Mathematical Programming 62 (1993) 261–275
North-Holland

Convergence of some algorithms for convex minimization

Rafael Correa
Departamento de Matematicas, Universidad de Chile, Santiago, Chile

Claude Lemaréchal
INRIA, Le Chesnay, France

Received 23 March 1992
Revised manuscript received 30 November 1992

An important research work of Phil Wolfe's concerned convex minimization. This paper is dedicated to him, on the occasion of his 65th birthday, in appreciation of his creative and pioneering work.

We present a simple and unified technique to establish convergence of various minimization methods. These contain the (conceptual) proximal point method, as well as implementable forms such as bundle algorithms, including the classical subgradient relaxation algorithm with divergent series.

AMS Subject Classification: 65K05, 90C25.

Key words: Nondifferentiable optimization, convex programming, proximal point method, bundle algorithms, global convergence.

Introduction

To establish convergence of algorithms for convex minimization, a usual assumption is the inf-compactness of the objective function, or at least the existence of a minimum. The aim of this paper is to remove any such assumption in a number of methods, namely: (i) in the prox-iteration, (ii) in its implementable approximations, which include in particular (iii) bundle methods, and finally (iv) in the classical subgradient optimization scheme.

Compactness assumptions were already removed in 1983 by K.C. Kiwiel for (iii); more recently, O. Güler and B. Lemaire removed them for (i) as well; as for (iv), the original presentation of B.T. Poljak in 1967 was already free of any compactness. This paper actually stresses two aspects: *clarification*, and *unification*; starting from a technical result, we obtain a simple proof-pattern, applicable to each of the above-mentioned methods.

Correspondence to: Claude Lemaréchal, Domaine de Voluceau-Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France.

1. Basic results

Let $f: X \to \mathbb{R} \cup \{+\infty\}$ be a proper closed convex function on a Hilbert space X and denote by

$$\bar{f} := \inf_{x \in X} f(x)$$

its infimal value (possibly $-\infty$). We are interested in estimating $\bar{f}$, and also in identifying a minimum point, if any.

We consider sequences $\{x_n\}$ constructed according to the formula

$$x_{n+1} = x_n - t_n \gamma_n \quad \text{for } n = 1, 2, \ldots, \tag{1.1}$$

where $t_n > 0$ is a stepsize and γ_n is an approximate subgradient at x_n,

$$\gamma_n \in \partial_{\varepsilon_n} f(x_n) \quad \text{for } n = 1, 2, \ldots. \tag{1.2}$$

Recall that (1.2) means

$$f(y) \geqslant f(x_n) + \langle \gamma_n, y - x_n \rangle - \varepsilon_n \quad \text{for all } y \in X,$$

so each x_n has to be in the domain of $f: f(x_n) < +\infty$, and also $\varepsilon_n \geqslant 0$. Thus, we consider minimization algorithms of the form (1.1), characterized by $\{t_n, \varepsilon_n, \gamma_n\}$; this last triple will be either given a priori or constructed by the algorithm itself.

Our study is entirely based on the following simple inequality, essentially due to [7].

Lemma 1.1. *With the notations* (1.1), (1.2), *there holds for all* $y \in X$ *and* $n = 1, 2, \ldots,$

$$\|x_{n+1} - y\|^2 \leqslant \|x_n - y\|^2 + t_n^2 \|\gamma_n\|^2 + 2t_n[f(y) - f(x_n) + \varepsilon_n]. \tag{1.3}$$

Proof. In the development

$$\|x_{n+1} - y\|^2 = \|x_{n+1} - x_n\|^2 + 2\langle x_{n+1} - x_n, x_n - y \rangle + \|x_n - y\|^2,$$

use (1.1) and (1.2); then (1.3) comes out directly. □

Proposition 1.2. *With the notations* (1.1), (1.2), *if*

$$\sum_{n=1}^{\infty} t_n = +\infty, \tag{1.4}$$

$$\varepsilon_n \to 0, \tag{1.5}$$

$$t_n \|\gamma_n\|^2 \to 0, \tag{1.6}$$

then

$$\liminf_{n \to \infty} f(x_n) = \bar{f}. \tag{1.7}$$

Proof. Assume for contradiction that there are $\delta>0$, $n_0\in\mathbb{N}$ and $y\in X$ such that

$$f(y)<f(x_n)-\delta \quad \text{for all } n\geqslant n_0.$$

Use this y in (1.3). In view of (1.5), (1.6) we may assume that n_0 is so large that $t_n\|\gamma_n\|^2+2\varepsilon_n$ is then smaller than δ, from which we obtain

$$\|x_{n+1}-y\|^2\leqslant\|x_n-y\|^2+t_n[t_n\|\gamma_n\|^2+2\varepsilon_n-2\delta]\leqslant\|x_n-y\|^2-\delta t_n$$

for all $n\geqslant n_0$. Summing up, this gives

$$0\leqslant\|x_n-y\|^2\leqslant\|x_{n_0}-y\|^2-\delta\sum_{i=n_0}^{n-1}t_i \tag{1.8}$$

for all $n>n_0$; letting $n\to+\infty$, (1.4) is contradicted. □

Proposition 1.3. *Let a sequence $\{x_n\}\subset X$ have a cluster point $\bar{x}$ satisfying*

$$\|x_{n+1}-\bar{x}\|^2\leqslant\|x_n-\bar{x}\|^2+\delta_n, \tag{1.9}$$

where $\{\delta_n\}$ is a nonnegative sequence such that

$$\sum_{n=1}^{\infty}\delta_n<+\infty.$$

Then the whole sequence $\{x_n\}$ converges to $\bar{x}$.

