Mathematical Programming 46 (1990) 105-122
North-Holland

PROXIMITY CONTROL IN BUNDLE METHODS FOR CONVEX NONDIFFERENTIABLE MINIMIZATION

Krzysztof C. KIWIEL

Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

Received 11 November 1987
Revised 11 July 1988

Proximal bundle methods for minimizing a convex function f generate a sequence $\{x^k\}$ by taking x^{k+1} to be the minimizer of $\hat{f}^k(x)+u^k|x-x^k|^2/2$, where $\hat{f}^k$ is a sufficiently accurate polyhedral approximation to f and $u^k>0$. The usual choice of $u^k=1$ may yield very slow convergence. A technique is given for choosing $\{u^k\}$ adaptively that eliminates sensitivity to objective scaling. Some encouraging numerical experience is reported.

Key words: Nondifferentiable minimization, convex programming, numerical methods, descent methods.

1. Introduction

This paper describes a proximal bundle method for minimizing a (possibly nondifferentiable) convex function $f:\mathbb{R}^N\to\mathbb{R}$. We assume that at each $y\in\mathbb{R}^N$ we can compute $f(y)$ and an arbitrary subgradient $g(y)\in\partial f(y)$ that defines the linearization of f at y

$$\bar{f}(x;y)=f(y)+\langle g(y),x-y\rangle \quad \forall x\in\mathbb{R}^N.$$

The method is a proximal point algorithm (see, e.g. Rockafellar [19]), i.e. it generates for any starting point z^1 a sequence $\{z^l\}$ in $\mathbb{R}^N$ by the approximate rule

$$z^{l+1}\approx \operatorname{argmin}\{f(z)+u^l|z-z^l|^2/2\colon z\in\mathbb{R}^N\},$$

where $u^l>0$. It is a bundle method (see Lemarechal [16]) in the sense that for implementability the calculation of z^{l+1} involves replacing f with its piecewise linear (polyhedral) approximation of the form

$$\hat{f}^l(z)=\max\{\bar{f}(z;y)\colon y\in Y^l\},$$

where Y^l is a finite set.

Our method has the following background. Lemarechal in his pioneering work [13] proposed an algorithm in which

$$z^{l+1}=\operatorname{argmin}\{\hat{f}^l(z)+\langle A^l(z-z^l),z-z^l\rangle/2\colon z\in\mathbb{R}^N\},$$

This research was supported by Project CPBP.02.15.

where the $N \times N$ matrix A^l was intended to accumulate information about the curvature of f around z^l. Next, Mifflin [18] showed that if f is inf-compact and the matrices A^l stay uniformly bounded and positive definite then at least one cluster point of $\{z^l\}$ is optimal. However, updating A^l by the BFGS secant formula gave disappointing numerical results (see Lemarechal [15]), and the method was hardly implementable because it required storing all the past linearizations. To avoid the first difficulty and to tackle the second one, Kiwiel [6] showed that the use of $A^l \equiv I$ (the identity matrix) suffices for bounding storage through subgradient selection or aggregation; the resulting method is globally convergent (and finitely convergent in the polyhedral case; see [9, 11]). Following the suggestion of Lemarechal [14], Kiwiel gave three different extensions to contrained problems: feasible point methods [9], exact penalty methods [8] and constraint linearization methods [11]. These methods seem to be quite efficient for well-scaled problems (cf. [8, 9, 10, 11]). Yet in practice, for instance, Kiwiel's method [6] for unconstrained minimization can converge very slowly, since it is sensitive to the scaling of f (multiplication of f by a positive constant). Auslender [1] studied convergence of modifications of Kiwiel's algorithms with $A^l = u^l I$ and bounded $\{u^l\}$ (for an inf-compact f), but he neither established global convergence to a solution nor gave practical rules for choosing u^l (see also Fukushima [4]). Moreover, he also observed that the general theoretical results of Rockafellar [19], which might ensure global convergence, can hardly be applied in practice when f is neither smooth nor polyhedral.

Thus so far it has not been known how to choose the weights u^l so that the resulting method is both globally convergent in theory and is more efficient in practice than methods with $u^l \equiv 1$.

In this paper we propose choosing the weight u^l by safeguarded quadratic interpolation so that it estimates the curvature of f between z^l and z^{l-1}. Our preliminary computational experience indicates that this technique can significantly decrease the number of objective evaluations required to reach a desired accuracy in the optimal f-value (this measure of efficiency is suitable for typical applications; see Lemarechal [15]). Thus our motivation is mainly practical, since our global convergence results are the same as in Kiwiel [9], whereas results on rate of convergence are still missing. At the same time, our global convergence analysis is general enough to leave room for other (hopefully) more efficient weighting techniques.

It is worth adding that our weighting technique can be incorporated in the methods for constrained minimization problems from Kiwiel [7, 8, 9, 11]. From lack of space, we leave details to the interested readers, observing only that our quadratic interpolation (cf. (2.13)) should use values of a merit function for the constrained problem (an exact penalty function or an improvement function). We mention these extensions because they are useful in calculations and thus motivate the present work.

We refer the reader to Lemarechal [16] for a review of approaches to the construction of superlinearly convergent methods. So far they have produced implementable methods only in the one-dimensional case.

The paper is organized as follows. The algorithm is derived in Section 2. Its global convergence is studied in Section 3. Some modifications and extensions are described in Section 4. Section 5 gives comparisons with related works of Auslender [1], Mifflin [18] and Rockafellar [19]. In Section 6 some numerical experience is reported. Finally, we have a conclusions section.

We use the following notation. We denote by $\langle \cdot , \cdot \rangle$ and $|\cdot|$, respectively, the usual inner product and norm in $\mathbb{R}^N$. We use x_i to denote the ith component of the vector x. Superscripts are used to denote different vectors. All vectors are column vectors, but (x, y) may denote $(x^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$, where x^{T} is the transpose of x. For $\varepsilon \geqslant 0$, the ε-subdifferential of f at x is defined by

$$\partial_\varepsilon f(x) = \{p \in \mathbb{R}^N : f(y) \geqslant f(x) + \langle p, y - x\rangle - \varepsilon \ \forall y \in \mathbb{R}^N\}.$$

We denote by ∂f the ordinary subdifferential $\partial_0 f$. The mapping ∂f is locally bounded and f is locally Lipschitz continuous (see Clarke [3]). A convex function $\phi : \mathbb{R}^N \to \mathbb{R}$ is called strongly convex with modulus $u > 0$ (see Rockafellar [19]) if

$$\phi(y) \geqslant \phi(x) + \langle \bar{g}, y - z\rangle + u|y - x|^2/2 \quad \forall x, y \in \mathbb{R}^N, \qquad \bar{g} \in \partial\phi(x),$$

where we may take $\bar{g} = 0$ if x minimizes ϕ.

2. Derivation of the method

The algorithm given below generates a sequence $\{x^k\}_{k=1}^{\infty} \subset \mathbb{R}^N$ that should converge to a minimizer of f, and a sequence of trial points $\{y^k\} \subset \mathbb{R}^N$ at which the linearizations of f are calculated, where $x^1 = y^1$ is a given starting point.

