Noname manuscript No.
(will be inserted by the editor)

Level methods uniformly optimal for composite and structured nonsmooth convex optimization

Guanghui Lan

the date of receipt and acceptance should be inserted later

Abstract The main goal of this paper is to develop uniformly optimal first-order methods for large-scale convex programming (CP). By uniform optimality we mean that the first-order methods themselves do not require the input of any problem parameters, but can still achieve the best possible iteration complexity bounds. To this end, we provide a substantial generalization of the accelerated level method by Lan [15] and demonstrate that it can uniformly achieve the optimal iteration complexity for solving a class of generalized composite CP problems, which covers a wide range of CP problems, including the nonsmooth, weakly smooth, smooth, minmax, composite and regularized problems etc. Then, we present two variants of this level method for solving a class of structured CP problems with a bilinear saddle point structure due to Nesterov [36]. We show that one of these variants can achieve the $\mathcal{O}(1/\epsilon)$ iteration complexity without requiring the input of any problem parameters. We illustrate the significant advantages of these level methods over some existing first-order methods for solving certain important classes of semidefinite programming (SDP) and two-stage stochastic programming (SP) problems.

Keywords: Convex Programming, Complexity, Level methods, Optimal methods, Semidefinite programming, Stochastic programming

1 Introduction

Consider the convex programming (CP) of

$$f^* := \min_{x \in X} f(x), \tag{1.1}$$

where X is a convex compact set and $f : X \to \mathbb{R}$ is a closed convex function. In the classic black-box setting, f is represented by a first-order oracle which, given an input point $x \in X$, returns $f(x)$ and $f'(x) \in \partial f(x)$, where $\partial f(x)$ denotes the subdifferential of f at $x \in X$.

If f is a general Lipschitz continuous convex function, then, by the classic complexity theory for CP [29], the number of calls to the first-order oracle for finding an ϵ-*solution* of (1.1) (i.e., a point $\bar{x} \in X$ s.t. $f(\bar{x}) - f^* \leq \epsilon$), cannot be smaller than $\mathcal{O}(1)(1/\epsilon^2)$ when n is sufficiently large, where $\mathcal{O}(1)$ denotes an absolute constant. This lower complexity bound can be achieved, for example, by the simple subgradient descent or mirror descent method [29]. If f is a smooth function with Lipschitz continuous gradient, Nesterov in a seminal work [32] presented an algorithm with the iteration complexity bounded by $\mathcal{O}(1)(1/\epsilon^{\frac{1}{2}})$, which, by [29], is also optimal for smooth convex optimization if n is sufficiently large. Moreover, if f is a weakly smooth function with Hölder continuous gradient, i.e., $\exists$ constants $\nu \in (0, 1)$ and $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|^\nu, \quad \forall\, x, y \in X,$$

Guanghui Lan
University of Florida, Gainesville, FL 32611, USA, E-mail: glan@ise.ufl.edu, research of this author was partially supported by NSF Grant: CMMI-1000347.

then the optimal iteration complexity bound is given by $\mathcal{O}(1)(1/\epsilon^{\frac{2}{1+3\nu}})$ (see [31,33,39]).

To accelerate the solutions of large-scale CP problems, much effort has recently been directed to exploiting the problem's structure, in order to identify possibly some new classes of CP problems with stronger convergence performance guarantee. One such example is given by the composite CP problems with the objective function given by $f(x) = \Psi(\phi(x))$. Here Ψ is a relatively simple nonsmooth convex function such as $\Psi(\cdot) = \|\cdot\|_1$ or $\Psi(\cdot) = \max\{y_1, \ldots, y_k\}$ (see Subsection 2.1 for more examples) and ϕ is an k-dimensional vector function, see [37,43,30,20,14,8]. In most of these studies, the components of ϕ are assumed to be smooth convex functions. In this case, the iteration complexity can be improved to $\mathcal{O}(1)(1/\epsilon^{\frac{1}{2}})$ by properly modifying Nesterov's optimal smooth method, see for example, [37,43,30]. Another important example is given by a class of structured CP problems due to Nesterov [36]. In its simplest form, f is given by

$$f(x) = \max_{y \in Y} \langle Ax, y \rangle,$$

where $Y \subseteq \mathbb{R}^m$ is a convex compact set and A denotes a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Nesterov shows that f can be closely approximated by a class of smooth convex functions and that the iteration complexity for solving this class of problems can be improved to $\mathcal{O}(1)(1/\epsilon)$. It is noted in [11] that this bound is unimprovable, for example, if Y is given by an Euclidean ball and the algorithm can only have access to A and A^* (the adjoin operator or A). These problems were later studied in [27,35,3,38,23,40,16] and found many interesting applications, for example, in [7,22,4].

Most approaches used to derive the aforementioned complexity bounds are (sub)gradient-type methods. In each iteration of these methods, we search for new feasible solutions of (1.1) along a given (sub)gradient (or accumulated gradient) direction with a certain stepsize. To attain the best possible theoretical complexity bounds for these algorithms, one need to specify a stepsize policy which explicitly requires the input of a number of problem parameters, such as ν, L, the operator norm $\|A\|$, or the diameter of the set X and Y (see Subsection 2). Since these parameters describe the CP problems in a global scope and hence possibly in a conservative way, this type of approaches are inherently worst-case orientated. In practice, the success of these first-order methods usually depends on how well we fine-tune the selection of the stepsizes, for example, by using certain on-line adjustments and/or a variety of heuristics.

The main goal this paper is to develop uniformly optimal methods for solving large-scale CP problems. By uniform optimality we mean that the first-order methods themselves do not require the input of any problem parameters, but can still achieve the best possible iteration complexity bounds. To this end, we focus on a different type of first-order methods, namely: the level methods. Evolving from the well-known bundle methods [12,13,18,25], level methods were first proposed by Lemaréchal, Nemirovskii and Nesterov [19] in 1995. Level methods or their certain variants [6,5] are known to exhibit the optimal iteration complexity and superior practical performance (to subgradient or mirror descent methods) for solving general non-smooth CP problems. While the original level methods were designed for solving general non-smooth CP problems, Lan [15] presented accelerated level methods which are uniformly optimal for solving, not only nonsmooth, but also smooth CP problems.

Our contribution in this paper mainly consists of the following aspects. Firstly, we present a class of generalized composite CP problems with the objective given by $f(x) = \Psi(\phi(x))$, where $\phi_i(x)$, $i \geq 1$, can be a mixture of smooth, non-smooth, weakly smooth or affine components. Such a formulation covers a wide range of CP problems, including the nonsmooth, weakly smooth, smooth, minmax, and regularized CP problems etc. (see Subsection 2.1 for more discussions). We develop the accelerated prox-level (APL) method, a substantial generalization of the method in [15], and show that it can achieve the optimal iteration complexity for solving this class of composite problems without requiring any global information on the inner functions, such as the smoothness level and the size of Lipschitz constant. Consider the black-box CP problems as a special case with $\psi(y) = y$ and $y \in \mathbb{R}$. We show that the APL method can automatically achieve the optimal iteration complexity, i.e., $\mathcal{O}(1)(1/\epsilon^{\frac{2}{1+3\nu}})$, $\nu \in [0, 1]$, for nonsmooth, weakly smooth and smooth problems, without knowing which class of problems it deals with.

Secondly, we present two smoothing level methods, namely: the basic smoothing level (BSL) and uniform smoothing level (USL) methods, for solving the aforementioned class of structured CP problems with a bilinear saddle point structure [36]. We show that both methods can find an ϵ-solution of these CP problems in at most $\mathcal{O}(1/\epsilon)$ iterations. While the BSL method still requires some information about the diameter of the set Y, the USL method is completely problem-parameter free. It is interesting to note that the smoothing level methods only differs slightly from the APL method, as they share a similar "gap reduction procedure" used to compute a sequence of lower and upper bounds converging to the optimal value f^* of (1.1).

Finally, we investigate the APL method and the smoothing level methods applied to solve certain important classes of semidefine programming (SDP) and two-stage stochastic programming (SP) problems. We demonstrate that these

algorithms can significantly outperform the existing algorithms with similar iteration complexity bounds. We also compare the performance of the three aforementioned level methods and point out the situations where one algorithm can outperform the other.

This paper is organized as follows. We introduce two classes of CP problems, namely: the generalized composite CP and the structured CP in Section 2. In Section 3, we present a generic gap reduction procedure for accelerated level methods and establish its convergence. By using this gap reduction procedure, we propose the APL method for generalized composite CP and the smoothing level methods for structured CP in Sections 4 and 5, respectively. Finally in Section 6, we present our numerical results for solving certain classes of SDP and SP problems.

1.1 Notation

A function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ is said to have $L(f)$-Lipschitz-continuous gradient with respect to $\|\cdot\|$ if it is differentiable and

$$\|f'(x) - f'(z)\|_* \leq L(f)\|x - x\| \quad \forall x, \tilde{x} \in X, \tag{1.2}$$

where $\|\cdot\|_*$ denotes the conjugate norm of $\|\cdot\|$. If $f(\cdot)$ has $L(f)$-Lipschitz-continuous gradient with respect to $\|\cdot\|$, then

$$f(x) \leq f(z) + \langle \nabla f(z), x - z\rangle + \frac{L(f)}{2}\|x - z\|^2. \tag{1.3}$$

Consider a convex compact set $X \subseteq \mathbb{R}^n$, a function $\omega : X \to \mathbb{R}$ is called a prox-function of X with modulus $\sigma > 0$, if it is differentiable and strongly convex with modulus σ, i.e.,

$$\langle \nabla\omega(x) - \nabla\omega(z), x - z\rangle \geq \sigma\|x - z\|^2, \quad \forall x, z \in X.$$

Moreover, we denote *the size of X with respect to ω* by

$$\mathcal{D}_\omega \equiv \mathcal{D}_{\omega,X} := \max_{x,z\in X} \{\omega(x) - \omega(z) - \langle \nabla\omega(z), x - z\rangle\}. \tag{1.4}$$

and the *prox-center of ω* by

$$c_\omega \equiv c_\omega(X) := \mathrm{argmin}_{x\in X}\omega(x). \tag{1.5}$$

Clearly, we have

$$\|x - z\|^2 \leq \frac{2}{\sigma}\mathcal{D}_\omega, \quad \forall x, z \in X. \tag{1.6}$$

2 The problems of interest

In this section, we present two classes of CP problems to be studied in this paper. More specifically, we introduce in Subsection 2.1 the generalized composite CP problem that covers a wide range of CP problems, such as, the non-smooth, smooth, weakly smooth, minmax, regularized CP problems. We then review in Subsection 2.2 another important class of composite CP problems possessing a special bilinear saddle point structure, which were first introduced by Nesterov [36] and later studied in [27,35,38,23,40,43].

2.1 Generalized composite CP problems

Consider the CP problem (1.1) with f given by:

$$f(x) := \Psi(\phi(x)), \tag{2.1}$$

where the outer function $\Psi : \mathbb{R}^k \to \mathbb{R}$ is Lipschitz continuous and convex, and the inner function given by $\phi(x) = (\phi_1(x), \ldots, \phi_k(x))$ is a k-dimensional vector-function with Lipschitz continuous and convex components ϕ_i, $i = 1, \ldots, k$. For the sake of notational convenience, we refer to this class of problems as problem (1.1)-(2.1). We assume that the structure of Ψ is relatively simple in comparison with ϕ (see Examples 1-4) and known to the iterative schemes for solving (2.1), while the inner functions ϕ_i, $i = 1, \ldots, k$, are represented by the black-box first-order oracles. These first-order oracles return, given an input point $x \in X$, the function values $\phi_i(x)$ and (sub)gradients of $\phi_i'(x)$. The following two additional assumptions are made about ϕ and Ψ.

Assumption 1 *$\exists\ \nu_i \in [0,1]$ and $L_i \equiv L(\phi_i) \geq 0$ such that:*

$$\|\phi_i'(x) - \phi_i'(y)\|_* \leq L_i \|x-y\|^{\nu_i}, \quad \forall\, x, y \in X, \tag{2.2}$$

where $\phi_i'(x) \in \partial\phi_i(x)$.

Observe that relation (2.2) holds with $\nu_i = 1$ for a smooth convex component ϕ_i and with $\nu_i = 0$ for a non-smooth convex component ϕ_i. If ϕ_i is a differentiable convex function with Hölder continuous gradient (c.f. [31,33,39]), then (2.2) holds for some $0 < \nu_i < 1$. The latter class of functions will be referred to as weakly smooth functions in this paper.

Clearly, if $L_i = 0$ for some $1 \leq i \leq k$, then the component ϕ_i must be affine. Otherwise, ϕ_i must be nonlinear. To fix the notation, let us assume throughout the paper that, for a given $1 \leq k_0 \leq k$, the first k_0 components of ϕ are nonlinear, i.e., $L_i > 0$ for any $1 \leq i \leq k_0$, while the remaining $k - k_0$ components are affine, i.e., $L_i = 0$ for any $k_0 + 1 \leq i \leq k$. We make the following assumption regarding the monotonicity of Ψ with respect to these nonlinear components.

Assumption 2 *The map*

$$y_i \mapsto \Psi(y_1, \cdots, y_i, \cdots, y_k)$$

is monotonically nondecreasing for any $1 \leq i \leq k_0$.

In view of Assumption 2, $\Psi(\cdot)$ is monotonically nondecreasing over the projection of $\mathbb{R}^k$ onto $\mathbb{R}^{k_0}$ obtained by fixing $y_{k_0+1}, \cdots, y_k$.

Many CP problems can be written in the form of problem (1.1)-(2.1). We give a few interesting examples as follows.

Example 1 **Nonsmooth, weakly smooth and smooth problems.** Let $k = 1$ and $\Psi(y) = y$. Then, problem (1.1)-(2.1) covers the usual nonsmooth, weakly smooth and smooth CP problems, for which condition (2.2) is satisfied with $\nu = 0$, $\nu \in (0,1)$ and $\nu = 1$, respectively.

Example 2 **Minmax problems.** Let $\Psi(y) = \max\{y_1, \ldots, y_k\}$. With this outer function, problem (1.1)-(2.1) becomes the minmax problem to minimize the maximum of a finite number of convex functions. It can be used to, for example, solve a system of smooth convex inequalities $\phi_i(x) \leq 0$, $i = 1, \ldots, k$, where $\phi_i(x)$ are convex functions satisfying (2.2) with $\nu_1 = \ldots = \nu_k = 1$. It can also be used to solve a system of mixed smooth and nonsmooth convex inequalities if $\nu_i = 0$ or 1, $i = 1, \ldots, k$.

Example 3 **Regularized problems.** Consider the problem of $\min_{x\in X} \phi_1(x) + \rho\|x\|_*$, where ϕ_1 is a smooth convex function with Lipschitz continuous gradient and $\|x\|_*$ is is a continuous, nonnegative, usually nonsmooth convex function. We can put this problem in the form of (1.1)-(2.1) by setting $\phi(x) = (\phi_1(x), x) \in \mathbb{R}^{n+1}$ and $\Psi(y, x) = y + \rho\|x\|_*$. Note that if $\|x\|_* = \|x\|_1$ and $\phi_1(x) = \|Ax - b\|^2$, this problem becomes the well-known l_1 sparse optimization problem.

