Iteration-complexity of first-order penalty methods for convex programming *

Guanghui Lan † Renato D.C. Monteiro ‡

July 24, 2008

Abstract

This paper considers a special but broad class of convex programing (CP) problems whose feasible region is a simple compact convex set intersected with the inverse image of a closed convex cone under an affine transformation. We study two first-order penalty methods for solving the above class of problems, namely: the quadratic penalty method and the exact penalty method. In addition to one or two gradient evaluations, an iteration of these methods requires one or two projections onto the simple convex set. We establish the iteration-complexity bounds for these methods to obtain two types of near optimal solutions, namely: near primal and near primal-dual optimal solutions. Finally, we present variants, with possibly better iteration-complexity bounds than the aforementioned methods, which consist of applying penalty-based methods to the perturbed problem obtained by adding a suitable perturbation term to the objective function of the original CP problem.

Keywords: Convex programming, quadratic penalty method, exact penalty method, Lagrange multiplier

AMS 2000 subject classification: 90C25, 90C06, 90C22, 49M37

1 Introduction

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

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

where $f : X \to \mathbf{R}$ is a convex function with Lipschitz continuous gradient, $X \subseteq \Re^n$ is a sufficiently simple closed convex set, $\mathcal{A} : \Re^n \to \Re^m$ is an affine function, and $\mathcal{K}^*$ denotes the dual cone of a closed convex cone $\mathcal{K} \subseteq \Re^m$, i.e., $\mathcal{K}^* := \{s \in \Re^m : \langle s, x\rangle \geq 0, \, \forall x \in \mathcal{K}\}$.

For the case where the feasible region consists only of the set X, or equivalently $\mathcal{A} \equiv 0$, Nesterov ([2, 4]) developed a method which finds a point $x \in X$ such that $f(x) - f^* \leq \epsilon$ in at most $\mathcal{O}(\epsilon^{-1/2})$

*The work of both authors were partially supported by NSF Grants CCF-0430644 and CCF-0808863 and ONR Grant N00014-08-1-0033.

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

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

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) but this approach would require the computation of projections onto the feasible region $X \cap \{x : \mathcal{A}(x) \in \mathcal{K}^*\}$, which for most practical problems is as expensive as solving the original problem itself.

In this paper, we are interested in alternative first-order methods for (1) whose iterations consist of, in addition to a couple of gradient evaluations, only projections onto the simple set X. More specifically, we present two penalty-based approaches for solving (1), namely: the quadratic and the exact penalty methods, and establish their iteration-complexity bounds for computing two types of near optimal solutions of (1). It is well-known that the penalty parameters of the penalization problems for the above penalty-based approaches must be chosen sufficiently large so that near optimal solutions of the penalization problems yield near optimal solutions for the original problem (1). Accordingly, we establish theoretical lower bounds on the penalty parameters that depend not only on the desired solution accuracies but also on the size $\|\lambda^*\|$ of the minimum norm Lagrange multiplier associated with the constraint $\mathcal{A}(x) \in \mathcal{K}^*$. Theoretically, setting the penalty parameters to their corresponding lower bounds would yield the lowest provably iteration-complexity bounds, and hence would be the best strategy to pursue. But since $\|\lambda^*\|$ is not known a priori, we present alternative penalty-based approaches based on simple "guess-and-check" procedures for these penalty parameters whose iteration-complexity bounds are of the same order as the approach in which $\|\lambda^*\|$ is known as priori, or equivalently, the theoretical lower bound on the penalty parameter is available. Finally, we present variants of the aforementioned methods, with possibly better iteration-complexity bounds, which consist of applying penalty-based methods to the problem obtained by adding a suitable perturbation term to the objective function of the original CP problem.

The paper is organized as follows. In Section 2, we give a more detailed description of problem (1) and definitions of two types of approximate solutions of (1). Section 3 discusses some technical results that will be used in our analysis and reviews two first order algorithms due to Nesterov [2, 4] for solving special classes of CP problems. In Section 4, we establish the iteration-complexity of quadratic penalty based methods for solving (1). More specifically, we present the iteration-complexity of the quadratic penalty method applied to (1) in Subsection 4.1, and of a variant, which consists of applying the quadratic penalty approach to the perturbed problem obtained by adding a suitable perturbation term to the objective function of (1), in Subsection 4.2. In Section 5, we establish the iteration-complexity of exact penalty based methods for solving (1). More specifically, we present the iteration-complexity of the exact penalty applied directly to (1) in Subsection 5.1, and of a variant, which consists of applying the exact penalty method to the perturbed problem corresponding to (1), in Subsection 5.2.

1.1 Notation and terminology

We denote the p-dimensional Euclidean space by $\mathbf{R}^p$. Also, $\mathbf{R}^p_+$ and $\mathbf{R}^p_{++}$ denote the nonnegative and the positive orthants of $\mathbf{R}^p$, respectively. In this paper, we use the notation $\Re^p$ to denote a p-dimensional vector space inherited with a inner product space $\langle \cdot, \cdot \rangle$. The dual norm $\|\cdot\|_*$ associated with an arbitrary norm $\|\cdot\|$ on $\Re^p$ is defined as

$$\|s\|_* := \max_x \{\langle s, x\rangle : \|x\| \leq 1\}, \quad \forall s \in \Re^p.$$

Note that if the norm $\|\cdot\|$ is the one induced by the inner product, i.e. $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$, then $\|\cdot\|_* \equiv \|\cdot\|$.

Given a closed convex set $\mathcal{C} \in \Re^p$, we define the distance function $d_{\mathcal{C}} : \Re^p \to \mathbf{R}$ to $\mathcal{C}$ with respect to $\|\cdot\|$ as $d_{\mathcal{C}}(u) := \min\{\|u-c\| : c \in \mathcal{C}\}$ for every $u \in \Re^p$. It is well-known that this minimum is always achieved at some $c \in \mathcal{C}$. Moreover, this minimizer is unique whenever $\|\cdot\|$ is a inner product norm. In such a case, we denote this unique minimizer by $\Pi_{\mathcal{C}}(u)$, i.e., $\Pi_{\mathcal{C}}(u) = \operatorname{argmin}\{\|u-c\| : c \in \mathcal{C}\}$ for every $u \in \Re^p$. The support function of a set $C \subset \Re^p$ is defined as $\sigma_C(u) := \sup\{\langle u, c\rangle : c \in C\}$.

2 Problem of interest

Consider inner product spaces $\Re^n$ and $\Re^m$ endowed with arbitrary norms, both denoted simply by $\|\cdot\|$. In this paper, we consider the CP problem (1) where $f : X \to \mathbf{R}$ is a convex function with L_f-Lipschitz-continuous gradient, i.e.:

$$\|\nabla f(\tilde{x}) - \nabla f(x)\|_* \leq L_f\|\tilde{x} - x\|, \quad \forall x, \tilde{x} \in X. \tag{2}$$

We define the norm of the map $\mathcal{A}$ as being the operator norm of its linear part $\mathcal{A}_0 := \mathcal{A}(\cdot) - \mathcal{A}(0)$, i.e.:

$$\|\mathcal{A}\| := \|\mathcal{A}_0\| = \max\{\|\mathcal{A}_0(x)\|_* : \|x\| \leq 1\} = \max\{\|\mathcal{A}(x) - \mathcal{A}(0)\|_* : \|x\| \leq 1\}.$$

The Lagrangian dual function and value function associated with (1) are defined as

$$d(\lambda) := \inf\{f(x) + \langle\lambda, \mathcal{A}(x)\rangle : x \in X\}, \quad \forall\lambda \in -\mathcal{K}, \tag{3}$$

$$v(u) := \inf\{f(x) : \mathcal{A}(x) + u \in \mathcal{K}^*,\, x \in X\}, \quad \forall u \in \Re^m. \tag{4}$$

We make the following assumptions throughout the paper:

Assumption 1 *A.1) the set X is bounded (and hence $f^* \in \mathbf{R}$);*

A.2) there exists a Lagrange multiplier for (1), i.e., a vector $\lambda^* \in -\mathcal{K}$ such that $f^* = d(\lambda^*)$.

Throughout this paper, we assume that the norm $\|\cdot\|$ in $\Re^m$ is induced by the inner pr. First note that x^* is an optimal solution of (1) if, and only if, $x^* \in X$, $\mathcal{A}(x^*) \in \mathcal{K}^*$ and $f(x^*) \leq f^*$. This observation leads us to our first definition of a near optimal solution $\tilde{x} \in X$ of (1), which essentially requires the primal infeasibility measure $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}))$ and the primal optimality gap $[f(\tilde{x}) - f^*]^+$ to be both small.

Definition 1 *For a given pair $(\epsilon_p, \epsilon_o) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, $\tilde{x} \in X$ is called an (ϵ_p, ϵ_o)-primal solution for (1) if*

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p \text{ and } f(\tilde{x}) - f^* \leq \epsilon_0. \tag{5}$$

One drawback of the above notion of near optimality of $\tilde{x}$ is that it says nothing about the size of $[f(\tilde{x}) - f^*]^-$. We will now show that this quantity can be bounded as $[f(\tilde{x}) - f^*]^- \leq \epsilon_p\|\lambda^*\|$, where λ^* is an arbitrary Lagrange multiplier for (1). We start by stating the following result whose proof is given in the Appendix.

Proposition 1 *Let $\mathcal{K} \subseteq \Re^m$ be a closed convex cone. Then, the following statements hold:*

a) $d_{\mathcal{K}^*} = \sigma_C$, where $C := (-\mathcal{K}) \cap B(0,1)$, where $B(0,1) := \{u \in \Re^m : \|u\| \leq 1\}$;

b) for every $u \in \Re^m$ and $\lambda \in \mathcal{K}$, we have $\langle u, \lambda \rangle \geq -\|\lambda\| \, d_{\mathcal{K}^*}(u)$.

As a consequence of the above result, we obtain the following technical inequality which will be frequently used in our analysis.

Corollary 2 *Let λ^* be an Lagrange multiplier for (1). Then, for every $x \in X$, we have $f(x) - f^* \geq -\|\lambda^*\| \, d_{\mathcal{K}^*}(\mathcal{A}(x))$.*

Proof. It is well-known that our assumptions imply that v is a convex function such that $\lambda^* \in \partial v(0)$, and hence that

$$v(u) - v(0) \geq \langle \lambda^*, u \rangle = \langle -\lambda^*, -u \rangle \geq -\| -\lambda^* \| \, d_{\mathcal{K}^*}(-u) = -\|\lambda^*\| \, d_{\mathcal{K}^*}(-u), \ \forall u \in \Re^m,$$

where the inequality follows from Proposition 1(b) and the fact that $-\lambda^* \in \mathcal{K}$. Now, let $x \in X$ be given. Since x is clearly feasible for problem (4) with $u = -\mathcal{A}(x)$, the definition of $v(\cdot)$ in (4) implies that $v(u) \leq f(x)$. Hence,

$$f(x) - f^* \geq v(u) - v(0) \geq -\|\lambda^*\| \, d_{\mathcal{K}^*}(-u) = -\|\lambda^*\| \, d_{\mathcal{K}^*}(\mathcal{A}(x)).$$

■

Corollary 3 *If $\tilde{x} \in X$ is an ϵ_p-feasible solution, i.e., if $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p$, then $f(\tilde{x}) - f^* \geq -\epsilon_p \, \|\lambda^*\|$, where λ^* is an arbitrary Lagrange multiplier for (1).*

Another way of defining a near optimal solution of (1) is based on the following optimality conditions: $x^* \in X$ is an optimal solution of (1) and $\lambda^* \in -\mathcal{K}$ is a Lagrange multiplier for (1) if, and only if, $(\tilde{x}, \tilde{\lambda}) = (x^*, \lambda^*)$ satisfies

$$\begin{aligned} &\mathcal{A}(\tilde{x}) \in \mathcal{K}^*, \quad \langle \tilde{\lambda}, \mathcal{A}(\tilde{x}) \rangle = 0, \\ &\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \, \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}), \end{aligned} \tag{6}$$

where $\mathcal{N}_X(\tilde{x}) := \{s \in \Re^n : \langle s, x - \tilde{x} \rangle \leq 0, \, \forall x \in X\}$ denotes the normal cone of X at $\tilde{x}$. Based on this observation, we can introduce our second definition of a near optimal solution of (1).

Definition 2 *For a given pair $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$, $(\tilde{x}, \tilde{\lambda}) \in X \times (-\mathcal{K})$ is called an (ϵ_p, ϵ_d)-primal-dual solution of (1) if*

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p, \quad \langle \tilde{\lambda}, \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \rangle = 0, \tag{7}$$

$$\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \, \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}) + \mathcal{B}(\epsilon_d), \tag{8}$$

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

In Sections 4 and 5, we will study the iteration-complexities of well-known penalty methods, namely the quadratic and the exact penalty methods, for computing the two types of near optimal solutions defined in this section.

3 Technical results

This section discusses some technical results that will be used in our analysis and reviews two first order algorithms due to Nesterov [2, 4] for solving special classes of CP problems. It consists of three subsections. The first one develops several technical results involving projected gradients. The second subsection reviews Nesterov's optimal method for solving a class of smooth CP problems. The third subsection reviews Nesterov's smooth approximation scheme for solving a class of non-smooth CP problems that can be reformulated as smooth convex-concave saddle point (i.e., min-max) problems.

3.1 Projected gradient and the optimality conditions

In this subsection, we consider an inner product space $\Re^n$ endowed with the norm $\|\cdot\|$ associated with its inner product. We consider the CP problem

$$\phi^* := \min_{x \in X} \phi(x), \tag{9}$$

where $X \subset \Re^n$ is a convex set and $\phi : X \to \mathbf{R}$ is a convex function that has L_ϕ-Lipschitz-continuous gradient over X with respect to the norm $\|\cdot\|$.

It is well-known that $x^* \in X$ is an optimal solution of (9) if and only if $\nabla\phi(x^*) \in -\mathcal{N}_X(x^*)$. Moreover, this optimality condition is in turn related to the projected gradient of the function ϕ over X defined as follows.

Definition 3 *Given a fixed constant $\tau > 0$, we define the projected gradient of ϕ at $\tilde{x} \in X$ with respect to X as (see, for example, [3])*

$$\nabla\phi(\tilde{x})]_X^\tau := \frac{1}{\tau}\left[\tilde{x} - \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))\right], \tag{10}$$

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

The following proposition relates the projected gradient to the aforementioned optimality condition.

Proposition 4 *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 $\nabla\phi(\tilde{x}^+) \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}\left((1 + \tau L_\phi)\epsilon\right)$.

Proof. To simplify notation, define $v := \nabla\phi(\tilde{x})]_X^\tau$. By (10) and well-known properties of the projection operator Π_X, we have

$$\begin{aligned}
v = \nabla\phi(\tilde{x})]_X^\tau \quad &\Leftrightarrow \quad \tilde{x} - \tau v = \Pi_X\left(\tilde{x} - \tau\nabla\phi(\tilde{x})\right) \\
&\Leftrightarrow \quad \langle \tilde{x} - \tau\nabla\phi(\tilde{x}) - (\tilde{x} - \tau v)\,,\, y - (\tilde{x} - \tau v)\rangle \leq 0, \quad \forall y \in X \\
&\Leftrightarrow \quad \langle v - \nabla\phi(\tilde{x}), y - (\tilde{x} - \tau v)\rangle \leq 0, \quad \forall y \in X \\
&\Leftrightarrow \quad \langle v - \nabla\phi(\tilde{x}), y - \tilde{x}^+\rangle \leq 0, \quad \forall y \in X \\
&\Leftrightarrow \quad \nabla\phi(\tilde{x}) - v \in -\mathcal{N}_X(\tilde{x}^+).
\end{aligned}$$

from which statement a) clearly follows. To show b), assume that $\|v\| \leq \epsilon$. Then, we have

$$\|\tilde{x}^+ - \tilde{x}\| \leq \|\Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x})) - \tilde{x}\| = \|(\tilde{x} - \tau v) - \tilde{x}\| = \tau\|v\| \leq \tau\epsilon.$$

which, together with the assumption that ϕ has L_ϕ-Lipschitz-continuous gradient, implies that $\|\nabla\phi(\tilde{x}^+) - \nabla\phi(\tilde{x})\| \leq \tau L_\phi \epsilon$. Statement b) now follows immediately from the latter conclusion and statement a). ∎

The following lemma summarizes some interesting properties of the projected gradient.