At the kth iteration the polyhedral approximation to f,

$$\hat{f}^k(x) = \max\{\bar{f}(x; y^j) : j \in J^k\} \quad \text{for all } x \tag{2.1}$$

with $J^k \subset \{1, \ldots, k\}$ satisfies $\hat{f}^k \leqslant f$ and $\hat{f}^k(y^j) = f(y^j)$ for all $j \in J^k$. The next trial point is chosen as

$$y^{k+1} = \operatorname{argmin}\{\hat{f}^k(x) + u^k|x - x^k|^2/2 : x \in \mathbb{R}^N\}, \tag{2.2}$$

where $u^k > 0$ is intended to keep y^{k+1} in the region where $\hat{f}^k$ should be close to f. A *serious step* from x^k to $x^{k+1} = y^{k+1}$ will occur if y^{k+1} is significantly better than x^k in the sense that

$$f(y^{k+1}) \leqslant f(x^k) + m_{\mathrm{L}} v^k, \tag{2.3}$$

where $m_{\mathrm{L}} \in (0, 0.5)$ is a parameter and

$$v^k = \hat{f}^k(y^{k+1}) - f(x^k) \tag{2.4}$$

is the predicted descent (if $v^k \geqslant 0$ the algorithm may stop with an optimal x^k; see below). Otherwise, a *null step* $x^{k+1} = x^k$ will improve the polyhedral approximation

$\hat f^{k+1}$. Namely, $\hat f^{k+1}$ will be selected with $J^{k+1}=\{k+1\}\cup J_s^k$ and $J_s^k\subset J^k$ so that *a posteriori*

$$\hat f_s^k(x)=\max\{\bar f(x;y^j): j\in J_s^k\}\quad\text{for all } x \tag{2.5}$$

may replace $\hat f^k$ in (2.2) and (2.4), i.e. $\hat f_s^k$ will incorporate all the active linearizations, and the inactive ones may be dropped to save storage without imparing convergence.

More specifically, the necessary and sufficient optimality condition for the strongly convex problem of (2.2)

$$0\in\partial\hat f^k(y^{k+1})+u^k(y^{k+1}-x^k)$$

holds if and only if there exist multipliers $\lambda_j^k\geqslant 0$, $j\in J^k$, summing up to 1 such that $\lambda_j^k\;[\bar f(y^{k+1};y^j)-\hat f^k(y^{k+1})]=0$ $\forall j\in J^k$ and such that the direction $d^k=y^{k+1}-x^k$ satisfies

$$p^k+u^kd^k=0, \tag{2.6}$$

where $p^k\in\partial\hat f^k(y^{k+1})$ equals $\sum_{j\in J^k}\lambda_j^k g(y^j)$. Replacing J^k in these conditions by any J_s^k containing all j with $\lambda_j^k\neq 0$, we see that indeed (2.2) and (2.4) hold with $\hat f^k$ replaced by $\hat f_s^k$. Moreover, since $p^k\in\partial\hat f_s^k(y^{k+1})$, the aggregate linearization

$$\tilde f^k(x)=\hat f^k(y^{k+1})+\langle p^k,x-y^{k+1}\rangle$$

minorizes $\hat f_s^k$, $\hat f^k$ and f, so that

$$f(x)\geqslant f(x^k)+\langle p^k,x-x^k\rangle-\tilde\alpha_p^k\quad\forall x\in\mathbb{R}^N, \tag{2.7}$$

with $\tilde\alpha_p^k=f(x^k)-\tilde f^k(x^k)\geqslant 0$, and hence (cf. (2.4) and (2.6))

$$v^k=-\{u^k|d^k|^2+\tilde\alpha_p^k\}=-\{|p^k|^2/u^k+\tilde\alpha_p^k\} \tag{2.8}$$

is non-positive and yields the optimality estimate

$$f(x)\geqslant f(x^k)-|u^kv^k|^{1/2}|x-x^k|+v^k\quad\forall x\in\mathbb{R}^N. \tag{2.9}$$

It is worth observing that $\tilde f^k=\sum_j\lambda_j^k\bar f(\,\cdot\,,y^j)$ and

$$(p^k,\tilde f_p^k,\tilde\alpha_p^k)=\sum_{j\in J^k}\lambda_j^k(g^j,f_k^j,\alpha_j^k), \tag{2.10}$$

where $g^j=g(y^j)$, $f_j^k=\bar f(x^k,y^j)$, $\tilde f_p^k=\tilde f^k(x^k)$, and $\alpha_j^k=f(x^k)-f_j^k=\alpha(x^k,y^j)$, where $\alpha(\,\cdot\,,\,\cdot\,)\geqslant 0$ is the linearization error

$$\alpha(x,y)=f(x)-\bar f(x;y).$$

Note that (d^k,v^k) and λ_j^k are the solution and Lagrange multipliers of the quadratic program

$$\begin{aligned}&\text{minimize}\quad && v+u^k|d|^2/2\ \text{ over all } (d,v)\in\mathbb{R}^{N+1}\\ &\text{satisfying}\quad && -\alpha_j^k+\langle g^j,d\rangle\leqslant v\ \text{ for all } j\in J^k.\end{aligned} \tag{2.11}$$

It remains to specify the choice of u^k. To illustrate the danger of keeping u^k fixed, suppose temporarily that $u^k = \bar{u} > 0$ for all k and consider relations (2.1)-(2.6), (2.8) and (2.10). If $\bar{u}$ is very large, we shall have small $|v^k|$ and $|d^k|$, almost all steps serious and slow descent. On the other hand, a small $\bar{u}$ will produce large $|v^k|$ and $|d^k|$, and each serious step will be followed by many null steps. (Both cases arise in practice for the methods of Kiwiel [9, Chapter 2] when $\bar{u} = 1$ is not suitable for a given f.)

Thus we need tests for deciding whether u^k is too large or too small. The former case may be detected by

$$f(y^{k+1}) \leqslant f(x^k) + m_R v^k \tag{2.12}$$

with $m_R \in (m_L, 1)$ (cf. (2.3)), i.e. u^k may be decreased if $\hat{f}^k$ is close to f at y^{k+1} (if $\hat{f}^k$ equaled f then $u^k = 0$ would be best in (2.2)!). To update u^k, assume temporarily that $N = 1$, f is quadratic and strictly convex, and $k = 1$, so that $v^k = \langle g(x^k), d^k \rangle = -u^k |d^k|^2$. Then simple calculations yield

$$f(y^{k+1}) = f(x^k) + v^k + a|d^k|^2/2,$$

where the Hessian $a > 0$ of f equals

$$u_{\text{int}}^{k+1} = 2u^k(1 - [f(y^{k+1}) - f(x^k)]/v^k), \tag{2.13}$$

and if $u^2 = u_{\text{int}}^2$ then $\hat{f}^2 + u^2|\cdot - x^2|^2/2 = f$, y^3 is optimal and $f(y^3) = f(x^2) + v^2/2$, so that (2.3) with $m_{\text{L}} \in (0, \frac{1}{2})$ gives $x^3 = y^3$. For a general f, (2.12) with $m_R \in (\frac{1}{2}, 1)$ will ensure that $u_{\text{int}}^{k+1}/u^k \leqslant 2(1 - m_R) < 1$. Also it will be useful to safeguard our quadratic interpolation by letting

$$u^{k+1} = \max\{u_{\text{int}}^{k+1}, u^k/10, u_{\min}\}, \tag{2.14}$$

where $u_{\min}$ is a small positive constant.

Next, consider the case when u^k seems to be too small. Improving $\hat{f}^k$ through consecutive null steps is enhanced if $|f(y^{k+1}) - f(x^k)|$ and the errors $\alpha(x^k, y^{k+1})$ of new linearizations are not much larger than the "variation"

$$V^k = f(x^k) - \min\{f(x) \colon |x - x^k| \leqslant 1\}.$$

Replacing $f(x)$ above by the right-hand side of (2.7), we get

$$V^k \leqslant |p^k| + \tilde{\alpha}_p^k. \tag{2.15}$$

On the other hand, v^k is an under-estimate for $f(x^k) - \min\{f(x) \colon |x - x^k| \leqslant |d^k|\}$ (cf. (2.2) and (2.4)). Hence one may use the test

$$\max\{|f(y^{k+1}) - f(x^k)|, \alpha(x^k, y^{k+1})\} > \max\{|p^k| + \tilde{\alpha}_p^k, -10v^k\} \tag{2.16}$$

for deciding that u^k should be increased.

We may now state the method in detail.

Algorithm 2.1

Step 0 (*Initialization*). Select an initial point $x^1 \in \mathbb{R}^N$, a final accuracy tolerance $\varepsilon_s \geqslant 0$, two improvement parameters $m_L \in (0, \frac{1}{2})$ and $m_R \in (m_L, 1)$, an initial weight $u^1 > 0$, a lower bound for weights $u_{\min} > 0$, and the maximum number of stored subgradients $M_g \geqslant N+2$. Set $y^1 = x^1$, $J^1 = \{1\}$, $f_1^1 = f(y^1)$, and $g^1 = g(y^1)$. Set the variation estimate $\varepsilon_v^1 = +\infty$. Set the counters $k=1$, $l=0$ and $k(0)=1$.