Example 4 **Composite smooth and nonsmooth problems.** Consider $\min_{x\in X} \psi(x) = \phi_1(x) + \phi_2(x)$, where ϕ_1 is a smooth component and ϕ_2 is a nonsmooth component. Clearly, we can write the problem in the form of (2.1) by setting $\phi(x) = (\phi_1(x), \phi_2(x))$ and $\Psi(y_1, y_2) = y_1 + y_2$. For this problem, we have $\nu_1 = 1$ and $\nu_2 = 0$. The applications can be found, for example, in certain penalization approaches for solving nonsmooth CP problems [14].

Since problem (1.1)-(2.1) covers nonsmooth, weakly smooth and smooth CP as certain special cases and k_0 is a given constant, in view of [29,31,33,39], a lower bound on the iteration complexity for solving this class of generalized composite problems is given by

$$\max_{i=1,\ldots,k_0} \left(\frac{L_i}{\epsilon}\right)^{\frac{2}{1+3\nu_i}}. \tag{2.3}$$

The composite CP problem presented in this subsection generalizes a few other composite CP problems existing in the literature (see, for example, Nesterov [34,37], Tseng [43], Lewis and Wright [20], and Nemirovski [30]). More specifically, Nesterov [37] and Tseng [43] studied composite problems similar to those in Example 3. In the composite problems studied by Lewis and Wright [20] (see also [42]), the inner functions ϕ_i, $i = 1, \ldots, k$, are smooth, but Ψ and ϕ_i are not necessarily convex. A little earlier, Nesterov [34] discussed an optimal method for solving the minmax

problem with smooth components (see Example 2) and Nemirovski [30] also presented a class of composite problems of the form (2.1) but with more restrictive assumptions. In particular, in the composite CP problems studied in [30], ϕ_i, $i = 1, \ldots, k$, are smooth convex functions and Ψ is monotone over its entire domain. As a result, this class of composite CP problems does cover the nonsmooth, weakly smooth and regularized problems.

In Section 4, we will present an accelerated prox-level method uniformly optimal for solving (2.1) without requiring any global information about the inner functions ϕ_i, such as, the smoothness levels ν_i and the Lipschitz constants L_i, for all $i = 1, \ldots, k_0$.

2.2 Structured non-smooth CP problems

In the subsection, we discuss a class of composite CP problem where the inner function contains a certain bilinear saddle point structure (c.f. [36,35]). Consider problem (1.1) with f given by

$$f(x) := \hat{f}(x) + F(x), \tag{2.4}$$

where $\hat{f} : X \to \mathbb{R}$ is a simple Lipschitz continuous convex function with constant $M_{\hat{f}}$ and

$$F(x) := \max_{y \in Y} \{\langle Ax, y\rangle - \hat{g}(y)\}. \tag{2.5}$$

Here, $Y \subseteq \mathbb{R}^m$ is a compact convex set, $\hat{g} : Y \to \mathbb{R}$ is a continuous convex function on Y and A denotes a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Clearly, (2.4) can be viewed as a special case of problem (1.1)-(2.1) with $\Psi(y_1, y_2) = y_1 + y_2$ and $\phi_1(x) = \hat{f}(x)$ and $\phi_2(x) = F(x)$. However, now the inner functions are not represented by black-box oracles anymore and their structures can be used explicitly in the iterative schemes (c.f. [27]). Observe also that problem (1.1)-(2.4) can be written in an adjoint form:

$$\max_{y \in Y} \{g(y) := -\hat{g}(y) + G(y)\}, \quad G(y) := \min_{x \in X} \left\{\langle Ax, y\rangle + \hat{f}(x)\right\}. \tag{2.6}$$

While the function F given by (2.5) is a non-smooth convex function in general (F is smooth, for example, if $\hat{g}$ is strongly convex), Nesterov in an important work [36] shows that it can be closely approximated by a class of smooth convex functions. We now briefly describe Nesterov's smoothing scheme as follows. Let $v(y)$ be a prox-function of Y with modulus σ_v and prox-center c_v (see (1.5)). Also let us denote

$$V(y) := v(y) - v(c_v) - \langle \nabla v(c_v), y - c_v\rangle,$$

and, for some $\eta > 0$,

$$F_\eta(x) := \max_y \left\{\langle Ax, y\rangle - \hat{\phi}(y) - \eta\, V(y) : \ y \in Y\right\}, \tag{2.7}$$

$$f_\eta(x) := \hat{f}(x) + F_\eta(x). \tag{2.8}$$

It is shown in [36] that $F_\eta(\cdot)$ has $\mathcal{L}_\eta$-Lipschitz-continuous gradient, where

$$\mathcal{L}_\eta \equiv \mathcal{L}(F_\eta) := \frac{\|A\|^2}{\eta \sigma_v} \tag{2.9}$$

and $\|A\|$ denote the operator norm of A. Moreover, the "closeness" of $F_\eta(\cdot)$ to $F(\cdot)$ depends linearly on the parameter η. In particular, we have, for every $x \in X$,

$$F_\eta(x) \le F(x) \le F_\eta(x) + \eta\, \mathcal{D}_v, \tag{2.10}$$

and, as a consequence,

$$f_\eta(x) \le f(x) \le f_\eta(x) + \eta\, \mathcal{D}_v, \tag{2.11}$$

where $\mathcal{D}_v$ given by (1.4).

Nesterov shows [36] that one can obtain an ϵ-solution of problem (1.1)-(2.4) in at most $\mathcal{O}(1/\epsilon)$ iterations, by applying a variant of his optimal smooth method [32,36] to $\min_{x \in X} f_\eta(x)$, for a properly chosen $\eta > 0$. This result is

significantly better than the iteration-complexity for the black-box non-smooth convex optimization techniques applied to (2.4).

To implement Nesterov's approximation scheme, it is necessary to know a number of problem parameters a priori, including $\|A\|$, σ_v and $\mathcal{D}_v$, and the total number of iterations N. To eliminate the requirement that N should be given in advance, Nesterov in [35] presented an excessive gap procedure where the above smoothing technique is applied to both the primal and dual problem (2.6). However, to apply the excessive gap procedure, one need to know a few more parameters, including $\|A\|$, σ_v, $\mathcal{D}_v$, σ_ω and $\mathcal{D}_\omega$, where σ_ω is the modulus of a given prox-function ω of X and $\mathcal{D}_\omega$ is defined in (1.4). In [27], Nemirovski proposed a prox-method with $\mathcal{O}(1/\epsilon)$ iteration-complexity bound for solving a slightly more general class of CP problems than (2.4). To attain the best possible iteration complexity in [27], it is still necessary to know the parameters $\|A\|$, σ_v, $\mathcal{D}_v$, σ_ω and $\mathcal{D}_\omega$ explicitly.

In Section 5, we will present two new smoothing methods for solving problem (1.1)-(2.4) and show that they can achieve the same iteration-complexity bounds as those in [36,35,27]. In particular, we demonstrate that one of these smoothing methods does not require any of those problem parameters mentioned above.

3 A gap reduction procedure for accelerated level methods

In this section, we consider the generic CP problem (1.1), and show how to compute a sequence of lower and upper bounds on f^* and to reduce the gap between these lower and upper bounds. This gap reduction procedure will be used later for solving the generalized composite CP and structured CP in Sections 4 and 5, respectively.

This section is organized as follows. We first present a few important ingredients, such as, the minorant functions, level sets and localizers, used for computing the lower bounds on f^* in Subsection 3.1. We then introduce in Subsections 3.2 and 3.3, respectively, the clustered subsets and the accelerated level descent steps for generating a sequence of feasible solutions of (1.1) and hence upper bounds on f^*. Finally, we describe the gap reduction procedure for accelerated level methods and discuss its convergence properties in Subsection 3.4. Our development substantially generalizes the techniques presented in [6,5,15] for solving the black-box CP problems.

3.1 Minorants, level sets and localizers

It is well-known that the linear function $h_f(z,\cdot)$ given by

$$h_f(z,x) := f(z) + \langle f'(z), x - z\rangle.$$

is a support hyperplane of f, where $f'(z) \in \partial f(z)$. Moreover, if f is a smooth convex function with $L(f)$-Lipschitz continuous gradient, then

$$0 \le f(x) - h_f(z,x) \le \frac{L(f)}{2}\|x - z\|^2, \quad \forall\, x \in X.$$

Motivated by the support hyperplane of a smooth convex function, we define the notion of minorant and M-smooth support function as follows.

Definition 1 A function $h : X \to \mathbb{R}$ is called a (convex) minorant of f, if h is convex and $h(x) \le f(x)$ for any $x \in X$. If, in addition, $h(z) = f(z)$ for some $z \in X$, then it is called a support function of f at z, denoted by $h(z,\cdot)$. Moreover, if

$$0 \le f(x) - h(z,x) \le \frac{M}{2}\|x - z\|^2, \quad \forall\, x \in X, \tag{3.1}$$

then $h(z,\cdot)$ is called a M-smooth support function of f at z.

Note that in view of the above definition, a minorant of f is not necessarily affine with respect to x.

If f assumes a lower level of smoothness, i.e., $\exists\ \nu \in [0,1)$ such that $\|f'(x) - f'(z)\|_* \le L(f)\|x - z\|^\nu$, then

$$0 \le f(x) - h_f(x,z) \le \frac{L(f)}{1+\nu}\|x - z\|^{1+\nu} \quad \forall\, x \in X. \tag{3.2}$$

One can provide a generalization of the support hyperplane of a weakly smooth or nonsmooth convex function. Our generalization below considers also the the case when f is given as a certain composition (say, summation) of a few components with different smoothness levels.

Definition 2 Let $\rho_i \in [0,1]$ and $M_i > 0$, $i = 1, \ldots, m$, be given. A support function $h(z,\cdot)$ of f at z has smoothness level (ρ_i, M_i), $i = 1, \ldots, m$, if

$$0 \le f(x) - h(z,x) \le \sum_{i=1}^{m} \frac{M_i}{1+\rho_i} \|x - z\|^{1+\rho_i}, \quad \forall\, x \in X. \tag{3.3}$$

We can compute a valid lower bound on f^* by using the minorant and the level set of f. More specifically, let $h(\cdot)$ be a minorant of f and $\mathcal{E}_f(l)$ denote the level set of f given by

$$\mathcal{E}_f(l) := \{x \in X : f(x) \le l\}. \tag{3.4}$$

Also let us denote

$$\bar{h} := \min\left\{h(x) : x \in \mathcal{E}_f(l)\right\}. \tag{3.5}$$

Then, we have

$$f(x) \ge \min\{l, \bar{h}\}, \quad \forall x \in X. \tag{3.6}$$

Indeed, by (3.4), we have $f(x) \ge l$ for any $x \in X \backslash \mathcal{E}_f(l)$. Moreover, by (3.5) and Definition 1, we have $\bar{h} \le h(x) \le f(x)$ for any $x \in \mathcal{E}_f(l)$.

Note, however, that to solve problem (3.5) is usually as difficult as to solve the original problem (1.1). To compute a convenient lower bound of f^*, one has to resort to certain relaxations of (3.5). For this purpose, we replace $\mathcal{E}_f(l)$ in (3.5) by a convex and compact set $\bar{\mathcal{E}}_f(l)$ satisfying

$$\mathcal{E}_f(l) \subseteq \bar{\mathcal{E}}_f(l) \subseteq X. \tag{3.7}$$

The set $\bar{\mathcal{E}}_f(l)$ will be referred to as a *localizer of the level set* $\mathcal{E}_f(l)$ in the remaining part of this paper. The following result shows the computation of a lower bound on f^* by solving a relaxation of (3.5).

Lemma 1 *Let $\bar{\mathcal{E}}_f(l)$ be a localizer of the level set $\mathcal{E}_f(l)$ for some $l \in \mathbb{R}$ and $h(\cdot)$ be a minorant of f. Denote*

$$h^* := \min\left\{h(x) : x \in \bar{\mathcal{E}}_f(l)\right\}. \tag{3.8}$$

We have

$$f(x) \ge \min\{l, h^*\}, \quad \forall x \in X. \tag{3.9}$$

Proof. Note that if the feasible set $\bar{\mathcal{E}}_f(l)$ in (3.8) is empty, then $h^* = +\infty$. In this case, we have $f(x) \ge l$ for any $x \in X$. Now assume that $\bar{\mathcal{E}}_f(l) \ne \emptyset$. By (3.5), (3.7) and (3.8), we have $h^* \le \bar{h}$, which together with (3.6), then clearly imply (3.9). ∎

3.2 The clustered subsets

In this subsection, we introduce the notion of a sequence of clustered convex subsets. We will use such convex subsets when generating a sequence of feasible solutions of (1.1).

Definition 3 Let ω be a given prox-function of X with modulus σ_ω, a sequence of nonempty convex subsets $X_t \subseteq X$, $t = 1, 2, \ldots$, are clustered with respect to ω, if,

$$c_\omega(X_t) \in X_{t-1}, \quad \forall\, t = 2, 3, \ldots, \tag{3.10}$$

where $c_\omega(X)$ is defined in (1.5).

In other words, in a sequence of clustered subsets X_t, $t = 1, \ldots$, the prox-center of the current subset X_t always falls into the previous subset X_{t-1}. In fact, as shown by the following result, the squared distances between the prox-centers of any two subsequent subsets are actually summable.

Lemma 2 *Let ω be a given prox-function of X with modulus σ_ω. If a sequence of nonempty convex subsets $X_t \subseteq X$, $t = 1, 2, \ldots$, are clustered with respect to ω, then*

$$\sum_{\tau=2}^{t} \|c_\tau - c_{\tau-1}\|^2 \leq \frac{2}{\sigma_\omega} [\omega(c_t) - \omega(c_1)], \quad t = 2, 3, \ldots, \tag{3.11}$$

where $c_t := c_\omega(X_t)$ and $c_\omega(\cdot)$ is defined in (1.5). As a consequence, denoting $c_0 := c_\omega(X)$, we have

$$\sum_{\tau=1}^{t} \|c_\tau - c_{\tau-1}\|^2 \leq \frac{2}{\sigma_\omega} [\omega(c_t) - \omega(c_0)], \quad t = 1, 2, \ldots. \tag{3.12}$$

Proof. By the optimality condition of (1.5), we have

$$\langle \nabla\omega(c_{t-1}), x - c_{t-1} \rangle \geq 0, \quad \forall x \in X_{t-1},$$

for any $t \geq 2$, which, in view of (3.10), implies that $\langle \nabla\omega(c_{t-1}), c_t - c_{t-1} \rangle \geq 0$ for any $t \geq 2$. Using the previous conclusion and the strong convexity of ω, we have

$$\frac{\sigma_\omega}{2}\|c_t - c_{t-1}\|^2 \leq \omega(c_t) - \omega(c_{t-1}) - \langle \nabla\omega(c_{t-1}), c_t - c_{t-1} \rangle \leq \omega(c_t) - \omega(c_{t-1}), \quad t \geq 2.$$

Summing up the above inequalities, we arrive at (3.11). Now, relation (3.12) immediately follows from (3.11) and the observation that the sequence of subsets, $X, X_1, X_2, \ldots$, is also clustered with respect to ω. ∎

3.3 Accelerated level descent steps

For a given initial point $x_0^u \in X$ and a level $l < f(x_0^u)$, our goal in this subsection is to present the accelerated level descent steps to reduce the gap given by $f(x_0^u) - l$. Here, the superscript u in x_0^u indicates that this point will be used to generate an upper bound on f^*.

