Computational and Applied Mathematics
Vol. 22, N. 1, pp. 37–52, 2003

On the convergence properties of the projected gradient method for convex optimization

A. N. IUSEM*

Instituto de Matemática Pura e Aplicada (IMPA)
Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, RJ, Brazil
E-mail: iusp@impa.br

Abstract. When applied to an unconstrained minimization problem with a convex objective, the steepest descent method has stronger convergence properties than in the noncovex case: the whole sequence converges to an optimal solution under the only hypothesis of existence of minimizers (i.e. without assuming e.g. boundedness of the level sets). In this paper we look at the projected gradient method for constrained convex minimization. Convergence of the whole sequence to a minimizer assuming only existence of solutions has also been already established for the variant in which the stepsizes are exogenously given and square summable. In this paper, we prove the result for the more standard (and also more efficient) variant, namely the one in which the stepsizes are determined through an Armijo search.

Mathematical subject classification: 90C25, 90C30.

Key words: projected gradient method, convex optimization, quasi-Fejér convergence.

1 Introduction

1.1 The problem

We are concerned in this paper with the following smooth optimization problem:

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

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

#541/02. Received: 3/IV/02.

*The work of this author was partially supported by CNPq grant no. 301280/86.

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $C \subset \mathbb{R}^n$ is closed and convex. Each iteration of the projected gradient method, which we describe formally in subsection 1.3, basically consists of two stages: starting from the k-th iterate $x^k \in \mathbb{R}^n$, first a step is taken in the direction of $-\nabla f(x^k)$, and then the resulting point is projected onto C, possibly with additional one-dimensional searches in either one of the stages. The classical convergence result establishes that cluster points of $\{x^k\}$ (if any) are stationary points for (1)–(2), i.e. they satisfy the first order optimality conditions, but in general neither existence nor uniqueness of cluster points is guaranteed. In this paper we prove a much stronger result for the case in which f is convex, namely that the whole sequence $\{x^k\}$ converges to a solution of (1)–(2) under the only assumption of existence of solutions. An analogous result is known to hold for the steepest descent method for unconstrained optimization, which we describe in the next subsection.

1.2 The steepest descent method

Given a continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$, the *steepest descent* method generates a sequence $\{x^k\} \in \mathbb{R}^n$ through

$$x^{k+1} = x^k - \beta_k \nabla f(x^k), \tag{3}$$

where β_k is some positive real number. Several choices are available for β_k. The first one is to set the β_k's exogenously, and a relevant option is

$$\beta_k = \frac{\alpha_k}{\left\|\nabla f(x^k)\right\|}, \tag{4}$$

with

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

Other options consist of performing an exact line minimization, i.e.

$$\beta_k = \operatorname{argmin}_{\beta > 0} f(x^k - \beta \nabla f(x^k)), \tag{6}$$

or an inexact linesearch, e.g. following an Armijo rule, namely

$$\beta_k = \bar{\beta} 2^{-\ell(k)} \tag{7}$$

with

$$\ell(k) = \min\Big\{ j \in \mathbb{Z}_{\geq 0} : f(x^k - \bar{\beta}2^{-j}\nabla f(x^k)) \\ \leq f(x^k) - \sigma\bar{\beta}2^{-j} \left\|\nabla f(x^k)\right\|^2 \Big\}, \tag{8}$$

for some $\bar{\beta} > 0$, $\sigma \in (0, 1)$.

The basic convergence result on this method (and in general on descent direction methods), under either exact linesearches or Armijo searches, derives from Zangwill's global convergence theorem (see [16]) and establishes that every cluster point $\bar{x}$ of $\{x^k\}$, if any, is stationary, i.e. such that $\nabla f(\bar{x}) = 0$. In order to ensure existence of cluster points, it is necessary to assume that the starting iterate x^0 belongs to a bounded level set of f (see [15] for this and other related results). The situation is considerably better when f is convex: it is possible to prove convergence of the whole sequence to a minimizer of f under the sole asumption of existence of minimizers (i.e. without any additional assumption on boundedness of level sets). Results of this kind for the convex case can be found in [8], [10] and [14] for the method with exogenously given β_k's satisfying (4)–(5), in [12] for the method with exact lineasearches as in (6), and in [2], [7] and [13] for the method with the Armijo rule (7)–(8). We observe that in the case of β_k's given by (4)–(5) the method is not in general a descent one, i.e. it is not guaranteed that $f(x^{k+1}) \leq f(x^k)$ for all k.

