Mathematical Programming 81 (1998) 23–35

On the projected subgradient method for nonsmooth convex optimization in a Hilbert space

Ya.I. Alber [1], A.N. Iusem [*,2], M.V. Solodov [2]

Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ, CEP 22460-320, Brazil

Received 18 January 1996; revised manuscript received 1 January 1997

Abstract

We consider the method for constrained convex optimization in a Hilbert space, consisting of a step in the direction opposite to an ε_k-subgradient of the objective at a current iterate, followed by an orthogonal projection onto the feasible set. The normalized stepsizes α_k are exogenously given, satisfying $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, and ε_k is chosen so that $\varepsilon_k \leqslant \mu\alpha_k$ for some $\mu > 0$. We prove that the sequence generated in this way is weakly convergent to a minimizer if the problem has solutions, and is unbounded otherwise. Among the features of our convergence analysis, we mention that it covers the nonsmooth case, in the sense that we make no assumption of differentiability of f, and much less of Lipschitz continuity of its gradient. Also, we prove weak convergence of the whole sequence, rather than just boundedness of the sequence and optimality of its weak accumulation points, thus improving over all previously known convergence results. We present also convergence rate results. © 1998 The Mathematical Programming Society, Inc. Published by Elsevier Science B.V.

Keywords: Convex optimization; Nonsmooth optimization; Projected gradient method; Steepest descent method; Weak convergence; Convergence rate

1. Introduction

We consider in this paper an extension of the projected subgradient method for convex optimization in a Hilbert space H. Let C be a closed and convex subset of H and $f : H \to \mathbb{R}$ a convex and continuous function. The problem under consideration is

* Corresponding author. Fax: +55-21 529 5129; e-mail: solodov@impa.br.

[1] Permanent address: Department of Mathematics, The Technion – Israel Institute of Technology, 32000 Haifa, Israel.

[2] Research of this author was partially supported by CNPq grant nos. 301280/86 and 300734/95-6.

S0025-5610/98/$19.00 © 1998 The Mathematical Programming Society, Inc.
Published by Elsevier Science B.V.
PII S0025-5610(97)00031-2

$$\min \quad f(x) \tag{1}$$

$$\text{s.t.} \quad x \in C. \tag{2}$$

The projected subgradient method consists of generating a sequence $\{x^k\}$, by taking from x^k a step in the direction opposite to a subgradient of f at x^k and then projecting the resulting vector orthogonally onto C. When $C = H$ and f is differentiable this is just the steepest descent method. Different variants of the method arise according to the rule used to choose the stepsizes. Frequently, these are chosen so as to ensure functional decrease at each iteration, e.g. through either exact one dimensional minimization or an Armijo-type search. The first option cannot be implemented in actual computation and the second one works only when f is smooth. In the nonsmooth case the only reasonable alternative seems to be exogenously given stepsizes. In this paper we use stepsizes α_k satisfying $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. This selection rule has been considered several times in the literature (e.g. [1,2]).

We also generalize the projected subgradient method by allowing inexact computation of the subgradient: the kth direction need not be a subgradient of f at x^k but rather an ε_k-subgradient, where $\{\varepsilon_k\}$ is a nonincreasing sequence of nonnegative numbers satisfying $\varepsilon_k \leqslant \mu\alpha_k$ for some $\mu > 0$. We remark that these two features (exogenously given stepsizes and inexact subgradients) have as a consequence that the sequence of functional values need not be decreasing, which provokes considerable complications in the convergence analysis. Nevertheless, we establish that the sequence $\{x^k\}$ is always a "minimizing" one (in the sense that $\liminf_{k\to\infty} f(x^k) = \inf_{x\in C} f(x)$), that it is weakly convergent to a solution of (1) and (2) when this problem has solutions, and that it is unbounded otherwise.

We emphasize three features of our convergence analysis.

1. We make no differentiability assumptions on f, and much less on Lipschitz continuity of its gradient. Convexity and continuity of f are enough. We also need no boundedness assumption either on C or on the level sets of f; in fact the solution set might even be unbounded.
2. We prove weak convergence of the whole sequence to a solution (provided that a solution exists) rather than just boundedness of the sequence and optimality of all its weak accumulation points.
3. All our results hold in a Hilbert space (of course, in the finite dimensional case we get strong, rather than weak, convergence).

In Section 2, after a formal statement of the algorithm, we will compare our result with other related results in the literature, particularly in connection with the features mentioned above.

2. Statement of the algorithm and discussion of related results

Let H be a Hilbert space, C a closed and convex subset of H, and $f : H \to \mathbb{R}$ a convex and continuous function. We assume that f is finite valued, so that its effec-

tive domain is H. We remind that for $\varepsilon \geqslant 0$ the ε-subdifferential of f at x is the set $\partial_\varepsilon f(x)$ defined by

$$\partial_\varepsilon f(x) = \{u \in H\colon f(y) - f(x) \geqslant \langle u, y - x\rangle - \varepsilon \text{ for all } y \in H\}. \tag{3}$$

Since f is convex and continuous, and its effective domain is H, $\partial_\varepsilon f(x)$ is nonempty for all $\varepsilon \geqslant 0$ and all $x \in H$ [3], Lemma, p. 174 and Theorem 9, p. 112. We also mention that a sufficient condition for continuity of a convex function f at any x in H is boundedness of f at some neighborhood of some $\bar{x} \in H$ [3], Theorem 8, p. 110. We need the following boundedness assumption on $\partial_\varepsilon f$.

(A) $\partial_\varepsilon f$ is bounded on bounded sets, i.e. $\bigcup_{x\in B} \partial_\varepsilon f(x)$ is bounded for any bounded subset B of H.

