Journal of Optimization Theory and Applications
(accepted for publication)

ON APPROXIMATIONS WITH FINITE PRECISION IN BUNDLE METHODS FOR NONSMOOTH OPTIMIZATION*

M. V. Solodov†

December 20, 2001 (Revised July 3, 2002)

ABSTRACT

We consider the proximal form of a bundle algorithm for minimizing a nonsmooth convex function, assuming that the function and subgradient values are evaluated approximately. We show how those approximations should be controlled in order to satisfy the desired optimality tolerance. This is relevant, for example, in the context of Lagrangian relaxation, where obtaining exact information about the function and subgradient values involves solving exactly a certain optimization problem, which can be relatively costly (and, as we show, is in any case unnecessary). We exhibit that approximation with some finite precision is sufficient in this setting, and give an explicit characterization of this precision. Alternatively, our result can be viewed as a stability analysis of standard proximal bundle methods, as it answers the following question: for a given approximation error, what kind of an approximate solution can be obtained and how does it depend on the magnitude of the perturbation?

Key words. Nonsmooth optimization, convex optimization, bundle methods, stability analysis, perturbation.
AMS subject classifications. 90C30, 90C33.

---

* Research of the author is partially supported by CNPq Grant 300734/95-6, by PRONEX–Optimization, and by FAPERJ.
† Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil.
Email : `solodov@impa.br` .

1 Introduction

We consider the problem

$$\min_x \{f(x) \mid x \in \Re^n\}, \tag{1}$$

where $f : \Re^n \to \Re$ is a convex (in general, nondifferentiable) function. Bundle methods (see [8], [7, Chapters XIV, XV] and [4, Chapter 7]) are perhaps the most practical and efficient algorithms for solving problems of this class. We are interested in the case where at each given point we have available only an approximate value of f and an approximate value of one of its subgradients. For example, this can be considered a typical situation in the context of Lagrangian relaxation [2, 12, 1, 4], where one would need an exact solution of a certain optimization problem to obtain exact information about f and its subgradient. To make this motivation a little more specific, consider the (primal) problem

$$\max_\xi \{q(\xi) \mid \xi \in P,\, h(\xi) = 0\}, \tag{2}$$

where P is a compact subset of $\Re^m$, and $q : \Re^m \to \Re$, $h : \Re^m \to \Re^n$. Lagrangian relaxation (of the equality constraints in this problem) leads to problem (1), where

$$f(x) = \max_\xi \{q(\xi) + \langle x, h(\xi)\rangle \mid \xi \in P\} \tag{3}$$

is the dual function. Trying to solve problem (2) by means of solving its dual (1) makes sense in many situations. One example is when some separable structure is present in (2), so that the maximization problem in (3) is decomposable (e.g., see [4, Chapter 8] and [1, Chapter 6]). The approach of Lagrangian relaxation gives rise to a convex nonsmooth problem (1) independently of any assumptions about (2). Of course, without further assumptions, there can be a duality gap between the primal and the dual problems. Another important issue is recovering primal solutions after solving the dual. Here, we shall not deal with those questions. At the very least, the optimal value of (1) always provides an upper bound for the optimal value of (2), which can be already useful by itself in many applications.

As is well known and easy to see, if for a given $x \in \Re^n$ some $\xi(x)$ is a solution of the optimization problem in (3) then we have not only the dual function value $f(x) = q(\xi(x)) + \langle x, h(\xi(x))\rangle$ but also a subgradient $h(\xi(x)) \in \partial f(x)$. The point relevant for the present paper is that evaluating the function value $f(x)$ and a subgradient $y \in \partial f(x)$ requires solving *exactly* the optimization problem in (3). In some cases, computing an exact solution can be cost-prohibitive or, at least, cost-inefficient. As we shall show, it is actually in some sense unnecessary. One can also argue that computing an exact solution is computationally unrealistic to start with, as any numerical method used for maximization in (3) would return an approximate solution according to some (finite) termination criterion. At any rate, it is probably desirable not to spend more time on solving (3) than really necessary. The question we ask is therefore the following:

Given the desired optimality tolerance $\Delta_{\text{opt}} > 0$ for problem (1), to which precision should we evaluate the values of f and its subgradients in order to guarantee that the bundle method using this inexact data would terminate with a point satisfying the given tolerance Δ_{opt}?

We also pose the following related *stability* questions. Given some nonzero (and not necessarily tending to zero!) approximation error, what are the convergence properties of the bundle method? What kind of approximate solutions of (1) can be obtained, and how do they depend on the approximation error? Stability analysis of computational algorithms with non-vanishing perturbations (for example, induced by noisy data) is of independent interest and importance, e.g., [21, 20].

