Computational Optimization and Applications, 20, 245–266, 2001

A Proximal Bundle Method Based on Approximate Subgradients

MICHAEL HINTERMÜLLER michael.hintermueller@kfunigraz.ac.at
Institute of Mathematics, Karl-Franzens-University of Graz, Heinrichstr. 36, A-8010 Graz, Austria

Received March 18, 1999; Revised April 18, 2000; Accepted October 24, 2000

Abstract. In this paper a proximal bundle method is introduced that is capable to deal with approximate subgradients. No further knowledge of the approximation quality (like explicit knowledge or controllability of error bounds) is required for proving convergence. It is shown that every accumulation point of the sequence of iterates generated by the proposed algorithm is a well-defined approximate solution of the exact minimization problem. In the case of exact subgradients the algorithm behaves like well-established proximal bundle methods. Numerical tests emphasize the theoretical findings.

Keywords: approximate subgradient, convex programming, nonsmooth optimization, proximal bundle method

1. Introduction

In the area of continuous optimization many real-life problems can be written as

$$\text{minimize} \quad f(x) \quad \text{over} \quad x \in \mathbb{R}^n \tag{1}$$

with a convex and not necessarily differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We refer to [12, 15] for examples and references. This fact gave reason for studying and developing algorithms to solve (1) numerically; see [4, 5, 7, 15, 16] and the references given there. Frequently used methods in practice comprise bundle-type methods which have (among other issues) the advantage of an implementable stopping rule; see [7, 10, 15]. While some quasi-second order methods are difficult to implement [13], or implementable forms are difficult to analyze [11], bundle methods based on the stabilized cutting plane idea are numerically and theoretically well understood [7, 9, 12, 15]. The stabilizing term which is similar to a prox-type regularization of f (and thus is also called the *proximal term*) may be considered to imitate (in a simple way) the second order behavior of f. Since this term involves an additional parameter that has to be adjusted appropriately several authors [7, 9, 15] have developed automatic adjustment schemes.

All aforementioned methods rely on the common requirement of having a finite process (subroutine or black box more general) at one's disposal which provides at a given $x \in \mathbb{R}^n$ the function value $f(x)$ and one (arbitrary) subgradient $g(x) \in \partial f(x)$, where (with $v^T w = \sum_{i=1}^n v_i w_i$ for $v, w \in \mathbb{R}^n$ and $\|v\| = \sqrt{v^T v}$) $\partial f(x) = \{g \in \mathbb{R}^n \mid g^T(z - x) \leq f(z) - f(x)$ for all $z \in \mathbb{R}^n\}$ is the subdifferential of f, i.e. the generalization of $\nabla f(x)$ in case that the convex f is not differentiable at x. For details on ∂f we refer to [3, 14].

In this paper we propose and analyze an algorithm that is capable to deal with approximate subgradients, i.e. $g_\varepsilon(x) \in \partial_\varepsilon f(x)$ for $\varepsilon > 0$, with

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n \mid g^T(z-x) \leq f(z) - f(x) + \varepsilon \quad \text{for all } z \in \mathbb{R}^n\} \tag{2}$$

which is an approximation of $\partial f(x)$ in the sense that it assembles all of the (sub)gradient information of a ball around x with radius (of at least) $\varepsilon/2L$. Here, L denotes the Lipschitz rank of f near x. Then it is well-known that an ε-*optimal* solution $\bar{x}$ of (1) is characterized by

$$0 \in \partial_\varepsilon f(\bar{x}) \quad \Longleftrightarrow \quad f(\bar{x}) \leq f(x) + \varepsilon \quad \text{for all } x \in \mathbb{R}^n\,.$$

There are several reasons for dealing with approximate (instead of exact) subgradients. If, for instance, a subgradient $g(x) \in \partial f(x)$ is expensive to compute, then one may take an already computed subgradient $g(\tilde{x})$ of f at some $\tilde{x}$ near x. Then

$$\begin{aligned} f(x) + g(\tilde{x})^T(z-x) &= f(\tilde{x}) + g(\tilde{x})^T(z-\tilde{x}) + \alpha_o(x,\tilde{x}) \\ &\leq f(z) + \alpha_o(x,\tilde{x}) \quad \text{for all } z \in \mathbb{R}^n, \end{aligned}$$

with $0 \leq \alpha_o(x,\tilde{x}) := f(x) - f(\tilde{x}) - g(\tilde{x})^T(x-\tilde{x})$. Thus, $g(\tilde{x}) \in \partial_{\alpha_o(x,\tilde{x})} f(x)$.

Another type of application arises in the following context: Assume that f is strongly convex with modulus $\mu > 0$, i.e.

$$f(x) + g(x)^T(z-x) + \frac{\mu}{2}\|z-x\|^2 \leq f(z) \quad \text{for all } x, z \in \mathbb{R}^n,\ g(x) \in \partial f(x),$$

and that $f(x) = w(v(x))$, with $v : \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable and $w : \mathbb{R}^m \to \mathbb{R}$ convex. By the chain rule [3, Theorem 2.3.9] we have $\partial f(x) = \{\sum_{i=1}^m \xi_i \nabla v_i(x) \mid \xi \in \partial w(v(x))\}$. Now assume that we have an approximation $\nabla_h v(x)$ of $\nabla v(x)$ such that $\|\nabla_h v(x) - \nabla v(x)\| \leq \kappa(h)$, $h > 0$. Such an approximation may be obtained by using finite differences. In this case, typically $\kappa(h) \to 0$ for $h \to 0$. Let $g_h(x) := \sum_{i=1}^m \xi_i \nabla_h v_i(x)$, $\xi \in \partial w(v(x))$. Then, we have

$$\begin{aligned} f(x) + g_h(x)^T(z-x) &\leq f(x) + g(x)^T(z-x) + \|\xi\|\|\nabla_h v(x) - \nabla v(x)\|\,\|z-x\| \\ &\leq f(z) - \frac{\mu}{2}\|z-x\|^2 + \kappa(h)\|\xi\|\,\|z-x\| \end{aligned} \tag{3}$$

for all $x, z \in \mathbb{R}^n$ and $g(x) = \sum_{i=1}^m \xi_i \nabla v_i(x) \in \partial f(x)$. Some simple manipulations show that

$$-\frac{\mu}{2}\|z-x\|^2 + \kappa(h)\|\xi\|\,\|z-x\| \leq \frac{1}{2\mu}\|\xi\|^2\kappa(h)^2 =: \varepsilon_h \quad \text{for all } x, z \in \mathbb{R}^n.$$

Note that by the definition of ξ the bound ε_h depends on x. We can continue (3) and obtain

$$f(x) + g_h(x)^T(z-x) \leq f(z) + \varepsilon_h \quad \text{for all } z \in \mathbb{R}^n.$$

From the local boundedness of $\partial w(v(x))$ (which implies that $\|\xi\|$ is bounded) we infer that $\varepsilon_h > 0$ is locally bounded. Thus, $g_h(x)$ is an ε_h-subgradient of f at x. If x remains in a bounded subset of $\mathbb{R}^n$, then there even exists a uniform bound $\varepsilon > 0$ such that $\varepsilon_h \leq \varepsilon$.

In light of the above discussion, our aim is to develop an implementable algorithm for which we assume to have a black box providing at $x \in \mathbb{R}^n$

$$f(x) \quad \text{and one (arbitrary)} \quad g_\varepsilon(x) \in \partial_\varepsilon f(x). \tag{4}$$

Note that we *do not* assume knowledge of $\varepsilon \geq 0$.

Under an assumption similar to (4), i.e. $f(x)$ replaced by $f_\varepsilon(x)$ with $|f(x) - f_\varepsilon(x)| \leq \varepsilon$, in [6] an implementable algorithm to find a solution of (1) is proposed. In contrast to our requirements, not only controllability of ε but also explicit knowledge of the error is assumed. Moreover, the stabilization of the cutting plane model lacks an adjustable parameter. However, this last drawback is overcome in [8].

The subsequent sections are organized as follows: In Section 2 we motivate the proximal bundle method that forms the basis of the algorithm proposed in this paper. The problem of finding an adequate search direction is discussed in Section 3. Moreover, optimality aspects are considered. The core of the algorithm, i.e. the inner loop implementing the proximal bundle strategy, is the main goal of Section 4. Section 5 comprises the overall algorithm and its convergence analysis. Finally, in Section 6 we report on numerical testing of the new algorithm.

2. Proximal bundle method

As already pointed out in Section 1, the bundle concept proved to be a useful idea for tackling the numerical minimization of $f(x)$ over $x \in \mathbb{R}^n$. Bundle methods typically proceed in two phases: (i) The first phase makes use of the bundle of information $(f(x^k), g_\varepsilon(x^k))$, $(f(x^{k-1}), g_\varepsilon(x^{k-1})), \ldots$ assembled so far in order to establish a model of f at the actual iterate x^k. (ii) Due to the ambiguity of the differential information in nonsmooth optimization and therefore due to the kinky structure of f, the model possibly is not yet adequate. Then even more information around the actual iterate x^k is mobilized to obtain a more reliable model. For more details on bundle methods we refer to [4, 5, 7, 15] and the references therein.