The following simple result shows that one can potentially reduce the aforementioned gap by making use of some properties of the support functions.

Lemma 3 *Let $(x_{t-1}, x_{t-1}^u) \in X \times X$ be given at the t-th iteration, $t \geq 1$, of an iterative scheme. Also let $h(z, \cdot)$ denote a support function of $f(\cdot)$ at z and suppose that the pair of new search points $(x_t, x_t^u) \in X \times X$ satisfy that, for some $l \in \mathbb{R}$ and $\alpha_t \in (0, 1]$,*

$$h(\alpha_t x_{t-1} + (1 - \alpha_t)x_{t-1}^u, x_t) \leq l, \tag{3.13}$$

$$x_t^u = \alpha_t x_t + (1 - \alpha_t)x_{t-1}^u. \tag{3.14}$$

Then, if $h(z, \cdot)$ is a support function of f at z with smoothness level (ρ_i, M_i), $i = 1, \ldots, m$, we have

$$f(x_t^u) - l \leq (1 - \alpha_t)[f(x_{t-1}^u) - l] + \sum_{i=1}^{m} \frac{M_i \alpha_t^{1+\rho_i}}{1 + \rho_i}\|x_t - x_{t-1}\|^{1+\rho_i}. \tag{3.15}$$

As a consequence, if $h(z, \cdot)$ is an M-smooth support function of f at z, then

$$f(x_t^u) - l \leq (1 - \alpha_t)[f(x_{t-1}^u) - l] + \frac{M\alpha_t^2}{2}\|x_t - x_{t-1}\|^2. \tag{3.16}$$

Proof. Denote $z_t = \alpha_t x_{t-1} + (1-\alpha_t)x_{t-1}^u$. It can be easily seen from (3.14) that $x_t^u - z_t = \alpha_t(x_t - x_{t-1})$. Using this observation, (3.3), (3.14), and the convexity of f and $h(z_t,\cdot)$, we have

$$\begin{aligned} f(x_t^u) &\le h(z_t, x_t^u) + \sum_{i=1}^m \frac{M_i}{1+\rho_i}\|x_t^u - z_t\|^{1+\rho_i} = h(z_t, x_t^u) + \sum_{i=1}^m \frac{M_i\alpha_t^{1+\rho_i}}{1+\rho_i}\|x_t - x_{t-1}\|^{1+\rho_i} \\ &\le (1-\alpha_t)h(z_t, x_{t-1}^u) + \alpha_t h(z_t, x_t) + \sum_{i=1}^m \frac{M_i\alpha_t^{1+\rho_i}}{1+\rho_i}\|x_t - x_{t-1}\|^{1+\rho_i} \\ &\le (1-\alpha_t)f(x_{t-1}^u) + \alpha_t l + \sum_{i=1}^m \frac{M_i\alpha_t^{1+\rho_i}}{1+\rho_i}\|x_t - x_{t-1}\|^{1+\rho_i}. \end{aligned}$$

Subtracting l from both sides of the above inequality, we obtain (3.15). ∎

In view of Lemma 3, if the distance $\|x_t - x_{t-1}\|$ is small enough and α_t is properly chosen, then $f(x_t^u) - l$ can become smaller than $f(x_{t-1}^u) - l$.

In order to control the distance $\|x_t - x_{t-1}\|$, $t \ge 1$, we assume in the sequel that there exist a sequence of convex subsets $X_t \subseteq X$, $t = 1, \ldots$, such that

$$x_t = c_\omega(X_t),\ t = 1, 2, \ldots, \quad \text{and} \quad x_t \in X_{t-1},\ t = 2, 3, \ldots, \tag{3.17}$$

where c_ω is defined in (1.5). In other words, x_t, $t \ge 1$, are the prox-centers of a sequence of clustered subsets with respect to a given prox-function ω. Moreover, denoting

$$\gamma_t := \begin{cases} 1, & t = 1, \\ \gamma_{t-1}(1-\alpha_t), & t \ge 2, \end{cases} \tag{3.18}$$

and

$$\Gamma_t(\rho) := \left(\gamma_1^{-1}\alpha_1^{1+\rho}, \ldots, \gamma_t^{-1}\alpha_t^{1+\rho}\right),\ \forall\, \rho \in [0,1], \tag{3.19}$$

we assume that $\alpha_t \in (0,1]$, $t \ge 1$, are chosen such that

$$\alpha_1 = 1 \ \text{ and } \ \gamma_t \|\Gamma_t(\rho)\|_{\frac{2}{1-\rho}} \le \mathcal{C}_\rho\, t^{-\frac{1+3\rho}{2}}, \forall\, \rho \in [0,1], \tag{3.20}$$

where $\mathcal{C}_\rho : [0,1] \to \mathbb{R}^+$ depends only on ρ.

Lemma 4 below states certain explicit rules for specifying α_t, $t \ge 1$, so that (3.20) holds. If the conditions (3.13), (3.14), (3.17) and (3.20) are satisfied, then we call the iteration from (x_{t-1}, x_{t-1}^u) to (x_t, x_t^u) an *accelerated level descent step*.

Lemma 4 *a) If α_t, $t = 1, 2, \ldots$, are set to*

$$\alpha_t = \frac{2}{t+1}, \tag{3.21}$$

then condition (3.20) holds with $\mathcal{C}_\rho = 2^{1+\rho}\, 3^{-\frac{1-\rho}{2}}$;

b) if α_t, $t = 1, 2, \ldots$, are computed recursively by

$$\alpha_1 = \gamma_1 = 1, \quad \alpha_t^2 = (1-\alpha_t)\gamma_{t-1} = \gamma_t, \tag{3.22}$$

then we have $\alpha_t \in (0,1]$ for any $t \ge 2$. Moreover, condition (3.20) holds with $\mathcal{C}_\rho = 4 \cdot 3^{-\frac{1-\rho}{2}}$.

Proof. We first show part a). Clearly, by (3.18) and (3.21), we have

$$\gamma_t = \frac{2}{t(t+1)} \ \text{ and } \ \gamma_t^{-1}\alpha_t^{1+\rho} = \left(\frac{2}{t+1}\right)^\rho t \le 2^\rho t^{1-\rho}. \tag{3.23}$$

If $\rho = 1$, it can be easily seen from (3.19) and (3.23) that $\|\Gamma_t(1)\|_\infty \leq 2$ and hence that $\gamma_t\|\Gamma_t(1)\|_\infty \leq 4t^{-2}$. Now suppose that $\rho \in [0,1)$. Using (3.19), (3.23) and the simple observation that $\sum_{\tau=1}^t \tau^2 = t(t+1)(2t+1)/6 \leq t(t+1)^2/3$, we have

$$\begin{aligned}\gamma_t\|\Gamma_t(\rho)\|_{\frac{2}{1-\rho}} &\leq \gamma_t\left[\sum_{\tau=1}^t \left(2^\rho\tau^{1-\rho}\right)^{\frac{2}{1-\rho}}\right]^{\frac{1-\rho}{2}} = 2^\rho\gamma_t\left(\sum_{\tau=1}^t \tau^2\right)^{\frac{1-\rho}{2}} \leq 2^\rho\gamma_t\left[\frac{t(t+1)^2}{3}\right]^{\frac{1-\rho}{2}} \\ &= \left(2^{1+\rho}\,3^{-\frac{1-\rho}{2}}\right)\left[t^{-\frac{1+\rho}{2}}(t+1)^{-\rho}\right] \leq \left(2^{1+\rho}\,3^{-\frac{1-\rho}{2}}\right)t^{-\frac{1+3\rho}{2}}.\end{aligned}$$

We now show that part b) holds. Note that by (3.22), we have $\alpha_t = \left(-\gamma_{t-1} + \sqrt{\gamma_{t-1}^2 + 4\gamma_{t-1}}\right)/2$, $t \geq 2$, which clearly implies that $\alpha_t > 0$ for any $t \geq 2$. It can also be easily seen from the previous observation that $\alpha_t \leq 1$ and $\gamma_t \leq 1$ by using induction. Now let us bound $1/\sqrt{\gamma_t}$ for any $t \geq 2$. First observe that, by (3.22), we have, for any $t \geq 2$,

$$\frac{1}{\sqrt{\gamma_t}} - \frac{1}{\sqrt{\gamma_{t-1}}} = \frac{\sqrt{\gamma_{t-1}} - \sqrt{\gamma_t}}{\sqrt{\gamma_{t-1}\gamma_t}} = \frac{\gamma_{t-1} - \gamma_t}{\sqrt{\gamma_{t-1}\gamma_t}\left(\sqrt{\gamma_{t-1}} + \sqrt{\gamma_t}\right)} = \frac{\alpha_t\gamma_{t-1}}{\gamma_{t-1}\sqrt{\gamma_t} + \gamma_t\sqrt{\gamma_{t-1}}}.$$

Using the above identity, (3.22) and the fact that $\gamma_t \leq \gamma_{t-1}$ due to (3.22), we conclude that

$$\frac{1}{\sqrt{\gamma_t}} - \frac{1}{\sqrt{\gamma_{t-1}}} \geq \frac{\alpha_t}{2\sqrt{\gamma_t}} = \frac{1}{2} \quad \text{and} \quad \frac{1}{\sqrt{\gamma_t}} - \frac{1}{\sqrt{\gamma_{t-1}}} \leq \frac{\alpha_t}{\sqrt{\gamma_t}} = 1,$$

which, in view of the fact that $\gamma_1 = 1$, then implies that $(t+1)/2 \leq 1/\sqrt{\gamma_t} \leq t$. Using the previous inequality and (3.22), we conclude that

$$\gamma_t \leq \frac{4}{(t+1)^2} \quad \text{and} \quad \gamma_t^{-1}\alpha_t^{1+\rho} = \left(\frac{1}{\sqrt{\gamma_t}}\right)^{1-\rho} \leq t^{1-\rho}.$$

Noting that the above relation is similar to (3.23), we can complete the proof of part b) similarly to that of part a). ∎

Proposition 1 below states some convergence properties for the accelerated level descent steps.

Proposition 1 *Let an initial point $x_0^u \in X$ be given. Also let us denote $x_0 = c_\omega(X)$ for a given prox-function ω of X with modulus σ_ω, where $c_\omega(X)$ is defined in (1.5). Suppose that $(x_t, x_t^u) \in X \times X$, $t \geq 1$, are generated such that relations (3.13), (3.14) and (3.17) hold and that the stepsizes α_t, $t \geq 1$, are chosen such that (3.20) holds. Then, if $h(z_t,\cdot)$, $t \geq 1$, are support functions of f with smoothness level (ρ_i, M_i), $i = 1,\ldots,m$, we have*

$$f(x_t^u) - l \leq \sum_{i=1}^m \left\{\frac{\mathcal{C}_{\rho_i}M_i}{(1+\rho_i)\,t^{\frac{1+3\rho_i}{2}}}\left[\frac{2}{\sigma_\omega}\left(\omega(x_t) - \omega(x_0)\right)\right]^{\frac{1+\rho_i}{2}}\right\}, \quad \forall\, t \geq 1. \tag{3.24}$$

As a consequence, if $h(z_t,\cdot)$, $t \geq 1$, are M-smooth support functions of f, then,

$$f(x_t^u) - l \leq \frac{\mathcal{C}_1 M}{\sigma_\omega t^2}\left[\omega(x_t) - \omega(x_0)\right], \quad \forall\, t \geq 1. \tag{3.25}$$

Proof. Letting γ_t be defined in (3.18) and dividing both sides of (3.15) by γ_t, we obtain

$$\begin{aligned}\frac{1}{\gamma_1}[f(x_1^u) - l] &\leq (1-\alpha_1)[f(x_0^u) - l] + \sum_{i=1}^m \frac{M_i\alpha_1^{1+\rho_i}}{(1+\rho_i)\gamma_1}\|x_1 - x_0\|^{1+\rho_i}, \\ \frac{1}{\gamma_t}[f(y_t) - l] &\leq \frac{1}{\gamma_{t-1}}[f(y_{t-1}) - l] + \sum_{i=1}^m \frac{M_i\alpha_t^{1+\rho_i}}{(1+\rho_i)\gamma_t}\|x_t - x_{t-1}\|^{1+\rho_i},\ t \geq 2.\end{aligned}$$

Adding up these inequalities, we have

$$\frac{1}{\gamma_t}[f(x_t^u) - l] \leq (1-\alpha_1)[f(x_0^u) - l] + \sum_{i=1}^m \left[\frac{M_i}{1+\rho_i}\left(\sum_{\tau=1}^t \frac{\alpha_\tau^{1+\rho_i}}{\gamma_\tau}\|x_t - x_{t-1}\|^{1+\rho_i}\right)\right].$$

Also observing that, by (3.19) and Hölder's inequality, we have

$$\sum_{\tau=1}^{t} \frac{\alpha_\tau^{1+\rho_i}}{\gamma_\tau}\|x_t - x_{t-1}\|^{1+\rho_i} \le \|\Gamma_t(\rho_i)\|_{\frac{2}{1-\rho_i}} \left(\sum_{\tau=1}^{t}\|x_\tau - x_{\tau-1}\|^2\right)^{\frac{1+\rho_i}{2}}$$
$$\le \|\Gamma_t(\rho_i)\|_{\frac{2}{1-\rho_i}} \left[\frac{2}{\sigma_\omega}(\omega(x_t)-\omega(x_0))\right]^{\frac{1+\rho_i}{2}},$$

where the last inequality follows from (3.12). Combining the above two inequalities and using (3.20), we obtain

$$f(x_t^u) - l \le \gamma_t(1-\alpha_1)\left[f(x_0^u)-l\right] + \sum_{i=1}^{m}\left\{\frac{M_i}{(1+\rho_i)}\gamma_t\|\Gamma_t(\rho_i)\|_{\frac{2}{1-\rho_i}}\left[\frac{2}{\sigma_\omega}(\omega(x_t)-\omega(x_0))\right]^{\frac{1+\rho_i}{2}}\right\}$$
$$\le \sum_{i=1}^{m}\left\{\frac{\mathcal{C}_{\rho_i}M_i}{(1+\rho_i)\,t^{\frac{1+3\rho_i}{2}}}\left[\frac{2}{\sigma_\omega}(\omega(x_t)-\omega(x_0))\right]^{\frac{1+\rho_i}{2}}\right\}.$$

■

It should be noted, however, that the accelerated level descent steps presented in this subsection are essentially conceptual only. Firstly, without properly specifying the value of l, we are not sure if it is feasible to generate a point $x_t \in X$ satisfying (3.13). Secondly, we have not specified how to choose a sequence of clustered subsets X_t, $t = 0, 1, \ldots$, such that (3.17) hold. We will address these issues in the next subsection.

