Mathematical Programming manuscript No.
(will be inserted by the editor)

Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming

Guanghui Lan · Zhaosong Lu · Renato D. C. Monteiro

Received: date / Accepted: date

Abstract In this paper we consider the general cone programming problem, and propose primal-dual convex (smooth and/or nonsmooth) minimization reformulations for it. We then discuss first-order methods suitable for solving these reformulations, namely, Nesterov's optimal method [10,11], Nesterov's smooth approximation scheme [11], and Nemirovski's prox-method [9], and propose a variant of Nesterov's optimal method which has outperformed the latter one in our computational experiments. We also derive iteration-complexity bounds for these first-order methods applied to the proposed primal-dual reformulations of the cone programming problem. The performance of these methods is then compared using a set of randomly generated linear programming and semidefinite programming (SDP) instances. We also compare the approach based on the variant of Nesterov's optimal method with the low-rank method proposed by Burer and Monteiro [2,3] for solving a set of randomly generated SDP instances.

Keywords Cone programming · primal-dual first-order methods · smooth optimal method · non-smooth method · prox-method · linear programming · semidefinite programming

Mathematics Subject Classification (2000) 65K05 · 65K10 · 90C05 · 90C22 · 90C25

The work of the first and third authors was partially supported by NSF Grants CCF-0430644 and CCF-0808863 and ONR Grants N00014-05-1-0183 and N00014-08-1-0033. The second author was supported in part by SFU President's Research Grant and NSERC Discovery Grant.

Guanghui Lan
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA
emailglan@isye.gatech.edu

Zhaosong Lu
Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
emailzhaosong@sfu.ca

Renato D. C. Monteiro
School of ISyE, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
E-mail: monteiro@isye.gatech.edu

1 Introduction

In [10,11], Nesterov proposed an optimal algorithm for solving convex programming (CP) problems of the form

$$\bar{f} \equiv \inf\{f(u) : u \in \mathcal{U}\}, \tag{1}$$

where f is a convex function with Lipschitz continuous derivative and $\mathcal{U}$ is a sufficiently simple closed convex set. It is shown that his method has $\mathcal{O}(\epsilon^{-1/2})$ iteration-complexity bound, where $\epsilon > 0$ is the absolute precision of the final objective function value. A proximal-point-type algorithm for (1) having the same complexity above has also been proposed more recently by Auslender and Teboulle [1].

For general minimization problems of the above form, where f is Lipschitz continuous, the classical subgradient method is known to be optimal with iteration-complexity bounded by $\mathcal{O}(1/\epsilon^2)$. In a more recent and very relevant work, Nesterov [11] presents a first-order method to solve CP problems of the form (1) for an important and broad class of non-smooth convex objective functions with iteration-complexity bounded by $\mathcal{O}(1/\epsilon)$. Nesterov's approach consists of approximating an arbitrary function f from the class by a sufficiently close smooth one with Lipschitz continuous derivative and applying the optimal smooth method in [10,11] to the resulting CP problem with f replaced by its smooth approximation. In a subsequent paper, Nemirovski [9] proposed a proximal-point-type first-order method for solving a slightly more general class of CP problems than the one considered by Nesterov [11] and also established an $\mathcal{O}(1/\epsilon)$ iteration-complexity bound for his method.

These first-order methods due to Nesterov [10,11] and Nemirovski [9] have recently been applied to certain semidefinite programming (SDP) problems with some special structures (see Lu et al. [8], Nesterov [12] and d'Aspremont [4]). More recently, Hoda et al. [6] have used Nesterov's smooth method [10,11] to successfully solve a special class of large-scale linear programming problems. However, the authors are not aware of any paper which use the first-order methods presented in [10,11,9] to solve the general cone programming problem.

In this paper we consider the general cone programming problem, and propose primal-dual convex (smooth and/or nonsmooth) minimization reformulations for it of the form (1). We then discuss the three first-order methods mentioned above, namely, Nesterov's optimal method [10,11], Nesterov's smooth approximation scheme [11], and Nemirovski's prox-method [9], and propose a variant of Nesterov's optimal method which has outperformed the latter one in our computational experiments. We also establish the iteration-complexity bounds of these first-order methods for solving the proposed primal-dual reformulations of the cone programming problem. The performance of these methods is then compared using a set of randomly generated linear programming (LP) and semidefinite programming (SDP) instances. The main conclusion of our comparison is that the approach based on the variant of Nesterov's optimal method outperforms all the others. We also compare the approach based on the variant of Nesterov's optimal method with the low-rank method proposed by Burer and Monteiro [2,3] for solving a set of randomly generated SDP instances. The main conclusion of this last comparison is that while the approach based on the variant of Nesterov's optimal method is comparable to the low-rank method to obtain solutions with low accuracies, the latter one substantially outperforms the first one when the goal is to compute highly accurate solutions.

The paper is organized as follows. In Section 2, we propose primal-dual convex minimization reformulations of the cone programming problem. In Subsection 2.1, we discuss reformulations with smooth objective functions which are suitable for Nesterov's optimal method or its variant and, in Subsection 2.2, we discuss those with nonsmooth objective functions and Nemirovski's prox-method. In Section 3, we discuss Nesterov's optimal method and also propose a variant of his method. We also derive the iteration-complexity bound of both methods applied to a class of smooth reformulations of the cone programming problem. In Section 4, we discuss Nesterov's smooth approximation scheme (Subsection 4.1) and Nemirovski's prox-method (Subsection 4.2), and derive their corresponding iteration-complexity bound for solving a class of nonsmooth reformulations of the cone programming problem. In Section 5, the performance of the first-order methods discussed in this paper and the low-rank method is compared on a set of randomly generated LP and SDP instances. Finally, we present some concluding remarks in Section 6.

1.1 Notation

In this paper, all vector spaces are assumed to be finite dimensional. Let U be a normed vector space whose norm is denoted by $\|\cdot\|_U$. The dual space of U, denoted by U^*, is the normed vector space consisting of all linear functionals of $u^* : U \to \Re$, endowed with the dual norm $\|\cdot\|_U^*$ defined as

$$\|u^*\|_U^* = \max_u \{\langle u^*, u\rangle : \ \|u\|_U \le 1\}, \quad \forall u^* \in U^*, \tag{2}$$

where $\langle u^*, u\rangle := u^*(u)$ is the value of the linear functional u^* at u.

If V denotes another normed vector space with norm $\|\cdot\|_V$, and $\mathcal{E} : U \to V^*$ is a linear operator, the adjoint of $\mathcal{E}$ is the linear operator $\mathcal{E}^* : V \to U^*$ defined by

$$\langle \mathcal{E}^* v, u\rangle = \langle \mathcal{E}u, v\rangle, \quad \forall u \in U, \ v \in V.$$

Moreover, the operator norm of $\mathcal{E}$ is defined as

$$\|\mathcal{E}\|_{U,V} = \max_u \{\|\mathcal{E}u\|_V^* : \|u\|_U \le 1\}. \tag{3}$$

It can be easily seen that

$$\|\mathcal{E}\|_{U,V} = \|\mathcal{E}^*\|_{V,U} \tag{4}$$

and

$$\|\mathcal{E}u\|_V^* \le \|\mathcal{E}\|_{U,V}\,\|u\|_U \ \text{ and } \ \|\mathcal{E}^* v\|_U^* \le \|\mathcal{E}\|_{U,V}\,\|v\|_V, \ \ \forall u \in U, \ v \in V. \tag{5}$$

Given $u^* \in U^*$ and $v^* \in V^*$, let $(u^*, v^*) : U \times V \to \Re$ denote the linear functional defined by

$$(u^*, v^*)(u, v) := \langle u^*, u\rangle + \langle v^*, v\rangle, \ \ \forall u \in U, \ v \in V.$$

A function $f : \Omega \subseteq U \to \Re$ is said to have L-Lipschitz-continuous gradient with respect to $\|\cdot\|_U$ if it is differentiable and

$$\|f'(u) - f'(\tilde{u})\|_U^* \le L\|u - \tilde{u}\|_U \quad \forall u, \tilde{u} \in \Omega. \tag{6}$$

Given a closed convex set $\mathcal{C} \subseteq U$ and an arbitrary norm $\|\cdot\|$ on U, let $\mathrm{dist}_{\mathcal{C}} : U \to \Re$ denote the distance function to $\mathcal{C}$ measured in terms of $\|\cdot\|$, i.e.

$$\mathrm{dist}_{\mathcal{C}}(u) := \inf_{\tilde{u} \in \mathcal{C}} \|u - \tilde{u}\|, \quad \forall u \in U. \tag{7}$$

2 The reformulations for cone programming

In this section, we propose primal-dual convex minimization reformulations for cone programming. In particular, we present the convex smooth and nonsmooth minimization reformulations in Subsections 2.1 and 2.2, respectively.

Assume that X and Y are two normed vector spaces. Given a linear operator $\mathcal{A} : X \to Y^*$, vectors $c^* \in X^*$ and $b^* \in Y^*$ and a closed convex cone $\mathcal{L} \subseteq X$, consider the cone programming problem

$$\begin{array}{ll} \min\limits_{x} & \langle c^*, x\rangle \\ \text{s.t.} & \mathcal{A}x = b^*, \quad x \in \mathcal{L}, \end{array} \tag{8}$$

and its associated dual problem

$$\begin{array}{ll} \max\limits_{(y,s^*)} & \langle b^*, y\rangle \\ \text{s.t.} & \mathcal{A}^*y + s^* = c^*, \quad s^* \in \mathcal{L}^*, \end{array} \tag{9}$$

where $\mathcal{L}^* := \{s^* \in X^* : \langle s^*, x\rangle \geq 0, \forall x \in \mathcal{L}\}$ is the dual cone of $\mathcal{L}$. We make the following assumption throughout the paper.

Assumption 1 *The pair of cone programming problems (8) and (9) have optimal solutions and their associated duality gap is zero.*

In view of the above assumption, a primal-dual optimal solution of (8) and (9) can be found by solving the following constrained system of linear equations:

$$\begin{array}{rl} \mathcal{A}^*y + s^* - c^* &= 0 \\ \mathcal{A}x - b^* &= 0\,, \quad (x, y, s^*) \in \mathcal{L} \times Y \times \mathcal{L}^*. \\ \langle c^*, x\rangle - \langle b^*, y\rangle &= 0 \end{array} \tag{10}$$

Alternatively, eliminating the variable s^* and using the weak duality lemma, the above system is equivalent to system:

$$\begin{array}{rcl} -\mathcal{A}^*y + c^* & \in & \mathcal{L}^* \\ \mathcal{A}x - b^* & = & 0\ , \quad (x, y) \in \mathcal{L} \times Y. \\ \langle c^*, x\rangle - \langle b^*, y\rangle & \leq & 0 \end{array} \tag{11}$$

In fact, there are numerous ways in which one can characterize a primal-dual solution of (8) and (9). For example, let $\mathcal{B}_+$ and $\mathcal{B}_-$ be closed convex cones in Y^* such that $\mathcal{B}_+ \cap \mathcal{B}_- = \{0\}$. Then, (10) and (11) are both equivalent to

$$\begin{array}{rcl} -\mathcal{A}^*y + c^* & \in & \mathcal{L}^* \\ \mathcal{A}x - b^* & \in & \mathcal{B}_+ \\ \mathcal{A}x - b^* & \in & \mathcal{B}_-\ , \quad (x, y) \in \mathcal{L} \times Y. \\ \langle c^*, x\rangle - \langle b^*, y\rangle & \leq & 0 \end{array} \tag{12}$$

Note that in the latter formulation we might choose $\mathcal{B}_+$ to be a pointed closed convex cone and set $\mathcal{B}_- := -\mathcal{B}_+$.