Step 1 (*Direction finding*). Find the solution (d^k, v^k) of (2.11) and its multipliers λ_j^k such that the set $\hat{J}^k = \{j \in J^k \colon \lambda_j^k \neq 0\}$ satisfies $|\hat{J}^k| \leqslant N+1$. Compute $|p^k|$ and $\tilde{\alpha}_p^k$ from (2.6) and (2.8).

Step 2 (*Stopping criterion*). If $v^k \geqslant -\varepsilon_s$, terminate; otherwise, continue.

Step 3 (*Descent test*). Set $y^{k+1} = x^k + d^k$. If (2.3) holds, set $t_L^k = 1$, $k(l+1) = k+1$ and increase the counter of serious steps l by 1; otherwise, set $t_L^k = 0$ (*null step*). Set $x^{k+1} = x^k + t_L^k d^k$.

Step 4 (*Linearization updating*). Select a set J_s^k such that $\hat{J}^k \subset J_s^k \subset J^k$ and $|J_s^k| \leqslant M_g - 1$, and set $J^{k+1} = J_s^k \cup \{k+1\}$. Set $g^{k+1} = g(y^{k+1})$, $f_{k+1}^{k+1} = \bar{f}(x^{k+1}; y^{k+1})$ and $f_j^{k+1} = f_j^k + \langle g^j, x^{k+1} - x^k \rangle$ for $j \in J^{k+1} \backslash \{k+1\}$.

Step 5 (*Weight updating*). If $x^{k+1} \neq x^k$, select u^{k+1} in $[u_{\min}, u^k]$ (e.g. by (2.14)) and set $\varepsilon_v^{k+1} = \max\{\varepsilon_v^k, -2v^k\}$; otherwise, i.e. if $x^{k+1} = x^k$, set $\varepsilon_v^{k+1} = \min\{\varepsilon_v^k, |p^k| + \tilde{\alpha}_p^k\}$, and either set $u^{k+1} = u^k$ or choose u^{k+1} in $[u^k, 10u^k]$ (e.g. $u^{k+1} = \min\{u_{\text{int}}^{k+1}, 10u^k\}$) if

$$\alpha(x^k, y^{k+1}) > \max\{\varepsilon_v^{k+1}, -10v^k\}. \tag{2.17}$$

Step 6. Increase k by 1 and go to Step 1.

A few comments on the method are in order.

Step 1 may use the dual quadratic programming method of Kiwiel [12], which can solve efficiently sequences of related subproblems (2.11) with varying u^k.

Step 2 is justified by the optimality estimate (2.9).

By the rules of Step 3,

$$x^k = x^{k(l)} \quad \text{if } k(l) \leqslant k < k(l+1), \tag{2.18}$$

where, for theoretical purposes, we may let $k(l+1) = +\infty$ if the number l of serious steps stays bounded.

At Step 4 one may let $J^{k+1} = J^k \cup \{k+1\}$ and then, if necessary, drop from J^{k+1} an index $j \in J^k \backslash \hat{J}^k$ with the largest error α_j^{k+1}.

The general criteria of Step 5 may cover interpolation formulae other than (2.13). Our implementation of Step 5 uses the following procedure, in which i_u^k counts serious or null steps since the latest change of u^k, and $i_u^1 = 0$.

Procedure 2.2 (*Weight Updating*)

(a) Set $u = u^k$.

(b) If $x^{k+1} = x^k$ go to (f).

(c) If (2.12) holds and $i_u^k > 0$ set $u = u_{\text{int}}^{k+1}$ and go to (e).

(d) If $i_u^k > 3$ set $u = u^k/2$.

(e) Set $u^{k+1} = \max\{u, u^k/10, u_{\min}\}$, $\varepsilon_v^{k+1} = \max\{\varepsilon_v^k, -2v^k\}$ and $i_u^{k+1} = \max\{i_u^k + 1, 1\}$. If $u^{k+1} \neq u^k$ set $i_u^{k+1} = 1$. Exit.

(f) Set $\varepsilon_v^{k+1} = \min\{\varepsilon_v^k, |p^k| + \tilde{\alpha}_p^k\}$. If (2.17) holds and $i_u^k < -3$, set $u = u_{\mathrm{int}}^{k+1}$. Set $u^{k+1} = \min\{u, 10u^k\}$ and $i_u^{k+1} = \min\{i_u^k - 1, -1\}$. If $u^{k+1} \neq u^k$ set $i_u^{k+1} = -1$. Exit.

The counter i_u^k introduces some inertia, which smooths out the weight updating. Step (d) may decrease the weight after four consecutive serious steps even if condition (2.12) (which ensures the decrease for interpoation) does not hold. Our test (2.17) is similar to (2.16), but neglects $|f(y^{k+1}) - f(x^k)|$ (which is of the same order as $\alpha(y^{k+1}, x^k)$) and uses $\varepsilon_v^{k+1} \leq |p| + \tilde{\alpha}_p^k$. To ensure that $\varepsilon_v^{k+1} \geq V^k$ for all k (cf. (2.15)), we should set $\varepsilon_v^{k+1} = +\infty$ at step (e), but we prefer smaller changes of this estimate. Of course, our procedure is just an example and there is still room for improvement.

It is worthwhile to observe that (in theory) our algorithm can be made invariant to the scaling of f. To this end, suppose that $g(x^1) \neq 0$, set $u^1 = |g(x^1)|$ (so that $|d^1| = 1$) and let $u_{\min}$ be a small multiple of u^1 (e.g. $u_{\min} = 10^{-10}u^1$). Also replace the stopping criterion by $v^k \geq -\varepsilon_s |f(x^k)|$. Of course, this test is impractical when $\min f \approx 0$; our implementation uses the test

$$v^k \geq -\varepsilon_s(1 + |f(x^k)|), \tag{2.19}$$

so that only large scaling factors are accounted for. Without attaching too much importance to the theoretical scale-invariance, we stress that in practice our method is much less sensitive to the scaling of f than that of Kiwiel [9].

3. Convergence

In this section we show that the sequence $\{x^k\}$ generated by the method converges to a point in the set $X = \operatorname{Argmin} f$ if $X \neq \emptyset$. We assume, of course, that the tolerance $\varepsilon_s = 0$. Then (2.9) implies that upon termination $x^k \in X$. Hence we may suppose that the algorithm does not terminate.

Consider the following condition:

$$f(x^k) \geq f(\tilde{x}) \quad \text{for some fixed } \tilde{x} \text{ and all } k, \tag{3.1}$$

which holds if $X \neq \emptyset$ or $\tilde{x}$ is a cluster point of $\{x^k\}$.

Lemma 3.1. *If* (3.1) *holds then*

$$\sum_{k=1}^{\infty} t_{\mathrm{L}}^k |v^k| \leq [f(x^1) - f(\tilde{x})]/m_{\mathrm{L}}. \tag{3.2}$$

and $v^k \xrightarrow{K} 0$ *if* $K = \{k: t_{\mathrm{L}}^k = 1\}$ *is infinite. Moreover,* $x^k \to \bar{x}$ *for some* $\bar{x} \in \mathbb{R}^N$.

