Bundle-type methods uniformly optimal for smooth and nonsmooth convex optimization *†

Guanghui Lan ‡

December 7, 2010

Abstract

The bundle-level method and their certain variants are known to exhibit an optimal rate of convergence, i.e., $\mathcal{O}(1/\sqrt{t})$, and also excellent practical performance for solving general non-smooth convex programming (CP) problems. However, this rate of convergence is significantly worse than the optimal one for solving smooth CP problems, i.e., $\mathcal{O}(1/t^2)$. In this paper, we present new bundle-type methods which possess the optimal rate of convergence for solving, not only non-smooth, but also smooth CP problems. Interestingly, these optimal rates of convergence can be obtained without even knowing whether a CP problem is smooth or not. Moreover, given that the problem is smooth, the developed methods do not require any smoothness information, such as, the size of the Lipschitz constant. To the best of our knowledge, this is the first time that uniformly optimal algorithms of this type have been presented in the literature.

Keywords: Convex Programming, Complexity, Bundle-level, Optimal methods

1 Introduction

Consider the basic convex programming (CP) problem of

$$f^* := \min_{x \in X} f(x), \tag{1}$$

where X is a convex compact set and $f : X \to \Re$ is a closed convex function such that

$$|f(x) - f(y)| \leq \mathcal{M}\|x - y\|, \ \forall x, y \in X. \tag{2}$$

Moreover, $f(\cdot)$ is represented by a first-order oracle which, upon requests, returns $f(x)$ and $f'(x) \in \partial f(x)$, where $\partial f(x)$ denotes the subdifferential of $f(\cdot)$ at $x \in X$. Different optimization techniques, including subgradient descent, mirror descent and bundle-type methods, have been developed for solving these general non-smooth CP problems [26, 4, 28]. In particular, the subgradient descent method [24] exhibits an $\mathcal{O}(1/\sqrt{t})$ rate of convergence, which is optimal if n is sufficiently large, i.e.,

*The manuscript is available on www.optimization-online.org.

†The author of this paper was partially supported by NSF Grant: CMMI-1000347.

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

$n \geq \mathcal{O}(t)$, where t denotes the number of calls to the oracle. Moreover, some bundle-type algorithms, such as, the bundle-level method and some of its variants [19, 4, 3], also exhibit $\mathcal{O}(1/\sqrt{t})$ rate of convergence. While the subgradient descent method is known for their slow convergence in practice, the bundle-level method and their certain variants often converge linearly, with an experimental rate given by $\mathcal{O}(\exp(-t/n))$, for solving many non-smooth CP problems in practice [19, 4, 23, 17].

One can improve the theoretical rate of convergence for solving (1), only by considering more specialized classes of CP problems. In particular, for minimizing smooth CP problems satisfying

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

Nesterov in a seminal work [25] presented an algorithm with the rate of convergence bounded by $\mathcal{O}(1/t^2)$. By the complexity theory for convex optimization [24], Nesterov's method is optimal for smooth convex optimization when $n \geq \mathcal{O}(t)$. Some variants of this method were further studied in [26], [1], [27], [16] and [30]. The basic idea of these methods is to construct and then to minimize a series of quadratic approximations of $f(\cdot)$ whose Hessians are given by the identity. By far, Nesterov's method and its variants are the only known optimal algorithms for smooth convex optimization. Recently, certain generalized versions of Nesterov's method [15, 7] were shown to be optimal for solving smooth, nonsmooth and stochastic CP problems, provided that some global information about $f(\cdot)$, such as L and $\mathcal{M}$, is given explicitly as the input of these algorithms. Note, however, that by definition, these constants describe the problem in a global scope and thus possibly in a very conservative way. For example, a general non-smooth function may turn out to be smooth locally or even, in the extreme case, along the whole trajectory of the algorithm. Moreover, these constants are sometimes difficult to compute and one often has to resort to some of their conservative estimates, which may slow down the algorithms dramatically.

Our work is motivated by the following three closely related questions: i) in view of good performance of the aforementioned bundle-type methods for non-smooth CP, should we use them for solving smooth CP problems as well? ii) could we achieve the optimal rate of convergence for solving smooth CP problems by using bundle-type methods? iii) given that first-order information of f is obtained via an oracle, should the optimization algorithms really need to know any smoothness information, such as, whether a problem is smooth, or the size of the Lipschitz constants L and $\mathcal{M}$?

Towards answering these questions, we study a class of CP problems of the form (1), where $f(\cdot)$ satisfies

$$f(y) - f(x) - \langle f'(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2 + M\|y - x\|, \quad \forall x, y \in X, \tag{3}$$

for some $L, M \geq 0$ and $f'(x) \in \partial f(x)$. Clearly, this class of problems covers non-smooth (with $L = 0$ and $M = 2\mathcal{M}$), smooth (with $L > 0$ and $M = 0$), and composite (with $L > 0$ and $M > 0$) CP problems [15]. Our major contribution consists of the following aspects. Firstly, we present a new bundle-type algorithm, namely: the accelerated bundle-level (ABL) method, and show that it is optimal for solving both non-smooth and smooth CP problems. Hence, we substantially improve the rate of convergence of bundle-type methods, when applied for solving smooth CP problems, from $\mathcal{O}(1/\sqrt{t})$ to $\mathcal{O}(1/t^2)$. Observe that this algorithm is different from Nesterov's methods since it works with a series of non-smooth approximations to the objective function $f(\cdot)$. Moreover, we show that the ABL method is uniformly optimal for minimizing smooth and non-smooth CP problems, in the sense that it does not require any smoothness information, such as whether a problem is smooth, or the size of the Lipschitz constants. To the best of our knowledge, this is the first time that uniformly optimal algorithms of this type have been proposed in the literature.

Secondly, one apparent problem for the ABL method is that, as the algorithm proceeds, its subproblems become more difficult to solve. As a result, each step of the ABL method becomes computationally more and more expensive. To remedy this issue, we present another uniformly optimal method for solving smooth and non-smooth CP problems, namely: the accelerated prox-level (APL) method, and demonstrate that one can actually take full control of its iteration costs. Moreover, the non-Euclidean prox-functions can be employed in the APL method in order to make use of the geometry of the feasible set X to obtain (nearly) dimension-independent rate of convergence. We also show that the APL method can be easily extended for solving strongly convex problems, sometimes with little modifications. Finally, our preliminary numerical experiments indicate that the APL method can significantly outperform the existing optimal methods for solving smooth CP problems, especially when the Lipschitz constant L is big and/or the desired accuracy is high.

The paper is organized as follows. In Section 2, we give a brief review of a few bundle-type methods, including the bundle-level method. We present the ABL method and the APL method, respectively, in Sections 3 and 4, and show that they are uniformly optimal for solving smooth and non-smooth CP problems. We then investigate in Section 5 how to extend the APL method for solving strongly convex CP problems. In Section 6, we present some numerical results for our methods. Section 7 is devoted to proving the main results of this paper. Finally, some concluding remarks are made in Section 8.

2 Review of bundle-level methods

In this section, we provide a review of bundle-level methods. As discussed in Section 1, these methods were designed for solving general nonsmooth CP problems of the form (1), where $f(\cdot)$ satisfies (3) with $L = 0$ and $M > 0$. Note that our review is by no means exhaustive due to a large body of literature existing in this area.

Given a sequence of search points $x_1, x_2, \ldots, x_t \in X$, an important construct, namely, the cutting plane model, of the objective function of problem (1) is given by

$$m_t(x) := \max \{h(x_i, x) : 1 \le i \le t\}, \tag{4}$$

where

$$h(z, x) := f(z) + \langle f'(z), x - z \rangle. \tag{5}$$

We call each $h(x_i, x)$, $i \ge 1$, a *component of the cutting plane model* $m_t(x)$.

In the classic cutting plane method [6, 10], one updates the search points by

$$x_{t+1} \in \mathrm{Argmin}_{x \in X} m_t(x).$$

This scheme converges slowly, both theoretically and practically [24]. Some significant progresses [11, 12, 18, 21] were made under the name of bundle methods. In these methods, a prox-term is introduced into the objective function of the above subproblem and hence the search points are updated by

$$x_{t+1} \in \mathrm{Argmin}_{x \in X} \left\{ m_t(x) + \frac{\rho_t}{2} \|x - x_t^+\|^2 \right\},$$

where the current prox-center x_t^+ is a certain point from $\{x_1, \ldots, x_t\}$ and ρ_t is the current penalty. Moreover, the prox-center for the next iterate, i.e., x_{t+1}^+, will be set to x_{t+1} if $f(x_{t+1})$ is sufficiently

smaller than $f(x_t)$. Otherwise, x_{t+1}^+ will be the same as x_t^+. The penalty ρ_t reduces the influence of the model m_t's inaccuracy and hence the instability of the algorithm. Note, however, that the determination of the penalty ρ_t usually requires certain on line adjustments or line-search. In the closely related trust-region technique [28, 20], the prox-term is put into the constraints of the subproblem instead of its objective function and the search points are then updated according to

$$x_{t+1} \in \text{Argmin}_{x \in X} \left\{ m_t(x) : \|x - x_t^+\|^2 \leq R_t \right\}.$$

However, this approach also encounters similar difficulties for determining the size of R_t.

In a seminal work [19], Lemaréchal, Nemirovskii and Nesterov introduced the idea of incorporating level sets into bundle-type methods. The basic version of their bundle-level method is described as follows:

- let f^t be the best objective value found so far and compute a lower bound on f^* by

$$f_t = \min_{x \in X} m_t(x); \tag{6}$$

- set $l_t = \lambda f^t + (1 - \lambda) f_t$ for some $\lambda \in (0, 1)$;
- update the search point x_t by

$$x_{t+1} \in \text{Argmin}_{x \in X} \left\{ \|x - x_t\|^2 : m_t(x) \leq l_t \right\}. \tag{7}$$

It is shown in [19] that the above scheme can find *an ϵ-solution of (1)*, i.e., a point $\bar{x} \in X$ s.t. $f(\bar{x}) - f^* \leq \epsilon$ in at most

$$\mathcal{O}(1) C(\lambda) \frac{M^2 D_X^2}{\epsilon^2}, \tag{8}$$

steps, where $C(\lambda)$ is a constant depending on λ and

$$D_X := \max_{x,y \in X} \|x - y\|. \tag{9}$$

Moreover, it turns out that the level sets give a stable description about the objective function $f(\cdot)$ and, as a consequence, very good practical performance [19, 4, 17] has been observed for the bundle-level method.

In the bundle-level method described above, the number of components appearing the cutting plane model $m_t(x)$ increases linearly with t. Hence, subproblems (6) and (7) become more and more difficult to solve as the algorithm proceeds. To address this issue, Kiwiel [13] presented novel rules of updating the prox-center, the bundle and the level, so as to eliminate the requirement of using full memory in $m_t(x)$. Similar ideas were used by Ben-tal and Nemirovski [4, 3] in their Truncated Proximal Bundle-level method and Non-Euclidean Restricted Memory Level Method (NERML). They show that the rates of convergence of these restricted memory versions of bundle-level method are still in the same order of magnitude as the one stated in (8).

3 The accelerated bundle-level method

In this section, we present a substantially enhanced version of the bundle-level method, namely: the accelerated bundle-level (ABL) method. We demonstrate that the ABL method can achieve the optimal rate of convergence for solving not only nonsmooth, but also smooth CP problems without requiring any smoothness information.

We first state a simple but crucial result which motivates the development of the ABL method. Note that this result originates from Lemma 5 of [15], in the context of studying a generalized version of Nesterov's method.

Lemma 1 *Let $(x, y, z) \in X \times X \times X$ be given. Suppose that, for some $\alpha \in [0, 1]$, the triple (x^+, y^+, z^+) satisfies the following condition:*

$$\begin{aligned} z^+ &= (1-\alpha)y + \alpha x, && (10)\\ h(z^+, x^+) &\le l, && (11)\\ y^+ &= \alpha x^+ + (1-\alpha)y, && (12) \end{aligned}$$

where $h(\cdot,\cdot)$ is defined in (5). Then,

$$f(y^+) \le (1-\alpha)f(y) + \alpha\, l + \frac{L\alpha^2}{2}\|x^+ - x\|^2 + M\alpha\|x^+ - x\|. \quad (13)$$

Proof. It can be easily seen from (10) and (12) that $y^+ - z^+ = \alpha(x^+ - x)$. Using this observation, (3), (5), (10), (11) and the convexity of f, we have

$$\begin{aligned} f(y^+) &\le h(z^+, y^+) + \frac{L}{2}\|y^+ - z^+\|^2 + M\|y^+ - z^+\| \\ &= (1-\alpha)h(z^+, y) + \alpha h(z^+, x^+) + \frac{L}{2}\|y^+ - z^+\|^2 + M\|y^+ - z^+\| \\ &= (1-\alpha)h(z^+, y) + \alpha h(z^+, x^+) + \frac{L\alpha^2}{2}\|x^+ - x\|^2 + M\alpha\|x^+ - x\| \\ &\le (1-\alpha)f(y) + \alpha h(z^+, x^+) + \frac{L\alpha^2}{2}\|x^+ - x\|^2 + M\alpha\|x^+ - x\| \\ &\le (1-\alpha)f(y) + \alpha l + \frac{L\alpha^2}{2}\|x^+ - x\|^2 + M\alpha\|x^+ - x\|. \end{aligned}$$

■

Observe that in Lemma 1, the function $h(z^+, \cdot)$ and the scaler l in (11), respectively, can be viewed as a simple linear model function and a level. Clearly, $h(z^+, \cdot)$ can be replaced by a more complicated cutting plane model in the form of (4) as long as it contains $h(z^+, \cdot)$ as a component. In view of Lemma 1, if $\alpha \in (0, 1]$, the level l and the distance $\|x^+ - x\|$ can be properly controlled, then the function value $f(y^+)$ will be sufficiently decreased from $f(y)$. Since one would like to control the distance $\|x^+ - x\|$ when searching for a new point x^+, the point x can be viewed as the current prox-center (c.f. (7)).