In our situation, feature (i) naturally leads to the *ε-cutting plane* (ε-CP) *approximation* of f at x^k. Let J^k denote the index set at the kth iteration with each $j \in J^k$ representing $(y^j, f(y^j), g_\varepsilon(y^j))$. Due to the ε-subgradient inequality, i.e. the defining inequality in (2), we have for all $x \in \mathbb{R}^n$ and for all $j \in J^k$

$$f(x) \geq f(y^j) + g_\varepsilon(y^j)^T(x - y^j) - \varepsilon = f(x^k) + g_\varepsilon(y^j)^T(x - x^k) - \alpha_{k,j} - \varepsilon, \tag{5}$$

where

$$\alpha_{k,j} := \alpha(x^k, y^j) = f(x^k) - f(y^j) - g_\varepsilon(y^j)^T(x^k - y^j) \geq -\varepsilon. \tag{6}$$

Since $\alpha_{k,j}$ corresponds to the error at x^k when linearizing at y^j, we will refer to $\alpha_{k,j}$ as *linearization error*. Let $d := x - x^k$, then for all d

$$f(x^k + d) \geq f(x^k) + \max_{j \in J^k} \left\{ g_\varepsilon(y^j)^T d - \alpha_{k,j} \right\} - \varepsilon =: f_\varepsilon^{\mathrm{CP}}(x^k; d), \tag{7}$$

where $f_\varepsilon^{\mathrm{CP}}(x^k; d)$ corresponds to the ε-CP approximation of f.

The ε-CP model usually becomes more and more crude an approximation of f the farther away from x^k. For this reason the proximal term $\frac{1}{2t^k}\|d\|^2$, $t^k > 0$, is introduced. A suitable adjustment of t^k shall bound the ε-CP model to the area where it is a reliable approximation of f. Then

$$d^k = \operatorname{argmin}\{ f_\varepsilon^{\mathrm{CP}}(x^k; d) + \frac{1}{2t^k}\|d\|^2 \mid d \in \mathbb{R}^n \} \tag{8}$$

is expected to be a descent direction for f.

However, due to the kinky structure of f the ε-CP model may be such a poor approximation of f that d^k is not a descent direction for f. In this situation, by feature (ii) the bundle is enriched with information around the actual iterate. As a result a more useful direction is computed subsequently. For more details we refer to [7, 9, 15].

3. Quadratic programming

We start by considering the minimization problem in (8). For convenience the shortened forms f^k and g_ε^j are used instead of $f(x^k)$ and $g_\varepsilon(y^j)$. At first we will skip the constant terms f^k and ε in $f_\varepsilon^{\mathrm{CP}}(x^k; d)$. The ε-CP approximation therefore reduces to

$$\hat{f}_\varepsilon^k(d) := \max_{j \in J^k} \left\{ {g_\varepsilon^j}^T d - \alpha_{k,j} \right\}.$$

Note that $\hat{f}_\varepsilon^k(0) \leq \varepsilon$ due to (6). Let $t > 0$ denote a suitable parameter. Then the search direction $d(t)$ will be computed by

$$d(t) := \operatorname{argmin}\left\{ \hat{f}_\varepsilon^k(d) + \frac{1}{2t}\|d\|^2 \mid d \in \mathbb{R}^n \right\}.$$

This problem can equivalently be written as a quadratic programming problem in $\mathbb{R} \times \mathbb{R}^n$ by

$$\begin{aligned} & v + \frac{1}{2t}\|d\|^2 = \min! \\ & {g_\varepsilon^j}^T d - \alpha_{kj} \leq v \quad \text{for all } j \in J^k. \end{aligned} \tag{9}$$

Let $(v(t), d(t))$ denote the optimal solution of (9) at t. It is easily seen that

$$v(t) = \hat{f}_\varepsilon^k(d(t)) = f_\varepsilon^{\mathrm{CP}}(x^k; d(t)) - f^k + \varepsilon.$$

Due to the convexity of $f_\varepsilon^{\mathrm{CP}}$ and $\hat{f}_\varepsilon^k(0) \le \varepsilon$ we find for $\tau \in [0, 1]$

$$f_\varepsilon^{\mathrm{CP}}(x^k; \tau d(t)) \le (1-\tau) f_\varepsilon^{\mathrm{CP}}(x^k; 0) + \tau f_\varepsilon^{\mathrm{CP}}(x^k; d(t)) \le f^k + \tau v(t).$$

Therefore $v(t)$ estimates the descent obtained from our model and will subsequently serve as expected descent of f.

Next we will inspect properties of (9).

Lemma 1. *(a) The unique solution $(v(t), d(t))$ of (9) always exists. (b) The pair $(v(t), d(t))$ solves (9) if and only if there exist a Lagrange multiplier $\lambda(t) \in \mathbb{R}^{|J^k|}$ with components $\lambda_j(t) \in \mathbb{R}$, $j \in J^k$, a vector $\zeta(t) \in \mathbb{R}^n$ and a scalar $\sigma(t) \in \mathbb{R}$ satisfying*

$$\lambda_j(t) \ge 0 \quad \text{for all } j \in J^k \text{ and } \sum_{j \in J^k} \lambda_j(t) = 1, \tag{10}$$

$$\left(g_\varepsilon^{j^T} d(t) - \alpha_{k,j} - v(t)\right)\lambda_j(t) = 0 \quad \text{for all } j \in J^k, \tag{11}$$

$$g_\varepsilon^{j^T} d(t) - \alpha_{k,j} \le v(t) \quad \text{for all } j \in J^k, \tag{12}$$

$$\zeta(t) = \sum_{j \in J^k} \lambda_j(t) g_\varepsilon^j, \tag{13}$$

$$\sigma(t) = \sum_{j \in J^k} \lambda_j(t) \alpha_{k,j}, \tag{14}$$

$$d(t) = -t\zeta(t), \tag{15}$$

$$v(t) = -\frac{1}{t}\|d(t)\|^2 - \sigma(t). \tag{16}$$

Proof: For the proof we refer to [5, Lemma 2.1]. □

The following continuity result on $(v(t), d(t))$ holds by strict convexity of the objective function [15]:

$$\text{The optimal solution } (v(t), d(t)) \text{ of (9) depends continuously on } t \in (0, \infty). \tag{17}$$

From (5) we obtain

$$g_\varepsilon^{j^T}(x - x^k) \le f(x) - f^k + \varepsilon + \alpha_{k,j} \quad \text{for all } x \in \mathbb{R}^n. \tag{18}$$

Multiplication of (18) with $\lambda_j(t)$, $j \in J^k$, defined by (10), and summing up over j yields (cf. (13)–(14)) for all $x \in \mathbb{R}^n$

$$\zeta(t)^T(x - x^k) \le f(x) - f^k + \varepsilon + \sigma(t). \tag{19}$$

For $x = x^k$ (19) yields

$$\sigma(t) \ge -\varepsilon, \tag{20}$$

and hence by (16) we have $v(t) \le \varepsilon$. We immediately find from (19)

Lemma 2. *Suppose for some* $\varepsilon_s \geq 0$,

$$\|\zeta(t)\| \leq \varepsilon_s \quad \textit{and} \quad \sigma(t) \leq \varepsilon_s. \tag{21}$$

Then x^k *is* ε-ε_s*-optimal, i.e.*

$$f^k \leq f(x) + \varepsilon_s \|x - x^k\| + (\varepsilon + \varepsilon_s) \quad \textit{for all } x \in \mathbb{R}^n.$$

Hence, $\varepsilon_s = 0$ implies ε-optimality of x^k which is most one can hope for when using ε-subgradients. If we assume momentarily that $\varepsilon = 0$ then $v(t) = 0$ implies optimality. But, however, in case of $\varepsilon > 0$ the condition $v(t) = 0$ need no longer imply optimality.

Now we will investigate relations between the optimal solution $(v(t), d(t))$ of (9) and the negative of the optimal value of (9)

$$w(t) := -v(t) - \frac{1}{2t}\|d(t)\|^2 = \frac{t}{2}\|\zeta(t)\|^2 + \sigma(t) \geq \sigma(t) \geq -\varepsilon, \tag{22}$$

where the second equality stems from (15)–(16) and the final inequality from (20). Hence, for $\varepsilon = 0$ we obtain $w(t) \geq 0$. For the inner iteration developed in the next section the following lemma is needed.