1.3 The projected gradient method

In this subsection we deal with problem (1)–(2). Convexity of C makes it posible to use the orthogonal projection onto $C, P_C : \mathbb{R}^n \to C$, for obtaining feasible directions which are also descent ones; namely a step is taken from x^k in the direction of $-\nabla f(x^k)$, the resulting vector is projected onto C, and the direction from x^k to this projection has the above mentioned properties. We remind that a point $z \in C$ is stationary for problem (1)–(2) iff $\nabla f(z)^t(x - z) \geq 0$ for all $x \in C$. A formal description of the algorithm, called the *projected gradient* method, is the following:

Initialization: Take $x^0 \in C$.

Iterative step: If x^k is stationary, then stop. Otherwise, let

$$z^k = P_C(x^k - \beta_k \nabla f(x^k)), \tag{9}$$

$$x^{k+1} = x^k + \gamma_k(z^k - x^k), \tag{10}$$

where β_k, γ_k are positive stepsizes, for which, again, several choices are possible. Before discussing them, we mention that in the unconstrained case, i.e. $C = \mathbb{R}^n$, then method given by (9)–(10) with $\gamma_k = 1$ for all k reduces to (3). Following [3], we will focus in three strategies for the stepsizes:

i) Armijo search along the feasible direction: $\{\beta_k\} \subset [\tilde{\beta}, \hat{\beta}]$ for some $0 < \tilde{\beta} \leq \hat{\beta}$ and γ_k determined with an Armijo rule, namely

$$\gamma_k = 2^{-\ell(k)} \tag{11}$$

with

$$\ell(k) = \min\left\{ j \in \mathbb{Z}_{\geq 0} : f(x^k - 2^{-j}(z^k - x^k)) \leq f(x^k) - \sigma 2^{-j} \nabla f(x^k)^t (x^k - z^k) \right\}, \tag{12}$$

for some $\sigma \in (0, 1)$.

ii) Armijo search along the boundary of C: $\gamma_k = 1$ for all k and β_k determined through (7) and the following two equations instead of (8):

$$\ell(k) = \min\left\{ j \in \mathbb{Z}_{\geq 0} : f(z^{k,j}) \leq f(x^k) - \sigma \nabla f(x^k)^t (x^k - z^{k,j}) \right\}, \tag{13}$$

with

$$z^{k,j} = P_C(x^k - \bar{\beta} 2^{-j} \nabla f(x^k)). \tag{14}$$

iii) Exogenous stepsize before projecting: β_k given by (4)–(5) and $\gamma_k = 1$ for all k.

Several comments are in order. First, observe that in the unconstrained case ($C = \mathbb{R}^n$) options (i) and (ii) reduce to the steepest descent method (3) with the Armijo rule given by (7)–(8), while option (iii) reduces to (3) with the β_k's given by (4)–(5). Secondly, note that option (ii) requires a projection onto C for each step of the inner loop resulting from the Armijo search, i.e. possibly many projections for each k, while option (i) demands only one projection for each outer step, i.e. for each k. Thus, option (ii) is competitive only when P_C is very easy to compute (e.g. when C is a box or a ball). Third, we mention that option (iii), as its counterpart in the unconstrained case, fails to be a descent method. Finally, it it easy to show that for option (iii) it holds that $\left\| x^{k+1} - x^k \right\| \leq \alpha_k$ for all k, with α_k as in (4). In view of (5), this means that all stepsizes are "small", while options (i) and (ii) allow for occasionally long steps. Thus option (iii) seems rather undesirable. Its redeeming feature is that its good convergence properties hold also in the nonsmooth case, when $\nabla f(x^k)$ is replaced by a subgradient ξ^k of f at x^k. Subgradients do not give raise to descent directions, so that Armijo searches are not ensured to succeed, and therefore exogenous stepsizes seem to be the only available alternative. This is the case analyzed in [1]. We will not be concerned with option (iii) in the sequel.

Without assuming convexity of f, the convergence results for these methods closely mirror the ones for the steepest descent method in the unconstrained case: cluster points may fail to exist, even when (1)–(2) has solutions, but if they exist, they are stationary and feasible, i.e. $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$, $\bar{x} \in C$ for all cluster point $\bar{x}$ of $\{x^k\}$ and all $x \in C$. These results can be found in Section 2.3.2 of [3]; for the case of option (ii), they are based upon the results in [9].