Proof. At Step 3 $0 \leq -m_{\mathrm{L}} t_{\mathrm{L}}^k v^k \leq f(x^k) - f(x^{k+1})$, and adding these inequalities for $k = 1, 2, \ldots$ we obtain the first assertion. Next, since $\langle p^k, \tilde{x} - x^k \rangle \leq \tilde{\alpha}_p^k$ from (2.7) with $x = \tilde{x}$, $x^{k+1} - x^k = t_{\mathrm{L}}^k d^k = -t_{\mathrm{L}}^k p^k / u^k$ from (2.6), and $t_{\mathrm{L}}^k \in \{0, 1\}$, we deduce that

$$\begin{aligned} |\tilde{x} - x^{k+1}|^2 &= |\tilde{x} - x^k|^2 + 2\langle \tilde{x} - x^k, x^k - x^{k+1} \rangle + |x^{k+1} - x^k|^2 \\ &\leq |\tilde{x} - x^k|^2 + 2 t_{\mathrm{L}}^k \tilde{\alpha}_p^k / u^k + t_{\mathrm{L}}^k |d^k|^2 \end{aligned}$$

and using (3.2), (2.8) and the bound $u^k \geq u_{\min}$ we get

$$|\tilde{x} - x^n|^2 \leq |\tilde{x} - x^k|^2 + \sum_{i=k}^{\infty} 2 t_{\mathrm{L}}^i |v^i| / u_{\min} < \infty \quad \text{if } n > k.$$

Hence $\{x^k\}$ is bounded and has an accumulation point $\bar{x}$, so we may set $\tilde{x} = \bar{x}$ above to deduce from (3.2) for any $\varepsilon > 0$ the esistence of k such that $|\bar{x} - x^k|^2 \leq \varepsilon/2$ and $|\bar{x} - x^n|^2 \leq \varepsilon$ for all $n > k$, i.e. $x^k \to \bar{x}$. □

Let $\hat{\phi}^k(x) = \hat{f}^k(x) + u^k |x - x^k|^2/2$, $\hat{\phi}_s^k(x) = \hat{f}_s^k(x) + u^k |x - x^k|^2/2$ and

$$\eta^k = \min \hat{\phi}^k = \hat{f}^k(y^{k+1}) + u^k |y^{k+1} - x^k|^2/2. \tag{3.3}$$

Note that $\eta^k \leq \hat{\phi}^k(x^k) \leq f(x^k)$. From our discussion in Section 2,

$$y^{k+1} = \operatorname{argmin} \hat{\phi}_s^k \quad \text{and} \quad \hat{f}_s^k(y^{k+1}) = \hat{f}^k(y^{k+1}),$$

so $\eta^k = \min \hat{\phi}_s^k$ and the strong convexity of $\hat{\phi}_s^k$ implies

$$\hat{\phi}_s^k(x) \geq \eta^k + u^k |x - y^{k+1}|^2/2 \quad \forall x \in \mathbb{R}^N. \tag{3.4}$$

Setting $x = x^k$ with $\hat{\phi}_s^k(x) \leq f(x)$ we get

$$u^k |y^{k+1} - x^k|^2/2 \leq f(x^k) - \eta^k. \tag{3.5}$$

If $x^{k+1} = x^k$, then $\hat{f}^{k+1} \geq \hat{f}_s^k$ and $u^{k+1} \geq u^k$, so $\phi^{k+1} \geq \hat{\phi}_s^k$ and

$$\eta^k + u^k |y^{k+2} - y^{k+1}|^2/2 \leq \eta^{k+1} \leq f(x^k) \quad \text{if } x^{k+1} = x^k \tag{3.6}$$

from (3.4). Letting $w^k = f(x^k) - \eta^k$, we get from (2.4) and (2.8),

$$w^k = u^k |d^k|^2/2 + \tilde{\alpha}_p^k = |p^k|^2/2u^k + \tilde{\alpha}_p^k, \tag{3.7}$$

$$v^k \leq -w^k \leq v^k/2 \leq 0. \tag{3.8}$$

Lemma 3.2. (i) *If* $k(l) \leq k \leq n < k(l+1)$ *then*

$$w^n \leq w^k \leq |g(x^{k(l)})|^2 / 2u^{k(l)} \leq |g(x^{k(l)})|^2 / 2u_{\min}.$$

(ii) *If* (3.1) *holds then there exists* $C < \infty$ *such that*

$$\alpha(x^k, y^{k+1}) \leq C/\sqrt{u^k} \quad \textit{for all } k. \tag{3.9}$$

Proof. (i) If $k=k(l)$ then $w^k \leqslant |g(x^k)|^2/2u^k$, since $k \in J^k$ and

$$\begin{aligned}\eta^k &\geqslant \min\{\bar{f}(x;y^k)+u^k|x-x^k|^2/2\colon x\in\mathbb{R}^N\}\\ &=\min\{f(x^k)+\langle g^k, x-x^k\rangle+u^k|x-x^k|^2/2\colon x\in\mathbb{R}^N\}\\ &=f(x^k)-|g^k|^2/2u^k,\end{aligned}$$

and assertion (i) follows from (3.6) and $u^k \geqslant u_{\min}$.

(ii) By Lemma 3.1, $x^k \to \bar{x}$. If $k(l) \leqslant k < k(l+1)$ then by (3.7)

$$u^k|d^k|^2/2 \leqslant w^k \leqslant |g(x^{k(l)})|^2/2u_{\min},$$

$|d^k| \leqslant |g(x^{k(l)})|/(u^k u_{\min})^{1/2}$, $u^k \geqslant u_{\min}$, $y^{k+1}=x^{k(l)}+d^k$ and

$$\alpha(x^k, y^{k+1}) \leqslant |f(x^{k(l)})-f(y^{k+1})|+|g(y^{k+1})||d^k|,$$

so we may use (2.18) with $x^k \to \bar{x}$, the local boundedness of g and the local Lipschitz continuity of f to complete the proof. □

We shall now use (3.9) to show that the possible unbounded growth of $\{u^k\}$ cannot impair asymptotic optimality of $\{x^k\}$.

Lemma 3.3. *If* (3.1) *holds and* $\liminf_{k\to\infty}|v^k|=0$ *then* $\{x^k\}$ *converges to some* $\bar{x}\in X$.

Proof. In view of Lemma 3.1, we need only show that $\bar{x}\in X$. If $\{u^k\}$ is bounded, we may pass to the limit in (2.9) with a subsequence such that v^k vanishes to obtain $\bar{x}\in X$. If $\{u^k\}$ is unbounded then (3.9) and the rules of Step 5 imply the existence of a subsequence of k on which $|p^k|+\tilde{\alpha}_p^k$ vanishes, since the only way to diminish ε_v^{k+1} (so that $\liminf \varepsilon_v^k=0$) is by setting it to $|p^k|+\tilde{\alpha}_p^k$, and passing to the limit in (2.7) shows that $\bar{x}\in X$. □

Note that Lemmas 3.1 and 3.3 imply that $x^k \to \bar{x}\in X$ if $X\neq\emptyset$ and the number l of serious steps is unbounded. It remains to analyze the case of a bounded l.

Lemma 3.4. *If* $x^k=x^{k(l)}=\bar{x}$ *for some fixed* l *and all* $k\geqslant k(l)$, *then* $w^k \downarrow 0$ *and* $v^k\to 0$.

Proof. (i) By the rules of Step 5 and Lemma 3.2, $u^{k+1}\geqslant u^k$, $w^{k+1}\leqslant w^k$ and $\varepsilon_v^{k+1}\leqslant\varepsilon_v^k$ for all $k\geqslant k(l)$.

(ii) If $u^k\uparrow+\infty$ then (2.17) and (3.9) show that $\liminf_{k\to\infty}\{|p^k|+\tilde{\alpha}_p^k\}=0$, and hence $w^k\downarrow 0$ from (3.7).

(iii) Suppose that $u^k\uparrow\bar{u}\in(0,\infty)$. By (3.6), $\eta^k\uparrow\bar{\eta}\leqslant f(\bar{x})$. Hence (3.5) shows that the sequence $\{y^k\}$ is bounded, while (3.6) yields $|y^{k+2}-y^{k+1}|\to 0$.