Lemma 3. *(a) For all* $t \in (0, \infty)$ *we have that* $w(t) > 0$ *implies* $v(t) < 0$. *(b) If* $w(t_i) \leq 0$ *for a sequence* (t_i) *with* $t_i > 0$ *and* $t_i \to \infty$ *then*

$$-\varepsilon \leq \sigma(t_i) \leq 0 \quad \textit{for all } i \textit{ and } \|\zeta(t_i)\| \to 0,$$

hence x^k *is* ε*-optimal.*

Proof: (a) From (22) it follows that

$$w(t) = -\left(v(t) + \frac{1}{2t}\|d(t)\|^2\right) > 0 \quad \Longrightarrow \quad v(t) + \frac{1}{2t}\|d(t)\|^2 < 0.$$

Hence, the required $v(t) < 0$ is obtained.

(b) Observe that by (22) and (20)

$$0 \geq w(t_i) = \frac{t_i}{2}\|\zeta(t_i)\|^2 + \sigma(t_i) \quad \Longrightarrow \quad -\varepsilon \leq \sigma(t_i) \leq 0 \quad \text{for all } i \in \mathbb{N}.$$

And $w(t_i) \leq 0$ further implies

$$\|\zeta(t_i)\| \leq \sqrt{\frac{-2\sigma(t_i)}{t_i}} \leq \sqrt{\frac{2\varepsilon}{t_i}}.$$

Hence $\|\zeta(t_i)\| \to 0$ for $i \to \infty$ which in view of (19) establishes ε-optimality of x^k. □

Remark 1. For part (b) of Lemma 3 the preceding proof shows that the stopping criterion (21) with $\varepsilon_s > 0$ is satisfied for $t_i \geq 2\varepsilon/\varepsilon_s^2$.

4. Inner iteration

In this section, we design a strategy for suitably adjusting t such that either a *serious step* yields a sufficient descent or a *null step* enriches our model. It will also be possible to detect ε-ε_s-optimal points.

At first we shall discuss several criteria by which we decide whether a serious step or a null step is taken, or t has to be modified. We already mentioned that $v(t)$ represents the descent of the ε-CP-model and will therefore serve as expected descent. Let $y_i := x^k + d_i$ denote a trial point, where the subscript i is the (running) index of the inner loop and (v_i, d_i) is the optimal solution of (9) at t_i. Provided that $v_i < 0$, a suitable descent test is

$$\text{SS(1)} \quad f(y_i) - f^k \leq \nu_1 v_i \,,$$

where $0 < \nu_1 \leq \frac{1}{2}$. If, in addition, the test

$$\text{SS(2) (a)} \quad g_\varepsilon(y_i)^T d_i \geq \nu_2 v_i \quad \text{or} \quad \text{(b)} \quad t_i \geq T_i - t_\varepsilon$$

holds then a serious step occurs, i.e. the $\boxed{\text{SS}}$ exit of the left branch of the inner loop displayed in figure 1 is taken. Here $\nu_2 \in (\nu_1, 1]$, $T_i > 0$ is the current upper bound on t_i, and $t_\varepsilon \in (0, T_i)$ is a small threshold. Note that SS(2)(a) takes care of a substantial change in the ε-CP model. If SS(1) is satisfied, but SS(2)(a) does not hold and t_i is still smaller than $T_i - t_\varepsilon$, then some larger t-value is tried.

On the contrary, if (again provided that $v_i < 0$)

$$\begin{aligned} &\text{NS(1)} \quad f(y_i) - f^k > \nu_1 v_i \quad \text{and} \\ &\text{NS(2)} \quad \text{(a)}\ \alpha(x^k, y_i) \leq \nu_3 |\sigma^{k-1}| \quad \text{or} \\ &\qquad\quad \text{(b)}\ |f^k - f(y_i)| \leq \max\left\{\|\zeta^{k-1}\|,\ \sigma^{k-1}\right\} \end{aligned}$$

a null step will improve the model. The inner loop is left at the $\boxed{\text{NS}}^1$ exit of the middle branch in figure 1. Here $0 < \nu_3 \leq 1$, $\zeta^0 := g_\varepsilon^1$ and $\sigma^0 := 0$.

If NS(1) holds with $v_i < 0$, then for all $j \in J^k$

$$g_\varepsilon(y_i)^T d_i - \alpha(x^k, y_i) = f(y_i) - f^k > \nu_1 v_i > v_i \geq g_\varepsilon^{j^T} d_i - \alpha_{k,j}.$$

Thus the bundle model augmented with $(y_i, f(y_i), g_\varepsilon(y_i))$ changes significantly.

In NS(2)(b) $\max\{\|\zeta^{k-1}\|, \sigma^{k-1}\} > 0$ can be assumed, because otherwise Lemma 2 applies with $\varepsilon_s = 0$ implying ε-optimality of x^{k-1}. The criterion NS(2) ensures that only information at points y_i near the actual iterate x^k will be added to the bundle. Therefore, if NS(2) does not hold, then t should be decreased. However, to ensure finite termination of the t-search, a decrease of t is not allowed in the following two cases. The first case occurs if SS(1) was satisfied in an earlier cycle with step-size $\tilde{t}$; then a serious step is taken with t_i reset to $\tilde{t}$ and d_i to $d(\tilde{t})$. This corresponds to the $\boxed{\text{SS}(\tilde{t})}$ exit of the middle branch in figure 1. The second case, in which a null step is enforced *even when* NS(2) fails, arises if $w(t) \leq 0$

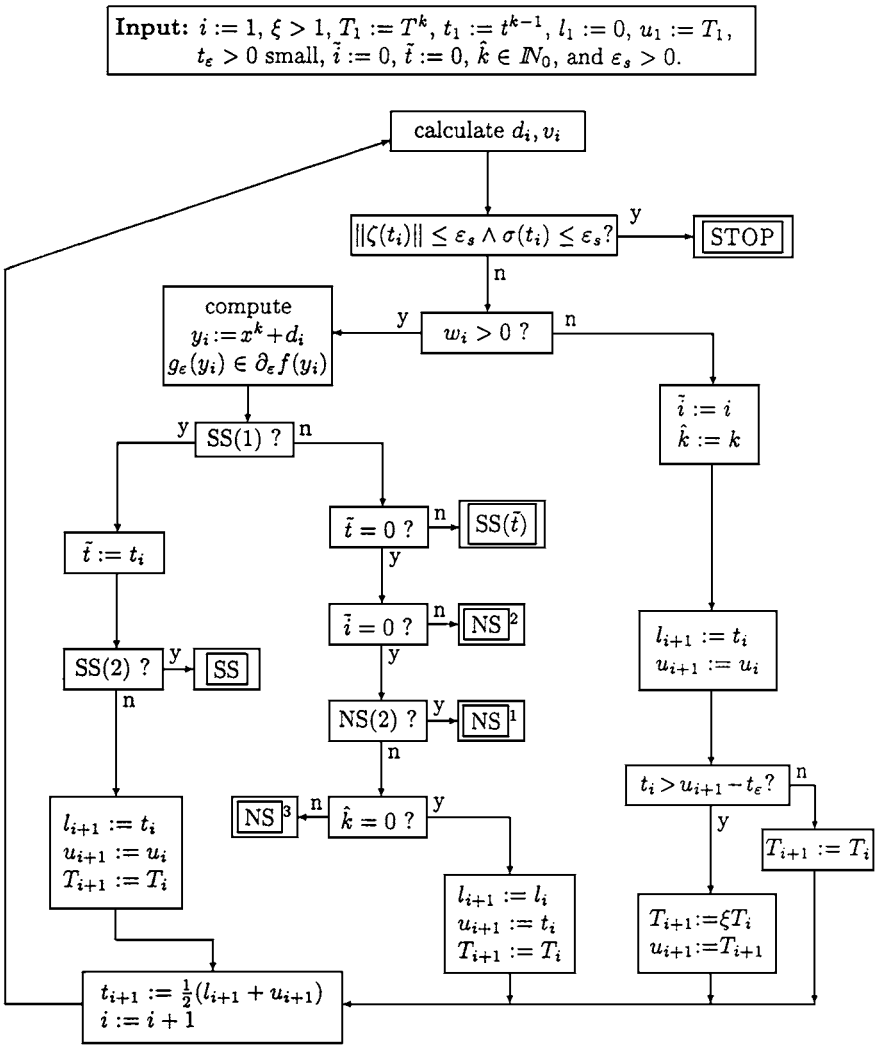

Figure 1. Inner iteration.

for some t generated either in the current inner loop or at an earlier iteration after which only null steps were taken. In the first situation, the exit NS 2 is taken, and in the latter case the exit NS 3 terminates the inner loop of figure 1. A detailed discussion is given below.

In the previous section we already mentioned that v_i may be nonnegative. As we would expect, this situation causes problems because the criteria SS and NS become meaningless. These difficulties are resolved as follows. Let $w_i := w(t_i)$ (cf. (22)). Lemma 3 implies that no problems occur whenever $w_i > 0$ because then $v_i < 0$. On the other hand, if $w_i \le 0$ then v_i could be nonnegative. When $v_i \ge 0$, a linear piece of our polyhedral approximation

exceeds f and prevents descent. Then the search region should be enlarged in order to find ε-subgradients that either yield a descent direction or confirm that the current iterate is ε-optimal. This case is detected via the indicator $\hat{k}$: if $\hat{k} \neq 0$ then $w_i \leq 0$ occurred in the inner loop of the $\hat{k}$th iteration after which only null steps were taken; each serious step sets $\hat{k} = 0$ to indicate that t may be reduced if necessary.