The problem setting we consider here is essentially the same as in [11]. Specifically, given $x \in \Re^n$ and $\tilde{\varepsilon} \geq 0$ (as a practical matter, actually $\tilde{\varepsilon} > 0$), we assume that we can find some $\tilde{f} \in \Re$ and $y \in \Re^n$ such that

$$f(x) \geq \tilde{f} \geq f(x) - \tilde{\varepsilon},$$

$$f(\zeta) \geq l(\zeta) := \tilde{f} + \langle y, \zeta - x \rangle \quad \forall\, \zeta \in \Re^n.$$

As discussed in [11], this setting is realistic in many applications. In the context of Lagrangian relaxation, those conditions amount to computing some $\xi \in P$ which is $\tilde{\varepsilon}$-optimal in the maximization problem in (3), and then taking $\tilde{f} = q(\xi) + \langle x, h(\xi) \rangle$, $y = h(\xi)$. As is easily seen, the stated conditions mean that $y \in \partial f_{\tilde{\varepsilon}}(x)$, where

$$\partial f_{\varepsilon}(x) = \{ y \in \Re^n \mid f(\zeta) \geq f(x) + \langle y, \zeta - x \rangle - \varepsilon \}$$

is the ε-subdifferential of f at x (with ∂f_0 being the exact subdifferential ∂f).

In [11] it was established that the proximal bundle method based on the inexact linearizations of f described above converges to a solution if the approximations of f and its subgradients are *asymptotically exact*, that is if $\tilde{\varepsilon} \to 0$ over the course of iterations. This means that for a given optimality tolerance $\Delta_{\text{opt}} > 0$, the method would terminate finitely with an acceptable approximate solution of (1) (also, at termination $\tilde{\varepsilon}$ would still be at some nonzero value). Note that this result says nothing about how to control the accuracy parameter $\tilde{\varepsilon}$ and, in particular, about how small $\tilde{\varepsilon}$ needs to be, or about how large it can be allowed to be, in order to ensure the finite termination. In the context of Lagrangian relaxation, it presumes that we can solve the optimization problem in (3) with an arbitrarily high precision. While this can be accepted as realistic in some cases, it may not be so in others. Also, it may be wasteful to approximate the objective and subgradient values (e.g., solve (3)) with higher precision than really necessary. Intuition suggests that it is reasonable to set this precision based on the required final optimality tolerance Δ_{opt}, especially at the later stages of computations. Making this relation explicit is the subject of this paper. In particular, we give an explicit expression for $\tilde{\varepsilon}$ in terms of Δ_{opt}, see (6).

Alternatively, our results provide stability analysis of the proximal bundle method, answering the following questions. Given some nonzero (and not tending to zero!) approximation error $\tilde{\varepsilon}$, what kind of approximate solutions of (1) can be obtained? How do they depend on the approximation error $\tilde{\varepsilon}$? We comment on those issues further at the end of the paper.

As other work on bundle methods with approximate data, we mention [6, 18]. In [6], it is assumed that for each $x \in \Re^n$ the exact value $f(x)$ is available while the subgradient is computed approximately, $y \in \partial f_\varepsilon(x)$. The attractive feature of the analysis in [6] is that, unlike in [11] and the present paper, no knowledge of ε is assumed. On the other hand, the exact value of f is needed, and so this setting is not suitable for some important applications of bundle methods, and in particular, for the Lagrangian relaxation. Indeed, as discussed above, in that case evaluating $f(x)$ and $y \in \partial f(x)$ is the same task, which is computing a solution $\xi(x)$ of the maximization problem in (3). In [18], an inexact proximal bundle method for the maximum eigenvalue function is developed. As in [11] and the present paper, both the function and subgradient values are approximated. Furthermore, certain rules to control the approximation are given. Those rules are constructive and are preferable to merely saying that $\tilde{\varepsilon} \to 0$. But as in [11], stability issues and finite (not tending to zero) precision of approximation are not considered in [18].

Our notation is standard. By $\langle \cdot, \cdot \rangle$ we denote the usual inner product in the given (finite-dimensional) space, with $\| \cdot \|$ being the associated norm. For an arbitrary set A, we denote by $|A|$ its cardinality.

2 The Method

Apart from being based on inexact data and a special stopping test, the algorithm we consider here is a fairly standard proximal bundle method, with the size of the model controlled by the aggregation technique. Nevertheless, it is worth to explain our notation.

It is convenient for our purposes to keep separate track of the usual linear approximations to the function and of the approximations obtained after the aggregation. The former will be collected into the set B_k^c and denoted by $l_i(x)$, $i \leq k$. The latter will be collected into the set B_k^a and denoted by $l_i^a(x)$, $i \leq k$. Thus B_k^c and B_k^a are subsets in the space of linear functions. The two sets I_k^c and I_k^a collect the iteration indices of the current members of B_k^c and B_k^a, respectively.

When the number of elements in $B_k^c \cup B_k^a$ reaches the prescribed upper bound $B_{\max}$, two or more of those elements are deleted from the bundle, being replaced by the so-called aggregate piece (Step 8 of Algorithm 2.1). This controls the complexity of the cutting-planes approximation of f, i.e., φ_k given by (8), and thus keeps subproblems (7) in Step 2 of Algorithm 2.1 manageable. Subproblem (7) is solved via some quadratic programming method applied to its dual (see (14) and Lemma 1), whose dimensionality is precisely $|B_k^c| + |B_k^a|$. It makes sense to use specialized highly effective QP solvers developed for the structure in (14), e.g., [9].