When f is convex, the stronger results for the unconstrained case, with β_k's given by (4)–(5), have also been extended to the projected gradient method under option (iii): it has been proved in [1] that in such a case, the whole sequence $\{x^k\}$ converges to a solution of problem (1)–(2) under the sole assumption of existence of solutions. On the other hand, the current situation is rather worse for options (i) and (ii): as far as we know, neither existence nor uniqueness of cluster points for options (i) and (ii) has been proved, assuming only convexity of f. We will prove both for option (i) in the following two sections. The corresponding results for the less interesting option (ii) have very similar proofs, and we sketch them in Section 4.

Of course, results of this kind are immediate under stronger hypotheses on the problem: cluster points of $\{x^k\}$ certainly exist if the intersection of C with some level set of f is nonempty and bounded, and strict convexity of f ensures uniqueness of the cluster point.

2 Preliminaries

This section contains some previously established results needed in our analysis. We prove them in order to make the paper closer to being self-contained. We start with the so called quasi-Fejér convergence theorem (see [7], Theorem 1).

Proposition 1. *Let $T \subset \mathbb{R}^n$ be a nonempty set and $\{a^k\} \subset \mathbb{R}^n$ a sequence such that*

$$\left\|a^{k+1} - z\right\|^2 \leq \left\|a^k - z\right\|^2 + \epsilon_k \tag{15}$$

for all $z \in T$ and all k, where $\{\epsilon_k\} \subset \mathbb{R}_+$ is a summable sequence. Then

i) $\{a^k\}$ is bounded.

ii) If a cluster point $\bar{a}$ of $\{a^k\}$ belongs to T, then the whole sequence $\{a^k\}$ converges to $\bar{a}$.

Proof.

i) Fix some $z \in T$. Applying iteratively (15) we get

$$\left\|a^k - z\right\|^2 \leq \left\|a^0 - z\right\|^2 + \sum_{j=0}^{k-1} \epsilon_j \leq \left\|a^0 - z\right\|^2 + \sum_{j=0}^{\infty} \epsilon_j .$$

Since $\{\epsilon_k\}$ is summable, it follows that $\{a^k\}$ is bounded.

ii) Let now $\bar{a} \in T$ be a cluster point of $\{a^k\}$ and take $\delta > 0$. Let $\{a^{\ell_k}\}$ be a subsequence of $\{a^k\}$ convergent to $\bar{a}$. Since $\{\epsilon_k\}$ is summable, there exists

k_0 such that $\sum_{j=k_0}^{\infty} \epsilon_j < \delta/2$, and there exists k_1 such that $\ell_{k_1} \geq k_0$ and $\left\| a^{\ell_k} - \bar{a} \right\|^2 < \frac{\delta}{2}$ for any $k \geq k_1$. Then, for any $k > \ell_{k_1}$ we have:

$$\left\| a^k - \bar{a} \right\|^2 \leq \left\| a^{\ell_{k_1}} - \bar{a} \right\|^2 + \sum_{j=\ell_{k_1}}^{k-1} \epsilon_j \leq \left\| a^{\ell_{k_1}} - \bar{a} \right\|^2 + \sum_{j=\ell_{k_1}}^{\infty} \epsilon_j < \delta/2 + \delta/2 = \delta.$$

We conclude that $\lim_{k\to\infty} a^k = \bar{a}$. □

Next we show that the linesearch for option (i) is always successful. We start with an immediate fact on descent directions.

Proposition 2. *Take $\sigma \in (0,1)$, $x \in C$ and $v \in \mathbb{R}^n$ such that $\nabla f(x)^t v < 0$. Then there exists $\bar{\gamma} < 1$ such that $f(x + \gamma v) < f(x) + \sigma\gamma\nabla f(x)^t v$ for all $\gamma \in (0, \bar{\gamma}]$.*

Proof. The result follows from the differentiability of f. □

Next we prove that the direction $z^k - x^k$ in option (i) is a descent one.

Proposition 3. *Take x^k and z^k as defined by* (9)–(12). *Then*

i) x^k belongs to C for all k.

ii) If $\nabla f(x^k) \neq 0$, then $\nabla f(x^k)^t (z^k - x^k) < 0$.