When examining the flow chart of the inner iteration presented in figure 1, one will find that the t-adjustment is realized by a simple bisection rule. Note that the left and the right branch increase the t-value by putting $l_{i+1} = t_i$ and $u_{i+1} \geq u_i$, while the middle branch decreases the t-value by $l_{i+1} = l_i$ and $u_{i+1} = t_i$. Then $t_{i+1} = \frac{1}{2}(l_{i+1} + u_{i+1})$ is computed. Of course, one may use a more sophisticated strategy. Moreover, to detect the cases where t_i cannot be decreased, the middle branch of figure 1 employs the indicators $\tilde{t}$ and $\hat{k}$ discussed above, as well as $\tilde{i}$ that may be updated by the right branch when $w_i \leq 0$. Details on how serious and null steps are carried out are discussed in Section 5.1.

As initial values may serve $t^0 := 1/\|g_\varepsilon^1\|$, which yields $\|d^1\| = 1$, $T^1 := 10\,t^0$, and instead of constant t_ε we use $t_\varepsilon^k := 0.05\,T^k$.

The main result of this section is the following

Theorem 1. *The inner iteration* (*represented by figure* 1) *stops after finitely many cycles, either with a serious step or a null step, or with the information that* x^k *is* ε-ε_s*-optimal.*

Proof: Suppose the cycle is not finite. Then we consider the following two cases:

i. $w_i \leq 0$ for infinitely many i. Then there exists a subsequence $(i(l))_{l\in\mathbb{N}}$ such that $w_{i(l)} \leq 0$ and $T_{i(l)} \to +\infty$ and hence $t_{i(l)} \to +\infty$ (otherwise we would have $T_{i+1} = T_i$ for all large i, $t_i - u_i \to 0$ by bisection and thus eventually $t_i > u_{i+1} - t_\varepsilon$ in the right branch, a contradiction). So the conclusion follows from Remark 1 with $\varepsilon_s > 0$.

ii. $w_i > 0$ for all large i. We consider the following cases concerning l_i:

- $l_i = 0$ for all i. Hence, NS(1) is always satisfied. This implies by the bisection rule that $t_i \downarrow 0$. By (15), (13) and (10), $d_i \to 0$, so $y_i \to x^k$ and the continuity of f yield $f(y_i) \to f^k$; therefore, NS(2)(b) holds for all large i. Thus the cycle must terminate in $\boxed{\text{NS}}^1$ with a null step, a contradiction.
- l_{i+1} becomes positive for some $i = \bar{i}$. Then, since either $\tilde{t}$ or $\tilde{i}$ become positive in the left or right branches, respectively, the middle branch can't be entered for $i > \bar{i}$ (otherwise it would terminate either in $\boxed{\text{SS}(\tilde{t})}$ or $\boxed{\text{NS}}^2$, a contradiction). Since by assumption the right branch is not executed for large i, the left branch produces by bisection $t_i - u_i \to 0$. Therefore, if $u_i = T_i$ for large i, then eventually SS(2)(b) is met and the inner loop is left at $\boxed{\text{SS}}$, a contradiction. However, $u_i < T_i$ for large i is possible only if for $i = \bar{i} - 1$ the middle branch sets $u_{\bar{i}} = t_{\bar{i}-1}$ and the other branches keep $u_i = u_{\bar{i}}$ for $i > \bar{i}$, so $t_i \to t_{\bar{i}-1}$. Hence, since NS(1) holds for $i = \bar{i} - 1$, the continuity of f and $(d(t), v(t))$ (cf. (17)) imply that NS(1) is satisfied for all large i, i.e., the middle branch is entered, a contradiction. □

5. Algorithm and convergence analysis

In this section we present the bundle-type algorithm which includes (in the inner iteration) the proximal bundle method and a strategy to keep the needed memory bounded. Afterwards a detailed convergence analysis is given. The main statement comprises the fact that each cluster point of the sequence of iterates (x^k) generated by the algorithm is ε-optimal for f on $\mathbb{R}^n$. Finally, we briefly discuss the case $\varepsilon = 0$.

5.1. The algorithm

Before we briefly summarize the overall algorithm with reset strategy, we have to specify how serious steps (SS) and null steps (NS) take place: Set $d^k := d_i$, where i is the index at which the inner iteration terminates, and then

$$\begin{aligned} \text{SS}: &\quad x^{k+1} := x^k + d^k,\ y^{k+1} := x^{k+1}, \\ \text{NS}: &\quad x^{k+1} := x^k,\ y^{k+1} := y_i. \end{aligned} \tag{23}$$

Moreover, let in both cases $T^{k+1} := T_i$, $t^k := t_i$, $v^k := v_i$, $w^k := w_i$, $\zeta^k := \zeta(t_i)$, and $\sigma^k := \sigma(t_i)$.

Next the algorithm is displayed.

Algorithm 1. *Choose a starting point $x^1 \in \mathbb{R}^n$ and parameters $0 < \nu_1 < \nu_2 < 1$, $0 < \nu_3 < 1$, $J_{\max} \geq 3$, $J_{\max} \in \mathbb{N}$, and a stopping tolerance $\varepsilon_s > 0$.*

1. *Initialization. Set $k := 1$. Compute $f(x^1)$, $g_\varepsilon(x^1) \in \partial_\varepsilon f(x^1)$ and put $y^1 := x^1$, $J^1 := \{1\}$, $\alpha_{1,1} := 0$. Choose $t^0 > 0$ and $T^1 > t^0$. Put $\hat{k} := 0$.*
2. *Inner iteration. Compute x^{k+1}, y^{k+1}, $g_\varepsilon(y^{k+1})$, T^{k+1}, t^k corresponding to (23), and update $\hat{k}$ by the inner loop in figure 1, respectively, or stop with $\varepsilon-\varepsilon_s$-optimal x^k.*
3. *Reset. If $|J^k| < J_{\max}$ go to step 4, otherwise choose $J \subset J^k$ with $|J| \leq J_{\max} - 2$ and $\max\{j \mid j \in J^k \text{ and } \alpha_{k,j} \leq 0\} \in J$. Introduce some additional index $\tilde{k}$ and define $g_\varepsilon^{\tilde{k}} := \zeta^k$, $\alpha_{k,\tilde{k}} := \sigma^k$, $J := J \cup \{\tilde{k}\}$.*
4. *Update. In case of a serious step put $\hat{k} := 0$ and*

$$\alpha_{k+1,j} := \alpha_{k,j} + f^{k+1} - f^k - g_\varepsilon^{j^T} d^k \quad \text{for all } j \in J, \alpha_{k+1,k+1} := 0.$$

In case of a null step put

$$\alpha_{k+1,j} := \alpha_{k,j} \quad \text{for all } j \in J, \alpha_{k+1,k+1} := \alpha(x^k, y^{k+1}).$$

Set $J^{k+1} := J \cup \{k+1\}$, $k = k+1$ go to step 2.

For details on how to choose $J \subset J^k$ in the reset step we refer to [5, 15].

5.2. *Convergence analysis*

The remainder of this section is devoted to the convergence analysis of Algorithm 1. Naturally, our convergence results assume that $\varepsilon_s = 0$. We shall now show that cluster points x^* (if any) of the sequence $(x^k)_{k\in\mathbb{N}}$ satisfy

$$x^* \in \mathcal{M}_\varepsilon \quad \text{with} \quad \mathcal{M}_\varepsilon := \{x \in \mathbb{R}^n \mid f(x) \le \bar{f} + \varepsilon\} \quad \text{and} \quad \bar{f} = \inf f(x)\,.$$

Here $\mathcal{M}_\varepsilon$ denotes the set of ε-optimal solutions of (1).

For $\varepsilon_s = 0$, the inner loop of figure 1 may be infinite. In this case the following lemma applies.

Lemma 4. *Suppose the inner loop is infinite for some k. Then $x^k \in \mathcal{M}_\varepsilon$.*

Proof: This follows from Lemma 3 and part (i) of the proof of Theorem 1. □

From now on we will assume that the inner iteration is finite for all k and that no termination occurs (otherwise the final x^k is in $\mathcal{M}_\varepsilon$). Several of our subsequent results need the following assumption:

$$\text{The sequence } (x^k)_{k\in\mathbb{N}} \text{ is bounded.} \tag{24}$$

The aim is to prove that

$$\lim_{k\to\infty} f^k \;\le\; \bar{f} + \varepsilon. \tag{25}$$

When (25) holds then, since the sequence (f^k) is monotonically decreasing by construction and f is continuous, each cluster point of (x^k) is an ε-optimal solution of (1). In view of (19), (25) will hold if (24) is satisfied and we find a subsequence $(k(l))_{l\in\mathbb{N}}$ for which

$$\lim_{l\to\infty} \zeta^{k(l)} = 0 \quad \text{and} \quad \limsup_{l\to\infty} \sigma^{k(l)} \le 0\,. \tag{26}$$

It is convenient to precede the main convergence result (Theorem 2) by several lemmas. We will discuss separately the case of unbounded T-growth, the case that one makes infinitely many serious steps or only finitely many serious steps, and the case that (t^k) has a zero cluster point.