Proof. For given $\delta>0$, take some n_1 satisfying

$$\|x_{n_1}-\bar{x}\|^2\leqslant\tfrac{1}{2}\delta \quad\text{and}\quad \sum_{n=n_1}^{\infty}\delta_n\leqslant\tfrac{1}{2}\delta;$$

conclude by summation in (1.9),

$$\|x_{n_2+1}-\bar{x}\|^2\leqslant\|x_{n_1}-\bar{x}\|^2+\sum_{n=n_1}^{n_2}\delta_n\leqslant\delta \quad\text{for all } n_2\geqslant n_1. \qquad \square$$

Naturally, this last result is motivated by (1.3): to prove that $\{x_n\}$ satisfying (1.7) does converge to a minimum point $\bar{x}$ of f, our aim will be to establish

$$\sum_{n=1}^{\infty}\{t_n^2\|\delta_n\|^2+2t_n[f(\bar{x})-f(x_n)+\varepsilon_n]\}<+\infty.$$

2. Convergence of the prox-iteration

For given $t_n>0$ and $x_n\in X$, consider the following perturbation of f:

$$\tilde{f}(y):=f(y)+\frac{1}{2t_n}\|y-x_n\|^2, \tag{2.1}$$

which is inf-compact, strongly convex, and has the subdifferential

$$\partial \tilde{f}(y) = \partial f(y) + \frac{y - x_n}{t_n}.$$

In the prox-iteration [2, 15, 18], $\{t_n\}$ is chosen a priori and $\{x_n\}$ is constructed by

$$x_{n+1} = p_f(x_n), \tag{2.2}$$

where

$$p_f(x_n) := \operatorname{argmin}\{\tilde{f}(y) : y \in X\}. \tag{2.3}$$

Observe that $p_f(x_n)$ is well-defined, and the prox-sequence $\{x_n\}$ is characterized by $0 \in \partial f(x_{n+1}) + (x_{n+1} - x_n)/t_n$; in fact

$$x = p_f(x_n) \Leftrightarrow \frac{x_n - x}{t_n} \in \partial f(x). \tag{2.4}$$

Said otherwise, the prox-iteration can be put in the form (1.1), with

$$\gamma_n \in \partial f(x_{n+1}) \quad \text{for } n = 1, 2, \ldots. \tag{2.5}$$

Needless to say, computing γ_n or x_{n+1} are two equally difficult tasks; they are usually impossible, unless f has an amenable structure, for example quadratic or piecewise linear; the next sections, precisely, will be devoted to numerical aspects.

To establish convergence of the above (conceptual) prox-iteration along the lines of Section 1, a direct use of (2.5) instead of (1.2) is not easy: exchanging x_n and x_{n+1} in (1.3), the inequality sign will be reverted, and the summation mechanism (1.8) will result in nothing. The key is a "transfer" from x_{n+1} to x_n, given by part (i) of the next result:

Lemma 2.1. *Let* $\{x_n\} \subset \operatorname{dom} f$ *be constructed by* (1.1).

(i) *If* γ_n *satisfies* (2.5), *then* (1.2) *holds with*

$$\varepsilon_n = f(x_n) - f(x_{n+1}) - t_n \|\gamma_n\|^2 \geq 0. \tag{2.6}$$

(ii) *Conversely, if* (1.2) *holds with* ε_n *given by* (2.6), *then* (2.5) *holds*: $x_{n+1} = p_f(x_n)$.

Proof. For arbitrary $y \in X$, the subgradient inequality expressing (2.5) can be written

$$\begin{aligned} f(y) &\geq f(x_{n+1}) + f(x_n) - f(x_n) + \langle \gamma_n, y - x_n + x_n - x_{n+1} \rangle \\ &= f(x_n) + \langle \gamma_n, y - x_n \rangle - \varepsilon_n. \end{aligned}$$

Simply set $y = x_n$ to observe that $\varepsilon_n \geq 0$.