In connection with (A) we mention that $\partial_\varepsilon f$ is always locally bounded (i.e. for any $\overline{X} \in H$ there exists a neighborhood V of $\overline{X}$ such that $\bigcup_{x\in V} \partial_\varepsilon f(x)$ is bounded). This follows from local boundedness of ∂f [4], Theorem 1 and the fact that for all bounded B, $\operatorname{diam}(\bigcup_{x\in B} \partial_\varepsilon f(x)) \leqslant \operatorname{diam}(\bigcup_{x\in B} \partial f(x)) + \varepsilon/\operatorname{diam}(B)$, where $\operatorname{diam}(B) = \sup\{\|x - y\|\colon x, y \in B\}$ [5], Lemma 1. In finite dimension, this result implies, through an easy compactness argument, that (A) always holds, but this is not the case in a Hilbert space, as the following example shows: let $H = \ell^2$ and $f(x) = \sum_{n=1}^{\infty} (2n)^{-1} (x_n)^{2n}$. It is easy to check that f is well defined, convex and differentiable, with $\nabla f(x)_n = (x_n)^{2n-1}$. Take now $e^j \in \ell^2$ defined as $e^j_n = 2\delta_{jn}$ (Kronecker's delta) and observe that $\|e^j\| = 2$ for all j while $\|\nabla f(e^j)\| = 2^{2j-1}$, i.e. ∇f is unbounded in the ball with center at 0 and radius 2. A sufficient (indeed also necessary) condition for (A) to hold is that $|f|$ is bounded on bounded sets: in order to prove that $\bigcup_{x\in B} \partial_\varepsilon f(x)$ is bounded, take $u \in \partial_\varepsilon f(x)$, let $x' = x + u/\|u\|$ and get, by definition of $\partial_\varepsilon f$, $\|u\| = \langle u, x' - x\rangle \leqslant f(x') - f(x) + \varepsilon \leqslant |f(x')| + |f(x)| + \varepsilon$. Let $B' = \{x \in H\colon \|y - x\| \leqslant 1 \text{ for some } y \in B\}$. Then B' is bounded and $x' \in B'$, so that we get a bound of $\|u\|$ in terms of ε and the bounds of $|f|$ on B and B'. We also remind that the subdifferential $\partial f(x)$ of f coincides with $\partial_0 f(x)$, i.e. the right hand side of Eq. (3) with $\varepsilon = 0$.

Take a sequence $\{\alpha_k\}$ of nonnegative real numbers satisfying

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \tag{4}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty \tag{5}$$

and a nonincreasing sequence of nonnegative real numbers $\{\varepsilon_k\}$ such that there exists $\mu > 0$ satisfying

$$\varepsilon_k \leqslant \mu\alpha_k \tag{6}$$

for all k. Let $P_C : H \to C$ be the orthogonal projection onto C. The algorithm is defined as follows.

Initialization

$$x^0 \in H. \tag{7}$$

Iterative step. Given x^k, if $0 \in \partial f(x^k)$ then stop. Otherwise, take $u^k \in \partial_{\varepsilon_k} f(x^k)$, $u^k \neq 0$, let $\eta_k = \max\{1, \|u^k\|\}$ and define

$$x^{k+1} = P_C\left(x^k - \frac{\alpha_k}{\eta_k} u^k\right) \tag{8}$$

with α_k, ε_k, satisfying Eqs. (4)–(6).

We make now some remarks on Eqs. (7) and (8). First, note that $\alpha_k = \varepsilon_k = 1/k$ satisfies Eqs. (4)–(6). Secondly, in connection with the stopping criterion, it is usually assumed, in the nonsmooth case, that an "oracle" is available which can provide an ε-subgradient of f at any $x \in H$. Our stopping criterion requires a little bit more: besides the oracle, we assume that we have a procedure which decides whether a given vector (the null vector in our case) is or is not a subgradient of f at x. This looks reasonable, since checking the subgradient inequality for a given vector should be easier than finding a vector which satisfies it, but if such a procedure is not available, then the iterative step should be rewritten as: "Take $u^k \in \partial_{\varepsilon_k} f(x^k)$; if $u^k = 0$ stop, otherwise let $\eta_k = \cdots$". In this case two consequences follow. First, in the stopping case we can only ensure that x^k is an ε_k-solution (meaning that $f(x^k) \leqslant f(x^*) + \varepsilon_k$, where x^* is a solution of (1) and (2)). Secondly, the sequence can hit an exact solution at iteration k and nevertheless continue, converging eventually to the same solution or to another one.

We discuss next convergence results on algorithms related to (7) and (8). First we mention that assuming differentiability of f, finite dimension of H and Lipschitz continuity of ∇f with constant L, it is rather straightforward to prove that $\{x^k\}$ is bounded and all its accumulation points are solutions of (1) and (2), when the problem has solutions. In this case (5) can be relaxed to $\alpha_k < 2L^{-1}$ (see e.g. [2] for the finite dimensional case, i.e. $H = \mathbb{R}^n$). The unrestricted case (i.e. $C = H$) in a Hilbert space with nonsmooth f and exact subgradients (i.e. $x^{k+1} = x^k - \alpha_k u^k / \|u^k\|$, with $u^k \in \partial f(x^k)$) is considered in [6], where it is proved mainly that the sequence is minimizing (i.e. $\liminf_{k\to\infty} f(x^k) = \inf_{x\in H} f(x)$) but no results are given on convergence of $\{x^k\}$. In this work Eq. (5) is relaxed to $\lim_{k\to\infty} \alpha_k = 0$.

The constrained case with exact subgradients and the same rule for α_k (i.e. with $\lim_{k\to\infty} \alpha_k = 0$ instead of (5)) is studied in [7], but only in the case of finite dimensional H. In addition, boundedness of C or of the level sets of f is assumed. The result is the same as in [6], namely that $\{x^k\}$ is a minimizing sequence.