These observations motivate us to modify the bundle-level method as follows. Firstly, we will define three different (but related) search sequences, with one to build the cutting plane model for obtaining the lower bounds on f^*, one to act as the prox-centers and one to compute the best

objective values, i.e., the upper bounds on f^*. Secondly, to control the "closeness" of the prox-centers, we will group the steps of the algorithm into subsequent segments, so that the prox-centers in each segment will be "close" enough to each other (c.f. Lemma 7). Note that the latter idea has been used in the analysis of the bundle-level method, but not in the algorithm itself [19, 4].

More specifically, each step t, $t = 0, 1, \ldots$, of segment s, $s = 1, 2, \ldots$, of the ABL method updates three intertwined search points, namely, $x^l_{s,t}$, $x_{s,t}$ and $x^u_{s,t}$, which, respectively, denote the search point to compute a lower bound of f^*, the prox-center and an upper bound of f^*. Given an initial point $p_0 \in X$, we start the algorithm by setting $x^l_{1,0} = p_0$ and computing a lower bound by

$$\mathrm{lb}_{1,0} = \min_{x \in X} h(x^l_{1,0}, x), \tag{14}$$

where $h(\cdot, \cdot)$ is defined in (5). Moreover, we choose an arbitrary optimal solution of (14) as $x^u_{1,0}$ and set [1]

$$\begin{aligned} \mathrm{ub}_{1,0} &= f(x^u_{1,0}), & (15)\\ \Delta_{1,0} &= \mathrm{ub}_{1,0} - \mathrm{lb}_{1,0}. & (16) \end{aligned}$$

Observe that by (3), (5), (14), (15) and (16), we have

$$\begin{aligned} \Delta_{1,0} &= f(x^u_{1,0}) - \left[f(x^l_{1,0}) + \langle f'(x^l_{1,0}), x^u_{1,0} - x^l_{1,0} \rangle \right] \\ &\le \frac{L}{2} \|x^u_{1,0} - x^l_{1,0}\|^2 + M \|x^u_{1,0} - x^l_{1,0}\| \le \frac{L D_X^2}{2} + M D_X, \end{aligned} \tag{17}$$

where D_X is defined in (9). Finally, let $x_{1,0} \in X$ be arbitrarily chosen, say, $x_{1,0} = p_0$.

At step t, $t = 1, 2, \ldots$, of segment s, $s = 1, 2, \ldots$, we already have the triple $(x^l_{s,t-1}, x_{s,t-1}, x^u_{s,t-1})$ and the bounds $(\mathrm{lb}_{s,t-1}, \mathrm{ub}_{s,t-1})$ in our disposal. We first determine if a new segment starts. In particular, if the gap between $\mathrm{ub}_{s,t-1}$ and $\mathrm{lb}_{s,t-1}$ has been sufficiently decreased, i.e.,

$$\Delta_{s,t-1} := \mathrm{ub}_{s,t-1} - \mathrm{lb}_{s,t-1} < \lambda \Delta_{s,0}, \tag{18}$$

then we set

$$\begin{aligned} (x^l_{s+1,0}, x_{s+1,0}, x^u_{s+1,0}) &= (x^l_{s,t-1}, x_{s,t-1}, x^u_{s,t-1}), & (19)\\ \mathrm{lb}_{s+1,0} &= \mathrm{lb}_{s,t-1} & (20)\\ \mathrm{ub}_{s+1,0} &= \mathrm{ub}_{s,t-1} & (21)\\ \Delta_{s+1,0} &= \Delta_{s,t-1} & (22)\\ T_s &= t - 1, & (23) \end{aligned}$$

and pass to segment $s + 1$, where T_s is used to count the number of steps performed in segment s. Otherwise, if condition (18) is not satisfied, we update $(x^l_{s,t-1}, x_{s,t-1}, x^u_{s,t-1})$ into $(x^l_{s,t}, x_{s,t}, x^u_{s,t})$ by going through the following procedure.

[1] Essentially, the point $x^u_{1,0} \in X$ can be arbitrarily chosen, since the initial gap $\Delta_{1,0}$ only logarithmically affects the rate of convergence of the ABL method, see Theorem 3.

- Set $x^l_{s,t} = (1-\alpha_t)\, x^u_{s,t-1} + \alpha_t\, x_{s,t-1}$ for certain given stepsize $\alpha_t \in (0,1]$, compute $f(x^l_{s,t})$, $f'(x^l_{s,t})$ and build the model

$$m_{s,t}(x) := \max\left\{h(x^l_{i,j}, x) : j = 0, 1 \ldots, T_i - 1, \forall\, i = 1, \ldots, s-1; j = 0, 1, \ldots, t \text{ for } i = s\right\}, \tag{24}$$

where $h(\cdot,\cdot)$ is defined in (5). Note that the model $m_{s,t}(x)$ is defined as as the maximum of all the linear approximations to the function $f(\cdot)$ at the points $x^l_{i,j}$ that have been generated so far. We then update the lower bound $\mathrm{lb}_{s,t}$ by

$$\mathrm{lb}_{s,t} = \min_{x\in X} m_{s,t}(x). \tag{25}$$

- Compute the level $l_{s,t} = \lambda\, \mathrm{lb}_{s,t} + (1-\lambda)\, \mathrm{ub}_{s,t-1}$ for some $\lambda \in (0,1)$ and set

$$x_{s,t} = \mathrm{argmin}_x \left\{\|x - x_{s,t-1}\|^2 : x \in X, m_{s,t}(x) \le l_{s,t}\right\}. \tag{26}$$

- Choose $x^u_{s,t} \in X$ such that

$$f(x^u_{s,t}) \le \min\left\{\mathrm{ub}_{s,t-1}, f(\alpha_t x_t + (1-\alpha_t) x^u_{t-1})\right\}. \tag{27}$$

In particular, denoting $\tilde{x}^u_{s,t} \equiv \alpha_t x_{s,t} + (1-\alpha_t) x^u_{s,t-1}$, we can set $x^u_{s,t} = \tilde{x}^u_{s,t}$ if $f(\tilde{x}^u_{s,t}) \le \mathrm{ub}_{s,t-1}$. Otherwise, we set $x^u_{s,t} = x^u_{s,t-1}$.

- Update the upper bound $\mathrm{ub}_{s,t}$ of f^* by

$$\mathrm{ub}_{s,t} = f(x^u_{s,t}) \tag{28}$$

and set $\Delta_{s,t} := \mathrm{ub}_{s,t} - \mathrm{lb}_{s,t}$.

Now let us summarize the accelerated bundle-level method as follows.

The Accelerated Bundle-Level (ABL) method:

Input: $\epsilon > 0$, $\lambda \in (0,1)$, $p_0 \in X$ and $\alpha_t \in (0,1]$, $t = 1, 2, \ldots$.

Initialization:
Set $x^l_{1,0} = p_0$, compute $f(x^l_{1,0})$, $f'(x^l_{1,0})$ and $\mathrm{lb}_{1,0} := \min_{x\in X} f(x^l_{1,0}) + \langle f'(x^l_{1,0}), x - x^l_{1,0}\rangle$.
Let $x^u_{1,0} \in \mathrm{Argmin}_{x\in X} f(x^l_{1,0}) + \langle f'(x^l_{1,0}), x - x^l_{1,0}\rangle$ and $\mathrm{ub}_0 = f(x^u_{1,0})$.
Let $x_{1,0} \in X$ be arbitrarily chosen, say $x_{1,0} = p_0$. Also let $\Delta_{1,0} = \mathrm{ub}_{1,0} - \mathrm{lb}_{1,0}$.

For $s = 1, 2, \ldots$

For $t = 1, 2, \ldots$

If $\Delta_{s,t-1} \le \epsilon$, **Terminate** the algorithm.
Else if $\Delta_{s,t-1} < \lambda \Delta_{s,0}$,
Set $(x^l_{s+1,0}, x_{s+1,0}, x^u_{s+1,0}) = (x^l_{s,t-1}, x_{s,t-1}, x^u_{s,t-1})$.
Set $\mathrm{lb}_{s+1,0} = \mathrm{lb}_{s,t-1}$, $\mathrm{ub}_{s+1,0} = \mathrm{ub}_{s,t-1}$ and $\Delta_{s+1,0} = \Delta_{s,t-1}$.
Set $T_s = t - 1$ and pass to segment $s + 1$.

End if

Set $x^l_{s,t} = (1-\alpha_t)\, x^u_{s,t-1} + \alpha_t\, x_{s,t-1}$.

Compute $f(x^l_{s,t})$, $f'(x^l_{s,t}) \in \partial f(x^l_{s,t})$ and $\mathrm{lb}_{s,t} = \min\limits_{x\in X} m_{s,t}(x)$.

Compute the level $l_{s,t} = \lambda\, \mathrm{lb}_{s,t} + (1-\lambda)\, \mathrm{ub}_{s,t-1}$ and set
$x_{s,t} = \arg\min\limits_{x\in X} \big\{\|x - x_{s,t-1}\|^2 : m_{s,t}(x) \le l_{s,t}\big\}$.

Choose $x^u_{s,t} \in X$ such that $f(x^u_{s,t}) \le \min\big\{\mathrm{ub}_{s,t-1}, f(\alpha_t\, x_t + (1-\alpha_t) x^u_{t-1})\big\}$.

Set $\mathrm{ub}_{s,t} = f(x^u_{s,t})$ and $\Delta_{s,t} = \mathrm{ub}_{s,t} - \mathrm{lb}_{s,t}$.

End for

End for

Observe that by (21), (27) and (28), we have

$$\mathrm{ub}_{1,0} \ge \mathrm{ub}_{1,1} \ge \ldots \ge \mathrm{ub}_{1,T_1-1} \ge \mathrm{ub}_{2,0} \ge \mathrm{ub}_{2,1} \ge \ldots \ge \mathrm{ub}_{2,T_2-1} \ge \ldots \ge f^*. \tag{29}$$

Moreover, by (5), (24) and the convexity of f, we obtain, $\forall\, x \in X$,

$$m_{1,0}(x) \le m_{1,1}(x) \le \ldots \le m_{1,T_1-1}(x) \le m_{2,0}(x) \le m_{2,1}(x) \le \ldots \le m_{2,T_2-1}(x) \le \ldots \le f(x), \tag{30}$$

which, in view of (25), then implies that

$$\mathrm{lb}_{1,0} \le \mathrm{lb}_{1,1} \le \ldots \le \mathrm{lb}_{1,T_1-1} \le \mathrm{lb}_{2,0} \le \mathrm{lb}_{2,1} \le \ldots \le \mathrm{lb}_{2,T_2-1} \le \ldots \le f^*. \tag{31}$$

Combining (29) and (31), we have

$$\Delta_{1,0} \ge \Delta_{1,1} \ge \ldots \ge \Delta_{1,T_1-1} \ge \Delta_{2,0} \ge \Delta_{2,1} \ge \ldots \ge \Delta_{2,T_2-1} \ge \ldots \ge 0. \tag{32}$$

Note however that relation (32) does not necessarily imply that the sequence $\{\Delta_{s,t}\}$ converges to zero. To guarantee the convergence of the above scheme of the ABL method, we need to appropriately specify the stepsizes α_t, $t \ge 1$. More specifically, denoting

$$\Gamma_t := \begin{cases} 1 & t = 1 \\ (1-\lambda\,\alpha_t)\Gamma_{t-1} & t \ge 2 \end{cases}, \tag{33}$$

and assuming that the stepsizes $\alpha_t \in (0,1]$, $t \ge 1$, are chosen such that

$$\frac{\alpha_t^2}{\Gamma_t} \le C_1, \quad \Gamma_t \le \frac{C_2}{t^2} \quad \text{and} \quad \Gamma_t \left[\sum_{\tau=1}^{t} \left(\frac{\alpha_\tau}{\Gamma_\tau}\right)^2\right]^{\frac{1}{2}} \le \frac{C_3}{\sqrt{t}}, \quad \forall\, t \ge 1 \tag{34}$$

for some $C_1, C_2, C_3 \in \Re_{++}$, we will show in Theorem 3 that the ABL method can actually achieve the optimal rate of convergence for solving (1). The following result, whose proof can be found in Section 7.4, states two different ways of choosing the stepsizes α_t, $t \ge 1$, so that condition (34) holds. It is worth noting that none of these stepsize policies require the knowledge of L, M or D_X.