3.4 The gap reduction procedure and its convergence properties

Given an initial point $x_0^u \in X$ and an initial lower bound $\underline{f}_0$ on f^*, we will present in the subsection an iterative scheme to reduce the optimality gap $f(x_0^u) - \underline{f}_0$. More specifically, we will employ the accelerated level descent steps introduced in Subsection 3.3 with some properly chosen l and X_t, $t \ge 1$.

We first describe this procedure as follows.

The APL gap reduction procedure:

Input: initial point $x_0^u \in X$, lower bound $\underline{f}_0$, prox-function ω, and minorant $h(\cdot,\cdot)$.

Output: solution $\bar{x}^u \in X$ and lower bound $\underline{f}^+$.

Initialize: Set $\overline{f}_0 = f(x_0^u)$, $l = \theta\underline{f}_0 + (1-\theta)\overline{f}_0$, $\bar{x}^u = x_0^u$, and $x_0 = c_\omega(X)$, where $c_\omega(X)$ is defined in (1.5) and $\theta \in (0,1)$ is a user-defined parameter, say $\theta = 0.5$. Choose an initial localizer X_0 of the level set $\mathcal{E}_f(l)$, say $X_0 = X$. Also let $t = 1$;

1) Set $z_t = \alpha_t x_{t-1} + (1-\alpha_t)x_{t-1}^u$ and compute

$$h_t^* := \min\{h(z_t, x) : x \in X_{t-1}\} \quad \text{and} \quad \underline{f}_t = \max\left\{\underline{f}_{t-1}, \min\{l, h_t^*\}\right\}; \tag{3.26}$$

2) If $\underline{f}_t \ge l - \beta(l - \underline{f}_0)$, then go to Step 6, where $\beta \in (0,1)$ is a user-defined parameter (say $\beta = 0.5$);

3) Compute

$$x_t := \operatorname{argmin}_x\{\omega(x) : x \in X_{t-1}, h(z_t, x) \le l\}, \tag{3.27}$$
$$x_t^u := \alpha_t x_t + (1-\alpha_t)x_{t-1}^u. \tag{3.28}$$

Set $\overline{f}_t = \min\{\overline{f}_{t-1}, f(x_t^u)\}$ and $\bar{x}^u$ to be the best solution with $f(\bar{x}^u) = \overline{f}_t$;

4) If $\overline{f}_t \le l + \beta(\overline{f}_0 - l)$, then go to Step 6;

5) Choose a convex compact set X_t such that $\underline{X}_t \subseteq X_t \subseteq \overline{X}_t$, where

$$\underline{X}_t := \{x \in X_{t-1} : h(z_t, x) \le l\} \tag{3.29}$$
$$\overline{X}_t := \{x \in X : \langle \nabla\omega(x_t), x - x_t\rangle \ge 0\}. \tag{3.30}$$

Set $t = t + 1$ and go to Step 1;

6) Report **success** and terminate the procedure with output $\underline{f}^+ = \underline{f}_t$ and $\bar{x}^u$.

We make some remarks about the above gap reduction procedure. Firstly, whenever t increases by 1, we say that an iteration of the gap reduction procedure occurs. Secondly, note that x_0 is initialized as the prox-center of X, but not necessarily the prox-center of the X_0, which gives us some flexibility in choosing the initial localizer X_0.

We summarize in Lemma 5 a few more observations regarding the execution of the above gap reduction procedure.

Lemma 5 *The following statements hold for the above gap reduction procedure:*

a) $\{X_t\}_{t\geq 1}$ *is a sequence of clustered localizers of the level set* $\mathcal{E}_f(l)$ *satisfying (3.17);*
b) $\underline{f}_0 \leq \underline{f}_1 \leq \ldots \leq \underline{f}_t \leq f^*$ *and* $\overline{f}_0 \geq \overline{f}_1 \geq \ldots \geq \overline{f}_t \geq f^*$ *for any* $t \geq 1$*;*
c) Problem (3.27) is always feasible unless the procedure terminates;
d) $\emptyset \neq \underline{X}_t \subseteq \overline{X}_t$ *for any* $t \geq 1$ *and hence Step 5 is always feasible unless the procedure terminates.*

Proof. We first show part a). Firstly, noting that $\mathcal{E}_f(l) \subseteq X_0$, we can show that $\mathcal{E}_f(l) \subseteq X_t$, $t \geq 1$, by using induction. Suppose that X_{t-1} is a localizer of the level set $\mathcal{E}_f(l)$. Then, for any $x \in \mathcal{E}_f(l)$, we have $x \in X_{t-1}$. Moreover, by the definitions of the minorants and level sets, we have $h(z_t, x) \leq f(x) \leq l$ for any $x \in \mathcal{E}_f(l)$. Using these two observations and the definition of $\underline{X}_t$ in (3.29), we have $\mathcal{E}_f(l) \subseteq \underline{X}_t$, which, in view of the fact that $\underline{X}_t \subseteq X_t$, implies $\mathcal{E}_f(l) \subseteq X_t$, i.e., X_t is a localizer of $\mathcal{E}_f(l)$. Secondly, by (3.27), we have $x_t \in X_{t-1}$, $t \geq 1$. Moreover, by (3.30), we have $\langle \nabla\omega(x_t), x - x_t\rangle \geq 0$ for any $x \in \overline{X}_t$, which, together with the fact that $X_t \subseteq \overline{X}_t$, then imply that $\langle \nabla\omega(x_t), x - x_t\rangle \geq 0$ for any $x \in X_t$ and hence that $x_t = c_\omega(X_t)$, $t \geq 1$. We have thus shown that (3.17) holds.

We now show part b). The first relation follows from Lemma 1, (3.26), and the fact that X_t, $t \geq 0$ are localizers of $\mathcal{E}_f(l)$ due to part a) and the initial assumption on X_0. The second relation of part b) follows immediately from the definition of $\overline{f}_t$, $t \geq 0$.

To show part c), suppose that problem (3.27) is infeasible. Then, by (3.26), we have $h^* = +\infty$ and $\overline{f}_t = l$, which implies that the condition stated in Step 2 is satisfied and the procedure will be terminated.

Now let us show part d). Note that by part c), the set $\underline{X}_t$ is nonempty. Moreover, by the optimality condition of (3.27) and the definition of $\underline{X}_t$ in (3.29), we have $\langle \nabla\omega(x_t), x - x_t\rangle \geq 0$ for any $x \in \underline{X}_t$, which then implies that $\underline{X}_t \subseteq \overline{X}_t$. ∎

In view of Lemma 5.d), we can choose any set X_t satisfying $\underline{X}_t \subseteq X_t \subseteq \overline{X}_t$ (the simplest way is to set $X_t = \underline{X}_t$ or $X_t = \overline{X}_t$). Observe also that, while the number of constraints defining $\underline{X}_t$ increases with t, the set $\overline{X}_t$ has only one more constraint than X. By choosing X_t between these two extremes, we can control the number of constraints in subproblems (3.26) and (3.27).

The following result shows that if the above gap reduction procedure terminates, then the optimality gap $f(x_0^u) - \underline{f}_0$ will be reduced by a constant factor, which only depends on θ and β.

Lemma 6 *Whenever the gap reduction procedure terminates, we have*

$$f(\bar{x}^u) - \underline{f}^+ \leq q\,[f(x_0^u) - \underline{f}_0], \tag{3.31}$$

where

$$q \equiv q(\beta, \theta) := 1 - (1-\beta)\min\{\theta, 1-\theta\}. \tag{3.32}$$

Proof. Denote $\underline{f}_0 = f(x_0^u)$. Suppose first that the procedure terminates at the tth iteration since the condition $\underline{f}_t \geq l - \beta(l - \underline{f}_0)$ in Step 2 holds. By using this condition, and the facts that $f(\bar{x}^u) \leq \overline{f}_0$ (see Lemma 5.b) and $l = \theta\underline{f}_0 + (1-\theta)\overline{f}_0$, we obtain

$$f(\bar{x}^u) - \underline{f}^+ = f(\bar{x}^u) - \underline{f}_t \leq \overline{f}_0 - [l - \beta(l - \underline{f}_0)] = [1 - (1-\beta)(1-\theta)](\overline{f}_0 - \underline{f}_0). \tag{3.33}$$

Now suppose that the procedure terminates at the tth iterations since the condition $\overline{f}_t \leq l + \beta(\overline{f}_0 - l)$ in Step 4 holds. By using this condition, and the facts that $\underline{f}^+ \geq \underline{f}_0$ (see Lemma 5.b) and $l = \theta\underline{f}_0 + (1-\theta)\overline{f}_0$, we have

$$f(\bar{x}^u) - \underline{f}^+ = \overline{f}_t - \underline{f}^+ \leq l + \beta(\overline{f}_0 - l) - \underline{f}_0 = [1 - (1-\beta)\theta](\overline{f}_0 - \underline{f}_0).$$

Combining the above two relations, we obtain (3.31). ∎

The following result provides a bound on the total number of iterations performed by the gap reduction procedure before it terminates.

Proposition 2 *Suppose that α_t, $t \geq 1$, are chosen such that (3.20) holds. If $h(z_t, \cdot)$, $t \geq 1$, are support functions of f at z_t with smoothness level (ρ_i, M_i), $i = 1, \ldots, m$, then the total number of iterations performed by the gap reduction procedure does not exceed*

$$T := \left\lceil \max_{i=1,\ldots,m} \left[\frac{m\, \mathcal{C}_{\rho_i}\, M_i}{(1+\rho_i)\, \beta\theta\, (\overline{f}_0 - \underline{f}_0)} \left(\frac{2D_{\omega,X}}{\sigma_\omega} \right)^{\frac{1+\rho_i}{2}} \right]^{\frac{2}{1+3\rho_i}} \right\rceil, \tag{3.34}$$

where

$$D_{\omega,X} := \max_{x \in X} \omega(x) - \min_{x \in X} \omega(x). \tag{3.35}$$

In particular, if $h(z_t, \cdot)$, $t \geq 1$, are M-smooth support functions of f at z_t, then the above bound reduces to

$$\left\lceil \left[\frac{\mathcal{C}_1 M D_{\omega,X}}{\sigma\beta\theta\, (\overline{f}_0 - \underline{f}_0)} \right]^{\frac{1}{2}} \right\rceil. \tag{3.36}$$

Proof. By the definition of $D_{\omega,X}$ and the fact that $x_0 = c_\omega(X)$, we have $\omega(x) - \omega(x_0) \leq D_{\omega,X}$ for any $x \in X$. Using this observation, (3.24) and (3.34), we obtain

$$f(x^u_{\lceil T \rceil}) - l \leq \sum_{i=1}^{m} \left[\frac{\mathcal{C}_{\rho_i} M_i}{(1+\rho_i)\, T^{\frac{1+3\rho_i}{2}}} \left(\frac{2D_{\omega,X}}{\sigma_\omega} \right)^{\frac{1+\rho_i}{2}} \right] \leq \sum_{i=1}^{m} \frac{\beta\theta(\overline{f}_0 - \underline{f}_0)}{m} = \beta\theta(\overline{f}_0 - \underline{f}_0) = \beta(\overline{f}_0 - l), \tag{3.37}$$

where the last equality follows from the definitions of $\overline{f}_0$ and l. The above inequality then implies that the procedure will terminate at iteration T since the condition in Step 4 holds. ∎

Observe that in the previous result, it is assumed that $h(z_t, \cdot)$, $t \geq 1$, are support functions of f at z_t. In the following result, we will relax this assumption in a way such that the minorants $h(z_t, \cdot)$, $t \geq 1$, are not necessarily support functions for f, but for a closely approximated function of f.

Proposition 3 *Suppose that α_t, $t \geq 1$, are chosen such that (3.20) holds. If $h(z_t, \cdot)$, $t \geq 1$, are M-smooth support functions of $\tilde{f}$, where $\tilde{f}$ satisfies*

$$\tilde{f}(x) \leq f(x) \leq \tilde{f}(x) + \frac{\beta\theta(\overline{f}_0 - \underline{f}_0)}{2}, \quad \forall\, x \in X, \tag{3.38}$$

then the total number of iterations performed by the gap reduction procedure does not exceed

$$\tilde{T} := \left\lceil \left[\frac{2\mathcal{C}_1 M D_{\omega,X}}{\sigma\beta\theta(\overline{f}_0 - \underline{f}_0)} \right]^{\frac{1}{2}} \right\rceil, \tag{3.39}$$

where $D_{\omega,X}$ is defined in (3.35).

Proof. Observe that, in view of (3.38), $h(z_t, \cdot)$ are minorants (not necessarily support functions) of f and hence still valid as the input for the gap reduction procedure. Noting that $h(z_t, \cdot)$ are M-smooth support function of $\tilde{f}$ at z_t, we conclude from (3.25), (3.39) and the fact that $\omega(x_t) - \omega(x_0) \leq D_{\omega,X}$ due to (3.35) that

$$\tilde{f}(x^u_{\tilde{T}}) - l \leq \frac{\mathcal{C}_1 M D_{\omega,X}}{\sigma \tilde{T}^2} \leq \frac{\beta\theta}{2}(\overline{f}_0 - \underline{f}_0),$$

which together with (3.38), then imply that $f(x^u_{\tilde{T}}) - l \leq \beta\theta(\overline{f}_0 - \underline{f}_0) = \beta(\overline{f}_0 - l)$. Hence, the gap reduction procedure will terminate at iteration $\tilde{T}$ since the termination criterion in Step 4 is satisfied. ∎

4 Minimize generalized composite CP problems

Our goal in this section is to present an accelerated prox-level (APL) method for solving the generalized composite CP problems (see Subsection 2.1) by using the gap reduction procedure introduced in Section 3.

Throughout this section, we assume that the objective function f in (1.1) is given by (2.1). We first show how to define the support functions of f in (2.1) and to compute the initial lower and upper bounds on the optimal value f^*. We then present the APL method and demonstrate that it can achieve the optimal rate of convergence for solving problem (1.1)-(2.1).

We first show how to define the support functions of f given in (2.1).