Lemma 5 *Let $\tilde{x} \in X$ be given and $\tilde{x}^+ := \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))$. Denoting $g(\cdot) := \nabla\phi(\cdot)$, $g_X(\cdot) := \nabla\phi(\cdot)]_X^\tau$. We have:*

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

b) For any $x \in X$, there holds

$$\langle g(\tilde{x}) - g_X(\tilde{x}), x - \tilde{x}^+\rangle \geq 0; \tag{11}$$

in particular, setting $x = \tilde{x}$, we obtain

$$\langle g(\tilde{x}) - g_X(\tilde{x}), g_X(\tilde{x})\rangle \geq 0; \tag{12}$$

c) For any $x \in X$, there holds

$$\phi(\tilde{x}^+) - \phi(x) \leq (1 + \tau L_\phi)\,\|g_X(\tilde{x})\|\,\|\tilde{x}^+ - x\|. \tag{13}$$

d) If $\tau = 1/L_\phi$ in definition (10), then

$$\phi(x) - \phi(x^*) \geq \frac{1}{2L_\phi}\|g_X(x)\|^2,\ \forall\, x \in X, \tag{14}$$

where $x^* \in \mathrm{Argmin}_{x\in X}\phi(x)$.

Proof. a) This statement is proved in p. 87 of [3].

b) Noting that $\tilde{x}^+ = \tilde{x} - \tau g_X(\tilde{x}) = \Pi_X(\tilde{x} - \tau g(\tilde{x}))$, it follows from well-known properties of the projection map Π_X that

$$\begin{aligned}\langle x - (\tilde{x} - \tau g_X), \tilde{x} - \tau g(\tilde{x}) - (\tilde{x} - \tau g_X(\tilde{x}))\rangle &= \langle x - (\tilde{x} - \tau g_X(\tilde{x})), \tau(g_X(\tilde{x}) - g(\tilde{x}))\rangle \\ &= \langle x - \tilde{x}^+, \tau(g_X(\tilde{x}) - g(\tilde{x}))\rangle \leq 0,\ \forall\, x \in X,\end{aligned}$$

which clearly implies statement b).

c) It follows from from the convexity of $\phi(\cdot)$, (11), the assumption that $\phi(\cdot)$ has L_ϕ-Lipschitz-continuous gradient, and definition (10) that

$$\begin{aligned}\phi(\tilde{x}^+) - \phi(x) &\leq \langle g(\tilde{x}^+), \tilde{x}^+ - x\rangle \\ &= \langle g(\tilde{x}) - g_X(\tilde{x}), \tilde{x}^+ - x\rangle + \langle g_X(\tilde{x}), \tilde{x}^+ - x\rangle + \langle g(\tilde{x}^+) - g(\tilde{x}), \tilde{x}^+ - x\rangle \\ &\leq \langle g_X(\tilde{x}), \tilde{x}^+ - x\rangle + \langle g(\tilde{x}^+) - g(\tilde{x}), \tilde{x}^+ - x\rangle \\ &\leq \langle g_X(\tilde{x}), \tilde{x}^+ - x\rangle + L_\phi\|\tilde{x}^+ - \tilde{x}\|\,\|\tilde{x}^+ - x\| \\ &= \langle g_X(\tilde{x}), \tilde{x}^+ - x\rangle + \tau L_\phi\|g_X(\tilde{x})\|\,\|\tilde{x}^+ - x\| \\ &\leq (1 + \tau L_\phi)\,\|g_X(\tilde{x})\|\,\|\tilde{x}^+ - x\|,\ \forall\, x \in X.\end{aligned}$$

d) Using the fact that $\Phi(x^*) \leq \Phi(x)$, $\forall x \in X$ and the assumption that $\phi(\cdot)$ has L_ϕ-Lipschitz-continuous gradient, we conclude that

$$\begin{aligned} \phi(x^*) &\leq \phi\left(x - \frac{1}{L_\phi} g_X(x)\right) \leq \phi(x) - \frac{1}{L_\phi}\langle g(x), g_X(x)\rangle + \frac{1}{2L_\phi}\|g_X(x)\|^2 \\ &\leq \phi(x) - \frac{1}{2L_\phi}\|g_X(x)\|^2 \end{aligned}$$

for any $x \in X$, where the last inequality follows from (12). ∎

3.2 Nesterov's Optimal Method

In this subsection, we discuss Nesterov's smooth first-order method for solving a class of smooth CP problems.

Our problem of interest is still the CP problem (9), which is assumed to satisfy the same assumptions mentioned in Subsection 3.1, except that now we allow $\|\cdot\|$ to be an arbitrary norm. Moreover, we assume throughout our discussion that the optimal value ϕ^* of problem (9) is finite and that its set of optimal solutions is nonempty. Let $h : X \to \mathbf{R}$ be a differentiable strongly convex function with modulus $\sigma_h > 0$ with respect to $\|\cdot\|$, i.e.,

$$h(x) \geq h(\tilde{x}) + \langle \nabla h(\tilde{x}), x - \tilde{x}\rangle + \frac{\sigma_h}{2}\|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{15}$$

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

$$d_h(x; \tilde{x}) \equiv h(x) - l_h(x; \tilde{x}), \quad \forall x, \tilde{x} \in X, \tag{16}$$

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

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

We are now ready to state Nesterov's smooth first-order method for solving (9). We use the superscript 'sd' in the sequence obtained by taking a steepest descent step and the superscript 'ag' (which stands for 'aggregated gradient') in the sequence obtained by using all past gradients.

Nesterov's Algorithm:

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

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

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

$$x_{k+1}^{sd} \in \operatorname{Argmin}\left\{ l_\phi(x; x_k) + \frac{L_\phi}{2}\|x - x_k\|^2 : x \in X \right\}, \tag{17}$$

$$x_{k+1}^{ag} \equiv \operatorname{argmin}\left\{ \frac{L_\phi}{\sigma_h} d_h(x; x_0) + \sum_{i=0}^{k} \frac{i+1}{2}\left[l_\phi(x; x_i)\right] : x \in X \right\}. \tag{18}$$

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

end

The main convergence result established by Nesterov [4] regarding the above algorithm is sum-

marized in the following theorem.

Theorem 6 *The sequence $\{x_k^{sd}\}$ generated by Nesterov's optimal method satisfies*

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

where x^* is an optimal solution of (9). As a consequence, given any $\epsilon > 0$, an iterate $x_k^{sd} \in X$ satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ can be found in no more than

$$\left\lceil\, 2\sqrt{\frac{d_h(x^*; x_0^{sd})L_\phi}{\sigma_h \epsilon}}\, \right\rceil \tag{19}$$

iterations.

The following result is as an immediate special case of Theorem 6.

Corollary 7 *Suppose that $\|\cdot\|$ is a inner product norm and $h : X \to \Re$ is chosen as $h(\cdot) = \|\cdot\|^2/2$ in Nesterov's optimal method. Then, for any $\epsilon > 0$, an iterate $x_k^{sd} \in X$ satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ can be found in no more than*

$$\left\lceil\, \|x_0^{sd} - x^*\| \sqrt{\frac{2L_\phi}{\epsilon}}\, \right\rceil \tag{20}$$

iterations, where x^* is an optimal solution of (9).

Proof. If $h(x) = \|x\|^2/2$, then (16) implies that $d_h(x^*; x_0^{sd}) = \|x_0^{sd} - x^*\|^2/2$. The corollary now follows from this bound and relation (19). ∎

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

$$\langle \nabla\phi(x) - \nabla\phi(\tilde{x}), x - \tilde{x}\rangle \geq \mu\|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{21}$$

Nesterov shows in Theorem 2.2.2 of [3] that, under the assumptions of Corollary 7, a variant of his optimal method finds a solution $x_k \in X$ satisfying $\phi(x_k) - \phi^* \leq \epsilon$ in no more than

$$\left\lceil\, \sqrt{\frac{L_\phi}{\mu}}\, \log \frac{L_\phi\, \|x_0^{sd} - x^*\|^2}{\epsilon}\, \right\rceil \tag{22}$$

iterations. The following result gives an iteration-complexity bound for Nesterov's optimal method that replaces the term $\log(L_\phi\|x_0^{sd} - x^*\|^2/\epsilon)$ in (22) with $\log(\mu\|x_0^{sd} - x^*\|^2/\epsilon)$. The resulting iteration-complexity bound is not only sharper but also more suitable since it makes it easier for us to compare the quality of the different bounds obtained in our analysis of first-order penalty methods.

Theorem 8 *Let $\epsilon > 0$ be given and suppose that the assumptions of Corollary 7 hold and that the function ϕ is strongly convex with modulus μ. Then, the variant where we restart Nesterov's optimal method, with proximal function $h(\cdot) = \|\cdot\|^2/2$, every*

$$K := \left\lceil\, \sqrt{\frac{8L_\phi}{\mu}}\, \right\rceil \tag{23}$$

iterations finds a solution $\tilde{x} \in X$ *satisfying* $\phi(\tilde{x}) - \phi^* \leq \epsilon$ *in no more than*

$$\left\lceil \sqrt{\frac{8\, L_\phi}{\mu}} \right\rceil \max\{1, \lceil \log Q \rceil\} \tag{24}$$

iterations, where

$$Q := \frac{\mu \, \|x_0^{sd} - x^*\|^2}{2\epsilon} \tag{25}$$

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

Proof. Denote $\mathcal{D} := \|x_0^{sd} - x^*\|$. First consider the case where $Q \leq 1$. Clearly we have $\epsilon \geq \mu \mathcal{D}^2/2$, which, in view of Corollary 7, implies that the number of iterations is bounded by $\lceil 2\sqrt{L_\phi/\mu} \rceil$ and hence bounded by (24).

We now show that bound (24) holds for the case when $Q > 1$. Let x^j be the iterate obtained at the end of $(j-1)$-th restart after K iterations of Nesterov's optimal method are performed. By Theorem 6 with $h(\cdot) = \|\cdot\|^2/2$ and inequality (21), we have

$$\begin{aligned} \phi(x^1) - \phi(x^*) &\leq \frac{2L_\phi \, \|x_0^{sd} - x^*\|^2}{K^2} \leq \frac{2L_\phi \, \mathcal{D}^2}{K^2} \\ \phi(x^j) - \phi(x^*) &\leq \frac{2L_\phi \, \|x^{j-1} - x^*\|^2}{K^2} \leq \frac{4L_\phi}{\mu \, K^2} \left[\phi(x^{j-1}) - \phi(x^*)\right], \ \forall \, j \geq 2. \end{aligned}$$

Using the above relations inductively, we conclude that

$$\phi(x^j) - \phi(x^*) \leq \frac{(4L_\phi)^j \, \mathcal{D}^2}{2\mu^{j-1} K^{2j}}.$$

Setting $j = \lceil \log Q \rceil$ and observing that

$$K^{2j} \geq \left[\left(\frac{8L_\phi}{\mu} \right)^{\frac{1}{2}} \right]^{2j} = 2^j \left(\frac{4L_\phi}{\mu} \right)^j \geq 2^{\log Q} \left(\frac{4L_\phi}{\mu} \right)^j \geq \frac{\mu \mathcal{D}^2}{2\epsilon} \frac{(4L_\phi)^j}{\mu^j} = \frac{\mathcal{D}^2}{2\epsilon} \frac{(4L_\phi)^j}{\mu^{j-1}},$$

we conclude that $\phi(x^j) - \phi(x^*) \leq \epsilon$. Hence, the overall number of iterations is bounded by $K \lceil \log Q \rceil$, or equivalently, by (24). ∎

It is interesting to compare the complexity bounds of Corollary 7 and Theorem 8. Indeed, it can be easily seen that there exists a threshold value $\bar{Q} > 0$ such that the condition $Q \geq \bar{Q}$, where Q is defined in (25), implies that the bound (20) is always greater than or equal to the bound (24). Moreover, as Q goes to infinity, the ratio between (20) and (24) converges to infinity. Hence, when the function ϕ is strongly convex and Q satisfies (25), we will always use the bound (24) in our analysis.

3.3 Nesterov's smooth approximation scheme

In this subsection, we discuss Nesterov's smooth approximation scheme for solving a class of non-smooth CP problems which can be reformulated as special smooth convex-concave saddle point (i.e., min-max) problems.

Our problem of interest in this subsection is still problem (9). We assume that $X \subseteq \Re^n$ is a closed convex set and $\phi : X \to \mathbf{R}$ is a convex function given by

$$\phi(x) := \hat{\phi}(x) + \sup\{\langle \mathcal{E}x, y\rangle - \psi(y) : y \in Y\}, \quad \forall x \in X, \tag{26}$$

where $\hat{\phi} : X \to \mathbf{R}$ is a convex function with $L_{\hat{\phi}}$-Lipschitz-continuous gradient with respect to an arbitrary norm $\|\cdot\|$ on $\Re^n$, $Y \subseteq \Re^m$ is a compact convex set, $\phi : Y \to \mathbf{R}$ is a continuous convex function, and $\mathcal{E} : \Re^n \to \Re^m$ is an affine operator.

Unless stated explicitly otherwise, we assume that the CP problem (9) referred to in this subsection is the one with its objective function given by (26). Also, we assume throughout our discussion that f^* is finite and that the set of optimal solutions of (9) is nonempty.

The function ϕ defined as in (26) is generally non-differentiable but can be closely approximated by a function with Lipschitz-continuous gradient using the following construction due to Nesterov [4]. Let $\tilde{h} : Y \to \mathbf{R}$ be a continuous strongly convex function with modulus $\sigma_{\tilde{h}} > 0$ with respect to an arbitrary norm $\|\cdot\|$ on $\Re^m$ satisfying $\min\{\tilde{h}(y) : y \in Y\} = 0$. For some *smoothness* parameter $\eta > 0$, consider the following function

$$\phi_\eta(x) = \hat{\phi}(x) + \max_y \left\{ \langle \mathcal{E}x, y\rangle - \psi(y) - \eta\tilde{h}(y) : \ y \in Y \right\}. \tag{27}$$

The next result, due to Nesterov [4], shows that ϕ_η is a function with Lipschitz-continuous gradient with respect to $\|\cdot\|$, whose "closeness" to ϕ depends linearly on the parameter η.

Proposition 9 *The following statements hold:*

a) For every $x \in X$, we have $\phi_\eta(x) \le \phi(x) \le \phi_\eta(x) + \eta D_{\tilde{h}}$, where

$$D_{\tilde{h}} := \max\left\{\tilde{h}(y) : y \in Y\right\}; \tag{28}$$

b) The function $\phi_\eta(x)$ has L_η-Lipschitz-continuous gradient with respect to $\|\cdot\|$, where

$$L_\eta := L_{\hat{\phi}} + \frac{\|\mathcal{E}\|^2}{\eta\sigma_{\tilde{h}}}. \tag{29}$$

We are now ready to state Nesterov's smooth approximation scheme to solve (9).

Nesterov's smooth approximation scheme:

0) Let $\epsilon > 0$ be given.