Conversely, the ε_n-subgradient inequality (1.2) with the value (2.6) boils down to

$$f(y) \geq f(x_{n+1}) + \langle \gamma_n, y - x_{n+1} \rangle \quad \text{for all } y \in X;$$

thus $\gamma_n [= (x_n - x_{n+1})/t_n] \in \partial f(x_{n+1})$; in view of (2.4), $x_{n+1} = p_f(x_n)$. □

Now write (2.6) as

$$0 \leqslant \varepsilon_n + t_n \|\gamma_n\|^2 = f(x_n) - f(x_{n+1});$$

unless $f(x_n) \to -\infty$, both ε_n and $t_n\|\gamma_n\|^2$ tend to 0 when $n \to +\infty$ and we are in the situation of Proposition 1.2: the minimizing property (1.7) holds under the sole assumption (1.4). We have here an argument similar to that in [5], [10]. In view of our care for numerical implementations, however, we consider a slightly more general situation:

Proposition 2.2. *With the notations* (1.1), (1.2), *let* $m>0$ *be given.*

(i) *If* (1.4), (1.5) *hold together with*

$$f(x_{n+1}) \leqslant f(x_n) - m t_n \|\gamma_n\|^2 \quad \text{for } n = 1, 2, \ldots, \tag{2.7}$$

then $f(x_n) \to \bar{f}$.

(ii) *Assume* X *is finite-dimensional. If, in addition,* $\{t_n\}$ *is bounded and* $\sum_{i=1}^{\infty} \varepsilon_n < +\infty$, *then* x_n *converges to a minimum point of* f *if there is some.*

Proof. (i) If the decreasing sequence $\{f(x_n)\}$ tends to $-\infty$, we are done. Otherwise, this sequence is bounded from below, (2.7) implies (1.6) and Proposition 1.2 applies.

(ii) Now assume that f has some minimum point $\bar{x}$; set $y = \bar{x}$ in (1.3) to obtain

$$\|x_{n+1} - \bar{x}\|^2 \leqslant \|x_n - \bar{x}\|^2 + t_n[t_n\|\gamma_n\|^2 + 2\varepsilon_n] \quad \text{for } n = 1, 2, \ldots. \tag{2.8}$$

Under our additional assumptions, $t_n[t_n\|\gamma_n\|^2 + 2\varepsilon_n]$ forms a convergent series, so $\{x_n\}$ is bounded and has some cluster point.

The lower semi-continuity of f and the first part of the proof imply that such a cluster point minimizes f; we can call it $\bar{x}$ and use this $\bar{x}$ in (2.8). Then Proposition 1.3 applies: the whole $\{x_n\}$ tends to $\bar{x}$. □

Corollary 2.3. *Let* $\{x_n\}$ *be constructed by* (2.2), (2.3) *and assume that* (1.4) *holds; then* $f(x_n) \to \bar{f}$. *Furthermore, assume* X *is finite-dimensional; then* $\{x_n\}$ *converges weakly to a minimum point of* f *if there is some.*

Proof. Since (2.2), (2.3) are equivalent to (1.1), (2.5), Lemma 2.1 shows that the prox-sequence $\{x_n\}$ satisfies (2.7) with $m = 1$. If $f(x_n) \to -\infty$, the proof is finished; otherwise, (2.6) implies (1.5) and Proposition 2.2 applies.

The rest is classical: let $\bar{x}$ minimize f and develop

$$\|x_n - \bar{x}\|^2 = \|x_n - x_{n+1}\|^2 + 2\langle x_n - x_{n+1}, x_{n+1} - \bar{x}\rangle + \|x_{n+1} - \bar{x}\|^2.$$

The property $f(\bar{x}) \leqslant f(x_{n+1})$ appended to the subgradient inequality expressing (2.5) shows that the cross-product above is nonnegative,

$$\|x_{n+1} - \bar{x}\|^2 \leqslant \|x_n - \bar{x}\|^2.$$

Terminate the proof as in Proposition 2.2(ii), this last inequality playing the role of (2.8) □

3. Implementable forms

For a concrete construction of the prox-sequence, some substitute must be found to (2.5) which, with the notation (1.1), characterizes the prox-mapping p_f. More precisely, two ingredients must now be provided: an algorithm to minimize $\tilde{f}$, and a rule to stop this algorithm; then the prox-center x will be updated and the next $\tilde{f}$ will be minimized.

The present section is devoted to the stopping rule, which we consider as the more important ingredient. Our point is that an accurate minimization of $\tilde{f}$ is *irrelevant*, as long as the real objective is f; said otherwise, an accurate computation of $p_f(x)$ is of little interest, as long as it has little to do with our original problem. As a result, the stopping criterion should operate early enough, possibly long before (2.3) is solved, even approximately. On the other hand, stopping too early kills the whole prox idea. A balance has to be found.

Of course, our stopping rule will be inspired by Proposition 2.2, and here comes a justification which we consider as very important: (2.7) can be written

$$f(x_{n+1}) \leqslant f(x_n) + m\langle \gamma_n, x_{n+1} - x_n \rangle$$

and, in view of (1.2) or (2.5), the term $\langle \gamma_n, x_{n+1} - x_n \rangle$ approximates $f(x_{n+1}) - f(x_n)$. A value $m < 1$ can then be interpreted as a tolerance coping with nonlinearity of f, and (2.7) is nothing but (an approximation of) the classical Armijo stepsize rule, widely used in numerical optimization (see [12]): to accept x_{n+1}, we require from f an improvement at least comparable to its "ideal" improvement $\langle \gamma_n, x_{n+1} - x_n \rangle$.