Before we can treat the critical situation $T^k \to +\infty$, the more convenient case of infinitely many serious steps is analyzed in

Lemma 5. *Suppose* (24) *holds and there exists $\underline{t} > 0$ such that infinitely many serious steps are taken with $t^k \ge \underline{t}$. Then for suitable subsequences* (26) *is satisfied.*

Proof: Let $(x^{k(l)})_{l\in\mathbb{N}}$ be a subsequence resulting in serious steps, i.e.

$$f\left(x^{k(l)+1}\right) - f\left(x^{k(l)}\right) \le \nu_1 v^{k(l)}.$$

Moreover, assume that $t^{k(l)} \geq \underline{t}$ for all l. Hence for all $m \geq 1$ we obtain

$$f\left(x^{k(m)+1}\right) - f\left(x^{k(1)}\right) \leq \nu_1 \sum_{l=1}^{m} v^{k(l)}. \tag{27}$$

Observe that by (24) $\inf f^k > -\infty$. Taking $m \to +\infty$ in (27) yields

$$-\infty < \inf f^k - f\left(x^{k(1)}\right) \leq \nu_1 \sum_{l=1}^{\infty} v^{k(l)} \leq 0$$

using $v^{k(l)} < 0$. Consequently, $v^{k(l)} \to 0$. In view of (15), (16) and (22), $w^{k(l)} = \frac{t^{k(l)}}{2}\|\zeta^{k(l)}\|^2 + \sigma^{k(l)} > 0$ implies

$$-2v^{k(l)} = 2t^{k(l)}\left\|\zeta^{k(l)}\right\|^2 + 2\sigma^{k(l)} \geq t^{k(l)}\left\|\zeta^{k(l)}\right\|^2 \geq \underline{t}\left\|\zeta^{k(l)}\right\|^2.$$

This yields $t^{k(l)}\|\zeta^{k(l)}\|^2 \to 0$ and $\zeta^{k(l)} \to 0$. From $v^{k(l)} = -t^{k(l)}\|\zeta^{k(l)}\|^2 - \sigma^{k(l)}$ we conclude $\sigma^{k(l)} \to 0$. □

We remark that Lemma 5 makes no assumption whether (T^k) is bounded or not. The only condition that has to be satisfied is that there exists a lower bound $\underline{t}$ of (t^k) corresponding to an infinite sequence of serious steps. The case of unbounded growth of the sequence (T^k) is considered next.

Lemma 6. *Suppose* (24) *holds and* $T^k \to +\infty$ *for* $k \to +\infty$. *Then* (26) *holds for suitable subsequences, or there exists* $\bar{k}$ *such that* $x^* = x^k = x^{\bar{k}}$ *for all* $k \geq \bar{k}$ *and* $x^* \in \mathcal{M}_\varepsilon$; *in this case for a suitable subsequence we have*

$$\lim_{l\to\infty} \zeta\left(\bar{t}^{k(l)}\right) = 0 \quad \textit{and} \quad \limsup_{l\to\infty} \sigma\left(\bar{t}^{k(l)}\right) = 0\,,$$

with $(\zeta(\bar{t}^{k(l)}), \sigma(\bar{t}^{k(l)}))$ *computed in iteration* $k(l)$ *with* $\bar{t}^{k(l)}$ *possibly different from* $t^{k(l)}$.

Proof: We consider two cases:

i. Infinitely many serious steps are taken. Assume that $(k(l))_{l\in\mathbb{N}}$ is a subsequence such that $T^{k(l)+1} > T^{k(l)}$, i.e., the right branch of figure 1 is executed at iteration $k(l)$. Then, since t can't be reduced until the next serious step that occurs at iteration $k_+(l) \geq k(l)$, we have $t^{k_+(l)} \geq t^{k(l)} \geq \frac{1}{2}T^{k(l)} \geq \frac{1}{2}T^1 > 0$ for all l. Hence, Lemma 5 applies (to $(k_+(l))_{l\in\mathbb{N}}$) with $\underline{t} := \frac{1}{2}T^1 > 0$, yielding the first assertion.

ii. Finitely many serious steps are taken, i.e., there exists an index $\bar{k}$ such that $x^k = x^{\bar{k}}$ for all $k \geq \bar{k}$. Defining $k(l)$ as in (i), pick $\bar{l}$ such that $k(\bar{l}) > \bar{k}$. Consider the inner loop of figure 1 at iteration $k = k(l)$ with $l > \bar{l}$. Then for some $i = \bar{i}_k$, in the right branch we have $w_{\bar{i}_k} \leq 0$ and $t_{\bar{i}_k} > u_{\bar{i}_k+1} - t_\varepsilon$ with $u_{\bar{i}_k+1} = T_i = T^{k(l)}$ (since $u_1 = T_1 = T^{k(l)}$

and the middle branch can't decrease u_i due to $\hat{k} > 0$). Further, by (22), $w_{\bar{i}_k} \leq 0$ yields $\sigma(t_{\bar{i}_k}) \leq 0$ and

$$\left\|\zeta\left(t_{\bar{i}_k}\right)\right\|^2 \leq \frac{-2\sigma\left(t_{\bar{i}_k}\right)}{t_{\bar{i}_k}} \leq \frac{2\varepsilon}{t_{\bar{i}_k}}.$$

Since $T^{k(l)} \to \infty$, passing to the limit in (19) yields the conclusion. □

From now on we can assume that $T^k \leq \bar{T} < +\infty$ for all k, i.e. there exists an index k^* such that $T^k = T^{k^*} =: \bar{T}$ for all $k \geq k^*$. If we assume the sequences (x^k) and (T^k) to be bounded, then we can prove that several auxiliary sequences are bounded, too.

Lemma 7. *Suppose* (24) *holds and* $t^k \leq \bar{T} < +\infty$ *for all* $k \in \mathbb{N}$. *Then the sequences of* $w^k, \zeta^k, \sigma^k, d^k, y^k, g_\varepsilon^k$ *and* $\alpha_{k,k}$, $k \in \mathbb{N}$, *are bounded.*

Proof: Upon termination of the inner iteration, $w^k = w_i > 0$. Hence by (22)

$$0 > -w^k = \max_{j \in J^k} \left\{ g_\varepsilon^{j^T} d^k - \alpha_{k,j} \right\} + \frac{1}{2t^k} \|d^k\|^2.$$

Due to our reset strategy there always exists $j(k) \in J^k$ such that $\alpha_{k,j(k)} \leq 0$. This yields the estimate

$$0 > -w^k \geq \min_{d \in \mathbb{R}^n} \left\{ g_\varepsilon^{j(k)^T} d + \frac{1}{2t^k} \|d\|^2 \right\} = -\frac{t^k}{2} \left\| g_\varepsilon^{j(k)} \right\|^2 \geq -\frac{\bar{T}}{2} \left\| g_\varepsilon^{j(k)} \right\|^2. \tag{28}$$

Observe that $\alpha_{k,j(k)} \leq 0$ implies that $g_\varepsilon^{j(k)} \in \partial_\varepsilon f(x^k)$ (cf. (18)). Then the boundedness of $(x^k)_{k \in \mathbb{N}}$ and the local boundedness of the map $x \mapsto \partial_\varepsilon f(x)$ yield the boundedness of $(g_\varepsilon^{j(k)})_{k \in \mathbb{N}}$. Hence by (28) the sequence $(w^k)_{k \in \mathbb{N}}$ is bounded. Then (22) and (15) yield boundedness of σ^k and d^k (using $t^k \leq \bar{T}$), and hence of y^k and $g_\varepsilon^k \in \partial_\varepsilon f(y^k)$ (since $\partial_\varepsilon f$ is locally bounded), so ζ^k is bounded by (13) and (10). Combined with the continuity of f, this yields boundedness of $\alpha_{k,k}$ (cf. (6)). □

Of course, there are situations where $t^k \to 0$. In this case the crucial bound $t^k \geq \underline{t} > 0$ of Lemma 5 does not hold, and therefore there has to be a special treatment.