1) Let $\eta := \epsilon/(2D_{\tilde{h}})$ and consider the approximation problem

$$\phi_\eta^* \equiv \inf\{\phi_\eta(x) : x \in X\}. \tag{30}$$

2) Apply Nesterov's optimal method (or its variant) to (30) and terminate whenever an iterate x_k^{sd} satisfying $\phi(x_k^{sd}) - \phi^* \le \epsilon$ is found.

end

The following result describes the convergence behavior of the above scheme.

Theorem 10 *Nesterov's smooth approximation scheme generates an iterate x_k^{sd} satisfying $\phi(x_k^{sd}) - \phi^* \leq \epsilon$ in no more than*

$$\left\lceil \sqrt{\frac{8D_h(x_0^{sd})}{\sigma_h \epsilon}\left(\frac{2D_{\tilde{h}}\|\mathcal{E}\|^2}{\sigma_{\tilde{h}}\epsilon} + L_{\hat{\phi}}\right)} \right\rceil \tag{31}$$

iterations, where $D_h(x_0^{sd}) := \max_{x\in X} d_h(x; x_0^{sd})$ and $D_{\tilde{h}}$ is defined in (28).

4 The quadratic penalty method

The goal of this section is to establish the iteration-complexity (in terms of Nesterov's optimal method iterations) of the quadratic penalty method for solving (1). Specifically, we present the iteration-complexity of the classical quadratic penalty method in Subsection 4.1, and a variant of this method, for which a perturbation term is added into the objective function of (1), is discussed and analyzed in Subsection 4.2.

The basic idea underlying penalty methods is rather simple, namely: instead of solving problem (1) directly, we solve certain relaxations of (1) obtained by penalizing some violation of the constraint $\mathcal{A}(x) \in \mathcal{K}^*$. More specifically, in the case of the quadratic penalty method, given a penalty parameter $\rho > 0$, we solve the relaxation

$$\Psi_\rho^* := \inf_{x\in X}\left\{\Psi_\rho(x) := f(x) + \frac{\rho}{2}\left[d_{\mathcal{K}^*}(\mathcal{A}(x))\right]^2\right\}. \tag{32}$$

In order to guarantee that the function Ψ_ρ is smooth, we assume throughout this section that the distance function $d_{\mathcal{K}^*}$ is defined with respect to an inner product norm $\|\cdot\|$ in $\Re^m$. Note that in this case, we have $\|\cdot\|_* = \|\cdot\|$.

We will now see that the objective function Ψ_ρ of (32) has Lipschitz continuous gradient. We first state the following well-known result which guarantees that the distance function has Lipschitz continuous gradient (see for example Proposition 15 of [1] for its proof).

Proposition 11 *Given a closed convex set $\mathcal{C} \subseteq \Re^m$, consider the distance function $d_{\mathcal{C}} : \Re^m \to \mathsf{R}$ to $\mathcal{C}$ with respect some inner product norm $\|\cdot\|$ on $\Re^m$. Then, the function $\psi : \Re^m \to \Re$ defined as $\psi(u) = [d_{\mathcal{C}}(u)]^2$ is convex and its gradient is given by $\nabla\psi(u) = 2[u - \Pi_C(u)]$ for every $u \in \Re^m$. Moreover, $\|\nabla\psi(\tilde{u}) - \nabla\psi(\tilde{u})\| \leq 2\|\tilde{u} - u\|$ for every $u, \tilde{u} \in \Re^m$.*

As an immediate consequence of Proposition 11, we obtain the following result.

Corollary 12 *The function Ψ_ρ has M_ρ-Lipschitz continuous gradient where $M_\rho := L_f + \rho\|\mathcal{A}\|^2$.*

Proof. The differentiability of Ψ_ρ follows immediately from the assumption that f is differentiable and Proposition 11. Moreover, it easily follows from the chain rule that $\nabla\Psi_\rho(x) = \nabla f(x) + \rho\mathcal{A}_0^*\nabla(d_{\mathcal{K}}^2)(\mathcal{A}(x))/2$, which together with (2) and Proposition 11, then imply that

$$\begin{aligned}
\|\nabla\Psi_\rho(x_1) - \nabla\Psi_\rho(x_2)\|_* &\leq \|\nabla f(x_1) - \nabla f(x_2)\|_* + \frac{\rho}{2}\|\mathcal{A}_0^*\nabla(d_{\mathcal{K}}^2)(\mathcal{A}(x_1)) - \mathcal{A}_0^*\nabla(d_{\mathcal{K}}^2)(\mathcal{A}(x_2))\|_* \\
&\leq L_f\|x_1 - x_2\| + \frac{\rho}{2}\|\mathcal{A}_0^*\|\|\nabla d_{\mathcal{K}}^2(\mathcal{A}(x_1)) - \nabla d_{\mathcal{K}}^2(\mathcal{A}(x_2))\| \\
&\leq L_f\|x_1 - x_2\| + \rho\|\mathcal{A}_0^*\|\|\mathcal{A}_0(x_1 - x_2)\| \\
&\leq L_f\|x_1 - x_2\| + \rho\|\mathcal{A}_0^*\|\|\mathcal{A}_0\|\|x_1 - x_2\| \;=\; M_\rho\|x_1 - x_2\|,
\end{aligned}$$

for every $x_1, x_2 \in X$, where the last equality follows from the fact that $\|\mathcal{A}\| = \|\mathcal{A}_0\| = \|\mathcal{A}_0^*\|$. ∎

Therefore, problem (32) can be solved, for example, by Nesterov's optimal method or one of its variants (see for example [1, 2, 4]).

4.1 Quadratic penalty method applied to the original problem

In this subsection, we consider the application of the quadratic penalty method directly to the original problem (1), which consists of solving the penalized problem (32) for a suitably chosen penalty parameter ρ. This subsection contains two subsubsections. The first one considers the complexity of obtaining an (ϵ_p, ϵ_o)-primal solution of (1) while the second one discusses the complexity of computing an (ϵ_p, ϵ_d)-primal-dual solution of (1).

4.1.1 Computation of a primal approximate solution

In this subsection, we are interested in deriving the iteration-complexity involved in the computation of an (ϵ_p, ϵ_o)-primal solution of problem (1) by approximately solving the penalized problem (32) for some value of the penalty parameter ρ.

Given an approximate solution $\tilde{x} \in X$ for the penalized problem (32), the following result provides bounds on the primal infeasibility and optimality gap of $\tilde{x}$ with respect to (1).

Proposition 13 *If $\tilde{x} \in X$ is a δ-approximate solution of (32), i.e., it satisfies*

$$\Psi_\rho(\tilde{x}) - \Psi_\rho^* \leq \delta, \tag{33}$$

then

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \frac{2}{\rho}\|\lambda^*\| + \sqrt{\frac{2\delta}{\rho}} \tag{34}$$

$$f(\tilde{x}) - f^* \leq \delta, \tag{35}$$

where λ^* is an arbitrary Lagrange multiplier associated with (1).

Proof. Using the fact that $v(0) = f^* \geq \Psi_\rho^*$, Corollary 2 and assumption (33), we conclude that

$$\begin{aligned}\delta &\geq \Psi_\rho(\tilde{x}) - \Psi_\rho^* = f(\tilde{x}) + \frac{\rho}{2}[d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}))]^2 - \Psi_\rho^* \\ &\geq f(\tilde{x}) - f^* + \frac{\rho}{2}[d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}))]^2 \geq -\|\lambda^*\|\, d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) + \frac{\rho}{2}[d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}))]^2,\end{aligned}$$

which clearly implies both (34) and (35). ∎

For the sake of future reference, we state the following simple result.

Lemma 14 *Let α_i, $i = 0, 1, 2$, be given positive constants. Then, the only positive scalar ρ satisfying the equation $\alpha_2\rho^{-1} + \alpha_1\rho^{-1/2} = \alpha_0$ is given by*

$$\rho = \left(\frac{\alpha_1 + \sqrt{\alpha_1^2 + 4\alpha_0\alpha_2}}{2\alpha_0}\right)^2.$$

With the aid of Proposition 13, we can now derive an iteration-complexity bound for Nesterov's method applied to the quadratic penalty problem (32) to compute an (ϵ_p, ϵ_o)-primal solution of (1).

Theorem 15 *Let λ^* be an arbitrary Lagrange multiplier for (1) and let $(\epsilon_p, \epsilon_o) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. If*

$$\rho = \rho_p(t) := \left(\frac{\sqrt{\epsilon_o} + \sqrt{\epsilon_o + 4\epsilon_p\, t}}{\sqrt{2}\epsilon_p} \right)^2 \tag{36}$$

for some $t \geq \|\lambda^*\|$, then Nesterov's optimal method applied to problem (32) finds an (ϵ_p, ϵ_o)-primal solution of (1) in at most

$$N_p(t) := \left\lceil 2\sqrt{\frac{D_h(x_0^{sd})}{\sigma_h}} \left\{ \sqrt{\frac{L_f}{\epsilon_o}} + \frac{\sqrt{2}\|\mathcal{A}\|}{\epsilon_p} + \|\mathcal{A}\| \sqrt{\frac{2t}{\epsilon_p \epsilon_o}} \right\} \right\rceil, \tag{37}$$

iterations, where $D_h(x_0^{sd})$ is defined in Theorem 10. In particular, if $\|\cdot\|$ is the inner product norm on $\Re^n$ and $h(\cdot) = \|\cdot\|^2/2$, then the above bound reduces to

$$N_p(t) := \left\lceil \sqrt{2} D_X \left\{ \sqrt{\frac{L_f}{\epsilon_o}} + \frac{\sqrt{2}\|\mathcal{A}\|}{\epsilon_p} + \|\mathcal{A}\| \sqrt{\frac{2t}{\epsilon_p \epsilon_o}} \right\} \right\rceil, \tag{38}$$

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

Proof. Let $\tilde{x} \in X$ satisfies (33) with $\delta = \epsilon_o$. Proposition 13, the assumption that $t \geq \|\lambda^*\|$, relation (36) and Lemma 14 with $\alpha_0 = \epsilon_p$, $\alpha_1 = \sqrt{\epsilon_o}$ and $\alpha_2 = t$ then imply that $f(\tilde{x}) - f^* \leq \epsilon_o$ and

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \frac{2}{\rho_p(t)} \|\lambda^*\| + \sqrt{2\frac{\epsilon_o}{\rho_p(t)}} \;\leq\; \frac{2t}{\rho_p(t)} + \sqrt{\frac{2\epsilon_o}{\rho_p(t)}} \;=\; \epsilon_p,$$

and hence that $\tilde{x}$ is an (ϵ_p, ϵ_o)-primal solution of (1). In view of Corollary 7, Nesterov's optimal method finds an approximate solution $\tilde{x}$ as above in at most

$$\left\lceil 2\sqrt{\frac{D_h(x_0^{sd})\, M_\rho}{\sigma_h \epsilon_0}} \right\rceil$$

iterations, where $M_\rho := L_f + \rho\|\mathcal{A}\|^2$. Substituting the value of ρ given by (36) in this iteration bound and using some trivial majorization, we obtain bound (37). ∎

We now make a few observations regarding Theorem 15. First, the choice of ρ given by (36) requires that $t \geq \|\lambda^*\|$ so as to guarantee that an ϵ_o-approximate solution $\tilde{x}$ of (32) satisfies $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p$. Second, note that the iteration-complexity $N_p(t)$ obtained in Theorem 15 achieves its minimum possible value over the interval $t \geq \|\lambda^*\|$ exactly when $t = \|\lambda^*\|$. However, since the quantity $\|\lambda^*\|$ is not known a priori, it is necessary to use a "guess and check" procedure for t so as to develop a scheme for computing an (ϵ_p, ϵ_o)-primal solution of (1) whose iteration-complexity has the same order of magnitude as the ideal one in which $t = \|\lambda^*\|$.

We now describe the aforementioned "guess and check" procedure for t.

Search Procedure 1:

1) Set $k = 0$ and define

$$\beta_0 := 2\sqrt{\frac{D_h(x_0^{sd})}{\sigma_h}}\left(\sqrt{\frac{L_f}{\epsilon_o}} + \frac{\sqrt{2}\|\mathcal{A}\|}{\epsilon_p}\right), \quad \beta_1 := 2\|\mathcal{A}\|\sqrt{\frac{2D_h(x_0^{sd})}{\sigma_h\epsilon_p\epsilon_o}}, \quad t_0 := \left[\frac{\max(1,\beta_0)}{\beta_1}\right]^2. \tag{39}$$

2) Set $\rho = \rho_p(t_k)$, and perform at most $\lceil N_p(t_k)\rceil$ iterations of Nesterov's optimal method applied to problem (32). If an iterate $\tilde{x}$ is obtained such that (33) with $\delta = \epsilon_o$ and $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p$ are satisfied, then **stop**; otherwise, go to step 3;

3) Set $t_{k+1} = 2t_k$, $k = k + 1$, and go to step 2.

Before establishing the iteration-complexity of the above procedure, we state a technical result, namely Lemma 17, which will be used more than once in the paper. Lemma 16 states a simple inequality which is also used in a few times in our presentation.

Lemma 16 *For any scalars $\tau \geq 0$, $x > 0$, and $\alpha \geq 0$, we have $\tau x + \alpha \leq (\tau + \alpha)\lceil x\rceil$.*

Lemma 17 *Let a positive scalar p be given. Then, there exists a constant $C = C(p)$ such that for any positive scalars $\beta_1, \beta_2, \bar{t}$, we have*

$$\sum_{k=0}^{K}\lceil\beta_1 + \beta_2 t_k^p\rceil \leq C\lceil\beta_1 + \beta_2\bar{t}^p\rceil, \tag{40}$$

where $t_k = t_0 2^k$ for $k = 1, \ldots, K$, and

$$t_0 := \left(\frac{\max(\beta_1, 1)}{\beta_2}\right)^{1/p}, \quad K := \max\left\{0, \left\lceil\log\left(\frac{\bar{t}}{t_0}\right)\right\rceil\right\}. \tag{41}$$

Proof. Assume first that $\bar{t} \leq t_0$. Due to the definition of K in (41), we have $K = 0$ in this case, and hence that

$$\sum_{k=0}^{K}\lceil\beta_1 + \beta_2 t_k^p\rceil \;=\; \lceil\beta_1 + \beta_2 t_0^p\rceil \;=\; \lceil\beta_1 + \max(\beta_1, 1)\rceil \;\leq\; 2\beta_1 + 2 \;\leq\; 4\lceil\beta_1\rceil \;\leq\; 4\lceil\beta_1 + \beta_2\bar{t}^p\rceil,$$

where the second equality follows from the definition of t_0. Hence, in the case where $\bar{t} \leq t_0$, inequality (40) holds with $C = 4$.

Assume now that $\bar{t} > t_0$. By the definition of K in (41), we have $K = \lceil\log(\bar{t}/t_0)\rceil$, from which we conclude that $K < \log(\bar{t}/t_0) + 1$, and hence that $t_0 2^{K+1} < 4\bar{t}$. Using these relations, the inequality

$\log x = (\log x^p)/p \leq x^p/p$ for any $x > 0$, and the definition of t_0 in (41), we obtain

$$\begin{aligned}
\sum_{k=0}^{K} \lceil \beta_1 + \beta_2 t_k^p \rceil &\leq \sum_{k=0}^{K} 1 + \beta_1 + \beta_2\, t_0^p\, 2^{pk} \;\leq\; (1+\beta_1)(1+K) + \beta_2 t_0^p \frac{2^{(K+1)p}}{2^p - 1} \\
&\leq (1+\beta_1)\left[2 + \log\left(\frac{\bar{t}}{t_0}\right)\right] + \beta_2 \frac{(4\bar{t})^p}{2^p - 1} \\
&\leq (1+\beta_1)\left[2 + \frac{1}{p}\left(\frac{\bar{t}}{t_0}\right)^p\right] + \frac{4^p}{2^p - 1}\beta_2 \bar{t}^p \\
&\leq (1+\beta_1)\left[2 + \frac{1}{p}\left(\frac{\beta_2 \bar{t}^p}{\max(\beta_1, 1)}\right)\right] + \frac{4^p}{2^p - 1}\beta_2 \bar{t}^p \\
&\leq 2(1+\beta_1) + \frac{2}{p}\beta_2 \bar{t}^p + \frac{4^p}{2^p - 1}\beta_2 \bar{t}^p \\
&\leq 2 + \max\left\{2, \frac{2}{p} + \frac{4^p}{2^p - 1}\right\}(\beta_1 + \beta_2 \bar{t}^p),
\end{aligned}$$

which, in view of Lemma 16, implies that (40) holds with

$$C = C(p) := 2 + \max\left\{2, \frac{2}{p} + \frac{4^p}{2^p - 1}\right\}.$$

■

The following result gives the iteration-complexity of Search Procedure 1 for obtaining an (ϵ_p, ϵ_o)-primal solution of (1).