In this section, we are therefore given $x = x_n \in X$, $t = t_n > 0$, and some mechanism generating approximations of $p_f(x)$, call them y. The idea is to update $x_{n+1} = y$, so as to fit with Proposition 2.2: first (2.7) must hold, which we write as

$$f(y) \leqslant f(x) - \frac{m}{t}\|y - x\|^2 \quad \text{for some } m \in]0, 1]. \tag{3.1}$$

Naturally, this does not suffice to imply the minimizing property (1.7) (note, incidentally, that (3.1) defines a convex set containing $p_f(x)$, but also x). We must also have (1.5), namely

$$\frac{x - y}{t} \in \partial_\varepsilon f(x) \tag{3.2}$$

for some $\varepsilon = \varepsilon_n$ which will have to tend to 0.

Lemma 3.1. *With the notations above, suppose that y satisfies* (3.2) *with* ε *defined by*

$$\varepsilon(y)=\kappa\left[f(x)-f(y)-\frac{m}{t}\|y-x\|^2\right] \quad \text{for some } \kappa>0; \tag{3.3}$$

then (3.1) *holds.*

Proof. Trivial since (3.2), (3.3) implies $\varepsilon(y)\geqslant 0$. □

The need to test (3.1) is therefore eliminated, a minor advantage since this test is so simple; but more importantly, an obvious connection is established with the value (2.6).

Theorem 3.2. *With the notations above, suppose that f is strongly coercive, i.e.*

$$f(z)/\|z\| \to +\infty \quad \text{when } \|z\| \to \infty;$$

furthermore, take $m\leqslant 1$, $\kappa\geqslant 1$, *one of these inequalities being strict. If x does not minimize f, there is* $\delta>0$ *such that* (3.2), (3.3) *hold for any y satisfying*

$$\tilde{f}(y)\leqslant\tilde{f}(p_f(x))+\delta.$$

Proof. Assume for contradiction that there is a sequence $\{y^k\}$ such that

$$\tilde{f}(y^k)\to\tilde{f}(p_f(x)) \quad \text{and} \quad \frac{x-y^k}{t}\notin\partial_{\varepsilon(y_k)}f(x).$$

The strong convexity of $\tilde{f}$ implies that $y_k\to p_f(x)$ and

$$\varepsilon(y^k)\to\kappa\left[f(x)-f(p_f(x))-\frac{m}{t}\|p_f(x)-x\|^2\right]=:\varepsilon',$$

$$\frac{x-y^k}{t}\to\frac{x-p_f(x)}{t}\in\partial_{\varepsilon_0}f(x),$$

where

$$\varepsilon_0:=f(x)-f(p_f(x))-\frac{1}{t}\|x-p_f(x)\|^2.$$

Because x does not minimize f, the values of κ and m imply $0<\varepsilon_0<\varepsilon'$.

Now the approximate subdifferentials of f are tilted level-sets of its conjugate f^*, more precisely,

$$\partial_\varepsilon f(x)=\{\gamma\in X:\ f^*(\gamma)-\langle\gamma, x\rangle\leqslant f(x)+\varepsilon\}. \tag{3.4}$$

Because f is strongly coercive, f^* is finite everywhere, and $f^* - \langle \cdot, x \rangle$ is continuous on X ([16], Proposition 3.3); since $\varepsilon' > \varepsilon_0$, (3.4) then implies

$$\partial_{\varepsilon_0} f(x) \subset \operatorname{int} \partial_{\varepsilon'} f(x),$$

establishing the required contradiction. □

The way is thus open towards effective implementations of (2.2), (2.3), based on Proposition 2.2. For this, we just need an algorithm to generate a minimizing sequence $\{y^k\}$ of $\tilde{f}$, for fixed $x = x_n$.

Algorithm 3.3. Fix for example $\kappa > 1$ and $m \in]0, 1[$. Start from $x_1 \in X$, set $n = 1$.

Step 1. Set $k = 1$; start from some $y^k = y^1$.

Step 2. Set

$$\varepsilon = \kappa \left[f(x_n) - f(y^k) - \frac{m}{t_n} \|y^k - x_n\|^2 \right].$$

If

$$\frac{x_n - y^k}{t_n} \in \partial_\varepsilon f(x_n)$$

then go to Step 3; otherwise compute y^{k+1}, increase k by 1 and execute Step 2 again.

Step 3. Set $x_{n+1} = y^k$, increase n by 1 and loop to Step 1.

Under the assumptions of Theorem 3.2, it should now be clear that:

- for each n, Step 2 eventually exits to Step 3, unless x_n is already optimal (then $y^k \to x_n$);
- the sequence $\{x_n\}$ thus generated satisfies the minimizing property (1.7).

However, this is still theoretical: normally, there is no way of knowing whether a given γ is in a given ε-subdifferential at a given x; so how can we check (3.2)? A possible answer is provided by bundle methods, which somehow revert the logic contained in Lemma 3.1: they organize the calculations so that, when the *easy and natural* test (3.1) is satisfied, (3.2), (3.3) automatically holds.