Lemma 8. *Suppose* (24) *holds and 0 is a cluster point of* $(t^k)_{k \in \mathbb{N}}$. *Then, for suitable subsequences* (26) *is satisfied.*

Proof: We can assume that $T^k \leq \bar{T} < +\infty$, because otherwise Lemma 6 applies. Since $\liminf_k t^k = 0$, there exists a subsequence $(k(l))$ such that $t^{k(l)+1} < t^{k(l)}$ and $t^{k(l)} \to 0$. Consider the inner loop of figure 1 at iteration $k = k(l) + 1$. Since $t^{k(l)+1} < t^{k(l)} =: t_1$, for some i in the middle branch we have

$$|f(x^k) - f(y_i)| > \max\left\{ \left\|\zeta^{k(l)}\right\|, \sigma^{k(l)} \right\}$$

(since NS(2) does not hold) with $t_i \leq t^{k(l)}$ and $y_i = x^k + d_i = x^k - t_i\zeta(t_i)$ (cf. (15)). But x^k is bounded by (24), and so is $\zeta(t_i)$ by (10), (13), and Lemma 7, whereas f is continuous, so the conclusion follows from $t_i \to 0$ and $f(x^k) - f(y_i) \to 0$ as $l \to \infty$. □

The remaining case of only finitely many serious steps is the subject of

Lemma 9. *Suppose only finitely many serious steps are taken. Then, for suitable subsequences* (26) *is satisfied.*

Proof: Of course, (x^k) is bounded. In view of Lemmas 6 and 8, we may assume the existence of $\underline{t} > 0$ and k^* such that $\underline{t} \leq t^k \leq T^k = T^{k^*}$ for all $k \geq k^*$. Further, there exists $\bar{k} > k^*$ such that $x^k = x^{\bar{k}}$ for all $k \geq \bar{k}$. We claim that the right branch of the inner iteration of figure 1 is entered only finitely often at iterations $k \geq \bar{k}$. Indeed, otherwise the middle branch would be accepting the current $t_i \geq t_1 := t^{k-1}$ as t^k due to $\hat{k} \neq 0$, the right branch would keep $u_i = T^{k^*}$, and the bisection would eventually produce $t_i > u_{i+1} - t_\varepsilon$ and hence $T^{k+1} > T^{k^*}$, a contradiction. Thus increasing $\bar{k}$ if necessary, we may assume that t^k is non-increasing for $k \geq \bar{k}$; from now on we only consider such iterations.

We need the following notation for the QP subproblem (9). Each approximate subgradient g_ε^j computed at y^j defines the linearization $\bar{f}_j(x) := f(y^j) + g_\varepsilon^{j^T}(x - y^j)$ of f at y^j such that $f(x) + \varepsilon \geq \bar{f}_j(x)$ for all x; equivalently, it may be defined by

$$\bar{f}_j(x) := f(x^k) + g_\varepsilon^{j^T}(x - x^k) - \alpha_{k,j}.$$

Similarly, with the dummy index $\tilde{k}$ introduced in Step 3 we associate the aggregate linearization

$$\bar{f}_{\tilde{k}}(x) := f(x^k) + g_\varepsilon^{\tilde{k}^T}(x - x^k) - \alpha_{k,\tilde{k}}$$

stemming from (cf. (13), (14))

$$\big(g_\varepsilon^{\tilde{k}}, \alpha_{k,\tilde{k}}\big) := (\zeta^k, \sigma^k) = \sum_{j \in J^k} \lambda_j^k \big(g_\varepsilon^j, \alpha_{k,j}\big),$$

where $\lambda_j^k := \lambda_j(t^k) \geq 0$ are the Lagrange multipliers of (9), with $\sum_{j \in J^k} \lambda_j^k = 1$. In other words, $\bar{f}_{\tilde{k}}(x)$ is a convex combination of $\bar{f}_j(x)$, $j \in J^k$, and hence $f(x) + \varepsilon \geq \bar{f}_{\tilde{k}}(x)$. This aggregate construction has the following property: the solution of (9) does not change if J^k is replaced by $J_s^k := J^{k+1} \setminus \{k+1\}$, i.e., $J_s^k = J$ at Step 4. (The property is trivial if $J_s^k = J^k$; for $J_s^k \ni \tilde{k}$ it follows from the KKT conditions for (9).)

Recalling that $y^{k+1} = x^k + d^k$ with $d^k = d(t^k)$, this property may be expressed in terms of the original approximation $\hat{f}^k(x) := \max_{j \in J^k} \bar{f}_j(x)$ and the a posteriori approximation $\hat{f}_s^k(x) := \max_{j \in J_s^k} \bar{f}_j(x)$, using

$$\Phi^k(x) := \hat{f}^k(x) + \frac{1}{2t^k}\|x - x^k\|^2 \quad \text{and} \quad \Phi_s^k(x) := \hat{f}_s^k(x) + \frac{1}{2t^k}\|x - x^k\|^2.$$

Namely, we have $y^{k+1} = \operatorname{argmin}\Phi^k = \operatorname{argmin}\Phi_s^k$ and $\hat{f}^k(y^{k+1}) = \hat{f}_s^k(y^{k+1})$. Thus $\Phi_s^k(y^{k+1}) = \Phi^k(y^{k+1}) \le \Phi^k(x^k) = \hat{f}^k(x^k) \le f^k + \varepsilon$ and $\Phi_s^k(x^k) \le f^k + \varepsilon$ by the linearization properties. Further, the convexity of $\hat{f}_s^k$ and the strong convexity of $\frac{1}{2t^k}\|\cdot - x^k\|^2$ imply strong convexity of Φ_s^k and hence

$$\Phi_s^k(x) \ge \Phi^k(y^{k+1}) + \frac{1}{2t^k}\|x - y^{k+1}\|^2 \quad \text{for all} \quad x \in \mathbb{R}^n. \tag{29}$$

Setting $x = x^k = x^{\bar{k}}$ yields

$$f(x^{\bar{k}}) + \varepsilon \ge \Phi_s^k(x^k) \ge \Phi^k(y^{k+1}) + \frac{1}{2t^k}\|x^k - y^{k+1}\|^2. \tag{30}$$

For $k \ge \bar{k}$ we have $x^{k+1} = x^k$ and $t^{k+1} \le t^k$. Further, $J^{k+1} \supset J_s^k$ yields $\hat{f}^{k+1} \ge \hat{f}_s^k$, so

$$\begin{aligned}\Phi^{k+1}(x) &:= \hat{f}^{k+1}(x) + \frac{1}{2t^{k+1}}\|x - x^{k+1}\|^2 \\ &\ge \hat{f}_s^k(x) + \frac{1}{2t^k}\|x - x^k\|^2 = \Phi_s^k(x),\end{aligned} \tag{31}$$

for all x. Using (29)–(31), we get

$$\Phi^k(y^{k+1}) + \frac{1}{2t^k}\|y^{k+2} - y^{k+1}\|^2 \le \Phi_s^k(y^{k+2}) \le \Phi^{k+1}(y^{k+2}) \le f(x^{\bar{k}}) + \varepsilon.$$

Hence there exists Φ^* such that

$$\Phi^k(y^{k+1}) \uparrow \Phi^* \le f(x^{\bar{k}}) + \varepsilon, \tag{32}$$

and $0 < t^{k+1} \le t^k$ yields

$$\lim_{k\to\infty} \|y^{k+2} - y^{k+1}\| = 0. \tag{33}$$

Next consider

$$\begin{aligned}&f(y^{k+1}) - \hat{f}^k(y^{k+1}) \\ &\quad = \bar{f}_{k+1}(y^{k+1}) - \hat{f}^k(y^{k+1}) \\ &\quad = \bar{f}_{k+1}(y^{k+2}) - \hat{f}^k(y^{k+1}) - g_\varepsilon(y^{k+1})^T(y^{k+2} - y^{k+1}) \\ &\quad \le \max\{\bar{f}_j(y^{k+2}) \mid j \in J^{k+1}\} - \hat{f}^k(y^{k+1}) + \|g_\varepsilon(y^{k+1})\| \cdot \|y^{k+2} - y^{k+1}\| \\ &\quad = \hat{f}^{k+1}(y^{k+2}) - \hat{f}^k(y^{k+1}) + \|g_\varepsilon(y^{k+1})\| \cdot \|y^{k+2} - y^{k+1}\|.\end{aligned}$$