In [8] the unrestricted case is considered in a Hilbert space with f differentiable and ∇f Lipschitz or Hölder continuous. α_k is given by explicit formulae in terms of the Lipschitz or Hölder constants. Under these hypotheses it is proved that if the problem has a solution then $\{x^k\}$ is bounded and all its weak accumulation points are optimal. For finite dimensional H, convergence of the whole sequence $\{x^k\}$ to a solution is established. The algorithm uses exact gradients (i.e. $u^k = \nabla f(x^k)$). This work, as well as [9], also presents convergence rate results, which are related to our results in Section 3. The constrained case with nonsmooth f, and α_k satisfying (4) and $\lim_{k\to\infty} \alpha_k = 0$ (instead of (5)), is studied in [9]. The iterative step

is given by a formula similar to (8) with $u^k \in \partial f(x^k)$, but an error term is allowed, as in [10], which we discuss below. It is assumed that ∂f is uniformly monotone, i.e. $\langle u - v, x - y\rangle \geqslant \varphi(\|x - y\|)$ for all $x, y \in H$, $u \in \partial f(x)$, $v \in \partial f(y)$, where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies $\varphi(0) = 0$ and some additional regularity conditions. In fact, the algorithm is discussed in the context of variational inequalities and a general operator T is used instead of ∂f. We mention that uniform monotonicity of ∂f does not imply differentiability of f, but it implies uniqueness of the solution. Under this rather strong assumption on ∂f, it is possible to prove strong convergence of $\{x^k\}$ to the unique solution of the problem.

The unconstrained case with exact subgradients and our rule for α_k (i.e. Eqs. (4) and (5)) is studied in [1]. The iteration is of the form $x^{k+1} = x^k - \alpha_k u^k / \|u^k\|$ with $u^k \in \partial f(x^k)$. In this work convergence of the whole sequence to a solution is proved (provided that the problem has solutions) without further assumptions on f, like those used in [8,9], but the result is obtained only in the finite dimensional case, previously considered in [11,12]. In infinite dimension, it is only proved in [1] that $\{x^k\}$ is a minimizing sequence, like in [7,6].

The case of inexact subgradients is considered in [9,10]. The iteration in [9,10] is of the form $x^{k+1} = P_C(x^k - \alpha_k(u^k + v^k))$ with $u^k \in \partial f(x^k)$. In [9] it is assumed that $\lim_{k\to\infty} v^k = 0$. In [10], on the other hand, the hypothesis is that $\|v^k\| \leqslant \tau$, i.e. an error of magnitude less that τ is allowed in the computation of the subgradient. In [10] f is not assumed to be convex, just locally Lipschitzian, but in the convex case this algorithm is virtually identical to (7) and (8), since $\partial_\varepsilon f(x)$ contains and is contained in the image through ∂f of appropriate balls around x. Ref. [10] characterizes the set of attractors of the sequence $\{x^k\}$, which is shown to consist of approximate solutions of the problem.

Finally we mention briefly some similar results for the unconstrained and smooth case with an Armijo-type rule for the α_k's. It is rather straightforward to prove boundedness of the sequence and optimality of the accumulation points assuming boundedness of the level sets of f and Lipschitz continuity of its gradient, and this result can be found in several text books (e.g. [13,2]). Without such assumptions, i.e. assuming just convexity and continuous differentiability of f, together with existence of solutions, convergence of the whole sequence to one solution has been established in [14–16] for finite dimensional spaces, and in [17] for Hilbert spaces. Our analysis in this paper uses several results which appear in [15,17], particularly the notion of quasi-Fejér convergence.

Some of the results presented here have been further extended by the authors in the subsequent paper [21], where convergence analysis for subgradient-type methods is developed for uniformly smooth and uniformly convex Banach spaces.

3. Convergence analysis

We need first two preliminary results unrelated to the algorithm. The first one is needed mainly to ensure uniqueness of the weak accumulation point of $\{x^k\}$. The sec-

ond one, on series of nonnegative real numbers, is related to conditions (4) and (5) on the α_k's. Loosely speaking, condition (5) ensures that the stepsizes are small enough to guarantee boundedness of $\{x^k\}$, while Eq. (4) ensures that they are not too small, in which case $\{x^k\}$ could get stuck midway to the solution set, i.e. converge to a point which is not a solution. Our second preliminary result will be used together with Eq. (5) to establish that $\{x^k\}$ is a minimizing sequence, i.e. that $\liminf_{k\to\infty} f(x^k) = \inf_{x\in C} f(x)$. In order to state our first result we need a definition.

Definition 1. Let H be a Hilbert space and V a nonempty subset of H. A sequence $\{x^k\} \subset H$ is said to be quasi-Fejér convergent to V iff for all $\bar{x} \in V$ there exists $\tilde{k} \geqslant 0$ and a sequence $\{\delta_k\} \subset \mathbb{R}_+$ such that $\sum_{k=0}^{\infty} \delta_k < \infty$ and $\|x^{k+1} - \bar{x}\|^2 \leqslant \|x^k - \bar{x}\|^2 + \delta_k$ for all $k \geqslant \tilde{k}$.

This definition originates in [18] and has been further elaborated in [19].

Proposition 1. *If $\{x^k\}$ is quasi-Fejér convergent to V then:*

1. *$\{x^k\}$ is bounded,*
2. *$\{\|x^k - \bar{x}\|^2\}$ converges for all $\bar{x} \in V$,*
3. *if all weak accumulation points of $\{x^k\}$ belong to V then $\{x^k\}$ is weakly convergent, i.e. it has a unique accumulation point.*

Proof. (i) Using recurrently Definition 1 for $k > \tilde{k}$, $\|x^k - \bar{x}\|^2 \leqslant \|x^{\tilde{k}} - \bar{x}\|^2 + \sum_{j=\tilde{k}}^{k-1} \delta_j \leqslant \|x^{\tilde{k}} - \bar{x}\|^2 + \sum_{j=0}^{\infty} \delta_j$. So the tail of the sequence, i.e. $\{x^k\}_{k>\tilde{k}}$, is contained in a certain ball centered at $\bar{x}$, and the result follows.