Proof.

i) By induction. It holds for $k = 0$ by the initialization step. Assume that $x^k \in C$. By (9), $z^k \in C$. By (12), $\gamma_k \in [0,1]$. By (10), $x^{k+1} \in C$.

ii) A well known elementary property of orthogonal projections states that $\langle v - u, P_C(u) - u \rangle \geq 0$ for all $u \in \mathbb{R}^n$, $v \in C$. By (i), $x^k \in C$. Thus, in view of (9),

$$\begin{aligned} 0 &\leq \left\langle x^k - \beta_k \nabla f(x^k) - x^k, P_C(x^k - \beta_k \nabla f(x^k)) - x^k + \beta_k \nabla f(x^k) \right\rangle \\ &= -\beta_k \langle \nabla f(x^k), z^k - x^k \rangle - \beta_k^2 \left\| \nabla f(x^k) \right\|^2 . \end{aligned} \tag{16}$$

Since $\beta_k > 0$ and $\nabla f(x^k) \neq 0$, it follows from (16) that $\nabla f(x^k)^t (z^k - x^k) \leq -\beta_k \left\|\nabla f(x^k)\right\|^2 < 0$. □

Corollary 1. *If* $\nabla f(x^k) \neq 0$, *then* γ_k *is well defined for the algorithm* (9)–(12).

Proof. Consider Proposition 2 with $x = x^k$, $v = z^k - x^k$. By Proposition 3(ii), $\nabla f(x)^t v < 0$. Thus the assumption of Proposition 2 holds, and the announced $\bar{\gamma}$ exists, so that the inequality in (12) holds for all j such that $2^{-j} \leq \bar{\gamma}$. It follows that both $\ell(k)$ and γ_k are well defined. □

Finally, we prove stationarity of the cluster points of $\{x^k\}$, if any.

Proposition 4. *Let* $\{x^k\}$ *be the sequence defined by* (9)–(12). *If* $\{x^k\}$ *is infinite,* $\bar{x}$ *is a cluster point of* $\{x^k\}$ *and Problem* (1)–(2) *has solutions, then* $\bar{x}$ *is stationary for Problem* (1)–(2).

Proof. Since C is closed, $\bar{x}$ belongs to C by Proposition 3(i). Let $\{x^{j_k}\}$ be a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{j_k} = \bar{x}$. Observe that $\{\gamma_k\} \subset [0, 1]$ by (11) and Corollary 1, and that $\{\beta_k\} \subset [\tilde{\beta}, \hat{\beta}]$. Thus, we may assume without loss of generality that $\lim_{k\to\infty} \gamma_{j_k} = \widehat{\gamma} \in [0, 1]$, $\lim_{k\to\infty} \beta_{j_k} = \bar{\beta} > \tilde{\beta} > 0$. By (10), (11) and (12),

$$0 < -\sigma \gamma_k \nabla f(x^k)^t (z^k - x^k) \leq f(x^k) - f(x^{k+1}). \tag{17}$$

It follows from (17) that $\{f(x^k)\}$ is a decreasing sequence. Since $\{x^k\} \subset C$ by Proposition 3(i) and Problem (1)–(2) has solutions, $\{f(x^k)\}$ is bounded below, hence convergent, so that $\lim_{k\to\infty}[f(x^k) - f(x^{k+1})] = 0$. Taking limits in (17) along the subsequence, and taking into account (9), we get

$$0 \leq -\sigma \widehat{\gamma} \nabla f(\bar{x})^t [P_C(\bar{x} - \bar{\beta} \nabla f(\bar{x})) - \bar{x}] \leq 0, \tag{18}$$

using also continuity of both ∇f and P_C. Now we consider two cases. Suppose first that $\widehat{\gamma} > 0$. Let $\bar{u} = \bar{x} - \bar{\beta}\nabla f(\bar{x})$. Then, it follows from (18) that

$$0 = \nabla f(\bar{x})^t [P_C(\bar{u}) - \bar{x}] = \bar{\beta}^{-1}(\bar{x} - \bar{u})^t [P_C(\bar{u}) - \bar{x}], \tag{19}$$

implying that $0 = (\bar{u} - \bar{x})^t [P_C(\bar{u}) - \bar{x}]$. Since $\bar{x}$ belongs to C, an elementary property of orthogonal projections implies that $\bar{x} = P_C(\bar{u}) = P_C(\bar{x} - \bar{\beta}\nabla f(\bar{x}))$, and, taking into account that $\bar{\beta} > 0$, it follows easily that $\nabla f(\bar{x})^t (x - \bar{x}) \geq 0$ for all $x \in C$, i.e. $\bar{x}$ is stationary for Problem (1)–(2).