Lemma 2 *a) if* $0 < \lambda \leq \frac{1}{2}$ *and* α_t, $t \geq 1$, *are given by*

$$\alpha_t = \frac{2}{\lambda(t+3)}, \tag{35}$$

then we have $\alpha_t \in (0, 1]$ *for any* $t \geq 1$. *Moreover, condition (34) is satisfied with*

$$C_1 = \frac{2}{3\lambda^2}, \quad C_2 = 6 \quad \textit{and} \quad C_3 = \frac{1}{3\sqrt{3}\,\lambda}; \tag{36}$$

b) if α_t, $t \geq 1$, *are recursively defined by*

$$\alpha_1 = \Gamma_1 = 1, \quad \Gamma_t = \alpha_t^2 = (1 - \lambda\alpha_t)\Gamma_{t-1}, \quad \forall\, t \geq 2, \tag{37}$$

then we have $\alpha_t \in (0, 1)$ *for any* $t \geq 1$. *Moreover, condition (34) is satisfied with*

$$C_1 = \frac{4}{\lambda^2}, \quad C_2 = 1 \quad \textit{and} \quad C_3 = \frac{4}{\sqrt{3\lambda}}. \tag{38}$$

We are now ready to describe the main convergence properties of the above ABL method.

Theorem 3 *Suppose that the stepsizes* $\alpha_t \in (0, 1]$, $t = 1, 2, \ldots$, *are chosen such that condition (34) holds. Then, the total number of steps performed by the ABL method before termination does not exceed* $\mathcal{N}(\epsilon) + \mathcal{S}(\epsilon)$, *where*

$$\mathcal{N}(\epsilon) := \frac{1}{1 - \lambda^{\frac{1}{2}}} \left(\frac{3C_1C_2LD_X^2}{2\lambda\,\epsilon} \right)^{\frac{1}{2}} + \frac{1}{1 - \lambda^2} \left(\frac{3C_3MD_X}{\lambda\,\epsilon} \right)^2, \tag{39}$$

$$\mathcal{S}(\epsilon) := \left[1 + \left(\frac{3C_2}{\lambda} \right)^{\frac{1}{2}} \right] \left\lceil \log_{\frac{1}{\lambda}} \left(\frac{LD_X^2}{2\epsilon} + \frac{MD_X}{\epsilon} \right) \right\rceil, \tag{40}$$

D_X *is defined in (9),* L *and* M *are given by (3).*

We now add a few remarks about Theorem 3. First, suppose that the ABL method terminates at step t of segment s for some $t \geq 1$ and $s \geq 1$. Then by (28), (18) and (31), we have

$$f(x_{s,t-1}^u) - f^* \leq \mathrm{ub}_{s,t-1} - \mathrm{lb}_{s,t-1} \leq \Delta_{s,t-1} \leq \epsilon,$$

which implies that $x_{s,t-1}^u$ is an ϵ-solution of (1). Second, it can be easily seen that $S(\epsilon) = \mathcal{O}(N(\epsilon))$. By setting $M = 0$ (resp. $L = 0$) in (39), we obtain the optimal iteration-complexity bound for smooth (resp. non-smooth) convex optimization. Hence, the ABL method achieves uniformly the optimal rate of convergence for solving smooth, non-smooth and composite CP problems (see [15] for a discussion about the lower bounds on the rate of convergence for solving these classes of CP problems). Third, the total number of steps stated in Theorem 3 is estimated for solving composite CP problems with $L > 0$ and $M > 0$. It should be noted that, without modifying the algorithm, we can slightly improve the iteration-complexity bound by a constant factor if either M or L is set to 0.

4 The accelerated prox-level method

Observe that one problem for the ABL method is that, as the algorithm proceeds, its subproblems, namely (25) and (26), become more and more difficult to solve. In this section, we present a variant of the ABL method, namely, the accelerated prox-level (APL) method, which has the following two desirable properties: i) the computational complexity of its subproblems does not increase as the algorithm proceeds; ii) the non-Euclidean prox-functions can be employed to make use of the geometry of the feasible set X. Note that similar mechanisms have been developed for the bundle-type algorithms for non-smooth convex minimization. In particular, our algorithm can be viewed as an "accelerated" version of the Non-Euclidean Restricted Memory Level (NERML) method by Ben-Tal and Nemirovski [4, 3].

The steps of the APL method are divided into subsequent phases, corresponding to segments in the ABL method. Phase s, $s = 1, 2, \ldots$, is associated with a prox-center c_s and a level $\tilde{l}_s$ such that

- the values of $f(c_s)$ and $f'(c_s)$ are known when phase s starts;
- $\tilde{l}_s = \lambda \tilde{\mathrm{lb}}_s + (1-\lambda)\tilde{\mathrm{ub}}_s$, where $\tilde{\mathrm{ub}}_s$ and $\tilde{\mathrm{lb}}_s$, respectively, are the smallest objective value and the largest lower bound on f^* found when phase s starts, and λ is a parameter of the algorithm.

Note that the prox-center c_1 of the very first phase can be chosen arbitrarily.

Let $\omega : X \to \Re$ be a given differentiable and strongly convex function with modulus σ (e.g., $\omega(x) = \|x\|^2/2$). We define the prox-function at phase s of the APL method as

$$\omega_s(x) \equiv \omega(x) - [\omega(c_s) + \nabla\omega(c_s)^T(x - c_s)]. \tag{41}$$

The prox-function $\omega_s(\cdot, \cdot)$ is also called Bregman's distance, which was initially studied by Bregman [5] and later by many others (see [1, 2, 14, 29, 23] and references therein). Observe that $\omega_s(x)$ is also differentiable and strongly convex with modulus σ, i.e.,

$$\omega_s(y) \geq \omega_s(x) + \langle \nabla\omega_s(x), y - x\rangle + \frac{\sigma}{2}\|y - x\|^2. \tag{42}$$

Moreover, we have $\nabla\omega_s(c_s) = 0$ and hence that $c_s = \arg\min_{x\in X} \omega_s(x)$. More discussions on the determination of prox-functions for different feasible sets can be found, for example, in [4, 3, 23, 17].

We start the whole process by computing a valid lower bound on f^*. More specifically, given an initial point $p_0 \in X$, we set

$$\tilde{\mathrm{lb}}_1 = \min_{x\in X} \left\{ f(p_0) + \langle f'(p_0), x - p_0\rangle \right\}. \tag{43}$$

Moreover, let $\tilde{p}_0$ be an optimal solution of (43), we compute an upper bound on f^* by setting [2]

$$\tilde{\mathrm{ub}}_1 = f(\tilde{p}_0). \tag{44}$$

Let us denote $\tilde{\Delta}_s := \tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s$ and

$$\mathcal{D}^2_{\omega,X} := \max_{x,y\in X} \{\omega(y) - \omega(x) - \langle \nabla\omega(x), y - x\rangle\}. \tag{45}$$

[2] Essentially, the upper bound $\tilde{\mathrm{ub}}_1$ can be the function value of an arbitrary feasible point in X, since the initial gap $\tilde{\mathrm{ub}}_1 - \tilde{\mathrm{lb}}_1$ only logarithmically affects the rate of convergence of the APL method, see Theorem 5.

Then, similar to (17), we have

$$\begin{aligned}\tilde{\Delta}_1 &= f(\tilde{p}_0) - \big[f(p_0) + \langle f'(p_0), \tilde{p}_0 - p_0\rangle\big] \le \frac{L}{2}\|\tilde{p}_0 - p_0\|^2 + M\|\tilde{p}_0 - p_0\| \\ &\le \frac{L\mathcal{D}_{\omega,X}^2}{\sigma} + \sqrt{\frac{2}{\sigma}}M\mathcal{D}_{\omega,X}, \end{aligned} \tag{46}$$

where the last inequality follows from (45) and the strong-convexity of ω.

Similarly to the ABL method, each step t, $t = 0, 1, \ldots$, of phase s, $s = 1, 2, \ldots$, updates three intertwined search points, namely, $(x_{s,t}^l, x_{s,t}, x_{s,t}^u)$. When generating $(x_{s,t}^l, x_{s,t}, x_{s,t}^u)$, we have already in our disposal the search points $(x_{s,t-1}^l, x_{s,t-1}, x_{s,t-1}^u)$, a valid lower bound $\mathrm{lb}_{s,t-1}$ on f^*, and a convex compact set $X_{s,t-1}$ in X, referred to as the *localizer*, such that

$$\mathcal{L}_s := \Big\{x \in X : f(x) \le \tilde{l}_s\Big\} \subseteq X_{s,t-1}. \tag{47}$$

In the beginning of phase s, we set $x_{s,0} = c_s$ and $\mathrm{lb}_{s,0} = \tilde{\mathrm{lb}}_s$. $X_{s,0}$ can be chosen as any set satisfying (47), say, X. It is worth noting that other choices of $X_{s,0}$ might be computationally more advantageous (see Section 6). Note also that the points $x_{s,0}^u \in X$ can be chosen arbitrarily, say c_s.

To update $(x_{s,t-1}^l, x_{s,t-1}, x_{s,t-1}^u, X_{s,t-1})$ into $(x_{s,t}^l, x_{s,t}, x_{s,t}^u, X_{s,t})$, the APL method first updates the lower bound on f^* by the following two steps:

1) Set $x_{s,t}^l = (1 - \alpha_t)x_{s,t-1}^u + \alpha_t x_{s,t-1}$ for certain given stepsize $\alpha_t \in (0, 1]$;

2) Solve the auxiliary problem

$$h^* := \min_{x \in X_{s,t-1}} h(x_{s,t}^l, x), \tag{48}$$

where $h(\cdot,\cdot)$ is defined in (5). Observe that the quantity

$$\hat{\mathrm{lb}} := \min\{h^*, \tilde{l}_s\} \tag{49}$$

is a lower bound on f^*. Indeed, by (47), we have $f(x) > \tilde{l}_s$ for any $x \in X \backslash X_{s,t-1}$. Moreover, by (5), (48) and the convexity of f, we have

$$h^* \le h(x_{s,t}^l, x) \le f(x), \quad \forall\, x \in X_{s,t-1}.$$

Hence, $\hat{\mathrm{lb}} := \min\{h^*, \tilde{l}_s\}$ underestimates $f(x)$ everywhere on X. Clearly, the quantity

$$\mathrm{lb}_{s,t} := \max\{\mathrm{lb}_{s,t-1}, \hat{\mathrm{lb}}\} \tag{50}$$

is also a lower bound on f^*.

Depending on the value of $\mathrm{lb}_{s,t}$ computed by (50), we consider the following two cases.

Case I: Significant progress on the lower bound. If

$$\mathrm{lb}_{s,t} \ge \tilde{l}_s - \theta(\tilde{l}_s - \tilde{\mathrm{lb}}_s), \tag{51}$$

where $\theta \in (0, 1)$ is a parameter of the APL method, we terminate phase s, set

$$\tilde{\mathrm{lb}}_{s+1} = \mathrm{lb}_{s,t}, \qquad \tilde{\mathrm{ub}}_{s+1} = \min\Big\{\tilde{\mathrm{ub}}_s, \min_{0 \le \tau \le t-1} f(x_{s,\tau}^u)\Big\}$$

and pass to phase $s + 1$. The prox-center $c_{s+1} \in X$ can be chosen arbitrarily, for example, $c_{s+1} = x_{s,t-1}^u$;

Case II: No significant progress on the lower bound, i.e., condition (51) is not satisfied. We first compute

$$x_{s,t} \equiv \operatorname{argmin}_x \left\{ \omega_s(x) : x \in X_{s,t-1}, h(x_t^l, x) \leq \tilde{l}_s \right\}. \tag{52}$$

Note that problem (52) must be feasible. Otherwise, the quantity h^* computed in (48) would be $+\infty$ and hence we should have $\mathrm{lb}_{s,t} = \hat{\mathrm{lb}} = \tilde{l}_s$. Therefore, Case I, rather than Case II, would have occurred since condition (51) was satisfied. After computing $x_{s,t}$, we choose $x_{s,t}^u \in X$ such that

$$f(x_{s,t}^u) \leq f(\alpha_t x_{s,t} + (1-\alpha_t) x_{s,t-1}^u). \tag{53}$$

The simplest option is to set $x_{s,t}^u = \alpha_t x_{s,t} + (1-\alpha_t) x_{s,t-1}^u$. Then, we check if the progress on the the objective value is significant. More specifically, we consider two subcases.

Case II.a): Significant progress on upper bound. If

$$f(x_{s,t}^u) - \tilde{l}_s \leq \theta(\tilde{\mathrm{ub}}_s - \tilde{l}_s), \tag{54}$$

we terminate phase s, set

$$\tilde{\mathrm{lb}}_{s+1} = \mathrm{lb}_{s,t}, \qquad \tilde{\mathrm{ub}}_{s+1} = \min \left\{ \tilde{\mathrm{ub}}_s, \min_{0 \leq \tau \leq t} f(x_{s,\tau}^u) \right\}$$

and pass to phase $s+1$.

Case II.b): No significant progress on upper bound. If condition (54) does not hold, we continue phase s and update the localizer $X_{s,t}$ as an arbitrary convex compact set such that $\underline{X}_{s,t} \subseteq X_{s,t} \subseteq \overline{X}_{s,t}$, where

$$\underline{X}_{s,t} \equiv \left\{ x \in X_{s,t-1} : h(x_{s,t}^l, x) \leq \tilde{l}_s \right\} \quad \text{and} \quad \overline{X}_{s,t} \equiv \left\{ x \in X : \langle \nabla \omega_s(x_{s,t}), x - x_{s,t} \rangle \geq 0 \right\}. \tag{55}$$

Note that in Case II.b), problem (48) is feasible and hence that the set $\underline{X}_{s,t}$ is nonempty. Moreover, by the optimality condition of (52), we have $\langle \nabla \omega_s(x_{s,t}), x - x_{s,t} \rangle \geq 0$ for any $x \in \underline{X}_{s,t}$, which then implies that $\underline{X}_{s,t} \subseteq \overline{X}_{s,t}$. Finally, let $\mathcal{L}_s$ be defined in (47) and suppose that $\mathcal{L}_s \subseteq X_{s,t-1}$. Then, by (55) and the fact that $h(x_{s,t}^l, x) \leq f(x)$ for any $x \in X$, we have $\mathcal{L}_s \subseteq \underline{X}_{s,t} \subseteq X_{s,t}$. Hence, by induction, condition (47) will be guaranteed given that $\mathcal{L}_s \subseteq X_{s,0}$. Therefore, we can always choose any $X_{s,t}$ satisfying $\underline{X}_{s,t} \subseteq X_{s,t} \subseteq \overline{X}_{s,t}$ (the simplest way is to set $X_{s,t} = \underline{X}_{s,t}$ or $X_t = \overline{X}_{s,t}$). Note that, while the number of constraints defining $\underline{X}_{s,t}$ increases with t, the set $\overline{X}_{s,t}$ has only one more constraint than X. By choosing $X_{s,t}$ in between these two extremes, we can control the number of constraints for subproblems (48) and (52).