Lemma 7 *Let $\phi_i'(z) \in \partial\phi_i(z)$ and denote $\phi'(z)d := (\langle\phi_1'(z), d\rangle; \ldots; \langle\phi_k'(z), d\rangle)$ for any $z \in X$ and $d \in \mathbb{R}^n$. The function $h_\Psi(z,\cdot)$ given by*

$$h_\Psi(z,x) := \Psi(\phi(z) + \phi'(z)(x-z)) \tag{4.1}$$

is a support function of f given in (2.1) at z with smoothness level (ν_i, L_0L_i), $i = 1,\ldots,k_0$, where ν_i and L_i, $i = 1,\ldots,k_0$ are defined in Assumptions 1 and 2, and

$$L_0 := \sup_{y\in\mathbb{R}^k, \delta\in\mathbb{R}^k_+} \left\{ \frac{\Psi(y+\delta) - \Psi(y)}{\|\delta\|_1} : \delta_i = 0, \forall\, k_0+1 \le i \le k. \right\}. \tag{4.2}$$

Proof. Clearly, we have $h_\Psi(z,z) = \psi(z)$ for any $z \in X$. Moreover, it follows from (2.2) and (3.2) that

$$\phi_i(z) + \langle\phi_i(z), x-z\rangle \le \phi_i(x) \le \phi_i(z) + \langle\phi_i'(z), x-z\rangle + \frac{L_i\|x-z\|^{\nu_i}}{1+\nu_i},$$

for any $i = 1,\ldots,k_0$ and that $\phi_i(x) = \phi_i(z) + \langle\nabla\phi_i(z), x-z\rangle$ for any $i = k_0+1,\ldots,k$. Using these observations, Assumption 2 and the definition of L_0 in (4.2), we have

$$\begin{aligned}\Psi(\phi(z) + \phi'(z)(x-z)) &\le \Psi(\phi(x)) \le \Psi(\phi(z) + \phi'(z)(x-z) + \frac{L_i}{1+\nu_i}\|x-z\|^{1+\nu_i})\\ &\le \Psi(\phi(z) + \phi'(z)(x-z)) + L_0\sum_{i=1}^{k_0}\frac{L_i}{1+\nu_i}\|x-z\|^{1+\nu_i}.\end{aligned}$$

■

It is worth noting that, in view of (4.2), we have $L_0 = 1$ for Examples 1,2,3 and 4 in Subsection 2.1.

Lemma 8 below shows how to compute the initial lower and upper bounds on f^* of problem (1.1)-(2.1).

Lemma 8 *Let $p_1 \in X$ be given and $\omega(\cdot)$ be a prox-function of X with modulus σ_ω. Also let f be given in (2.1) and denote*

$$\mathrm{lb}_1 := \min_{x\in X} h_\Psi(p_1,x) \quad \textit{and} \quad \mathrm{ub}_1 := \min\{f(p_1), f(\tilde{p}_1)\}, \tag{4.3}$$

where $\tilde{p}_1 \in Argmin_{x\in X} h_\Psi(p_1,x)$. Then, we have $\mathrm{lb}_1 \le f^*$ and

$$\mathrm{ub}_1 - \mathrm{lb}_1 \le \sum_{i=1}^{k_0}\frac{L_0L_i}{1+\nu_i}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}} =: \bar{\Delta}_\Psi. \tag{4.4}$$

Proof. Clearly, we have $\mathrm{lb}_1 \le f^*$ by minimizing both sides of $h_\Psi(p_1,x) \le f(x)$ w.r.t. $x \in X$. Moreover, it follows from Lemma 7 that,

$$\mathrm{ub}_1 - \mathrm{lb}_1 \le f(\tilde{p}_1) - h_\Psi(p_1,\tilde{p}_1) \le \sum_{i=1}^{k_0}\frac{L_0L_i}{1+\nu_i}\|p_1-\tilde{p}_1\|^{1+\nu_i} \le \sum_{i=1}^{k_0}\frac{L_0L_i}{1+\nu_i}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}},$$

where the last inequality follows (1.6). ■

We are now ready to describe the APL method for solving problem (2.1).

The APL method for generalized composite optimization:

Input: $p_1 \in X$, prox-function ω of X with modulus σ_ω, tolerance $\epsilon > 0$.

1) Compute lb_1 and ub_1 by (4.3). Set $s = 1$;
2) If $\mathrm{ub}_s - \mathrm{lb}_s \le \epsilon$, **terminate**;
3) Choose an arbitrary prox-center $o_s \in X$ (say, $o_s = p_1$ or $o_s = p_s$), define the prox-function

$$\omega_s(x) = \omega(x) - [\omega(o_s) + \langle \nabla\omega(o_s), x - o_s\rangle]; \tag{4.5}$$

4) Call the gap reduction procedure with input $y_0 = p_s$, $\underline{f}_0 = \mathrm{lb}_s$, $\omega = \omega_s$, and $h = h_\Psi$. Set $p_{s+1} = \bar{y}$, $\mathrm{ub}_{s+1} = f(\bar{y})$ and $\mathrm{lb}_{s+1} = \underline{f}^+$, where $\bar{y}$ and $\underline{f}^+$ are the output of the gap reduction procedure;
5) Set $s = s + 1$ and go to step 2.

We now add a few remarks about the APL method described above. First, we say that a phase of the APL method occurs whenever s increments by 1. Second, it can be easily seen from (4.5) that ω_s are prox-functions of X with modulus σ_ω and prox-center o_s. Moreover, we have $\omega_s(o_s) = 0$ and, by (3.35) and (1.4),

$$D_{\omega_s,X} = \max_{x\in X}\omega_s(x) - \min_{x\in X}\omega_s(x) = \max_{x\in X}\omega_s(x) = \max_{x\in X}\{\omega(x) - \omega(o_s) - \langle\nabla\omega(o_s), x - o_s\rangle\} \le \mathcal{D}_\omega, \tag{4.6}$$

for any $s \ge 1$. By defining the prox-function ω_s for each phase s, $s \ge 1$, we can control the prox-center o_s, and hence the starting point of the sequence of prox-centers $\{x_t\}$ used in the gap reduction procedure.

We describe in Theorem 1 the convergence behavior of the above APL method.

Theorem 1 *The following statement holds for the APL method applied to problem (1.1)-(2.1).*

a) the number of phases can be bounded by

$$\mathcal{S}_\Psi(\epsilon) \equiv \mathcal{S}(\bar{\Delta}_\Psi, \epsilon, q) := 1 + \max\left\{0, \log_{\frac{1}{q}}\frac{\bar{\Delta}_\Psi}{\epsilon}\right\}, \tag{4.7}$$

where q and $\bar{\Delta}_\psi$ are defined in (3.32) and (4.4), respectively;

b) the total number of gap reduction iterations can be bounded by $\mathcal{S}_\Psi(\epsilon) + \mathcal{N}_\Psi(\epsilon)$, where

$$\mathcal{N}_\Psi(\epsilon) := \sum_{i=1}^{k_0}\left[\frac{k_0\tilde{\mathcal{C}}_i L_0 L_i}{\beta\theta\,\epsilon}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}}\right]^{\frac{2}{1+3\nu_i}}, \quad \tilde{\mathcal{C}}_i \equiv \tilde{\mathcal{C}}(\nu_i, q) := \frac{\mathcal{C}_{\nu_i}}{(1+\nu_i)(1-q^{\frac{2}{1+3\nu_i}})}, \tag{4.8}$$

$\mathcal{D}_\omega$, $\mathcal{C}_\rho$, and L_0 are defined in (1.4), (3.20), and (4.2), respectively.

Proof. Denote $\Delta_s \equiv \mathrm{ub}_s - \mathrm{lb}_s$, $s \ge 1$. Without loss of generality, we assume that $\Delta_1 > \epsilon$, since otherwise the statements are obviously true. Note that by (3.31) and the origin of ub_s and lb_s, we have

$$\Delta_{s+1} \le q\Delta_s, \tag{4.9}$$

which clearly implies part a). Now, suppose that the APL method terminates at phase S for some $1 \le S \le \mathcal{S}_\Psi(\epsilon)$. Observe that $\Delta_S > \epsilon$, which together with (4.9) then imply that

$$\Delta_s \ge q^{s-S}\Delta_S > q^{s-S}\epsilon, s = 1, \ldots, S. \tag{4.10}$$

Using this observation, Lemma 7, Proposition 2 and relation (4.6), we conclude that the number of gap reduction iterations performed at phase s, $s \ge 1$, is bounded by

$$1 + \max_{i=1,\ldots,k_0}\left[\frac{k_0\mathcal{C}_{\nu_i}L_0L_i}{(1+\nu_i)\beta\theta\Delta_s}\left(\frac{2D_{\omega_s,X}}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}}\right]^{\frac{2}{1+3\nu_i}} \le 1 + \sum_{i=1}^{k_0}\left[\frac{k_0\mathcal{C}_{\nu_i}L_0L_i}{(1+\nu_i)\beta\theta\,\epsilon\,q^{s-S}}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}}\right]^{\frac{2}{1+3\nu_i}}.$$

Hence, the total number of gap reduction iterations is bounded by

$$S + \sum_{i=1}^{k_0}\left[\left(\frac{k_0\mathcal{C}_{\nu_i}L_0L_i}{(1+\nu_i)\beta\theta\,\epsilon\,q^{s-S}}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_i}{2}}\right)^{\frac{2}{1+3\nu_i}}\sum_{s=1}^{S} q^{\frac{2(S-s)}{1+3\nu_i}}\right],$$

which together with the fact that $\sum_{s=1}^{S} q^{\frac{2(S-s)}{1+3\nu_i}} = \sum_{t=1}^{S-1} q^{\frac{2t}{1+3\nu_i}} \le 1/(1-q^{\frac{2}{1+3\nu_i}})$, then clearly imply our result. ∎

We now add a few comments about the results obtained in Theorem 1. Firstly, it can be easily seen from (4.4), (4.7), and (4.8) that $\mathcal{S}_\Psi(\epsilon) = \mathcal{O}(\mathcal{N}_\Psi(\epsilon))$. Secondly, if there exists only one nonlinear component in the inner function $\phi(\cdot)$, i.e., $k_0 = 1$, then the bound $\mathcal{S}_\Psi(\epsilon) + \mathcal{N}_\Psi(\epsilon)$ reduces to

$$\mathcal{O}(1)\left[\frac{L_1}{\epsilon}\left(\frac{\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_1}{2}}\right]^{\frac{2}{1+3\nu_1}}.$$

Hence, the APL method is uniformly optimal for smooth, weakly smooth and non-smooth CP problems without requiring any information on L_1 and ν_1. Moreover, for a given $k_0 > 1$, we can see from (2.3) and (4.8) that the complexity bound in (4.8) is optimal, up to a constant factor depending on k_0, for solving the generalized composite CP problems. Finally, as shown by the following result, under certain special cases, one can improve the dependence of the iteration-complexity bound on the number of components of ϕ. It is worth noting that no modification to the APL algorithm is required for such an improvement.

Corollary 1 *Suppose that $\nu_1 = \nu_2 = \ldots = \nu_{k_0}$ in Assumption 1. Let us denote*

$$\tilde{L} := \sup_{y\in\mathbb{R}^k, t>0} \frac{\Psi(y+t\delta_L)-\Psi(y)}{t}, \quad \delta_L := (L_1,\ldots,L_{k_0},0,\ldots,0). \tag{4.11}$$

Then, the total number of gap reduction iterations performed by the APL method applied to problem (1.1)-(2.1) can be bounded by $\mathcal{S}_\Psi(\epsilon) + \tilde{\mathcal{N}}_\Psi(\epsilon)$, where

$$\tilde{\mathcal{N}}_\Psi(\epsilon) := \left[\frac{2\tilde{\mathcal{C}}_1\tilde{L}}{\sigma_\omega\beta\theta\epsilon}\left(\frac{2\mathcal{D}_\omega}{\sigma_\omega}\right)^{\frac{1+\nu_1}{2}}\right]^{\frac{2}{1+3\nu_1}}, \tag{4.12}$$

$\mathcal{D}_\omega$ and $\tilde{\mathcal{C}}_1$ are defined in (1.4) and (4.8), respectively.

Proof. Similarly to Lemma 7, we can show that $h_\Psi(z,\cdot)$ given by (4.1) is a support function of f in (2.1) with smoothness level $(\nu_1, \tilde{L})$. The rest of the proof is similar to that of Theorem 1 and hence the details are skipped. ∎

Consider a special case of problem (1.1)-(2.1) where $k = k_0$, $\nu_1 = \ldots = \nu_k$ and $\Psi(y) = \max_{1\le i\le k} y_i$ (see Example 2). We can easily see from (4.11) that $\tilde{L} = \max_{1\le i\le k_0} L_i$ and hence that, by Corollary (1), the iteration-complexity bound of the APL method merely depends on the number of components in the inner function $\phi(\cdot)$.

5 Minimize structured nonsmooth CP problems

In this section, we consider the structured non-smooth CP problem discussed in Subsection 2.2, where the objective function f in (1.1) is given by (2.4).

One possible approach for solving problem (1.1)-(2.4) would be to apply the APL method, which is shown to be optimal for smooth convex optimization (see Section 4), to the smooth approximation $\min_{x\in X} f_\eta(x)$ for some $\eta > 0$, similarly to Nesterov's smoothing scheme [36]. Note however, that this approach would require the number of iterations (or the target accuracy) and the problem parameter $\mathcal{D}_v$ (see (2.11)) given a prior. Our goal in this section is to present two new smoothing techniques, namely: the *basic and uniform smoothing level method*, obtained by properly modifying the APL method. In these methods, we adjust the smoothing parameter η dynamically during their execution rather than fix it in advance. While the first one of them requires some information about $\mathcal{D}_v$, the second one is completely "parameter-free", but can still achieve essentially the same iteration-complexity bound as that of the former one.

We start by describing how to define the minorants of f in (2.4).

Lemma 9 *Let f_η be defined in (2.8) for some $\eta > 0$. Then, for any $z \in X$, the function*

$$h_\eta(z,x) \equiv h_{f_\eta}(z,x) := \hat{f}(x) + F_\eta(z) + \langle \nabla F_\eta(z), x - z\rangle, \tag{5.1}$$

is a minorant of f. Moreover, $h_\eta(z,\cdot)$ is a $\mathcal{L}_\eta$-smooth support function of f_η at z, where $\mathcal{L}_\eta$ is defined in (2.9).

Proof. Clearly, by (5.1), the convexity of F_η and the first relation in (2.10), we have

$$h_\eta(z,x) \le \hat{f}(x) + F_\eta(x) \le \hat{f}(x) + F(x) = f(x),$$

which implies that $h_\eta(z,\cdot)$ is a minorant of f for any $z \in X$. It then follows from the above relation and (2.8) that $h_\eta(z,x) \le f_\eta(x)$ for any $x \in X$ and $h_\eta(z,z) = f_\eta(z)$. Moreover, by (2.8), (5.1) and the fact that F_η has $\mathcal{L}_\eta$-Lipschitz continuous gradient, we obtain

$$f_\eta(x) - h_\eta(z,x) = F_\eta(x) - [F_\eta(z) + \langle \nabla F_\eta(z), x - z\rangle] \le \frac{\mathcal{L}_\eta}{2}\|x - z\|^2.$$

Hence, $h_\eta(z,\cdot)$ is a $\mathcal{L}_\eta$-smooth support function of f_η at z. ∎

Before showing how to compute the initial lower and upper bounds on f^*, we first present a technical result which will be used to provide a convenient estimate on the gap between the initial lower and upper bounds on f^*. The proof of this result is given in the Appendix.