Consider finally the case of $0 = \widehat{\gamma} = \lim_{k\to\infty} \gamma_{j_k}$. Fix $q \in \mathbb{N}$. Since $\gamma_{j_k} = 2^{-\ell(j_k)}$, there exists k such that $\ell(j_k) > q$, so that, in view of (12),

$$f\left(x^{j_k} - 2^{-q}(z^{j_k} - x^{j_k})\right) > f(x^{j_k}) - \sigma 2^{-q} \nabla f(x^{j_k})^t (x^{j_k} - z^{j_k}). \tag{20}$$

Taking limits in (20) with $k \to \infty$, and defining $\bar{z} = P_C(\bar{x} - \bar{\beta}\nabla f(\bar{x}))$, we get, for an arbitrary $q \in \mathbb{N}$,

$$f(\bar{x} - 2^{-q}(\bar{z} - \bar{x})) \geq f(\bar{x}) + \sigma 2^{-q} \nabla f(\bar{x})^t (\bar{z} - \bar{x}). \tag{21}$$

Combining (21) with Proposition 2, we conclude that $\nabla f(\bar{x})^t (\bar{z} - \bar{x}) \geq 0$. Using now Proposition 3(ii), we get that $0 = \nabla f(\bar{x})^t (\bar{z} - \bar{x}) = \nabla f(\bar{x})^t (P_C(\bar{u}) - \bar{x})$, i.e. (19) holds also in this case, and the conclusion is obtained with the same argument as in the previous case. □

Two comment are in order. First, no result proved up to this point requires convexity of f. Second, all these results are rather standard and well known (see e.g. Section 2.3.2 in [3] or [16]) The novelty of this paper occurs in the following sections.

3 Convergence properties in the convex case

In this section we prove that when f is convex then the sequence generated by variant (i) of the projected gradient method (i.e. (9)–(12)) converges to a solution of Problem (1)–(2), under the only assumption of existence of solutions.

Theorem 1. *Assume that Problem* (1)–(2) *has solutions. Then, either the algorithm given by* (9)–(12) *stops at some iteration k, in which case x^k is a solution of Problem* (1)–(2), *or it generates an infinite sequence $\{x^k\}$, which converges to a solution x^* of the problem.*

Proof. In the case of finite stopping, the stopping rule states that x^k is stationary. Since f is convex, stationary points are solutions of Problem (1)–(2). We assume in the sequel that the algorithm generates an infinite sequence $\{x^k\}$.

Let $\widehat{x}$ be any solution of Problem (1)–(2). Using (10) and elementary algebra:

$$\begin{aligned} &\left\|x^{k+1}-x^k\right\|^2+\left\|x^k-\widehat{x}\right\|^2-\left\|x^{k+1}-\widehat{x}\right\|^2 \\ &= 2\langle x^k-x^{k+1}, x^k-\widehat{x}\rangle = 2\gamma_k\langle z^k-x^k, \widehat{x}-x^k\rangle. \end{aligned} \tag{22}$$

The already used elementary property of orthogonal projections can be restated as $\langle P_C(u)-u, v-P_C(u)\rangle \geq 0$ for all $u \in \mathbb{R}^n$ and all $v \in C$. In view of (9)

$$\begin{aligned} &0 \leq \langle z^k-x^k+\beta_k\nabla f(x^k), \widehat{x}-z^k\rangle = \\ &\langle z^k-x^k+\beta_k\nabla f(x^k), \widehat{x}-x^k\rangle + \langle z^k-x^k+\beta_k\nabla f(x^k), x^k-z^k\rangle. \end{aligned} \tag{23}$$

By (23),

$$\begin{aligned} \langle z^k-x^k, \widehat{x}-x^k\rangle &\geq \beta_k\langle\nabla f(x^k), x^k-\widehat{x}\rangle - \langle z^k-x^k+\beta_k\nabla f(x^k), x^k-z^k\rangle \\ &\geq \beta_k[f(x^k)-f(\widehat{x})] + \langle z^k-x^k+\beta_k\nabla f(x^k), z^k-x^k\rangle \\ &\geq \langle z^k-x^k+\beta_k\nabla f(x^k), z^k-x^k\rangle = \left\|z^k-x^k\right\|^2 + \beta_k\langle\nabla f(x^k), z^k-x^k\rangle, \end{aligned} \tag{24}$$

using the gradient inequality for the convex function f in the second inequality, and feasibility of x^k, resulting from Proposition 3(i), together with optimality of $\widehat{x}$ and positivity of β_k, in the third one. Combining now (22) and (24), and taking into account (10),

$$\begin{aligned} &\left\|x^{k+1}-x^k\right\|^2+\left\|x^k-\widehat{x}\right\|^2-\left\|x^{k+1}-\widehat{x}\right\|^2 \\ &\geq 2\gamma_k\left[\left\|z^k-x^k\right\|^2+\beta_k\langle\nabla f(x^k), z^k-x^k\rangle\right] \\ &= 2\gamma_k^{-1}\left\|x^{k+1}-x^k\right\|^2+2\gamma_k\beta_k\langle\nabla f(x^k), z^k-x^k\rangle. \end{aligned} \tag{25}$$

