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

Guanghui Lan † Renato D.C. Monteiro ‡

July 24, 2008 (Revised July 10, 2011, March 9, 2012)

Abstract

This paper considers a special but broad class of convex programming (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. It studies the computational complexity of quadratic penalty based methods for solving the above class of problems. An iteration of these methods, which is simply an iteration of Nesterov's optimal method (or one of its variants) for approximately solving a smooth penalization subproblem, consists of one or two projections onto the simple convex set. Iteration-complexity bounds expressed in terms of the latter type of iterations are derived for two quadratic penalty based variants, namely: one which applies the quadratic penalty method directly to the original problem and another one which applies the latter method to a perturbation of the original problem obtained by adding a small quadratic term to its objective function.

Keywords: Convex programming, quadratic 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

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

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 [6, 8] developed a method which finds a point $x \in X$ such that $f(x) - f^* \leq \epsilon$ in at most $\mathcal{O}(\epsilon^{-1/2})$ iterations. Moreover, each iteration of his method requires one gradient evaluation of f and computation

*The work of the first author was partially supported by NSF Grants CCF-0430644 and CMMI-1000347. The work of the second author was partially supported by NSF Grants CCF-0430644, CCF-0808863 and CMMI-0900094 and ONR Grants N00014-08-1-0033 and N00014-11-1-0062.

†Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611. (email: glan@ise.ufl.edu).

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

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. An alternative and natural approach is to investigate first-order methods for solving problem (1) whose iterations consist of, in addition to a couple of gradient evaluations, only projections onto the simple set X.

In this paper, we consider one penalty-based approach for solving (1), namely: the quadratic penalty method. Clearly, it is possible to develop different iteration-complexity bounds for this method depending on the (possibly, many) adopted termination criteria. In our presentation, we adopt a certain natural primal-dual termination criterion which can be described as follows. For the purpose of this local discussion only, assume for simplicity that $\mathcal{K} = \Re^m$ and hence that $\mathcal{K}^* = \{0\}$. Motivated by the optimality condition of (1), define an (ϵ_p, ϵ_d)-*primal-dual* solution of (1) as a pair $(\tilde{x}, \tilde{\lambda}) \in X \times \Re^m$ satisfying

$$\|\mathcal{A}(\tilde{x})\| \leq \epsilon_p, \tag{2}$$ primal_inf0

$$\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}) + \mathcal{B}(\epsilon_d). \tag{3}$$ dual_inf0

Here, $\mathcal{A}_0 := \mathcal{A}(\cdot) - \mathcal{A}(0)$ is the linear part of $\mathcal{A}$, $\mathcal{N}_X(\tilde{x}) := \{s \in \Re^n : \langle s, x - \tilde{x}\rangle \leq 0, \forall x \in X\}$ denotes the normal cone of X at $\tilde{x}$ and $\mathcal{B}(\eta) := \{x \in \Re^n : \|x\| \leq \eta\}$ for every $\eta \geq 0$. The main goal of this paper is to derive the computational-complexity of a variant of the quadratic penalty approach in which the penalty subproblems are solved by Nesterov's optimal method (or, some of its variants), and the overall complexity is expressed in terms of the number of iterations of the latter method.

It is well-known that the penalty parameter of the penalization problem for the above penalty-based approach must be chosen larger than some threshold value so as to ensure that (near) optimal solutions of the penalization problem yield (near) optimal solutions for the original problem (1). Accordingly, we develop a threshold penalty parameter value which depends 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 parameter to this threshold value would yield the best provably iteration-complexity bound. But since $\|\lambda^*\|$, and hence the aforementioned threshold penalty parameter value, is not known a priori, we present an alternative penalty-based approach based on a simple "guess-and-check" procedure for the penalty parameter whose iteration-complexity bound is of the same order as the one for the penalty approach with *known* threshold penalty parameter value. Finally, we present a variant of this guess-and-check quadratic penalty method which consists of applying the guess-and-check quadratic penalty method to the perturbed problem obtained by adding a suitable quadratic perturbation term to the objective function of (1), and show that its iteration-complexity is better than the one for the guess-and-check quadratic penalty method applied directly to (1). More specifically, we show that the iteration-complexity of the variant, after disregarding a few constant factors, is given by

$$\mathcal{O}\left(\frac{1}{(\epsilon_p\epsilon_d)^{\frac{1}{2}}}\log\left(\epsilon_p\epsilon_d\right)^{-\frac{1}{2}}\right), \tag{4}$$ conv_res

while the one for the guess-and-check quadratic penalty method applied directly to (1) is bounded by $\mathcal{O}\left(1/(\epsilon_p\epsilon_d)\right)$.

It is worth mentioning a few other possible approaches for solving problem (1). We first discuss dual methods for solving (1). A natural method is to consider the Lagrangian dual

$$d^* := \max_{\lambda \in \Re^m} d(\lambda) := \min\{f(x) + \langle \lambda, \mathcal{A}(x)\rangle : x \in X\}, \tag{5}$$ dp

and use the subgradient method to solve it. It is well-known that the subgradient method requires $\mathcal{O}(1/\varepsilon^2)$ subgradient evaluations to compute $\lambda \in \Re^m$ such that $d^* - d(\lambda) \leq \varepsilon$. Noting that the computation of a subgradient of d at λ requires us to find a solution of the subproblem $\min\{f(x) + \langle \lambda, \mathcal{A}(x)\rangle : x \in X\}$, it follows that one still has to account for the work involved in computing such solution (or an approximation of it) in the general case where it cannot be expressed in closed form. Another potential dual approach to solve (1) is to use Nesterov's smoothing approximation technique to solve (5). In this case, one would add a small strongly convex perturbation term $h(x)$ to the objective function of the above subproblem, thereby forcing the perturbed dual function $d_h(\lambda) := \min\{f(x) + h(x) + \langle \lambda, \mathcal{A}(x)\rangle : x \in X\}$ to be smooth with Lipschitz continuous gradient. One can then use Nesterov's optimal method for solving the perturbed dual problem in order to obtain an approximate primal-dual solution of (1). We observe that, for the above two methods, one has to deal with the technicality of solving the subproblems only approximately, and hence of working with approximate subgradients of the dual function or its perturbed version. We observe that the overall complexity of these methods taking into account this technical issue has not been fully studied in the literature.