We do not give any specific rule for choosing the proximal parameter γ_k in Step 2 of Algorithm 2.1. This choice is important for practical efficiency of any bundle method, but it is not central to the subject of this paper. For some rules to control γ_k, see [10, 19]. Also, it is possible (and can be desirable) to use more sophisticated quadratic terms in (7). For some possibilities, see [3, 13, 17, 14, 5, 15].

In Step 3 of Algorithm 2.1 the value of the "predicted descent" is computed, and the solution of (7) is accepted as the next "best iterate" if the actual descent obtained with this point is at least a σ-fraction of the predicted (Step 6 of Algorithm 2.1). The indices of such "descent steps" are collected into the set K_d. The current-best (approximate) value of f is recorded as $\bar{f}_k$.

According to Proposition 2 below, the stopping criterion (10) in Step 4 of Algorithm 2.1 ensures that if the method terminates with some iterate x^k, then we have

$$d^k \in \partial_{\varepsilon_k} f(x^k) \tag{4}$$

such that

$$\Delta_k := \frac{1}{2\gamma_k}\|d^k\|^2 + \varepsilon_k \leq \Delta_{\text{opt}}, \tag{5}$$

where d^k is computed using the solution of (14) (dual to (7), see Lemma 1). Relations (4), (5) essentially constitute the usual stopping test in bundle methods, e.g., [4, Chapter 7]. In Theorem 3 below, we show that our stopping test is guaranteed to be eventually satisfied if the approximations are controlled according to the following rule:

$$\frac{1-\sigma}{2(2-\sigma)}\Delta_{\text{opt}} > \limsup_k \left(\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}\right), \tag{6}$$

where $\sigma \in (0,1)$ is the relaxation parameter used in the descent test (Step 6 of Algorithm 2.1). In other words, if the approximation of the objective and subgradient values is (asymptotically) accurate enough in the sense of (6), then Algorithm 2.1 terminates finitely with a point x^k satisfying (4), (5). We emphasize that the required precision is finite ($\tilde{\varepsilon}_k$ need not tend to zero) and that we show precisely how small/large it needs to be for computing the desired approximate solution of (1). At the end of the paper, our results are further translated into the language of stability analysis.

Algorithm 2.1 *Choose some* $\sigma \in (0,1)$, $x^0 \in \Re^n$, *and set* K_d, B_{-1}^c, B_0^a, $I_0^a := \emptyset$. *Choose an integer* $B_{\max} \geq 2$. *Compute some* $\tilde{f}_0 \in \Re$ *and* $y^0 \in \Re^n$ *such that* $f(x^0) \geq \tilde{f}_0 \geq f(x^0) - \tilde{\varepsilon}_0$, $\tilde{\varepsilon}_0 \geq 0$, $f(x) \geq l_0(x) := \tilde{f}_0 + \langle y^0, x - x^0\rangle \; \forall\, x \in \Re^n$. *Set* $z^0 = x^0$, $\bar{f}_0 := \tilde{f}_0$ *and* $k := 0$.

1. `Add the new piece to the approximation of` f.
Set

$$B_k^c := B_{k-1}^c \cup \{l_k(x)\}, \quad l_k(x) := \tilde{f}_k + \langle y^k, x - z^k \rangle,$$
$$I_k^c := \{0 \le i \le k \mid l_i(x) \in B_k^c\}.$$

2. `Minimize the regularized cutting planes approximation of` f.
Choose $\gamma_k > 0$ *and compute* z^{k+1} *as the solution of*

$$\min_x \{\varphi_k(x) + \frac{\gamma_k}{2}\|x - x^k\|^2 \mid x \in \Re^n\}, \tag{7}$$

where

$$\varphi_k(x) := \max\{\max\{l_i(x) \mid i \in I_k^c\}, \max\{l_i^a(x) \mid i \in I_k^a\}\}. \tag{8}$$

3. `Compute the value of ‘‘predicted descent’’.`

$$\delta_k := \bar{f}_k - \varphi_k(z^{k+1}) - \frac{\gamma_k}{2}\|z^{k+1} - x^k\|^2. \tag{9}$$

4. `Stopping test.`
Stop if

$$\delta_k + 2\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\} \le \Delta_{\text{opt}}. \tag{10}$$

Otherwise, go to step 5.