(iv) Let $k\geqslant k(l)$ and $\varepsilon^k=f(y^{k+1})-\hat{f}^k(y^{k+1})\geqslant 0$. Then

$$\begin{aligned}\varepsilon^k &= \bar{f}(y^{k+1};y^{k+1})-\hat{f}^k(y^{k+1})\\ &=\bar{f}(y^{k+2};y^{k+1})-\hat{f}^k(y^{k+1})-\langle g^{k+1}, y^{k+2}-y^{k+1}\rangle\\ &\leqslant \hat{f}^{k+1}(y^{k+2})-\hat{f}^k(y^{k+1})+|g^{k+1}||y^{k+2}-y^{k+1}|\\ &=\eta^{k+1}-\eta^k+u^k|y^{k+1}-\bar{x}|^2/2-u^{k+1}|y^{k+2}-\bar{x}|^2/2+|g(y^{k+1})||y^{k+2}-y^{k+1}|,\end{aligned}$$

since $k+1 \in J^{k+1}$ and (3.5) holds with $x^{k+1}=x^k=\bar{x}$. Hence $\varepsilon^k \to 0$ from part (iii) and the local boundedness of g.

(v) Since $f(y^{k+1})-f(x^k) > m_L v^k$ for $k \geqslant k(l)$, relation

$$\begin{aligned}\varepsilon^k &= f(y^{k+1})-f(x^k)-[\hat{f}^k(y^{k+1})-f(x^k)] \\ &\geqslant m_L v^k - v^k = (1-m_L)|v^k| \end{aligned} \tag{3.10}$$

with $m_L \in (0,1)$ and $\varepsilon^k \to 0$ imply $v^k \to 0$.

(vi) Since in both cases $w^k \downarrow 0$ and $v^k \to 0$ by (3.8), the proof is complete. □

Remark 3.5. The above proof is of interest in its own right. First, it can handle the nonconvex case (see Kiwiel [9]) with

$$\varepsilon^k = -\alpha_{k+1}^{k+1} + \langle g^{k+1}, d^k \rangle - \max\{-\alpha_j^k + \langle g^j, d^k \rangle : j \in J^k\},$$

while a less general derivation of part (iv) may use

$$\begin{aligned}\varepsilon^k &= f(y^{k+1})-f(y^k)+f(y^k)-\hat{f}^k(y^{k+1}) \\ &\leqslant |f(y^{k+1})-f(y^k)| + |\hat{f}^k(y^k)-\hat{f}^k(y^{k+1})| \leqslant 2L|y^{k+1}-y^k|,\end{aligned}$$

where $\hat{f}^k(y^k)=f(y^k)$ and L is the Lipschitz constant of f on a bounded set containing $\{y^k\}$. Secondly, $\varepsilon^k = f(y^{k+1})-\hat{f}^k(y^{k+1})$ and $f \geqslant \hat{f}^k$ imply (cf. (3.3)),

$$f(x)+u^k|x-x^k|^2/2 \geqslant \eta^k = f(y^{k+1})+u^k|y^{k+1}-x^k|^2/2-\varepsilon^k \quad \forall x, \tag{3.11}$$

hence in part (iv) we may let $\bar{y}$ be a cluster point of $\{y^k\}$ and pass to the limit above to deduce that

$$y^k \to \bar{y} = \operatorname{argmin}\{f(x)+\bar{u}|x-\bar{x}|^2/2:\ x \in \mathbb{R}^N\},$$

since $\{y^k\}$ is bounded, and that, by (2.4),

$$v^k = f(y^{k+1})-f(x^k)+\varepsilon^k \to \bar{v} = f(\bar{y})-f(\bar{x}).$$

These asymptotic results (which complement those of Auslender [1]) indicate what would occur after many null steps, e.g. if we used m_L close to 1 in (2.3).

We may now state our principal result.

Theorem 3.6. *Either $x^k \to \bar{x} \in X$ or $X = \emptyset$ and $|x^k| \to +\infty$. In both cases $f(x^k) \downarrow \inf f$.*

Proof. If (3.1) holds, e.g. $X \neq \emptyset$ or $\{x^k\}$ has a cluster point, then the preceding results imply that $x^k \to \bar{x} \in X$ and $f(x^k) \downarrow f(\bar{x}) = f(\tilde{x})$, so $\bar{x} \in X$ and the definition of $\inf f$ yields the desired conclusion. □

The next result justifies our stopping criterion.

Lemma 3.7. *If $\inf f > -\infty$ then $v^k \to 0$.*

Proof. If the number l of serious steps stays bounded then $\{v^k\}$ tends to zero from Lemma 3.4. Hence suppose that $l \to +\infty$. Replacing $f(\tilde{x})$ by $\inf f$ in (3.1) and (3.2), we get $v^k \to^K 0$. Let $k \in K$. Since $\tilde{f}^k \leqslant \hat{f}_s^k \leqslant \hat{f}^{k+1}$ and $\tilde{f}^k(x) = \tilde{f}^k(x^k) + \langle p^k, x - x^k \rangle$, by (3.3),

$$\begin{aligned}\eta^{k+1} &\geqslant \min\{\tilde{f}^k(x) + u^{k+1}|x - x^{k+1}|^2/2 \colon x \in \mathbb{R}^N\}\\ &= \tilde{f}^k(x^{k+1}) - |p^k|^2/2u^{k+1}\end{aligned}$$

and $w^{k+1} = f(x^{k+1}) - \eta^{k+1}$ satisfies

$$\begin{aligned}w^{k+1} \leqslant\ & f(x^{k+1}) - f(x^k) + \tilde{f}^k(x^k) - \tilde{f}^k(x^{k+1})\\ & + f(x^k) - \tilde{f}^k(x^k) + |p^k|^2/2u^{k+1}.\end{aligned}$$

Since $t_{\mathrm{L}}^k = 1$, we have $f(x^{k+1}) \leqslant f(x^k)$, $\tilde{f}^k(x^k) - \tilde{f}^k(x^{k+1}) = u^k|d^k|^2$ from (2.6) and $u^k/10 \leqslant u^{k+1} \leqslant u^k$ at Step 5, so $|p^k|^2/2u^{k+1} \leqslant 5|p^k|^2/u^k$. Since $f(x^k) - \tilde{f}^k(x^k) = \tilde{\alpha}_p^k$, combining these relations with (2.8) we obtain $0 \leqslant w^{k+1} \leqslant -6v^k$, and hence $\lim_{l \to \infty} w^{k(l)} \to 0$. Then Lemma 3.2(i) and (3.8) yield the desired conclusion. □

Thus the method must terminate if $\inf f > -\infty$ and $\varepsilon_s > 0$.

To complete our analysis, we show that if f is inf-compact (i.e. $X \neq \emptyset$ and the level set $S(x^1) = \{x \in \mathbb{R}^N \colon f(x) \leqslant f(x^1)\}$ are bounded) then even without the bound $u^k \geqslant u_{\min} > 0$ we still have $f(x^k) \downarrow \min f$, although the question whether $\{x^k\}$ converges remains open.

Theorem 3.8. *Suppose that at Step 5 of Algorithm 2.1 we choose* $u^{k+1} \in (0, \max\{u^k, u_{\max}\}]$ *if* $x^{k+1} \neq x^k$, *where* $u_{\max} > 0$ *is fixed. If* f *is inf-compact and the algorithm does not terminate then* $f(x^k) \downarrow \min f$.