4. The bundle variant

Bundle methods need the following assumption:

The convex function f is finite-valued over the whole of X, (4.1)

and they aim at solving our minimization problem with the sole help of the following information:

Given $x \in X$, the value $f(x)$ and some $g = g(x) \in \partial f(x)$ are available. (4.2)

Using the language of the present paper, these methods can be sketched as follows: (i) f in (2.3) is replaced by some simpler function, say φ; (ii) a solution, say y, of the modified problem (2.3) is computed; (iii) a test inspired from (2.7) decides whether the prox-center can be updated to this y, or must be kept fixed for the next iteration (which will be performed with an updated φ). This gives the following algorithm:

Algorithm 4.1 (prox-form of bundle methods). An initial point x_1 is given, together with a tolerance $m \in]0, 1[$ and a positive sequence $\{t_n\}$; set $n=k=1$.

Step 1. Choose a convex function $\varphi^k: X \to \mathbb{R}$. Solve for y

$$\min\left\{\varphi^k(y)+\frac{1}{2t_n}\|y-x_n\|^2\right\} \tag{4.3}$$

to obtain the unique optimal solution y^k, as well as $\gamma^k := (x_n-y^k)/t_n \in \partial\varphi^k(y^k)$.

Step 2. Compute $f(y^k)$. If a good decrease is obtained, namely if

$$f(x_n)-f(y^k) \geqslant m[f(x_n)-\varphi^k(y^k)], \tag{4.4}$$

then set $x_{n+1}=y^k$ and increase n by 1.

Step 3. Increase k by 1 and go to Step 1.

As compared to Algorithm 3.3, the index k still denotes an internal iteration, but we prefer not to reset it when the prox-center x_n is updated. This notation is better suited to our development. The expression of γ^k in Step 1 (called the aggregate subgradient by K.C. Kiwiel) corresponds to the characterization of $y^k=p_{\varphi^k}(x_n)$ via (2.4), (2.5). When the prox center is updated in Step 2, we say that a *descent-step* has been made, playing the role of one prox-iteration (2.2); otherwise, a *null-step* has been made, the new φ^{k+1} in the next iteration will supposedly improve the approximation of the true $p_f(x_n)$.

A key-object is then the *aggregate* affine function

$$y \mapsto l^k(y) := \varphi^k(y^k)+\langle \gamma^k, y-y^k\rangle; \tag{4.5}$$

because $\gamma^k \in \partial\varphi^k(y^k)$, it is easy to see that

$$l^k \leqslant \varphi^k \quad \text{and} \quad y^k=p_{\varphi^k}(x_n)=p_{l^k}(x_n). \tag{4.6}$$

Naturally, convergence of Algorithm 4.1 cannot hold for arbitrary $\{\varphi^k\}$; indeed, we require the properties (4.7)–(4.9) below:

$$\varphi^k \leqslant f \quad \text{for } k=1, 2, \ldots, \tag{4.7}$$

$$\left.\begin{aligned} &l^k \leqslant \varphi^{k+1} \quad \text{and} \\ &f(y^k)+\langle g^k, \cdot -y^k\rangle \leqslant \varphi^{k+1} \end{aligned}\right\} \text{if the } k\text{th iteration made a null-step.} \tag{4.8, 4.9}$$

In (4.9), $g^k=g(y^k)$ is the subgradient computed according to (4.2).

Remark 4.2. Bundle methods take piecewise affine functions φ^k. Then (4.3) is a linear-quadratic problem, for which there is no lack of efficient software, see for example [8]. For example, the following choices do satisfy (4.7)–(4.9):

- the "maximal" choice with k affine pieces, as is done in the cutting-plane algorithm of [3], [6]:

$$\varphi^{k+1}(y) := \max\{f(y^i) + \langle g^i, y - y^i\rangle : i = 1, \ldots, k\}$$

(with the notation $y^0 = x_1$, for a correct definition of φ^1);

- the choice with only 2 affine pieces:

$$\varphi^{k+1}(y) := \max\{l^k(y), f(y^k) + \langle g^k, y - y^k\rangle\} \tag{4.10}$$

which is "minimal" after a null-step;

- intermediate choices: select some index set $I^k \subset \{1, \ldots, k\}$ containing k, and then

$$\varphi^{k+1}(y) := \max\{l^k(y), f(y^i) + \langle g^i, y - y^i\rangle, i \in I^k\}.$$

All these choices, based on the cutting-plane idea, are arguably not very attractive; unfortunately, nothing better is known. It is amusing to note that, if iteration k has made a descent step, φ^{k+1} can be chosen arbitrary satisfying (4.7): null-steps will correct it if necessary. An apparently reasonable choice would be for example

$$\varphi^{k+1}(y) = f(x_n) + \langle g^k, y - x_n\rangle,$$

but the $(k+1)$st iteration would then be made along a subgradient (steepest descent if f is differentiable at $x_n = y^k$). This strategy is therefore not recommended: generally speaking, a richer I^k should improve the performance.