5. `Approximate the values of` f `and of its subgradient at` z^{k+1}.
Compute some $\tilde{f}_{k+1} \in \Re$ *and* $y^{k+1} \in \Re^n$ *such that*

$$\begin{aligned} f(z^{k+1}) \ge \tilde{f}_{k+1} \ge f(z^{k+1}) - \tilde{\varepsilon}_{k+1}, \quad \tilde{\varepsilon}_{k+1} \ge 0, \\ f(x) \ge l_{k+1}(x) := \tilde{f}_{k+1} + \langle y^{k+1}, x - z^{k+1} \rangle \; \forall\, x \in \Re^n. \end{aligned} \tag{11}$$

6. `Descent test.`
If

$$\bar{f}_k - \tilde{f}_{k+1} - \tilde{\varepsilon}_{k+1} \ge \sigma \delta_k, \tag{12}$$

set $x^{k+1} := z^{k+1}$, $\bar{f}_{k+1} := \tilde{f}_{k+1}$, $K_d := K_d \cup \{k+1\}$, *and go to Step 8.*

7. `Null step.`
Set $x^{k+1} := x^k$ *and* $\bar{f}_{k+1} := \bar{f}_k$.

8. `Managing the complexity of the approximation of` f.
If $|B_k^c| + |B_k^a| < B_{\max}$, *then set* $B_{k+1}^a := B_k^a$ *and go to Step 9.*
Otherwise, choose some $C_k \subset B_k^c \cup B_k^a$ *such that* $|C_k| \ge 2$, $l_{k_0}(x) \notin C_k$, *where* $k_0 = \max\{i \mid i \in K_d\}$. *Set*

$$B_k^c := B_k^c \setminus C_k, \quad B_{k+1}^a := (B_k^a \setminus C_k) \cup \{l_k^a(x)\},$$
$$l_k^a(x) := \varphi_k(z^{k+1}) + \gamma_k\|z^{k+1} - x^k\|^2 + \gamma_k\langle x^k - z^{k+1}, x - x^k \rangle,$$
$$I_{k+1}^a := \{0 \le i \le k+1 \mid l_i^a(x) \in B_{k+1}^a\}.$$

9. Set $k := k+1$ and go to step 1.

Note that the value of δ_k given by (9) is not necessarily nonnegative in the given setting. Another observation is that the stopping condition (10) is also somewhat unusual. According to Proposition 2, we could alternatively just use the more traditional (4), (5). However, this would require already at this stage all the objects defined in Lemma 1 below. We thus prefer the more compact form of (10).

3 Convergence Properties

We start with the dual characterization of the solution of (7), which is similar to what can be found in the literature on bundle methods, e.g., [4, Chapter 7]. Here, we adapt it to our purposes and give a proof for completeness, as some of the expressions will be needed later on in any case.

Let $(\alpha_i, v^i, u^i) \in \Re \times \Re^n \times \Re^n$, $i = 1, \ldots, |B_k^c| + |B_k^a|$, be the data defining φ_k in (8). We shall use the representation

$$\alpha_i + \langle v^i, x - u^i \rangle = \begin{cases} l_i(x) \in B_k^c, & i \in I_k^c, \\ l_i^a(x) \in B_k^a, & i \in I_k^a. \end{cases} \tag{13}$$

Note that formally speaking, there are overlapping indices in (13), since $I_k^c \cap I_k^a \neq \emptyset$. While keeping in mind this detail, we prefer to use the representation (13), as it does not lead to any real confusion, while simplifying notation. Let I_k be the union of all the indices defining φ_k ($|I_k| = |B_k^c| + |B_k^a|$), and consider the quadratic program

$$\max_{\lambda} \left\{ -\frac{1}{2\gamma_k} \| \sum_{i \in I_k} \lambda_i v^i \|^2 + \sum_{i \in I_k} \lambda_i (\alpha_i + \langle v^i, x^k - u^i \rangle) \;\middle|\; \lambda \geq 0, \sum_{i \in I_k} \lambda_i = 1 \right\}, \tag{14}$$

where $\lambda \in \Re^{|I_k|}$.

Lemma 1 *Let $\bar{\lambda}^k$ be a solution of* (14)*, and denote*

$$d^k := \sum_{i \in I_k} \bar{\lambda}_i^k v^i.$$

It holds that

$$z^{k+1} = x^k - \frac{1}{\gamma_k} d^k, \tag{15}$$

$$d^k \in \partial \varphi_k(z^{k+1}), \tag{16}$$

$$d^k \in \partial_{\varepsilon_k} f(x^k), \tag{17}$$

where $\varepsilon_k = \varepsilon_k^c + \varepsilon_k^a \geq 0$,

$$\varepsilon_k^c = \sum_{i \in I_k^c} \bar\lambda_i^k (f(x^k) - f(z^i) - \langle y^i, x^k - z^i \rangle + \tilde\varepsilon_i) \geq 0,$$

$$\varepsilon_k^a = \sum_{i \in I_k^a} \bar\lambda_i^k (f(x^k) - \varphi_i(z^{i+1}) - \frac{1}{\gamma_i}\|d^i\|^2 - \langle d^i, x^k - x^i \rangle) \geq 0.$$

Proof. Problem (7) is equivalent to the convex quadratic program

$$\min_{x,t} \{t + \frac{\gamma_k}{2}\|x - x^k\|^2 \mid l_i(x) \leq t, i \in I_k^c,\ l_i^a(x) \leq t, i \in I_k^a\}.$$

The dual of this program (e.g., [16]) is

$$\max_{x,t,\lambda} \left\{ \begin{array}{l} t + \frac{\gamma_k}{2}\|x - x^k\|^2 \\ + \sum_{i \in I_k} \lambda_i(\alpha_i + \langle v^i, x^k - u^i \rangle - t) \end{array} \;\middle|\; \begin{array}{l} \gamma_k(x - x^k) + \sum_{i \in I_k} \lambda_i v^i = 0, \\ \sum_{i \in I_k} \lambda_i = 1,\ \lambda \geq 0. \end{array} \right\}, \tag{18}$$

with $(\alpha_i, v^i, u^i) \in \Re \times \Re^n \times \Re^n$, $i \in I_k$, defined according to (13). It can be easily seen that problem (18) reduces to (14) after eliminating the variables x and t via some simple transformations.