Lemma 10 *Let F be defined in (2.5) and v be a prox-function of Y with modulus σ_v. We have*

$$F(x_0) - F(x_1) - \langle F'(x_1), x_0 - x_1\rangle \le 2\left(\frac{2\|A\|^2\mathcal{D}_v}{\sigma_v}\right)^{\frac{1}{2}}\|x_0 - x_1\|, \quad \forall x_0, x_1 \in \mathbb{R}^n, \tag{5.2}$$

where $F'(x_1) \in \partial F(x_1)$ and $\mathcal{D}_v$ is defined in (1.4).

Lemma 11 below shows how to compute an initial lower bound on f^*.

Lemma 11 *Let $p_1 \in X$ be given and ω be a prox-function of X with modulus σ_ω. Also let f be defined in (2.4) and denote*

$$\mathrm{lb}_1 := \min_{x\in X}\left\{h_0(p_1,x) := \hat{f}(x) + F(p_1) + \langle F'(p_1), x - p_1\rangle\right\} \quad \textit{and} \quad \mathrm{ub}_1 := \min\{f(p_1), f(\tilde{p}_1)\}, \tag{5.3}$$

where $\tilde{p}_1 \in \mathrm{Argmin}_{x\in X} h_0(p_1,x)$. Then, we have $\mathrm{lb}_1 \le f^*$ and

$$\mathrm{ub}_1 - \mathrm{lb}_1 \le 4\left(\frac{\|A\|^2\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}\right)^{\frac{1}{2}} =: \bar{\Delta}_F. \tag{5.4}$$

Proof. By the convexity of F, (2.8) and (5.3), we can easily see that $\mathrm{lb}_1 \le f^*$. Moreover, we conclude from (2.8), (5.2) and (5.3) that

$$\begin{aligned}\mathrm{ub}_1 - \mathrm{lb}_1 &\le f(\tilde{p}_1) - \mathrm{lb}_1 = F(\tilde{p}_1) - F(p_1) - \langle F'(p_1), \tilde{p}_1 - p_1\rangle \\ &\le 2\left(\frac{2\|A\|^2\mathcal{D}_v}{\sigma_v}\right)^{\frac{1}{2}}\|\tilde{p}_1 - p_1\|,\end{aligned}$$

which, in view of (1.6), then implies the result. ∎

Assuming that an estimate $Q \ge \mathcal{D}_v$ is available, we now present the *basic smoothing level (BSL) method* for solving (2.4). The BSL method can be viewed as a variant of the APL method in Section 4 incorporated with the following two modifications:

a) The initial lower bound lb_1 and upper bound ub_1 are now given by (5.3);
b) In Step 4, the minorant h_Ψ is replaced by $h_{\eta_s(Q)}$ for some $Q \ge \mathcal{D}_v$, where h_η is defined in (5.1),

$$\eta_s(Q) := \frac{\beta\theta\Delta_s}{2Q} \quad \text{and} \quad \Delta_s \equiv \mathrm{ub}_s - \mathrm{lb}_s. \tag{5.5}$$

Observe that, by Lemma 9, $h_{\eta_s(Q)}(z,\cdot)$ is a $\mathcal{L}_{\eta_s(Q)}$-smooth support function for $f_{\eta_s(Q)}$ at z. Moreover, by (2.11), (5.5) and the fact that $Q \ge D_v$, we have

$$0 \le f(x) - f_{\eta_s(Q)}(x) \le \eta_s(Q)\mathcal{D}_v = \frac{\beta\theta\Delta_s\mathcal{D}_v}{2Q} \le \frac{\beta\theta\Delta_s}{2}. \tag{5.6}$$

In other words, $f_{\eta_s(Q)}$ is a close approximation of f and the error linearly depends on Δ_s.

We now describe the convergence properties of the BSL method.

Theorem 2 *The following statements hold for the BSL method applied to problem (1.1)-(2.4):*

a) the number of phases performed by the BSL method is bounded by $\mathcal{S}_F(\epsilon) \equiv \mathcal{S}(\bar{\Delta}_F, \epsilon, q)$*, where* $\mathcal{S}(\cdot,\cdot,\cdot)$ *and* $\bar{\Delta}_F$ *are defined in (4.7) and (5.4), respectively;*

b) the total number of gap reduction iterations performed by the BSL method does not exceed $\mathcal{S}_F(\epsilon) + \mathcal{N}_F(\epsilon, Q)$*, where*

$$\mathcal{N}_F(\epsilon, Q) := \frac{2\|A\|}{\beta\theta(1-q)\,\epsilon}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q}{\sigma_\omega\sigma_v}}. \tag{5.7}$$

Proof. Part a) follows from an argument similar to the one used in the proof of Theorem 1.a). Now suppose that the BSL method terminates at phase S for some $1 \le S \le \mathcal{S}_F(\epsilon)$. We conclude from (5.6) and Proposition 3 with $\tilde{f} = f_{\eta_s(Q)}$, $D_{\omega,X} = D_{\omega_s,X}$, $\overline{f}_0 - \underline{f}_0 = \Delta_s$ and $M = \mathcal{L}_{\eta_s(Q)}$ that the number of gap reduction iterations performed at phase s cannot exceed

$$1 + \sqrt{\frac{2\mathcal{C}_1\mathcal{L}_{\eta_s(Q)}D_{\omega_s,X}}{\sigma_\omega\beta\theta\Delta_s}} = 1 + \frac{2\|A\|}{\beta\theta\Delta_s}\sqrt{\frac{\mathcal{C}_1 D_{\omega_s,X}Q}{\sigma_\omega\sigma_v}} \le 1 + \frac{2\|A\|}{\beta\theta\Delta_s}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q}{\sigma_\omega\sigma_v}}, \tag{5.8}$$

where the second equality follows from (2.9) and (5.5), and the last inequality follows from (4.6). We then conclude from the previous conclusion, part a) and (4.10) that the total number of gap reduction iterations can be bounded by

$$S + \frac{2\|A\|}{\beta\theta}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q}{\sigma_\omega\sigma_v}}\sum_{s=1}^{S}\Delta_s^{-1} \le S + \frac{2\|A\|}{\beta\theta\epsilon}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q}{\sigma_\omega\sigma_v}}\sum_{s=1}^{S}q^{S-s} \le S + \frac{2\|A\|}{\beta\theta(1-q)\,\epsilon}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q}{\sigma_\omega\sigma_v}},$$

where the last inequality follows from the fact that $\sum_{s=1}^{S} q^{S-s} = \sum_{t=1}^{S-1} q^t \le 1/(1-q)$. ∎

We now add a few comments about the results obtained in Theorem 3. Firstly, provided $Q \ge \mathcal{D}_v$, the iteration complexity bound in (5.7) increases with Q. Hence, one can obtain the smallest possible bound by setting $Q = \mathcal{D}_v$. Secondly, we can easily see from (4.7) and (5.4) that

$$\mathcal{S}_F(\epsilon) = \mathcal{O}\left(\frac{\|A\|}{\epsilon}\sqrt{\frac{\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}}\right).$$

Hence, the iteration complexity bound $\mathcal{S}_F(\epsilon) + \mathcal{N}_F(\epsilon, Q)$ (with $Q = \mathcal{D}_v$) is in the same order of magnitude as those in [36] and [27]. Thirdly, observe that, if $Q < \mathcal{D}_v$ in (5.5), then relation (5.6) and thus Proposition 3 cannot apply, which in turn implies that we cannot guarantee the termination of the gap reduction procedure. However, in many practical situation, the value of $\mathcal{D}_v$ is not known exactly and one then has to use certain conservative estimation of this unknown quantity, which may slow down the BSL algorithm. To resolve the aforementioned problem, we present below a variant of the BSL method, referred to as the *uniform smoothing level (USL) method*, which does not require $\mathcal{D}_v$ and any other problem parameters a prior. We show that its iteration-complexity has the same order of magnitude as the one for the BSL method with $Q = \mathcal{D}_v$.

We first modify the gap reduction procedure in Subsection 3.4 by replacing step 5 with step 5' stated below, so as to guarantee its termination for the case when $Q < \mathcal{D}_v$ (see Lemma 12). The resulting gap reduction procedure will be called *the modified gap reduction procedure*. Note that we need one additional input, namely: $\tilde{f}$, which is a "close" approximation to the original function f.

5') If the following condition

$$\tilde{f}(x_t^u) - l \le \frac{\beta}{2}(\overline{f}_0 - l) \tag{5.9}$$

holds, report **failure** and terminate the procedure with output $\underline{f}^+ = \underline{f}_t$ and $\bar{x}^u$.

Otherwise, choose a convex compact set X_t such that $\underline{X}_t \subseteq X_t \subseteq \overline{X}_t$, set $t = t+1$ and go to Step 1, where $\underline{X}_t$ and $\overline{X}_t$ are defined in (3.29) and (3.30), respectively.

We are now ready to describe the USL method.

The uniform smoothing level method:

Input: $p_1 \in X$, prox-function ω of X with modulus σ_ω, tolerance $\epsilon > 0$ and initial estimate $Q_1 \in (0, \mathcal{D}_v]$.

1) Compute lb_1 and ub_1 by (5.3). Set $s = 1$;
2) If $\mathrm{ub}_s - \mathrm{lb}_s \le \epsilon$, **terminate**;
3) Choose an arbitrary prox-center $o_s \in X$ (say, $o_s = p_0$ or $o_s = p_s$) and define ω_s as in (4.5);
4) Call the modified gap reduction procedure with input $y_0 = p_s$, $\underline{f}_0 = \mathrm{lb}_s$, $\omega = \omega_s$, $h = h_{\eta(Q_s)}$ and $\tilde{f} = f_{\eta_s(Q_s)}$.
   - Set $p_{s+1} = \bar{y}$, $\mathrm{ub}_{s+1} = f(\bar{y})$, $\mathrm{lb}_{s+1} = \underline{f}^+$, where $\bar{y}$ and $\underline{f}^+$ are the output from the modified gap reduction procedure,
   - if the modified gap reduction procedure reports failure, set $Q_{s+1} = 2Q_s$, otherwise, set $Q_{s+1} = Q_s$;
5) Set $s = s + 1$ and go to step 2).

We now make a few remarks about the USL method described above. Firstly, each phase s, $s \ge 1$, of the USL method is associated with an estimation Q_s on $\mathcal{D}_v$, and we assume that $Q_1 \in (0, \mathcal{D}_v]$ is given. Note that such a Q_1 can be easily obtained due to the definition of $\mathcal{D}_v$. Secondly, we differentiate two types of phases: a phase is called *significant* if the gap reduction procedure reports success upon termination, otherwise, it is called *non-significant*. Thirdly, we increase the value of Q_s by a factor of 2 in a non-significant phase. We make a few more observations about these two types of phases in Lemma 12.

Lemma 12 *The following statements hold for the USL method:*

a) if phase s, $s = 1, 2, \ldots$, is non-significant, then $Q_s < \mathcal{D}_v$;
b) if phase s, $s = 1, 2, \ldots$, is significant, then $Q_s < 2\mathcal{D}_v$.
c) the number of gap reduction iterations performed at phase s, $s \ge 1$, cannot exceed

$$\left\lceil \hat{T}_s(Q_s) := \frac{2\|A\|}{\beta\theta\Delta_s}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q_s}{\sigma_\omega\sigma_v}} \right\rceil; \tag{5.10}$$

Proof. We first show part a). Assume for the contradiction that $Q_s \ge \mathcal{D}_v$ for a non-significant phase s. Note that in the modified gap reduction procedure, we have $\tilde{f} = f_{\eta_s(Q_s)}$ and $\overline{f}_0 - l = \theta\Delta_s$. Observe also that (5.9) holds for some $t \ge 1$ at the non-significant phase s. Using these observations, (2.11) and (5.6), we conclude that

$$f(x_t^u) - l \le f_{\eta_s(Q)}(x_t^u) - l + f(x_t^u) - f_{\eta_s(Q)}(x_t^u) \le \frac{\beta}{2}(\bar{f}_0 - l) + \frac{\beta\theta\Delta_s}{2} = \beta(\bar{f}_0 - l),$$

which implies that the modified gap reduction procedure must terminate with success. We have thus arrived at a contradiction. Part b) follows from part a) and the facts that $Q_1 < \mathcal{D}_v$ and that Q_s can be increased by a factor of 2 only in the non-significant phases. Now, denoting $\hat{T}_s \equiv \hat{T}_s(Q_s)$, similarly to (3.37), we can show that

$$\tilde{f}(x^u_{\lceil \hat{T}_s \rceil}) - l = f_{\eta_s(Q)}(x^u_{\lceil \hat{T}_s \rceil}) - l \le \frac{\beta}{2}(\bar{f}_0 - l),$$

which, in view of (5.9), implies that the procedure must terminate before or at iteration $\lceil \hat{T}_s \rceil$. ∎

As consequence of Lemma 12, if the initial estimate $Q_1 \ge \mathcal{D}_v$, then the non-significant phases will not occur and the USL method reduces to the BSL method. Theorem 3 below summarizes the main convergence properties of the USL method.

Theorem 3 *The following statements hold for the USL method applied to problem (1.1)-(2.4):*

a) the number of significant phases is bounded by $\mathcal{S}_F(\epsilon) \equiv \mathcal{S}(\bar{\Delta}_F, \epsilon, q)$, where $\mathcal{S}(\cdot,\cdot,\cdot)$ and $\bar{\Delta}_F$ are defined in (4.7) and (5.4), respectively, and the number of non-significant phases is bounded by $\tilde{\mathcal{S}}_F(Q_1) := \lceil \log \mathcal{D}_v/Q_1 \rceil$;
b) the total number of gap reduction iterations performed by the USL method does not exceed

$$\mathcal{S}_F(\epsilon) + \tilde{\mathcal{S}}_F(Q_1) + \frac{2\tilde{q}\|A\|}{\beta\theta\,\epsilon}\sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}}, \tag{5.11}$$

where $\tilde{q} := \sqrt{2}/(1-q) + \sqrt{2} + 1$.

Proof. Noting that (4.9) holds for a significant phase of the USL method, we can show the first claim in part a) by using an argument similar to the one used in the proof of Theorem 1.a). Moreover, the second claim in part a) immediately follows from the facts that a non-significant phase can occur only if $Q_1 \leq \mathcal{D}_v$ due to Lemma 12.a) and that Q_s, $s \geq 1$, is increased by a factor of 2 in each non-significant phase.