The primal-dual systems (10), (11) and (12) are all special cases of the constrained cone linear system (CCLS) described as follows. Let U and V denote two vector spaces. Given a linear operator $\mathcal{E} : U \to V^*$, a closed convex set $\mathcal{U} \subseteq U$, and a vector $e \in V^*$,

and a closed convex cone $\mathcal{K} \subseteq V$, the general CCLS consists of finding a vector $u \in U$ such that

$$\mathcal{E}u - e \in \mathcal{K}^*, \quad u \in \mathcal{U}, \tag{13}$$

where $\mathcal{K}^*$ denotes the dual cone of $\mathcal{K}$. For example, (10) can be viewed as a special case of (13) by letting $U \equiv X \times Y \times X^*$, $V \equiv X \times Y \times \Re$, $\mathcal{U} \equiv \mathcal{L} \times Y \times \mathcal{L}^*$, $\mathcal{K}^* = \{0\} \subset V^*$,

$$\mathcal{E} \equiv \begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ c^* & -b^* & 0 \end{pmatrix}, \quad u \equiv \begin{bmatrix} x \\ y \\ s^* \end{bmatrix}, \text{ and } e \equiv \begin{bmatrix} c^* \\ b^* \\ 0 \end{bmatrix}.$$

Henceforth, we will consider the more general problem (13), for which we assume a solution exists. Our approach to solve this problem will be to reformulate it as

$$\bar{f} \equiv \min\{f(u) := \psi(\mathcal{E}u - e) : u \in \mathcal{U}\} \tag{14}$$

where $\psi : V^* \to \Re$ is a convex function such that

$$\mathcal{K}^* = \operatorname{Argmin}\{\psi(v^*) : v^* \in V^*\}. \tag{15}$$

Note that finding an optimal solution of (14) is equivalent to solving (13) and that the optimal value $\bar{f}$ of (14) is known, namely, $\bar{f} = \psi(v^*)$ for any $v^* \in \mathcal{K}^*$. Note also that, to obtain a concrete formulation (14), it is necessary to specify the function ψ. Also, the structure of the resulting formulation clearly depends on the properties assumed of ψ. In the next two subsections, we discuss two classes of functions ψ for which (14) can be solved by one of the smooth and/or non-smooth first-order methods proposed by Nesterov [11] and Nemirovski [9], provided that $\mathcal{U}$ is a "simple enough" set.

2.1 Smooth formulation

For a given norm $\|\cdot\|_U$ on the space U, Nesterov [10] proposed a smooth first-order optimal method to minimize a convex function with Lipschitz-continuous gradient with respect to $\|\cdot\|_U$ over a simple closed convex set in U. The following simple result gives a condition on ψ so that (14) becomes a natural candidate for Nesterov's smooth first-order optimal method.

Proposition 1 *If $\psi : V^* \to \Re$ has L-Lipschitz-continuous gradient with respect to $\|\cdot\|_V^*$, where $\|\cdot\|_V$ is a norm on V, then $f : U \to \Re$ defined in (14) has $L\|\mathcal{E}\|_{U,V}^2$-Lipschitz-continuous gradient with respect to $\|\cdot\|_U$.*

Proof. Using (14), we see that $f'(u) = \psi'(\mathcal{E}u - e) \circ \mathcal{E}$. The remaining proof immediately follows from the assumption that $\psi(\cdot)$ has L-Lipschitz-continuous gradient with respect to $\|\cdot\|_V^*$, and relations (2), (4) and (5). ∎

Throughout our presentation, we will say that (14) is a smooth formulation whenever ψ is a convex function with Lipschitz-continuous gradient with respect to $\|\cdot\|_V^*$, where $\|\cdot\|_V$ is a norm on V.

A natural example of a convex function with Lipschitz-continuous gradient for which the set of global minima is $\mathcal{K}^*$ is the square of the distance function to $\mathcal{K}^*$ measured in terms of a scalar product norm. For the sake of future reference, we list this case as follows.

Example 1 Let $\|\cdot\|^*$ denote a scalar product norm on the vector space V^*. By Proposition 5 of the Appendix, the function $\psi \equiv (\mathrm{dist}_{\mathcal{K}^*})^2$, where $\mathrm{dist}_{\mathcal{K}^*}$ is the distance function to $\mathcal{K}^*$ measured in terms of the norm $\|\cdot\|^*$, is a convex function with 2-Lipschitz-continuous gradient with respect to $\|\cdot\|^*$. In this case, formulation (14) becomes

$$\min\left\{f(u) := (\mathrm{dist}_{\mathcal{K}^*}(\mathcal{E}u - e))^2 : u \in \mathcal{U}\right\} \tag{16}$$

and its objective function f has $2\|\mathcal{E}\|^2$-Lipschitz-continuous gradient with respect to $\|\cdot\|_U$, where $\|\mathcal{E}\|$ denotes the operator norm of $\mathcal{E}$ with respect to the pair of norms $\|\cdot\|_U$ and $\|\cdot\|^*$, i.e.:

$$\|\mathcal{E}\| := \max\{\|\mathcal{E}u\|^* : \|u\|_U \leq 1\}. \tag{17}$$

In particular, note that when $\mathcal{K}^* = \{0\}$, (16) reduces to

$$\min\{f(u) := (\|\mathcal{E}u - e\|^*)^2 : u \in \mathcal{U}\}. \tag{18}$$

It shall be mentioned that the above smooth formulations can be solved by Nesterov's optimal method and its variant (see Section 3).

2.2 Nonsmooth formulation

In his more recent paper [11], Nesterov proposed a way to approximate a class of non-smooth objective functions by sufficiently close ones with Lipschitz-continuous gradient. By applying the smooth first-order method in [10] to the resulting smooth formulation, he developed an efficient numerical scheme for obtaining a near optimal solution for the original minimization problem whose objective function belongs to the forementioned class of non-smooth functions. In this subsection, we describe a class of functions ψ for which formulation (14) can be solved by the above scheme proposed by Nesterov [11].

More specifically, in this subsection, we consider functions ψ which can be expressed in the form:

$$\psi(v^*) \equiv \max\{\langle v^*, v\rangle - \tilde{\psi}(v) : v \in \mathcal{V}\}, \quad \forall v^* \in V^*, \tag{19}$$

where $\mathcal{V} \subseteq V$ is a compact convex set and $\tilde{\psi} : V \to (-\infty, \infty]$ is a proper closed convex function such that $\mathcal{V} \cap \mathrm{dom}\,\tilde{\psi} \neq \emptyset$. Using the notation of conjugate functions, this means that $\psi = (\tilde{\psi} + I_{\mathcal{V}})^*$, where $I_{\mathcal{V}}$ denotes the indicator function of $\mathcal{V}$.

The following proposition gives a necessary and sufficient condition for ψ to be a suitable objective function for (14).

Proposition 2 *Assume that* $\mathrm{ri}\,\mathcal{V} \cap \mathrm{ri}\,(\mathrm{dom}\,\tilde{\psi}) \neq \emptyset$. *Then,* $\mathcal{K}^*$ *is the set of global minima of* ψ *if and only if* $\partial\tilde{\psi}(0) + N_{\mathcal{V}}(0) = \mathcal{K}^*$.

Proof. Note that $v \in V^*$ is a global minimizer of ψ if and only if $0 \in \partial\psi(v) = \partial(\tilde{\psi}+I_{\mathcal{V}})^*(v)$, which in turn is equivalent to $v \in \partial(\tilde{\psi}+I_{\mathcal{V}})(0) = \partial\tilde{\psi}(0)+N_{\mathcal{V}}(0)$. Hence, we conclude that $\mathcal{K}^*$ is the set of global minima of ψ if and only if $\partial\tilde{\psi}(0)+N_{\mathcal{V}}(0) = \mathcal{K}^*$. ■

The most natural example of a function ψ which has a representation of the form (19) satisfying the conditions of Proposition 2 is the distance function to the cone $\mathcal{K}^*$ measured in terms of the dual norm $\|\cdot\|^*$ associated with some norm $\|\cdot\|$ on V. We record this specific example below for future reference.

Example 2 Let $\|\cdot\|$ be an arbitrary norm on V and set $\tilde{\psi} \equiv 0$ and $\mathcal{V} = \{v \in V : \|v\| \leq 1\} \cap (-\mathcal{K})$. In this case, it is easy to verify that the function ψ given by (19) is equal to the distance function $\text{dist}_{\mathcal{K}^*}$ measured in terms of $\|\cdot\|^*$ and hence satisfies (15). Note that the equivalent conditions of Proposition 2 holds too since $\partial\tilde{\psi}(0) = 0$ and $N_{\mathcal{V}}(0) = \mathcal{K}^*$. Note also that formulation (14) becomes

$$\min_{u\in\mathcal{U}} \max_{\|v\|\leq 1,\, v\in -\mathcal{K}} \langle \mathcal{E}u - e, v\rangle = \min_{u\in\mathcal{U}}\{f(u) := \text{dist}_{\mathcal{K}^*}(\mathcal{E}u - e)\}. \tag{20}$$

In particular, if $\mathcal{K}^* = \{0\}$, then the function ψ given by (19) is identical to $\|\cdot\|^*$ and the above formulation reduces to $\min\{f(u) := \|\mathcal{E}u - e\|^* : u \in \mathcal{U}\}$.

It is interesting to observe that the representation (19) can also yield functions with Lipschitz-continuous gradient. Indeed, it can be shown as in Theorem 1 of [11] that, if $\tilde{\psi}$ is strongly convex over $\mathcal{V}$ with modulus $\tilde{\sigma} > 0$ with respect to $\|\cdot\|$, then, ψ has $(1/\tilde{\sigma})$-Lipschitz-continuous gradient with respect to $\|\cdot\|^*$. In fact, under the situation described in Example 1, the function $\psi = (\text{dist}_{\mathcal{K}^*})^2$ used in (16) can be expressed as in (19) by letting $\tilde{\psi} \equiv (\|\cdot\|^*)^2/4 = \langle\cdot,\cdot\rangle^*/4$ and $\mathcal{V} = -\mathcal{K}$.

Throughout the paper, we will say that (14) is a nonsmooth formulation whenever ψ is given in the form (19) and $\tilde{\psi}$ is not strongly convex over $\mathcal{V}$. (We observe that this definition does not imply that the objective function of a nonsmooth formulation is necessarily nonsmooth.) Two first-order algorithms for solving nonsmooth formulations will be briefly reviewed in Section 4. More specifically, we describe Nesterov's smooth approximation scheme in Subsection 4.1 and Nemirovski's prox-method in Subsection 4.2.

3 First-order methods for the smooth formulation

In this section, we discuss Nesterov's smooth first-order method for solving a class of smooth CP problems. We also present and analyze a variant that has consistently outperformed Nesterov's method in our computational experiments. The convergence behavior of these methods applied to the smooth formulation (16) is also discussed.

Let U be a normed vector space with norm denoted by $\|\cdot\|_U$ and let $\mathcal{U} \subseteq U$ be a closed convex set. Assume that $f : \mathcal{U} \to \Re$ is a differentiable convex function such that for some $L \geq 0$:

$$\|f'(u) - f'(\tilde{u})\|_U^* \leq L\|u - \tilde{u}\|_U, \quad \forall u, \tilde{u} \in \mathcal{U}. \tag{21}$$

Our problem of interest in this section is the CP problem (1).