Assertion (15) of the lemma is now evident from the constraints in (18), using also the uniqueness of z^{k+1} and strong duality. Assertion (16) then follows from the optimality condition for (7). We proceed to prove the last assertion.

We first verify, by induction, that

$$f(x) \geq \varphi_k(x) \quad \forall\, x \in \Re^n,\ \forall\, k. \tag{19}$$

For $k = 0$ this is obvious, because $\varphi_0(x) = l_0(x)$. Suppose now that (19) holds for some index k. If $|B_k^c| + |B_k^a| < B_{\max}$, then

$$\varphi_{k+1}(x) = \max\{\varphi_k(x),\ l_{k+1}(x)\} \leq f(x),$$

by (11) and (19). If $|B_k^c| + |B_k^a| = B_{\max}$, then

$$\varphi_{k+1}(x) \leq \max\{\varphi_k(x),\ l_k^a(x),\ l_{k+1}(x)\} \leq f(x),$$

because

$$l_k^a(x) = \varphi_k(z^{k+1}) + \langle d^k, x - z^{k+1} \rangle \leq \varphi_k(x),$$

by (16). Hence, (19) holds.

By convex programming duality (e.g., [16]), the optimal values of (7) and (14) are equal, from which we obtain that

$$\varphi_k(z^{k+1}) = -\frac{1}{\gamma_k}\|d^k\|^2 + \sum_{i \in I_k} \bar\lambda_i^k (\alpha_i + \langle v^i, x^k - u^i \rangle). \tag{20}$$

By (19), we further obtain that for all $x \in \Re^n$,

$$\begin{aligned}
f(x) &\geq \varphi_k(x) \geq \varphi_k(z^{k+1}) + \langle d^k, x - z^{k+1} \rangle \\
&= \langle d^k, x - x^k \rangle + \sum_{i \in I_k} \bar{\lambda}_i^k (\alpha_i + \langle v^i, x^k - u^i \rangle) \\
&= f(x^k) + \langle d^k, x - x^k \rangle - \sum_{i \in I_k} \bar{\lambda}_i^k (f(x^k) - \alpha_i - \langle v^i, x^k - u^i \rangle) \\
&\geq f(x^k) + \langle d^k, x - x^k \rangle \\
&\quad - \sum_{i \in I_k^c} \bar{\lambda}_i^k (f(x^k) - f(z^i) - \langle y^i, x^k - z^i \rangle + \tilde{\varepsilon}_i) \\
&\quad - \sum_{i \in I_k^a} \bar{\lambda}_i^k (f(x^k) - \varphi_i(z^{i+1}) - \frac{1}{\gamma_i} \|d^i\|^2 - \langle d^i, x^k - x^i \rangle) \\
&= f(x^k) + \langle d^k, x - x^k \rangle - \varepsilon_k^c - \varepsilon_k^a,
\end{aligned}$$

where the second inequality is by (16), the first equality is by (20) and (15), and the third inequality is by (13) and (11). The quantity ε_k^c is nonnegative, because $y^i \in \partial_{\tilde{\varepsilon}_i} f(z^i)$, by (11). To prove (17), it remains to estimate ε_k^a. We have that

$$\begin{aligned}
\varepsilon_k^a &= \sum_{i \in I_k^a} \bar{\lambda}_i^k (f(x^k) - \varphi_i(z^{i+1}) - \frac{1}{\gamma_i} \|d^i\|^2 - \langle d^i, x^k - x^i \rangle) \\
&= \sum_{i \in I_k^a} \bar{\lambda}_i^k (f(x^k) - \varphi_i(z^{i+1}) - \langle d^i, x^k - z^{i+1} \rangle) \\
&\geq \sum_{i \in I_k^a} \bar{\lambda}_i^k (f(x^k) - \varphi_i(x^k)) \\
&\geq 0,
\end{aligned}$$

where the second equality is by (15), the first inequality is by (16), and the last is by (19). ■

The next result shows that the stopping test of Algorithm 2.1 guarantees the desired accuracy in the solution of problem (1).