We now show that part b) holds. Let $B = \{b_1, b_2, \ldots, b_k\}$ and $N = \{n_1, n_2, \ldots, n_m\}$, respectively, denote the set of indices of the significant and non-significant phases. Note that $\Delta_{b_{t+1}} \leq q\,\Delta_{b_t}$, $t \geq 1$, and hence that $\Delta_{b_t} \geq q^{t-k}\Delta_{b_k} > \epsilon q^{t-k}$, $1 \leq t \leq k$. Note also that by Lemma 12, the number of gap reduction iterations in a significant phase b_t is bounded by $\lceil \hat{T}_{b_t}(2\mathcal{D}_v) \rceil$. Using these observations, we conclude that the total number of steps performed in the significant phases is bounded by

$$\begin{aligned}\sum_{t=1}^{k} \lceil \hat{T}_{b_t}(2\mathcal{D}_v) \rceil &\leq k + \sum_{t=1}^{k} \frac{2\|A\|}{\beta\theta\Delta_{b_t}} \sqrt{\frac{2\mathcal{C}_1\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}} \leq k + \frac{2\|A\|}{\beta\theta\epsilon} \sqrt{\frac{2\mathcal{C}_1\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}} \sum_{t=1}^{k} q^{k-t} \\ &\leq \mathcal{S}_F + \frac{2\|A\|}{\beta\theta(1-q)\epsilon} \sqrt{\frac{2\mathcal{C}_1\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}}, \end{aligned} \tag{5.12}$$

where the last inequality follows from part a) and the observation that $\sum_{t=1}^{k} q^{k-t} \leq 1/(1-q)$. Moreover, by Lemma 12, the number of steps performed in the non-significant phase n_r, $1 \leq r \leq m$, is bounded by $\hat{T}_{n_r}(Q_{n_r})$. Also note that $Q_{n_{r+1}} = 2Q_{n_r}$ for any $1 \leq r \leq m$. Using these observations and the fact that $\Delta_{n_r} > \epsilon$ for any $1 \leq r \leq m$, we conclude that the total number of gap reduction iterations performed in the non-significant phases is bounded by

$$\begin{aligned}\sum_{r=1}^{m} \lceil \hat{T}_{n_r}(Q_r) \rceil &= m + \sum_{r=1}^{m} \frac{2\|A\|}{\beta\theta\Delta_{n_r}} \sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q_r}{\sigma_\omega\sigma_v}} \leq m + \frac{2\|A\|}{\beta\theta\,\epsilon} \sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q_1}{\sigma_\omega\sigma_v}} \sum_{r=1}^{m} 2^{\frac{r-1}{2}} \\ &\leq \tilde{\mathcal{S}}_F + \frac{2\|A\|}{\beta\theta\,\epsilon} \sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega Q_1}{\sigma_\omega\sigma_v}} \sum_{r=1}^{\tilde{\mathcal{S}}_F} 2^{\frac{r-1}{2}} \leq \tilde{\mathcal{S}}_F + \frac{2\|A\|}{(\sqrt{2}-1)\beta\theta\,\epsilon} \sqrt{\frac{\mathcal{C}_1\mathcal{D}_\omega\mathcal{D}_v}{\sigma_\omega\sigma_v}}. \end{aligned} \tag{5.13}$$

Combining (5.12) and (5.13), we obtain (5.11). ∎

6 Applications and numerical illustration

In this section, we present some applications and preliminary numerical results to illustrate the effectiveness of the algorithms developed in Sections 4 and 5. In particular, we discuss the APL, BSL and USL methods applied to certain classes of semidefinite programming (SDP) and stochastic programming (SP) problems in Subsections 6.1 and 6.2, respectively.

6.1 A class of SDP problems

In this subsection, we consider the classic SDP problem of

$$\min_{x \in X} \lambda_1 \left(\mathcal{A}(x) \right), \tag{6.1}$$

where $X \subseteq \mathbb{R}^n$ is a convex and compact set, $\lambda_1 : \mathbb{R}^{m\times m} \to \mathbb{R}$ denotes the maximal eigenvalue of a symmetric matrix,

$$\mathcal{A}(x) = A_0 + \sum_{i=1}^{n} x_i A_i,$$

and A_i, $i = 1, \ldots, n$, are given $m \times m$ symmetric matrices.

Let a symmetric matrix $A \in \mathbb{R}^{m\times m}$ be given. It is well-known that the subdifferential of λ_1 at A is given by

$$\partial\lambda_1(A) = \mathrm{co}\left\{ uu^T : u^T u = 1, Au = \lambda_1(A)u \right\}, \tag{6.2}$$

where co$(\cdot)$ denotes the convex hull of all the rank-one matrices uu^T given in (6.2). Moreover, λ_1 is smooth (i.e., $\partial\lambda_1(A)$ is a singleton) if and only if the maximal eigenvalue of A has multiplicity 1.

One can solve problem (6.1) by using interior-point methods. However, due to the high iteration cost of interior-point methods, much effort has recently been directed to the development of first-order methods for solving problem (6.1). Since problem (6.1) is in general non-smooth, one can use general non-smooth convex optimization methods, such as, the Non-Euclidean Restricted Memory Level (NERML) algorithm [6,5], or the APL method presented in Section 4, for solving (6.1). These methods, in the worst case, require $\mathcal{O}(1/\epsilon^2)$ iterations to find an ϵ-solution of problem (6.1), and the major iteration costs of these methods consist of finding a maximal eigenvector u_x of $\mathcal{A}(x)$ for a given $x \in X$ and assembling the subgradient $\mathcal{A}^* u_x$, where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$.

In comparison with the NERML algorithm, a nice feature of the APL method is that it can automatically explore the local smoothness structures of a particular problem instance, as the objective function of (6.1) may be differentiable along certain parts of the trajectory of the algorithm. It should be noted that other bundle type methods, such as the spectral-bundle method by Helmberg and Rendl [9], have also been developed for solving problem (6.1). The spectral-bundle method is obtained by tailoring the well-known bundle method [12,13,18,25] to problem (6.1). By making use of the specific structure of problem (6.1), each iteration of this method requires the solution of a quadratic semidefinite programming problem. Note however, that there is no complexity results available for the aforementioned spectral bundle method.

Since problem (6.1) can also be written as a bilinear saddle point problem:

$$\min_{x\in X} \lambda_1(\mathcal{A}(x) = \min_{x\in X} \max_{y\in Y} \langle \mathcal{A}(x), y\rangle, \tag{6.3}$$

where $Y := \left\{y \in \mathbb{R}^{m\times m} | \mathrm{Tr}(y) = 1, y \succeq 0\right\}$, we can apply Nesterov's smoothing scheme (NEST-S) [36,38] (see Subsection 2.2), as well as the BSL and USL methods developed in Section 5, for solving (6.3). These methods can find an ϵ-solution of (6.3) in at most $\mathcal{O}(1/\epsilon)$ iterations. It should be noted that the iteration costs of these methods can slightly differ from each other. More specifically, both the BSL and USL method applied to (6.3) require a full eigenvalue decomposition and one computation of the adjoint operator $\mathcal{A}^*$ to define the minorant h_η (see (5.1)) in Step 1 of the gap reduction procedures. Moreover, while the BSL only requires to find a maximum eigenvalue to compute $f(x_t^u)$ in Step 3 of the gap reduction procedure, the USL method requires a full eigenvalue decomposition in Step 5' to compute the value of $\tilde{f}(x_t^u)$ (see (5.9)). On the other hand, each iteration of Nesterov's smoothing scheme [36] requires two (or one in some variants of Nesterov's method, see, e.g.,[14]) full eigenvalue decompositions and computations of the adjoint operator $\mathcal{A}^*$.

A few more details about our experiments are outlined as follows. We have implemented five different algorithms, namely: NERML, APL, BSL, USL and NEST-S, for solving (6.1). We assume that the feasible set X is a standard simplex given by $\left\{x \in \mathbb{R}^n | \sum_{i=1}^n x_i = 1, x_i \geq 0, \forall i\right\}$. The prox-function of X, as required by all these five algorithms, is set to $u(x) = \sum_{i=1}^n x_i \log x_i$ and the norm is set to $\|x\|_1$. The prox-function of Y, as required by the algorithms BSL, USL and NEST-S, is set to $v(y) = \sum_{i=1}^n \lambda_i(y) \log \lambda_i(y)$, and the norm is set to $\sum_{i=1}^n \lambda_i(y)$, where $\lambda_i(y)$, $i = 1, \ldots, n$, denote the eigenvalues of $y \in Y$.

- For the APL, BSL, USL and NERML algorithms, we define the localizer X_t as

$$X_t = \{x \in X : \langle \nabla\omega(z_t), x - x_t\rangle \geq 0\} \bigcap \mathcal{M}_t, \quad t \geq 0,$$

  where $\mathcal{M}_t$ denotes the intersection of totally at most B half spaces of the form $\{x : h(z_t, x) \leq l\}$ which have been generated most recently (see [15]). Note that the larger the size of B is, the more difficult the subproblems (3.26) and (3.27) are. On the other hand, a larger B might help to compute a better lower bound in (3.26). It is found from our initial experiments that different values of B within $[10, 50]$ perform almost equally well. Hence, we set B to 30 in our final experiments. The subproblems (3.26) and (3.27) are solved by Mosek [26]. Please see [15] for other possibilities to solve these subproblems. Moreover, we set $\beta = \theta = 1/2$, and use the $\{\alpha_t\}$ in (3.21) for the APL, BSL and USL methods.
- For the NEST-S scheme, we compute the Lipschitz constant $\mathcal{L}_\eta$ by (2.9), where the smoothing parameter η is set to

$$\frac{2\|\mathcal{A}\|}{N+1} \frac{D_\omega}{\sigma_\omega \sigma_v D_v}.$$

  and the operator norm $\|\mathcal{A}\|$ is computed according to [38]. Note however, that the resulting estimate of $\mathcal{L}_\eta$ can be rather conservative, which leads to the slow convergence of the NEST-S scheme. We had also implemented a variant

of Nesterov's method which can adaptively search for the Lipschitz constant $\mathcal{L}_\eta$ ([37]). However, our preliminary experiments indicate that the improvement from this approach is not significant. In our final experiments, we run NEST-S four times, and each time we multiply the Lipschitz constant $\mathcal{L}_\eta$ estimated above by a different factor: 10^{-1}, 10^{-2}, 10^{-3}, or 10^{-4}. We then report the best solutions, in terms of the objective value, obtained from these four runs of the NEST-S scheme.

– All the codes are implemented in MATLAB2007 under Windows Vista and the experiments were conducted in an INTEL 2.53 GHz labtop.

The experiments were conducted on a set of randomly generated instances (i.e., E1, E2 and E3) of problem (6.1), each of which has various size of A_i's for $i = 1, \ldots, n$. More details about these instances are shown in Table 1, where n is the dimension of x, ub_1 is the objective value at $p_1 = (1/n, \ldots, 1/n)$, and $\Delta_1 = \mathrm{ub}_1 - \mathrm{lb}_1$ denotes the initial gap with lb_1 given by (5.3). We run 200 iterations for all these algorithms and report the objective values obtained at the 100th and 200th iteration in Column 2 and Column 4, respectively. For the APL, BSL, USL and NERML algorithms, we also report the optimality gap Δ_s at the 100th and 200th iteration, respectively, in column 3 and 5 of Table 1. The CPU time (in Seconds) for running these algorithms are reported in column 6 of Table 1. It should be noted that we only report the CPU time for one run of the NEST-S algorithm, although we had run it for 4 times to find a good estimate of $\mathcal{L}_\eta$.

Table 1 Experiments with the maximal eigenvalue problem

E1: $n = 1,000$, $m = 400$, $d = 2\%$, $\mathrm{ub}_1 = 6.329960$, $\Delta_1 = 5.08e-1$					
alg.	ub_{100}	Δ_{100}	ub_{200}	Δ_{200}	Time
BSL	6.026060	$8.48e-5$	6.026045	$1.28e-6$	174.10
USL	6.026066	$1.66e-4$	6.026045	$3.37e-6$	250.20
NEST-S	6.077040	-	6.076351	-	190.51
APL	6.026048	$4.23e-5$	6.026045	$1.22e-6$	109.59
NERML	6.026323	$7.97e-4$	6.026084	$2.39e-4$	100.37
E2: $n = 1,000$, $m = 600$, $d = 2\%$, $\mathrm{ub}_1 = 7.788735$, $\Delta_1 = 6.45e-1$					
alg.	ub_{100}	Δ_{100}	ub_{200}	Δ_{200}	Time
BSL	7.458582	$3.39e-4$	7.458538	$5.07e-6$	364.93
USL	7.458558	$2.63e-4$	7.458538	$3.34e-6$	596.56
NEST-S	7.539811	-	7.538583	-	552.12
APL	7.458561	$8.96e-5$	7.458537	$1.96e-6$	166.27
NERML	7.458801	$1.07e-3$	7.458557	$8.35e-5$	142.16
E3: $n = 1,000$, $m = 800$, $d = 2\%$, $\mathrm{ub}_1 = 8.855385$, $\Delta_1 = 5.57e-1$					
alg.	ub_{100}	Δ_{100}	ub_{200}	Δ_{200}	Time
BSL	8.555496	$1.35e-4$	8.555475	$4.08e-6$	799.30
USL	8.555529	$1.19e-4$	8.555475	$4.50e-6$	1377.67
NEST-S	8.635632	-	8.635473	-	1347.26
APL	8.555484	$6.61e-5$	8.555475	$2.05e-6$	275.22
NERML	8.555743	$7.45e-4$	8.555494	$1.24e-4$	213.25

We can draw a few conclusions from our experiments with the maximum eigenvalue problem. Firstly, among the three methods with $\mathcal{O}(1/\epsilon)$ convergence, both BSL and USL can significantly outperform NEST-S: the former two algorithms can reach 6 accuracy digits after 200 iterations while the latter algorithm reaches at most 2 accuracy digits for these instances. Also note that the solution quality of the USL is comparable to that of the BSL algorithm although its computational time is larger, since it involves one more full eigenvalue decomposition in each iteration. Secondly, for the two non-smooth methods, the APL consistently outperform the NERML algorithm in terms its solution quality while the computational time is comparable to the latter one. Thirdly, while the solution quality of the BSL and USL methods are significantly better than the one of the NERML algorithm, it is interesting to notice that the solution quality of the APL algorithm is comparable or better than that of the BSL and USL. One plausible explanation is that the problems to be solved, due to the inherent randomness, are smooth along most part of the trajectory of the APL algorithm. We will see in next subsection that the USL method can indeed outperform the APL method in some cases.

6.2 A class of two-stage stochastic programming problems

In this subsection, we consider the classic two-stage stochastic linear programming given by

$$\min_{x\in X}\left\{f(x)=c^Tx+\mathbb{E}[V(x,\xi)]\right\}, \tag{6.4}$$

with

$$V(x,\xi)=\min\left\{q^T\pi : W\pi=h+Tx,\pi\geq 0\right\}. \tag{6.5}$$

Here, $x\in\mathbb{R}^{n_1}$ and $\pi\in\mathbb{R}^{n_2}$, respectively, are the first and second-stage decision variables, $X\subseteq\mathbb{R}^{n_1}$ is a non-empty convex compact set, and $\xi\equiv(q,h,T)$ is a random vector with a known distribution supported on $\Xi\subseteq\mathbb{R}^{n_2+m_2+m_2\times n_1}$. We assume that problem (6.5) is feasible for every possible realization of ξ, i.e., problem (6.4) has a complete recourse. Moreover, for the purpose of illustrating the effectiveness of the algorithms developed in this paper, we assume that ξ is a discrete random vector and the number of possible realizations of ξ (or the sample space) is not too big.