We assume throughout our discussion that the optimal value $\bar{f}$ of problem (1) is finite and that its set of optimal solutions is nonempty. Let $h_U : \mathcal{U} \to \Re$ be a continuous strongly convex function with modulus $\sigma_U > 0$ with respect to $\|\cdot\|_U$, i.e.,

$$h_U(u) \geq h_U(\tilde{u}) + \langle h_U{}'(\tilde{u}), u - \tilde{u}\rangle + \frac{\sigma_U}{2}\|u - \tilde{u}\|_U^2, \quad \forall u, \tilde{u} \in \mathcal{U}. \tag{22}$$

The Bregman distance $d_{h_U} : \mathcal{U} \times \mathcal{U} \to \Re$ associated with h_U is defined as

$$d_U(u; \tilde{u}) \equiv h_U(u) - l_{h_U}(u; \tilde{u}), \quad \forall u, \tilde{u} \in \mathcal{U}, \tag{23}$$

where $l_{h_U} : U \times \mathcal{U} \to \Re$ is the "linear approximation" of h_U defined as

$$l_{h_U}(u; \tilde{u}) = h_U(\tilde{u}) + \langle h_U{}'(\tilde{u}), u - \tilde{u}\rangle, \quad \forall (u, \tilde{u}) \in U \times \mathcal{U}.$$

We are now ready to state Nesterov's smooth first-order method for solving (1). We use the subscript 'sd' in the sequence obtained by taking a steepest descent step and the subscript 'ag' (which stands for 'aggregated gradient') in the sequence obtained by using all past gradients.

Nesterov's Algorithm:

0) Let $u_0^{sd} = u_0^{ag} \in \mathcal{U}$ be given and set $k = 0$
1) Set $u_k = \frac{2}{k+2} u_k^{ag} + \frac{k}{k+2} u_k^{sd}$ and compute $f(u_k)$ and $f'(u_k)$.
2) Compute $(u_{k+1}^{sd}, u_{k+1}^{ag}) \in \mathcal{U} \times \mathcal{U}$ as

$$u_{k+1}^{sd} \in \operatorname{Argmin} \left\{ l_f(u; u_k) + \frac{L}{2} \|u - u_k\|_U^2 : u \in \mathcal{U} \right\}, \tag{24}$$

$$u_{k+1}^{ag} \equiv \operatorname{argmin} \left\{ \frac{L}{\sigma_U} d_U(u; u_0) + \sum_{i=0}^{k} \frac{i+1}{2} [l_f(u; u_i)] : u \in \mathcal{U} \right\}. \tag{25}$$

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

end

Observe that the above algorithm is stated slightly differently than the way presented in Nesterov's paper [11]. More specifically, the indices of the sequences $\{u_k^{sd}\}$ and $\{u_k^{ag}\}$ are shifted by plus one in the above formulation. Moreover, the algorithm is stated in a different order (and with a different notation) to clearly point out its close connection to the variant discussed later in this section.

The main convergence result established by Nesterov [11] regarding the above algorithm is summarized in the following theorem.

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

$$f(u_k^{sd}) - f(u) \leq \frac{4L\, d_U(u; u_0^{sd})}{\sigma_U\, k(k+1)}, \quad \forall u \in \mathcal{U},\ k \geq 1.$$

In particular, if $\bar{u}$ denotes an optimal solution of (1), then

$$f(u_k^{sd}) - \bar{f} \leq \frac{4L\, d_U(\bar{u}; u_0^{sd})}{\sigma_U\, k(k+1)}, \quad \forall k \geq 1. \tag{26}$$

In this section, we explore the application of the above result to the situation where $\bar{f}$ is known (see Section 2). When such a priori knowledge is not available, in order to monitor the progress made by the algorithm, it becomes necessary to algorithmically estimate $\bar{f}$ or to directly estimate the right hand side of (26). However, to the best of our knowledge, these two options have been shown to be possible only under the condition that the set $\mathcal{U}$ is compact. Note that formulation (14) does not assume the latter condition but it has the property that $\bar{f}$ is known.

The following result is an immediate consequence of Theorem 1 regarding the behavior of Nesterov's optimal method applied to formulation (16).

Corollary 1 *Let $\{u_k^{sd}\}$ be the sequence generated by Nesterov's optimal method applied to problem (16). Given any $\epsilon > 0$, an iterate $u_k^{sd} \in \mathcal{U}$ satisfying* $\operatorname{dist}_{\mathcal{K}^*}(\mathcal{E} u_k^{sd} - e) \leq \epsilon$

can be found in no more than

$$\left\lceil \frac{2\sqrt{2}\|\mathcal{E}\|}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0^{sd})}{\sigma_U}} \right\rceil \tag{27}$$

iterations, where $\bar{u}$ is an optimal solution of (16) and $\|\mathcal{E}\|$ is defined in (17).

Proof. Noting that $f(u) = [\mathrm{dist}_{\mathcal{K}^*}(\mathcal{E}u - e)]^2$ for all $u \in \mathcal{U}$, $\bar{f} = 0$ and the fact that this function f has $2\|\mathcal{E}\|^2$-Lipschitz-continuous gradient with respect to $\|\cdot\|_U$, it follows from Theorem 1 that

$$\left[\mathrm{dist}_{\mathcal{K}^*}(\mathcal{E}u_k^{sd} - e)\right]^2 \leq \frac{8\|\mathcal{E}\|^2\, d_U(\bar{u};u_0^{sd})}{\sigma_U\, k(k+1)}, \quad \forall k \geq 1.$$

The corollary then follows immediately from the above relation. ∎

We next state and analyze a variant of Nesterov's optimal method which has consistently outperformed the latter one in our computational experiments.

Variant of Nesterov's algorithm:

0) Let $u_0^{sd} = u_0^{ag} \in \mathcal{U}$ be given and set $k = 0$.
1) Set $u_k = \frac{2}{k+2}u_k^{ag} + \frac{k}{k+2}u_k^{sd}$ and compute $f(u_k)$ and $f'(u_k)$.
2) Compute $(u_{k+1}^{sd}, u_{k+1}^{ag}) \in \mathcal{U} \times \mathcal{U}$ as

$$u_{k+1}^{sd} \in \mathrm{Argmin}\left\{ l_f(u;u_k) + \frac{L}{2}\|u - u_k\|_U^2 : u \in \mathcal{U} \right\}, \tag{28}$$

$$u_{k+1}^{ag} \equiv \mathrm{argmin}\left\{ \frac{k+2}{2}\, l_f(u;u_k) + \frac{L}{\sigma_U}\, d_U(u;u_k^{ag}) : u \in \mathcal{U} \right\}. \tag{29}$$

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

end

Note that the above variant differs from Nesterov's method only in the way the sequence $\{u_k^{ag}\}$ is defined. In the remaining part of this section, we will establish the following convergence result for the above variant.

Theorem 2 *The sequence $\{u_k^{sd}\}$ generated by the above variant satisfies*

$$f(u_k^{sd}) - \bar{f} \leq \frac{4Ld_U(\bar{u};u_0^{sd})}{\sigma_U k(k+2)}, \quad \forall k \geq 1, \tag{30}$$

where $\bar{u}$ is an optimal solution of (1).

Before proving the above result, we establish two technical results from which Theorem 2 immediately follows.

Let $\tau(u)$ be a convex function over a convex set $\mathcal{U} \in U$. Assume that $\hat{u}$ is an optimal solution of the problem $\min\{\tau_\eta(u) := \tau(u) + \eta\,\|u - \tilde{u}\|_U^2 : u \in \mathcal{U}\}$ for some $\tilde{u} \in \mathcal{U}$ and $\eta > 0$. Due to the well-known fact that the sum of a convex and a strongly convex function is also strongly convex, one can easily see that

$$\tau_\eta(u) \geq \min\{\tau_\eta(\check{u}) : \check{u} \in \mathcal{U}\} + \eta\|u - \hat{u}\|_U^2.$$

The next lemma generalizes this result to the case where the function $\|\cdot - \tilde{u}\|_U^2$ in the definition of $\tau_\eta(\cdot)$ is replaced with the Bregman distance $d_U(\cdot;\hat{u})$ associated with a convex function h_U. It is worth noting that the result described below does not assume the strong-convexity of the function h_U.

Lemma 1 *Let $\mathcal{U}$ be a convex set of a normed vector space U and $\tau, h_U : \mathcal{U} \to \Re$ be differentiable convex functions. Assume that $\hat{u}$ is an optimal solution of $\min\{\tau(u) + \eta\, d_U(u;\tilde{u}) : u \in \mathcal{U}\}$ for some $\tilde{u} \in \mathcal{U}$ and $\eta > 0$. Then,*

$$\tau(u) + \eta\, d_U(u;\tilde{u}) \geq \min\{\tau(\check{u}) + \eta\, d_U(\check{u};\tilde{u}) : \check{u} \in \mathcal{U}\} + \eta\, d_U(u;\hat{u}), \quad \forall u \in \mathcal{U}.$$

Proof. The definition of $\hat{u}$ and the fact that $\tau(\cdot) + \eta\, d_U(\cdot;\tilde{u})$ is a differentiable convex function imply that

$$\langle \tau'(\hat{u}) + \eta\, d'_U(\hat{u};\tilde{u}), u - \hat{u}\rangle \geq 0, \quad \forall u \in \mathcal{U},$$

where $d'_U(\hat{u};\tilde{u})$ denotes the gradient of $d_U(\cdot;\tilde{u})$ at $\hat{u}$. Using the definition of the Bregman distance (23), it is easy to verify that

$$d_U(u;\tilde{u}) = d_U(\hat{u};\tilde{u}) + \langle d'_U(\hat{u};\tilde{u}), u - \hat{u}\rangle + d_U(u;\hat{u}), \quad \forall u \in \mathcal{U}.$$

Using the above two relations and the assumption that τ is convex, we then conclude that

$$\begin{aligned}
\tau(u) + \eta\, d_U(u;\tilde{u}) &= \tau(u) + \eta\,[d_U(\hat{u};\tilde{u}) + \langle d'_U(\hat{u};\tilde{u}), u - \hat{u}\rangle + d_U(u;\hat{u})] \\
&\geq \tau(\hat{u}) + \eta\, d_U(\hat{u};\tilde{u}) + \langle \tau'(\hat{u}) + \eta\, d'_U(\hat{u};\tilde{u}), u - \hat{u}\rangle + \eta\, d_U(u;\hat{u}) \\
&\geq \tau(\hat{u}) + \eta\, d_U(\hat{u};\tilde{u}) + \eta\, d_U(u;\hat{u}),
\end{aligned}$$

and hence that the lemma holds. ∎

Before stating the next lemma, we mention a few facts that will be used in its proof. It follows from the convexity of f and assumptions (21) and (22) that

$$l_f(u,\tilde{u}) \leq f(u) \leq l_f(u,\tilde{u}) + \frac{L}{2}\,\|u - \tilde{u}\|_U^2, \quad \forall u, \tilde{u} \in \mathcal{U}, \tag{31}$$

and

$$d_U(u;\tilde{u}) \geq \frac{\sigma_U}{2}\|u - \tilde{u}\|_U^2, \quad \forall u, \tilde{u} \in \mathcal{U}. \tag{32}$$

Moreover, it can be easily verified that

$$l_f(\alpha u + \alpha' u';\tilde{u}) = \alpha l_f(u;\tilde{u}) + \alpha' l_f(u';\tilde{u}) \tag{33}$$

for any $u, u' \in U$, $\tilde{u} \in \mathcal{U}$ and $\alpha, \alpha' \in \Re$ such that $\alpha + \alpha' = 1$.