Another possible approach for solving problem (1) is to reformulate it as a monotone variational inequality and use some specific methods for solving the latter problem. More specifically, (1) is equivalent to finding $w^* \in \Omega := X \times \Re^m$ such that

$$\langle w - w^*, F(w^*)\rangle \geq 0, \quad \forall w \in \Omega, \tag{6}$$ eq:def.vip0

where

$$F(w) = F(x, \lambda) := \begin{pmatrix} \nabla f(x) + \mathcal{A}_0^* \lambda \\ -\mathcal{A}(x) \end{pmatrix}. \tag{7}$$ def:F

One can use Korpelevich's algorithm or Tseng's modified forward-backward splitting method whose iteration-complexities have been more recently studied by Nemirovski [5] and Monteiro and Svaiter [3, 4]. Another possibility is to use Nesterov's method proposed in [9]. It is worth mentioning that the complexities derived in [5, 9] assume that the set Ω is bounded and are based on a different weaker termination criterion than the one used in this paper. It is not clear to us how the complexities derived in [5, 9] can be applied to our specific case in which $\Omega = X \times \Re^m$ is unbounded. On the other hand, by considering slightly more general termination criteria, Monteiro and Svaiter [3, 4] develop iteration-complexity bounds for Korpelevich's algorithm as well as Tseng's modified forward-backward splitting method for the case when Ω is unbounded, and hence to our specific case in this paper.

This paper is organized as follows. Section 2 describes the assumptions imposed on (1), introduces the definition of an approximate primal-dual solution of (1) for a general closed convex cone $\mathcal{K}$ and derives a few basic properties of (1). Section 3 discusses some technical results that will be used in our analysis and reviews a first-order algorithm due to Nesterov [6, 8] for solving CP problems with simple feasible sets, and a restarting version of it for solving CP problems with strongly convex objective functions. Section 4 establishes iteration-complexity bounds for quadratic penalty based methods for solving (1). More specifically, Subsection 4.1 presents an iteration-complexity bound for the

quadratic penalty method applied directly to (1) and Subsection 4.2 establishes a sharper iteration-complexity bound for a variant of the above method, which consists of applying the quadratic penalty method to a perturbed problem obtained by adding a small quadratic term to the objective function of (1). Finally, Section 5 compares the best iteration-complexity obtained in this paper with that derived in Monteiro and Svaiter [3, 4] and shows that the first one is generally better than the latter one.

1.1 Notation and terminology

notation

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 an inner product space $\langle \cdot, \cdot \rangle$.

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 a given norm $\|\cdot\|$ 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 an 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

intr_pen

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

$$\|\nabla f(\tilde{x}) - \nabla f(x)\| \le L_f \|\tilde{x} - x\|, \quad \forall x, \tilde{x} \in X. \tag{8}$$ lips_f

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.:

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

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

$$d(\lambda) := \inf\{f(x) + \langle \lambda, \mathcal{A}(x)\rangle : x \in X\}, \quad \forall \lambda \in -\mathcal{K}, \tag{9}$$ lagr

$$v(u) := \inf\{f(x) : \mathcal{A}(x) + u \in \mathcal{K}^*, x \in X\}, \quad \forall u \in \Re^m. \tag{10}$$ value

We make the following assumptions throughout this paper:

assump1 **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 $f^* = d(\lambda^*)$.

We define a near optimal solution of (1) 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

$$\begin{aligned} &\mathcal{A}(\tilde{x}) \in \mathcal{K}^*, \quad \langle \tilde{\lambda}, \mathcal{A}(\tilde{x})\rangle = 0, \\ &\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}), \end{aligned} \tag{11}$$ cp_opt

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

Definition 1 *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*

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p, \quad \langle \tilde{\lambda}, \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}))\rangle = 0, \tag{12}$$ pd_sol1

$$\nabla f(\tilde{x}) + (\mathcal{A}_0)^* \tilde{\lambda} \in -\mathcal{N}_X(\tilde{x}) + \mathcal{B}(\epsilon_d), \tag{13}$$ pd_sol2

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

In the remaining part of this section, we describe some properties for the distance function $d_{\mathcal{K}^*}(\cdot)$ used in (12). The proof of the following result is given in the Appendix.

max-rep **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 $C := (-\mathcal{K}) \cap B(0,1)$ and $B(0,1) := \{u \in \Re^m : \|u\| \leq 1\}$;

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 used in our analysis.

cor-dist **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 (10) with $u = -\mathcal{A}(x)$, the definition of $v(\cdot)$ in (10) implies that $v(u) \leq f(x)$. Hence,

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

■

3 Technical results

This section discusses some technical results that will be used in our analysis and reviews a first-order algorithm due to Nesterov [6, 8] for solving CP problems with simple feasible sets, and a restarting version of it for solving CP problems with strongly convex objective functions. It consists of two subsections. The first one develops several technical results involving projected gradients. The second subsection reviews Nesterov's optimal method and its restarting variant for solving CP problems with strongly convex objective functions.