We now summarize the APL method as follows.

The Accelerated Prox-Level (APL) Algorithm:

Input: $\epsilon > 0$, $\lambda \in (0,1)$, $\theta \in (0,1)$, $\alpha_t \in (0,1]$ for any $t \geq 1$, initial point $p_0 \in X$.

Initialization:
Compute $f(p_0)$, $f'(p_0) \in \partial f(p_0)$, $\tilde{p}_0 \in \mathrm{Argmin}_{x \in X} f(p_0) + \langle f'(p_0), x - p_0 \rangle$.
Set $\tilde{\mathrm{lb}}_1 = f(p_0) + \langle f'(p_0), \tilde{p}_0 - p_0 \rangle$ and $\tilde{\mathrm{ub}}_1 = f(\tilde{p}_0)$.
Choose the initial prox-center $c_1 \in X$ arbitrarily, e.g., $\tilde{p}_0$.

For $s = 1, 2, \ldots$ (start phase s),

If $\tilde{\text{ub}}_s - \tilde{\text{lb}}_s \leq \epsilon$, **terminate** the algorithm.

Set $\tilde{l}_s = \lambda\, \tilde{\text{lb}}_s + (1-\lambda)\, \tilde{\text{ub}}_s$.

Set $x_{s,0} = x^u_{s,0} = c_s$, $\text{lb}_{s,0} = \tilde{\text{lb}}_s$, choose $X_{s,0}$ s.t. (47) holds.

For $t = 1, 2, \ldots$ (start step t of phase s),

Set $x^l_{s,t} = (1-\alpha_t)x^u_{s,t-1} + \alpha_t x_{s,t-1}$.

Compute $f(x^l_{s,t})$, $f'(x^l_{s,t}) \in \partial f(x^l_{s,t})$, and $h^* = \min_{x \in X_{s,t-1}} h(x^l_{s,t}, x)$.

Set $\text{lb}_{s,t} = \max\Big\{\text{lb}_{s,t-1}, \min(\tilde{l}_s, h^*)\Big\}$.

If $\text{lb}_{s,t} \geq \tilde{l}_s - \theta(\tilde{l}_s - \tilde{\text{lb}}_s)$,

Set $\tilde{\text{lb}}_{s+1} = \text{lb}_{s,t}$, $\tilde{\text{ub}}_{s+1} = \min\Big\{\tilde{\text{ub}}_s, \min_{0 \leq \tau \leq t-1} f(x^u_\tau)\Big\}$,

choose $c_{s+1} \in X$ and pass to phase $s+1$.

Else

Compute $x_{s,t} = \text{argmin}_x \Big\{\omega_s(x) : x \in X_{s,t-1}, h(x^l_{s,t}, x) \leq \tilde{l}_s\Big\}$.

Find $x^u_{s,t} \in X$ such that $f(x^u_{s,t}) \leq f(\alpha_t x_{s,t} + (1-\alpha_t)x^u_{s,t-1})$.

If $f(x^u_{s,t}) - \tilde{l}_s \leq \theta(\tilde{\text{ub}}_s - \tilde{l}_s)$,

Set $\tilde{\text{lb}}_{s+1} = \text{lb}_{s,t}$, $\tilde{\text{ub}}_{s+1} = \min\Big\{\tilde{\text{ub}}_s, \min_{0 \leq \tau \leq t} f(x^u_{s,\tau})\Big\}$,

Choose $c_{s+1} \in X$ and pass to phase $s+1$,

Else

Choose $X_{s,t}$ s.t. $\underline{X}_{s,t} \subseteq X_{s,t} \subseteq \overline{X}_{s,t}$ and pass to step $t+1$ of phase s.

End if

End if

End for (finish step t of phase s)

End for (finish phase s)

Similarly to the ABL method, we still need to properly specify the stepsizes $\{\alpha_t\}_{t \geq 1}$ in order to guarantee the optimal convergence of the above APL method. More specifically, denoting

$$\tilde{\Gamma}_t := \begin{cases} 1, & t = 1 \\ \tilde{\Gamma}_{t-1}(1-\alpha_t), & t \geq 2 \end{cases}, \tag{56}$$

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

$$\alpha_1 = 1, \quad \frac{\alpha_t^2}{\tilde{\Gamma}_t} \leq \tilde{C}_1, \quad \tilde{\Gamma}_t \leq \frac{\tilde{C}_2}{t^2} \quad \text{and} \quad \tilde{\Gamma}_t \left[\sum_{\tau=1}^{t} \left(\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\right)^2\right]^{\frac{1}{2}} \leq \frac{\tilde{C}_3}{\sqrt{t}}, \quad \forall\, t \geq 1, \tag{57}$$

for some $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \in \Re_{++}$. The following lemma, whose proof can be found in Section 7.4, states two different ways to specify the stepsizes α_t, $t \geq 1$. It is worth noting that none of these stepsize policies require the knowledge of L, M or $\mathcal{D}_{\omega,X}$.

Lemma 4 *a) if α_t, $t \geq 1$, are set to*

$$\alpha_t = \frac{2}{t+1}, \tag{58}$$

then condition (57) holds with $\tilde{C}_1 = 2$, $\tilde{C}_2 = 2$ and $\tilde{C}_3 = 2/\sqrt{3}$;

b) if α_t, $t \geq 1$, are computed recursively by

$$\alpha_1 = \tilde{\Gamma}_1 = 1, \quad \alpha_t^2 = (1-\alpha_t)\tilde{\Gamma}_{t-1} = \tilde{\Gamma}_t, \ \forall\, t \geq 2, \tag{59}$$

then we have $\alpha_t \in (0,1]$ for any $t \geq 2$. Moreover, condition (57) is satisfied with $\tilde{C}_1 = 1$, $\tilde{C}_2 = 4$ and $\tilde{C}_3 = 4/\sqrt{3}$.

We are now ready to describe the main convergence properties of the above APL method.

Theorem 5 *Suppose that $\alpha_t \in (0,1]$, $t = 1, 2, \ldots$, in the APL method are chosen such that condition (57) holds. Also denote*

$$q \equiv q(\theta, \lambda) := 1 - (1-\theta)\min\{\lambda, 1-\lambda\}. \tag{60}$$

Then, the total number of steps performed by the APL method before termination does not exceed $\tilde{\mathcal{N}}(\epsilon) + \tilde{\mathcal{S}}(\epsilon)$, where

$$\tilde{\mathcal{N}}(\epsilon) = \frac{1}{1-\sqrt{q}}\left(\frac{2\tilde{C}_1\tilde{C}_2 L\mathcal{D}_{\omega,X}^2}{\sigma\theta\lambda\,\epsilon}\right)^{\frac{1}{2}} + \frac{1}{1-q^2}\left(\frac{2\sqrt{2}\tilde{C}_3 M\mathcal{D}_{\omega,X}}{\sqrt{\sigma}\theta\lambda\,\epsilon}\right)^2, \tag{61}$$

$$\tilde{\mathcal{S}}(\epsilon) := 1 + \max\left\{0, \log\left(\frac{L\mathcal{D}_{\omega,X}^2}{\sigma\epsilon} + \sqrt{\frac{2}{\sigma}}\frac{M\mathcal{D}_{\omega,X}}{\epsilon}\right)\right\}. \tag{62}$$

and $\mathcal{D}_{\omega,X}$ is defined in (45).

Similar to the remarks made after Theorem 3, we can easily see from Theorem 5 that the APL method is also uniformly optimal for minimizing smooth, nonsmooth and composite CP problems. Moreover, the iteration-complexity bound in Theorem 5 can be slightly improved by assuming either M or L is set to 0 in (3).

5 The APL method for minimizing strongly convex problems

In this section, we still consider (1), but now the objective function f is strongly convex, i.e., for some $\mu > 0$:

$$f(y) - f(x) - \langle f'(x), y - x\rangle \geq \frac{\mu}{2}\|y-x\|^2, \quad \forall\, x, y \in X. \tag{63}$$

Our goal in this section is to show that the APL method in section 4, after certain modifications, can also achieve the optimal rate of convergence for minimizing strongly convex problems.

Throughout this section, we assume that the prox-function $\omega_s(x)$ (c.f. (41)) at stage s, $s \geq 1$, of the APL method is growing quadratically, i.e., there exists a constant $\mathcal{Q}$ such that $\omega_s(x) \leq \mathcal{Q}\|x - c_s\|^2/2$ for any $x \in X$. The smallest constant $\mathcal{Q}$ satisfying the previous relation is called the

quadratic growth constant. Without loss of generality, we assume that $\mathcal{Q} = 1$ for the prox-function $\omega_s(x)$ if it grows quadratically, i.e.,

$$\omega_s(x) \leq \frac{1}{2}\|x - c_s\|^2, \quad \forall x \in X. \tag{64}$$

Indeed, if $\mathcal{Q} \neq 1$, we can multiply the corresponding distance generating function ω by $1/\mathcal{Q}$ and the resulting prox-function will satisfy (64). More discussions on the quadratically growing prox-functions can be found, for example, in [9] and [7].

The following result states certain conditions under which the APL method is optimal for minimizing strongly convex functions.

Theorem 6 *Suppose that conditions (63) and (64) hold. If the prox-center c_s and the localizer $X_{s,t}$ of the APL method are chosen such that either one of the following two conditions holds:*

$$\textit{i)} \qquad f(c_s) = \mathrm{ub}_s, \forall s \geq 1 \quad \textit{and} \quad x^* \in X_{s,t}, \quad \forall s \geq 1, t \geq 1, \tag{65}$$

$$\textit{ii)} \qquad \underline{X}_{s,t} \subseteq X_{s,t} \subseteq \overline{X}_{s,t} \cap \left\{ x \in X : \|x - c_s\|^2 \leq \frac{2\tilde{\Delta}_s}{\mu} \right\}, \quad \forall s \geq 1, t \geq 1, \tag{66}$$

where $\tilde{\Delta}_s = \tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s$, $\underline{X}_{s,t}$ and $\overline{X}_{s,t}$ are defined in (55), then the total number of steps performed by the APL method applied to (1) can be bounded by

$$\tilde{\mathcal{S}}(\epsilon) \left(1 + \sqrt{\frac{2\tilde{C}_1\tilde{C}_2 L}{\sigma\theta\lambda\mu}} \right) + \frac{8\tilde{C}_3^2 M^2}{(1-q)\,\sigma\theta^2\lambda^2\mu\epsilon}, \tag{67}$$

where $\tilde{\mathcal{S}}(\epsilon)$ is given by (62).

We now add a few remarks on the results obtained in Theorem 6. First, it can be easily seen from (62) and (67) that, if f is a smooth convex function, i.e., $M = 0$, then the total number of steps performed by the APL method is bounded by

$$\mathcal{O}(1)\sqrt{\frac{L}{\mu}} \left[1 + \max\left(0, \log \frac{L D_{\omega,X}^2}{\sigma\epsilon} \right) \right].$$

Moreover, the total number of steps can be bounded by $\mathcal{O}(1)M^2/(\mu\epsilon)$ if f is a non-smooth convex function, i.e., $L = 0$. Therefore, Theorem 6 establishes the optimal rate of convergence for solving either smooth or non-smooth strongly convex problems.

Second, to ensure the first relation in condition (65), we should choose the prox-center c_s as the best solution found so far when phase s starts. However, the second relation in (65) can be satisfied only under certain specific assumptions, for example, when the optimal value of (1) is known and the initial lower bound lb_1 is set to $f(x^*)$. In this case, we have $\tilde{l}_s \geq f(x^*)$ and hence

$$h(x_{s,t}^l, x^*) \leq f(x^*) \leq \tilde{l}_s, \quad \forall\, t = 1, 2, \ldots,$$

which implies that $x^* \in \underline{X}_{s,t} \in X_{s,t}$. It is worth noting that, under the aforementioned circumstances, we do not need to modify the definition of $X_{s,t}$.

Third, while condition (65) can only be guaranteed in some special cases, we can always ensure condition (66) by incorporating an additional constraint, namely, $\|x - c_s\|^2 \leq 2\tilde{\Delta}_s^2/\mu$, into the definition of $X_{s,t}$. The basic idea is to shrink the feasible set whenever a new phase starts. Similar techniques have been used for solving strongly convex problems (c.f. [7]). Note however that, incorporating one more constraint will slightly increase the difficulty for solving subproblems (48) and (52). It should also be noted that, if condition (66) holds, assumption (64) on the quadratic growing prox-function can be relaxed, by properly modifying the definition of $\omega(\cdot)$ (c.f. [9]).