Proof. Let $\bar{u} = \liminf_{k \to \infty} u^k$. If $\bar{u} > 0$ then $u^k \geqslant \bar{u}/2$ for large k and the preceding analysis applies. If $\bar{u} = 0$ then $l \to +\infty$ and there exist $\tilde{x} \in \mathbb{R}^N$ and an infinite set $K' \subset \{1, 2, \ldots\}$ such that $u^k \to^{K'} 0$, $x^k \to^{K'} \tilde{x}$ and $v^k \to^{K'} 0$, since $\{x^k\} \subset S(x^1)$ is bounded and the first assertion of Lemma 3.1 holds. Letting $k \in K'$ tend to infinity in (2.9), we get $\tilde{x} \in X$ and hence $f(x^k) \downarrow \min f$ from the monotonicity of $\{f(x^k)\}$ and the continuity of f. □

4. Modifications and extensions

To trade off storage and work per iteration for speed of convergence, one may replace subgradient selection with aggregation as in Kiwiel [9]. To this end, at Step 0 let $\tilde{f}^0(x) = \bar{f}(x; y^1)$ be the initial aggregate linearization. Having $\tilde{f}^{k-1}(x) = f_p^k + \langle p^{k-1}, x \rangle$ at the k-th iteration, use the approximation

$$\hat{f}^k(x) = \max\{\bar{f}(x; y^j\} \colon j \in J^k; \tilde{f}^{k-1}(x)\} \quad \forall x$$

in subproblem (2.2), or equivalently append to subproblem (2.11) the aggregate constraint

$$-\alpha_p^k + \langle p^{k-1}, d \rangle \leqslant v$$

with Lagrange multiplier λ_p^k, where $\alpha_p^k = f(x^k) - f_p^k$. Then

$$(p^k, \tilde f_p^k, \tilde\alpha_p^k) = \sum_{j\in J^k} \lambda_j^k (g^j, f_j^k, \alpha_j^k) + \lambda_p^k (p^{k-1}, f_p^k, \alpha_p^k) \tag{4.1}$$

defines $\tilde f^k(x) = \tilde f_p^k + \langle p^k, x - x^k\rangle$ as in (2.10). At Step 4 set $f_p^{k+1} = \tilde f^k(x^{k+1})$ to close the recursion, and let J_s^k be any subset of J^k with $|J_s^k| \leq M_g - 1$, where now $M_g \geq 1$ (in contrast with the previous requirement $M_g \geq N+2$). Replacing (2.5) by

$$\hat f_s^k(x) = \max\{\bar f(x; y^j) : j \in J_s^k; \tilde f^k(x)\},$$

one may verify all the preceding convergence results. In practice convergence can be slow if M_g is too small.

Next, consider the problem

$$\text{minimize} \quad f(x) \text{ over all } x \in S_h, \tag{4.2}$$

where $S_h \neq \emptyset$ is a closed convex subset of $\mathbb{R}^N$. Let δ_h denote the indicator function of S_h, i.e. $\delta_h(x) = 0$ if $x \in S_h$, $\delta_h(x) = +\infty$ otherwise. Of course (4.2) is equivalent to the problem

$$\text{minimize} \quad f(x) + \delta_h(x) \quad \text{over all } x \in \mathbb{R}^N.$$

An extension of Algorithm 2.1 for this problem is obtained by choosing $x^1 \in S_h$ and adding $\delta_h(x)$ to $\hat f^k(x)$ in (2.2), so that $\{y^k\} \subset S_h$ and $\{x^k\} \subset S_h$. By adding (where necessary) δ_h to f and its approximations $\hat f^k$, $\hat f_s^k$, $\tilde f^k$ and $\bar f(\cdot\,; y^j)$, one may verify Theorem 3.6 for $X = \operatorname{Argmin}\{f + \delta_h\}$. (Hint: use the fact that δ_h vanishes on $\{x^k\}$ and $\{y^k\}$.) Moreover, one may replace $\mathbb{R}^N$ with S_h in (2.7) and (2.9), since (2.6) holds with $p^k \in \partial(\hat f^k + \delta_h)(y^{k+1})$. Incidentally, $p^k = p_f^k + p_h^k$ with $p_f^k \in \partial \hat f^k(y^{k+1})$ and $p_h^k \in \partial\delta_h(y^{k+1})$, so that the sum of

$$\tilde f^k(x) = \hat f^k(y^{k+1}) + \langle p_f^k, x - y^{k+1}\rangle \quad \text{and} \quad \tilde\delta_h^k(x) = \langle p_h^k, x - y^{k+1}\rangle$$

minorizes $f + \delta_h$, and $\tilde\alpha_p^k = \tilde\alpha_f^k + \tilde\alpha_h^k$ with $\tilde\alpha_f^k = f(x^k) - \tilde f^k(x^k)$ and $\tilde\alpha_h^k = -\tilde\delta^k(x^k)$. Thus $(p_f^k, \tilde\alpha_f^k)$ replaces $(p^k, \tilde\alpha_p^k)$ in (2.10) (or in (4.1) if subgradient aggregation is used).

If $S_h = \{x \in \mathbb{R}^N : h_i(x) \leq 0 \ \forall i \in I_h\}$, where h_i are affine functions and I_h is finite, the above extension amounts to augmenting subproblem (2.11) with the constraints

$$h_i(x^k) + \langle \nabla h_i, d\rangle \leq 0 \quad \forall i \in I_h.$$

Then, in theory, the algorithm can be made invariant to the scaling of both f and the linear constraints. Note that f and g need not be evaluated outside S_h.

Let us now consider the case when

$$f(x) = \sum_{i=1}^{n} f_i(x) \quad \text{for all } x \in \mathbb{R}^N,$$

where $f_i : \mathbb{R}^N \to \mathbb{R}$ are convex functions with subgradients $g_{fi}(x) \in \partial f_i(x)$, for $i = 1, \ldots, n$. By exploiting the structure of f one may increase the speed of convergence at the cost of more storage and work per iteration. To this end, replace $\hat{f}^k$ in subproblem (2.2) by the approximations

$$\hat{f}^k(x) = \sum_{i=1}^{n} \hat{f}_i^k(x),$$

$$\hat{f}_i^k(x) = \max\{\bar{f}_i(x; y^j): j \in J_i^k\}$$

constructed from the linearizations of f_i,

$$\bar{f}_i(x; y) = f_i(y) + \langle g_{fi}(y), x - y\rangle$$

as in (2.1), where the sets J_i^k satisfying $\sum_{i=1}^n |J_i^k| \leq M_g$ with $M_g \geq N + 2n$ are selected by finding at most $N + n$ nonzero Lagrange multipliers λ_{ij}^k, $j \in J_i^k$, $i = 1, \ldots, n$, of the corresponding extension of (2.11) (see Kiwiel [10]). Thus $J_i^{k+1} = J_{si}^k \cup \{k+1\}$, where $\{j \in J_i^k: \lambda_{ij}^k \neq 0\} \subset J_{si}^k \subset J_i^k$, $i = 1, \ldots, n$. This extension of our subgradient selection strategy can be analyzed as is Sections 2 and 3. Moreover, our technique for the additional constraint of (4.2) can be employed. We note that Ruszczyński [20] gives encouraging numerical results for a similar method (with $u^k \equiv 1$) for polyhedral problems

5. Relations with other methods

One may show that the sequence $z^l = x^{k(l)}$ for $l = 0, 1, \ldots$ is generated as in the proximal point method, since for $\phi^k(x) = f(x) + u^k|x - x^k|^2/2$ we have

$$\phi^k(y^{k+1}) \leq \min \phi^k + \varepsilon^k$$

with $0 \leq \varepsilon^k = f(y^{k+1}) - \hat{f}^k(y^{k+1}) \leq -(1 - m_{\mathrm{L}})v^k$ if $k = k(l+1) - 1$ (cf. (3.10), (3.11)) and $|y^{k+1} - \operatorname{argmin} \phi^k| \leq (2\varepsilon^k/u^k)^{1/2}$ (see Auslender [1]). This connection, however, does not tell us much about our method, which does not seem to satisfy the conditions for convergence of Rockafellar [19], while the results of Auslender [1] would only show that $f(z^l) \downarrow \min f$ if $l \to \infty$, $\{u^k\}$ is bounded, $u^{k+1} = u^k$ if $x^{k+1} = x^k$, and $X \neq \emptyset$ is bounded.

Next, we note that Auslender [1] considers four different criteria for serious steps. His criteria (1) and (2) do not ensure descent, while criteria (3) and (4), which have the form

$$f(y^{k+1}) \leq f(x^k) + m_{\mathrm{L}}v^k - u|d^k|^2/2,$$

$$f(y^{k+1}) \leq f(x^k) - m_{\mathrm{L}}w^k,$$

do not seem to be more efficient than (2.3) and can be analyzed as in Section 3 (since $-v^k/2 \leq w^k \leq -v^k$ and $m_{\mathrm{L}}v^k - u^k|d^k|^2/2 \geq (m_{\mathrm{L}} + \frac{1}{2})v^k$ by (2.8) and (3.7)). Moreover, Theorem 3.8 subsumes the corresponding results of Auslender [1] (who supposes that f is inf-compact, $0 < u^k \leq u_{\max}$ and $u^{k+1} = u^k$ if $x^{k+1} = x^k$), which do not establish convergence of $\{x^k\}$.