Proposition 2 *Suppose that at some iteration k the stopping criterion* (10) *is satisfied. Then x^k is an approximate solution of* (1) *in the sense of* (4), (5), *where d^k is defined in Lemma 1.*

Proof. Let $k_0 = \max\{i \mid i \in K_d\}$ be the index of the last descent step before (10) was satisfied. By (9) and taking into account that $\bar{f}_k = \bar{f}_{k_0}$, we have that

$$\delta_k = \bar{f}_k - \varphi_k(z^{k+1}) - \frac{1}{2\gamma_k} \|d^k\|^2$$

$$\begin{aligned} &= \bar{f}_{k_0} + \frac{1}{2\gamma_k}\|d^k\|^2 - \sum_{i\in I_k} \bar{\lambda}_i^k(\alpha_i + \langle v^i, x^k - u^i\rangle) \\ &\geq \frac{1}{2\gamma_k}\|d^k\|^2 + \sum_{i\in I_k} \bar{\lambda}_i^k(f(x^{k_0}) - \alpha_i - \langle v^i, x^k - u^i\rangle) - \tilde{\varepsilon}_{k_0}, \end{aligned}$$

where the second equality is by (20) and the inequality is by (11). Since $x^{k_0} = x^k$, we further obtain

$$\begin{aligned} \sum_{i\in I_k^c} \bar{\lambda}_i^k(f(x^{k_0}) - \alpha_i - \langle v^i, x^k - u^i\rangle) &= \sum_{i\in I_k^c} \bar{\lambda}_i^k(f(x^k) - \tilde{f}_i - \langle y^i, x^k - z^i\rangle) \\ &\geq \sum_{i\in I_k^c} \bar{\lambda}_i^k(f(x^k) - f(z^i) - \langle y^i, x^k - z^i\rangle) \\ &= \varepsilon_k^c - \sum_{i\in I_k^c} \bar{\lambda}_i^k \tilde{\varepsilon}_i \\ &\geq \varepsilon_k^c - \max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}, \end{aligned}$$

where the first inequality is by (11). Taking also into account that

$$\sum_{i\in I_k^a} \bar{\lambda}_i^k(f(x^k) - \alpha_i - \langle v^i, x^k - u^i\rangle) = \varepsilon_k^a,$$

we conclude that

$$\begin{aligned} \delta_k &\geq \frac{1}{2\gamma_k}\|d^k\|^2 + \varepsilon_k^c + \varepsilon_k^a - \tilde{\varepsilon}_{k_0} - \max\{\tilde{\varepsilon}_i \mid i \in I_k^c\} \\ &\geq \frac{1}{2\gamma_k}\|d^k\|^2 + \varepsilon_k - 2\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}, \end{aligned}$$

where the last inequality follows from the fact that $k_0 \in I_k^c$, by Step 8 of Algorithm 2.1. The last relation implies (5) whenever (10) holds. ■

Given Proposition 2, to validate our claims it remains to prove that the stopping test (10) is necessarily activated if the approximation precision satisfies condition (6).

Theorem 3 *Suppose that the optimal value of* (1) *is finite, and in Algorithm 2.1* $\gamma_{k+1} \geq \gamma_k$ *and* $\tilde{\varepsilon}_{k+1} \leq \tilde{\varepsilon}_k$ *for all* k *sufficiently large,* $k \notin K_d$*.*

Then the following two events are mutually exclusive:

1. *Algorithm 2.1 performs an infinite number of iterations.*

2. *Condition* (6) *holds.*

Proof. Suppose that $\{x^k\}$ is infinite, which means that (10) is never satisfied, i.e.,

$$\delta_k + 2\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\} > \Delta_{\text{opt}} \quad \forall\, k. \tag{21}$$

Consider first the case where there is an infinite number of descent steps. For all $k \in K_d$, (12) and (11) imply that

$$\begin{aligned}\sigma\delta_k &\le \bar{f}_k - (\tilde{f}_{k+1} + \tilde{\varepsilon}_{k+1}) \\ &\le f(x^k) - f(x^{k+1}).\end{aligned}$$

Hence,

$$\sigma \sum_{k\in K_d} \delta_k \le f(x^0) - \min_x \{f(x) \mid x \in \Re^n\} < +\infty.$$

In particular, it follows that $\liminf_{k\in K_d} \delta_k \le 0$. By (21), we then have that

$$\limsup_{k\in K_d} (\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}) \ge \frac{\Delta_{\text{opt}}}{2},$$

which contradicts (6).

Suppose now that the number of descent steps is finite and $k_0 = \max\{i \mid i \in K_d\}$ is the index of the last descent iterate, so that $x^k = x^{k_0}$, $\bar{f}_k = \bar{f}_{k_0}$ for all $k \ge k_0$. For all k large enough, say $k \ge k_1$, we have that

$$\begin{aligned}\bar{f}_k - \delta_k + \frac{\gamma_k}{2}\|z^{k+2} - z^{k+1}\|^2 &= \varphi_k(z^{k+1}) + \frac{\gamma_k}{2}(\|z^{k+1} - x^{k_0}\|^2 + \|z^{k+2} - z^{k+1}\|^2) \\ &= \varphi_k(z^{k+1}) + \frac{\gamma_k}{2}(2\langle x^{k_0} - z^{k+1}, z^{k+2} - z^{k+1}\rangle + \|z^{k+2} - x^{k_0}\|^2) \\ &\le \varphi_k(z^{k+1}) + \langle d^k, z^{k+2} - z^{k+1}\rangle + \frac{\gamma_{k+1}}{2}\|z^{k+2} - x^{k_0}\|^2,\end{aligned} \tag{22}$$