Corollary 18 *Let λ^* be the minimum norm Lagrange multiplier for (1). Then, the overall number of iterations of Search Procedure 1 for obtaining an (ϵ_p, ϵ_o)-primal solution of (1) is bounded by $\mathcal{O}(N_p(\|\lambda^*\|))$, where $N_p(\cdot)$ is defined in (37).*

Proof. In view of Theorem 15, the iteration count k in Search Procedure 1 can not exceed $K := \max\{0, \lceil \log(\|\lambda^*\|/t_0)\rceil$, and hence its overall number of inner (i.e., Nesterov's optimal method) iterations is bounded by $\sum_{k=0}^{K} N_p(t_k) = \sum_{k=0}^{K} \lceil \beta_0 + \beta_1 t_k^{1/2} \rceil$, where β_0 and β_1 are defined by (39). The result now follows from the definition of t_0 in (39) and (40) with $p_1 = 1/2$ and $\bar{t} = \|\lambda^*\|$. ■

4.1.2 Computation of a primal-dual approximate solution

In this subsection, we are interested in deriving the iteration-complexity involved in the computation of an (ϵ_p, ϵ_d)-primal-dual solution of problem (1) by approximately solving the penalized problem (32) for certain value of the penalty parameter ρ. We assume throughout this subsection that the norms on $\Re^n$ and $\Re^m$ are the ones associated with their respective inner products.

Given an approximate solution $\tilde{x} \in X$ for the penalized problem (32), the following result shows that there exists a pair $(\tilde{x}^+, \lambda) \in X \times (-\mathcal{K})$ depending on $\tilde{x}$ that approximately satisfies the optimality conditions (6).

Proposition 19 *If $\tilde{x} \in X$ is a δ-approximate solution of (32), i.e., it satisfies (33), then the pair $(\tilde{x}^+, \lambda)$ defined as*

$$\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\Psi_\rho(\tilde{x})/M_\rho), \tag{42}$$

$$\lambda := \rho[\mathcal{A}(\tilde{x}^+) - \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+))] \tag{43}$$

is in $X \times (-\mathcal{K})$ and satisfies the second relation in (7) and the relations

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+)) \leq \frac{1}{\rho}\|\lambda^*\| + \sqrt{\frac{\delta}{\rho}}, \tag{44}$$

$$\nabla f(\tilde{x}^+) + (\mathcal{A}_0)^*\lambda \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(2\sqrt{2M_\rho\delta}), \tag{45}$$

where λ^* is an arbitrary Lagrange multiplier for (1).

Proof. It follows from Lemma 5(a) with $\phi = \Psi_\rho$ and $\tau = M_\rho$ that $\Psi_\rho(\tilde{x}^+) \leq \Psi_\rho(\tilde{x})$, and hence that $\Psi_\rho(\tilde{x}^+) - \Psi_\rho^* \leq \delta$ in view of (33). By Proposition 13, this implies that (44) holds. Moreover, by Lemma 5(d) with $\phi = \Psi_\rho$ and $L_\phi = M_\rho$ and assumption (33), we have

$$\|\nabla\Psi_\rho(\tilde{x})]_X^{1/M_\rho}\| \leq \sqrt{2M_\rho[\Psi_\rho(\tilde{x}) - \Psi_\rho^*]} \leq \sqrt{2M_\rho\delta}.$$

Relation (45) now follows from this inequality, Proposition 4(b) with $L_\phi = M_\rho$, $\tau = 1/M_\rho$ and $\epsilon = \sqrt{2M_\rho\delta}$, and the fact that $\nabla\Psi_\rho(\tilde{x}^+) = \nabla f(\tilde{x}^+) + (\mathcal{A}_0)^*\lambda$, where λ is given by (43). Finally, $\lambda \in (-\mathcal{K})$ and the second relation of (7) holds in view of the definition of λ and well-known properties of the projection operator onto a cone. ∎

With the aid of Proposition 19, we can now derive an iteration-complexity bound for Nesterov's method applied to the quadratic penalty problem (32) to compute an (ϵ_p, ϵ_d)-primal-dual solution of (1).

Theorem 20 *Let λ^* be an arbitrary Lagrange multiplier for (1) and let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. If*

$$\rho = \rho_{pd}(t) := \frac{1}{\epsilon_p}\left(t + \frac{\epsilon_d}{\sqrt{8}\|\mathcal{A}\|}\right) \tag{46}$$

for some $t \geq \|\lambda^*\|$, then Nesterov's optimal method with $h(\cdot) = \|\cdot\|^2/2$ applied to problem (32) finds an (ϵ_p, ϵ_d)-primal-dual solution of (1) in at most

$$N_{pd}(t) := \left\lceil 4D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|^2 t}{\epsilon_p\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{8}\epsilon_p}\right)\right\rceil \tag{47}$$

iterations, where D_X is defined in Theorem 15.

Proof. Let $\tilde{x} \in X$ be a δ-approximate solution of (32) where $\delta := \epsilon_d^2/(8M_\rho)$. Noting that $\delta \leq \epsilon_d^2/(8\rho\|\mathcal{A}\|^2)$ in view of the fact that $M_\rho := L_f + \rho\|\mathcal{A}\|^2 \geq \rho\|\mathcal{A}\|^2$, we conclude from Proposition 19 that the pair $(\tilde{x}^+, \lambda)$ defined as $\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\Psi_\rho(\tilde{x})/M_\rho)$ and $\lambda := \rho[\mathcal{A}(\tilde{x}^+) - \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+))]$ satisfies $\nabla f(\tilde{x}^+) + \mathcal{A}^*\lambda \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d)$ and

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+)) \leq \frac{1}{\rho}\|\lambda^*\| + \sqrt{\frac{\delta}{\rho}} \leq \frac{1}{\rho}\left(\|\lambda^*\| + \frac{\epsilon_d}{\sqrt{8}\|\mathcal{A}\|}\right) \leq \epsilon_p,$$

where the last inequality is due to (46) and the assumption that $t \geq \|\lambda^*\|$. We have thus shown that $(\tilde{x}^+, \lambda)$ is an (ϵ_p, ϵ_d)-primal-dual solution of (1). In view of Corollary 7, Nesterov's optimal method finds an approximate solution $\tilde{x}$ as above in at most

$$\left\lceil \sqrt{2} D_X \left(\frac{M_\rho}{\delta} \right)^{1/2} \right\rceil = \left\lceil \sqrt{2} D_X \left(\frac{8 M_\rho^2}{\epsilon_d^2} \right)^{1/2} \right\rceil = \left\lceil 4 D_X \frac{M_\rho}{\epsilon_d} \right\rceil = \left\lceil 4 D_X \frac{L_f + \rho \|\mathcal{A}\|^2}{\epsilon_d} \right\rceil$$

iterations. Substituting the value of ρ given by (36) in the above bound, we obtain bound (37). ∎

The following result states the iteration-complexity of a "guess and check" procedure based on Theorem 20. Its proof is based on similar arguments as those used in the proof of Corollary 18.

Corollary 21 *Let λ^* be the minimum norm Lagrange multiplier for (1). Consider the "guess and check" procedure similar to Search Procedure 1 where the functions N_p and ρ_p are replaced by the functions N_{pd} and ρ_{pd} defined in Theorem 20, δ is set to $\epsilon_d^2/(8M_\rho)$, and t_0 is set to $[\max(\beta_0, 1)/\beta_1]$ with*

$$\beta_0 = 4 D_X \left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{8}\epsilon_p} \right), \quad \beta_1 = \frac{4 D_X \|\mathcal{A}\|^2}{\epsilon_p \epsilon_d}.$$

Then, the overall number of iterations of this "guess and check" procedure for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1) is bounded by $\mathcal{O}(N_{pd}(\|\lambda^*\|))$, where $N_{pd}(\cdot)$ is defined in (47).

4.2 Quadratic penalty method applied to a perturbed problem

Throughout this subsection, we assume that the norm on $\Re^n$ is the one associated with its inner product. (Recall that throughout Section 4, we are also assuming that the norm on $\Re^m$ is the one associated with its inner product.) Consider the perturbation problem

$$f_\gamma^* := \min\{f_\gamma(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 : \mathcal{A}(x) \in \mathcal{K}^*, x \in X\}, \tag{48}$$

where x_0 is a fixed point in X and $\gamma > 0$ is a pre-specified perturbation parameter. It is well-known that if γ is sufficiently small, then an approximate solution of (48) will also be an approximate solution of (1).

In this subsection, we will derive the iteration-complexity (in terms of Nesterov's optimal method iterations) of computing an approximate solution of (1) by applying the quadratic penalty approach to the perturbation problem (48) for a conveniently chosen perturbation parameter $\gamma > 0$. The subsection is divided into two subsubsections: the computation of a primal approximate solution of (1) is treated in the first subsubsection while the computation of a primal-dual approximate solution of (1) is treated in the second one.

4.2.1 Computation of a primal approximate solution

The quadratic penalty problem associated with (48) is given by

$$\Psi_{\rho,\gamma}^* := \min_{x \in X} \left\{ \Psi_{\rho,\gamma}(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 + \frac{\rho}{2} d_{\mathcal{K}^*}(\mathcal{A}(x))^2 \right\}. \tag{49}$$

It can be easily seen that the function $\Psi_{\rho,\gamma}$ has $M_{\rho,\gamma}$-Lipschitz continuous gradient where

$$M_{\rho,\gamma} := L_f + \rho\|\mathcal{A}\|^2 + \gamma. \tag{50}$$

The following simple lemma relates the optimal values of the perturbation problem (48) and the original problem (1).

Lemma 22 *Let f^*, Ψ_ρ^*, f_γ^*, and $\Psi_{\rho,\gamma}^*$ be the optimal values defined in (1), (32), (48), and (49), respectively. Then,*

$$0 \le f_\gamma^* - f^* \le \gamma D_X^2/2 \tag{51}$$

$$0 \le \Psi_{\rho,\gamma}^* - \Psi_\rho^* \le \gamma D_X^2/2, \tag{52}$$

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

Proof. The first inequalities in both relations (51) and (52) follow immediately from the fact that $f_\gamma \ge f$ and $\Psi_{\rho,\gamma} \ge \Psi_\rho$. Now, let x^* and x_γ^* be optimal solutions of (1) and (48), respectively. Then,

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

from which the second inequality in (51) immediately follows. The second inequality in (52) can be similarly shown. ∎

The following theorem gives the iteration-complexity of computing an (ϵ_p, ϵ_o)-primal solution of (1) by applying the quadratic penalty approach to the perturbation problem (48) with an appropriate choice of perturbation parameter $\gamma > 0$.

Theorem 23 *Let λ_γ^* be an arbitrary Lagrange multiplier for (48) and let $(\epsilon_p, \epsilon_o) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. If*

$$\rho = \tilde{\rho}_p(t) := \frac{4t}{\epsilon_p} \quad \text{and } \gamma = \frac{\epsilon_o}{D_X^2}, \tag{53}$$

for some $t \ge \|\lambda_\gamma^*\|$, then the variant of Nesterov's optimal method of Theorem 8 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ applied to problem (49) finds an (ϵ_p, ϵ_o)-primal solution of (1) in at most

$$\tilde{N}_p(t) := \left\lceil 2\sqrt{2} D_X \left(\sqrt{\frac{L_f}{\epsilon_o}} + 2\|\mathcal{A}\| \sqrt{\frac{t}{\epsilon_p \epsilon_o}} \right) + 1 \right\rceil \max\left\{ 1, \left\lceil \log \frac{\epsilon_o}{t\epsilon_p} \right\rceil \right\} \tag{54}$$

iterations, where D_X is defined in Theorem 15.

Proof. Assume that ρ and γ are given by (53) for some $t \ge \|\lambda_\gamma^*\|$. Let $\delta := \min\{\epsilon_o/4, \epsilon_p t/2\}$ and $\tilde{x} \in X$ be a δ-approximate solution of (49), i.e., $\Psi_{\rho,\gamma}(\tilde{x}) - \Psi_{\rho,\gamma}^* \le \delta$. Then, Proposition 13 with $f = f_\gamma$, $\Psi_\rho = \Psi_{\rho,\gamma}$ and $\rho = \tilde{\rho}_p(t)$ and the assumption that $t \ge \|\lambda_\gamma^*\|$ imply that

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \;\le\; \frac{2}{\tilde{\rho}_p(t)}\|\lambda_\gamma^*\| + \sqrt{\frac{2\delta}{\tilde{\rho}_p(t)}} \;\le\; \frac{\epsilon_p\|\lambda_\gamma^*\|}{2t} + \sqrt{\frac{t\epsilon_p}{4t/\epsilon_p}} \le \frac{\epsilon_p}{2} + \frac{\epsilon_p}{2} = \epsilon_p$$

and $f_\gamma(\tilde{x}) - f_\gamma^* \leq \delta \leq \epsilon_o/4$. The later inequality together with (51) and (53) in turn imply that

$$f(\tilde{x}) - f^* \leq f_\gamma(\tilde{x}) - f^* \leq [f_\gamma(\tilde{x}) - f_\gamma^*] + [f_\gamma^* - f^*] \leq \frac{\epsilon_o}{4} + \frac{\epsilon_o}{2} < \epsilon_o.$$

We have thus proved that $\tilde{x}$ is an (ϵ_p, ϵ_o)-primal solution of (1). Now, using Theorem 8 with $\phi = \Psi_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \delta$, and noting that the definition of γ in (53) and the definition of δ imply that

$$Q = \frac{\mu \|x_0^{sd} - x^*\|^2}{2\epsilon} = \frac{\gamma \|x_0^{sd} - x^*\|^2}{2\delta} \leq \frac{\gamma D_X^2}{2\delta} = \frac{\epsilon_o}{\min\{\epsilon_o/2, \epsilon_p t\}} = \max\left\{2, \frac{\epsilon_o}{\epsilon_p t}\right\},$$

we conclude that the number of iterations performed by Nesterov's optimal method (with the restarting feature) for finding a δ-approximate solution $\tilde{x}$ as above is bounded by

$$\left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \lceil \log Q \rceil = \left\lceil \sqrt{\frac{8M_{\rho,\gamma}}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\epsilon_o}{t\epsilon_p} \right\rceil \right\}.$$

Now, using relations (50) and (53) and some trivial majorization, we easily see that the latter bound is majorized by (54). ∎

Note that the complexity bound (54) derived in Theorem 23 is guaranteed only under the assumption that $t \geq \|\lambda_\gamma^*\|$, where λ_γ^* is the minimum norm Lagrange multiplier for (48). Since the bound (54) is a monotonically increasing function of t, the ideal (theoretical) choice of t would be to set $t = \|\lambda_\gamma^*\|$. Without assuming any knowledge of this Lagrange multiplier, the following result shows that a "guess and check" procedure similar to Search Procedure 1 still has an $\mathcal{O}(\tilde{N}_p(\|\lambda_\gamma^*\|))$ iteration-complexity bound, where $\tilde{N}_p(\cdot)$ is defined in (54). Before establishing the iteration-complexity of this "guess and check" procedure, we state the following technical result which substantially generalizes the one stated in Lemma 17. The proof of this result is given in the Appendix.