Interpreting φ^k as some approximation of f, (4.4) means that the prox-center x_n is updated when f decreases by at least a fraction of the number $f(x_n) - \varphi^k(y^k)$, which can be interpreted as a wished decrease of f, and is positive by necessity (unless $y^k = x_n$): from (4.7) and because $\gamma_k \in \partial\varphi^k(y^k)$,

$$f(x_n) - \varphi^k(y^k) \geqslant \varphi^k(x_n) - \varphi^k(y^k) \geqslant t_n \|\gamma^k\|^2.$$

As in (2.1), we consider the perturbations $\tilde{\varphi}^k$ and $\tilde{l}^k$ associated with φ^k and l^k; say from (4.5),

$$\tilde{l}^k(y) := l^k(y) + \frac{1}{2t_n}\|y - x_n\|^2;$$

$\tilde{l}$ being quadratic, the following equality is easy to establish (actually, we will need only an inequality, characterizing the strong convexity of $\tilde{l}$):

$$\tilde{l}^k(y) = \tilde{l}^k(y^k) + \frac{1}{2t_n}\|y - y^k\|^2 \quad \text{for all } y \text{ and } k. \tag{4.11}$$

The following result confirms that null-steps aim at computing $p_f(x_n)$. This was demonstrated in [4], see also [1]; but we follow the proof technique of [9]: it is made necessary by our abstract framework, and it lends itself to various generalizations.

Proposition 4.3. *Consider Algorithm* 4.1, *applied to a function satisfying* (4.1). *Suppose that, after a certain prox-center x_n has been reached, the descent test* (4.4) *is suppressed: only null-steps are made. If* (4.7)–(4.9) *hold, then*

$$f(y^k)-\varphi^k(y^k)\to 0, \qquad y^k\to p_f(x_n) \quad \textit{strongly}. \tag{4.12}$$

Proof. Call k_0 the iteration that has produced x_n: a null-step is made for each $k>k_0$. All the inequalities below will be understood "for $k>k_0$" – i.e. for k large enough.

Start from (4.7) and the definition of the $\check{}$-operation:

$$\begin{aligned} f(x_n)\geqslant\varphi^{k+1}(x_n)&=\tilde{\varphi}^{k+1}(x_n)\\ &\geqslant\tilde{\varphi}^{k+1}(y^{k+1}) \quad \text{(definition of } y^{k+1})\\ &=\tilde{l}^{k+1}(y^{k+1}) \quad \text{((4.5) and the } \check{}\text{-operation)}\\ &\geqslant\tilde{l}^{k}(y^{k+1}) \quad \text{(because of (4.8))}\\ &\geqslant\tilde{l}^{k}(y^{k})+\frac{1}{2t_n}\|y^{k+1}-y^k\|^2 \quad \text{(set } y=y^{k+1} \text{ in (4.11)).} \end{aligned}$$

Because x_n is fixed, these relations show that $\{\tilde{l}^k(y^k)\}$ is convergent and

$$y^{k+1}-y^k\to 0. \tag{4.13}$$

Furthermore, fix y in (4.11) to obtain with (4.6), (4.7) and the convergence of $\{\tilde{l}^k(y^k)\}$:

$$\{y^k\} \text{ is bounded.} \tag{4.14}$$

Now from (4.7) and (4.9),

$$f(y^{k+1})-f(y^k)\geqslant\varphi^{k+1}(y^{k+1})-f(y^k)\geqslant\langle g^k, y^{k+1}-y^k\rangle;$$

in view of (4.13)–(4.14) and the local Lipschitz property of f ([16], Theorem 2.28), $\varphi^{k+1}(y^{k+1})-f(y^k)\to 0$. Using (4.13) once more, this establishes (4.12).

Finally write the subgradient inequality $\gamma^k=(x_n-y^k)/t_n\in\partial\varphi^k(y^k)$:

$$f(y)\geqslant\varphi^k(y)\geqslant\varphi^k(y^k)+\langle\gamma^k, y-y^k\rangle \quad \text{for all } y \text{ and } k=1,2,\ldots, \tag{4.15}$$

and extract a subsequence such that $y^k\to\bar{y}$ weakly. Take first $y=\bar{y}$ to obtain

$$t_n[f(\bar{y})-\varphi^k(y^k)]\geqslant\langle x_n-y^k, \bar{y}-y^k\rangle=\langle x_n-\bar{y}, \bar{y}-y^k\rangle+\|\bar{y}-y^k\|^2$$

and pass to the limit: from (4.12), $y^k\to\bar{y}$ strongly. Then pass to the limit in (4.15) to see that $(x_n-\bar{y})/t_n\in\partial f(\bar{y})$; remembering (2.4), $\bar{y}=p_f(x_n)$. □

With this key result, the convergence properties of bundle methods can now be established:

Theorem 4.4. *In Algorithm* 4.1, *let* (4.7)–(4.9) *hold. Then there are two cases*:

- *Either some prox-center x_n is reached and, from then on, n remains fixed: only null-steps are made; then x_n actually minimizes f.*
- *Or $n \to +\infty$; then $f(x_n) \to \bar{f}$ if* (1.4) *holds. Assume X is finite-dimensional. If, in addition, $\{t_n\}$ is bounded, x_n converges to a minimum point of f if there is some.*

Proof. Suppose that, for some n, x_n is never updated: in view of (4.4),

$$f(x_n) - f(y^k) < m[f(x_n) - \varphi^k(y^k)]$$

(for all k larger than the k_0 of Proposition 4.3). Letting $k \to +\infty$, we obtain with Proposition 4.3,

$$(1-m)[f(x_n) - f(p_f(x_n))] \leqslant 0;$$

because $m < 1$, this implies $f(x_n) \leqslant f(p_f(x_n))$, i.e. $x_n = p_f(x_n)$: x_n minimizes f.