Lemma 2 *Let $(u_k^{sd}, u_k^{ag}) \in \mathcal{U} \times \mathcal{U}$ and $\alpha_k \geq 1$ be given and set $u_k \equiv (1 - \alpha_k^{-1})u_k^{sd} + \alpha_k^{-1} u_k^{ag}$. Let $(u_{k+1}^{sd}, u_{k+1}^{ag}) \in \mathcal{U} \times \mathcal{U}$ be a pair computed according to (28) and (29). Then, for every $u \in \mathcal{U}$:*

$$\alpha_k^2(f(u_{k+1}^{sd}) - f(u)) + \frac{L}{\sigma_U}\, d_U(u;u_{k+1}^{ag}) \leq (\alpha_k^2 - \alpha_k)(f(u_k^{sd}) - f(u)) + \frac{L}{\sigma_U}\, d_U(u;u_k^{ag}).$$

Proof. By the definition (28) and the second inequality in (31), it follows that

$$f(u_{k+1}^{sd}) \leq \min_{\tilde{u}\in\mathcal{U}} \left\{ l_f(\tilde{u};u_k) + \frac{L}{2}\|\tilde{u}-u_k\|_U^2 \right\}$$
$$\leq \min_{u\in\mathcal{U}} \left\{ l_f\left(\left(1-\alpha_k^{-1}\right)u_k^{sd} + \alpha_k^{-1}u;u_k\right) + \frac{L}{2}\left\|\left(1-\alpha_k^{-1}\right)u_k^{sd} + \alpha_k^{-1}u - u_k\right\|_U^2 \right\},$$

where the last inequality follows from the fact that every point of the form $\tilde{u} = \left(1-\alpha_k^{-1}\right)u_k^{sd} + \alpha_k^{-1}u$ with $u \in \mathcal{U}$ is in $\mathcal{U}$ due to the convexity of $\mathcal{U}$ and the assumption that $\alpha_k \geq 1$. The above inequality together with the definition of u_k, and relations (33), (31) and (32) then imply that

$$\alpha_k^2 f(u_{k+1}^{sd}) \leq \alpha_k^2 \min_{u\in\mathcal{U}} \left\{ \left(1-\alpha_k^{-1}\right) l_f(u_k^{sd};u_k) + \alpha_k^{-1} l_f(u;u_k) + \frac{L}{2}\alpha_k^{-2}\|u-u_k^{ag}\|_U^2 \right\}$$
$$= (\alpha_k^2-\alpha_k) l_f(u_k^{sd};u_k) + \min_{u\in\mathcal{U}} \left\{ \alpha_k l_f(u;u_k) + \frac{L}{2}\|u-u_k^{ag}\|_U^2 \right\}$$
$$\leq (\alpha_k^2-\alpha_k) f(u_k^{sd}) + \min_{u\in\mathcal{U}} \left\{ \alpha_k l_f(u;u_k) + \frac{L}{\sigma_U} d_U(u;u_k^{ag}) \right\}.$$

It then follows from this inequality, relation (29), Lemma 1 with $\tilde{u} = u_k^{ag}$, $\hat{u} = u_{k+1}^{ag}$, $\eta = L/\sigma_U$, $\tau(\cdot) \equiv \alpha_k l_f(\cdot;u_k)$ and the first inequality in (31) that for every $u \in \mathcal{U}$:

$$\alpha_k^2 f(u_{k+1}^{sd}) - (\alpha_k^2-\alpha_k) f(u_k^{sd}) + \frac{L}{\sigma_U} d_U(u;u_{k+1}^{ag})$$
$$\leq \min_{u\in\mathcal{U}} \left\{ \alpha_k l_f(u;u_k) + \frac{L}{\sigma_U} d_U(u;u_k^{ag}) \right\} + \frac{L}{\sigma_U} d_U(u;u_{k+1}^{ag})$$
$$\leq \alpha_k l_f(u;u_k) + \frac{L}{\sigma_U} d_U(u;u_k^{ag}) \;\leq\; \alpha_k f(u) + \frac{L}{\sigma_U} d_U(u;u_k^{ag}).$$

The conclusion of the lemma now follows by subtracting $\alpha_k^2 f(u)$ from both sides of the above relation and rearranging the resulting inequality. ∎

We are now ready to prove Theorem 2.

Proof of Theorem 2: Let $\bar{u}$ be an optimal solution of (1). Noting that the iterate u_k in our variant of Nesterov's algorithm can be written as $u_k \equiv (1-\alpha_k^{-1})u_k^{sd} + \alpha_k^{-1}u_k^{ag}$ with $\alpha_k = (k+2)/2$ and that the sequence $\{\alpha_k\}$ satisfies $\alpha_k \geq 1$ and $\alpha_k^2 \geq \alpha_{k+1}^2 - \alpha_{k+1}$, it follows from Lemma 2 with $u = \bar{u}$ that, for any $k \geq 0$,

$$(\alpha_{k+1}^2-\alpha_{k+1})(f(u_{k+1}^{sd}) - f(\bar{u})) + \frac{L}{\sigma_U} d_U(\bar{u};u_{k+1}^{ag})$$
$$\leq \alpha_k^2(f(u_{k+1}^{sd}) - f(\bar{u})) + \frac{L}{\sigma_U} d_U(\bar{u};u_{k+1}^{ag}) \leq (\alpha_k^2-\alpha_k)(f(u_k^{sd}) - f(\bar{u})) + \frac{L}{\sigma_U} d_U(\bar{u};u_k^{ag}),$$

from which it follows inductively that

$$(\alpha_k^2-\alpha_k)(f(u_k^{sd}) - f(\bar{u})) + \frac{L}{\sigma_U} d_U(\bar{u};u_k^{ag}) \leq (\alpha_0^2-\alpha_0)(f(u_0^{sd}) - f(\bar{u})) + \frac{L}{\sigma_U} d_U(\bar{u};u_0^{ag})$$
$$= \frac{L}{\sigma_U} d_U(\bar{u};u_0^{sd}), \quad \forall k \geq 1,$$

where the equality follows from the fact that $\alpha_0 = 1$ and $u_0^{ag} = u_0^{sd}$. Relation (30) now follows from the above inequality and the facts that $d_U(\bar{u}; u_k^{ag}) \geq 0$ and $\alpha_k^2 - \alpha_k = k(k+2)/4$.

We observe that in the above proof the only property we used about the sequence $\{\alpha_k\}$ was that it satisfies $\alpha_0 = 1 \leq \alpha_k$ and $\alpha_k^2 \geq (\alpha_{k+1}^2 - \alpha_{k+1})$ for all $k \geq 0$. Clearly, one such sequence is given by $\alpha_k = (k+2)/2$, which yields the variant of Nesterov's algorithm considered above. However, a more aggressive choice for the sequence $\{\alpha_k\}$ would be to compute it recursively using the identity $\alpha_k^2 = \alpha_{k+1}^2 - \alpha_{k+1}$ and the initial condition $\alpha_0 = 1$. We note that the latter choice was first suggested in the context of Nesterov's optimal method [10].

4 First-order methods for the nonsmooth formulation

In this section, we discuss Nesterov's smooth approximation scheme (see Subsection 4.1) and Nemirovski's prox-method (see Subsection 4.2) for solving an important class of non-smooth CP problems which admit special smooth convex-concave saddle point (i.e., min-max) reformulations. We also analyze the convergence behavior of each method in the particular context of the nonsmooth formulation (20).

4.1 Nesterov's smooth approximation scheme

Assume that U and V are normed vector spaces. Our problem of interest in this subsection is still problem (1) with the same assumption that $\mathcal{U} \subseteq U$ is a closed convex set. But now we assume that its objective function $f : \mathcal{U} \to \Re$ is a convex function given by

$$f(u) := \hat{f}(u) + \sup\{\langle \mathcal{E}u, v\rangle - \phi(v) : v \in \mathcal{V}\}, \quad \forall u \in \mathcal{U}, \tag{34}$$

where $\hat{f} : \mathcal{U} \to \Re$ is a convex function with $L_{\hat{f}}$-Lipschitz-continuous gradient with respect to a given norm $\|\cdot\|_U$ in U, $\mathcal{V} \subseteq V$ is a compact convex set, $\phi : \mathcal{V} \to \Re$ is a continuous convex function, and $\mathcal{E}$ is a linear operator from U to V^*.

Unless stated explicitly otherwise, we assume that the CP problem (1) referred to in this subsection is the one with its objective function given by (34). Also, we assume throughout our discussion that $\bar{f}$ is finite and that the set of optimal solutions of (1) is nonempty.

The function f defined as in (34) is generally non-differentiable but can be closely approximated by a function with Lipschitz-continuous gradient using the following construction due to Nesterov [11]. Let $h_V : \mathcal{V} \to \Re$ be a continuous strongly convex function with modulus $\sigma_V > 0$ with respect to a given norm $\|\cdot\|_V$ on V satisfying $\min\{h_V(v) : v \in \mathcal{V}\} = 0$. For some *smoothness* parameter $\eta > 0$, consider the following function

$$f_\eta(u) = \hat{f}(u) + \max_v \{\langle \mathcal{E}u, v\rangle - \phi(v) - \eta h_V(v) : \ v \in \mathcal{V}\}. \tag{35}$$

The next result, due to Nesterov [11], shows that f_η is a function with Lipschitz-continuous gradient with respect to $\|\cdot\|_U$ whose "closeness" to f depends linearly on the parameter η.

Theorem 3 *The following statements hold:*

a) For every $u \in \mathcal{U}$, we have $f_\eta(u) \le f(u) \le f_\eta(u) + \eta D_V$, where

$$D_V \equiv D_{h_V} := \max\{h_V(v) : v \in \mathcal{V}\}; \tag{36}$$

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

$$L_\eta := L_{\hat{f}} + \frac{\|\mathcal{E}\|_{U,V}^2}{\eta \sigma_V}. \tag{37}$$

We are now ready to state Nesterov's smooth approximation scheme to solve (1).

Nesterov's smooth approximation scheme:

0) Assume $\bar{f}$ is known and let $\epsilon > 0$ be given.
1) Let $\eta := \epsilon/(2D_V)$ and consider the approximation problem

$$\bar{f}_\eta \equiv \inf\{f_\eta(u) : u \in \mathcal{U}\}. \tag{38}$$

2) Apply Nesterov's optimal method (or its variant) to (38) and terminate whenever an iterate u_k^{sd} satisfying $f(u_k^{sd}) - \bar{f} \le \epsilon$ is found.

end

Note that the above scheme differs slightly from Nesterov's original scheme presented in [11] in that its termination is based on the exact value of $\bar{f}$, while the termination of the latter one is based on a lower bound estimate of $\bar{f}$ computed at each iteration of the optimal method applied to (38). The following result describes the convergence behavior of the above scheme. Since the proof of this result given in [11] assumes that $\mathcal{U}$ is bounded, we provide below another proof which is suitable to the case when $\mathcal{U}$ is unbounded.

Theorem 4 *Nesterov's smooth approximation scheme generates an iterate u_k^{sd} satisfying $f(u_k^{sd}) - \bar{f} \le \epsilon$ in no more than*

$$\left\lceil \sqrt{\frac{8 d_U(\bar{u}; u_0)}{\sigma_U \epsilon} \left(\frac{2 D_V \|\mathcal{E}\|_{U,V}^2}{\sigma_V \epsilon} + L_{\hat{f}} \right)} \right\rceil \tag{39}$$

iterations, where $\bar{u}$ is an optimal solution of (1) and $\|\mathcal{E}\|_{U,V}$ is defined in (3).