6 Implementation issues and numerical results

Our objective in this section is to discuss a few implementation issues and present the results of our preliminary numerical experiments.

We start by detailing a few implementation issues for the bundle-type methods presented in Section 3 and 4.

- *Solve the ABL subproblems (25) and (26).* If the set X is simple, one can use the interior point methods (IPM) to solve subproblems (25) and (26). Moreover, from our numerical experiments, choosing only n most recently generated components in $m_{s,t}(x)$ will not slow down the convergence of the ABL method.

- *Define the APL subproblems (48) and (52).* In our implementation, we set

$$X_{s,t} = \{x \in X : \langle \nabla \omega_s(x_t), x - x_{s,t} \rangle \geq 0\} \bigcap \mathcal{M}_{s,t}, \quad t \geq 0,$$

  where $\mathcal{M}_{s,t}$ denotes the polyhedron defined by the intersection of B half spaces of the form $\left\{x : h(x_{i,j}^l, x) \leq \tilde{l}_s\right\}$ which have been generated most recently and B is set to 10 in our experiments.

- *Solve the APL subproblems (48) and (52).* These problems can be solved, for example, by using interior point methods. Notice that the Lagrangian duals of these subproblems only have a very small number of variables and the first-order information of these Lagrangian duals can be easily computed if X is simple enough. We can then solve the Lagrangian duals of these subproblems by any efficient algorithms for lower-dimensional CP, such as the Ellipsoid algorithm (see [3] for similar strategies used in the NERML algorithm). It can also be easily verified that the Lagrangian duals of (48) and (52) are, respectively, non-smooth and smooth CP problems. In our implementation, the truncated ABL method (with at most n most recently generated components in $m_{s,t}(x)$) is used to solve the Lagrangian duals of these subproblems.

- *Determine other parameters.* We set $\lambda = 1/2$ and use the stepsizes in (35) (for its simplicity) in our implementation of the ABL method. Also we set $\lambda = 1/2$, $\theta = 1/2$, and use the stepsizes in (58) in our implementation of the APL method.

The main objective of our numerical experiments is to investigate the performance of APL method for solving smooth CP problems. For this purpose, we consider the quadratic programming (QP) problem of

$$\min_{\|x\| \leq 1} \|Ax - b\|^2, \tag{68}$$

where $A \in \Re^{m\times n}$ and $b \in \Re^m$. We compare the following four algorithms: i) the APL method with subproblems solved by Mosek [22], an efficient software package implementing the IPMs for linear and second-order cone programming; ii) the APL method with subproblems solved by the ABL method; iii) NERML [4] with subproblem solved by Mosek; iv) Nesterov's optimal method [27]. For the bundle-type methods (APL and NERML), we also compare their performance when different initial lower bounds, i.e., $-\infty$ or 0, are provided. All of these algorithms were implemented in MATLAB2007 under Windows Vista and the experiments were conducted in an INTEL 2.53 GHz labtop. It is worth noting that, while many algorithms, including some specialized algorithms (c.f. [8]), have been developed for solving (68), Nesterov's method and their variants are the only previously known optimal algorithms for solving (68). In this paper, we are not intending to conduct an exhaustive comparison on all different algorithms developed for solving (1), although it might be of great interest from the computational point of view.

Our first experiment is conducted for some worst-case QP instances provided by A. Nemirovski, which were constructed to demonstrate the worst-case convergence behavior of first-order algorithms for (68). The Lipschitz constants for these instances are relatively small (≈ 1.0). From our experiments, both the APL method and Nesterov's method exhibit similar sublinear convergence for solving these instances, see Table 6 for the number of steps performed the APL method (with subproblems solved by Mosek) for achieving a specified accuracy, i.e., the difference between the objective value and the optimal value. Hence, the APL method is not more advantageous over Nesterov's method for solving these worst-case instances.

Table 1: Number of steps performed by the APL method for bad instances

Acc.	Bdata 1 1036×1036	Bdata 2 2062×2062	Bdata 3 2062×2062	Bdata 4 2062×4124
$1.0e-4$	32	30	30	30
$1.0e-5$	126	113	118	114
$1.0e-6$	392	406	402	368
$1.0e-7$	980	1,178	1,224	1,114

In our second experiment, we assume that A and b of (68) are generated randomly, so that the optimal value of (68) is given by 0. The Lipschitz constants of these instances are much bigger (of order 10^6) and the initial errors (starting with $x = 0$) are of order 10^4 (see the description of instances LS1, LS2, LS3 in Table 6). The number of steps, CPU Time (in seconds) and actual accuracy were reported in columns $3-5$ and columns $6-8$ of Table 6, respectively, for target accuracy $1.0e-5$ and $1.0e-7$. The following observations can be made from these results. Firstly, the APL methods can significantly outperform Nesterov's method in terms of both the number of iterations and the computational time for solving these randomly generated instances. It is worth noting that the fact that all the algorithms were implemented in MATLAB seems to be in favor of Nesterov's algorithm, since the major cost of this method is the matrix-vector multiplication. We believe that the APL method can be made much faster with an implementation in C. Secondly, the supply of a good initial lower bound help the convergence of the APL and NERML algorithms, especially for the latter one. In particular, if the initial bound is set to $-\infty$, the NERML algorithm converges very slowly after 500 iterations and it cannot achieve accuracy smaller than $1.0e-5$ for solving these QP instances.

Table 2: Tests for randomly generated instances

LS1: $n = 4000$, $m = 2000$, $L = 2.00e6$ and $err_0 = 3.85e4$							
Alg.	LB	Iter.	Time	Acc.	Iter.	Time	Acc.
APL	0	70	59.85	$8.76e-6$	95	81.66	$8.24e-8$
(Mosek)	$-\infty$	190	154.20	$6.21e-6$	373	317.53	$8.30e-8$
APL	0	69	54.76	$9.08e-6$	156	122.59	$9.09e-8$
(ABL)	$-\infty$	175	185.35	$2.27e-6$	309	270.97	$7.45e-8$
NERML	0	142	125.13	$7.94e-6$	176	155.09	$9.95e-8$
(Mosek)	$-\infty$	500	419.30	$1.14e-4$	-	-	-
NEST	-	9,000	432.21	$9.79e-6$	30,400	1,467.15	$1.00e-7$
LS2: $n = 6000$, $m = 3000$, $L = 4.50e6$ and $err_0 = 3.85e4$							
Alg.	LB	Iter.	Time	Acc.	Iter.	Time	Acc.
APL	0	67	116.12	$8.31e-6$	92	159.45	$8.22e-8$
(Mosek)	$-\infty$	227	385.45	$3.64e-6$	399	677.50	$9.63e-8$
APL	0	91	84.19	$9.80e-6$	225	208.16	$9.28e-8$
(ABL)	$-\infty$	217	190.23	$6.21e-6$	384	336.54	$2.00e-8$
NERML	0	132	221.32	$9.40e-6$	167	279.77	$8.84e-8$
(Mosek)	$-\infty$	500	835.86	$3.02e-4$	-	-	-
NEST	-	12,700	1,304.81	$9.79e-6$	41,551	4,269.03	$1.00e-7$
LS3: $n = 8000$, $m = 4000$, $L = 8.0e6$ and $err_0 = 3.2e6$							
Alg.	LB	Iter.	Time	Acc.	Iter.	Time	Acc.
APL	0	66	191.41	$9.91e-6$	90	261.02	$8.59e-8$
(Mosek)	$-\infty$	218	631.14	$8.45e-6$	384	1,111.71	$8.10e-8$
APL	0	70	132.90	$9.44e-6$	152	288.57	$9.23e-8$
(ABL)	$-\infty$	217	446.69	$5.40e-6$	360	741.05	$6.46e-8$
NERML	0	134	395.50	$8.74e-6$	168	495.85	$9.26e-8$
(Mosek)	$-\infty$	500	1,403.62	$1.39e-4$	-	-	-
NEST	-	16,200	2,825.51	$9.80e-6$	53,746	9,374.02	$1.00e-7$

7 Convergence analysis

In this section, we provide the proofs of our main results presented in Sections 3, 4 and 5.

7.1 Convergence analysis for the ABL method

The goal of this subsection is to prove Theorem 3, which describes the main convergence properties of the ABL method. Before proving this result, we first need to show a few technical results.

The following lemma shows that the subproblems (26) in each segment of the ABL method have at least one common feasible point and hence that the prox-centers in each segment will be "close" enough to each other. This result resembles the corresponding one of the bundle-level method (see, for example, [4]).

Lemma 7 *For each segment s, $s \geq 1$, of the ABL method, the level sets given by*

$$\mathcal{L}_{s,t} := \{x \in X : m_{s,t}(x) \leq l_{s,t}\}, \quad t = 1, \ldots, T_s - 1 \tag{69}$$

have a point in common, where $m_{s,t}(\cdot)$ is defined in (25) and $l_{s,t} := \lambda \, \mathrm{lb}_{s,t} + (1-\lambda)\mathrm{ub}_{s,t-1}$. As a consequence, we have

$$\sum_{\tau=1}^{t} \|x_{s,\tau} - x_{s,\tau-1}\|^2 \leq D_X^2, \quad t = 1, \ldots, T_s - 1, \tag{70}$$

where T_s denotes the number of steps in segment s and D_X is defined in (9).

Proof. Denote $r = T_s - 1$ and let $u \in \mathrm{Argmin}_{x \in X} m_{s,r}(x)$. Note that by (18) and the definition of T_s, we have, for any $t = 1, \ldots, r$, $\Delta_{s,t} = \mathrm{ub}_{s,t} - \mathrm{lb}_{s,t} \geq \lambda \Delta_{s,0}$. Using this observation, (30) and (25), we conclude that, for any $t = 1, \ldots, r$,

$$m_{s,t}(u) \leq m_{s,r}(u) = \mathrm{lb}_{s,r} = \mathrm{ub}_{s,r} - \Delta_{s,r} \leq \mathrm{ub}_{s,r} - \lambda \Delta_{s,0} \leq \mathrm{ub}_{s,t-1} - \lambda \Delta_{s,0},$$

which, in view of the facts that $\Delta_{s,0} \geq \Delta_{s,t-1}$ and that $\mathrm{lb}_{s,t-1} \leq \mathrm{lb}_{s,t}$ for any $t = 1, 2, \ldots, r$, then implies that

$$m_{s,t}(u) \leq \mathrm{ub}_{s,t-1} - \lambda \Delta_{s,t-1} = (1-\lambda)\, \mathrm{ub}_{s,t-1} + \lambda \, \mathrm{lb}_{s,t-1} \leq (1-\lambda)\, \mathrm{ub}_{s,t-1} + \lambda \, \mathrm{lb}_{s,t} = l_{s,t}.$$

We have thus shown that $u \in \mathcal{L}_{s,t}$ for any $t = 1, \ldots, r$. Now observe that by (26), we have

$$\|x_{s,\tau} - u\|^2 + \|x_{s,\tau-1} - x_{s,\tau}\|^2 \leq \|x_{s,\tau-1} - u\|^2, \tau = 1, \ldots, t,$$

for any $t = 1, \ldots, r$. Summing up the above inequalities and using (9), we obtain

$$\|x_{s,t} - u\|^2 + \sum_{\tau=1}^{t} \|x_{s,\tau-1} - x_{s,\tau}\|^2 \leq \|x_{s,0} - u\|^2 \leq D_X^2,$$

which clearly implies (70). ∎

The following result establishes an important recursion of the ABL method.

Lemma 8 *Let $(x_{s,t}^l, x_{s,t}, x_{s,t}^u)$, $s = 1, 2, \ldots$; $t = 0, 1, \ldots, T_s - 1$, be the search points computed by the ABL method at step t of segment s. Let Γ_t be defined in (33) and suppose that the stepsizes $\alpha_t \in (0, 1]$, $t \geq 1$, are chosen such that the first relation of (34) holds. Then, for any $t = 1, \ldots, T_s - 1$,*

$$\frac{\Delta_{s,t}}{\Gamma_t} \leq (1 - \alpha_1 \lambda)\Delta_{s,0} + \frac{C_1 L D_X^2}{2} + M D_X \left[\sum_{\tau=1}^{t} \left(\frac{\alpha_\tau}{\Gamma_\tau} \right)^2 \right]^{\frac{1}{2}}. \tag{71}$$

Proof. First observe that conditions (10), (11) and (12) hold with $(x, y, z) = (x_{s,t-1}, x_{s,t-1}^u, x_{s,t-1}^l)$, $(x^+, y^+, z^+) = (x_{s,t}, x_{s,t}^u, x_{s,t}^l)$, $l = l_{s,t}$ and $\alpha = \alpha_t$. Hence, by Lemma 1, we have

$$f(x_{s,t}^u) \leq (1 - \alpha_t) f(x_{s,t-1}^u) + \alpha_t l_{s,t} + \frac{L \alpha_t^2}{2} \|x_{s,t} - x_{s,t-1}\|^2 + M \alpha_t \|x_{s,t} - x_{s,t-1}\|.$$