We may add that our convergence results do not follow from those of Mifflin [18], since we do not assume that f is inf-compact and $J^k=\{1,\ldots,k\}$ for all k (this would require unbounded storage).

6. Numerical examples

We shall now report on computational testing of the algorithm with a double precision Fortran code on an IBM PC/XT microcomputer with relative accuracy $\varepsilon_M \approx 2.2\times 10^{-16}$ $(=2.2E-16)$. The parameters had values $m_{\mathrm{L}}=0.1$, $m_{\mathrm{R}}=0.5$, $u_{\min}=1E-10$ and $\varepsilon_s=1E-6$ (cf. (2.19)).

For convex constrained problems of the form

$$\text{minimize}\quad f(x) \text{ over all } x\in\mathbb{R}^N \tag{6.1a}$$

$$\text{satisfying}\quad F_i(x)\leqslant 0 \text{ for } i=1,\ldots,m_I, \tag{6.1b}$$

$$h_i(x)\leqslant 0 \text{ for } i\in I_h, \tag{6.1c}$$

with convex F_i and affine h_i we solved the exact penalty formulation with a fixed penalty coefficient $c>0$:

$$\text{minimize}\quad e(x)\equiv f(x)+c\sum_{i=1}^{m_I}\max\{F_i(x),0)$$

$$\text{subject to}\quad h_i(x)\leqslant 0 \text{ for } i\in I_h.$$

To save space, we shall only sketch the structure of some problems; details can be found in the quoted references. We give values of f and $F_+(\cdot)=\max\{0, F_i(\cdot)$: $i=1,\ldots,m_I\}$ at the initial point x^1 and solution $\bar{x}$.

Test 1. Shor's minimax problem [21, p. 138] with

$$f(x)=\max\left\{b_i\sum_{j=1}^{5}(x_j-a_{ij})^2\colon i=1,\ldots,10\right\}.$$

$N=5$, $f(\bar{x})=22.60016$, $f(x^1)=80$.

Test 2. Lemarechal's minimax problem MAXQUAD [17, p. 151],

$$f(x)=\max\{x^{\mathrm{T}}A^ix-x^{\mathrm{T}}b^i\colon i=1,\ldots,5\}, \tag{6.2}$$

$N=10$, $f(\bar{x})=-0.841408$, $f(x^1)=5337$.

Test 3. Goffin's polyhedral problem,

$$f(x)=N\max\{x_i\colon i=1,\ldots,N\}-\sum_{i=1}^{N}x_i,$$

$N=50$, $\bar{x}=0$, $f(\bar{x})=0$, $x_i^1=i-(N+1)/2$ for $i=1,\ldots,N$, $f(x^1)=1225$.

Test 4. Lemarechal's polyhedral problem TR48 [17],

$$f(x)=\sum_{j=1}^{N}d_j\max\{x_i-a_{ij}\colon i=1,\ldots,N\}-\sum_{i=1}^{N}s_ix_i,$$

$N=48$, $f(\bar{x})=-638565$, $f(x^1)=-464816$. (This is the dual of a transportation problem.)

Test 5. The polyhedral problem with

$$f(x)=\sum_{i=1}^{N}\left|\sum_{j=1}^{N}(x_j-1)/(i+j-1)\right|,$$

$N=50$, $\bar{x}=(1,\ldots,1)^{\mathrm{T}}$, $f(\bar{x})=0$, $x^1=0$, $f(x^1)=68.817$. (This corresponds to solving $Ax=b$ with an ill-conditioned Hilbert matrix A.)

Test 6. Ill-conditioned linear programming,

$$f(x)=\sum_{i=1}^{N}c_i(x_i-1),$$

$$F_i(x)=\langle a^i,x\rangle-b_i,\quad i=1,\ldots,m_I=N,$$

where $a_j^i=1/(i+j)$, $b_i=\sum_{j=1}^{N}a_j^i$, $c_i=-(b_i+1/(1+i))$, $N=30$, $\bar{x}=(1,\ldots,1)^{\mathrm{T}}$, $f(\bar{x})=0$, $x^1=0$, $f(x^1)=40.81$, $F_+(x^1)=0$. (Here the linear constraints are treated as nonlinear.)

Test 7. This problem has the objective (6.2) and the linear constraints $\sum_{i=1}^{10}x_i\leqslant 0.05$, $-0.05\leqslant x_i\leqslant 0.05$ for $i=1,\ldots,10$, $f(\bar{x})=-0.36816644175$, $x^1=0$, $f(x^1)=0$ (cf. [10]).

Test 8. Streit's problem no. 1 [22],

$$\begin{aligned}&\text{minimize} && \|Az-b\|_\infty \text{ over all } z\in\mathbb{C}^n\\ &\text{satisfying} && |(z-e)_i|\leqslant d_i,\ i=1,\ldots,n,\\ &&& |(Bz-g)_i|\leqslant c_i,\ i=1,\ldots,r,\end{aligned}\tag{6.3}$$

where $A\in\mathbb{C}^{m\times n}$, $b\in\mathbb{C}^m$, $e\in\mathbb{C}^n$, $d\in\mathbb{C}^n$, $B\in\mathbb{C}^{r\times n}$, $g\in\mathbb{C}^r$, $c\in\mathbb{C}^r$, and $\|b\|_\infty=\max\{|b_i|: i=1,\ldots,r\}$. Letting $z_i=x_{2i-1}+\sqrt{-1}x_{2i}$ for $i=1,\ldots,n$ and $N=2n$, we get problem (6.1a, b) with $m_I=n+r$. The first Streit problem has $n=2$, $m=5$, $r=2$, $N=4$, $m_I=4$, $f(\bar{x})=\sqrt{2}/2$, $x^1=0$, $f(x^1)=\sqrt{2}$, $F_+(x^1)=0$. (This is a constrained complex approximation problem.)

Test 9. Streit's problem no. 2 has the form (6.3) with $N=4$, $m_I=4$, $f(\bar{x})=\sqrt{2}-0.4$, $x^1=0$, $f(x^1)=\sqrt{2}$, $F_+(x^1)=0$.

Test 10. Streit's problem no. 3 has the form (6.3) with $n=3$, $m=101$, $r=0$, $N=6$, $m_I=3$, $f(\bar{x})=0.01470631$, $x^1=0$, $f(x^1)=1$, $F_+(x^1)=0$.

Test 11. The first Colville problem (Shell Primal) [5, p. 105],

$$f(x)=\sum_{j=1}^{5}d_jx_j+\sum_{i=1}^{5}\sum_{j=1}^{5}c_{ij}x_ix_j+\sum_{j=1}^{5}e_jx_j,$$

$$F_i=b_i-\sum_{j=1}^{5}a_{ij}x_j,\quad i=1,\ldots,m_I=10,$$

with constraints $x_i\geqslant 0$, $i=1,\ldots,N=5$, $f(\bar{x})=-32.348679$, $f(x^1)=20$, $F_+(x^1)=0$.

Test 12. The Rosen–Suzuki problem [5, p. 66] with $N=4$, $m_I=3$, $f(\bar{x})=-44$, $f(x^1)=0$ and $F_+(x^1)=0$.

Test 13. The minimax location problem [2],

$$f(x)=\max\{w_{i1}\|a^i-(x_1,x_2)\|_{p_{i1}},\ |(x_1,x_2)-(x_3,x_4)|,$$
$$w_{i2}\|a^i-(x_3,x_4)\|_{p_{i2}}\colon i=1,\ldots,9\},$$
$$F_1(x)=|(x_1,x_2)-a^1|^2-144,\quad F_2(x)=|(x_1,x_2)-a^4|^2-121,$$
$$F_3(x)=|(x_3,x_4)-a^1|^2-225,\quad F_4(x)=|(x_3,x_4)-a^2|^2-144,$$

$N=4$, $m_{\mathrm{I}}=4$, $f(\bar{x})=23.886767$, $f(x^1)=30.53$, $F_+(x^1)=0$.