The error between the function f and its model $\hat{f}^k$ at y^{k+1} is denoted by $\epsilon^k := f(y^{k+1}) - \hat{f}^k(y^{k+1})$. Therefore

$$\epsilon^k \le \hat{f}^{k+1}(y^{k+2}) - \hat{f}^k(y^{k+1}) + \|g_\varepsilon(y^{k+1})\| \cdot \|y^{k+2} - y^{k+1}\|$$

$$
\begin{aligned}
&= \Phi^{k+1}(y^{k+2}) - \frac{1}{2t^{k+1}}\|y^{k+2} - x^{k+1}\|^2 - \Phi^k(y^{k+1}) + \frac{1}{2t^k}\|y^{k+1} - x^k\|^2 \\
&\quad + \|g_\varepsilon(y^{k+1})\| \cdot \|y^{k+2} - y^{k+1}\| \\
&\leq (\Phi^{k+1}(y^{k+2}) - \Phi^k(y^{k+1})) + \frac{1}{2t^k}(\|y^{k+1} - x^k\|^2 - \|y^{k+2} - x^{k+1}\|^2) \\
&\quad + L\|y^{k+2} - y^{k+1}\| \\
&= (\Phi^{k+1}(y^{k+2}) - \Phi^k(y^{k+1})) + \frac{1}{2t^k}(\|y^{k+1} - y^{k+2}\|^2 \\
&\quad - 2(y^{k+1} - y^{k+2})^T(x^* - y^{k+2})) + L\|y^{k+2} - y^{k+1}\|, \qquad (34)
\end{aligned}
$$

where $L < \infty$ majorizes $\|g_\varepsilon(y^{k+1})\|$ (cf. Lemma 7). Consider the right hand side in (34), (32) and (33), then we obtain

$$\limsup_k \epsilon^k \leq 0. \qquad (35)$$

Since, for $k \geq \bar{k}$, we solely make null steps, criterion NS(1) is always satisfied. This yields

$$
\begin{aligned}
\epsilon^k &= f(y^{k+1}) - \hat{f}^k(y^{k+1}) = f(y^{k+1}) - f^k - (\hat{f}^k(y^{k+1}) - f^k) \\
&> \nu_1 v^k - v^k = (1 - \nu_1)|v^k| \geq 0. \qquad (36)
\end{aligned}
$$

Hence (35) gives $v^k \to 0$. Finally, since (cf. (22)) $-v^k = w^k + \frac{t^k}{2}\|\zeta^k\|^2$ with $w^k = \frac{t^k}{2}\|\zeta^k\|^2 + \sigma^k \geq 0$ and $t^k \geq \underline{t}$, we have $\zeta^k \to 0$ and $\sigma^k \to 0$, as required. □

Next we prove that $(x^k)_{k\in\mathbb{N}}$ is bounded if (25) does not hold. This implies that our preceding results (i.e., Lemmas 5 through 9) apply.

Lemma 10. *Suppose that* (25) *does not hold, i.e.,* $\inf_k f^k > \bar{f} + \varepsilon$. *Then* $(x^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof: If (25) is not satisfied, then there exist $\bar{x} \in \mathbb{R}^n$ and $\gamma > 0$ such that

$$f(\bar{x}) + \varepsilon + \gamma \leq f^k \quad \text{for all } k. \qquad (37)$$

Relation (19) becomes in terms of ζ^k and σ^k for $x := \bar{x}$

$$\zeta^{k^T}(\bar{x} - x^k) \leq f(\bar{x}) - f^k + \varepsilon + \sigma^k. \qquad (38)$$

If we put

$$\delta^k := \begin{cases} 1, & \text{if } k \to k+1 \text{ is a serious step,} \\ 0, & \text{if } k \to k+1 \text{ is a null step,} \end{cases}$$

then $x^{k+1} - x^k = \delta^k d^k = -\delta^k t^k \zeta^k$ for all k and thus

$$-(\bar{x} - x^k)^T (x^{k+1} - x^k) = \delta^k t^k (\bar{x} - x^k)^T \zeta^k \leq \delta^k t^k (f(\bar{x}) - f^k + \varepsilon + \sigma^k)$$

due to (38). Therefore, it follows that

$$\begin{aligned} \|\bar{x} - x^{k+1}\|^2 &= \|\bar{x} - x^k\|^2 + \|x^k - x^{k+1}\|^2 - 2(\bar{x} - x^k)^T (x^{k+1} - x^k) \\ &\leq \|\bar{x} - x^k\|^2 + \delta^k (t^k)^2 \|\zeta^k\|^2 + 2\delta^k t^k (f(\bar{x}) - f^k + \varepsilon + \sigma^k) \\ &= \|\bar{x} - x^k\|^2 + 2\delta^k t^k (f(\bar{x}) - f^k + \varepsilon + t^k \|\zeta^k\|^2 + \sigma^k) - \delta^k (t^k)^2 \|\zeta^k\|^2. \end{aligned} \tag{39}$$

Since $v^k = -t^k \|\zeta^k\|^2 - \sigma^k$ we conclude from (39) that

$$\|\bar{x} - x^{k+1}\|^2 - \|\bar{x} - x^k\|^2 \leq 2\delta^k t^k (f(\bar{x}) - f^k + \varepsilon - v^k). \tag{40}$$

By SS(1) we have

$$f^{k+1} \leq f^k + \nu_1 \delta^k v^k \quad \text{for all } k.$$

Hence, for arbitrary $m > 1$

$$\begin{aligned} f(x^1) - f(x^m) &= f(x^1) - f(x^2) + f(x^2) - \cdots + f(x^{m-1}) - f(x^m) \\ &\geq -\nu_1 \sum_{l=1}^{m} \delta^l v^l. \end{aligned}$$

For $m \to \infty$ we obtain from (37)

$$+\infty > f(x^1) - f(\bar{x}) \geq -\nu_1 \sum_{l=1}^{\infty} \delta^l v^l \geq 0.$$

Thus $\delta^l v^l$ becomes arbitrarily small for l sufficiently large. If we consider $(\delta^k)^2 = \delta^k$ and remodel (40), then for sufficiently large k such that $-\delta^k v^k < \gamma$

$$\|\bar{x} - x^{k+1}\|^2 - \|\bar{x} - x^k\|^2 \leq 2\delta^k t^k (f(\bar{x}) + \varepsilon + \gamma - f^k) \leq 0.$$

This proves $(x^k)_{k \in \mathbb{N}}$ to be bounded. □

Our convergence result now follows easily.

Theorem 2. *The sequence $(f^k)_{k \in \mathbb{N}}$ generated by Algorithm 1 satisfies*

$$\lim_{k \to \infty} f^k \leq \bar{f} + \varepsilon$$

implying that each cluster point x^* of $(x^k)_{k \in \mathbb{N}}$ is in $\mathcal{M}_\varepsilon$.

Proof: If (25) failed to hold, then (x^k) would be bounded by Lemma 10, so Lemmas 5 through 9 combined with (19) would imply (25), a contradiction. □

5.3. The case $\varepsilon = 0$

As we would expect, in case of exact subgradients the convergence results can be strengthened substantially. First, observe that $w_i \geq 0$ and $v_i \leq 0$ for all i in inner iterations. This implies that the right branch in figure 1 will never be entered. Hence $(t^k)_{k\in\mathbb{N}}$ is bounded by the initial T^1 as in [15]. Additionally, we may set $\tilde{t} := 0$ constantly in the inner iteration. Then the proposed algorithm can be analyzed like the algorithms developed in [7, 15].

6. Numerical tests

We shall now report on computational testing of Algorithm 1 with a double precision FORTRAN-code on a DECstation 5000/25 with relative accuracy $\varepsilon_M \approx 1.1 \times 10^{-16}$. All test examples except Fermat, Ms, Mml, Msl can be found in [12]. The examples Fermat, Ms (minisum), Mml (minimax location), Msl (unconstrained minisum location) can be found in [2]. In Table 1 n denotes the dimension of the problem, $\bar{f}$ is the (known) optimal value.

We calculate an ε-subgradient at x by $g_\varepsilon(x) = \lambda g(x) + (1-\lambda)g(x_1)$, where $g(x)$ is a subgradient at x and $g(x_1)$ is a subgradient at a point x_1 such that $0 < \alpha_o(x, x_1) = f(x) - f(x_1) - g(x_1)^T(x - x_1) \leq \varepsilon$. Here $x_1 \in B_r(x) = \{z \in \mathbb{R}^n | \|z - x\| \leq r\}$ and $\lambda \in [0, 1]$ are randomly chosen. The radius r is adjusted iteratively in the following way: If we find the linearization error $\alpha_o(x, x_1) > \varepsilon$ then r is reduced by a multiple smaller than one. On the other hand, if $\alpha_o(x, x_1)$ is significantly smaller than ε, then r is increased by a multiple greater than one.

The parameters have values $\nu_1 = 0.1$, $\nu_2 = 0.2$, $\nu_3 = 0.9$, $\xi = 10$ and $\varepsilon_s = 10^{-6}$. We use $J_{\max} = 5$ for CB2, CB3, QL, $J_{\max} = 10$ for Fermat, Mifflin1, Mifflin2, Rosen, Shor, Ms, Mml, $J_{\max} = 20$ for Msl and $J_{\max} = 100$ for Maxq.

In the subsequent tables, $\#f/g_\varepsilon$ denotes the number of function and ε-subgradient evaluations. Moreover, $\sigma^* = \sigma^k$ and $f^* = f^k$ where k indicates the iteration in which the stopping criterion (21) is satisfied.

Table 1. Test problems.

Problem	n	$\bar{f}$	Problem	n	$\bar{f}$
CB2	2	1.952225	Rosen	4	−44
CB3	2	2	Ms	4	67.23856
Fermat	2	562.86055	Mml	4	23.886767
QL	2	7.2	Shor	5	22.60016
Mifflin1	2	−1	Msl	6	68.829556
Mifflin2	2	−1	Maxq	20	0

Table 2. Performance for exact subgradients, i.e. $\varepsilon = 0$.