3.1 Projected gradient and the optimality conditions

2.1

We consider the CP problem

$$\phi^* := \min_{x \in X} \phi(x), \tag{14}$$ optphi

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 (14) 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 2 *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, [7])*

$$\nabla\phi(\tilde{x})]_X^\tau := \frac{1}{\tau}\left[\tilde{x} - \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))\right], \tag{15}$$ def_prj

where $\Pi_X(\cdot)$ is the projection map onto X defined in terms of $\|\cdot\|$.

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

opt_cond **Proposition 3** *Let $\tilde{x} \in X$ be given and define $\tilde{x}^+ := \Pi_X(\tilde{x} - \tau\nabla\phi(\tilde{x}))$. Then, for any given $\epsilon \geq 0$ and $\tau > 0$, 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 (15) 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.

grad_prop **Lemma 4** *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 $x \in X$ and $\tau > 0$, there holds

$$\langle g(\tilde{x}) - g_X(\tilde{x}), x - \tilde{x}^+\rangle \geq 0; \tag{16}$$ prj_rel

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

$$\langle g(\tilde{x}) - g_X(\tilde{x}), g_X(\tilde{x})\rangle \geq 0; \tag{17}$$ prj_g_size

c) For any $x \in X$ and $\tau > 0$, there holds

$$\phi(\tilde{x}^+) - \phi(x) \leq (1 + \tau L_\phi)\,\|g_X(\tilde{x})\|\,\|\tilde{x}^+ - x\|. \tag{18}$$ prj_rel1

d) If $\tau = 1/L_\phi$ in definition (15), then

$$\phi(x) - \phi(x^*) \geq \frac{1}{2L_\phi}\|g_X(x)\|^2,\ \forall\, x \in X, \tag{19}$$ bnd_prj_grad

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

Proof. a) This statement is proved in p. 87 of [7].

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 $\phi(\cdot)$, (16), the assumption that $\phi(\cdot)$ has L_ϕ-Lipschitz-continuous gradient, and definition (15) 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 (17). ∎

3.2 Nesterov's Optimal Method

3.2

In this subsection, we discuss Nesterov's smooth first-order method for solving a class of smooth CP problems. This method or its restarting version described in this subsection will be used by the quadratic penalty method to solve the penalization subproblems. We observe, however, that any variant of Nesterov's method with the same optimal complexity, see, for example, [6, 7, 8, 1, 2, 10], could also be used to solve the penalization subproblems without changing any of the main results of the paper.

Our problem of interest is still the CP problem (14), which is assumed to satisfy the same assumptions mentioned in Subsection 3.1. Moreover, we assume throughout our discussion that the optimal value ϕ^* of problem (14) 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.,

$$h(x) \geq h(\tilde{x}) + \langle \nabla h(\tilde{x}), x - \tilde{x} \rangle + \frac{\sigma_h}{2}\|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{20}$$ strong_h

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

$$d_h(x; \tilde{x}) \equiv h(x) - l_h(x; \tilde{x}), \quad \forall x, \tilde{x} \in X, \tag{21}$$ Bregdist

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 (14). 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

$$x_{k+1}^{sd} \equiv \operatorname{argmin} \left\{ l_\phi(x; x_k) + \frac{L_\phi}{2} \|x - x_k\|^2 : x \in X \right\}, \tag{22}$$ Ne1

$$x_{k+1}^{ag} \equiv \operatorname{argmin} \left\{ \frac{L_\phi}{\sigma_h} d_h(x; x_0) + \sum_{i=0}^{k} \frac{i+1}{2} [l_\phi(x; x_i)] : x \in X \right\}. \tag{23}$$ Ne2

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

end

The main convergence result established by Nesterov [8] regarding the above algorithm is summarized in the following theorem.

ptimalmethod **Theorem 5** *The sequence $\{x_k^{sd}\}$ generated by Nesterov's optimal method satisfies*

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

where x^* is an optimal solution of (14). 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

$$\left\lceil 2\sqrt{\frac{d_h(x^*; x_0^{sd})L_\phi}{\sigma_h \epsilon}} \right\rceil \tag{24}$$ op_eq_1

iterations.

The following result is as an immediate special case of Theorem 5.

cor_opt **Corollary 6** *Suppose that $h : X \to \Re$ 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*

$$\left\lceil \|x_0^{sd} - x^*\| \sqrt{\frac{2L_\phi}{\epsilon}} \right\rceil \tag{25}$$ op_eq_3

iterations, where x^* is an optimal solution of (14).