By subtracting $\mathrm{lb}_{s,t}$ from both sides of the above inequality, and observing that in the ABL method, $f(x^u_{s,t}) - \mathrm{lb}_{s,t} = \mathrm{ub}_{s,t} - \mathrm{lb}_{s,t} = \Delta_{s,t}$ and that

$$\begin{aligned}(1-\alpha_t)f(x^u_{s,t-1}) - \mathrm{lb}_{s,t} + \alpha_t l_{s,t} &= (1-\alpha_t)\mathrm{ub}_{s,t-1} - \mathrm{lb}_{s,t} + \alpha_t\left[\lambda\,\mathrm{lb}_{s,t} + (1-\lambda)\,\mathrm{ub}_{s,t-1}\right] \\ &= (1-\alpha_t)\mathrm{ub}_{s,t-1} + \alpha_t(1-\lambda)\,\mathrm{ub}_{s,t-1} - (1-\alpha_t\lambda)\mathrm{lb}_{s,t} \\ &\leq (1-\alpha_t)\mathrm{ub}_{s,t-1} + \alpha_t(1-\lambda)\,\mathrm{ub}_{s,t-1} - (1-\alpha_t\lambda)\mathrm{lb}_{s,t-1} \\ &= (1-\alpha_t\lambda)\Delta_{s,t-1},\end{aligned}$$

where the inequality follows from the fact $\mathrm{lb}_{s,t} \geq \mathrm{lb}_{s,t-1}$, we then have, for any $t \geq 1$,

$$\Delta_{s,t} \leq (1-\alpha_t\lambda)\Delta_{s,t-1} + \frac{L\alpha_t^2}{2}\|x_{s,t} - x_{s,t-1}\|^2 + M\alpha_t\|x_{s,t} - x_{s,t-1}\|. \tag{72}$$

Now, letting $t = 1$ and dividing both sides of the above inequality by Γ_t, we have

$$\begin{aligned}\frac{\Delta_{s,1}}{\Gamma_1} &\leq \frac{1-\alpha_1\lambda}{\Gamma_1}\Delta_{s,0} + \frac{L\alpha_1^2}{2\Gamma_1}\|x_{s,1} - x_{s,0}\|^2 + \frac{M\alpha_1}{\Gamma_1}\|x_{s,1} - x_{s,0}\| \\ &= (1-\alpha_1\lambda)\Delta_{s,0} + \frac{L\alpha_1^2}{2\Gamma_1}\|x_{s,1} - x_{s,0}\|^2 + \frac{M\alpha_1}{\Gamma_1}\|x_{s,1} - x_{s,0}\|,\end{aligned} \tag{73}$$

where the last inequality follows from the fact that $\Gamma_1 = 1$. Similarly, we have

$$\begin{aligned}\frac{\Delta_{s,t}}{\Gamma_t} &\leq \frac{1-\alpha_t\lambda}{\gamma_t}\Delta_{s,t-1} + \frac{L\alpha_t^2}{2\Gamma_t}\|x_{s,t} - x_{s,t-1}\|^2 + \frac{M\alpha_t}{\Gamma_t}\|x_{s,t} - x_{s,t-1}\| \\ &= \frac{\Delta_{s,t-1}}{\Gamma_{t-1}} + \frac{L\alpha_t^2}{2\Gamma_t}\|x_{s,t} - x_{s,t-1}\|^2 + \frac{M\alpha_t}{\Gamma_t}\|x_{s,t} - x_{s,t-1}\|, \quad \forall\, t \geq 2,\end{aligned} \tag{74}$$

where the last identity follows from the definition of Γ_t in (33). Now summing up the concluding inequalities in (73) and (74), we obtain

$$\begin{aligned}\frac{\Delta_{s,t}}{\Gamma_t} &\leq (1-\alpha_1\lambda)\Delta_{s,0} + \frac{L}{2}\sum_{\tau=1}^{t}\frac{\alpha_\tau^2}{\Gamma_t}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\sum_{\tau=1}^{t}\frac{\alpha_\tau}{\Gamma_\tau}\|x_{s,\tau} - x_{s,\tau-1}\| \\ &\leq (1-\alpha_1\lambda)\Delta_{s,0} + \frac{L}{2}\sum_{\tau=1}^{t}\frac{\alpha_\tau^2}{\Gamma_t}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\Gamma_\tau}\right)^2\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2\right]^{\frac{1}{2}},\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality. The previous inequality, together with (70) and the first relation in (34), then imply that

$$\begin{aligned}\frac{\Delta_{s,t}}{\Gamma_t} &\leq (1-\alpha_1\lambda)\Delta_{s,0} + \frac{C_1L}{2}\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\Gamma_\tau}\right)^2\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2\right]^{\frac{1}{2}} \\ &\leq (1-\alpha_1\lambda)\Delta_{s,0} + \frac{C_1LD_X^2}{2} + MD_X\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\Gamma_\tau}\right)^2\right]^{\frac{1}{2}}.\end{aligned}$$

■

The following lemma bounds the number of steps in each segment of the ABL method.

Lemma 9 *Suppose that the stepsizes α_t, $t = 1, 2, \ldots$, of the ABL method are chosen such that condition (34) hold. Then, the number of steps in segment s is bounded by N_s, where*

$$N_s := 1 + \left(\frac{3C_2}{\lambda}\right)^{\frac{1}{2}} + \left(\frac{3C_1C_2LD_X^2}{2\lambda\Delta_{s,0}}\right)^{\frac{1}{2}} + \left(\frac{3C_3MD_X}{\lambda\Delta_{s,0}}\right)^2. \tag{75}$$

Proof. Let T_s denote the total number of steps performed in segment s by the ABL method. Assume for contradiction $T_s > N_s$. Clearly, we have $\Delta_{s,T_s-1} \geq \lambda\Delta_{s,0}$ since (18) does not hold. By the second two relations of (34) and (71), we obtain

$$\begin{aligned}
\Delta_{s,T_s-1} &\leq \Gamma_{T_s-1}(1-\alpha_1\lambda)\Delta_{s,0} + \frac{C_1\Gamma_{T_s-1}LD_X^2}{2} + MD_X\Gamma_{T_s-1}\left[\sum_{\tau=1}^{T_s-1}\left(\frac{\alpha_\tau}{\Gamma_\tau}\right)^2\right]^{\frac{1}{2}}. \\
&\leq \frac{\mathcal{C}_2(1-\alpha_1\lambda)\Delta_{s,0}}{(T_s-1)^2} + \frac{\mathcal{C}_1\mathcal{C}_2LD_X^2}{2(T_s-1)^2} + \frac{\mathcal{C}_3\,MD_X}{\sqrt{T_s-1}} \\
&< \frac{\mathcal{C}_2\Delta_{s,0}}{(N_s-1)^2} + \frac{\mathcal{C}_1\mathcal{C}_2LD_X^2}{2(N_s-1)^2} + \frac{\mathcal{C}_3\,MD_X}{\sqrt{N_s-1}},
\end{aligned}$$

where the last inequality follows from the facts that $T_s > N_s$, $\alpha_1 \geq 0$ and $\lambda \in (0,1)$. Using the above inequality and (75), we have $\Delta_{s,T_s-1} < \lambda\Delta_{s,0}$, which contradicts with $\Delta_{s,T_s-1} \geq \lambda\Delta_{s,0}$. ∎

We also need the following simple technical result.

Lemma 10 *Let the constants $r \in (0,1)$ and $v > 0$ be given. Then,*

$$\sum_{s=1}^{S}\gamma^{v(S-s)} \leq \frac{1}{1-r^v},\ \forall S \geq 1. \tag{76}$$

Proof. Clearly, we have

$$\sum_{s=1}^{S}\gamma^{v(S-s)} = \sum_{t=0}^{S-1}\gamma^{vt} \leq \frac{1}{1-r^v}.$$

∎

We now provide the proof of Theorem 3.

Proof of Theorem 3: Obviously, the ABL method will terminate in one step if $\Delta_{1,0} \leq \epsilon$. Without loss of generality, let us assume that $\Delta_{1,0} > \epsilon$. Observe that by (18) and (22), we have

$$\Delta_{s+1,0} < \lambda\Delta_{s,0},\quad \forall\, s \geq 1. \tag{77}$$

Letting S be the total number of segments performed by the ABL method, we can easily see from the above relation and (17) that

$$S \leq \left\lceil \log_{\frac{1}{\lambda}}\frac{\Delta_{1,0}}{\epsilon}\right\rceil \leq \left\lceil \log_{\frac{1}{\lambda}}\left(\frac{LD_X^2}{2\epsilon} + \frac{MD_X}{\epsilon}\right)\right\rceil. \tag{78}$$

It also follows from (77) that $\Delta_{s,0} > \epsilon\lambda^{s-S}$, $s = 1, \ldots, S$, since $\Delta_{S,0} > \epsilon$ due to the origin of S. Using this observation and (76), we obtain

$$\sum_{s=1}^{S} \Delta_{s,0}^{-\frac{1}{2}} < \sum_{s=1}^{S} \frac{\lambda^{\frac{1}{2}(S-s)}}{\epsilon^{\frac{1}{2}}} \leq \frac{1}{\epsilon^{\frac{1}{2}}(1-\lambda^{\frac{1}{2}})} \quad \text{and} \quad \sum_{s=1}^{S} \Delta_{s,0}^{-2} < \sum_{s=1}^{S} \frac{\lambda^{2(S-s)}}{\epsilon^2} \leq \frac{1}{\epsilon^2(1-\lambda^2)}.$$

Now by Lemma 9 and the above two inequalities, the total number of steps performed by the ABL method can be bounded by

$$\begin{aligned} \sum_{s=1}^{S} N_s &= S\left[1 + \left(\frac{3C_2}{\lambda}\right)^{\frac{1}{2}}\right] + \left(\frac{3C_1C_2LD_X^2}{2\lambda}\right)^{\frac{1}{2}} \sum_{s=1}^{S} \Delta_{s,0}^{-\frac{1}{2}} + \left(\frac{3C_3MD_X}{\lambda}\right)^2 \sum_{s=1}^{S} \Delta_{s,0}^{-2} \\ &\leq S\left[1 + \left(\frac{3C_2}{\lambda}\right)^{\frac{1}{2}}\right] + \frac{1}{1-\lambda^{\frac{1}{2}}}\left(\frac{3C_1C_2LD_X^2}{2\lambda\epsilon}\right)^{\frac{1}{2}} + \frac{1}{1-\lambda^2}\left(\frac{3C_3MD_X}{\lambda\epsilon}\right)^2. \end{aligned}$$

Combining (78) and the previous conclusion, we conclude that the total number of steps performed by the ABL method is bounded by $\mathcal{N}(\epsilon) + \mathcal{S}(\epsilon)$, where $\mathcal{N}(\epsilon)$ and $\mathcal{S}(\epsilon)$ are defined in (39) and (40), respectively. ∎

7.2 Convergence analysis for the APL method

The goal of this subsection is to prove Theorem 5, which describes the main convergence properties of the APL method. Our first lemma below states a result in place of Lemma 7 for the ABL method.

Lemma 11 *Let $x_{s,t}$ and $X_{s,t} \subseteq X$, respectively, denote the search point and the localizer computed at step t of phase s of the APL method. Then,*

a) $x_{s,t} = \operatorname{argmin}_x \{\omega_s(x) : x \in X_{s,t}\}$ *for any* $t \geq 1$*;*

b) $\frac{\sigma}{2}\|x_{s,t+1} - x_{s,t}\|^2 \leq \omega_s(x_{s,t+1}) - \omega_s(x_{s,t})$ *for any* $t \geq 0$*;*

c) $\frac{\sigma}{2}\sum_{\tau=1}^{t} \|x_{s,\tau} - x_{s,\tau-1}\|^2 \leq \omega_s(x_{s,t}) - \omega_s(x_{s,0})$ *for any* $t \geq 1$.

Proof. We first show part a). By the second identity of (55), we have $\langle \nabla\omega_s(x_{s,t}), x - x_{s,t}\rangle \geq 0$ for any $x \in \overline{X}_{s,t}$, which implies that $x_{s,t} = \operatorname{argmin}_{x\in\overline{X}_{s,t}} \omega_s(x)$. Moreover, noting that by (52) and the first identity of (55), we have $x_{s,t} \in \underline{X}_{s,t}$. Using these two observations and the fact that $\underline{X}_{s,t} \subseteq X_{s,t} \subseteq \overline{X}_{s,t}$, we have $x_{s,t} = \operatorname{argmin}_{x\in X_{s,t}} \omega_s(x)$ for any $t \geq 1$.

We now show part b). We first claim that $\langle \nabla\omega_s(x_{s,t}), x_{s,t+1} - x_{s,t}\rangle \geq 0$ for any $t \geq 0$. This claim is obviously true when $t = 0$ due to the facts that $x_{s,0} = c_s$ and $\nabla\omega_s(c_s) = 0$. Now suppose that $t \geq 1$. By (52), we have $x_{s,t+1} \in X_{s,t}$, which, in view of part a), then implies that $\langle \nabla\omega_s(x_{s,t}), x_{s,t+1} - x_{s,t}\rangle \geq 0$. Part b) now follows directly from our previous claim and the strong convexity of $\omega_s(x)$ (c.f. (42)). Moreover, part c) can be easily obtained by summing up the inequalities in part b). ∎

We now establish an important recursion of the APL method.