Proposition 24 *For some positive integer L, let positive scalars $p_1, p_2, \cdots, p_L$ be given. Then, there exists a constant $C = C(p_1, \cdots, p_L)$ such that for any positive scalars $\beta_0, \beta_1, \cdots, \beta_L$, ν, and $\bar{t}$, we have*

$$\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, t_k^{p_l}\right) \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil \right\} \leq C \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, \bar{t}^{p_l}\right) \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \right\}, \tag{55}$$

where

$$K := \max\left\{0, \left\lceil \log\left(\frac{\bar{t}}{t_0}\right) \right\rceil\right\}, \; t_0 := \min_{1 \leq l \leq L} \left(\frac{\max(\beta_0, 1)}{\beta_l}\right)^{1/p_l}, \; t_k = t_0 2^k, \; \forall k = 1, \ldots, K. \tag{56}$$

In particular, if $\nu = \bar{t}$, then (55) implies that

$$\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, t_k^{p_l}\right) \right\rceil \leq C \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l \, \bar{t}^{p_l}\right) \right\rceil. \tag{57}$$

We are now ready to discuss the iteration-complexity bound of the above-mentioned "guess and check" procedure.

Corollary 25 *Let λ_γ^* be the minimum norm Lagrange multiplier for (48). Consider the "guess and check" procedure similar to Search Procedure 1 where the functions N_p and ρ_p are replaced by the functions $\tilde{N}_p$ and $\tilde{\rho}_p$ defined in Theorem 23, δ is set to $\delta := \min\{\epsilon_o/4, \epsilon_p t_k/2\}$, and t_0 is set to $[\max(\beta_0, 1)/\beta_1]^2$ with*

$$\beta_0 = 2\sqrt{2}D_X \frac{L_f}{\epsilon_o} + 1, \quad \beta_1 = \frac{4\sqrt{2}D_X\|\mathcal{A}\|}{\sqrt{\epsilon_p\epsilon_o}}. \tag{58}$$

Then, the overall number of iterations of this "guess and check" procedure for obtaining an (ϵ_p, ϵ_o)-primal-dual solution of (1) is bounded by $\mathcal{O}(\tilde{N}_p(\|\lambda_\gamma^*\|))$, where $\tilde{N}_p(\cdot)$ is defined in (54).

Proof. In view of Theorem 23, the iteration count k in Search Procedure 1 can not exceed $K := \max\{0, \lceil \log(\|\lambda_\gamma^*\|/t_0)\rceil\}$, and hence its overall number of inner (i.e., Nesterov's optimal method) iterations is bounded by $\sum_{k=0}^{K} \tilde{N}_p(t_k) = \sum_{k=0}^{K} \lceil \beta_0 + \beta_1 t_k^{1/2}\rceil \max\{1, \lceil \log(\nu/t_k)\rceil\}$, where β_0 and β_1 are defined by (58), and $\nu = \epsilon_0/\epsilon_p$. The result now follows from the definition of t_0 and Proposition 24 with $L = 1$, $p_1 = 1/2$, and $\bar{t} = \|\lambda_\gamma^*\|$. ∎

It is interesting to compare the two functions $N_p(t)$ and $\tilde{N}_p(t)$ defined in Theorems 15 and 23, respectively. Indeed, both functions (up to some constant factor) are comparable for the case when $r_t := \epsilon_o/(t\epsilon_p) = \mathcal{O}(1)$. Otherwise, if $r_t = \Omega(1)$ and, we further assume that the term involving t dominates the other terms in the right hand side of (54), i.e.,

$$2\sqrt{2}D_X\sqrt{\frac{L_f}{\epsilon_o}} + 1 = \mathcal{O}\left(4\sqrt{2}D_X\|\mathcal{A}\|\sqrt{\frac{t}{\epsilon_p\epsilon_o}}\right),$$

then

$$\frac{\tilde{N}_p(t)}{N_p(t)} \leq \mathcal{O}(1)\left(\frac{\log r_t}{\sqrt{r_t}}\right),$$

for some universal constant $\mathcal{O}(1)$. Hence, in the latter case, we conclude that the function $\tilde{N}_p(t)$ is asymptotically smaller than the function $N_p(t)$.

4.2.2 Computation of a primal-dual approximate solution

In this subsection, we derive the iteration-complexity of computing a primal-dual approximate solution of (1) by applying the quadratic penalty method to the perturbation problem (48). We will show that the resulting approach has a substantially better iteration-complexity than the one discussed in Subsection 4.1.2, which consists of applying the quadratic penalty method directly to the original problem (1).

Theorem 26 *Let λ_γ^* be an arbitrary Lagrange multiplier for (48) and let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. Also assume that*

$$\rho = \tilde{\rho}_{pd}(t) := \frac{1}{\epsilon_p}\left(t + \frac{\epsilon_d}{\sqrt{32}\|\mathcal{A}\|}\right), \ \gamma = \frac{\epsilon_d}{2D_X}, \tag{59}$$

for some $t \geq \|\lambda_\gamma^*\|$. Then, the variant of Nesterov's optimal method of Theorem 8 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ applied to problem (49) finds an (ϵ_p, ϵ_d)-primal-dual solution of (1) in at most

$$\tilde{N}_{pd}(t) := \lceil 8\mathcal{S}(t)\rceil \ \lceil 2\log(2\mathcal{S}(t))\rceil, \tag{60}$$

where D_X is defined in Theorem 15 and

$$\mathcal{S}(t) := \left[2D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{32}\epsilon_p}\right) + 1\right]^{\frac{1}{2}} + \frac{\sqrt{2}D_X^{1/2}\|\mathcal{A}\|}{\sqrt{\epsilon_p\epsilon_d}}\, t^{1/2}. \tag{61}$$

Proof. Define $\delta := \epsilon_d^2/(32M_{\rho,\gamma})$ and let $\tilde{x} \in X$ be a δ-approximate solution for (49), i.e., $\Psi_{\rho,\gamma}(\tilde{x}) - \Psi^*_{\rho,\gamma} \leq \delta$. It then follows from Proposition 19 with $\Psi_\rho = \Psi_{\rho,\gamma}$, $f = f_\gamma$, and $M_\rho = M_{\rho,\gamma}$ that the pair $(\tilde{x}^+, \lambda)$ defined as $\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\Psi_{\rho,\gamma}(\tilde{x})/M_{\rho,\gamma})$ and $\lambda := \rho[\mathcal{A}(\tilde{x}^+) - \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+))]$ satisfies

$$\nabla f_\gamma(\tilde{x}^+) + \mathcal{A}^*\lambda \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(2\sqrt{2M_{\rho,\gamma}\delta}) = -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d/2).$$

This together with the fact that $\gamma\|\tilde{x}^+ - x_0\| \leq \gamma D_X \leq \epsilon_d/2$ then imply that

$$\begin{aligned}\nabla f(\tilde{x}^+) + \mathcal{A}^*\lambda &= [\nabla f_\gamma(\tilde{x}^+) - \gamma(\tilde{x}^+ - x_0)] + \mathcal{A}^*\lambda \\ &\in -\gamma(\tilde{x}^+ - x_0) - \mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d/2) \subseteq -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(\epsilon_d).\end{aligned}$$

Moreover, noting that $\delta := \epsilon_d^2/(32M_{\rho,\gamma}) \leq \epsilon_d^2/(32\rho\|\mathcal{A}\|^2)$, we conclude that Proposition 19 also implies that

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+)) \leq \frac{1}{\rho}\|\lambda^*_\gamma\| + \sqrt{\frac{\delta}{\rho}} \leq \frac{1}{\rho}\left(\|\lambda^*_\gamma\| + \frac{\epsilon_d}{\sqrt{32}\|\mathcal{A}\|}\right) \leq \epsilon_p,$$

where the last inequality is due to (59) and the assumption that $t \geq \|\lambda^*_\gamma\|$. We have thus shown that $(\tilde{x}^+, \lambda)$ is an (ϵ_p, ϵ_d)-primal-dual solution of (1). Now, using Theorem 8 with $\phi = \Psi_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \delta$, and noting that the definition of δ and the definition of γ in (59) imply that

$$Q = \frac{\mu\|x_0^{sd} - x^*\|^2}{2\epsilon} = \frac{\gamma\|x_0^{sd} - x^*\|^2}{2\delta} \leq \frac{\gamma D_X^2}{2\delta} = \frac{\gamma D_X^2}{\epsilon_d^2/(16M_{\rho,\gamma})} = \frac{4M_{\rho,\gamma}}{\gamma},$$

we conclude that the number of iterations performed by Nesterov's optimal method (with the restarting feature) for finding a δ-approximate solution $\tilde{x}$ as above is bounded by

$$\left\lceil 8\sqrt{\frac{M_{\rho,\gamma}}{\gamma}} \right\rceil \lceil \log Q \rceil = \left\lceil 8\sqrt{\frac{M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil \log\left(\frac{4M_{\rho,\gamma}}{\gamma}\right) \right\rceil. \tag{62}$$

Now, using (50) and the definitions of ρ and γ in (53), we easily see that the latter bound is majorized by (60). ∎

Note that the complexity bound (60) derived in Theorem 26 is guaranteed only under the assumption that $t \geq \|\lambda^*_\gamma\|$, where λ^*_γ is the minimum norm Lagrange multiplier for (48). Since the bound (60) is a monotonically increasing function of t, the ideal (theoretical) choice of t would be to set $t = \|\lambda^*_\gamma\|$. Without assuming any knowledge of this Lagrange multiplier, the following result shows that a "guess and check" procedure similar to Search Procedure 1 still has an $\mathcal{O}(\tilde{N}_{pd}(\|\lambda^*_\gamma\|))$ iteration-complexity bound, where $\tilde{N}_{pd}(\cdot)$ is defined in (60).

Corollary 27 *Let λ^*_γ be the minimum norm Lagrange multiplier for (48). Consider the "guess and check" procedure similar to Search Procedure 1 where the functions N_p and ρ_p are replaced by the functions $\tilde{N}_{pd}$ and $\tilde{\rho}_{pd}$ defined in Theorem 26, Nesterov's optimal method is replaced by its variant*

of Theorem 8 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ (see step 2 of Search Procedure 1), δ is set to $\epsilon_d^2/(32M_{\rho,\gamma})$, and t_0 is set to $[\max(\beta_0, 1)/\beta_1]^2$ with

$$\beta_0 = 8\left[2D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{32}\epsilon_p}\right) + 1\right]^{1/2}, \quad \beta_1 = \frac{8\sqrt{2}D_X^{1/2}\|\mathcal{A}\|}{(\epsilon_p\epsilon_d)^{1/2}}. \tag{63}$$

Then, the overall number of iterations of this "guess and check" procedure for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1) is bounded by $\mathcal{O}(\tilde{N}_{pd}(\|\lambda_\gamma^*\|))$, where $\tilde{N}_{pd}(\cdot)$ is defined in (60).

Proof. In view of Theorem 26, the iteration count k of the above "guess and check" (see Search Procedure 1) for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1) can not exceed $K := \max\{0, \lceil \log(\|\lambda^*\|/t_0)\rceil$, and hence its overall number of inner (i.e., Nesterov's optimal method) iterations is bounded by

$$\sum_{k=0}^{K} \tilde{N}_{pd}(t_k) = \sum_{k=0}^{K} \lceil 8\mathcal{S}(t_k)\rceil \, \lceil 2\log(2\mathcal{S}(t_k))\rceil \leq \lceil 2\log(2\mathcal{S}(t_K))\rceil \sum_{k=0}^{K} \lceil 8\mathcal{S}(t_k)\rceil \,. \tag{64}$$

Now, the definition of t_0 and relation (40) with $p = 1/2$ and $\bar{t} = \|\lambda_\gamma^*\|$ imply that

$$\sum_{k=0}^{K} \lceil 8\mathcal{S}(t_k)\rceil = \sum_{k=0}^{K} \left\lceil \beta_0 + \beta_1 t_k^{1/2}\right\rceil = \mathcal{O}\left(\left\lceil \beta_0 + \beta_1 \|\lambda_\gamma^*\|^{1/2}\right\rceil\right) = \mathcal{O}\left(\lceil 8\mathcal{S}(\|\lambda_\gamma^*\|)\rceil\right),$$

where β_0 and β_1 are defined by (63). Moreover, using the fact that $t_K \leq 2\|\lambda_\gamma^*\|$, we easily see that $\lceil 2\log(2\mathcal{S}(t_K))\rceil = \mathcal{O}(\lceil 2\log(2\mathcal{S}(\|\lambda_\gamma^*\|))\rceil)$. Substituting the last two bounds into (64), we then conclude that $\sum_{k=0}^{K} \tilde{N}_{pd}(t_k) = \mathcal{O}(\tilde{N}_{pd}(\|\lambda_\gamma^*\|))$, and hence that the corollary holds. ∎

It is interesting to compare the functions $N_{pd}(t)$ and $\hat{N}_{pd}(t)$ defined in Theorems 20 and 26, respectively. It follows from (47), (60), and (61) that

$$\frac{\tilde{N}_{pd}(t)}{N_{pd}(t)} \leq \frac{(\lceil 8\mathcal{S}(t)\rceil \, \lceil 2\log(2\mathcal{S}(t))\rceil)}{\left\lceil (\mathcal{S}(t))^2 - 1\right\rceil} \leq \mathcal{O}(1)\frac{\log \mathcal{S}(t)}{\mathcal{S}(t) - 1}, \tag{65}$$

where $\mathcal{O}(1)$ denotes an absolute constant. Hence, when $\mathcal{S}(t)$ is large, the bound $\tilde{N}_{pd}(t)$ can be substantially smaller than the bound $N_{pd}(t)$.

Note that we can not use the previous observation to compare the iteration-complexity of Corollary 21 with that obtained in Corollary 27 since the first one is expressed in terms of $\|\lambda^*\|$ and the later in terms of $\|\lambda_\gamma^*\|$. However, if $\|\lambda_\gamma^*\| = \mathcal{O}(\|\lambda^*\|)$, then it can be easily seen that (65) implies that

$$\frac{\tilde{N}_{pd}(\|\lambda_\gamma^*\|)}{N_{pd}(\|\lambda^*\|)} \leq \mathcal{O}(1)\frac{\log \mathcal{S}(\|\lambda^*\|)}{\mathcal{S}(\|\lambda^*\|) - 1}.$$

Hence, the first complexity is better than the second one whenever $\|\lambda_\gamma^*\| = \mathcal{O}(\|\lambda^*\|)$ and $\mathcal{S}(\|\lambda^*\|)$ is sufficiently large.

The following result describes an alternative way of choosing the penalty parameter ρ in subproblem (49) as a function of t, but requires that $t \geq \|\lambda^*\|$ instead of $t \geq \|\lambda_\gamma^*\|$.