After rearrangement, we obtain from (25),

$$\begin{aligned} &\left\|x^{k+1}-\widehat{x}\right\|^2 \leq \left\|x^k-\widehat{x}\right\|^2+(1-2\gamma_k^{-1})\left\|x^{k+1}-x^k\right\|^2 \\ &-2\gamma_k\beta_k\langle\nabla f(x^k), z^k-x^k\rangle \leq \left\|x^k-\widehat{x}\right\|^2-2\gamma_k\beta_k\langle\nabla f(x^k), z^k-x^k\rangle, \end{aligned} \tag{26}$$

using the fact that γ_k belongs to $[0,1]$ in the second inequality. Now we look at the specific way in which the γ_k's are determined. By (11), (12), for all j,

$$-\sigma\gamma_j\nabla f(x^j)^t(z^j-x^j) \leq f(x^j)-f(x^{j+1}). \tag{27}$$

Multiplying (27) by $(2\beta_k)/\sigma$, and defining

$$\epsilon_j = -2\beta_j \gamma_j \nabla f(x^j)^t (z^j - x^j), \tag{28}$$

we get, since $\{f(x^j)\}$ is nonincreasing,

$$\epsilon_j \le \frac{2\beta_j}{\sigma}[f(x^j) - f(x^{j+1})] \le \frac{2\hat{\beta}}{\sigma}[f(x^j) - f(x^{j+1})] \tag{29}$$

Summing (29) with j between 0 and k,

$$\sum_{j=0}^{k} \epsilon_j \le \frac{2\hat{\beta}}{\sigma}[f(x^0) - f(x^{k+1})] \le \frac{2\hat{\beta}}{\sigma}[f(x^0) - f(\widehat{x})], \tag{30}$$

and it follows from (30) that $\sum_{j=0}^{\infty} \epsilon_j < \infty$. By (26) and (28),

$$\left\| x^{k+1} - \widehat{x} \right\|^2 \le \left\| x^k - \widehat{x} \right\|^2 + \epsilon_k. \tag{31}$$

Let S^* be the set of solutions of Problem (1)–(2). Since $\widehat{x}$ is an arbitrary element of S^* and the ϵ_k's are summable, (31) means that $\{x^k\}$ is quasi-Fejér convergent to S^*. Since S^* is nonempty by assumption, it follows from Proposition 1(i) that $\{x^k\}$ is bounded, and therefore it has cluster points. By Proposition 4 all such cluster points are stationary. By convexity of f, they are solutions of Problem (1)–(2), i.e. they belong to S^*. By Proposition 1(ii), the whole sequence $\{x^k\}$ converges to a solution of Problem (1)–(2). □

4 Option (ii): search along an arc

We sketch in this section the analysis corresponding to option (ii), which we restate next:

Initialization: Take $x^0 \in C$.

Iterative step: If x^k is stationary, then stop. Otherwise, take

$$x^{k+1} = P_C(x^k - \beta_k \nabla f(x^k)), \tag{32}$$

where β_k is given by

$$\beta_k = \hat{\beta} 2^{-\ell(k)}, \tag{33}$$

with

$$\ell(k) = \min\Big\{ j \in \mathbb{Z}_{\geq 0} : f(z^{k,j}) \leq f(x^k) \\ - \sigma \nabla f(x^k)^t (x^k - z^{k,j}) \Big\}, \tag{34}$$

and

$$z^{k,j} = P_C(x^k - \hat{\beta} 2^{-j} \nabla f(x^k)). \tag{35}$$

Results for this variant follow closely those for option (i), developed in the previous two sections. Without assuming convexity of f, the following results hold:

Proposition 5. *If* $\{x^k\}$ *is the sequence generated by* (32)–(35)*, then*

i) $\{x^k\} \subset C$.

ii) β_k *is well defined by* (32)–(35).

iii) If $\nabla f(x^k) \neq 0$*, then* $\langle \nabla f(x^k), P_C(x^k - \beta_k \nabla f(x^k)) - x^k \rangle < 0$.

iv) If Problem (1)–(2) *has solutions and* $\bar{x}$ *is a cluster point of* $\{x^k\}$*, then* $\bar{x}$ *is stationary for Problem* (1)–(2).