Lemma 12 *Let $(x_{s,t}, x^u_{s,t}) \in X \times X$ be the search points computed at step t of phase s by the APL method. Let $\tilde{\Gamma}_t$ be defined in (56) and suppose that the stepsizes α_t, $t = 1, 2, \ldots$, are chosen such that the first two relations of (57) hold. Then, for any $t \geq 1$ and $s \geq 1$,*

$$\frac{1}{\tilde{\Gamma}_t}\left[f(x^u_{s,t}) - \tilde{l}_s\right] \leq \frac{\tilde{C}_1 L}{\sigma}\left[\omega_s(x_{s,t}) - \omega_s(x_{s,0})\right] + M\left[\frac{2[\omega_s(x_{s,t}) - \omega_s(x_{s,0})]}{\sigma}\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\right)^2\right]^{\frac{1}{2}}. \tag{79}$$

Proof. First observe that conditions (10), (11) and (12) hold with $(x, y, z) = (x_{s,t-1}, x^u_{s,t-1}, x^l_{s,t-1})$, $(x^+, y^+, z^+) = (x_{s,t}, x^u_{s,t}, x^l_{s,t})$, $l = \tilde{l}_s$ and $\alpha = \alpha_t$. Hence, by Lemma 1, we have

$$f(x^u_{s,t}) \leq (1-\alpha_t) f(x^u_{s,t-1}) + \alpha_t \tilde{l}_s + \frac{L\alpha_t^2}{2}\|x_{s,t} - x_{s,t-1}\|^2 + M\alpha_t\|x_{s,t} - x_{s,t-1}\|.$$

By subtracting $\tilde{l}_s$ from both sides of (13), we have, for any $t \geq 1$,

$$\begin{aligned} f(x^u_{s,t}) - \tilde{l}_s &\leq (1-\alpha_t)\left[f(x^u_{s,t-1}) - \tilde{l}_s\right] + \frac{L\alpha_t^2}{2}\|x_{s,t} - x_{s,t-1}\|^2 + M\alpha_t\|x_{s,t} - x_{s,t-1}\| \\ &\leq (1-\alpha_t)\left[f(x^u_{s,t-1}) - \tilde{l}_s\right] + \frac{\tilde{C}_1 L\tilde{\Gamma}_t}{2}\|x_{s,t} - x_{s,t-1}\|^2 + M\alpha_t\|x_{s,t} - x_{s,t-1}\|, \end{aligned}$$

where the last inequality follows from the second relation in (57). Dividing both sides of the above inequality by $\tilde{\Gamma}_t$, and using (56) and the first relation in (57), we obtain that

$$\begin{aligned} \frac{1}{\tilde{\Gamma}_1}\left[f(x^u_{s,1}) - \tilde{l}_s\right] &\leq \frac{1-\alpha_1}{\tilde{\Gamma}_1}\left[f(x^u_{s,0}) - \tilde{l}_s\right] + \frac{\tilde{C}_1 L}{2}\|x_{s,1} - x_{s,0}\|^2 + M\frac{\alpha_1}{\Gamma_1}\|x_{s,1} - x_{s,0}\| \\ &= \frac{\tilde{C}_1 L}{2}\|x_{s,1} - x_{s,0}\|^2 + M\frac{\alpha_1}{\Gamma_1}\|x_{s,1} - x_{s,0}\| \end{aligned}$$

and that, for any $\tau = 2, 3, \ldots, t$,

$$\begin{aligned} \frac{1}{\tilde{\Gamma}_\tau}\left[f(x^u_{s,\tau}) - \tilde{l}_s\right] &\leq \frac{1-\alpha_\tau}{\tilde{\Gamma}_\tau}\left[f(x^u_{s,\tau-1}) - \tilde{l}_s\right] + \frac{\tilde{C}_1 L}{2}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\|x_{s,\tau} - x_{s,\tau-1}\| \\ &= \frac{1}{\tilde{\Gamma}_{\tau-1}}\left[f(x^u_{s,\tau-1}) - \tilde{l}_s\right] + \frac{\tilde{C}_1 L}{2}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\|x_{s,\tau} - x_{s,\tau-1}\|. \end{aligned}$$

Adding up the above inequalities, we have, for any $t \geq 1$,

$$\begin{aligned} \frac{1}{\tilde{\Gamma}_t}\left[f(x^u_{s,t}) - \tilde{l}_s\right] &\leq \frac{\tilde{C}_1 L}{2}\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\sum_{\tau=1}^{t}\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\|x_{s,\tau} - x_{s,\tau-1}\| \\ &\leq \frac{\tilde{C}_1 L}{2}\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2 + M\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\tilde{\Gamma}_\tau}\right)^2\right]^{\frac{1}{2}}\left[\sum_{\tau=1}^{t}\|x_{s,\tau} - x_{s,\tau-1}\|^2\right]^{\frac{1}{2}}, \end{aligned} \tag{80}$$

where the last inequality follows from the Cauchy-Schwartz inequality. The result now immediately follows by combining (80) and Lemma 11.c). ∎

Our next result bounds the total number of steps performed in phase s, $s \geq 1$, of the APL method.

Lemma 13 *Suppose that the stepsize α_t, $t = 1, 2, \ldots$, are chosen such that (57) holds. Then the number of steps performed in phase s, $s = 1, 2, \ldots$, of the APL method is bounded by $\lceil \tilde{N}_s \rceil$, where*

$$\tilde{N}_s := \left(\frac{2\tilde{C}_1 \tilde{C}_2 L \mathcal{D}_{\omega,X}^2}{\sigma\theta\lambda\tilde{\Delta}_s} \right)^{\frac{1}{2}} + \frac{8\tilde{C}_3^2 M^2 \mathcal{D}_{\omega,X}^2}{\sigma\theta^2\lambda^2\tilde{\Delta}_s^2}, \tag{81}$$

$\tilde{\Delta}_s := \tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s$ and $\mathcal{D}_{\omega,X}$ is defined in (45).

Proof. Let $(x_{s,t}, x_{s,t}^u) \in X \times X$ be the search points computed at step t of phase s by the APL method. Observe that by (41), we have $\omega_s(x_{s,0}) \geq 0$ and hence that $\omega_s(x_{s,t}) - \omega_s(x_{s,0}) \leq \omega_s(x_{s,t})$. This observation, together with Lemma 12 and the last two relations in (57), then imply that

$$f(x_{s,t}^u) - \tilde{l}_s \leq \frac{\tilde{C}_1 L \tilde{\Gamma}_t}{\sigma} \omega_s(x_{s,t}) + M\tilde{\Gamma}_t \left[\frac{2\omega_s(x_{s,t})}{\sigma} \sum_{\tau=1}^{t} \left(\frac{\alpha_\tau}{\tilde{\Gamma}_\tau} \right)^2 \right]^{\frac{1}{2}} \leq \frac{\tilde{C}_1 \tilde{C}_2 L\, \omega_s(x_{s,t})}{\sigma t^2} + \frac{\tilde{C}_3 M \sqrt{2\omega_s(x_{s,t})}}{\sqrt{\sigma t}}. \tag{82}$$

By the above inequality, (81), the relation $\omega_s(x_{s,t}) \leq \mathcal{D}_{\omega,X}^2$ due to (45), and the fact that $\lceil \tilde{N}_s \rceil \geq \tilde{N}_s$, we have

$$f(x_{s,\lceil \tilde{N}_s \rceil}^u) - \tilde{l}_s \leq \frac{\tilde{C}_1 \tilde{C}_2 L \mathcal{D}_{\omega,X}^2}{\sigma \tilde{N}_s^2} + \frac{\sqrt{2}\tilde{C}_3 M \mathcal{D}_{\omega,X}}{\sqrt{\sigma \tilde{N}_s}} \leq \theta\lambda\tilde{\Delta}_s = \theta\lambda(\tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s) = \theta(\tilde{\mathrm{ub}}_s - \tilde{l}_s),$$

where the last equality follows from the fact that $\tilde{l}_s = \lambda\, \tilde{\mathrm{lb}}_s + (1-\lambda)\, \tilde{\mathrm{ub}}_s$. The previous conclusion, in view of (54), then clearly implies that the number of steps in phase s can be bounded by $\lceil \tilde{N}_s \rceil$. ∎

We are now ready to prove Theorem 5.

Proof of Theorem 5: Obviously, if $\tilde{\Delta}_1 \leq \epsilon$, the APL method will terminate in one step. Without loss of generality, let us assume that $\tilde{\Delta}_1 > \epsilon$. First, observe that

$$\tilde{\Delta}_{s+1} \leq q\, \tilde{\Delta}_s, \quad \forall\, s \geq 1, \tag{83}$$

where $q \equiv q(\theta, \lambda)$ is defined in (60). Indeed, if phase s is terminated according to (51), then

$$\tilde{\Delta}_{s+1} = \tilde{\mathrm{ub}}_{s+1} - \tilde{\mathrm{lb}}_{s+1} \leq \tilde{\mathrm{ub}}_s - [\tilde{l}_s - \theta(\tilde{l}_s - \tilde{\mathrm{lb}}_s)] = [1 - (1-\theta)(1-\lambda)]\tilde{\Delta}_s. \tag{84}$$

Otherwise, phase s must terminate when (54) holds. In this case, we have

$$\begin{aligned} \tilde{\Delta}_{s+1} &= \tilde{\mathrm{ub}}_{s+1} - \tilde{\mathrm{lb}}_{s+1} \leq \tilde{\mathrm{ub}}_{s+1} - \tilde{\mathrm{lb}}_s \leq \tilde{l}_s + \theta(\tilde{\mathrm{ub}}_s - \tilde{l}_s) - \tilde{\mathrm{lb}}_s \\ &= \theta\lambda\tilde{\Delta}_s + (1-\lambda)\tilde{\Delta}_s = [1 - (1-\theta)\lambda]\tilde{\Delta}_s. \end{aligned} \tag{85}$$

Combining (84) and (85), we obtain (83). Letting $\tilde{S}$ be the total number of phases performed by the APL method, we can easily see from (46), (62) and (83) that

$$\tilde{S} \leq \left\lceil \log_{\frac{1}{\gamma}} \frac{\tilde{\Delta}_1}{\epsilon} \right\rceil \leq 1 + \log_{\frac{1}{\gamma}} \frac{\tilde{\Delta}_1}{\epsilon} \leq \tilde{\mathcal{S}}(\epsilon). \tag{86}$$

It also follows from (83) that $\tilde{\Delta}_s > \epsilon q^{s-\tilde{S}}$, $s = 1, \ldots, \tilde{S}$, since $\tilde{\Delta}_{\tilde{S}} > \epsilon$ due to the origin of $\tilde{S}$. Using this observation and (76), we obtain

$$\sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-\frac{1}{2}} < \sum_{s=1}^{\tilde{S}} \frac{q^{\frac{1}{2}(\tilde{S}-s)}}{\epsilon^{\frac{1}{2}}} \leq \frac{1}{\epsilon^{\frac{1}{2}}(1-q^{\frac{1}{2}})} \quad \text{and} \quad \sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-2} < \sum_{s=1}^{\tilde{S}} \frac{q^{2(\tilde{S}-s)}}{\epsilon^2} \leq \frac{1}{\epsilon^2(1-q^2)}. \tag{87}$$

Now, by Lemma 9, (86) and the above two inequalities, the total number of steps performed by the APL method can be bounded by

$$\begin{aligned}
\sum_{s=1}^{\tilde{S}} \left\lceil \tilde{N}_s \right\rceil &\leq \tilde{S} + \sum_{s=1}^{\tilde{S}} \tilde{N}_s \leq \tilde{\mathcal{S}}(\epsilon) + \sum_{s=1}^{\tilde{S}} \tilde{N}_s \\
&\leq \tilde{\mathcal{S}}(\epsilon) + \left(\frac{2\tilde{C}_1 \tilde{C}_2 L \mathcal{D}_{\omega,X}^2}{\sigma\theta\lambda} \right)^{\frac{1}{2}} \sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-\frac{1}{2}} + \frac{8\tilde{C}_3^2 M^2 \mathcal{D}_{\omega,X}^2}{\sigma\theta^2\lambda^2} \sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-2} \\
&\leq \tilde{\mathcal{S}}(\epsilon) + \frac{1}{1-\sqrt{q}} \left(\frac{2\tilde{C}_1 \tilde{C}_2 L \mathcal{D}_{\omega,X}^2}{\sigma\theta\lambda\,\epsilon} \right)^{\frac{1}{2}} + \frac{1}{1-q^2} \left(\frac{2\sqrt{2}\tilde{C}_3 M \mathcal{D}_{\omega,X}}{\sqrt{\sigma}\theta\lambda\,\epsilon} \right)^2 \\
&= \tilde{\mathcal{S}}(\epsilon) + \tilde{\mathcal{N}}(\epsilon),
\end{aligned}$$

where the last identity follows from the definition of $\tilde{\mathcal{N}}(\epsilon)$ in (61). ∎

7.3 Minimizing strongly convex functions

Our goal in this subsection is to provide the proof of Theorem 6.