Proof. If $h(x) = \|x\|^2/2$, then (21) implies that $d_h(x^*; x_0^{sd}) = \|x_0^{sd} - x^*\|^2/2$. The corollary now follows from this bound and relation (24). ∎

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

$$\langle \nabla\phi(x) - \nabla\phi(\tilde{x}), x - \tilde{x}\rangle \geq \mu\|x - \tilde{x}\|^2, \quad \forall x, \tilde{x} \in X. \tag{26}$$ strongcx

Nesterov shows in Theorem 2.2.2 of [7] that, under the assumptions of Corollary 6, a variant of his optimal method finds a solution $x_k \in X$ satisfying $\phi(x_k) - \phi^* \leq \epsilon$ in no more than

$$\left\lceil \sqrt{\frac{L_\phi}{\mu}} \log \frac{L_\phi \|x_0^{sd} - x^*\|^2}{\epsilon} \right\rceil \tag{27}$$ op_eq_4

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 (27) 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.

timalmethod1 **Theorem 7** *Let $\epsilon > 0$ be given and suppose 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*

$$K := \left\lceil \sqrt{\frac{8L_\phi}{\mu}} \right\rceil \tag{28}$$ defK

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

$$\left\lceil \sqrt{\frac{8\,L_\phi}{\mu}} \right\rceil \max\{1, \lceil \log Q \rceil\} \tag{29}$$ mult_stage

iterations, where

$$Q := \frac{\mu \|x_0^{sd} - x^*\|^2}{2\epsilon} \tag{30}$$ defQ

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 6, implies that the number of iterations is bounded by $\lceil 2\sqrt{L_\phi/\mu} \rceil$, and hence bounded by (29).

We now show that bound (29) 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 5 with $h(\cdot) = \|\cdot\|^2/2$ and inequality (26), 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 (29). ∎

It is interesting to compare the complexity bounds of Corollary 6 and Theorem 7. 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 (30), implies that the bound (25) is always greater than or equal to the bound (29). Moreover, as Q goes to infinity, the ratio between (25) and (29) converges to infinity.

4 The quadratic penalty method

sec4

The goal of this section is to establish iteration-complexity bounds for quadratic penalty based methods for solving (1), expressed in terms of number of Nesterov's optimal method iterations performed for approximately solving quadratic penalty subproblems. It consists of two subsections. Subsection 4.1 derives an iteration-complexity bound for the quadratic penalty method applied directly to (1). Subsection 4.2 establishes a sharper iteration-complexity bound for a variant of the above method, which consists of applying the quadratic penalty method to a perturbed problem obtained by adding a small quadratic term to the objective function of (1).

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

$$\Psi_\rho^* := \inf_{x \in X} \left\{ \Psi_\rho(x) := f(x) + \frac{\rho}{2} \left[d_{\mathcal{K}^*}(\mathcal{A}(x)) \right]^2 \right\}. \tag{31}$$ cp_p

We will now see that the objective function Ψ_ρ of (31) 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 [2] for its proof).

lip_dist **Proposition 8** *Given a closed convex set $\mathcal{C} \subseteq \Re^m$, consider the distance function $d_{\mathcal{C}} : \Re^m \to \mathbf{R}$ to $\mathcal{C}$ with respect to $\|\cdot\|$ 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 8, we obtain the following result.

cor-lips **Corollary 9** *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 8. 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 (8) and Proposition 8, 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 $\|\mathcal{A}\| = \|\mathcal{A}_0\| = \|\mathcal{A}_0^*\|$. ∎

In view of Corollary 9, which establishes that Ψ_ρ has Lipschitz continuous gradient, it follows that iteration-complexity bounds for approximately solving (31) via Nesterov's optimal method (or one of its variants described for example in [2, 6, 8]) or its restarting variant (see Subsection 3.2) can now be easily obtained by means of Corollary 6 or Theorem 7.

4.1 Quadratic penalty method applied to the original problem

q_no_purt

In this subsection, we consider the quadratic penalty method applied directly to the original problem (1). It consists of approximately solving penalized subproblems of the form (31) for an increasing sequence of penalty parameters ρ, until eventually an (ϵ_p, ϵ_d)-primal-dual solution of (1) is obtained.

Given an approximate solution $\tilde{x} \in X$ for the penalized problem (31), 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 (11).

opt_pd **Proposition 10** *If $\tilde{x} \in X$ is a δ-approximate solution of (31), i.e., it satisfies*

$$\Psi_\rho(\tilde{x}) - \Psi_\rho^* \leq \delta, \tag{32}$$ app_sol

then the pair $(\tilde{x}^+, \lambda)$ *defined as*

$$\tilde{x}^+ := \Pi_X(\tilde{x} - \nabla\Psi_\rho(\tilde{x})/M_\rho), \quad (33)\ \texttt{xplus}$$

$$\lambda := \rho[\mathcal{A}(\tilde{x}^+) - \Pi_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+))] \quad (34)\ \texttt{lambdaplus}$$

is in $X \times (-\mathcal{K})$ *and satisfies the second relation in (12) and the relations*

$$d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x}^+)) \leq \frac{1}{\rho}\|\lambda^*\| + \sqrt{\frac{\delta}{\rho}}, \quad (35)\ \texttt{p\_res\_1}$$

$$\nabla f(\tilde{x}^+) + (\mathcal{A}_0)^*\lambda \in -\mathcal{N}_X(\tilde{x}^+) + \mathcal{B}(2\sqrt{2M_\rho\delta}), \quad (36)\ \texttt{d\_res}$$

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

Proof. It follows from Lemma 4(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 (32). Using the fact that $v(0) = f^* \geq \Psi_\rho^*$, Corollary 2 and assumption (32), 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 (35). Moreover, by Lemma 4(d) with $\phi = \Psi_\rho$ and $L_\phi = M_\rho$ and assumption (32), 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 (36) now follows from this inequality, Proposition 3(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 (34). Finally, $\lambda \in (-\mathcal{K})$ and the second relation of (12) holds in view of the definition of λ and well-known properties of the projection operator onto a cone. ∎

With the aid of Proposition 10, we can now derive an iteration-complexity bound for Nesterov's method applied to the quadratic penalty problem (31) to compute an (ϵ_p, ϵ_d)-primal-dual solution of (1).