(ii) The sequence $\{\|x^k - \bar{x}\|^2\}$ is bounded by (i). Assume that it has two accumulations points, say v and ξ. Take subsequences $\{x^{j_k}\}$ and $\{x^{\ell_k}\}$ of $\{x^k\}$ such that $\lim_{k\to\infty} \|x^{j_k} - \bar{x}\|^2 = v$, $\lim_{k\to\infty} \|x^{\ell_k} - \bar{x}\|^2 = \xi$. Fix $\lambda > 0$. Take $\hat{k}$ such that $\ell_{\hat{k}} > \tilde{k}$ and $\|x^{\ell_k} - \bar{x}\|^2 \leqslant \xi + \frac{1}{2}\lambda$ for all $k > \hat{k}$. Take $\bar{k} \geqslant \hat{k}$ such that $\sum_{i=\ell_{\bar{k}}}^{\infty} \delta_i \leqslant \frac{1}{2}\lambda$. Using recurrently Definition 1, we get, for all k such that $j_k > \ell_{\bar{k}}$,

$$\|x^{j_k} - \bar{x}\|^2 \leqslant \|x^{\ell_{\bar{k}}} - \bar{x}\|^2 + \sum_{i=\ell_{\bar{k}}}^{j_k-1} \delta_i \leqslant \|x^{\ell_{\bar{k}}} - \bar{x}\|^2 + \sum_{i=\ell_{\bar{k}}}^{\infty} \delta_i \leqslant \|x^{\ell_{\bar{k}}} - \bar{x}\|^2 + \frac{\lambda}{2}$$

$$\leqslant v + \frac{\lambda}{2} + \frac{\lambda}{2} \leqslant v + \lambda \tag{9}$$

using $\ell_{\hat{k}} > \tilde{k}$ in the leftmost inequality and $\bar{k} > \hat{k}$ in the rightmost inequality. Taking limits in Eq. (9) as k goes to ∞, we get $\xi \leqslant v + \lambda$ for all $\lambda > 0$. It follows that $\xi \leqslant v$. Reversing the roles of $\{x^{\ell_k}\}$ and $\{x^{j_k}\}$ a similar argument shows that $v \leqslant \xi$, and we conclude therefore that $\xi = v$. It follows that all accumulation points of $\{\|x^k - \bar{x}\|^2\}$ coincide, i.e. that $\{\|x^k - \bar{x}\|^2\}$ converges (not necessarily to 0).

(iii) Existence of weak accumulation points of $\{x^k\}$ follows from (i). Let $\tilde{x}, \hat{x}$ be two weak accumulation points of $\{x^k\}$ and $\{x^{\tilde{j}_k}\}$, $\{x^{\hat{j}_k}\}$ be two subsequences of $\{x^k\}$ weakly convergent to $\tilde{x}, \hat{x}$ respectively. Let $\pi = \lim_{k\to\infty} \|x^k - \tilde{x}\|^2$, $\zeta = \lim_{k\to\infty} \|x^k - \hat{x}\|^2$. π and ζ exist by (ii), since $\tilde{x}, \hat{x}$ belong to V by hypothesis. Let $\omega = \|\hat{x} - \tilde{x}\|^2$. Then:

$$\|x^{\hat{j}_k} - \tilde{x}\|^2 = \|x^{\hat{j}_k} - \hat{x}\|^2 + \|\hat{x} - \tilde{x}\|^2 + 2\langle x^{\hat{j}_k} - \hat{x}, \hat{x} - \tilde{x}\rangle, \tag{10}$$

$$\|x^{\tilde{j}_k} - \hat{x}\|^2 = \|x^{\tilde{j}_k} - \tilde{x}\|^2 + \|\tilde{x} - \hat{x}\|^2 + 2\langle x^{\tilde{j}_k} - \tilde{x}, \tilde{x} - \hat{x}\rangle. \tag{11}$$

Take limits in Eqs. (10) and (11) as k goes to ∞, observing that the inner products in the right hand sides of Eqs. (1) and (11) converge to 0 because $\hat{x}, \tilde{x}$ are the weak limits of $\{x^{\hat{j}_k}\}$, $\{x^{\tilde{j}_k}\}$ respectively, and get, using the definitions of π, ζ, ω,

$$\pi = \zeta + \omega, \tag{12}$$

$$\zeta = \pi + \omega. \tag{13}$$

From Eqs. (12) and (13), we get $\pi - \zeta = \omega = \zeta - \pi$, which implies $\omega = 0$, i.e. $\tilde{x} = \hat{x}$. It follows that all weak accumulation points of $\{x^k\}$ coincide, i.e. that $\{x^k\}$ is weakly convergent. □

A slightly stronger result holds in the finite dimensional case: it is enough to have one accumulation point in V in order to ensure convergence of $\{x^k\}$. The proof, much easier than in the Hilbert space case, can be found in [15]. In the finite dimensional case, as a consequence of the observation just made, item (ii) of Proposition 2 is not needed. The result of Proposition 1(ii) in the finite dimensional case appears in [12, Lemma 3.2.1].

Proposition 2. *Let* $\{\alpha_k\}$, $\{\beta_k\} \subset \mathbb{R}$. *Assume that* $\alpha_k \geqslant 0$ *for all* $k \geqslant 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k \beta_k < \infty$ *and there exists* $\tilde{k} \geqslant 0$ *such that* $\beta_k \geqslant 0$ *for* $k \geqslant \tilde{k}$. *Then:*

(i) *there exists a subsequence* $\{\beta_{i(k)}\}$ *of* $\{\beta_k\}$ *such that* $\lim_{k\to\infty} \beta_{i(k)} = 0$.

(ii) *If, additionally, there exists* $\theta > 0$ *such that* $\beta_{k+1} - \beta_k \leqslant \theta\alpha_k$ *for all* k *then* $\lim_{k\to\infty} \beta_k = 0$.