Test 14. The minisum location problem [2],

$$f(x)=\sum_{i=1}^{3}\sum_{j=1}^{5} w_{ij}\|(x_{2i-1},x_{2i})-a^j\|_p+\sum_{1\le i<j\le 3}\|(x_{2i-1},x_{2i})-(x_{2j-1},x_{2j})\|,$$
$$F_1(x)=x_5+x_6-3,$$

$N=6$, $m_{\mathrm{I}}=1$, $f(\bar{x})=68.82856$, $f(x^1)=175.6$, $F_+(x^1)=0$.

We used the penalty coefficient $c=50$ in test 11, and $c=10$ in tests 6, 8, 9, 10, 12, 13 and 14.

Our results are summarized in Table 1, in which k denotes the final iteration number (and the total number of problem function and subgradient evaluations), $e(x^k)=f(x^k)$ if $m_{\mathrm{I}}=0$, and NS and N0 denote the number of serious and null steps respectively. The value of $\varepsilon_s=10^{-6}$ corresponds to the relative accuracy in the optimal f-value of about 6 digits (including 0's after the decimal point). This value may be too large for finite convergence on polyhedral problems (cf. tests 3–6). For tests 5 and 6 values of $f(x^k)$ of order 10^{-14} are found with $\varepsilon_s=10^{-10}$.

Table 2 contains results for the usual choice $u^k\equiv 1$.

Table 1

Test results for the weighting technique

Test	k	NS	N0	$e(x^k)$	$f(\bar{x})$
1	29	19	9	22.600 162	22.600 16
2	41	23	17	−0.841 4074	−0.841 408
3	52	34	17	6.0E−13	0
4	180	65	114	−638 565.00	−638 565
5	16	12	3	3.6E−7	0
6	7	6	0	3.5E−9	0
7	23	11	11	−0.368 1664	−0.368 116 64
8	14	10	3	0.707 1074	0.707 1068
9	9	6	2	1.014 2141	1.014 2136
10	47	25	21	0.014 7064	0.014 7063
11	10	7	2	−32.348 679	−32.348 679
12	20	16	3	−44.999 961	−44
13	15	7	7	23.886 767	23.886 767
14	23	12	10	68.829 581	68.829 56

Table 2

Test results for constant $u^k = 1$

Test	k	NS	N0	$e(x^k)$	$f(\bar{x})$
1	36	21	34	22.600 163	22.600 16
2	208	37	170	−0.841 4078	−0.841 408
3	56	25	30	5.4E − 13	0
4*	252	135	116	−638 372.3	−638 565
5	86	65	20	$1.7E-5$	0
6	38	37	0	$1.9E-5$	0
7	40	14	25	−0 368 1659	−0.368 116 64
8	11	8	2	0.707 1073	0.707 1068
9	14	10	3	1.014 2145	1.014 2136
10	199	193	5	0.014 7157	0.014 7063
11	13	7	8	−32.348 673	−32.348 679
12	20	10	9	−43.999 970	−44
13	30	17	12	23.886 767	23.886 767
14	27	13	13	68.829 582	68.829 56

* For $\varepsilon_s = 10^{-7}$ we have finite convergence with $k = 579$, $\mathrm{NS} = 457$, $\mathrm{N0} = 121$, $f(x^k) = -638\,565$.

Of course no firm conclusions should be drawn from such limited experiments, but the interested reader may compare our results with those given in [2, 4, 9, 10, 11, 15, 21, 22, 23]. Such comparisons suggest that our method can find reasonably accurate solutions in about half the number of function evaluations required by other algorithms. Further improvement is expected from more sophisticated implementations.

7. Conclusions

We have presented a technique for choosing weights of proximal terms in bundle methods for convex nondifferentiable minimization of Kiwiel [7, 8, 9, 11]. Only the method for unconstrained minimization has been treated in detail, but extensions to constrained problems are straightforward.

Our limited computational experience suggests that the method can compete successfully with other descent algorithms.

Acknowledgment

I would like to acknowledge several discussions with Alfred Auslender on this subject. My thanks go to Claude Lemarechal for constructive criticism and constant encouragement, and to Andrzej Ruszczyński for reading an earlier draft of this paper. Moreover, the comments of a referee are appreciated.

References

[1] A. Auslender, "Numerical methods for nondifferentiable convex optimizations," *Mathematical Programming Study* 30 (1986) 102–126.

[2] J. Chatelon, D. Hearn and T.J. Lowe, "A subgradient algorithm for certain minimax and minisum problems," *SIAM Journal on Control and Optimization* 20 (1982) 455–469.

[3] F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983).

[4] M. Fukushima, "A descent algorithm for nonsmooth convex programming," *Mathematical Programming* 30 (1984) 163–175.

[5] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Economics and Mathematical Systems 187 (Springer, Berlin, 1981).

[6] K.C. Kiwiel, "An aggregate subgradient method for nonsmooth convex minimization," *Mathematical Programming* 27 (1983) 320–341.

[7] K.C. Kiwiel, "An algorithm for linearly constrained convex nondifferentiable minimization problems," *Journal of Mathematical Analysis and Applications* 105 (1985) 452–465.

[8] K.C. Kiwiel, "An exact penalty function method for nonsmooth constrained convex minimization problems," *IMA Journal of Numerical Analysis* 5 (1985) 111–119.

[9] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133 (Springer, Berlin, 1985).

[10] K.C. Kiwiel, "A method of linearizations for linearly constrained nonconvex nonsmooth optimization," *Mathematical Programming* 34 (1986) 175–187.

[11] K.C. Kiwiel, "A constraint linearization method for nondifferentiable convex minimization," *Numerische Mathematik* 51 (1987) 395–414.

[12] K.C. Kiwiel, "A dual method for solving certain positive semi definite quadratic programming problems," *SIAM Journal on Scientific and Statistical Computing* (to appear).

[13] C. Lemarechal, "Nonsmooth optimization and descent methods," Research Report RR-78-4, International Institute of Applied Systems Analysis (Laxenburg, Austria, 1977).

[14] C. Lemarechal, "Nonlinear programming and nonsmooth optimization—a unification," Rapport de Recherche No. 332, Institut de Recherche d'Informatique et d'Automatique (Rocquencourt, Le Chesnay, 1978).

[15] C. Lemarechal, "Numerical experiments in nonsmooth optimization," in: E.A. Nurminski, ed., *Progress in Nondifferentiable Optimization* (CP-82-S8, International Institute for Applied Systems Analysis (Laxenburg, Austria, 1982) pp. 61–84.

[16] C. Lemarechal, "Constructing bundle methods for convex optimization," in: J.B. Hiriart-Urruty, ed., *Fermat Days* 85: *Mathematics for Optimization* (North-Holland, Amsterdam, 1986) pp. 201–240.

[17] C. Lemarechal and R. Mifflin, eds., *Nonsmooth Optimization* (Pergamon Press, Oxford, 1978).

[18] R. Mifflin, "A modification and an extension of Lemarechal's algorithm for nonsmooth minimization," *Mathematical Programming Study* 17 (1982) 77–90.

[19] R.T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization* 14 (1976) 877–898.

[20] A. Ruszczyński, "A regularized decomposition method for minimizing a sum of polyhedral functions," *Mathematical Programming* 35 (1986) 309–333.

[21] N.Z. Shor, *Minimization Methods for Nondifferentiable Functions* (Springer, Berlin, 1985).

[22] R.L. Streit, "Solution of systems of complex linear equations in the l_∞ norm with constraints on the unknowns," *SIAM Journal on Scientific and Statistical Computing* 7 (1986) 132–149.

[23] J. Zowe, "Nondifferentiable optimization," in: K. Schittkowski, ed., *Computational Mathematical Programming* (Springer, Berlin, 1985) pp. 323–356.