pd_theorem **Theorem 11** *Let* λ^* *be an arbitrary Lagrange multiplier for (1) and let* $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ *be given. If*

$$\rho = \rho_{pd}(t) := \frac{1}{\epsilon_p}\left(t + \frac{\epsilon_d}{\sqrt{8}\|\mathcal{A}\|}\right) \quad (37)\ \texttt{opt\_pen\_pd}$$

for some $t \geq \|\lambda^*\|$, *then Nesterov's optimal method with* $h(\cdot) = \|\cdot\|^2/2$ *applied to problem (31) finds an* (ϵ_p, ϵ_d)*-primal-dual solution of (1) in at most*

$$N_{pd}(t) := \left\lceil 4D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|^2 t}{\epsilon_p\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{8}\epsilon_p}\right)\right\rceil \quad (38)\ \texttt{bound\_pd}$$

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

Proof. Let $\tilde{x} \in X$ be a δ-approximate solution of (31) 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 10 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 (37) 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 6, Nesterov's optimal method finds an approximate solution $\tilde{x}$ as above in at most

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

iterations. Substituting the value of ρ given by (37) in the above bound, we obtain bound (38). ∎

We now make a few observations regarding Theorem 11. First, the choice of ρ given by (37) requires that $t \geq \|\lambda^*\|$ so as to guarantee that an δ-approximate solution $\tilde{x}$ of (31) satisfies $d_{\mathcal{K}^*}(\mathcal{A}(\tilde{x})) \leq \epsilon_p$, where $\delta = \epsilon_d^2/(8M_\rho)$. Second, note that the iteration-complexity $N_{pd}(t)$ obtained in Theorem 11 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_d)-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

$$\beta_0 = 4D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{8}\epsilon_p}\right), \quad \beta_1 = \frac{4D_X\|\mathcal{A}\|^2}{\epsilon_p\epsilon_d}, \quad t_0 := \frac{\max(1,\beta_0)}{\beta_1}. \tag{39}$$ dd

2) Set $\rho = \rho_{pd}(t_k)$, and perform at most $\lceil N_{pd}(t_k)\rceil$ iterations of Nesterov's optimal method applied to problem (31). If an iterate $\tilde{x}$ is obtained such that (32) with $\delta = \epsilon_d^2/(8M_\rho)$ 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 13. Lemma 12 states a simple inequality used in the proof of Lemma 13.

simple_c **Lemma 12** *For any $\tau, \alpha \geq 0$ and $x \in \Re$, we have $\tau x + \alpha \leq (\tau + \alpha)\max\{1, \lceil x\rceil\}$.*

count **Lemma 13** *Let scalar $p > 0$ be given. Then, there exists a constant $C = C(p)$ such that for any $\beta_0 \in \Re$ and $\beta_1, \bar{t} > 0$, we have*

$$\sum_{k=0}^{K}\lceil\beta_0 + \beta_1 t_k^p\rceil \leq C\lceil\beta_0 + \beta_1\bar{t}^p\rceil, \tag{40}$$ deff1

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

$$t_0 := \left(\frac{\max(\beta_0, 1)}{\beta_1}\right)^{1/p}, \quad K := \max\left\{0, \left\lceil \log\left(\frac{\bar{t}}{t_0}\right)\right\rceil\right\}. \tag{41}$$ deff

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

$$\begin{aligned}\sum_{k=0}^{K} \lceil \beta_0 + \beta_1 t_k^p \rceil &= \lceil \beta_0 + \beta_1 t_0^p \rceil = \lceil \beta_0 + \max(\beta_0, 1) \rceil \leq 2\beta_0 + 2 \\ &\leq 4 \max\{\lceil \beta_0 \rceil, 1\} \leq 4 \lceil \beta_0 + \beta_1 \bar{t}^p \rceil,\end{aligned}$$

where the second equality follows from the definition of t_0 and the second inequality follows from Lemma 12. 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

$$\begin{aligned}\sum_{k=0}^{K} \lceil \beta_0 + \beta_1 t_k^p \rceil &\leq \sum_{k=0}^{K} 1 + \beta_0 + \beta_1 t_0^p 2^{pk} \leq (1 + \beta_0)(1 + K) + \beta_1 t_0^p \frac{2^{(K+1)p}}{2^p - 1} \\ &\leq (1 + \beta_0)\left[2 + \log\left(\frac{\bar{t}}{t_0}\right)\right] + \beta_1 \frac{(4\bar{t})^p}{2^p - 1} \\ &\leq (1 + \beta_0)\left[2 + \frac{1}{p}\left(\frac{\bar{t}}{t_0}\right)^p\right] + \frac{4^p}{2^p - 1}\beta_1 \bar{t}^p \\ &\leq (1 + \beta_0)\left[2 + \frac{1}{p}\left(\frac{\beta_1 \bar{t}^p}{\max(\beta_0, 1)}\right)\right] + \frac{4^p}{2^p - 1}\beta_1 \bar{t}^p \\ &\leq 2(1 + \beta_0) + \frac{2}{p}\beta_1 \bar{t}^p + \frac{4^p}{2^p - 1}\beta_1 \bar{t}^p \\ &\leq 2 + \max\left\{2, \frac{2}{p} + \frac{4^p}{2^p - 1}\right\}(\beta_0 + \beta_1 \bar{t}^p),\end{aligned}$$

which, in view of Lemma 12, 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_d)-primal-dual solution of (1).