Proof. (i) If the result does not hold then there exists $\sigma > 0$ and $\bar{k} \geqslant \tilde{k}$ such that $\beta_k \geqslant \sigma$ for all $k \geqslant \bar{k}$, so that $\infty > \sum_{k=\bar{k}}^{\infty} \alpha_k \beta_k \geqslant \sigma \sum_{k=\bar{k}}^{\infty} \alpha_k$, in contradiction with $\sum_{k=0}^{\infty} \alpha_k = \infty$.

(ii) By (i) there exists a subsequence $\{\beta_{i(k)}\}$ of $\{\beta_k\}$ such that $\lim_{k\to\infty} \beta_{i(k)} = 0$. If the result does not hold then there exists some $\sigma > 0$ and some other subsequence $\{\beta_{m(k)}\}$ of $\{\beta_k\}$ such that $\beta_{m(k)} \geqslant \sigma$ for all k. In this case, we can construct a third subsequence $\{\beta_{j(k)}\}$ of $\{\beta_k\}$, where the subindices $j(k)$ are chosen in the following way:

$$j(0) = \min\{\ell \geqslant 0 \colon \beta_\ell \geqslant \sigma\} \tag{14}$$

and, given $j(2k)$,

$$j(2k+1) = \min\{\ell \geqslant j(2k) \colon \beta_\ell \leqslant \tfrac{1}{2}\sigma\}, \tag{15}$$

$$j(2k+2) = \min\{\ell \geqslant j(2k+1) \colon \beta_\ell \geqslant \sigma\}. \tag{16}$$

Note that the existence of the subsequences $\{\beta_{i(k)}\}$, $\{\beta_{m(k)}\}$ guarantees that $j(k)$ is well defined for all $k \geqslant 0$. Observe also that, by Eqs. (15) and (16),

$$\beta_\ell \geqslant \frac{1}{2}\sigma \quad \text{for} \quad j(2k) \leqslant \ell \leqslant j(2k+1) - 1. \tag{17}$$

Then, since $\sum_{k=0}^{\infty} \alpha_k \beta_k < \infty$, we have, in view of Eq. (17),

$$\infty > \sum_{k=0}^{\infty} \alpha_k \beta_k \geqslant \sum_{k=0}^{\infty} \sum_{\ell=j(2k)}^{j(2k+1)-1} \alpha_\ell \beta_\ell \geqslant \frac{\sigma}{2} \sum_{k=0}^{\infty} \sum_{\ell=j(2k)}^{j(2k+1)-1} \alpha_\ell. \tag{18}$$

Let $\psi_k = \sum_{\ell=j(2k)}^{j(2k+1)-1} \alpha_\ell$. It follows from Eq. (18) that $\sum_{k=0}^{\infty} \psi_k < \infty$, implying

$$\lim_{k\to\infty} \psi_k = 0. \tag{19}$$

On the other hand, by Eqs. (15) and (16), we have $\beta_{j(2k)} \geqslant \sigma$, $\beta_{j(2k+1)} \leqslant \frac{1}{2}\sigma$, so that for all k,

$$\frac{\sigma}{2} \leqslant \beta_{j(2k)} - \beta_{j(2k+1)} = \sum_{\ell=j(2k)}^{j(2k+1)-1} (\beta_\ell - \beta_{\ell+1}) \leqslant \sum_{\ell=j(2k)}^{j(2k+1)-1} \theta\alpha_\ell = \theta\psi_k \tag{20}$$

using the hypothesis of (ii) in the rightmost inequality of Eq. (20). By Eq. (20), $\psi_k \geqslant \sigma/(2\theta)$ for all k, in contradiction with Eq. (19). The contradiction arises from assuming that there exists a subsequence of $\{\beta_k\}$ which is bounded away from 0, and therefore $\lim_{k\to\infty} \beta_k = 0$. □

To finish with the preliminaries, we gather in the following proposition two well known facts on orthogonal projections, to be used in the sequel.

Proposition 3. (i) $\|P_C(y) - P_C(z)\| \leqslant \|y - z\|$ *for all* $y, z \in H$.
(ii) $\langle y - \bar{y}, y - P_C(y)\rangle \geqslant 0$ *for all* $y \in H$, $\bar{y} \in C$.

Proof. See [2, p. 121]. □

The following lemma contains the main ideas of our result. It is written in an indirect way (with a hypothesis on existence of some $\tilde{x}$) in order to cover both the cases of nonempty and of empty solution set. For $x \in C$, let $L(x) = \{y \in C\colon f(y) \leqslant f(x)\}$.

Lemma 1. *If the algorithm generates an infinite sequence and there exists* $\tilde{x} \in C$ *and* $\tilde{k} \geqslant 0$ *such that* $f(\tilde{x}) \leqslant f(x^k)$ *for all* $k \geqslant \tilde{k}$, *then:*
(i) $\{x^k\}$ *is quasi-Fejér convergent to* $L(\tilde{x})$,
(ii) $\{f(x^k)\}$ is a convergent sequence, and $\lim_{k\to\infty} f(x^k) = f(\tilde{x})$,
(iii) *the sequence* $\{x^k\}$ *is weakly convergent to some* $\bar{x} \in L(\tilde{x})$.

Proof. (i) Take any $x \in L(\tilde{x})$. Let $z^k = x^k - (\alpha_k/\eta_k)u^k$, $\beta_k = f(x^k) - f(x)$. It follows from Eq. (8) that $x^k \in C$ for all $k \geqslant 1$, so that $P_C(x^k) = x^k$ and therefore

$$\|x^{k+1} - x^k\| = \|P_C(z^k) - P_C(x^k)\| \leqslant \|z^k - x^k\| = \frac{\alpha_k}{\eta_k}\|u^k\| \leqslant \alpha_k \tag{21}$$

using Proposition 3(i) and $\|u^k\| \leqslant \eta_k$. We proceed to prove that $\{x^k\}$ is quasi-Fejér convergent to $L(\tilde{x})$. In the following chain of equalities and inequalities, where we

establish a summable upper bound of $\alpha_k\beta_k/\eta_k$, the equalities are trivial and the inequalities are justified immediately below. We have