Theorem 28 *Let λ^* be an arbitrary Lagrange multiplier for (1) and let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. Assume that*

$$\rho = \hat{\rho}_{pd}(t) := \left(\frac{\sqrt{\epsilon_d D_X/2} + \sqrt{(\epsilon_d D_X/2) + 4\alpha(t)\epsilon_p}}{2\epsilon_p} \right)^2, \qquad \gamma = \frac{\epsilon_d}{2D_X}, \tag{66}$$

where $\alpha(t) := t + \epsilon_d/(\sqrt{32}\,\|\mathcal{A}\|)$ for some $t \geq \|\lambda^*\|$. Then, the variant of Nesterov's optimal method of Theorem 8 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ applied to problem (49) finds an (ϵ_p, ϵ_d)-primal-dual solution of (1) in at most

$$\hat{N}_{pd}(t) := \left\lceil 8\hat{\mathcal{S}}(t) \right\rceil \left\lceil 2\log(2\hat{\mathcal{S}}(t)) \right\rceil, \tag{67}$$

where D_X is defined in Theorem 15 and

$$\hat{\mathcal{S}}(t) := \sqrt{\frac{2L_f D_X}{\epsilon_d}} + \|\mathcal{A}\| \left(\frac{D_X}{\epsilon_p} + \sqrt{\frac{2tD_X}{\epsilon_p \epsilon_d}} \right) + \frac{\sqrt{\|\mathcal{A}\| D_X}}{8^{1/4}\sqrt{\epsilon_p}} + 1. \tag{68}$$

Proof. Define $\delta := \epsilon_d^2/(32M_{\rho,\gamma})$ and let $\tilde{x} \in X$ be a δ-approximate solution for (49), i.e., $\Psi_{\rho,\gamma}(\tilde{x}) - \Psi^*_{\rho,\gamma} \leq \delta$. As shown in the proof of Theorem 26, the pair $(\tilde{x}^+, \lambda)$ defined as $\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\Psi_{\rho,\gamma}(\tilde{x})/M_{\rho,\gamma})$ and $\lambda := \rho[\mathcal{A}(\tilde{x}^+) - \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+))]$ satisfies

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

Moreover, it follows from Lemma 5(a) with $\phi = \Psi_{\rho,\gamma}$ and $\tau = M_{\rho,\gamma}$ that $\Psi_{\rho,\gamma}(\tilde{x}^+) \leq \Psi_{\rho,\gamma}(\tilde{x})$, and hence that $\Psi_{\rho,\gamma}(\tilde{x}^+) - \Psi^*_{\rho,\gamma} \leq \Psi_{\rho,\gamma}(\tilde{x}) - \Psi^*_{\rho,\gamma} \leq \delta$. This observation together with (32), (49), (52) and (66) then imply that

$$\begin{aligned} \Psi_\rho(\tilde{x}^+) - \Psi^*_\rho &= \Psi_{\rho,\gamma}(\tilde{x}^+) - \frac{\gamma}{2}\|\tilde{x}^+ - x_0\|^2 - \Psi^*_\rho \leq [\Psi_{\rho,\gamma}(\tilde{x}^+) - \Psi^*_{\rho,\gamma}] + [\Psi^*_{\rho,\gamma} - \Psi^*_\rho] \\ &\leq \delta + \gamma D_X^2 \leq \frac{\epsilon_d^2}{32\rho\|\mathcal{A}\|^2} + \frac{\epsilon_d D_X}{2}, \end{aligned}$$

where the last inequality follows from the definition of δ and the fact that $M_\rho \geq \rho\|\mathcal{A}\|^2$. The above inequality together with Proposition 13, the assumption that $t \geq \|\lambda^*\|$, relation (66) and Lemma 14 with $\alpha_0 = \epsilon_p$, $\alpha_1 = \sqrt{\epsilon_d D_X/2}$ and $\alpha_2 = t + \epsilon_d/(\sqrt{32}\,\|\mathcal{A}\|) =: \alpha(t)$ then imply that

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+)) \leq \frac{1}{\rho}\|\lambda^*\| + \sqrt{\frac{\epsilon_d^2}{32\rho^2\|\mathcal{A}\|^2} + \frac{D_X\epsilon_d}{2\rho}} \leq \frac{1}{\rho}\left(t + \frac{\epsilon_d}{\sqrt{32}\,\|\mathcal{A}\|}\right) + \sqrt{\frac{D_X\epsilon_d}{2\rho}} = \epsilon_p.$$

We have thus shown that $(\tilde{x}^+, \lambda)$ is an (ϵ_p, ϵ_d)-primal-dual solution of (1). Now, using Theorem 8 with $\phi = \Psi_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \delta$, and noting that the definition of δ and the definition of γ in (66) imply that

$$Q = \frac{\mu\|x_0^{sd} - x^*\|^2}{2\epsilon} = \frac{\gamma\|x_0^{sd} - x^*\|^2}{2\delta} \leq \frac{\gamma D_X^2}{2\delta} = \frac{\gamma D_X^2}{\epsilon_d^2/(16M_{\rho,\gamma})} = \frac{4M_{\rho,\gamma}}{\gamma},$$

we conclude that the number of iterations performed by Nesterov's optimal method (with the restarting feature) for finding a δ-approximate solution $\tilde{x}$ as above is bounded by (62). Since relations (50) and (66) and the definition of $\alpha(t)$ imply that

$$\sqrt{\frac{M_{\rho,\gamma}}{\gamma}} = \sqrt{\frac{L_f + \rho\|\mathcal{A}\|^2 + \gamma}{\gamma}} \leq \sqrt{\frac{L_f}{\gamma}} + \|\mathcal{A}\|\sqrt{\frac{\rho}{\gamma}} + 1 = \sqrt{\frac{2L_f D_X}{\epsilon_d}} + \|\mathcal{A}\|\sqrt{\frac{\rho}{\gamma}} + 1$$

and

$$\sqrt{\frac{\rho}{\gamma}} \leq \left(\frac{\sqrt{\epsilon_d D_X/2}}{\epsilon_p} + \sqrt{\frac{\alpha(t)}{\epsilon_p}}\right)\sqrt{\frac{2D_X}{\epsilon_d}} = \frac{D_X}{\epsilon_p} + \sqrt{\frac{t+\epsilon_d/(\sqrt{32}\|\mathcal{A}\|)}{\epsilon_p}}\sqrt{\frac{2D_X}{\epsilon_d}}$$
$$\leq \left(\frac{D_X}{\epsilon_p} + \sqrt{\frac{2tD_X}{\epsilon_p\epsilon_d}} + \sqrt{\frac{D_X}{\sqrt{8}\epsilon_p\|\mathcal{A}\|}}\right),$$

it follows that (62) can be bounded by (67). ∎

The following result states the iteration-complexity of a "guess and check" procedure based on Theorem 28. Its proof is based on similar arguments as those used in the proof of Corollary 27.

Corollary 29 *Let λ^* be the minimum norm Lagrange multiplier for (1). Consider the "guess and check" procedure similar to Search Procedure 1 where the functions N_p and ρ_p are replaced by the functions $\hat{N}_{pd}$ and $\hat{\rho}_{pd}$ defined in Theorem 26, Nesterov's optimal method is replaced by its variant of Theorem 8 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ (see step 2 of Search Procedure 1), δ is set to $\epsilon_d^2/(32M_{\rho,\gamma})$, and t_0 is set to $[\max(\beta_0,1)/\beta_1]^2$ with*

$$\beta_0 = 8\left(\sqrt{\frac{2L_f D_X}{\epsilon_d}} + \frac{\|\mathcal{A}\|D_X}{\epsilon_p} + \sqrt{\frac{\|\mathcal{A}\|D_X}{\sqrt{8}\epsilon_p}} + 1\right), \quad \beta_1 = 8\|\mathcal{A}\|\sqrt{\frac{2D_X}{\epsilon_p\epsilon_d}}.$$

Then, the overall number of iterations of this "guess and check" procedure for obtaining an (ϵ_p, ϵ_d)-primal-dual solution of (1) is bounded by $\mathcal{O}(\hat{N}_{pd}(\|\lambda^*\|))$, where $\hat{N}_{pd}(\cdot)$ is defined in (67).

It is interesting to compare the functions $N_{pd}(t)$ and $\hat{N}_{pd}(t)$ defined in Theorems 20 and 26, respectively. When the second term in the right hand side of (68) is dominated by the other terms, i.e.,

$$\frac{\|\mathcal{A}\|D_X}{\epsilon_p} = \mathcal{O}\left(\frac{\|\mathcal{A}\|t^{1/2}D_X^{1/2}}{\sqrt{\epsilon_p\epsilon_d}} + \sqrt{\frac{L_f D_X}{\epsilon_d}}\right), \tag{69}$$

then it can be easily seen that

$$\frac{\hat{N}_{pd}(t)}{N_{pd}(t)} \leq \mathcal{O}(1)\frac{\tilde{N}_{pd}(t)}{N_{pd}(t)} \leq \mathcal{O}(1)\frac{\log \mathcal{S}(t)}{\mathcal{S}(t)-1}.$$

Hence, if (69) holds and $\mathcal{S}(t)$ is large, then $\hat{N}_{pd}(t)$ can be considerably smaller than $N_{pd}(t)$. It is also interesting to observe when $\epsilon_p = \epsilon_d = \epsilon$, the dependence on ϵ of the function $\hat{N}_{pd}$ is $\mathcal{O}(1/\epsilon)\log(1/\epsilon)$ while that of N_{pd} is $\mathcal{O}(1/\epsilon^2)$.

5 The exact penalty method

The goal of this section is to establish the iteration-complexity (in terms of Nesterov's optimal method iterations) of first-order exact penalty methods for computing a ϵ_p-primal solution of (1). It consists of two subsections. The first one considers the complexity of a first-order exact penalty method applied directly to the original problem (1) while the second one deals with the complexity of a first-order exact penalty method applied to the perturbation problem (48) for some suitably chosen perturbation parameter θ.

5.1 Exact penalty method applied to the original problem

In this subsection, we consider the complexity of a first-order exact penalty method applied directly to the original problem (1).

Instead of using the quadratic penalty term $\rho[d_{\mathcal{K}^*}(\mathcal{A}(x))]^2/2$, the exact penalty method employs the penalty term given by $\theta d_{\mathcal{K}^*}(\mathcal{A}(x))$ for some $\theta > 0$, and solves the following relaxation problem

$$\Phi_\theta^* := \inf_{x \in X} \Phi_\theta(x) := f(x) + \theta d_{\mathcal{K}^*}(\mathcal{A}(x)) \tag{70}$$

Unless explicitly mentioned, we assume throughout section 5 that the norm used to define the distance function $d_{\mathcal{K}^*}$ is arbitrary.

Given an approximate solution $\tilde{x} \in X$ for the penalized problem (70), the following result provides bounds on the primal infeasibility and optimality gap of $\tilde{x}$ with respect to (1).

Proposition 30 *If $\tilde{x} \in X$ is a δ-approximate solution of (70), i.e., it satisfies*

$$\Phi_\theta(\tilde{x}) - \Phi_\theta^* \leq \delta, \tag{71}$$

for some $\theta > \|\lambda^*\|$, where λ^* is an arbitrary Lagrange multiplier for (1), then

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \frac{\delta}{\theta - \|\lambda^*\|}, \tag{72}$$

$$f(\tilde{x}) - f(x^*) \leq \delta. \tag{73}$$

Proof. Using the fact that $v(0) = f^* \geq \Phi_\theta^*$, Corollary 2 and assumption (71), we conclude that

$$\begin{aligned} \delta &\geq \Phi_\theta(\tilde{x}) - \Phi_\theta^* = f(\tilde{x}) + \theta\, d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) - \Phi_\theta^* \\ &\geq f(\tilde{x}) - f^* + \theta\, d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \geq -\|\lambda^*\|\, d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) + \theta\, d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})), \end{aligned}$$

which clearly implies both (72) and (73). ∎

Note that in view of Proposition 1(a) we can reformulate the penalized problem (70) as the smooth saddle-point problem

$$\min_{x \in X} f(x) + \theta \max_{y \in Y} \langle \mathcal{A}(x), y \rangle, \tag{74}$$

where $Y := (-\mathcal{K}) \cap B(0,1)$. Clearly, this problem can be solved by means of Nesterov's smooth approximation scheme discussed in Subsection 3.3. Using Proposition 30 and applying the results of Subsection 3.3 to the above reformulation, we can now derive an iteration-complexity bound for Nesterov's smooth approximation scheme to compute an (ϵ_p, ϵ_o)-primal solution of (1) by approximately solving the saddle point problem (74) for some value of the penalty parameter θ.

Theorem 31 *Let λ^* be an arbitrary Lagrange multiplier for (1) and let $(\epsilon_p, \epsilon_o) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. If*

$$\theta = \theta(t) := \frac{\epsilon_o}{\epsilon_p} + t, \tag{75}$$

for some $t \geq \|\lambda^*\|$, then Nesterov's smooth approximation scheme applied to problem (74) finds an (ϵ_p, ϵ_o)-primal solution of (1) in at most

$$N_e(t) := \left\lceil \sqrt{\frac{8D_h(x_0^{sd})}{\sigma_h}} \left\{ \sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \sqrt{\frac{2D_{\tilde{h}}}{\sigma_{\tilde{h}}}} \left(\frac{1}{\epsilon_p} + \frac{t}{\epsilon_o} \right) \right\} \right\rceil. \tag{76}$$

iterations, where $D_h(x_0^{sd})$ and $D_{\tilde{h}}$ are defined in Theorem 10 and relation (28), respectively. In particular, if the norm $\|\cdot\|$ on $\Re^m$ is the inner product norm and $\tilde{h}(\cdot) = \|\cdot\|^2/2$, then the above bound reduces to

$$N_e(t) := \left\lceil \sqrt{\frac{8D_h(x_0^{sd})}{\sigma_h}} \left\{ \sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \left(\frac{1}{\epsilon_p} + \frac{t}{\epsilon_o} \right) \right\} \right\rceil . \tag{77}$$

Proof. Let $\tilde{x} \in X$ be a point satisfying (71) with $\delta = \epsilon_o$. It follows from (75), the assumption that $t \geq \|\lambda^*\|$ and Proposition 30 that $f(\tilde{x}) - f^* \leq \epsilon_o$ and

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \frac{\delta}{\theta - \|\lambda^*\|} \leq \frac{\epsilon_o}{\epsilon_o/\epsilon_p} = \epsilon_p,$$

and hence that $\tilde{x}$ is an (ϵ_p, ϵ_o)-primal solution of (1). Using Theorem 10 with $\|\mathcal{E}\| = \theta\|\mathcal{A}\|$ and $L_{\hat{\phi}} = L_f$, we conclude that Nesterov's smooth approximation scheme finds an approximate solution $\tilde{x}$ as above in at most

$$\sqrt{\frac{8D_h(x_0^{sd})}{\sigma_h \epsilon_o} \left(\frac{2\theta^2 D_{\tilde{h}} \|\mathcal{A}\|^2}{\sigma_{\tilde{h}} \epsilon_o} + L_f \right)},$$

iterations. Substituting the value of θ given by (75) in this iteration bound and using some trivial majorization, we obtain bound (76). The last statement follows from the fact $\sigma_{\tilde{h}} = 1$ and $D_{\tilde{h}} = 1/2$ when the norm $\|\cdot\|$ on $\Re^m$ is the inner product norm and $\tilde{h}(\cdot) = \|\cdot\|^2/2$. ∎

We now make a few observations about the iteration-complexity bound $N_e(t)$ obtained in Theorem 31. First, note that if the norm on $\Re^n$ is the inner product one, then the bound $N_e(t)$ on (77) reduces to

$$N_e(t) := \left\lceil 2D_X \left\{ \sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \left(\frac{1}{\epsilon_p} + \frac{t}{\epsilon_o} \right) \right\} \right\rceil . \tag{78}$$

Second, it is interesting to compare the bound $N_e(t)$ in (77) with the bound $N_p(t)$ obtained in Theorem 15. Indeed, if

$$\chi_t := \sqrt{\frac{t\epsilon_p}{\epsilon_o}} = \mathcal{O}\left(1 + \frac{\epsilon_p}{\|\mathcal{A}\|} \sqrt{\frac{L_f}{\epsilon_o}} \right),$$

then it is easy to see that $N_p(t)/N_e(t) = \mathcal{O}(1)$. Otherwise, we have

$$\frac{N_p(t)}{N_e(t)} = \frac{\mathcal{O}(1)}{\chi_t}.$$

Moreover, observations similar to ones made in the paragraph following Theorem 15 applies to the above result as well. In particular, a "guess and check" procedure similar to the Search Procedure 1 in Subsection 4.1.1 can be stated as follows.