cor-pd **Corollary 14** *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_d)*-primal-dual solution of (1) is bounded by* $\mathcal{O}(N_{pd}(\|\lambda^*\|))$, *where* $N_{pd}(\cdot)$ *is defined in (38).*

Proof. In view of Theorem 11, the iteration count k in Search Procedure 1 cannot 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 $\sum_{k=0}^{K} N_{pd}(t_k) = \sum_{k=0}^{K} \lceil \beta_0 + \beta_1 t_k \rceil$, 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 $p = 1$ and $\bar{t} = \|\lambda^*\|$. ■

q_purt

4.2 Quadratic penalty method applied to a perturbed problem

Consider the perturbation problem

$$f_\gamma^* := \min\{f_\gamma(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 : \mathcal{A}(x) \in \mathcal{K}^*, x \in X\}, \tag{42}$$ pe_cp

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 (42) 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 a primal-dual approximate solution of (1) by applying the quadratic penalty approach to the perturbation problem (42) for a conveniently chosen perturbation parameter $\gamma > 0$.

Note that the quadratic penalty problem associated with (42) is given by

$$\Psi_{\rho,\gamma}^* := \min_{x \in X} \left\{ \Psi_{\rho,\gamma}(x) := f(x) + \frac{\gamma}{2}\|x - x_0\|^2 + \frac{\rho}{2} d_{\mathcal{K}^*}(\mathcal{A}(x))^2 \right\}. \tag{43}$$ cp_pert

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

$$M_{\rho,\gamma} := L_f + \rho\|\mathcal{A}\|^2 + \gamma. \tag{44}$$ Mrhogamma

The following simple lemma relates the optimal values of the perturbation problem (42), the penalty problem (43) and the original problem (1).

Lemma 15 *Let f^*, Ψ_ρ^*, f_γ^*, and $\Psi_{\rho,\gamma}^*$ be the optimal values defined in (1), (31), (42), and (43), respectively. Then,*

$$0 \le f_\gamma^* - f^* \le \gamma D_X^2/2 \tag{45}$$ close

$$0 \le \Psi_{\rho,\gamma}^* - \Psi_\rho^* \le \gamma D_X^2/2, \tag{46}$$ close1

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

Proof. The first inequalities in both relations (45) and (46) 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 (42), 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 (45) immediately follows. The second inequality in (46) can be similarly shown. ∎

Theorem 16 below describes the iteration-complexity of computing a primal-dual approximate solution of (1) by applying the quadratic penalty method to the perturbation problem (42). It shows that the resulting approach has a substantially better iteration-complexity than the one discussed in Subsection 4.1, which consists of applying the quadratic penalty method directly to the original problem (1).

_theorem_per **Theorem 16** *Let λ_γ^* be an arbitrary Lagrange multiplier for (42) and let $(\epsilon_p, \epsilon_d) \in \mathbf{R}_{++} \times \mathbf{R}_{++}$ be given. Also assume that*

$$\rho = \tilde{\rho}_{pd}(t) := \frac{1}{\epsilon_p}\left(t + \frac{\epsilon_d}{\sqrt{32}\|\mathcal{A}\|}\right), \ \gamma = \frac{\epsilon_d}{2D_X}, \tag{47}$$ pert_para_pd

for some $t \geq \|\lambda_\gamma^*\|$. Then, the variant of Nesterov's optimal method of Theorem 7 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ applied to problem (43) finds an (ϵ_p, ϵ_d)-primal-dual solution of (1) in at most

$$\tilde{N}_{pd}(t) := \lceil 8\mathcal{S}(t) \rceil \ \lceil 2\log(2\mathcal{S}(t)) \rceil, \tag{48}$$ bound_pert_pd

where D_X is defined in Theorem 11 and

$$\mathcal{S}(t) := \left[2D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{32}\epsilon_p}\right) + 1\right]^{\frac{1}{2}} + \frac{\sqrt{2}D_X^{1/2}\|\mathcal{A}\|}{\sqrt{\epsilon_p\epsilon_d}}\, t^{1/2}. \tag{49}$$ def_S

Proof. Define $\delta := \epsilon_d^2/(32M_{\rho,\gamma})$ and let $\tilde{x} \in X$ be a δ-approximate solution for (43), i.e., $\Psi_{\rho,\gamma}(\tilde{x}) - \Psi_{\rho,\gamma}^* \leq \delta$. It then follows from Proposition 10 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 10 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 (47) 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 7 with $\phi = \Psi_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \delta$, and noting that the definition of δ and the definition of γ in (47) 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

$$\left\lceil 8\sqrt{\frac{M_{\rho,\gamma}}{\gamma}} \right\rceil \lceil \log Q \rceil = \left\lceil 8\sqrt{\frac{M_{\rho,\gamma}}{\gamma}} \right\rceil \left\lceil \log\left(\frac{4M_{\rho,\gamma}}{\gamma}\right) \right\rceil. \tag{50}$$ rell

Now, using (44) and the definitions of ρ and γ in (47), we easily see that the latter bound is majorized by (48). ∎

Note that the complexity bound (48) derived in Theorem 16 is guaranteed only under the assumption that $t \geq \|\lambda_\gamma^*\|$, where λ_γ^* is the minimum norm Lagrange multiplier for (42). Since the

bound (48) 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 (48).

cor-pert-pd **Corollary 17** *Let λ_γ^* be the minimum norm Lagrange multiplier for (42). Consider the "guess and check" procedure similar to Search Procedure 1 where the functions N_{pd} and ρ_{pd} are replaced by the functions $\tilde{N}_{pd}$ and $\tilde{\rho}_{pd}$ defined in Theorem 16, Nesterov's optimal method is replaced by its variant of Theorem 7 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*

$$\beta_0 = 8\left[2D_X\left(\frac{L_f}{\epsilon_d} + \frac{\|\mathcal{A}\|}{\sqrt{32}\epsilon_p}\right) + 1\right]^{1/2}, \quad \beta_1 = \frac{8\sqrt{2}D_X^{1/2}\|\mathcal{A}\|}{(\epsilon_p\epsilon_d)^{1/2}}. \tag{51}$$ betas_per_pd

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 (48).