$$\begin{aligned}
\alpha_k^2 + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 &\geqslant \|x^{k+1} - x^k\|^2 + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 \\
&= 2\langle x^k - x, x^k - x^{k+1}\rangle = 2\langle x^k - x, x^k - z^k\rangle + 2\langle x^k - x, z^k - x^{k+1}\rangle \\
&= 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle + 2\langle x^k - z^k, z^k - x^{k+1}\rangle + 2\langle z^k - x, z^k - x^{k+1}\rangle \\
&= 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle + 2\langle x^k - z^k, z^k - x^{k+1}\rangle + 2\langle z^k - x, z^k - P_C(z^k)\rangle \\
&\geqslant 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle + 2\langle x^k - z^k, z^k - x^{k+1}\rangle \\
&= 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle + 2\langle x^k - z^k, z^k - x^k\rangle + 2\langle x^k - z^k, x^k - x^{k+1}\rangle \\
&\geqslant 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle - 2\|x^k - z^k\|^2 - 2\|x^k - z^k\|\|x^k - x^{k+1}\| \\
&\geqslant 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle - 2\frac{\alpha_k^2}{\eta_k^2}\|u^k\|^2 - 2\frac{\alpha_k^2}{\eta_k}\|u^k\| \\
&\geqslant 2\frac{\alpha_k}{\eta_k}\langle u^k, x^k - x\rangle - 4\alpha_k^2 \geqslant 2\frac{\alpha_k}{\eta_k}[f(x^k) - f(x) - \varepsilon_k] - 4\alpha_k^2 \\
&= 2\frac{\alpha_k}{\eta_k}\beta_k - 2\frac{\alpha_k}{\eta_k}\varepsilon_k - 4\alpha_k^2 \\
&\geqslant 2\frac{\alpha_k}{\eta_k}\beta_k - 2\alpha_k\varepsilon_k - 4\alpha_k^2 \geqslant 2\frac{\alpha_k}{\eta_k}\beta_k - (2\mu + 4)\alpha_k^2
\end{aligned} \tag{22}$$

using Eq. (21) in the first inequality, Proposition 3(ii) in the second one, Cauchy–Schwartz inequality in the third one, Eq. (21) again in the fourth one, $\|u^k\| \leqslant \eta_k$ in the fifth one, definition of $\partial_{\varepsilon_k} f(x^k)$ in the sixth one, $\eta_k \geqslant 1$ in the seventh one and Eq. (6) in the eighth one.

Since $x \in L(\tilde{x})$ we have, for $k \geqslant \tilde{k}$, $f(x) \leqslant f(\tilde{x}) \leqslant f(x^k)$, so that $\beta_k = f(x^k) - f(x) \geqslant 0$. Therefore we get from Eq. (22),

$$0 \leqslant 2\frac{\alpha_k}{\eta_k}\beta_k \leqslant \|x^k - x\|^2 - \|x^{k+1} - x\|^2 + (2\mu + 5)\alpha_k^2 \tag{23}$$

for $k \geqslant \tilde{k}$. Let $\delta_k = (2\mu + 5)\alpha_k^2$, $\gamma = \sum_{k=0}^{\infty}\alpha_k^2$. By Eq. (5), $\gamma < \infty$ so that $\sum_{k=0}^{\infty}\delta_k = (2\mu + 5)\gamma < \infty$. Since $\|x^{k+1} - x\|^2 \leqslant \|x^k - x\|^2 + \delta_k$ for $k \geqslant \tilde{k}$ by Eq. (23) and x is an arbitrary element of $L(\tilde{x})$, we conclude that $\{x^k\}$ is quasi-Fejér convergent to $L(\tilde{x})$, and therefore (i) holds.

(ii) $\{x^k\}$ is bounded by (i) and Proposition 1(i). Let B be a bounded set containing $\{x^k\}$ and $\bar{\varepsilon} = \sup\{\varepsilon_k\}$. Then

$$u^k \in \partial_{\varepsilon_k} f(x^k) \subset \partial_{\bar{\varepsilon}} f(x^k) \subset \bigcup_{y \in B}\partial_{\bar{\varepsilon}} f(y). \tag{24}$$

By Eq. (24) and assumption (A), $\{u^k\}$ is bounded, so that there exists $\rho > 1$ such that $\|u^k\| \leqslant \rho$ for all k. Therefore $\eta_k = \max\{1, \|u^k\|\} \leqslant \max\{1, \rho\} = \rho$. By Eq. (23)

$$0 \leqslant \frac{2}{\rho}\alpha_k\beta_k \leqslant \|x^k - x\|^2 - \|x^{k+1} - x\|^2 + (2\mu + 5)\alpha_k^2 \tag{25}$$

for $k \geqslant \tilde{k}$ and $x \in L(\tilde{x})$. Summing up Eq. (25) from $k = \tilde{k}$ to n

$$\frac{2}{\rho}\sum_{k=\tilde{k}}^{n} \alpha_k \beta_k \leqslant \|x^0 - x\|^2 - \|x^{n+1} - x\|^2 + (2\mu + 5)\sum_{k=\tilde{k}}^{n} \alpha_k^2$$
$$\leqslant \|x^0 - x\|^2 + (2\mu + 5)\gamma. \tag{26}$$

By Eq. (26), $\sum_{k=\tilde{k}}^{\infty} \alpha_k \beta_k \leqslant (\frac{1}{2}\rho)(\|x^0 - x\|^2 + (2\mu + 5)\gamma) < \infty$, implying

$$\sum_{k=0}^{\infty} \alpha_k \beta_k < \infty. \tag{27}$$

Up to now, x is any element of $L(\tilde{x})$. Take now $x = \tilde{x}$ so that $\beta_k = f(x^k) - f(\tilde{x})$. Observe that

$$\beta_{k+1} - \beta_k = f(x^{k+1}) - f(x^k) \leqslant \langle u^{k+1}, x^{k+1} - x^k \rangle + \varepsilon_{k+1}$$
$$\leqslant \|u^{k+1}\| \|x^{k+1} - x^k\| + \varepsilon_k \leqslant (\mu + \rho)\alpha_k \tag{28}$$

using definition of $\partial_{\varepsilon_{k+1}} f(x^{k+1})$ in the first inequality, $\varepsilon_{k+1} \leqslant \varepsilon_k$ and Cauchy–Schwartz inequality in the second one and Eq. (6) together with Eq. (21) in the third one. Let $\theta = \mu + \rho$. Since $\beta_k \geqslant 0$ for $k \geqslant \tilde{k}$, we are, in view of (27) and (28), within the hypotheses of Proposition 2, and we can conclude that $\lim_{k\to\infty} \beta_k = 0$, i.e. that $\lim_{k\to\infty} f(x^k) = f(\tilde{x})$.