Proof. Item (i) follows immediately from (32); for the remaining items see Proposition 2.3.3 and Lemma 2.3.1 in [3]. □

For convex f, we have the following result.

Theorem 2. *Assume that Problem* (1)–(2) *has solutions. Then, either the algorithm given by* (31)–(35) *stops at some iteration* k*, in which case* x^k *is a solution of Problem* (1)–(2)*, or it generates an infinite sequence* $\{x^k\}$*, which converges to a solution* x^* *of the problem.*

Proof. The result for the case of finite termination follows from the stopping criterion and the convexity of f. For the case of an infinite sequence, we observe that the computations in the proof of Theorem 1 up to (26) do not use the specific form of the β_k's and the γ_k's, so that they hold for the sequence under consideration, where now $\gamma_k = 1$ for all k and β_k is given by (33)–(35). Thus, for all solution $\widehat{x}$ of Problem (1)–(2), we have

$$\left\|x^{k+1} - \widehat{x}\right\|^2 \leq \left\|x^k - \widehat{x}\right\|^2 + \varepsilon_k, \tag{36}$$

with

$$\varepsilon_k = 2\beta_k \nabla f(x^k)^t \left[x^k - P_C(x^k - \beta_k \nabla f(x^k))\right] \tag{37}$$

Note that $\varepsilon_k \geq 0$ for all k by Proposition 5(iii). We prove next that $\{\varepsilon_k\}$ is summable.

In view of (32)–(35) (particularly the criterion of the arc-search), we have

$$\sigma \nabla f(x^k)^t \left[x^k - P_C(x^k - \beta_k \nabla f(x^k))\right] \leq f(x^k) - f(x^{k+1}). \tag{38}$$

Combining (37) and (38), and using then (33),

$$\varepsilon_k \leq \frac{2\beta_k}{\sigma}[f(x^k) - f(x^{k+1})] \leq \frac{2\hat{\beta}}{\sigma}[f(x^k) - f(x^{k+1})]. \tag{39}$$

By (39), $\sum_{k=0}^{\infty} \varepsilon_k \leq \frac{2\hat{\beta}}{\sigma}[f(x^0) - f(\widehat{x})] < \infty$. In view of (36) it follows, as in Theorem 1, that $\{x^k\}$ is quasi-Fejér convergent to the solution set, and then Proposition 1(i) implies that $\{x^k\}$ is bounded, so that it has cluster points. By Proposition 5(iv) and convexity of f, all of them solve Problem (1)–(2). Finally, Proposition 1(ii) implies that the whole sequence $\{x^k\}$ converges to a solution. □

5 Final remarks

1. The purpose of this paper is theoretical, and it consists of determining the convergence properties of the projected gradient method in the case of a convex objective. We make no claims whatsoever on the advantages and/or drawbacks of this algorithm viz-a-viz others.

2. Despite the previous remark, we mention that some variants of the projected gradient methods have been proved to be quite successful from a computational point of view, particularly the *spectral projected gradient* method (SPG); see e.g.

[5], [6], [4]. In this method β_k is taken as a safeguarded spectral parameter, with the following meaning. Let

$$\eta_k = \frac{\left\|x^k - x^{k-1}\right\|^2}{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}.$$

We mention that when f is twice differentiable η_k^{-1} is the Rayleigh quotient asociated with the averaged Hessian matrix $\int_0^1 \nabla^2 f(tx^k + (1-t)x^{k-1})dt$, thus justifying the denomination of *spectral parameter* given to η_k. Then β_k is taken as follows: β_0 is exogenously given; for $k \geq 1$, $\beta_k = \beta_{k-1}$ if $\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle \leq 0$. Otherwise, β_k is taken as the median between $\tilde{\beta}$, η_k and $\hat{\beta}$ ($\tilde{\beta}$ and $\hat{\beta}$ act as the ''safeguards'' for the spectral parameter). Since our strategy (i) encompasses such choice of β_k, the result of Theorem 1 holds for this variant. On the other hand, SPG as presented in [5], [6], includes another feature, namely a backtracking procedure generating a possibly nonmonotone sequence of functional values $\{f(x^k)\}$. In fact, (12) is replaced by