Proof. In view of Theorem 16, 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) cannot 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

$$\sum_{k=0}^{K}\tilde{N}_{pd}(t_k) = \sum_{k=0}^{K}\lceil 8\mathcal{S}(t_k)\rceil\ \lceil 2\log(2\mathcal{S}(t_k))\rceil \le \lceil 2\log(2\mathcal{S}(t_K))\rceil\sum_{k=0}^{K}\lceil 8\mathcal{S}(t_k)\rceil\,. \tag{52}$$ oo1

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 (51). Moreover, using the fact that $t_K \le 2\|\lambda_\gamma^*\|$, 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 (52), 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 11 and 16, respectively. It follows from (38), (48), and (49) that

$$\frac{\tilde{N}_{pd}(t)}{N_{pd}(t)} \le \frac{(\lceil 8\mathcal{S}(t)\rceil\ \lceil 2\log(2\mathcal{S}(t))\rceil)}{\left\lceil(\mathcal{S}(t))^2 - 1\right\rceil} \le \mathcal{O}(1)\frac{\log\mathcal{S}(t)}{\mathcal{S}(t) - 1}, \tag{53}$$ comp1

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 cannot use the previous observation to compare the iteration-complexity of Corollary 14 with that obtained in Corollary 17 since the first one is expressed in terms of $\|\lambda^*\|$ and the latter in terms of $\|\lambda_\gamma^*\|$. However, if $\|\lambda_\gamma^*\| = \mathcal{O}(\|\lambda^*\|)$, then it can be easily seen that (53) implies that

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

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

Observe that the iteration-complexity bound given in Corollary 17 is expressed in terms of the minimum norm Lagrange multiplier for the perturbed problem (42). A natural question is whether an alternative bound can be obtained in terms of the minimum norm Lagrange multiplier for the original problem (1). Indeed, Theorem 19 and Corollary 20 derive these alternative bounds.

Before presenting these results, we first state the following simple result.

simple **Lemma 18** *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*

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

theorem_per1 **Theorem 19** *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*

$$\rho = \hat{\rho}_{pd}(t) := \left(\frac{\sqrt{\epsilon_d D_X/2} + \sqrt{(\epsilon_d D_X/2) + 4\alpha(t)\epsilon_p}}{2\epsilon_p} \right)^2, \quad \gamma = \frac{\epsilon_d}{2D_X}, \tag{54}$$ pert_para_pd1

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 7 with $\mu = \gamma$ and $L_\phi = M_{\rho,\gamma}$ applied to problem (43) finds an (ϵ_p, ϵ_d)-primal-dual solution of (1) in at most

$$\hat{N}_{pd}(t) := \left\lceil 8\hat{\mathcal{S}}(t) \right\rceil \left\lceil 2\log(2\hat{\mathcal{S}}(t)) \right\rceil, \tag{55}$$ bound_pert_pd

where D_X is defined in Theorem 11 and

$$\hat{\mathcal{S}}(t) := \sqrt{\frac{2L_f D_X}{\epsilon_d}} + \|\mathcal{A}\| \left(\frac{D_X}{\epsilon_p} + \sqrt{\frac{2tD_X}{\epsilon_p\epsilon_d}} \right) + \frac{\sqrt{\|\mathcal{A}\| D_X}}{8^{1/4}\sqrt{\epsilon_p}} + 1. \tag{56}$$ def_Sh

Proof. Define $\delta := \epsilon_d^2/(32M_{\rho,\gamma})$ and let $\tilde{x} \in X$ be a δ-approximate solution for (43), i.e., $\Psi_{\rho,\gamma}(\tilde{x}) - \Psi_{\rho,\gamma}^* \leq \delta$. As shown in the proof of Theorem 16, 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 4(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 (31), (43), (46) and (54) 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 10, the assumption that $t \geq \|\lambda^*\|$, relation (54) and Lemma 18 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 7 with $\phi = \Psi_{\rho,\gamma}$, $\mu = \gamma$ and $\epsilon = \delta$, and noting that the definition of δ and the definition of γ in (54) 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 (50). Since relations (44) and (54) 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

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

it follows that (50) can be bounded by (55). ∎

The following result states the iteration-complexity of a "guess and check" procedure based on Theorem 19. Its proof is based on similar arguments as those used in the proof of Corollary 17.

cor1-pert-pd **Corollary 20** *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_{pd} and ρ_{pd} are replaced by the functions $\hat{N}_{pd}$ and $\hat{\rho}_{pd}$ defined in Theorem 16, Nesterov's optimal method is replaced by its variant of Theorem 7 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 (55).