(iii) Let $\bar{x}$ be a weak accumulation point of $\{x^k\}$, which exists by (i) and Proposition 1(i). If $\{x^{j_k}\}$ is a subsequence of $\{x^k\}$ whose weak limit is $\bar{x}$, then we have, since convex functions are weakly lower semicontinuous,

$$f(\bar{x}) \leqslant \liminf_{k\to\infty} f(x^{j_k}) = \lim_{k\to\infty} f(x^k) = f(\tilde{x}). \tag{29}$$

It follows from Eq. (29) that $\bar{x} \in L(\tilde{x})$, noting that $\bar{x} \in C$ because C is closed and convex, henceforth weakly closed. We have proved that all weak accumulation points of $\{x^k\}$ belong to $L(\tilde{x})$. By (i) and Proposition 1(iii), we conclude that there exists only one accumulation point, i.e. that $\{x^k\}$ is weakly convergent to some $\bar{x} \in L(\tilde{x})$. □

Finally, we state and prove our main convergence result.

Theorem 1. (i) *If Algorithm Eqs. (7) and (8) generates an infinite sequence then* $\liminf_{k\to\infty} f(x^k) = \inf_{x\in C} f(x)$.

(ii) *If the set S^* of solutions of Problem Eqs. (1) and (2) is nonempty then either Algorithm Eqs. (7) and (8) stops at some iteration k, in which case $x^k \in S^*$, or it generates an infinite sequence which converges weakly to some $\bar{x} \in S^*$.*

(iii) *If S^* is empty then $\{x^k\}$ is unbounded.*

Proof. (i) Let $f^* = \inf_{x\in C} f(x)$ (possibly $f^* = -\infty$). Since $x^k \in C$ for all $k \geqslant 1$, we have $\liminf_{k\to\infty} f(x^k) \geqslant f^*$. Assume $\liminf_{k\to\infty} f(x^k) > f^*$. Then there exists $\tilde{x}$ such that

$$\liminf_{k\to\infty} f(x^k) > f(\tilde{x}). \tag{30}$$

It follows from Eq. (30) that there exists $\tilde{k}$ such that $f(x^k) \geqslant f(\tilde{x})$ for all $k \geqslant \tilde{k}$. By Lemma 1(ii) $\lim_{k\to\infty} f(x^k) = f(\tilde{x})$, in contradiction with Eq. (30). The result follows.

(ii) Since $S^* \neq \emptyset$, take any $x^* \in S^*$, in which case $L(x^*) = S^*$. By optimality of x^*, $f(x^k) \geqslant f(x^*)$ for all k. Apply Lemma 1(iii) with $\tilde{x} = x^*$, $\tilde{k} = 0$, and conclude that $\{x^k\}$ converges weakly to some $\bar{x} \in S^*$.

(iii) Assume that S^* is empty but $\{x^k\}$ is bounded. Let $\{x^{j_k}\}$ be a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} f(x^{j_k}) = \liminf_{k\to\infty} f(x^k)$. Since $\{x^{j_k}\}$ is bounded, without loss of generality (i.e. refining $\{x^{j_k}\}$ if necessary), we may assume that $\{x^{j_k}\}$ converges weakly to some $\bar{x} \in C$. By weak lower semicontinuity of f,

$$f(\bar{x}) \leqslant \liminf_{k\to\infty} f(x^{j_k}) = \lim_{k\to\infty} f(x^{j_k}) = \liminf_{k\to\infty} f(x^k) = f^* \tag{31}$$

using (i) in the equality. By Eq. (31), $\bar{x}$ belongs to S^*, in contradiction with the hypothesis. It follows that $\{x^k\}$ is bounded. □

We make a few comments on the results of Theorem 1. To our knowledge, this is the first proof of convergence of the whole sequence generated by Eqs. (7) and (8) to a unique weak limit, without assuming finite dimensionality (as in [1]) or uniform monotonicity of ∂f (as in [9]). Additionally, our analysis includes the feature of approximate subgradients. The result of Theorem 1(i), on the other hand, is similar to the result in [1], except for the inclusion of inexact subgradients and constrained problems, which are not considered in [1]. We remark also that Eq. (5) (i.e. summability of α_k^2) is needed only to establish quasi-Féjer convergence of $\{x^k\}$ to S^* in the case of nonempty S^*. It is easy to check that if $\{x^k\}$ is known to be bounded beforehand (e.g. when C is bounded) then our results hold also with $\lim_{k\to\infty} \alpha_k = 0$ instead of Eq. (5).

Finally, we present a convergence rate result for the sequence of functional values $\{f(x^k)\}$.

