (1.1), the total number of iterations (in terms of the iterations of Nesterov's optimal method) performed by the I-AL method can be bounded by $\mathcal{O}(1/(\epsilon_p \epsilon_d^{\frac{3}{4}}))$, while the one by the quadratic penalty approach is bounded by $\mathcal{O}\left(1/(\epsilon_p \epsilon_d)\right)$. Moreover, the complexity of the modified I-AL method applied to the perturbed problem is given by $\mathcal{O}\left\{\sqrt{1/(\epsilon_p \epsilon_d)}\, \left[\log\left(1/(\epsilon_p \epsilon_d)\right)\right]^{\frac{3}{4}} \log\log\left(1/(\epsilon_p \epsilon_d)\right)\right\}$, while the one of the quadratic penalty method is given by $\mathcal{O}\left\{\sqrt{1/(\epsilon_p \epsilon_d)}\, \log\left(1/(\epsilon_p \epsilon_d)\right)\right\}$.

It should be noted, however, that while it is possible to derive a complexity bound depending on D_Λ instead of D_Λ^γ for the quadratic penalty method applied to the perturbed problem (see Corollary 20 of [11]), to obtain this type of complexity for the modified I-AL method seems to be more difficult and will be an interesting topic for future research. Observe that some other approaches, such as Nesterov's smoothing technique [17] and the extra-gradient methods (see [12,13]) can also be applied to problem (1.1). We refer to [11] for some discussions on the comparison between the penalty-based approaches and these alternative methods.

In this paper, we present guess-and-check procedures to deal with the case when the size of the Lagrange multiplier λ^* (or λ_γ^*) is unknown. However, our methods still require the input of the Lipschitz constant L_f in order to set up the stepsizes in Nesterov's method, the penalty parameter ρ in the I-AL method, and the accuracy ζ in the post-processing procedure. While the value of L_f is known for some important cases, e.g., when f is linear or quadratic, the following modifications can be incorporated

into our algorithmic scheme to cope with unknown L_f. First, we can use line search procedures in Nesterov's optimal method [18] or apply some recently developed uniformly optimal methods, e.g., those based on level methods [10], to solve subproblem (1.2). Second, we can use an arbitrary $\rho > 0$ and a summable sequence $\{\eta_k\}$, e.g., $\eta_k = \xi\rho\epsilon_p^2/[128(1+\xi)(k+1)^{1+\xi}]$ for some $\xi > 0$ in the I-AL method. Such a selection of $\{\eta_k\}$ will result in a slightly worse complexity (see Theorem 5.2.2 of [9]). Third, we can employ a guess-and-check procedure on ζ in the post-processing phase by using the results in Proposition 3 to check if ζ is small enough.

Observe that in this paper, we have used the classical augmented Lagrangian function (see (2.7)) mainly for its simplicity, for example, we know how to approximate its gradients (see Proposition 2). Recently, there have been some interesting theoretical developments that lead to new augmented Lagrangian functions with improved exactness properties [20]. It will be interesting to see if our complexity analysis can be generalized to the I-AL methods employed with these enhanced augmented Lagrangian functions.

References

1. Bertsekas, D.: Constrained Optimization and Lagrange Multiplier Methods, 1st edn. Academic Press, New York (1982)
2. Bertsekas, D.: Nonlinear Programming, 2nd edn. Athena Scientific, New York (1984)
3. Burer, S., Monteiro, R.D.C.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Math. Program. Series B **95**, 329–357 (2003)
4. Burer, S., Monteiro, R.D.C.: Local minima and convergence in low-rank semidefinite programming. Math. Program. **103**, 427–444 (2005)
5. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. SIAM J. Optim. **22**, 1469–1492 (2012)
6. Golshtein, E.G., Tretyakov, N.V.: Modified Lagrangians and Monotone Maps in Optimization. Springer, New York (1996)
7. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Appl. **4**, 303–320 (1969)
8. Jarre, F., Rendl, F.: An augmented primal-dual method for linear conic programs. Manuscript, Institut fur Mathematik, Universit at Dusseldorf, Germany, Austria, (2007)
9. Lan, G.: Convex optimization under inexact first-order information. Ph.D. Dissertation, School of Industrial Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, (2009)
10. Lan, G.: Bundle-level type methods uniformly optimal for smooth andnon-smooth convex optimization. Manuscript, Department of Industrialand Systems Engineering, University of Florida, Gainesville, FL32611, USA, January 2013. Revision submitted to MathematicalProgramming
11. Lan, G., Monteiro, R.D.C.: Iteration-complexity of first-order penalty methods for convex programming. Math. Program. **138**, 115–139 (2013)
12. Monteiro, R.D.C., Svaiter, B.F.: Complexity of variants of tseng's modified f-b splitting and korpelevich's methods for hemi-variational inequalities with applications to saddle-point and convex optimization problems. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, (2010)
13. Nemirovski, A.: 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 J. Optim. **15**, 229–251 (2004)
14. Nemirovski, A.S., Yudin, D.: Problem complexity and method efficiency in optimization. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, (1983)
15. Nesterov, Y.E.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Dokl. AN SSSR **269**, 543–547 (1983)
16. Nesterov, Y.E.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Massachusetts (2004)
17. Nesterov, Y.E.: Smooth minimization of nonsmooth functions. Math. Program. **103**, 127–152 (2005)

18. Nesterov, Y.E.: Gradient methods for minimizing composite objective functions. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (2007)
19. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (1999)
20. Pillo, G.D., Lucidi, S.: An augmented lagrangian function with improved exactness properties. SIAM J. Optim. **12**, 376–406 (2002)
21. Powell, M.M.D.: An efficient method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization, pp. 283–298. Academic Press, London (1969)
22. Ruszczynski, A.: Nonlinear Optimization, 1st edn. Princeton University Press, Princeton (2006)
23. Zhao, X., Sun, D., Toh, K.: A Newton-CG augmented Lagrangian method for semidefinite programming. Manuscript, National University of Singapore, Singapore (2008)