On the iterate convergence of descent methods for convex optimization

Clovis C. Gonzaga [†]

March 1, 2014

Abstract

We study the iterate convergence of strong descent algorithms applied to convex functions. We assume that the function satisfies a very simple growth condition around its minimizers, and then show that the trajectory described by the iterates generated by any such method has finite length, which proves that the sequence of iterates converge.

1 Introduction

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous convex function with a set of minimizers $X^* \neq \emptyset$, where it assumes the value f^*. Assume that an initial point $x^0 \in \mathbb{R}^n$ is given, $f(x^0) < +\infty$, and define the set $S = \left\{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\right\}$.

To each point $x \in S$ we associate a *slope*

$$F'(x) = \min\{\|\gamma\| \mid \gamma \in \partial f(x)\} = -\min\{f'(x,d) \mid d \in \mathbb{R}^n, \|d\| \leq 1\},$$

where $\partial f(x)$ stands for the subdifferential of f at x and $f'(x,d)$ is the directional derivative of f along the direction d. We know that $F'(x) = 0$ if and only if $x \in X^*$.

If f is differentiable, the slope is $\|\nabla f(x)\|$, equal to the directional derivative of f along the steepest ascent direction $\nabla f(x)/\,\|\nabla f(x)\|$.

We define a strong descent step from a point $x \in S$.

Definition 1. *A vector $\Delta x \in \mathbb{R}^n$ is a strong descent step from $x \in S$ if*

$$f(x + \Delta x) \leq f(x) - \alpha F'(x)\,\|\Delta x\|, \tag{1}$$

where $\alpha \in (0,1)$ is a given fixed number.

Algorithm 1. *Strong descent algorithm model*

Data: $x^0 \in \mathbb{R}^n$, $k = 0$
WHILE $F'(x^k) > 0$
  $x^{k+1} = x^k + d^k$, where d^k is a strong descent step from x^k
  $k = k + 1$.

A strong descent step from x is usually obtained by taking a step $\Delta x = \lambda d$ along a descent direction d, $\|d\| = 1$ such that $f'(x,d)$ is sufficiently negative, and then choosing λ to satisfy an Armijo condition. In the differentiable case, the step should satisfy $\nabla f(x)^T d \leq -r\,\|\nabla f(x)\|$ and $f(x + \lambda d) \leq f(x) + s\lambda\, d^T \nabla f(x)$ (Armijo condition), where $r, s \in (0,1)$. This will be a strong descent step with $\alpha = rs$.

[†]Department of Mathematics, Federal University of Santa Catarina, Florianópolis, SC, Brazil; e-mail: ccgonzaga1@gmail.com.
[¶]The author is supported by CNPq.

(H1) Growth condition. Assume that there exist numbers $\beta > 0, p \geq 1$ such that for all $x \in S$

$$f(x) \geq g(x) = f^* + \beta d(x, X^*)^p,$$

where $d(x, X^*) = \min\{\|x - z\| \,|\, z \in X^*\} = \|x - P(x)\|$ is the distance between x and the set X^* and $P(x)$ is the projection of x onto X^*.

In this paper we prove that with this condition any minimization algorithm with strong descent iterations generates convergent iterate sequences. The convergence proof is straightforward, using only very simple geometrical ideas. We are aware of the fact that the same result can be obtained by combining results in references [1, 3], and using the Lojaziewicz'z condition discussed below. As this mathematical treatment is no so widespread, we include both proofs in these notes .

Remark. We only study the convergence of the iterates for the algorithms. We do not prove that the limit points are solutions of the minimization problem: the responsibility for solving the problem depends on the specific methods (global convergence). Knowing that the iterates converge, the optimality proofs become easier, as we will discuss briefly in the end of this paper.

Previous results. The pointwise convergence of iterates generated by descent algorithms for nonlinear programming has been established in several situations, always based on some definition of "strong" descent steps.