Search Procedure 2:

1) Set $k = 0$ and define

$$\beta_1 := \sqrt{\frac{8D_h(x_0^{sd})}{\sigma_h}} \left(\sqrt{\frac{L_f}{\epsilon_o}} + \frac{\|\mathcal{A}\|}{\epsilon_p} \sqrt{\frac{2D_{\tilde{h}}}{\sigma_{\tilde{h}}}} \right), \quad \beta_2 := \frac{4\|\mathcal{A}\|}{\epsilon_o} \sqrt{\frac{D_h(x_0^{sd}) D_{\tilde{h}}}{\sigma_h \sigma_{\tilde{h}}}}, \quad t_0 := \frac{\max(1, \beta_1)}{\beta_2}.$$

2) Set $\theta = \theta(t_k)$, and perform at most $\lceil N_e(t_k) \rceil$ iterations of Nesterov's optimal method applied to problem (74). If an iterate $\tilde{x}$ is obtained such that (71) with $\delta = \epsilon_o$ and $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p$ are satisfied, then **stop**; otherwise, go to step 3;

3) Set $t_{k+1} = 2t_k$, $k = k + 1$, and go to step 2.

The following result, whose proof is similar to that of Corollary 18, gives the iteration-complexity of Search Procedure 2 for obtaining an (ϵ_p, ϵ_o)-primal solution of (1).

Corollary 32 *Let λ^* be the minimum norm Lagrange multiplier for (1). Then, the overall number of iterations of Search Procedure 2 for obtaining an (ϵ_p, ϵ_o)-primal solution of (1) is bounded by $\mathcal{O}(N_e(\|\lambda^*\|))$, where $N_e(\cdot)$ is defined in (37).*

5.2 Exact penalty method applied to a perturbed problem

In this subsection, we will derive the iteration-complexity of solving (1) by applying the exact penalty method to the perturbed problem (48) for a properly chosen perturbation parameter $\gamma > 0$. We assume throughout this subsection that the norm on $\Re^n$ is the inner product one.

The exact penalty problem associated with (48) is given by

$$\Phi^*_{\theta,\gamma} := \inf_{x \in X} \left\{ \Phi_{\theta,\gamma}(x) := f_\gamma(x) + \theta d_{\mathcal{K}^*}(\mathcal{A}(x)) = f(x) + \theta d_{\mathcal{K}^*}(\mathcal{A}(x)) + \frac{\gamma}{2}\|x - x_0\|^2 \right\}. \tag{79}$$

Problem (79) is a non-smooth optimization problem and it can be reformulated as the following smooth saddle-point problem

$$\inf_{x \in X} \left\{ \Phi_{\theta,\gamma}(x) = f_\gamma(x) + \max_{y \in Y} \langle \theta \mathcal{A}(x), y \rangle \right\}, \tag{80}$$

where $Y := (-\mathcal{K}) \cap B(0,1)$. The function $\Phi_{\theta,\gamma}$ is non-smooth but, in view of Nesterov's smooth approximation scheme, it can be closely approximated by the following function with Lipschitz-continuous gradient:

$$\bar{\Phi}^\eta_{\theta,\gamma} := \inf_{x \in X} \left\{ \Phi^\eta_{\theta,\gamma}(x) := f_\gamma(x) + \max_{y \in Y} \left\{ \langle \theta \mathcal{A}(x), y \rangle - \eta \tilde{h}(y) \right\} \right\}, \tag{81}$$

where $\eta > 0$ is the smooth parameter (see (27)) and $\tilde{h} : Y \to \mathbf{R}$ is as in Subsection 3.3. In view of Proposition 9(b) with $\mathcal{E} = \theta\mathcal{A}$, $\psi = 0$, $\hat{\phi} = f_\gamma$ and $\phi_\eta = \Phi^\eta_{\theta,\gamma}$, the function $\Phi^\eta_{\theta,\gamma}$ has $L^\eta_{\theta,\gamma}$-Lipschitz continuous gradient where

$$L^\eta_{\theta,\gamma} := L_f + \gamma + \frac{\theta^2 \|\mathcal{A}\|^2}{\sigma_{\tilde{h}} \eta}. \tag{82}$$

With the aid of reformulation (80) and Proposition 30, we can now derive an iteration-complexity bound for solving (1) by applying the exact penalty method to the perturbation problem (48) with an appropriately chosen parameter $\gamma > 0$.

Theorem 33 *Let λ^*_γ be an arbitrary Lagrange multiplier for (48) and let $(\epsilon_p, \epsilon_o) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. If*

$$\theta = \tilde{\theta}(t) := 2t, \ \eta = \frac{\min\{\epsilon_o/2, t\epsilon_p\}}{2D_{\tilde{h}}}, \ \text{and } \gamma = \frac{\epsilon_o}{D_X^2}, \tag{83}$$

for some $t \geq \|\lambda_\gamma^*\|$, then the variant of Nesterov's optimal method of Theorem 8 with $\mu = \gamma$ and $L_\phi = L_{\theta,\gamma}^\eta$ applied to problem (81) finds an (ϵ_p, ϵ_o)-primal solution of (1) in at most

$$\tilde{N}_e(t) := \left\lceil 8D_X \left[\sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \sqrt{\frac{2D_{\tilde{h}}}{\sigma_{\tilde{h}}}} \left(\frac{\sqrt{8}\, t}{\epsilon_o} + \frac{2\sqrt{t}}{\sqrt{\epsilon_p \epsilon_o}} \right) \right] + 8 \right\rceil \max \left\{ 1, \left\lceil \log \left(\frac{\epsilon_o}{t\epsilon_p} \right) \right\rceil \right\} \tag{84}$$

iterations, where D_X is defined in Theorem 15. In particular, if the norm on $\Re^m$ is the inner product one and $\tilde{h}(\cdot) = \|\cdot\|^2/2$, then the above bound reduces to

$$\tilde{N}_e(t) := \left\lceil 8D_X \left[\sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \left(\frac{\sqrt{8}\, t}{\epsilon_o} + \frac{2\sqrt{t}}{\sqrt{\epsilon_p \epsilon_o}} \right) \right] + 8 \right\rceil \max \left\{ 1, \left\lceil \log \left(\frac{\epsilon_o}{t\epsilon_p} \right) \right\rceil \right\}. \tag{85}$$

Proof. Define $\delta := \min\{\epsilon_o/2, t\epsilon_p\} = 2\eta D_{\tilde{h}}$, where the last equality is due to (75). Also, let $\tilde{x} \in X$ be an $\delta/2$-approximate solution of (81), i.e., $\Phi_{\theta,\gamma}^\eta(\tilde{x}) - \bar{\Phi}_{\theta,\gamma}^\eta \leq \delta/2$. Using the fact that $\eta D_{\tilde{h}} = \delta/2$ and Proposition 9(a) with $\phi = \Phi_{\theta,\gamma}$ and $\phi_\eta = \Phi_{\theta,\gamma}^\eta$, we easily see that that $\tilde{x}$ is a δ-approximate solution of (80), i.e., $\Phi_{\theta,\gamma}(\tilde{x}) - \Phi_{\theta,\gamma}^* \leq \delta$. It then follows from (83), the assumption $t \geq \|\lambda_\gamma^*\|$, the definition of δ, and Proposition 30 with $f = f_\gamma$ and $\Phi_\theta = \Phi_{\theta,\gamma}$ that

$$\begin{aligned} d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) &\leq \frac{\delta}{\theta - \|\lambda_\gamma^*\|} = \frac{\delta}{2t - \|\lambda_\gamma^*\|} \leq \frac{t\epsilon_p}{t} = \epsilon_p \\ f_\gamma(\tilde{x}) - f_\gamma^* &\leq \delta \leq \epsilon_o/2. \end{aligned}$$

The latter inequality together with (51) and (83) imply that

$$f(\tilde{x}) - f^* \leq f_\gamma(\tilde{x}) - f^* \leq [f_\gamma(\tilde{x}) - f_\gamma^*] + [f_\gamma^* - f^*] \leq \frac{\epsilon_o}{2} + \frac{\epsilon_o}{2} \leq \epsilon_o.$$

We have thus shown that $\tilde{x}$ is an (ϵ_p, ϵ_o)-primal solution of (1).

To end the proof, it remains to derive the iteration-complexity for the variant of Nesterov's optimal method of Theorem 8 with $\mu = \gamma$ and $L_\phi = L_{\theta,\gamma}^\eta$ to obtain a $\delta/2$-approximate solution of problem (81). First, observe that $\Phi_{\theta,\gamma}^\eta$ is strongly convex with modulus γ. Now, using Theorem 8 with $\phi = \Phi_{\theta,\gamma}^\eta$, $\mu = \gamma := \epsilon_o/D_X^2$ and $\epsilon = \delta/2$, and noting that the definition of δ and the definition of γ in (83) imply that

$$Q = \frac{\mu\|x_0^{sd} - x^*\|^2}{2\epsilon} = \frac{\gamma\|x_0^{sd} - x^*\|^2}{\delta} \leq \frac{\gamma D_X^2}{\delta} = \frac{\epsilon_o}{\min\{\epsilon_o/2, \epsilon_p t\}} = \max\left\{2, \frac{\epsilon_o}{\epsilon_p t}\right\},$$

we conclude that the number of iterations for the aforementioned variant to find a $\epsilon = \delta/2$-approximate solution $\tilde{x}$ as above is bounded by

$$\left\lceil 8\sqrt{\frac{L_{\theta,\gamma}^\eta}{\gamma}} \right\rceil \lceil \log Q \rceil = \left\lceil 8\sqrt{\frac{L_{\theta,\gamma}^\eta}{\gamma}} \right\rceil \max\left\{1, \left\lceil \log \frac{\epsilon_o}{t\epsilon_p} \right\rceil\right\}.$$

The result now follows by noting that (82) and (83) imply that

$$\begin{aligned} \sqrt{\frac{L_{\theta,\gamma}^\eta}{\gamma}} &= \sqrt{\frac{L_f}{\gamma} + 1 + \frac{\theta^2\|\mathcal{A}\|^2}{\gamma\sigma_{\tilde{h}}\eta}} = \sqrt{\frac{D_X^2 L_f}{\epsilon_o} + 1 + \frac{8t^2 D_X^2 D_{\tilde{h}} \|\mathcal{A}\|^2}{\epsilon_o \sigma_{\tilde{h}} \min\{\epsilon_o/2, t\epsilon_p\}}} \\ &\leq D_X \left[\sqrt{\frac{L_f}{\epsilon_o}} + \|\mathcal{A}\| \sqrt{\frac{2D_{\tilde{h}}}{\sigma_{\tilde{h}}}} \left(\frac{\sqrt{8}\, t}{\epsilon_o} + \frac{2\sqrt{t}}{\sqrt{\epsilon_p \epsilon_o}} \right) \right] + 1. \end{aligned}$$

■

The following result, whose proof is similar to that of Corollary 25, gives the iteration-complexity of Search Procedure 2 for obtaining an (ϵ_p, ϵ_o)-primal solution of (1).

Corollary 34 *Let λ^* be the minimum norm Lagrange multiplier for (1). Consider the "guess and check" procedure similar to Search Procedure 2 where the functions N_e and θ are replaced by the functions $\tilde{N}_e$ and $\tilde{\theta}$ defined in Theorem 33, δ is set to $\delta := \min\{\epsilon_o/2, t_k\epsilon_p\}$, and t_0 is set to $\min\{\max(\beta_0,1)/\beta_1, [\max(\beta_0,1)/\beta_2]^2\}$ with*

$$\beta_0 = 8D_X\frac{L_f}{\epsilon_o} + 8, \quad \beta_1 = \frac{8\sqrt{8}D_X\|\mathcal{A}\|}{\epsilon_o}, \quad \beta_2 = \frac{16D_X\|\mathcal{A}\|}{\sqrt{\epsilon_p\epsilon_o}}.$$

Then, the overall number of iterations of this "guess and check" procedure for obtaining an (ϵ_p, ϵ_o)-primal-dual solution of (1) is bounded by $\mathcal{O}(\tilde{N}_e(\|\lambda^*_\gamma\|))$, where $\tilde{N}_e(\cdot)$ is defined in (85).

It is interesting to compare the two functions $N_e(t)$ and $\tilde{N}_e(t)$ defined in (78) and (85), respectively. Indeed, both functions are comparable for the case when $r_t := \epsilon_o/(t\epsilon_p) = \mathcal{O}(1)$. Otherwise, if $r_t = \Omega(1)$ and, we further assume that the terms involving t dominate the other terms in the right hand side of (85), i.e.,

$$D_X\sqrt{\frac{L_f}{\epsilon_o}} + 1 = \mathcal{O}\left(D_X\|\mathcal{A}\|\left(\frac{\sqrt{8}\,t}{\epsilon_o} + \frac{2\sqrt{t}}{\sqrt{\epsilon_p\epsilon_o}}\right)\right),$$

then

$$\frac{\tilde{N}_e(t)}{N_e(t)} \leq \mathcal{O}(1)\left(\frac{\log r_t}{\sqrt{r_t}}\right),$$

where $\mathcal{O}(1)$ denotes an absolute constant. Hence, in the latter case, we conclude that the function $\tilde{N}_e(t)$ is asymptotically smaller than the function $N_e(t)$.

Appendix

In this section, we prove Propositions 1 and 24.

Proof of Proposition 1. Define $\mathcal{C} := \{(v,t) \in \Re^m \times \mathbf{R} : \|v\| \leq t\}$ and let $\mathcal{C}^*$ denote the dual cone of $\mathcal{C}$. It is easy to see that $\mathcal{C}^* = \{(\tilde{v},\tilde{t}) \in \Re^m \times \mathbf{R} : \|\tilde{v}\|_* \leq \tilde{t}\}$. By definition of $d_{\mathcal{K}^*}$ and conic duality, we have

$$\begin{aligned}
d_{\mathcal{K}^*}(u) &= \inf_{\tilde{k},\tilde{t}}\{\tilde{t} : \|u - \tilde{k}\|_* \leq \tilde{t},\ \tilde{k} \in \mathcal{K}^*\} \\
&= \inf_{\tilde{k},\tilde{t},\tilde{v}}\{\tilde{t} : \tilde{v} + \tilde{k} = u,\ (\tilde{v},\tilde{t}) \in \mathcal{C}^*,\ \tilde{k} \in \mathcal{K}^*\} \\
&= \sup_{(v,k,t)}\{\langle u, y\rangle : t = 1,\ y + v = 0,\ y + k = 0,\ (v,t) \in \mathcal{C},\ k \in \mathcal{K}\} \\
&= \sup\{\langle u, y\rangle : (-y, 1) \in \mathcal{C},\ -y \in \mathcal{K}\} \ = \ \sup\{\langle u, y\rangle : y \in (-\mathcal{K}) \cap B(0,1)\}.
\end{aligned}$$

Statement (a) follows from the above identity and the definition of the support function of a set (see Subsection 1.1).

To show statement (b), let $u \in \Re^m$ and $\lambda \in \mathcal{K}$ be given and assume without loss of generality that $\lambda \neq 0$. Now noting that $-\lambda/\|\lambda\| \in C := (-\mathcal{K}) \cap B(0,1)$, we conclude from the above identity that $d_{\mathcal{K}}(u) \geq \langle u, -\lambda/\|\lambda\| \rangle$, or equivalently, $\langle u, \lambda \rangle \geq -\|\lambda\| \, d_{\mathcal{K}}(u)$. ∎

Our goal in the remaining part of this section is to prove Proposition 24. We first start with an easy case of the result.