It should be noted that if ξ is a continuous random vector or the number of possible realizations of ξ is astronomically large, to solve problem (6.4) is highly challenging, due to the fundamental difficulty of computing the expectation to a high accuracy when the dimension of ξ is high, see [28,17] for a discussion on some recent advancements in this area. However, if the number of possible realizations of ξ is not astronomically large, it is possible to solve problem (6.4) to a high accuracy in a reasonable amount of time (especially under a parallel computing environment) by using more powerful algorithms. This is indeed what we intend to demonstrate in this subsection.

Since problem (6.4) is a general non-smooth CP problem, one can apply the the NERML method, or APL algorithm presented in Section 4 to find an ϵ-solution of (6.4) in at most $\mathcal{O}(1/\epsilon^2)$ iterations. Recently, Ahmed [1] shows that one can improve this complexity bound to $\mathcal{O}(1/\epsilon)$, by applying Nesterov's smoothing scheme to (6.4). The basic idea is as follows. Let $\mathcal{Y}(q):=\{W^Ty\leq q\}$ and $\mathcal{B}_{m_2}$ be the Euclidean ball in $\mathbb{R}^{m_2}$. Note that by strong duality, we have

$$V(x,\xi)=\max\left\{(h+Tx)^Ty : y\in\mathcal{Y}(q)\right\}. \tag{6.6}$$

Moreover, by Hoffman's Lemma [10], there exists a constant $\mathcal{R}_W>0$ depending on W such that

$$\mathcal{Y}(q)\subseteq\mathcal{Y}(0)+\mathcal{R}_W\|q\|\mathcal{B}_{m_2}=\mathcal{R}_W\|q\|\mathcal{B}_{m_2},$$

where the last identity follows from the fact that $\mathcal{Y}(0)=\{0\}$ due to the complete recourse assumption. In other words, the feasible region of (6.6) is bounded. We can then uniformly approximate $f(x)$ in (6.4) by $f_\eta(x):=c^Tx+\mathbb{E}[V_\eta(x,\xi)]$ for some $\eta>0$, where

$$V_\eta(x,\xi)=\max\left\{(h+Tx)^Ty-\eta\|y\|^2/2 : y\in\mathcal{Y}(q)\right\}. \tag{6.7}$$

To apply Nesterov's smoothing scheme, it is necessary to fine-tune a large number of problem parameters, including $\mathcal{R}_W$, $\|q\|$ and $\|T\|$, as well as $\mathcal{D}_\omega$. The preliminary numerical results of this approach were not very promising [2].

In our experiments, we have implemented three methods, namely: APL and NERML and USL, applied to problem (6.4). All these methods do not require the input of any problem parameters and the implementation details are similar to those in Subsection 6.1 except that we set the prox-function of X to $\|x\|^2/2$. We conduct our experiments on a few SP instances which have been studied by a few authors, namely: a telecommunication design (SSN) problem of Sen, Doverspike, and Cosares [41] and the motor freight carrier routing problem (20-term) of Mak, Morton, andWood [24]. The dimensions of these instances are shown in Table 2, please see [21] for more details about these instances.

Table 2 Dimension of the SP instances

	n_1	m_1	n_2	m_2
SSN	89	1	796	175
20-term	63	3	764	124

It is worth noting that here we assume that the number of possible realizations are fixed ($N=50$ or 100) and hence obtain four different instances, namely: SSN(50), SSN(100), 20-term(50) and 20-term(100). Noting that the initial

optimality gap for these instances are rather high (in order of 10^3 or 10^7), we run each algorithm for 400 iterations and the results are reported in Table 3. The structure of the table is similar to Table 1 (see Subsection 6.1).

We can make a few observations from the numerical results in Table 3. Firstly, the iteration cost of the USL method is larger than that of the APL method, which, in turn, is larger than that of the NERML algorithm. In particular, the major iteration cost of the NERML and APL algorithm consists of solving N and $2N$ second-stage LP problems respectively, while the one of the APL algorithm involves the solutions of N smoothed quadratic programming problems (see (6.7)). Secondly, both the APL and USL methods can significantly outperform the the NERML algorithm in terms of the solution quality. As we can see from Table 3, the NERML algorithm makes little progresses after 200 iterations for these SP instances. Thirdly, the solution quality of the APL method is worse than that of the USL method for solving the first two instances: SSN(50) and SSN(100), but it significantly outperforms the latter one for solving the last two instances: 20-term(50) and 20-term(100). One possible reason is that the sizes of $\mathcal{D}_v(\approx \mathcal{R}_W^2\|q\|^2)$ for the last two instances are significantly larger than those for the first two instances, see Table 4 for the estimates on $\mathcal{D}_v$ reported by the USL method (with $Q_1 = 1$).

Table 3 Experiments with the two-stage SP problem

SSN(50): $\mathrm{ub}_1 = 2.352586e+2$, $\Delta_1 = 3.923265e+3$					
alg.	ub_{200}	Δ_{200}	ub_{400}	Δ_{400}	Time
APL	4.839075	$2.437395e-3$	4.838074	$5.053628e-7$	366.30
USL	4.838125	$1.837968e-4$	4.838073	$6.599449e-7$	754.31
NERML	5.550903	$5.012402e+0$	5.086603	$1.303485e+0$	193.47
SSN(100): $\mathrm{ub}_1 = 2.407279e+2$, $\Delta_1 = 4.023982e+3$					
alg.	ub_{200}	Δ_{200}	ub_{400}	Δ_{400}	Time
APL	7.354770	$9.148243e-3$	7.352610	$4.198017e-6$	730.15
USL	7.354090	$2.683424e-3$	7.352610	$6.804606e-7$	1471.62
NERML	8.323381	$4.771295e+0$	7.578802	$1.079491e+0$	383.27
20-term(50): $\mathrm{ub}_1 = 7.718543e+5$, $\Delta_1 = 1.804693e+7$					
alg.	ub_{200}	Δ_{200}	ub_{400}	Δ_{400}	Time
APL	$2.549453e+5$	$1.229655e-3$	$2.549453e+5$	$2.405432e-7$	1056.82
USL	$2.551031e+5$	$1.896133e+3$	$2.549602e+5$	$3.310795e+2$	1209.53
NERML	$2.587140e+5$	$1.473815e+4$	$2.576649e+5$	$1.368899e+4$	301.03
20-term(100): $\mathrm{ub}_1 = 7.664067e+5$, $\Delta_1 = 1.801832e+7$					
alg.	ub_{200}	Δ_{200}	ub_{400}	Δ_{400}	Time
APL	$2.532875e+5$	$3.679608e-3$	$2.532875e+5$	$2.463930e-7$	1895.63
USL	$2.533441e+5$	$5.119095e+2$	$2.532923e+5$	$7.614912e+1$	2517.26
NERML	$2.581546e+5$	$2.171689e+4$	$2.540804e+5$	$3.754735e+3$	602.60

Table 4 Estimate on D_v from the USL method

	SSN(50)	SSN(100)	20-term(50)	20-term(100)
Q	64	128	$1.68e+7$	$1.68e+7$

7 Appendix

In this section, we provide the proof of Lemma 10.

Let F and F_η be defined in (2.5) and (2.7), respectively. Also let us denote, for any $\eta > 0$ and $x \in X$,

$$\psi_x(z) := F_\eta(x) + \langle \nabla F_\eta(x), z - x\rangle + \frac{\mathcal{L}_\eta}{2}\|z - x\|^2 + \eta\mathcal{D}_v, \tag{7.1}$$

where $\mathcal{D}_v$ and $\mathcal{L}_\eta$ are defined in (1.4) and (2.9), respectively. Clearly, in view of (1.3) and (2.10), ψ_x is a majorant of both F_η and f. Also let us define

$$Z_x := \left\{ z \in \mathbb{R}^n : \|z - x\|^2 = \frac{2}{\mathcal{L}_\eta}\left[\eta\mathcal{D}_v + F_\eta(x) - F(x)\right] \right\}. \tag{7.2}$$

Clearly, by the first relation in (2.10), we have

$$\|z - x\|^2 \leq \frac{2\eta \mathcal{D}_v}{\mathcal{L}_\eta}, \quad \forall\, z \in Z_x. \tag{7.3}$$

Moreover, we can easily check that, for any $z \in Z_x$,

$$\psi_x(z) + \langle \nabla \psi_x(z), x - z \rangle = F(x), \tag{7.4}$$

where $\nabla \psi_x(z) = \nabla F_\eta(x) + \mathcal{L}_\eta(z - x)$.

The following results provides the characterization of a subgradient direction of F.

Lemma 13 *Let $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$ be given. Then, $\exists z \in Z_x$ such that*

$$\langle F'(x), p \rangle \leq \langle \nabla \psi_x(z), p \rangle = \langle \nabla F_\eta(x) + \mathcal{L}_\eta(z - x), p \rangle.$$

where $F'(x) \in \partial F(x)$.

Proof. Let us denote

$$t = \frac{1}{\|p\|} \left\{ \frac{2}{\mathcal{L}_\eta} \left[\eta \mathcal{D}_v + F_\eta(x) - F(x) \right] \right\}^{\frac{1}{2}}.$$

and $z_0 = x + tp$. Clearly, in view of (7.2), we have $z_0 \in Z_x$. By convexity of F and (7.4), we have

$$\begin{aligned} F(x) + \langle F'(x), tp \rangle \leq F(x + tp) &= \psi_x(z_0) = F(x) + \langle \nabla \psi_x(z_0), z_0 - x \rangle \\ &= F(x) + t \langle \nabla \psi_x(z_0), p \rangle, \end{aligned}$$

which clearly implies the result. ∎

We are now ready to prove Lemma 10.

Proof of Lemma 10. First note that by the convexity of F, we have

$$F(x_0) - \big[F(x_1) + \langle F'(x_1), x_0 - x_1 \big] \rangle \leq \langle F'(x_0), x_0 - x_1 \rangle + \langle F'(x_1), x_1 - x_0 \rangle.$$

Moreover, by Lemma 13, $\exists z_0 \in Z_{x_0}$ and $z_1 \in Z_{x_1}$ s.t.

$$\begin{aligned} &\langle F'(x_0), x_0 - x_1 \rangle + \langle F'(x_1), x_1 - x_0 \rangle \\ &\leq \langle \nabla F_\eta(x_0) - \nabla F_\eta(x_1), x_0 - x_1 \rangle + \mathcal{L}_\eta \langle z_0 - x_0 - (z_1 - x_1), x_0 - x_1 \rangle \\ &\leq \mathcal{L}_\eta \|x_0 - x_1\|^2 + \mathcal{L}_\eta (\|z_0 - x_0\| + \|z_1 - x_1\|) \|x_0 - x_1\| \\ &\leq \mathcal{L}_\eta \|x_0 - x_1\|^2 + 2\mathcal{L}_\eta \left(\frac{2\eta \mathcal{D}_v}{\mathcal{L}_\eta} \right)^{\frac{1}{2}} \|x_0 - x_1\| \\ &= \frac{\|A\|^2}{\sigma_v \eta} \|x_0 - x_1\|^2 + 2 \left(\frac{2\|A\|^2 \mathcal{D}_v}{\sigma_v} \right)^{\frac{1}{2}} \|x_0 - x_1\|, \end{aligned}$$

where the last inequality and equality follow from (7.3) and (2.9), respectively. Combining the above two relations, we have

$$F(x_0) - \big[F(x_1) + \langle F'(x_1), x_0 - x_1 \rangle \big] \leq \frac{\|A\|^2}{\sigma_v \eta} \|x_0 - x_1\|^2 + 2 \left(\frac{2\|A\|^2 \mathcal{D}_v}{\sigma_v} \right)^{\frac{1}{2}} \|x_0 - x_1\|.$$

The result now follows by tending η to $+\infty$ in the above relation. ∎

References

1. S. Ahmed. Smooth minimization of two-stage stochastic linear programs. Manuscript, Georgia Institute of Technology, 2006.
2. S. Ahmed. personal communication, 2010.
3. A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.
4. S. Becker, J. Bobin, and E. Candes. Nesta: A fast and accurate first-order method for sparse recovery. Manuscript, California Institute of Technology, 2009.
5. A. Ben-Tal and A.Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming*, 102:407–456, 2005.
6. A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2000.
7. A. d'Aspremont, O. Banerjee, and L. El Ghaoue. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30:56–66, 2008.
8. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, July 2010.
9. C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10:673–696, 2000.
10. A. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards, Section B. Math. Sci.*, 49:263265, 1952.
11. A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. Manuscript, Georgia Institute of Technology, Atlanta, GA, 2008. submitted to SIAM Journal on Control and Optimization.
12. K.C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.
13. K.C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122, 1990.
14. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010. Forthcoming, Online first.
15. G. Lan. Bundle-type methods uniformly optimal for smooth and non-smooth convex optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, November 2010.
16. 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.
17. G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 2011. Forthcoming, online first.
18. C. Lemaréchal. An extension of davidon methods to non-differentiable problems. *Mathematical Programming Study*, 3:95–109, 1975.
19. C. Lemaréchal, A. Nemirovski, and Y. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–148, 1995.
20. A.S. Lewis and S.J. Wright. A proximal method for composite minimization. Manuscript, Cornell University, Ithaca, NY, 2009.
21. J. Linderoth, A. Shapiro, and S. Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 142:215–241, 2006.
22. Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19:1807–1827, 2009.
23. 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.
24. W. K. Mak, D.P. Morton, and R.K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.
25. R. Mifflin. A modification and an extension of lemaréchal's algorithm for nonsmooth minimization. *Mathematical Programming Study*, 17:77–90, 1982.
26. Mosek. The mosek optimization toolbox for matlab manual. version 6.0 (revision 93). `http://www.mosek.com`.
27. 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.
28. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
29. A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
30. A. S. Nemirovski. *Efficient methods in convex programming*. Lecture notes, Technion, 1994.
31. A. Nemirovskii and Y.Nesterov. Optimal methods for smooth convex minimization. *Zh. Vichisl. Mat. Fiz. (In Russian)*, 25:356–369, 1985.
32. 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.
33. Y. E. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomo. i. Mat. Metody*, 24:509–517, 1988.
34. Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

35. Y. E. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16:235–249, 2005.
36. Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.
37. Y. E. Nesterov. Gradient methods for minimizing composite objective functions. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, September 2007.
38. Y. E. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110:245–259, 2007.
39. O.Devolder, F.Glineur, and Yu.Nesterov. First-order methods of smooth convex optimization with inexact oracle. Manuscript, CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium, December 2010.
40. J. Peña. Nash equilibria computation via smoothing techniques. *Optima*, 78:12–13, 2008.
41. S. Sen, R.D. Doverspike, and S. Cosares. Network planning with random demand. *Telecommunication Systems*, 3:11–30, 1994.
42. A. Shapiro. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28:677–692, 2003.
43. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.