Proof. By Theorem 1, we have

$$f_\eta(u_k^{sd}) - f_\eta(\bar{u}) \le \frac{4 L_\eta}{k(k+1)\sigma_U} d_U(\bar{u}; u_0), \quad \forall k \ge 1.$$

Hence,

$$\begin{aligned} f(u_k^{sd}) - \bar{f} &= \Big(f(u_k^{sd}) - f_\eta(u_k^{sd}) \Big) + \Big(f_\eta(u_k^{sd}) - f_\eta(\bar{u}) \Big) + \big(f_\eta(\bar{u}) - \bar{f} \big) \\ &\le \eta D_V + \frac{4 L_\eta d_U(\bar{u}; u_0)}{\sigma_U k(k+1)}, \quad \forall k \ge 1, \end{aligned}$$

where in the last inequality we have used the facts $f(u_k^{sd}) - f_\eta(u_k^{sd}) \leq \eta D_V$ and $f_\eta(\bar{u}) - \bar{f} \leq 0$ implied by (a) of Theorem 3. The above relation together with the definition (37) and the fact $\eta = \epsilon/(2D_V)$ clearly imply that

$$\begin{aligned} f(u_k^{sd}) - \bar{f} &\leq \eta D_V + \frac{4d_U(\bar{u};u_0)}{\sigma_U k^2}\left(\frac{\|\mathcal{E}\|_{U,V}^2}{\eta\sigma_V} + L_{\hat{f}}\right) \\ &= \epsilon\left[\frac{1}{2} + \frac{4d_U(\bar{u};u_0)}{\sigma_U k^2 \epsilon}\left(\frac{2D_V\|\mathcal{E}\|_{U,V}^2}{\sigma_V\epsilon} + L_{\hat{f}}\right)\right] \end{aligned}$$

from which the claim immediately follows by rearranging the terms. ∎

Recall that the only assumption imposed on the function h_V is that it satisfies the condition that $\min\{h_V(v) : v \in \mathcal{V}\} = 0$. Letting $v^0 := \operatorname{argmin}\{h_V(v) : v \in \mathcal{V}\}$, it can be easily seen that $\min\{d_{h_V}(v;v^0) : v \in \mathcal{V}\} = 0$ and $\max\{d_{h_V}(v;v^0) : v \in \mathcal{V}\} \leq D_{h_V} := \max\{h_V(v) : v \in \mathcal{V}\}$. Thus, replacing the function h_V by the Bregman distance function $d_{h_V}(\cdot;v^0)$ only improves the iteration bound (39).

The following corollary describes the convergence result of Nesterov's smooth approximation scheme applied to the nonsmooth formulation (20). We observe that the norm used to measure the size of $\mathcal{E}u_k^{sd} - e$ is not necessarily equal to $\|\cdot\|_V^*$.

Corollary 2 *Nesterov's smooth approximation scheme applied to formulation (20) generates an iterate u_k^{sd} satisfying $\operatorname{dist}_{\mathcal{K}^*}(\mathcal{E}u_k^{sd} - e) \leq \epsilon$ in no more than*

$$\left\lceil \frac{4\|\mathcal{E}\|_{U,V}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)D_V}{\sigma_U\sigma_V}} \right\rceil \tag{40}$$

iterations, where $\bar{u}$ is an optimal solution of (20), $D_V := \max\{h_V(v) : \|v\| \leq 1, v \in -\mathcal{K}\}$, and $\|\mathcal{E}\|_{U,V}$ is defined in (3).

Proof. The bound (40) follows directly from Theorem 4 and the fact that $L_{\hat{f}} = 0$ and $\mathcal{V} = \{v : \|v\| \leq 1, v \in -\mathcal{K}\}$ for formulation (20). ∎

The next result essentially shows that, for the case when $\mathcal{K} = V$, and hence $\mathcal{K}^* = \{0\}$, the iteration-complexity bound (40) to obtain an iterate $u_k^{sd} \in \mathcal{U}$ satisfying $\|\mathcal{E}u_k^{sd} - e\|^* \leq \epsilon$ for Nesterov's smooth approximation scheme is always minorized by the iteration-complexity bound (27) for Nesterov's optimal method regardless of the norm $\|\cdot\|_V$ and strong convex function h_V chosen.

Proposition 3 *If $\mathcal{K}^* = \{0\}$, then the bound (40) is minorized by*

$$\left\lceil \frac{2\sqrt{2}\|\mathcal{E}\|}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)}{\sigma_U}} \right\rceil, \tag{41}$$

where $\|\mathcal{E}\|$ is defined in (17). Moreover, if $\|\cdot\|$ (or equivalently, $\|\cdot\|^*$) is a scalar product norm, $h_V \equiv \|\cdot\|^2/2$ and $\|\cdot\|_V \equiv \|\cdot\|$, then bound (40) is exactly equal to (41).

Proof. Let $v_0 := \operatorname{argmin}\{h_V(v) : v \in \mathcal{V}\}$. Using (36), the fact that h_V is strongly convex with modulus σ_V with respect to $\|\cdot\|_V$, and the assumption that $\mathcal{K} = V$, and hence $\mathcal{V} = \{v : \|v\| \leq 1\}$, we conclude that

$$\begin{aligned}\sqrt{\frac{D_V}{\sigma_V}} = \left(\frac{\max\{h_V(v) : \|v\| \leq 1\}}{\sigma_V}\right)^{1/2} &\geq \frac{1}{\sqrt{2}} \max\{\|v - v_0\|_V : \|v\| \leq 1\} \\ &\geq \frac{1}{\sqrt{2}} \max\{\max(\|v - v_0\|_V,\, \|v + v_0\|_V) : \|v\| \leq 1\} \\ &\geq \frac{1}{\sqrt{2}} \max\{\|v\|_V : \|v\| \leq 1\},\end{aligned}$$

where the second inequality is due to the fact that $\|v\| \leq 1$ implies $\|-v\| \leq 1$ and the last inequality follows from the triangle inequality for norms. The above relation together with relations (4) and (5) then imply that

$$\begin{aligned}\frac{4\|\mathcal{E}\|_{U,V}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)D_V}{\sigma_U\sigma_V}} &\geq \frac{2\sqrt{2}\,\max\{\|\mathcal{E}^*\|_{V,U}\|v\|_V : \|v\| \leq 1\}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)}{\sigma_U}} \\ &\geq \frac{2\sqrt{2}\,\max\{\|\mathcal{E}^*v\|_U^* : \|v\| \leq 1\}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)}{\sigma_U}} \\ &= \frac{2\sqrt{2}\,\|\mathcal{E}\|}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)}{\sigma_U}},\end{aligned}$$

where the last equality follows from (3) and (4) with $\|\cdot\|_V = \|\cdot\|$ and definition (17). The second part of the proposition follows from the fact that, under the stated assumptions, we have $\|\mathcal{E}\|_{U,V} = \|\mathcal{E}\|$, $D_V = 1/2$ and $\sigma_V = 1$, and hence that (40) is equal to (41). ∎

In words, Proposition 3 shows that when $\mathcal{K}^* = \{0\}$ and the norm $\|\cdot\|^*$ used to measure the size of $\mathcal{E}u_k^{sd} - e$ is a scalar product norm, then the best norm $\|\cdot\|_V$ to choose for Nesterov's smooth approximation scheme applied to the CP problem (20) is the norm $\|\cdot\|$.

Finally, observe that the bound (41) is the same as the bound derived for Nesterov's optimal method applied to the smooth formulation (16).

4.2 Nemirovski's Prox method for non-smooth CP

In this subsection, we briefly review Nemirovski's prox method for solving a class of saddle point (i.e., min-max) problems.

Let U and V be vector spaces and consider the product space $U \times V$ endowed with a norm denoted by $|||\cdot|||$. For the purpose of reviewing Nemirovski's prox-method, we still consider problem (1) with the same assumption that $\mathcal{U} \subseteq U$ is a closed convex set but now we assume that the objective function $f : \mathcal{U} \to \Re$ is a convex function having the following representation:

$$f(u) := \max\{\phi(u, v) : v \in \mathcal{V}\}, \quad \forall u \in \mathcal{U}, \tag{42}$$

where $\mathcal{V} \subseteq V$ is a compact convex set and $\phi : \mathcal{U} \times \mathcal{V} \to \Re$ is a function satisfying the following three conditions:

C1) $\phi(u,\cdot)$ is concave for every $u \in \mathcal{U}$;
C2) $\phi(\cdot,v)$ is convex for every $v \in \mathcal{V}$;
C3) ϕ has L-Lipschitz-continuous gradient with respect to $|||\cdot|||$.

Unless stated explicitly otherwise, we assume that the CP problem (1) referred in this subsection is the one with its objective function specialized in (42). For shortness of notation, in this subsection we denote the pair (u,v) by w and the set $\mathcal{U}\times\mathcal{V}$ by $\mathcal{W}$. For any $w \in \mathcal{W}$, let $\Phi(w) : U \times V \to \Re$ be the linear map (see Subsection 1.1) defined as

$$\Phi(w) := (\phi'_u(w), -\phi'_v(w)), \quad \forall w \in \mathcal{W},$$

where $\phi'_u(w) : U \to \Re$ and $\phi'_v(w) : V \to \Re$ denote the partial derivative of ϕ at w with respect to u and v. The motivation for considering the map Φ comes from the fact that a solution $w^* \in \mathcal{W}$ of the variational inequality $\langle \Phi(w^*), w - w^*\rangle \geq 0$ for all $w \in \mathcal{W}$, would yield a solution of the CP problem (1). It should be noted however that, in our discussion below, we do not make the assumption that this variational inequality has a solution but only the weaker assumption that the CP problem (1) has an optimal solution.

Letting $\mathcal{H} : \mathcal{W} \to \Re$ denote a differentiable strongly convex function with modulus $\sigma_{\mathcal{H}} > 0$ with respect to $|||\cdot|||$, the following algorithm for solving problem (1) has been proposed by Nemirovski [9].

Nemirovski's prox-method:

0) Let $w_0 \in \mathcal{W}$ be given and set $k = 0$.
1) Compute $\Phi(w_k)$ and let
$$w_k^{sd} = \arg\min_{w\in\mathcal{W}} \langle \Phi(w_k), w\rangle + \frac{\sqrt{2}\,L}{\sigma_{\mathcal{H}}}\, d_{\mathcal{H}}(w; w_k).$$
2) If the condition
$$\langle \Phi(w_k), w_k^{sd} - w_k\rangle + \frac{\sqrt{2}\,L}{\sigma_{\mathcal{H}}}\, d_{\mathcal{H}}(w_k^{sd}; w_k) \geq 0 \tag{43}$$
is satisfied, let $w_{k+1} = w_k^{sd}$; otherwise, compute $\Phi(w_k^{sd})$ and let
$$w_{k+1} = \arg\min_{w\in\mathcal{W}} \langle \Phi(w_k^{sd}), w\rangle + \frac{\sqrt{2}\,L}{\sigma_{\mathcal{H}}}\, d_{\mathcal{H}}(w; w_k). \tag{44}$$
3) Set $k \leftarrow k+1$ and go to step 1.

Although stated slightly differently here, the above algorithm is exactly the one for which Nemirovski obtained an iteration-complexity bound in [9]. We also observe that a slight variation of the above algorithm, where the test (43) is skipped and w_{k+1} is always computed by means of (44), has been first proposed by Korpelevich [7] for the special case of convex programming but no uniform rate of convergence is given. Note that each iteration of Nemirovski's prox-method requires at most two gradient evaluations and the solutions of at most two subproblems.