Test	#SS	#NS	$\#f/g_\varepsilon$	σ^*	f^*
CB2	20	7	28	0.673E-06	1.9522245
CB3	15	6	22	0.688E-10	2.0000000
Fermat	14	4	46	0.789E-07	562.8605512
QL	25	10	36	0.664E-13	7.2000000
Mifflin1	20	19	57	0.211E-09	−0.9999999
Mifflin2	13	11	32	0.195E-07	−0.9999999
Rosen	40	26	68	0.261E-13	−44.0000000
Shor	22	20	54	0.789E-08	22.6001621
Ms	23	5	31	0.270E-12	67.2385605
Mml	37	29	210	0.510E-08	23.8867670
Msl	26	32	64	0.532E-11	68.8295558
Maxq	52	99	156	0.991E-06	0.0000004

Table 2 shows computational results when applying the algorithm to the exact problem, i.e. $\varepsilon = 0$. Hence, we use exact subgradients. For reasons of comparison, we make use of the inner iteration as displayed in figure 1. Of course, Section 5.3 suggests that putting $\tilde{t} := 0$ constantly in the inner iteration is possible. Then the new algorithm is close to the one in [15]. The corresponding results can be found in Table 3.

Upon studying Table 2–6 the following conclusions may be drawn: In all cases, the algorithm stops with ε-ε_s-optimal solutions. For $\varepsilon > 0$ one finds that frequently $\sigma^* < 0$. The function values upon termination of the algorithm satisfy $f^* \leq \bar{f} + \varepsilon$ (for $\varepsilon > 0$).

Table 3. Performance for exact subgradients, i.e. $\varepsilon = 0$, with $\tilde{t} = 0$ constantly.

Test	#SS	#NS	$\#f/g_\varepsilon$	σ^*	f^*
CB2	20	7	28	0.673E-06	1.9522245
CB3	15	6	22	0.688E-10	2.0000000
Fermat	3	12	42	0.412E-08	562.8605511
QL	25	10	36	0.664E-13	7.2000000
Mifflin1	18	25	67	0.549E-11	−1.0000000
Mifflin2	14	16	52	0.150E-07	−0.9999999
Rosen	40	26	68	0.261E-13	−44.0000000
Shor	22	20	54	0.789E-08	22.6001621
Ms	23	5	31	0.270E-12	67.2385605
Mml	16	15	207	0.918E-11	23.8867670
Msl	26	31	67	0.526E-10	68.8295558
Maxq	49	109	171	0.812E-06	0.0000004

Table 4. Performance for ϵ-subgradients with $\epsilon := 1E\text{-}5$.

Text	#SS	#NS	$\#f/g_\epsilon$	σ^*	f^*	$f^* - \bar{f}$
CB2	13	10	32	0.386E-06	1.9522246	3.8E-6
CB3	13	9	33	0.733E-07	2.0000049	4.9E-6
Fermat	17	18	94	−0.171E-05	562.8605570	5.8E-6
QL	19	21	70	−0.971E-08	7.2000001	1.0E-7
Mifflin1	16	14	59	−0.506E-13	−0.9999989	1.1E-6
Mifflin2	10	6	24	−0.124E-05	−0.9999998	2.0E-7
Rosen	21	31	98	0.823E-07	−43.9999995	5.0E-7
Shor	23	33	135	−0.649E-08	22.6001633	3.3E-6
Ms	12	13	40	−0.108E-08	67.2385605	<1.0E-7
Mml	37	28	216	−0.264E-05	23.8867685	1.5E-6
Msl	17	20	103	−0.254E-07	68.8295574	1.6E-6
Maxq	171	148	329	−0.202E-11	0.0000000	0.0

Concerning #SS and #NS, i.e. the number of serious steps and null steps, one observes that for many test examples for increasing ε the numbers #SS and #NS decrease. However the number of function and ε-subgradient evaluations shows no clear tendency for varying ε. A possible explanation which is confirmed by our numerical test runs is as follows: For $\varepsilon > 0$ there are many runs that increase t several times in the last or next to the last iteration, and then terminate with ε-ε_s-optimal x^k. Here the strategy of increasing t in the inner loop (bisection strategy with upper bound fixed to T_i) is the drawback. One may incorporate a

Table 5. Performance for ϵ-subgradients with $\epsilon = 0.01$.

Test	#SS	#NS	$\#f/g_\epsilon$	σ^*	f^*	$f^* - \bar{f}$
CB2	8	3	32	−0.989E-06	1.9527006	4.76E-4
CB3	14	3	32	0.645E-07	2.0002261	2.26E-4
Fermat	9	7	78	−0.226E-05	562.8605660	1.60E-5
QL	19	3	59	−0.197E-05	7.2000912	9.15E-5
Mifflin1	13	14	58	0.170E-13	−0.9999990	1.00E-6
Mifflin2	7	7	22	−0.945E-06	−0.9998227	1.77E-4
Rosen	14	18	68	0.344E-06	−43.9997207	2.80E-4
Shor	13	28	128	−0.667E-04	22.6009756	8.16E-4
Ms	8	8	105	−0.460E-05	67.2388922	3.32E-4
Mml	34	32	201	0.169E-10	23.8873110	5.44E-4
Msl	19	11	74	−0.195E-05	68.8328302	3.27E-3
Maxq	177	85	267	−0.253E-08	0.0000010	1.00E-6

Table 6. Performance for ϵ-subgradients with $\epsilon = 0.1$.

Test	#SS	#NS	$\#f/g_\epsilon$	σ^*	f^*	$f^* - \bar{f}$
CB2	8	2	34	−0.835E-08	1.9882433	3.60E-2
CB3	14	3	32	0.644E-07	2.0002261	2.26E-4
Fermat	9	7	78	−0.228E-05	562.8605675	1.75E-5
QL	9	8	23	−0.267E-04	7.2150365	1.50E-2
Mifflin1	8	10	44	−0.210E-13	−0.9923110	7.69E-3
Mifflin2	8	7	24	−0.735E-06	−0.9997822	2.18E-4
Rosen	10	15	91	−0.121E-03	−43.9955174	4.48E-3
Shor	9	14	125	−0.595E-03	22.6082298	8.07E-3
Ms	8	12	86	−0.550E-03	67.2398573	1.30E-3
Mml	37	26	225	0.564E-07	23.8870238	2.57E-4
Msl	11	19	105	−0.816E-03	−68.8547624	2.52E-2
Maxq	63	115	184	0.637E-09	0.0000012	1.20E-6

routine which traces the t-enlargement in a single inner iteration, and chooses more and more progressive values for t.

Acknowledgment

The author acknowledges two anonymous referees and an associate editor for their careful reading and valuable comments.

References

1. B.M. Bell, "Global convergence of a semi-infinite optimization method," Applied Mathematics and Optimization, vol. 21, pp. 69–88, 1990.
2. J. Chatelon, D. Hearn, and T.J. Lowe, "A subgradient algorithm for certain minimax and minisum problems—the constrained case," SIAM Journal on Control and Optimization, vol. 20, pp. 455–469, 1982.
3. F.H. Clarke, Optimization and Nonsmooth Analysis, Wiley: New York, 1983.
4. J. Hiriart-Urruty and C. Lemaréchal, Convex Analysis and Minimization Algorithms I+II, Springer-Verlag: Berlin, 1993.
5. K.C. Kiwiel, Methods of Descent for Nondifferentiable Optimization, Lecture Notes in Mathematics, Springer-Verlag: Berlin, 1985.
6. K.C. Kiwiel, "An algorithm for nonsmooth convex minimization with errors," Mathematics of Computation, vol. 45, pp. 173–180, 1985.
7. K.C. Kiwiel, "Proximity control in bundle methods for convex nondifferentiable minimization," Mathematical Programming, vol. 46, pp. 105–122, 1990.
8. K.C. Kiwiel, "Approximations in proximal bundle methods and decomposition of convex programs," Journal of Optimization Theory and Applications, vol. 84, pp. 529–548, 1995.
9. K.C. Kiwiel, "Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization," SIAM Journal on Optimization, vol. 6, pp. 227–249, 1996.
10. C. Lemaréchal, "Nondifferentiable optimization," Handbook of OR & MS, vol.1, pp. 529–572, 1989.

11. C. Lemaréchal and C. Sagastizábal, "Variable metric bundle methods: from conceptual to implementable forms," Mathematical Programming, vol. 76, pp. 393–410, 1997.
12. M.M. Mäkelä and P. Neittaanmäki, Nonsmooth Optimization, World Scientific Publishing: Singapore, 1992.
13. R. Mifflin, "A quasi-second-order proximal bundle algorithm," Mathematical Programming, vol. 73, pp. 51–72, 1996.
14. R.T. Rockafellar, Convex Analysis, Princeton University Press: NJ, 1970.
15. 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, vol. 2, pp. 121–152, 1992.
16. N.Z. Shor, Minimization Methods for Non-Differentiable Functions, Springer-Verlag: Berlin, 1985.