It is interesting to compare the functions $N_{pd}(t)$ and $\hat{N}_{pd}(t)$ defined in Theorems 11 and 19, respectively. When the second term in the right hand side of (56) is dominated by the other terms, i.e.,

$$\frac{\|\mathcal{A}\|D_X}{\epsilon_p} = \mathcal{O}\left(\frac{\|\mathcal{A}\|t^{1/2}D_X^{1/2}}{\sqrt{\epsilon_p\epsilon_d}} + \sqrt{\frac{L_f D_X}{\epsilon_d}}\right), \tag{57}$$ noo

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 (57) 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 Concluding remarks

In this section, we compare the results obtained in this paper for the quadratic penalty method with another possible approach for solving variational inequalities (VI) studied in Nemirovski ([5]) for bounded sets, and Monteiro and Svaiter ([4]) for unbounded sets. For the sake of simplicity, we assume that $\mathcal{K} = \Re^m$ and hence that $\mathcal{K}^* = \{0\}$.

Given a closed convex set $\Omega \in \Re^p$ and a monotone continuous function $F : \Omega \to \Re^p$. The (monotone) VI problem with respect to the pair (F, Ω), denoted by $VI(F, \Omega)$, consists of finding w^* such that (6) holds. It is well-known that, under the assumption that F is monotone and continuous, (6) is equivalent to

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

Relaxing (6) and the above condition, we obtain the following two notions of approximate solutions of $VI(F, \Omega)$.

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

It is well-known that the CP problem (1) is equivalent to solving the $VI(F, \Omega)$, where $\Omega := X \times \Re^m$ and F given by (7). Moreover, defining the norm on $\Re^n \times \Re^m$ as $\|w\| := (\|x\|^2 + \|\lambda\|^2)^{1/2}$, then it is easy to see that an (ϵ_p, ϵ_d)-primal-dual solution $(\bar{x}, \bar{\lambda})$ is a $(\varrho, 0)$-strong solution, where $\varrho = \max\{\epsilon_p, \epsilon_d\}$. Disregarding the constants L_f, $\|\mathcal{A}\|$, D_X and $\|\lambda^*\|$, it has been shown in Monteiro and Svaiter ([4]) that, given $(\varrho, \epsilon) \in \Re_{++} \times \Re_{++}$, a variant of Korpelevich's method can find a (ϱ, ϵ)-strong solution for $VI(F, \Omega)$ in $\mathcal{O}(\varrho^{-2} + \epsilon^{-1})$ iterations. Also, when ϵ is chosen as $\epsilon = \mathcal{O}(\varrho^2)$, it can be shown by means of Proposition C.6 of Monteiro and Svaiter [3] that such solution can be cheaply purified to a $(\varrho, 0)$-strong solution of $VI(F, \Omega)$. Hence, we conclude that this variant of Korpelevich's method can find a $(\varrho, 0)$-strong solution for $VI(F, \Omega)$ in $\mathcal{O}(\varrho^{-2})$ iterations. On the other hand, we show in this paper that a $(\varrho, 0)$-strong solution can be found in

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

iterations by applying the guess-and-check procedure of Subsection 4.2 with $\epsilon_p = \epsilon_d = \varrho/\sqrt{2}$. Hence, the best complexity obtained in this paper for the quadratic penalty method is better than the one obtained in [4] for a variant of Korpelevich's method by at least a factor of $\varrho(\log \varrho^{-1})$.

It should be noted that [4] also shows that an (ϱ, ϵ)-weak solution for $VIP(F, \Omega)$ can be found in

$$\mathcal{O}(\varrho^{-1} + \epsilon^{-1}). \tag{58}$$ hope

It would be interesting to see whether our analysis in this paper can be modified to the context of finding a weak solution of $VI(F, \Omega)$ so as to obtain a better iteration-complexity bound than (58).

In this paper, we have studied the iteration-complexities of the quadratic penalty methods (see Theorems 11 and 16) under the assumption that the convex set X in (1) is bounded. It would be interesting to generalize these results to the situation where X is unbounded. We can immediately point out two difficulties if one is to pursue this task. First, the iteration-complexity bounds would have to be expressed in terms of a quantity related to the distances of x_0 to the optimal sets of subproblems (31) or (43) corresponding to different values of ρ since the quadratic penalty approach studied in this paper consists of applying Nesterov's method to these subproblems and the iteration-complexity of the latter method depends on these distances (see Theorems 5 and 7). Clearly, all these distances are simply majorized by D_X when X is bounded. Second, one would need to develop a suitable way to terminate Nesterov's method applied to the above subproblems based on easily computable stopping criteria for the case when X is unbounded. We observe that the two well-known ways of terminating Nesterov's method proposed in the literature only work when X is bounded or ϕ^* is known. Indeed, the first way consists of checking the optimality gap $\phi(x_k^{sd}) - \phi^*$. This can be accomplished when ϕ^* is known or by generating a sequence of lower bounds for ϕ^* (see [8]). The second way to terminate Nesterov's method is to perform a pre-specified number of iterations estimated by means of the complexity bounds (24) and (29). The latter way was exactly the one we have used in the guess and check procedures developed in this paper. However, none of them seem to be suitable when X is not assumed to be bounded.

Appendix

In this section, we prove Proposition 1.

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

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

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, $\langle u, \lambda \rangle \geq -\|\lambda\| \, d_{\mathcal{K}}(u)$. ∎

References

AuTe06-1 [1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.

LaLuMo11-1 [2] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126:1–29, 2011.

MonSva10-1 [3] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemi-variational inequalities with applications to saddle-point and convex optimization problems. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA, June 2010.

MonSva10-3 [4] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal projection method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20:2755–2787, 2010.

Nem05-1 [5] A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2005.

Nest83-1 [6] 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.

Nest04 [7] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

Nest05-1 [8] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

Nest07-2 [9] Y. E. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109:319–344, 2007.

tseng08-1 [10] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.