Lemma 35 *For some positive integer L, let positive scalars $p_1, p_2, \cdots, p_L$ be given and define*

$$C_0 := 2 + \max\left\{2, \max_{1 \leq l \leq L} \left(\frac{2}{p_l} + \frac{4^{p_l}}{2^{p_l} - 1} \right) \right\}.$$

Then, for any positive scalars $\beta_0, \beta_1, \cdots, \beta_L$, and $\bar{t} > t_0$, we have

$$\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil \leq C_0 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil, \tag{86}$$

where K and $t_0, \ldots, t_K$ are defined in (56).

Proof. Without loss of generality, assume that $t_0 = (\max(\beta_0, 1)/\beta_1)^{1/p_1}$. Clearly, due to the definition of K in (56) and the assumption $\bar{t} > t_0$, we have $K < \log(\bar{t}/t_0) + 1$, and hence that $t_0 2^{K+1} < 4\bar{t}$. Using these relations and the inequality $\log x = (\log x^p)/p \leq x^p/p$ for any $x > 0$, $p > 0$, we obtain

$$\begin{aligned}
\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil &\leq \sum_{k=0}^{K} \left(1 + \beta_0 + \sum_{l=1}^{L} \beta_l t_0^{p_l} 2^{p_l k} \right) \leq (1+\beta_0)(1+K) + \sum_{l=1}^{L} \beta_l t_0^{p_l} \frac{2^{(K+1)p_l}}{2^{p_l} - 1} \\
&\leq (1+\beta_0) \left[2 + \log\left(\frac{\bar{t}}{t_0} \right) \right] + \sum_{l=1}^{L} \beta_l \frac{(4\bar{t})^{p_l}}{2^{p_l} - 1} \\
&\leq (1+\beta_0) \left[2 + \frac{1}{p_1} \left(\frac{\bar{t}}{t_0} \right)^{p_1} \right] + \sum_{l=1}^{L} \frac{4^{p_l}}{2^{p_l} - 1} \beta_l \bar{t}^{p_l} \\
&\leq (1+\beta_0) \left[2 + \frac{1}{p_1} \left(\frac{\beta_2 \bar{t}^{p_1}}{\max(\beta_1, 1)} \right) \right] + \sum_{l=1}^{L} \frac{4^{p_l}}{2^{p_l} - 1} \beta_l \bar{t}^{p_l} \\
&\leq 2(1+\beta_0) + \frac{2}{p_1} \beta_1 \bar{t}^{p_1} + \sum_{l=1}^{L} \frac{4^{p_l}}{2^{p_l} - 1} \beta_l \bar{t}^{p_l} \\
&\leq 2 + \max\left\{ 2, \frac{2}{p_1} + \frac{4^{p_1}}{2^{p_1} - 1}, \max_{2 \leq l \leq L} \frac{4^{p_l}}{2^{p_l} - 1} \right\} (\beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l}).
\end{aligned}$$

Inequality (86) now clearly follows from the above conclusion, Lemma 16, and some trivial majorization. ∎

The following technical lemma provides an useful inequality to prove a more difficult case of our result.

Lemma 36 *Let the positive scalars a and p be given. For any $0 \le K \le a - 1/(p \ln 2) - 1$, we have*

$$\sum_{k=0}^{K} 2^{pk}(a-k) \le \frac{2^{p(K+1)}}{p \ln 2} \left[a - (K+1) + \frac{1}{p \ln 2} \right].$$

Proof. Noting that the function $\psi(s) := 2^{ps}(a-s)$ is non-decreasing for any $s \le a - 1/(p \ln 2)$, we obtain

$$\begin{aligned} \sum_{k=0}^{K} 2^{pk} (a-k) &\le \int_0^{K+1} 2^{ps} (a-s)\, ds = \frac{2^{ps}}{p \ln 2} \left(a - s + \frac{1}{p_l \ln 2} \right) \Big|_0^{K+1} \\ &\le \frac{2^{p(K+1)}}{p \ln 2} \left[a - (K+1) + \frac{1}{p_l \ln 2} \right]. \end{aligned}$$

■

We now prove a more difficult case of the result.

Lemma 37 *For some positive integer L, let positive scalars $p_1, p_2, \cdots, p_L$ be given and define*

$$C_1 := 1 + \frac{1}{(\ln 2)\, \min_{1 \le l \le L} p_l}, \tag{87}$$

$$C_2 := 5 + \max \left\{ 5, \max_{1 \le l \le L} \left(\frac{4}{p_l^2} + \frac{7}{p_l} + 2C_1 4^{p_l} \right) \right\}. \tag{88}$$

Then, for any positive scalars $\beta_0, \beta_1, \cdots, \beta_L$, and ν, we have

$$\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil \max \left\{ 1, \left\lceil \log \frac{\nu}{t_k} \right\rceil \right\} \le C_2 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil \max \left\{ 1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \right\} \tag{89}$$

for every $\bar{t} \in (t_0, \tilde{t}\,]$, where $\tilde{t} := 2^{-C_1}\nu$ and the scalars K and $t_0, \ldots, t_K$ are defined in (56).

Proof. Without loss of generality, assume that $t_0 = (\max(\beta_0, 1)/\beta_1)^{1/p_1}$. Clearly, due to the definition of K in (56) and the assumption $\bar{t} > t_0$, we have $\log(\bar{t}/t_0) \le K \le \log(\bar{t}/t_0) + 1$. Using these relations, the inequalities $\log x = (\log x^p)/p \le x^p/p$ for any $x > 0$, $p > 0$, and $(\log x)^2 = (2 \log x^{p/2})/p)^2 \le 4x^p/p^2$ for any $x \ge 1$, $p > 0$, and the facts $\log(\nu/\bar{t}) \ge C_1 \ge 1$ and $\bar{t}/t_0 \ge 1$ due to

the assumption $t_0 < \bar{t} \le 2^{-C_1}\nu$, we obtain

$$\begin{aligned}
\sum_{k=0}^{K}\left(\log\frac{\nu}{t_k}+1\right) &= \sum_{k=0}^{K}\left(\log\frac{\nu}{t_0}+1-k\right) \;=\; (K+1)\left(\log\frac{\nu}{t_0}+1-\frac{K}{2}\right)\\
&= (K+1)\left(\log\frac{\nu}{t_0}-K+1+\frac{K}{2}\right) \;\le\; (K+1)\left(\log\frac{\nu}{\bar{t}}+1+\frac{K}{2}\right)\\
&= \frac{1}{2}\left[2(K+1)\log\frac{\nu}{\bar{t}}+K^2+3K+2\right] \;\le\; \frac{1}{2}(K^2+5K+4)\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil\\
&\le \frac{1}{2}\left(\log^2\frac{\bar{t}}{t_0}+7\log\frac{\bar{t}}{t_0}+10\right)\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil\\
&\le \frac{1}{2}\left[\left(\frac{4}{p_1^2}+\frac{7}{p_1}\right)\left(\frac{\bar{t}}{t_0}\right)^{p_1}+10\right]\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil\\
&= \frac{1}{2}\left[\left(\frac{4}{p_1^2}+\frac{7}{p_1}\right)\frac{\beta_1\bar{t}^{p_1}}{\max\{1,\beta_0\}}+10\right]\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil\\
&\le \left[\left(\frac{4}{p_1^2}+\frac{7}{p_1}\right)\frac{\beta_1\bar{t}^{p_1}}{1+\beta_0}+5\right]\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil. \qquad (90)
\end{aligned}$$

Moreover, it follows from the fact $K \le \log(\bar{t}/t_0)+1$ and the assumption $t_0 < \bar{t} \le 2^{-C_1}\nu$ that

$$\begin{aligned}
K+1 &\le \log\frac{\bar{t}}{t_0}+2 = \log\frac{2\nu}{t_0}+\log\frac{\bar{t}}{2\nu}+2 \le \log\frac{2\nu}{t_0}-C_1+1\\
&= \log\frac{2\nu}{t_0}-\frac{1}{(\ln 2)\,\min_{1\le l\le L} p_l} \le \log\frac{2\nu}{t_0}-\frac{1}{(\ln 2)\,p_l},\;\; \forall\, 1\le l\le L,
\end{aligned}$$

which, together with Lemma 36 (with $a=\log(2\nu/t_0)$ and $p=p_l$) and the fact that $K \ge \log(\bar{t}/t_0)$, then imply that

$$\begin{aligned}
\sum_{k=0}^{K}2^{p_l k}\left(\log\frac{2\nu}{t_0}-k\right) &\le \frac{2^{p_l(K+1)}}{p_l\ln 2}\left(\log\frac{2\nu}{t_0}-(K+1)+\frac{1}{p_l\ln 2}\right)\\
&\le \frac{2^{p_l(K+1)}}{p_l\ln 2}\left(\log\frac{2\nu}{t_0}-(\log\frac{\bar{t}}{t_0}+1)+\frac{1}{p_l\ln 2}\right)\\
&= \frac{2^{p_l(K+1)}}{p_l\ln 2}\left(\log\frac{\nu}{\bar{t}}+\frac{1}{p_l\ln 2}\right) \;\le\; \frac{2}{p_l\ln 2}2^{p_l(K+1)}\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil\\
&\le 2C_1 2^{p_l(K+1)}\left\lceil\log\frac{\nu}{\bar{t}}\right\rceil, \qquad (91)
\end{aligned}$$

where the last two inequalities follow from the relation $\log(\nu/\bar{t}) \ge \log(\nu/\tilde{t}) = C_1 \ge 1/(p_l\ln 2)$.

By the definitions of K and $t_k, \forall k=1,\cdots,K$ in (56) and the assumption $\bar{t} \le 2^{-C_0}\nu$, we have $t_k = t_0 2^k \le t_0 2^{\log(\bar{t}/t_0)+1} = 2\bar{t} \le 2^{-C_0+1}\nu$, which implies that $\log(\nu/t_k) \ge C_0 - 1 > 0$ and hence that $\max\{1,\lceil\log(\nu/t_k)\rceil\} = \lceil\log(\nu/t_k)\rceil \le \log(\nu/t_k)+1$. Using these relations, (90), (91), the relation

$t_0 2^{K+1} < 4\bar{t}$ due to $K < \log(\bar{t}/t_0) + 1$, we obtain

$$\begin{aligned}
\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil\right\} &\leq \sum_{k=0}^{K} \left(1 + \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l}\right) \left(\log \frac{\nu}{t_k} + 1\right) \\
&= (1+\beta_0) \sum_{k=0}^{K} \left(\log \frac{\nu}{t_k} + 1\right) + \sum_{l=1}^{L} \sum_{k=0}^{K} \beta_l t_0^{p_l} 2^{k p_l} \left(\log \frac{\nu}{t_k} + 1\right) \\
&= (1+\beta_0) \sum_{k=0}^{K} \left(\log \frac{\nu}{t_k} + 1\right) + \sum_{l=1}^{L} \sum_{k=0}^{K} \beta_l t_0^{p_l} 2^{k p_l} \left(\log \frac{2\nu}{t_0} - k\right) \\
&\leq \left\{5(1+\beta_0) + \left(\frac{4}{p_1^2} + \frac{7}{p_1}\right) \beta_1 \bar{t}^{p_1} + 2C_0 \sum_{l=1}^{L} \beta_l t_0^{p_l} 2^{p_l (K+1)}\right\} \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \\
&\leq \left[5(1+\beta_0) + \left(\frac{4}{p_1^2} + \frac{7}{p_1}\right) \beta_1 \bar{t}^{p_1} + 2C_0 \sum_{l=1}^{L} 4^{p_l} \beta_l \bar{t}^{p_l}\right] \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \\
&\leq \left[5 + \max\left\{5, \frac{4}{p_1^2} + \frac{7}{p_1} + 2C_0 4^{p_1}, 2C_0 \max_{2 \leq l \leq L} 4^{p_l}\right\} \left(\beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l}\right)\right] \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil .
\end{aligned}$$

Inequality (89) now immediately follows from the above conclusion, Lemma 16, the fact that $\max\{1, \lceil \log(\nu/\bar{t}) \rceil\} = \lceil \log(\nu/\bar{t}) \rceil$ due to the assumption $\bar{t} \leq \nu 2^{-C_1}$, and some trivial majorization. ∎

We are now ready to prove Proposition 24.

Proof of Proposition 24. Assume first that $\bar{t} \leq t_0$. Due to the definition of K in (56), we have $K = 0$ in this case, which, in view of the definition of t_0 in (56) and Lemma 16, then imply that

$$\begin{aligned}
\sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil\right\} &= \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_0^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_0} \right\rceil\right\} \\
&\leq \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_0^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\} \leq \lceil \beta_0 + L \max\{\beta_0, 1\} \rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\} \\
&\leq \lceil (L+1)\beta_0 + L \rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\} \leq (2L+1) \lceil \beta_0 \rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\} \\
&\leq (2L+1) \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil\right\} .
\end{aligned}$$

Hence, in the case where $\bar{t} \leq t_0$, inequality (55) holds with $C = 2L + 1$.

Assume now that $\bar{t} > t_0$. Denoting $\tilde{t} := 2^{-C_1}\nu$, where C_1 is given in (87), we further consider two subcases: a) $\tilde{t} \geq \bar{t}$; b) $\tilde{t} \leq \bar{t}$. In subcase a), we have $\tilde{t} \geq \bar{t} > t_0$, which, in view of Lemma 37, clearly implies that inequality (55) holds with $C = C_2$.

We now consider the remaining subcase b) where $\tilde{t} \leq \bar{t}$. Denoting $\tilde{K} := \max\{0, \lceil \log(\tilde{t}/t_0) \rceil\}$, we have $t_k \geq \tilde{t}$ for any $k \geq \tilde{K}$ and hence that,

$$\max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil\right\} \leq \max\left\{1, \left\lceil \log \frac{\nu}{\tilde{t}} \right\rceil\right\} = \max\left\{1, \lceil C_1 \rceil\right\} \leq C_1 + 1, \, \forall k \geq \tilde{K},$$

which together with Lemma 35, then imply that

$$\begin{aligned}
&\sum_{k=\tilde{K}}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l\, t_k^{p_l}\right) \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{t_k} \right\rceil \right\} \\
&\quad \leq \quad (C_1+1) \sum_{k=\tilde{K}}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l\, t_k^{p_l}\right) \right\rceil \leq (C_1+1) \sum_{k=0}^{K} \left\lceil \beta_0 + \sum_{l=1}^{L} \left(\beta_l\, t_k^{p_l}\right) \right\rceil \\
&\quad \leq \quad C_0(C_1+1) \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \right\}.
\end{aligned}$$

Moreover, if $\tilde{K} \geq 1$, i.e., $\tilde{t} > t_0$, it then follows from Lemma 37 with $\bar{t} = \tilde{t}$ that

$$\begin{aligned}
\sum_{k=0}^{\tilde{K}-1} \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l t_k^{p_l} \right\rceil \left\lceil \log \frac{\nu}{t_k} \right\rceil \quad &\leq \quad C_2 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \tilde{t}^{p_l} \right\rceil \left\lceil \log \frac{\nu}{\tilde{t}} \right\rceil = C_2 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \tilde{t}^{p_l} \right\rceil \lceil C_1 \rceil \\
&\leq \quad C_2 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil \lceil C_1 \rceil \\
&\leq \quad (C_1+1) C_2 \left\lceil \beta_0 + \sum_{l=1}^{L} \beta_l \bar{t}^{p_l} \right\rceil \max\left\{1, \left\lceil \log \frac{\nu}{\bar{t}} \right\rceil \right\},
\end{aligned}$$

where the second inequality follows from the assumption $\tilde{t} \leq \bar{t}$. Combining the previous two conclusions, we conclude that (55) holds with $C = (C_1+1)(C_0+C_2)$. ∎

References

[1] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\wr(1/\epsilon)$ iteration-complexity for cone programming. Manuscript, School of Industrial Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, December 2006.

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

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

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