where the first equality is by (9), and the inequality follows from (15) and $\gamma_k \le \gamma_{k+1}$. Furthermore, if $|B_k^c| + |B_k^a| < B_{\max}$ then

$$\begin{aligned}\varphi_k(z^{k+1}) + \langle d^k, z^{k+2} - z^{k+1}\rangle &\le \varphi_k(z^{k+2}) \\ &\le \max\{\varphi_k(z^{k+2}), l_{k+1}(z^{k+2})\} \\ &= \varphi_{k+1}(z^{k+2}),\end{aligned}$$

where the first inequality is by (16). If $|B_k^c| + |B_k^a| = B_{\max}$ then

$$\varphi_k(z^{k+1}) + \langle d^k, z^{k+2} - z^{k+1}\rangle = l_k^a(z^{k+2}) \le \varphi_{k+1}(z^{k+2}).$$

Combining those two cases, (22) implies that

$$\begin{aligned}\bar{f}_k - \delta_k + \frac{\gamma_k}{2}\|z^{k+2} - z^{k+1}\|^2 &\leq \varphi_{k+1}(z^{k+2}) + \frac{\gamma_{k+1}}{2}\|z^{k+2} - x^{k_0}\|^2 \\ &= \bar{f}_k - \delta_{k+1},\end{aligned}$$

where the equality is by (9). Hence, for $k \geq k_1$,

$$\delta_k \geq \delta_{k+1} + \frac{\gamma_k}{2}\|z^{k+2} - z^{k+1}\|^2 \geq \delta_{k+1}. \tag{23}$$

In particular, the sequence $\{\delta_k\}$ is nonincreasing (for $k \geq k_1$), and it is bounded below, by (21) and (6). Therefore, it converges:

$$\bar{\delta} = \lim_{k\to\infty} \delta_k. \tag{24}$$

We next show that the sequence $\{z^k\}$ is bounded. We have that

$$\begin{aligned}\bar{f}_k - \delta_k + \frac{\gamma_k}{2}\|z^{k+1} - x^{k_0}\|^2 &= \varphi_k(z^{k+1}) + \gamma_k\|z^{k+1} - x^{k_0}\|^2 \\ &= \varphi_k(z^{k+1}) + \langle d^k, x^{k_0} - z^{k+1}\rangle \\ &\leq \varphi_k(x^{k_0}),\end{aligned}$$

where the inequality follows from (16). Hence, for k large enough ($k \geq k_1$), it holds that

$$\begin{aligned}\|z^{k+1} - x^{k_0}\|^2 &\leq \frac{2}{\gamma_k}\left(\delta_k + \varphi_k(x^{k_0}) - \bar{f}_{k_0}\right) \\ &\leq \frac{2}{\gamma_{k_1}}\left(\delta_{k_1} + f(x^{k_0}) - \bar{f}_{k_0}\right) \\ &\leq \frac{2}{\gamma_{k_1}}\left(\delta_{k_1} + \tilde{\varepsilon}_{k_0}\right),\end{aligned}$$

where the second inequality is by (19) and by the monotonicity of $\{\gamma_k\}$ and $\{\delta_k\}$ (for $k \geq k_1$), and the last is by (11). Thus $\{z^k\}$ is bounded.

Since for $k \geq k_0$ the descent test (12) is never satisfied, we have that

$$\bar{f}_k - \tilde{f}_{k+1} - \tilde{\varepsilon}_{k+1} < \sigma\delta_k.$$

Hence,

$$\begin{aligned}(1-\sigma)\delta_k &< \tilde{f}_{k+1} + \tilde{\varepsilon}_{k+1} - \bar{f}_k + \delta_k \\ &\leq f(z^{k+1}) + \tilde{\varepsilon}_{k+1} - \varphi_k(z^{k+1}) - \frac{\gamma_k}{2}\|z^{k+1} - x^{k_0}\|^2 \\ &\leq f(z^{k+1}) - \tilde{f}_k + \varphi_k(z^k) - \varphi_k(z^{k+1}) + \tilde{\varepsilon}_{k+1} \\ &\leq f(z^{k+1}) - f(z^k) + \tilde{\varepsilon}_k + \varphi_k(z^k) - \varphi_k(z^{k+1}) + \tilde{\varepsilon}_{k+1} \\ &\leq 2L\|z^{k+1} - z^k\| + 2\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\},\end{aligned}$$

where the second inequality is by (11) and (9), the third is by $\varphi_k(z^k) \geq l_k(z^k) = \tilde{f}_k$, the fourth by (11), and the last by the Lipschitz-continuity of f and φ_k (taking into account that $\{z^k\}$ is bounded).