The other alternative is that every x_n is eventually updated at some iteration $k = k(n)$. We set $\gamma_n = \gamma^{k(n)}$ to use the notation (1.1); then we refine the proof of Corollary 2.3.

First write the subgradient inequality $\gamma_n \in \partial\varphi^{k(n)}(x_{n+1})$: for all n and y,

$$\begin{aligned} f(y) &\geqslant \varphi^{k(n)}(y) \geqslant \varphi^{k(n)}(x_{n+1}) + \langle \gamma_n, y - x_{n+1} \rangle \\ &= f(x_n) + \varphi^{k(n)}(x_{n+1}) - f(x_n) + \langle \gamma_n, y - x_n \rangle + t_n \|\gamma_n\|^2, \end{aligned}$$

to see that (1.2) holds with

$$\varepsilon_n = f(x_n) - \varphi^{k(n)}(x_{n+1}) - t_n\|\gamma_n\|^2 \leqslant \frac{f(x_n) - f(x_{n+1})}{m} - t_n\|\gamma_n\|^2.$$

Thus, if $\{f(x_n)\}$ is bounded from below,

$$\sum_{n=1}^{\infty} [\varepsilon_n + t_n\|\gamma_n\|^2] < +\infty; \tag{4.16}$$

(1.5) and (1.6) hold, $f(x_n) \to \bar{f}$ because of Proposition 1.2.

Finally suppose that f has a minimum point $\bar{x}$; use Lemma 1.1 with $y = \bar{x}$,

$$\|x_{n+1} - \bar{x}\|^2 \leqslant \|x_n - \bar{x}\|^2 + t_n[t_n\|\gamma_n\|^2 + 2\varepsilon_n].$$

In view of (4.16), we are exactly in the situation of Proposition 2.2(ii). □

A variant of Algorithm 4.1 would be to update the prox-coefficient t_n even after null-steps. This would complicate the convergence analysis; furthermore, our proposed strategy fits better in the framework of the present paper: if y^k were computed with a varying prox-coefficient t^k—instead of a fixed t_n—we could hardly speak of $p_f(x_n)$.

Remark 4.6. A more technical proof of Proposition 4.3 is based on

$$\tilde{\varphi}^{k+1}(y^{k+1}) \geqslant \min_{y} \tilde{\varphi}_m(y),$$

where $\tilde{\varphi}_m$ is the "minimal" function obtained by taking $\tilde{\varphi}^{k+1}$ as in (4.10). The minimum of $\tilde{\varphi}_m$ can be explicitly computed; with the help of (4.4), the number of null-steps between two descent steps can then be majorized. Such a technique was actually used in earlier versions, see [20, 11, 7]; see also [14].

Needless to say, the convergence condition (1.4) makes little sense numerically (who is patient enough to check whether a series is divergent?). For efficient implementations, it must be replaced by an "on line" computation of t_n (or t^k), as in [9, 19].

Historically, Algorithm 4.1 has its roots in [13], in which the controlled parameter is ε_n, instead of t_n. More precisely, the algorithm of [13] is as follows:

- Given the prox-center x_n and the approximation φ^k, choose a parameter ε and project the origin onto $\partial_\varepsilon \varphi^k(x_n)$, i.e. solve

$$\min \tfrac{1}{2}\|\gamma\|^2 \quad \text{subject to } (\varphi^k)^*(\gamma) - \langle \gamma, x_n \rangle \leqslant \varphi^k(x_n) + \varepsilon. \tag{4.17}$$

- Then choose a stepsize τ and take the next iterate y^k as

$$x_n - \tau\gamma =: z.$$

With Remark 4.2 in mind, we mention that the conjugate $(\varphi^k)^*$ of the piecewise affine function φ^k can be computed explicitly. Under appropriate choices of the parameters, this algorithm is equivalent to 4.1:

Proposition 4.7. *If ε is chosen so that, a posteriori, t_n is inverse to a multiplier of the constraint in* (4.17), *and if $\tau = t_n$, then $z = p_{\varphi^k}(x_n)$.*

Proof. By assumption, we have

$$0 \in \gamma + \frac{\partial(\varphi^k)^*(\gamma) - x_n}{t_n}, \quad \text{i.e. } \gamma \in \partial\varphi^k(x_n - t_n\gamma), \quad \text{i.e. } \frac{x_n - z}{t_n} \in \partial\varphi^k(z).$$

In view of (2.4), this characterizes the prox-operation. □

Needless to say, the value of ε mentioned in Proposition 4.7 is that given in (2.6), namely

$$\varphi^k(x_n) - \varphi^k(p_{\varphi^k}(x_n)) - \frac{\|x_n - p_{\varphi^k}(x_n)\|^2}{t_n}.$$