The following result states the convergence properties of the above algorithm. We include a short proof extending the one of Theorem 3.2 of Nemirovski [9] to the case where the set $\mathcal{U}$ is unbounded.

Theorem 5 *Assume that $\phi : \mathcal{W} \to \Re$ is a function satisfying conditions C1)-C3) and that $\mathcal{H} : \mathcal{W} \to \Re$ is a differentiable strongly convex function with modulus $\sigma_{\mathcal{H}} > 0$ with*

respect to $|||\cdot|||$. *Assume also that* $\bar{u}$ *is an optimal solution of problem (1). Then, the sequence of iterates* $\{w_k^{sd} =: (u_k^{sd}, v_k^{sd})\}$ *generated by Nemirovski's algorithm satisfies*

$$f(u_k^{ag}) - \bar{f} \leq \frac{\sqrt{2}\, L\, D_{\mathcal{H}}(\bar{u}; w_0)}{\sigma_{\mathcal{H}}\, k}, \quad \forall k \geq 0,$$

where

$$u_k^{ag} := \frac{1}{k+1} \sum_{l=0}^{k} u_l^{sd}, \quad \forall k \geq 0,$$

$$D_{\mathcal{H}}(\bar{u}; w_0) := \max\left\{ d_{\mathcal{H}}(w; w_0) : w \in \{\bar{u}\} \times \mathcal{V} \right\}. \tag{45}$$

Proof. The arguments used in proof of Proposition 2.2 of Nemirovski [9] (specifically, by letting $\gamma_t \equiv \sigma_{\mathcal{H}}/(\sqrt{2}L)$ and $\epsilon_t \equiv 0$ in the inequalities given in the first seven lines on page 236 of [9]) imply that

$$\phi(u_k^{ag}, v) - \phi(u, v_k^{ag}) \leq \frac{\sqrt{2}\, L\, d_{\mathcal{H}}(w; w_0)}{\sigma_{\mathcal{H}}\, k}, \quad \forall w = (u, v) \in \mathcal{W},$$

where $v_k^{ag} := (\sum_{l=0}^{k} v_l^{sd})/(k+1)$. Using the definition (42), the inequality obtained by maximizing both sides of the above relation over the set $\{\bar{u}\} \times \mathcal{V}$, and the definition (45), we obtain

$$\begin{aligned} f(u_k^{ag}) - f(\bar{u}) &= \max_{v \in \mathcal{V}} \phi(u_k^{ag}, v) - \max_{v \in \mathcal{V}} \phi(\bar{u}, v) \\ &\leq \left[\max_{v \in \mathcal{V}} \phi(u_k^{ag}, v) \right] - \phi(\bar{u}, v_k^{ag}) \;\leq\; \frac{\sqrt{2}\, L\, D_{\mathcal{H}}(\bar{u}; w_0)}{\sigma_{\mathcal{H}}\, k}. \end{aligned}$$

■

Now, we examine the consequences of applying Nemirovski's prox-method to the nonsmooth formulation (20) for which $\phi(u, v) = \langle \mathcal{E}u - e, v \rangle$. For that end, assume that $\|\cdot\|_U$ and $\|\cdot\|_V$ are given norms for the vector spaces U and V, respectively, and consider the norm on the product space $U \times V$ given by

$$|||w||| := \sqrt{{\theta_U}^2 \|u\|_U^2 + {\theta_V}^2 \|v\|_V^2}\ , \tag{46}$$

where $\theta_U, \theta_V > 0$ are positive scalars. Assume also that $h_U : \mathcal{U} \to \Re$ and $h_V : \mathcal{V} \to \Re$ are given differentiable strongly convex functions with modulus σ_U and σ_V with respect to $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively, and consider the following function $\mathcal{H} : \mathcal{W} \to \Re$ defined as

$$\mathcal{H}(w) := \gamma_U h_U(u) + \gamma_V h_V(v), \quad \forall w = (u, v) \in \mathcal{W}, \tag{47}$$

where $\gamma_U, \gamma_V > 0$ are positive scalars. It is easy to check that $\mathcal{H}$ is a differentiable strongly convex function with modulus

$$\sigma_{\mathcal{H}} = \min\{\sigma_U \gamma_U {\theta_U}^{-2}, \sigma_V \gamma_V {\theta_V}^{-2}\} \tag{48}$$

with respect to the norm (46), and that the dual norm of (46) is given by

$$|||(u^*, v^*)|||^* := \sqrt{{\theta_U}^{-2} (\|u^*\|_U^*)^2 + {\theta_V}^{-2} (\|v^*\|_V^*)^2}, \quad \forall (u^*, v^*) \in U^* \times V^*. \tag{49}$$

The following result describes the convergence behavior of Nemirovski's prox-method applied to problem (20).

Corollary 3 *For some positive scalars $\theta_U, \theta_V, \gamma_U$ and γ_V, consider the norm $|||\cdot|||$ and function $\mathcal{H}$ defined in (46) and (47), respectively. Then, Nemirovski's prox-method applied to (20) generates a point $u_k^{ag} := \sum_{l=0}^{k} u_l^{sd}/(k+1) \in \mathcal{U}$ satisfying $\mathrm{dist}_{\mathcal{K}^*}(\mathcal{E}u_k^{ag} - e) \leq \epsilon$ in no more than*

$$\left\lceil \frac{\sqrt{2}\,{\theta_U}^{-1}{\theta_V}^{-1}\|\mathcal{E}\|_{U,V}\,(\gamma_U d_U(\bar{u};u_0) + \gamma_V D_V(v_0))}{\min\{\sigma_U\gamma_U{\theta_U}^{-2}, \sigma_V\gamma_V{\theta_V}^{-2}\}\,\epsilon} \right\rceil \tag{50}$$

iterations, where $\bar{u}$ is an optimal solution of (20), $\|\mathcal{E}\|_{U,V}$ is defined in (3), and

$$D_V(v_0) := \max\{d_{h_V}(v;v_0) : \|v\| \leq 1, v \in -\mathcal{K}\}. \tag{51}$$

Proof. Noting that $\phi(w) := \langle \mathcal{E}u - e, v\rangle$ and $\Phi(w) := (\mathcal{E}^*v, -\mathcal{E}u + d)$ for problem (20) and using relations (49), (5) and (46), we have

$$\begin{aligned}
|||\Phi(w_1) - \Phi(w_2)|||^* &= \sqrt{{\theta_U}^{-2}(\|\mathcal{E}^*(v_1 - v_2)\|_U^*)^2 + {\theta_V}^{-2}(\|\mathcal{E}(u_1 - u_2)\|_V^*)^2} \\
&\leq \sqrt{{\theta_U}^{-2}\|\mathcal{E}\|_{U,V}^2\|v_1 - v_2\|_V^2 + {\theta_V}^{-2}\|\mathcal{E}\|_{U,V}^2\|u_1 - u_2\|_U^2} \\
&= \sqrt{{\theta_U}^{-2}{\theta_V}^{-2}\|\mathcal{E}\|_{U,V}^2\left({\theta_V}^2\|v_1 - v_2\|_V^2 + {\theta_U}^2\|u_1 - u_2\|_U^2\right)} \\
&= {\theta_U}^{-1}{\theta_V}^{-1}\|\mathcal{E}\|_{U,V}|||w_1 - w_2|||,
\end{aligned}$$

from which we conclude that ϕ has L-Lipschitz-continuous gradient with respect to $|||\cdot|||$, where $L = {\theta_U}^{-1}{\theta_V}^{-1}\|\mathcal{E}\|_{U,V}$. Also, using relations (45), (47) and (51) and the fact that, for formulation (20), the set $\mathcal{V}$ is equal to $\{v \in V : \|v\| \leq 1, v \in \mathcal{K}\}$, we easily see that $D_{\mathcal{H}}(\bar{u};w_0) = \gamma_U d_U(\bar{u};u_0) + \gamma_V D_V(v_0)$. The result now follows from Theorem 5, the above two conclusions and relation (48). ∎

An interesting issue to examine is how the bound (50) compares with the bound (40) derived for Nesterov's smooth approximation scheme. The following result shows that the bound (40) minorizes, up to a constant factor, the bound (50).

Proposition 4 *Regardless of the values of positive scalars $\theta_U, \theta_V, \gamma_U$ and γ_V, the bound (50) is minorized by*

$$\left\lceil \frac{2\sqrt{2}\|\mathcal{E}\|_{U,V}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)D_V(v_0)}{\sigma_U\sigma_V}} \right\rceil. \tag{52}$$

Proof. Letting K denote the bound (50), we easily see that

$$K \geq \frac{\sqrt{2}\|\mathcal{E}\|_{U,V}}{\epsilon}\left(\frac{\theta_U d_U(\bar{u};u_0)}{\theta_V\sigma_U} + \frac{\theta_V D_V(v_0)}{\theta_U\sigma_V}\right) \geq \frac{2\sqrt{2}\|\mathcal{E}\|_{U,V}}{\epsilon}\sqrt{\frac{d_U(\bar{u};u_0)D_V(v_0)}{\sigma_U\sigma_V}},$$

where the last inequality follows from the fact that $\alpha^2 + \tilde{\alpha}^2 \geq 2\alpha\tilde{\alpha}$ for any $\alpha, \tilde{\alpha} \in \Re$. ∎

Observe that the lower bound (52) is equal to the bound (40) divided by $\sqrt{2}$ whenever the function h_V used in Nesterov's smooth approximation scheme is chosen as the Bregman distance function $d_{h_V}(\cdot; v_0)$, where v_0 is the v-component of w^0 (see also the discussion after the proof of Theorem 4). Observe also that the bound (50)

becomes equal to the lower bound (52) upon choosing the scalars $\theta_U, \theta_V, \gamma_U$ and γ_V as

$$\theta_U = \sqrt{\frac{\sigma_U}{d_U(\bar{u};u_0)}}, \quad \theta_V = \sqrt{\frac{\sigma_V}{D_V(v_0)}}, \quad \gamma_U = \frac{1}{d_U(\bar{u};u_0)}, \quad \gamma_V = \frac{1}{D_V(v_0)}.$$

Unfortunately, we can not use the above optimal values since the values for θ_U and γ_U depend on the unknown quantity $d_U(\bar{u};u_0)$. However, if an upper bound on $d_U(\bar{u};u_0)$ is known or a reasonable guess of this quantity is made, a reasonable approach is to replace $d_U(\bar{u};u_0)$ by its upper bound or estimate on the above formulas for θ_U and γ_U. It should be noted however that if the estimate of $d_U(\bar{u};u_0)$ is poor then the resulting iteration bound for Nemirovski's prox method may be significantly larger than the one for Nesterov's smooth approximation scheme.

5 Computational results

In this section, we report the results of our computational experiments where we compare the performance of the four first-order methods discussed in the previous sections applied to the CCLS (10) corresponding to two instance sets of cone programming problems, namely: linear programming and semidefinite programming. We also compare the performance of the above first-order methods applied to CCLS (10) against the low-rank method [2,3] on a set of randomly generated SDP instances.

5.1 Algorithm setup

As mentioned in Section 2, there are numerous ways to reformulate the general cone programming as a CCLS. In our computational experiments though, we only consider the CCLS (10) and its associated primal-dual formulations (16) and (20). In this subsection, we provide a detailed description of the norm $\|\cdot\|^*$ used in the primal-dual formulations (16) and (20), the norms $\|\cdot\|_U$ and $\|\cdot\|_V$ and functions h_U and h_V used by the several first-order methods and the termination criterion employed in our computational experiments. Since we only deal with linear programming and semidefinite programming, we assume throughout this subsection that the spaces $X = X^*$, $Y = Y^*$ are finite-dimensional Euclidean spaces endowed with the standard Euclidean norm which we denote by $\|\cdot\|_2$.