Combining the latter relation with (23), we obtain that

$$\delta_k - \delta_{k+1} \geq \frac{\gamma_k}{8L^2}((1-\sigma)\delta_{k+1} - 2\max\{\tilde{\varepsilon}_i \mid i \in I_{k+1}^c\})^2.$$

Since, by (24), $\delta_k - \delta_{k+1} \to 0$, and taking also into account that $\gamma_k \geq \gamma_{k_1} > 0$, the relation above implies that

$$\lim_{k\to\infty} \delta_k = \bar{\delta} = \frac{2}{1-\sigma} \lim_{k\to\infty} (\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}).$$

Passing onto the limit in (21), we then obtain

$$\left(2 + \frac{2}{1-\sigma}\right) \lim_{k\to\infty} (\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}) \geq \Delta_{\text{opt}},$$

in contradiction with (6). ■

As a final remark, we note that the presented analysis also provides a stability result for the proximal bundle algorithm. Indeed, instead of stating the question about how small should be the approximation errors $\tilde{\varepsilon}$ in order to guarantee that the iterates would eventually satisfy the given optimality tolerance Δ_{opt}, we could pose the following closely related questions. Given an upper bound $\epsilon > 0$ for the approximation errors $\tilde{\varepsilon}$, i.e.,

$$\epsilon \geq \limsup_k (\max\{\tilde{\varepsilon}_i \mid i \in I_k^c\}),$$

what are the convergence properties of the proximal bundle method? What kind of an approximate solution can be obtained? How does it depend on the value of ϵ? Our results show that using the bundle method we are guaranteed to find an approximate solution of (1) in the sense of (4), (5), with

$$\Delta_{\text{opt}} = \frac{2(2-\sigma)\epsilon}{1-\sigma} + t, \quad t > 0 \text{ arbitrarily small.}$$

Our analysis applies even when the value of ϵ can be relatively large. One example where the latter is possible are problems with noisy data, where the bound on the magnitude of perturbations is known, but usually not controllable (in particular, it cannot be driven to zero).

References

[1] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, Massachusetts, 1995.

[2] D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation.* Prentice–Hall, Inc, Englewood Cliffs, New Jersey, 1989.

[3] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. A family of variable metric proximal point methods. *Mathematical Programming*, 68:15–47, 1995.

[4] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Optimisation Numérique: aspects théoriques et pratiques.* Springer–Verlag, Berlin, Germany, 1997.

[5] X. Chen and M. Fukushima. Proximal quasi-Newton methods for nondifferentiable convex optimization. *Mathematical Programming*, 85:313–334, 1999.

[6] M. Hintermüller. A proximal bundle method based on approximate subgradients. *Computational Optimization and Applications*, 20:245–266, 2001.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms.* Springer–Verlag, Berlin, Germany, 1993.

[8] K.C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization.* Lecture Notes in Mathematics, Vol. 1133. Springer–Verlag, Berlin, Germany, 1985.

[9] K.C. Kiwiel. A method for solving certain quadratic programming problems arising in nonsmooth optimization. *IMA Journal of Numerical Analysis*, 6:137–152, 1986.

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

[11] K.C. Kiwiel. Approximations in proximal bundle methods and decomposition of convex programs. *Journal of Optimization Theory and Applications*, 84:529–548, 1995.

[12] C. Lemaréchal. Lagrangian decomposition and nonsmooth optimization: bundle algorithm, prox iteration, augmented Lagrangian. In F. Giannessi, editor, *Nonsmooth Optimization: Methods and Applications*, pages 201–216. Gordon and Breach, Philadelphia, PA, 1992.

[13] C. Lemaréchal and C. Sagastizábal. An approach to variable metric bundle methods. In J. Henry and J.-P. Yvon, editors, *Lecture notes in Control and Information Sciences No. 197, System Modelling and Optimization*, pages 144–162. Springer, Berlin, 1994.

[14] C. Lemaréchal and C. Sagastizábal. Variable metric bundle methods : From conceptual to implementable forms. *Mathematical Programming*, 76:393–410, 1997.

[15] L. Lukšan and J. Vlček. Globally convergent variable metric method for convex nonsmooth unconstrained optimization. *Journal of Optimization Theory and Applications*, 102:593–613, 1999.

[16] O.L. Mangasarian. *Nonlinear Programming.* McGraw–Hill, New York, 1969.

[17] R. Mifflin. A quasi-second-order proximal bundle algorithm. *Mathematical Programming*, 73:51–72, 1996.

[18] S.A. Miller. *An Inexact Bundle Method for Solving Large Structured Linear Matrix Inequalities.* PhD thesis, University of California, Santa Barbara, 2001.

[19] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2:121–152, 1992.

[20] M. V. Solodov and S. K. Zavriev. Error stabilty properties of generalized gradient-type algorithms. *Journal of Optimization Theory and Applications*, 98:663–680, 1998.

[21] M.V. Solodov. Convergence analysis of perturbed feasible descent methods. *Journal of Optimization Theory and Applications*, 93:337–353, 1997.