$$\begin{aligned} \ell(k) = \min\Big\{ j \in \mathbb{Z}_{\geq 0} : f(x^k - 2^{-j}(z^k - x^k)) \\ \leq \psi_k - \sigma 2^{-j} \nabla f(x^k)^t (x^k - z^k) \Big\}, \end{aligned}$$

with $\psi_k = \max_{0 \leq j \leq m} f(x^{k-j})$ for some fixed m. The proof of Theorem 1 does not work for this nonmonotone Armijo search: one gets $\epsilon_k \leq \frac{2\hat{\beta}}{\sigma}[\max_{0 \leq j \leq m} f(x^{k-j}) - f(x^{k+1})]$, but the right hand side of this inequality does not seem to be summable, as required for the application of the quasi-Fejér convergence result. The issue of the validity of Theorem 1 for SPG with this nonmonotone search remains as an open problem. We end this remark by mentioning that a variant of SPG with a search along the arc, similar to our option (ii), has also been developed in [5]. Our comments above apply to this variant in relation with Theorem 2.

3. When ∇f is Lipschitz continuous in C with known Lipschitz constant L, it is well known that the Armijo search can be avoided, by taking, as γ_k in option (i) or β_k in option (ii), a constant $\theta \in (0, 2/L)$, without affecting the convergence properties for the nonconvex case (i.e. Propositions 2–5). It is not difficult to verify that in the convex case Theorems 1 and 2 also remain valid for this choice of the stepsizes (in fact, the proofs are indeed simpler).

4. Unpublished results related to those in this paper were obtained by B.F. Svaiter in [16].

REFERENCES

[1] Alber, Ya.I., Iusem, A.N. and Solodov, M.V., *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space,* Mathematical Programming, **81** (1998), 23–37.

[2] Bereznyev, V.A., Karmanov, V.G. and Tretyakov, A.A., *The stabilizing properties of the gradient method,* USSR Computational Mathematics and Mathematical Physics, **26** (1986), 84–85.

[3] Bertsekas, D., *Nonlinear Programming,* Athena Scientific, Belmont (1995).

[4] Birgin, E.G. and Evtushenko, Y.G., *Automatic differentiation and spectral projected gradient methods for optimal control problems,* Optimization Methods and Software, **10** (1998), 125–146.

[5] Birgin, E.G., Martínez, J.M. and Raydan, M., *Nonmonotone spectral projected gradient methods on convex sets,* SIAM Journal on Control and Optimization, **10** (2000), 1196–1211.

[6] Birgin, E.G., Martínez, J.M. and Raydan, M., *SPG: software for convex constrained optimization (to be published in* ACM Transactions on Mathematical Software).

[7] Burachik, R., Graña Drummond, L.M., Iusem, A.N. and Svaiter, B.F., *Full convergence of the steepest descent method with inexact line searches,* Optimization, **32** (1995), 137–146.

[8] Correa, R. and Lemaréchal, C., *Convergence of some algorithms for convex minimization,* Mathematical Programming, **62** (1993), 261–275.

[9] Gafni, E.N. and Bertsekas, D., *Convergence of a gradient projection method,* Technical Report LIDS-P-1201, Laboratory for Information and Decision Systems, M.I.T. (1982).

[10] Golstein, E. and Tretyakov, N., *Modified Lagrange Functions.* Moscow (1989).

[11] Hiriart Urruty, J.-B. and Lemaréchal, C., *Convex Analysis and Minimization Algorithms,* Springer, Berlin (1993).

[11] Iusem, A.N. and Svaiter, B.F., *A proximal regularization of the steepest descent method,* RAIRO, Recherche Opérationelle, **29**, 2, (1995), 123–130.

[12] Kiwiel, K. and Murty, K.G., *Convergence of the steepest descent method for minimizing quasi convex functions,* Journal of Optimization Theory and Applications, **89** (1996), 221–226.

[13] Nesterov, Y.E., *Effective Methods in Nonlinear Programming,* Moscow (1989).

[14] Polyak, B.T., *Introduction to Optimization,* Optimization Software, New York (1987).

[15] Svaiter, B.F., *Projected gradient with Armijo search.* Unpublished manuscript.

[16] Zangwill, W.I., *Nonlinear Programming: a Unified Approach*, Prentice Hall, New Jersey (1969).