We assume that the norm $\|\cdot\|^*$ on $X^* \times Y^* \times \Re$ used in both formulations (16) and (20) is given by

$$\|(x^*, y^*, t)\|^* := \sqrt{\omega_d^2 \|x^*\|_2^2 + \omega_p^2 \|y^*\|_2^2 + \omega_o^2 |t|^2}, \quad \forall (x^*, y^*, t) \in X^* \times Y^* \times \Re, \quad (53)$$

where ω_d, ω_p and ω_o are prespecified positive constants. Reasonable choices for these constants will be discussed in the next two subsections.

We next describe the norm $\|\cdot\|_U$ and function h_U used in our implementation of Nesterov's optimal method and its variant proposed in this paper. We define $\|\cdot\|_U$ as

$$\|(x, y, s^*)\|_U := \sqrt{\theta_X^2 \|x\|_2^2 + \theta_Y^2 \|y\|_2^2 + \theta_S^2 \|s^*\|_2^2}, \quad \forall (x, y, s^*) \in U \equiv X \times Y \times X^*, \quad (54)$$

where θ_X, θ_Y, and θ_S are some positive scalars whose specific choice will be described below. We then set $h_U(\cdot) := \frac{1}{2}\|\cdot\|_U^2$. Note that this function h_U is strongly convex

with modulus $\sigma_U = 1$ with respect to $\|\cdot\|_U$. We use $u_0 = (x_0, y_0, s_0) = (0, 0, 0)$ as initial point for both algorithms.

The choice of the scalars θ_X, θ_Y and θ_S are made so as to minimize an upper bound on the iteration-complexity bound (27). It can be easily verified that

$$\|\mathcal{E}\| \leq \sqrt{\theta_X^{-2}\mathcal{F}_X^2 + \theta_Y^{-2}\mathcal{F}_Y^2 + \theta_S^{-2}\mathcal{F}_S^2},$$

where

$$\mathcal{F}_X := \sqrt{\omega_p^2\|\mathcal{A}\|_2^2 + \omega_o^2\|c^*\|_2^2}, \quad \mathcal{F}_Y := \sqrt{\omega_d^2\|\mathcal{A}\|_2^2 + \omega_o^2\|b^*\|_2^2}, \quad \mathcal{F}_S := \omega_d,$$

and $\|\mathcal{A}\|_2$ is the spectral norm of $\mathcal{A}$. Letting $\bar{u} =: (\bar{x}, \bar{y}, \bar{s}^*)$ be the optimal solution that appears in (27), we see that

$$d_U(\bar{u}; u_0) \leq \frac{\|\bar{y}\|^2}{2}\left(\theta_X^2\mathcal{Q}_X^2 + \theta_Y^2 + \theta_S^2\mathcal{Q}_S^2\right),$$

where $\mathcal{Q}_X$ and $\mathcal{Q}_Y$ are arbitrary constants satisfying

$$\mathcal{Q}_X \geq \|\bar{x}\|_2/\|\bar{y}\|_2, \quad \mathcal{Q}_S \geq \|\bar{s}^*\|_2/\|\bar{y}\|_2. \tag{55}$$

The above two relations together with the fact that $\sigma_U = 1$ then clearly imply that (27) can be bounded by

$$\frac{2\|\bar{y}\|}{\epsilon}\sqrt{\theta_X^{-2}\mathcal{F}_X^2 + \theta_Y^{-2}\mathcal{F}_Y^2 + \theta_S^{-2}\mathcal{F}_S^2}\sqrt{\theta_X^2\mathcal{Q}_X^2 + \theta_Y^2 + \theta_S^2\mathcal{Q}_S^2}.$$

In terms of the bounds $\mathcal{Q}_X$ and $\mathcal{Q}_S$, a set of scalars θ_X, θ_Y and θ_S that minimizes the above bound is

$$\theta_X = \left(\tfrac{\mathcal{F}_X}{\mathcal{Q}_X}\right)^{1/2}, \quad \theta_Y = \mathcal{F}_Y^{1/2}, \quad \theta_S = \left(\tfrac{\mathcal{F}_S}{\mathcal{Q}_S}\right)^{1/2}. \tag{56}$$

Note that the formulae for θ_X, θ_Y and θ_S depend on the bounds $\mathcal{Q}_X$ and $\mathcal{Q}_S$, which are required to satisfy (55). Since $(\bar{x}, \bar{y}, \bar{s}^*)$ is unknown, the ratios $\|\bar{x}\|_2/\|\bar{y}\|_2$ and $\|\bar{s}^*\|_2/\|\bar{y}\|_2$ can not be precisely computed. However, making the reasonable assumption that $\|\bar{y}\|_2 \geq 1$, then we can estimate the second ratio as

$$\frac{\|\bar{s}^*\|_2}{\|\bar{y}\|_2} \leq \frac{\|\mathcal{A}^*\bar{y} - c^*\|_2}{\|\bar{y}\|_2} \leq \frac{\|\mathcal{A}^*\|_2\|\bar{y}\|_2 + \|c^*\|_2}{\|\bar{y}\|_2} \leq \|\mathcal{A}\|_2 + \|c^*\|_2,$$

and hence set $\mathcal{Q}_S := \|\mathcal{A}\|_2 + \|c^*\|_2$. As for the other ratio, we do not know a sound way to majorize it and hence we simply employ the following adhoc scheme which defines $\mathcal{Q}_X$ as $\mathcal{Q}_X := \sqrt{p/q}$ where p and q are the dimensions of the spaces X and Y, respectively.

As for Nesterov's smooth approximation scheme and Nemirovski's prox-method, we set the norm $\|\cdot\|_U$ and the function h_U in exactly the same way as those for Nesterov's optimal method described above. Moreover, in view of Proposition 3, we choose the norm $\|\cdot\|_V \equiv \|\cdot\|$ and the function $h_V(\cdot) \equiv \|\cdot\|^2/2$ in order to minimize the iteration-complexity bound (40) or an estimated iteration-complexity bound (52), respectively, for Nesterov's smooth approximation scheme and Nemirovski's prox-method.

The following two termination criterion are used in our computational studies. For a given termination parameter $\epsilon > 0$, we want to find a point $(x, y, s^*) \in X \times Y \times X$ such that

$$\max\left\{\frac{\|\mathcal{A}^* y + s^* - c^*\|_2}{\max(1, \|c^*\|_2)}, \frac{\|\mathcal{A}x - b^*\|_2}{\max(1, \|b^*\|_2)}, \frac{|\langle c^*, x\rangle - \langle b^*, y\rangle|}{\max(1, (|\langle c^*, x\rangle| + |\langle b^*, y\rangle|)/2)}\right\} \leq \epsilon, \quad (57)$$

or a point $(x, y) \in X \times Y$ satisfying

$$\max\left\{\frac{d_{\mathcal{L}^*}(c^* - \mathcal{A}^* y)}{\max(1, \|c^*\|_2)}, \frac{\|\mathcal{A}x - b^*\|_2}{\|b^*\|_2}, \frac{|\langle c^*, x\rangle - \langle b^*, y\rangle|}{\max(1, (|\langle c^*, x\rangle| + |\langle b^*, y\rangle|)/2)}\right\} \leq \epsilon. \quad (58)$$

We observe that the termination criterion (57) is exactly the same as the one used by the code SDPT3 [13] for solving SDP problems. Observe also that the left hand side of the alternative criterion (58) is obtained by minimizing the left hand side of (57) for all $s^* \in \mathcal{L}^*$ and, as a result, it does not depend on s^*. The latter termination criterion is used when comparing the variant of Nesterov's method with Burer and Monteiro's low-rank method which, being a primal-only method, generates only x and a dual estimate y based on x.

5.2 Comparisons of three first-order methods for LP

In this subsection, we compare the performance of Nesterov's optimal method, Nesterov's smooth approximation scheme, and Nemirovski's prox-method on a set of randomly generated LP instances.

We use the algorithm setup as described in Subsection 5.1. To be more compatible with the termination criterion (57), we choose the following weights ω_p, ω_d and ω_o for the norm $\|\cdot\|^*$ defined in (53):

$$\omega_d = \frac{1}{\max(1, \|c\|_2)}, \quad \omega_p = \frac{1}{\max(1, \|b\|_2)}, \quad \omega_o = \frac{1}{\max(1, \|b\|_2 + \|c\|_2)}. \quad (59)$$

For our experiment, twenty seven LP instances were randomly generated as follows. First, we randomly generate a matrix $\mathcal{A} \in \Re^{m \times n}$ with prescribed density ρ (i.e., the percentage of nonzero elements of the matrix), and vectors $x^0, s^0 \in \Re^n_+$ and $y^0 \in \Re^m$. Then, we set $b = \mathcal{A}x^0$ and $c = \mathcal{A}^T y^0 + s^0$. Clearly, the resulting LPs have primal-dual optimal solutions. The number of variables n, the number of constraints m, and the density parameter ρ for these twenty seven LPs are listed in columns one to three of Table 5.2, respectively.

The codes for the aforementioned three first-order methods (abbreviated as NES-S, NES-N and NEM, respectively) used for this comparison are written in Matlab. The termination criterion (57) with $\epsilon = 0.01$ is used for all three methods. All computations are performed on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8. The performance of NES-S, NES-N and NEM for the above LP instances are presented in Table 5.2. The numbers of iterations of NES-S, NES-N and NEM are given in columns four to six, and CPU times (in seconds) are given in the last three columns, respectively. From Table 5.2, we conclude that NES-S, that is, Nesterov's optimal method, is the most efficient one among these three first-order methods for solving LPs.

Problem			Iter			Time		
n	m	ρ	NES-S	NES-N	NEM	NES-S	NES-N	NEM
1000	100	1%	1396	2532	7088	11.24	20.69	303.82
1000	100	5%	1340	2331	8473	12.32	21.74	398.22
1000	100	10%	1229	1995	6968	12.92	21.21	349.47
1000	500	1%	1019	2293	3780	11.17	25.38	235.76
1000	500	5%	839	2096	2162	14.14	35.45	158.74
1000	500	10%	647	1842	1574	15.97	45.48	134.37
1000	900	1%	1123	2778	4246	15.31	38.29	334.95
1000	900	5%	695	2247	2081	17.63	57.01	204.53
1000	900	10%	714	2122	2105	27.65	81.70	246.95
5000	500	1%	3289	5377	45657	163.41	270.56	11830.66
5000	500	5%	2235	3307	23297	204.64	303.21	7458.39
5000	500	10%	1094	2452	3391	157.60	350.88	1282.41
5000	2500	1%	1945	4683	11645	219.40	531.45	5114.41
5000	2500	5%	1248	4242	7093	450.26	1513.71	5189.31
5000	2500	10%	1335	3826	7177	890.61	2543.32	7977.18
5000	4500	1%	1499	5353	10229	297.23	1055.15	6434.42
5000	4500	5%	1632	4836	10950	1216.35	3654.01	15104.14
5000	4500	10%	1649	4568	10892	2358.79	6535.41	25046.20
10000	1000	1%	4096	6447	74335	557.05	876.63	44987.31
10000	5000	1%	1906	6507	14743	880.24	2977.75	18373.51
10000	9000	1%	2208	7747	20846	1825.14	6371.42	40180.74

Table 1 Comparison of the three methods on random LP problems

Group	m	n	ρ	$q \equiv m\rho/n$	Instances
group 1	1600	80	80%	16	sdp-d11, sdp-d12, sdp-d13
group 2	1600	80	60%	12	sdp-d21, sdp-d22, sdp-d23
group 3	2000	100	20%	4	sdp-d31, sdp-d32, sdp-d33
group 4	6000	150	2%	0.8	sdp-d41, sdp-d42, sdp-d43
group 5	10000	200	1%	0.5	sdp-d51, sdp-d52, sdp-d53

Table 2 SDP instances

5.3 Computational comparison on SDP problems

In this subsection, we compare the performance of Nesterov's optimal method, its variant proposed in Section 3, and the low-rank method [2,3] on a set of randomly generated SDP instances. The codes of these three methods are all written in ANSI C for this comparison. The experiments were conducted on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8.