5. Subgradient algorithms

The convergence condition (1.4) strongly connotes traditional subgradient methods, in which γ_n is directly given by (4.2) and is normalized: rather than (1.1), the recurrence is

$$x_{n+1} = x_n - \tau_n \frac{\gamma_n}{\|\gamma_n\|} \tag{5.1}$$

for some stepsize $\{\tau_n\}$. Indeed, our analysis applies to this case:

Proposition 5.1. *Assume* (4.1). *Let* $\{x_n\}$ *be generated by* (5.1), *with* $\gamma_n \in \partial f(x_n)$ *and*

$$\sum_{n=1}^{\infty} \tau_n = +\infty, \tag{5.2}$$

$$\sum_{n=1}^{\infty} \tau_n^2 < +\infty; \tag{5.3}$$

then $\liminf f(x_n) = \bar{f}$. *Furthermore, assume X is finite-dimensional. Then* x_n *tends to a minimum point of f if there is some.*

Proof. In (1.3), set y to some x^* such that $f(x^*) \leqslant f(x_n)$ for all n (if there is no such x^*, the proof is finished). We obtain

$$\|x_{n+1} - x^*\|^2 \leqslant \|x_n - x^*\|^2 + \tau_n^2 \tag{5.4}$$

and (5.3) implies that $\{x_n\}$ is bounded, so $\{\gamma_n\}$ is bounded as well. To use the notation (1.1), (1.2), we set $t_n = \tau_n / \|\gamma_n\|$; then (1.4) comes from (5.2), while (5.3) implies $\tau_n \to 0$, hence (1.6) holds; (1.5) is automatic; altogether, Proposition 1.2 applies. Finally, (5.3) and (5.4) allow the use of Proposition 1.3: just follow the pattern of Proposition 2.2(ii) to show that $\{x_n\}$ must converge to a minimum of f if there is one. □

To conclude, we recall here that the property $\liminf f(x_n) = \bar{f}$ was already proved in [17], with (5.3) replaced by the weaker requirement $\tau_n \to 0$. The technique was similar, based on a variant of (1.3).

Acknowledgment

We are indebted to the referees for a very careful and critical reading of this paper.

References

[1] A. Auslender, "Numerical methods for nondifferentiable convex optimization," in: B. Cornet, N.V. Hien and J.P. Vial, eds. *Nonlinear Analysis and Optimization. Mathematical Programming Study No. 30* (1987) pp. 102–126.

[2] R.E. Bellman, R.E. Kalaba and J. Lockett, *Numerical Inversion of the Laplace Transform* (Elsevier, New York, 1966) pp. 143–144.

[3] E.W. Cheney and A.A. Goldstein, "Newton's method for convex programming and Tchebycheff approximation," *Numerische Mathematik* 1 (1959) 253–268.

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

[5] O. Güler, "On the convergence of the proximal point algorithm for convex minimization," *SIAM Journal on Control and Optimization* 29 (1991) 403–419.

[6] J.E. Kelley, "The cutting plane method for solving convex programs," *Journal of the Society for Industrial and Applied Mathematics* 8 (1960) 703–712.

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

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

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

[10] B. Lemaire, "About the convergence of the proximal method," in: D. Pallaschke, ed., *Advances in Optimization. Lecture Notes in Economics and Mathematicsl Systems No. 382* (Springer, Berlin, 1992) pp. 39–51.

[11] C. Lemaréchal, "An extension of Davidon methods to non differentiable problems," in: M.L. Balinski and P. Wolfe, eds., *Nondifferentiable Optimization. Mathematical Programming Study No. 3* (1975) pp. 95–109.

[12] C. Lemaréchal, "A view of line-searches," in: A. Auslender, W. Oettli and J. Stoer, eds., *Optimization and Optimal Control. Lecture Notes in Control and Information Science No. 30* (Springer, Berlin, 1980) pp. 59–78.

[13] C. Lemaréchal, J.J. Strodiot and A. Bihain, "On a bundle algorithm for nonsmooth optimization," in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., *Nonlinear Programming 4* (Academic Press, New York, 1981) pp. 245–282.

[14] C. Lemaréchal, A.S. Nemirovskii and Yu.E. Nesterov, "New variants of bundle methods," INRIA Report No. RR1508 (Le Chesnay, 1991).

[15] B. Martinet, "Régularisation d'inéquations variationnelles par approximation successives," *Revue Française d'Informatique et Recherche Opérationnelle* R3 (1970) 154–158.

[16] R.R. Phelps, *Convex Functions, Monotone Operators and Differentiability. Lecture Notes in Mathematics No. 1364* (Springer, Berlin, 1989).

[17] B.T. Poljak, "A general method of solving extremum problems," *Soviet Mathematics Doklady* 8 (1967) 593–597.

[18] R.T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research* 1 (1976) 97–116.

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

[20] P. Wolfe, "A method of conjugate subgradients for minimizing nondifferentiable functions," in: M.L. Balinski and P. Wolfe, eds., *Nondifferentiable Optimization. Mathematical Programming Study No. 3* (1975) pp. 145–173.