Theorem 2. *If problem (1) has solutions and the sequence $\{x^k\}$ generated by Eqs. (7) and (8) is infinite, then there exists a subsequence $\{x^{\ell_k}\}$ of $\{x^k\}$ such that $f(x^{\ell_k}) - f(x^*) \leqslant (\sum_{j=0}^{\ell_k} \alpha_j)^{-1}$, where x^* is any solution of (1) and (2).*

Proof. We look at the proof of Lemma 1 with $\tilde{x} = x^* \in S^*$, $\tilde{k} = 0$, $\beta_k = f(x^k) - f(x^*)$. By Eq. (27), $0 \leqslant \sum_{k=0}^{\infty} \alpha_k \beta_k < \infty$. Let $s_k = \sum_{j=0}^{k} \alpha_j$, $N_1 = \{k\colon \beta_k \leqslant s_k^{-1}\}$, $N_2 = \{k\colon \beta_k > s_k^{-1}\}$. Suppose that N_1 is finite, then there exists $\bar{k}$ such that $x^k \in N_2$ for all $k \geqslant \bar{k}$, so that

$$\sum_{k=\bar{k}}^{\infty} \frac{\alpha_k}{s_k} \leqslant \sum_{k=\bar{k}}^{\infty} \alpha_k \beta_k \leqslant \sum_{k=0}^{\infty} \alpha_k \beta_k < \infty.$$

It follows that $\sum_{k=0}^{\infty} \alpha_k / s_k < \infty$. On the other hand, Abel–Dini's criterion for divergent series [see 20, §39] states that if $\sum_{n=0}^{\infty} \kappa_n = \infty$ then $\sum_{n=0}^{\infty} [\kappa_n / (\sum_{j=0}^{n} \kappa_j)] = \infty$. So we conclude, in view of (4), that $\sum_{k=0}^{\infty} \alpha_k / s_k = \infty$. This contradiction implies

that N_1 is infinite, and we can take $\{x^{\ell_k}\}$ as consisting precisely of those x^k with $k \in N_1$. □

This result does not give any information on the asymptotic behavior of $\{f(x^k)\}$ outside the subsequence $\{x^{\ell_k}\}$. If we assume that f is Gateaux differentiable, that its gradient is uniformly continuous and that $\varepsilon_k = 0$ for all k (i.e. $u^k = \nabla f(x^k)$ in Eq. (8)), then we can get results on the asymptotic behavior of the whole sequence $\{f(x^k)\}$. More precisely, if $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous and nondecreasing function such that $\varphi(0) = 0$ and $\|\nabla f(x) - \nabla f(y)\| \leqslant \varphi(\|x - y\|)$ for all $x, y \in H$, then we get, in addition to the result of Theorem 2, that $\beta_{\ell_k+1} \leqslant \beta_{\ell_k} + \alpha_{\ell_k}\varphi(\alpha_{\ell_k})$ and $\beta_i \leqslant \beta_{\ell_k+1}$ for all i such that $\ell_k + 1 \leqslant i < \ell_{k+1}$. The proof is rather involved and we will not develop it in this paper.

For the finite dimensional case, a sharper and nonasymptotic convergence rate result can be found in [12, Theorem 3.2.2].

Acknowledgements

The first author thanks the Institute for Pure and Applied Mathematics (IMPA), at Rio de Janeiro, Brazil, where he was a visiting professor while this paper was written.

References

[1] R. Correa, C. Lemaréchal, Convergence of some algorithms for convex minimization, Mathematical Programming 62 (1993) 261–275.
[2] B.T. Polyak, Introduction to Optimization, Optimization Software, New York, 1987.
[3] J.R. Giles, Convex analysis with applications in differentiation of convex functions, in: Research Notes in Mathematics, vol. 58, Pitman, Boston, MA, 1982.
[4] R.T. Rockafellar, Local boundedness of nonlinear monotone operators, Michigan Mathematical Journal 16 (1969) 397–407.
[5] A. Brønsted, R.T. Rockafellar, On the subdifferentiability of convex functions, Proceedings of the American Mathematical Society 16 (1965) 605–611.
[6] B.T. Polyak, A general method of solving extremum problems, Soviet Mathematics Doklady 8 (1967) 593–597.
[7] Yu.M. Ermoliev, Methods for solving nonlinear extremal problems, Cybernetics 2 (1966) 1–17.
[8] Ya.I. Alber, On minimization of smooth functional by gradient methods, USSR Computational Mathematics and Mathematical Physics 11 (1971) 752–758.
[9] Ya.I. Alber, Recurrence relations and variational inequalities, Soviet Mathematics Doklady 27 (1983) 511–517.
[10] M.V. Solodov, S.K. Zavriev, Error stability properties of generalized gradient-type algorithms Journal of Optimization Theory and Applications 98 (1998), to appear.
[11] E. Golstein, N. Tretyakov, Modified Lagrangian Functions, Nauka, Moscow, 1989.
[12] Yu. Nesterov, Effective Methods in Nonlinear Programming, Nauka, Moscow, 1989.
[13] M. Minoux, Mathematical Programming, Theory and Algorithms, Wiley, New York, 1986.
[14] V.A. Bereznyev, V.G. Karmanov, A.A. Tretyakov, The stabilizing properties of the gradient method, USSR Computational Mathematics and Mathematical Physics 26 (1986) 84–85.

[15] R. Burachik, L.M. Graña Drummond, A.N. Iusem, B.F. Svaiter, Full convergence of the steepest descent method with inexact line searches, Optimization 32 (1995) 137–146.
[16] K. Kiwiel, K. Murty, Convergence of the steepest descent method for minimizing quasi convex functions, Journal of Optimization Theory and Applications 89 (1996) 221–226.
[17] B.F. Svaiter, Steepest descent method in Hilbert spaces with Armijo search (to be published).
[18] Yu.M. Ermoliev, On the method of generalized stochastic gradients and quasi-Fejér sequences, Cybernetics 5 (1969) 208–220.
[19] A.N. Iusem, B.F. Svaiter, M. Teboulle, Entropy-like proximal methods in convex programming, Mathematics of Operations Research 19 (1994) 790–814.
[20] K. Knopp, Theory and Application of Infinite Series, Dover, New York, 1990.
[21] Ya.I. Alber, A.N. Iusem, M.V. Solodov, Minimization of nonsmooth convex functionals in Banach spaces, Journal of Convex Analysis; to appear.