Table 2 describes five groups of SDP instances which were randomly generated in a similar way as the LP instances of Subsection 5.2. Within each group, we generated three SDP instances having the same number of constraints m, the same dimension $n(n+1)/2$ of the variable x, and density ρ as listed in columns two to four, respectively. Column five displays the quantity $q \equiv m\rho/n$, which is proportional to the ratio between the arithmetic complexity of an iteration of the low-rank method and that of Nesterov's optimal method or its variant. Indeed, the arithmetic complexity of an iteration of Nesterov's optimal method and/or its variant can be bounded by $\mathcal{O}(mn^2\rho + n^3)$ (see Section 3), while that of the low-rank method can be bounded by $\mathcal{O}(n^2r + mn^2\rho)$, where r is a positive number satisfying $r(r+1)/2 \leq m$ (see the discussions in [2,3]).

In Subsection 5.2, we have already seen that Nesterov's optimal method computationally outperforms the other two first-order methods, namely, Nesterov's smooth

Instance	Nesterov's method	The variant	Improvement
sdp-d11	642	492	23.36%
sdp-d12	742	590	20.49%
sdp-d13	847	658	22.31%
sdp-d21	815	612	24.91%
sdp-d22	913	713	21.91%
sdp-d23	817	605	25.95%
sdp-d31	24,900	21,100	15.26%
sdp-d32	28,866	24,086	16.56%
sdp-d33	26,150	22,377	14.43%
sdp-d41	135,084	108,121	19.96%
sdp-d42	136,592	109,013	20.19%
sdp-d43	127,647	104,851	17.86%
sdp-d51	93,164	78,154	16.11%
sdp-d52	96,450	81,088	15.93%
sdp-d53	87,388	73,651	15.72%

Table 3 Comparison of Nesterov's method and its variant

approximation scheme and Nemirovski's prox-method for solving LP problems. In addition, our preliminary computational study shows that the same conclusion also holds for SDP problems. In the first experiment of this subsection, we compare the best of the three aforementioned methods, namely, Nesterov's optimal method with its variant. We use the same algorithm setup as described in Subsection 5.1 for these two methods. The termination criterion (57) with $\epsilon = 2.0e-3$ is used for both methods. The computational results are reported in Table 3. The instance names and the number of iterations of both methods are given in columns one to three, respectively. In the last column, we report the percentage of improvement of the variant compared to the original method in terms of number of iterations. Note that since both methods have similar computational cost per iteration, the number of iterations is a good measure for comparing both methods. The results given in Table 3 show that the variant consistently outperforms the original method.

In the second experiment, we compare the variant of Nesterov's optimal method with the low-rank method. We divide this comparison into two parts. In the first part, the stopping criterion (58) with $\epsilon = 2.0e-2$ is used for both methods. The computational results are presented in Table 4. The instance names are listed in column one. The CPU times (in seconds) for both methods are presented in columns two and four, and the number of iterations for both methods are given in columns three and five, respectively. Table 4 shows that the performance of both methods is comparable. In the second part, the termination criterion (58) with $\epsilon = 1.0e-5$ is used for both methods. Also, each method is given an upper bound of two hours (or 7,200 seconds) of computation time on each instance. The computational results are reported in Table 5. Column one gives the instance names. The final accuracies that these methods achieve are given in columns two and four, the CPU times (in seconds) are given in columns three and six, and the number of iterations are given in columns four and seven, respectively. From Table 5, we see that the low-rank method outperforms the variant of Nesterov's optimal method on all instances.

¿From the last experiment above, we conclude that the performance of the variant of Nesterov's optimal method and the low-rank method is comparable for obtaining solutions with low accuracy, but the low-rank method has much better performance when solutions with high accuracy are computed. We note however that, while the low-

	Low-rank		The new variant	
Instance	Time	Iterations	Time	Iterations
sdp-d11	2	10	10	60
sdp-d12	1	10	12	60
sdp-d13	2	10	11	60
sdp-d21	2	20	9	60
sdp-d22	2	20	9	60
sdp-d23	1	20	9	60
sdp-d31	5	60	11	90
sdp-d32	5	60	11	90
sdp-d33	4	50	10	90
sdp-d41	28	500	20	110
sdp-d42	51	880	20	110
sdp-d43	37	620	21	110
sdp-d51	52	440	62	170
sdp-d52	54	470	62	170
sdp-d53	49	400	69	190

Table 4 Comparison of the variant and low-rank method for $\epsilon = 0.02$

	Low Rank			P-D First-order		
Instance	Accuracy	Time	Iterations	Accuracy	Time	Iterations
sdp-d11	1.1e-4	7,200	56,940	5.8e-4	7,200	42,440
sdp-d12	1.4e-4	7,200	57,370	5.9e-4	7,200	42,590
sdp-d13	5.6e-5	7,200	56,980	5.8e-4	7,200	42,460
sdp-d21	5.5e-5	7,200	75,890	5.2e-4	7,200	54,010
sdp-d22	4.3e-5	7,200	75,890	4.9e-4	7,200	54,030
sdp-d23	3.8e-5	7,200	75,180	4.8e-4	7,200	54,070
sdp-d31	1.0e-5	594	9,510	1.2e-3	7,200	64,990
sdp-d32	1.0e-5	1,387	22,060	1.2e-3	7,200	65,950
sdp-d33	1.0e-5	590	9,440	1.3e-3	7,200	64,860
sdp-d41	1.0e-5	126	2,250	3.9e-3	7,200	42,530
sdp-d42	1.0e-5	161	2,810	3.8e-3	7,200	42,500
sdp-d43	1.0e-5	121	2,050	4.7e-3	7,200	42,480
sdp-d51	1.0e-5	227	1,960	9.9e-3	7,200	20,810
sdp-d52	1.0e-5	210	1,960	9.7e-3	7,200	20,470
sdp-d53	1.0e-5	263	2,100	1.0e-2	7,200	20,670

Table 5 Comparison of the variant and low-rank method for $\epsilon = 1.0e-5$

rank method has only been implemented for SDP problems, the variant of Nesterov's optimal method can potentially solve a large class of cone programming problems.

6 Concluding remarks

In this paper, we discussed primal-dual first-order methods, namely, Nesterov's optimal method and its variant, Nesterov's smooth approximation scheme, and Nemirovski's prox-method for cone programming problems. The computational results showed that the variant of Nesterov's optimal method is the best one among these methods. In this section, we propose some potential topics for future research.

All computational studies performed in this paper are based on the primal-dual formulation (10). As mentioned in Section 2, there are a variety of formulations that are suitable for the variant of Nesterov's optimal method proposed in this paper. We

would like to see how the performance of the variant of Nesterov's optimal method varies with the different formulations.

When applied to solve LP, an iteration of the methods discussed in this paper is fairly cheap since it only involves matrix-vector multiplications. Preliminary computational results also show that these methods are very promising for finding solutions with low accuracy of large-scale sparse and/or dense LPs. It would be interesting to compare these methods with other ones such as interior point and simplex methods on a collection of LP instances from Netlib. In addition, we expect these methods to be capable of finding solutions with low accuracy for large-scale second-order cone programming problems.

Finally, the primal-dual first-order methods discussed in this paper are presently only suitable for solving linear cone programming problems. A natural question is whether these methods can be extended to solve convex quadratic cone programming problems.

Appendix

Proposition 5 *Given a scalar product norm $\|\cdot\|$ on the vector space V and a closed convex set $\mathcal{C} \subseteq V$, the function $\psi : V \to \Re$ defined as $\psi(v) = (d_{\mathcal{C}}(v))^2$ is convex and has 2-Lipschitz-continuous gradient with respect to $\|\cdot\|$, where $d_{\mathcal{C}}$ is the distance function to $\mathcal{C}$ measured in terms of $\|\cdot\|$.*

Proof. The convexity of ψ follows directly from the facts that the distance function $d_{\mathcal{C}}$ is convex (see Example IV.1.3 (c) on page 153 of [5]) and that ψ is the post-composition of $d_{\mathcal{C}}$ with the increasing convex function $\tau(t) = t^2$, for $t \geq 0$ (see Proposition IV.2.1.8 in [5]). Moreover, it is shown in Example XI.3.4.4 on page 121 of [5] that ψ is differentiable with derivative ψ' given by

$$\psi'(v) = 2(v - \Pi_{\mathcal{C}}(v)), \quad \forall v \in V, \tag{60}$$

where $\Pi_{\mathcal{C}}(v) := \operatorname{argmin}_{\tilde{v} \in \mathcal{C}}\{\|v - \tilde{v}\|\}$. Also, denoting the scalar product associated with the norm $\|\cdot\|$ by $\langle \cdot, \cdot \rangle$, it follows from Proposition III.3.1.3 of [5] that

$$\langle \Pi_{\mathcal{C}}(v_1) - \Pi_{\mathcal{C}}(v_2), v_1 - v_2 \rangle \geq \|\Pi_{\mathcal{C}}(v_1) - \Pi_{\mathcal{C}}(v_2)\|^2, \quad \forall v_1, v_2 \in V,$$

which clearly implies that

$$\begin{aligned}
&\|(v_1 - \Pi_{\mathcal{C}}(v_1) - (v_2 - \Pi_{\mathcal{C}}(v_2))\|^2 \\
&= \|v_1 - v_2\|^2 - 2\langle \Pi_{\mathcal{C}}(v_1) - \Pi_{\mathcal{C}}(v_2), v_1 - v_2 \rangle + \|\Pi_{\mathcal{C}}(v_1) - \Pi_{\mathcal{C}}(v_2)\|^2 \\
&\leq \|v_1 - v_2\|^2 - \|\Pi_{\mathcal{C}}(v_1) - \Pi_{\mathcal{C}}(v_2)\|^2 \ \leq \ \|v_1 - v_2\|^2.
\end{aligned}$$

Hence, we conclude from identity (60) and the above relation that ψ has 2-Lipschitz-continuous gradient with respect to $\|\cdot\|$. ∎

References

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

2. S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 95:329–357, 2003.
3. S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103:427–444, 2005.
4. A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19:1171–1183, 2008.
5. J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization algorithms I*, volume 305 of *Comprehensive Study in Mathematics*. Springer-Verlag, New York, 1993.
6. S. Hoda, A. Gilpin, and J. Peña. A gradient-based approach for computing nash equilibria of large sequential games. Working paper, Tepper School of Business, Carnegie Mellon University, 2006.
7. G. Korpelevich. The extragradient method for finding saddle points and other problems. *Eknomika i Matematicheskie Metody*, 12:747–756, 1976.
8. Z. Lu, A. Nemirovski, and R. D. C. Monteiro. Large-scale semidefinite programming via saddle point mirror-prox algorithm. *Mathematical Programming*, 109:211–237, 2007.
9. 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.
10. 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.
11. Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.
12. Y. E. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110:245–259, 2006.
13. R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95:189–217, 2003.