Proof of Theorem 6: Obviously, the APL method will terminate in one step if $\tilde{\Delta}_1 \leq \epsilon$. Without loss of generality, we assume that $\tilde{\Delta}_1 > \epsilon$. We first claim that, under assumption (65) or (66),

$$\omega_s(x_{s,t}) \leq \frac{\tilde{\Delta}_s}{\mu}, \quad \forall\, s \geq 1, t \geq 1. \tag{88}$$

Indeed, by (52) and (65), we have $\omega_s(x_{s,t}) \leq \omega_s(x^*)$, which together with (63), (64) and (65), then imply that

$$\omega_s(x_{s,t}) \leq \frac{1}{2}\|c_s - x^*\|^2 \leq \frac{1}{\mu}[f(c_s) - f^*] = \frac{1}{\mu}(\mathrm{ub}_s - f^*) \leq \frac{\tilde{\Delta}_s}{\mu}.$$

Moreover, if condition (66) (rather than (65)) holds, then by (64), (66) and the fact that $x_{s,t} \in X_{s,t}$, we also have

$$\omega_s(x_{s,t}) \leq \frac{\|x_{s,t} - c_s\|^2}{2} \leq \frac{\tilde{\Delta}_s}{\mu}.$$

We now show that the number of steps performed at phase s of the APL method is bounded by $\lceil \hat{N}_s \rceil$ with

$$\hat{N}_s := \sqrt{\frac{2\tilde{C}_1 \tilde{C}_2 L}{\sigma\theta\lambda\mu}} + \frac{8\tilde{C}_3^2 M^2}{\sigma\theta^2\lambda^2\mu\tilde{\Delta}_s}. \tag{89}$$

Indeed, by (82), (88), (89) and the fact that $\lceil \hat{N}_s \rceil \geq \hat{N}_s$, we have

$$f(x^u_{s,\lceil \hat{N}_s \rceil}) - \tilde{l}_s \leq \frac{\tilde{C}_1 \tilde{C}_2 L \tilde{\Delta}_s}{\sigma \mu \hat{N}_s^2} + \frac{\tilde{C}_3 M \sqrt{2\tilde{\Delta}_s}}{\sqrt{\sigma \hat{N}_s}} \leq \theta \lambda \tilde{\Delta}_s = \theta \lambda (\tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s) = \theta (\tilde{\mathrm{ub}}_s - \tilde{l}_s),$$

where the last equality follows from the fact that $\tilde{l}_s = \tilde{\mathrm{lb}}_s + (1-\lambda)(\tilde{\mathrm{ub}}_s - \tilde{\mathrm{lb}}_s)$. Using the previous conclusion and (54), we can easily see that the number of steps in phase s can be bounded by $\lceil \hat{N}_s \rceil$. Let $\tilde{S}$ be the number of phases performed by the APL method. By (86), we have $\tilde{S} \leq \tilde{\mathcal{S}}(\epsilon)$, where $\tilde{\mathcal{S}}(\epsilon)$ is defined in (62). Moreover, using an argument similar to the one given in the proof of (87), we can show that

$$\sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-1} \leq \frac{1}{\epsilon(1-q)}.$$

Using these observations, we conclude that the total number of steps performed by the APL method is bounded by

$$\begin{aligned} \sum_{s=1}^{\tilde{S}} \left\lceil \hat{N}_s \right\rceil &\leq \tilde{S} \left(1 + \sqrt{\frac{2\tilde{C}_1 \tilde{C}_2 L}{\sigma \theta \lambda \mu}} \right) + \frac{8 \tilde{C}_3^2 M^2}{\sigma \theta^2 \lambda^2 \mu} \sum_{s=1}^{\tilde{S}} \tilde{\Delta}_s^{-1} \\ &\leq \tilde{\mathcal{S}}(\epsilon) \left(1 + \sqrt{\frac{2\tilde{C}_1 \tilde{C}_2 L}{\sigma \theta \lambda \mu}} \right) + \frac{1}{1-q} \frac{8 \tilde{C}_3^2 M^2}{\sigma \theta^2 \lambda^2 \mu \epsilon}. \end{aligned} \tag{90}$$

■

7.4 Determining the stepsizes

The goal of this subsection is to prove Lemmas 2 and 4.

The following technical result will be used in the proof of part b) of both Lemma 2 and Lemma 4.

Lemma 14 *Let $\alpha_1 = \gamma_1 = 1$. Also suppose that α_t, γ_t, $t \geq 2$, are computed recursively by*

$$\gamma_t = \alpha_t^2 = (1 - \beta \alpha_t) \gamma_{t-1} \tag{91}$$

for some $\beta \in (0, 1]$, then we have, for any $t \geq 1$,

$$\alpha_t \in (0, 1], \quad \gamma_t \leq \frac{4}{\beta^2 t^2} \quad \textit{and} \quad \gamma_t \left[\sum_{\tau=1}^{t} \left(\frac{\alpha_\tau}{\gamma_t} \right)^2 \right]^{\frac{1}{2}} \leq \frac{4}{\sqrt{3\beta t}}.$$

Proof. Note that by (91), we have

$$\alpha_t = \frac{1}{2} \left[-\beta \gamma_{t-1} + \sqrt{\beta^2 \gamma_{t-1}^2 + 4\gamma_{t-1}} \right], \quad t \geq 2, \tag{92}$$

which clearly implies that $\alpha_t > 0$, $t \geq 2$. We now show that $\alpha_t \leq 1$ and $\gamma_t \leq 1$ by induction. Indeed, by the inductive hypothesis and the fact $\beta \in (0, 1]$, we have $(1 - \beta)\, \gamma_{t-1} \leq 1$, which in view of (92),

then implies that

$$\begin{aligned}\alpha_t &\leq \frac{1}{2}\left[-\beta\gamma_{t-1}+\sqrt{\beta^2\gamma_{t-1}^2+4\gamma_{t-1}+4[1-(1-\beta)\gamma_{t-1}]}\right]\\ &= \frac{1}{2}\left[-\beta\gamma_{t-1}+\sqrt{\beta^2\gamma_{t-1}^2+4\beta\gamma_{t-1}+4}\right]=1. \end{aligned} \tag{93}$$

The previous conclusion, together with the fact that $\alpha_t^2=\gamma_t$ due to (91), then also imply that $\gamma_t\leq 1$. Now let us bound $1/\sqrt{\gamma_t}$ for any $t\geq 2$. First observe that, by (91), we have, for any $t\geq 2$,

$$\frac{1}{\sqrt{\gamma_t}}-\frac{1}{\sqrt{\gamma_{t-1}}}=\frac{\sqrt{\gamma_{t-1}}-\sqrt{\gamma_t}}{\sqrt{\gamma_{t-1}\gamma_t}}=\frac{\gamma_{t-1}-\gamma_t}{\sqrt{\gamma_{t-1}\gamma_t}\left(\sqrt{\gamma_{t-1}}+\sqrt{\gamma_t}\right)}=\frac{\beta\alpha_t\gamma_{t-1}}{\gamma_{t-1}\sqrt{\gamma_t}+\gamma_t\sqrt{\gamma_{t-1}}}. \tag{94}$$

Using the above identity, (91) and the fact that $\gamma_t\leq\gamma_{t-1}$ due to (91), we conclude that

$$\frac{1}{\sqrt{\gamma_t}}-\frac{1}{\sqrt{\gamma_{t-1}}}\geq\frac{\beta\alpha_t}{2\sqrt{\gamma_t}}=\frac{\beta}{2},$$

which, in view of the fact that $\gamma_1=1$, then implies that $1/\sqrt{\gamma_t}\geq 1+(t-1)\beta/2=[2+\beta(t-1)]/2$. By the previous inequality and the fact that $\beta\in(0,1]$, we have

$$\gamma_t\leq\frac{4}{[2+\beta(t-1)]^2}\leq\frac{4}{\beta^2t^2},\quad\forall\, t\geq 1. \tag{95}$$

Moreover, it also follows from (94) that $\frac{1}{\sqrt{\gamma_t}}-\frac{1}{\sqrt{\gamma_{t-1}}}\leq\frac{\beta\alpha_t}{\sqrt{\gamma_t}}=\beta$, which, in view of the fact $\gamma_1=1$ and (91), then implies that $\frac{\alpha_t}{\gamma_t}=\frac{1}{\sqrt{\gamma_t}}\leq 1+\beta(t-1)$. By using the previous conclusion and (95), we have

$$\begin{aligned}\gamma_t\left[\sum_{\tau=1}^t\left(\frac{\alpha_\tau}{\gamma_t}\right)^2\right]^{\frac{1}{2}} &\leq \frac{4}{[2+\beta(t-1)]^2}\left[\sum_{\tau=1}^t\left(1+\beta(\tau-1)\right)^2\right]^{\frac{1}{2}}\\ &\leq \frac{4}{[2+\beta(t-1)]^2}\left(\int_0^{2+\beta(t-1)}u^2du\right)^{\frac{1}{2}}=\frac{4}{\sqrt{3}[2+\beta(t-1)]^{\frac{1}{2}}}\leq\frac{4}{\sqrt{3\beta t}}.\end{aligned}$$

■

We are now ready to show Lemmas 2 and 4. Note that we need the following simple inequality inside these proofs.

$$\sum_{\tau=1}^t\tau^2=\frac{t(t+1)(2t+1)}{6}\leq\frac{t(t+1)^2}{3}. \tag{96}$$

Proof of Lemma 2: Part b) of the result follows directly from Lemma 14 with $\beta=\lambda$ and $\gamma_t=\Gamma_t$. We only need to show part a). Clearly, by (35) and the facts that $\lambda\in(0,1/2]$ and $t\geq 1$, we have $\alpha_t\in(0,1]$ for any $t\geq 1$. It can also be easily seen from (33) and (35) that

$$\Gamma_t=\frac{6}{(t+2)(t+3)}\leq\frac{6}{t^2},\quad t\geq 1.$$

Using the above relation, (35) and (96), we have

$$\begin{aligned}\Gamma_t\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\Gamma_t}\right)^2\right]^{\frac{1}{2}} &= \Gamma_t\left[\sum_{\tau=1}^{t}\left(\frac{t+2}{3\lambda}\right)^2\right]^{\frac{1}{2}} \le \frac{\Gamma_t}{3\lambda}\left(\sum_{\tau=1}^{t+2}\tau^2\right)^{\frac{1}{2}} \\ &\le \frac{\Gamma_t}{3\lambda}\left(\frac{(t+2)(t+3)^2}{3}\right)^{\frac{1}{2}} = \frac{1}{3\lambda\sqrt{3(t+2)}} \le \frac{1}{3\lambda\sqrt{3t}}.\end{aligned}$$

These observations together with (35) then imply that (34) holds with $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ given by (36). ■

Proof of Lemma 4: Part b) of the result follows directly from Lemma 14 with $\beta = 1$ and $\gamma_t = \tilde{\Gamma}_t$. We only need to show part a). Using (56) and (58), we have $\alpha_1 = 1$ and

$$\tilde{\Gamma}_t = \frac{2}{t(t+1)} \le \frac{2}{t^2}, \quad t \ge 1. \tag{97}$$

Clearly, we have $\alpha_t^2 \le 2\tilde{\Gamma}_t$. Moreover, by (97), (58) and (96), we have

$$\tilde{\Gamma}_t\left[\sum_{\tau=1}^{t}\left(\frac{\alpha_\tau}{\tilde{\Gamma}_t}\right)^2\right]^{\frac{1}{2}} = \tilde{\Gamma}_t\left(\sum_{\tau=1}^{t}\tau^2\right)^{\frac{1}{2}} \le \tilde{\Gamma}_t\left(\frac{t(t+1)^2}{3}\right)^{\frac{1}{2}} = \frac{2}{\sqrt{3t}}.$$

■

8 Concluding remarks

In this paper, we present two new bundle-type methods for convex programming. Our major theoretical contributions consist of the following aspects: i) substantially improve the rate of convergence of bundle-type methods, when applied to smooth CP problems, from $\mathcal{O}(1/\sqrt{t})$ to $\mathcal{O}(1/t^2)$; ii) present a class of uniformly optimal algorithms for solving smooth and non-smooth CP problems. In other words, given that the CP problem is represented by a first-order oracle, these algorithms do not require any global smoothness information to achieve the optimal rates of convergence. From the practical point of view, our contribution is to introduce certain promising alternative algorithms to Nesterov's methods (or its variants), which are the only previously known optimal algorithms for smooth convex optimization. In the future, we will investigate the optimization techniques presented in this paper for solving more structured non-smooth CP and variational inequality problems.

References

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

[2] H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Controal and Optimization*, 42:596–636, 2003.

[3] A. Ben-Tal and A.Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming*, 102:407–456, 2005.

[4] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2000.

[5] L.M. Bregman. The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.*, 7:200–217, 1967.

[6] E.W. Chenny and A.A. Goldstein. Newton's methods for convex programming and tchebytcheff approximation. *Numerische Mathematik*, 1:253–268, 1959.

[7] 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.

[8] M. Heinkenschloss. A trust region method for norm constrained problems. *SIAM Journal on Numerical Analysis*, 4:1594–1620, 1998.

[9] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Manuscript.

[10] J.E. Kelley. The cutting plane method for solving convex programs. *Journal of the SIAM*, 8:703–712, 1960.

[11] K.C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.

[12] K.C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122, 1990.

[13] K.C. Kiwiel. Proximal level bundle method for convex nondifferentable optimization, saddle point problems and variational inequalities. *Mathematical Programming, Series B*, 69:89–109, 1995.

[14] K.C. Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM Journal on Controal and Optimization*, 35:1142–1168, 1997.

[15] G. Lan. An optimal method for stochastic composite optimization. *submitted to Mathematical Programming*, 2009. *E-print available at:* `http://www.optimization-online.org`.

[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*, 2009. to appear.

[17] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of robust stochastic approximation method. *submitted to Mathematical Programming*, 2008. *E-print available at:* `http://www.optimization-online.org`.

[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] J. Linderoth and S. Wright. Decomposition algorithms for stochastic programming on a computational grid. *Computational Optimization and Applications*, 24:207–250, 2003.

[21] R. Mifflin. A modification and an extension of lemaréchal's algorithm for nonsmooth minimization. *Mathematical Programming Study*, 17:77–90, 1982.

[22] Mosek. The mosek optimization toolbox for matlab manual. version 6.0 (revision 93). `http://www.mosek.com`.

[23] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

[24] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[25] 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.

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

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

[28] Ruszczyński. *Nonlinear Optimization*. Princeton University Press, first edition, 2006.

[29] M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.

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