The steepest descent method with exact line searches is not in this class: it may generate multiple accumulation points even for convex continuously differentiable functions. An example is shown in [7] for a function with positive definite hessian at all non-optimal points. Strong descent steps are usually obtained by taking steps that satisfy an Armijo condition, either by using an Armijo line search or by solving a trust region or proximal point subproblem.

In the case of convex continuously differentiable problems, the convergence of the steepest descent algorithm with Armijo line search was independently proved by Burachik et al. [5] and by Kiwiel and Murty [10] for unconstrained problems and in Iusem [9] for the projected gradient method. These studies use no growth condition, but are limited to the steepest descent directions.

Non-smooth convex problems may be solved by $\epsilon-$subgradient methods, which take steps $x-\lambda d$ with $d \in \partial_\epsilon f(x)$. Imposing an upper bound on the step length λ (thus avoiding examples like the one in [7]), iterate convergence proofs are presented by Correa and Lemaréchal [6] and by Robinson [15].

The analysis of more general classes of algorithms has been made using properties of analytic, semi-algebraic and subanalytic functions, applying the differential properties devised by Łojasiewicz [14] and their extensions by Kurdyka [11]. Let us comment on three papers.

Absil, Mahony and Andrews [1] use the same definition of strong descent steps as we do for the minimization of analytic functions, proving the iterate convergence using Łojasiewicz's theorem.

Attouch, Bolte and Svaiter [2] use the Kurdyka-Łojasiewicz condition, an extension of Łojasiewicz's condition to a very general class non-smooth, non-convex problems. Their definition of strong descent differs from ours: it depends on properties of f at the point $x + \Delta x$, which may be more appropriate then ours for non-smooth problems. Our approach does not seem applicable to their setting.

Bolte, Daniliidis and Lewis [3] describe the extension of Łojasiewicz's theorem to subanalytic functions and apply it to the convergence of subgradient type dynamical systems. They also particularize the results to convex non-smooth problems, concluding that the relevant condition, which is satisfied by subanalytic convex functions in bounded sets is the following, using our notation:

(H2) Łojasiewicz's condition. There exist an exponent $\theta \in [0, 1)$ such that the function

$$x \in S \setminus X^* \mapsto \frac{(f(x) - f^*)^\theta}{F'(x)} \tag{2}$$

is bounded.

They show that if (H2) holds in a bounded set, then the same is true for (H1). For completeness, we show now that the converse is also true (without the limitation to bounded sets).

Lemma 1. *Assume that the growth condition (H1) holds. Then (H2) holds in S.*

Proof. Take a point $x \in S$. By (H1), $f(x) - f^* \geq \beta d(x, X*)^p$, with $\beta > 0$ and $p \geq 1$. Define $x^* = x - P(x)$ and the direction $h = (x - x^*)/\|x - x^*\|$. Then $f(x - d(x, X^*)h) = f^*$. By convexity of f,

$$f'(x, -h) \leq -\frac{f(x) - f^*}{d(x, X^*)}.$$

As $F'(x) \geq -f'(x, h)$ by definition, and $d(x, X^*) \leq (f(x) - f^*)^{1/p}/\beta^{1/p}$, we obtain by substitution

$$F'(x) \geq \beta^{1/p}(f(x) - f^*)^{1-1/p}.$$

Hence (1) holds with $\theta = 1 - 1/p$ completing the proof. □

So, in the particular case of convex problems, the growth condition implies the conditions used in the cited references.

2 Analysis of a strong descent algorithm

In this section we prove that the sequences generated by strong descent algorithms converge whenever (H1) is satisfied. We do it by two different paths, first reproducing the analysis made in [1] using the Łojasiewicz condition and then by a geometrical reasoning based on the growth condition.

2.1 Analysis based on Łojasiewicz condition

Let (x^k) be a sequence generated by a strong descent algorithm as above, and assume that (H2) holds. We shall use the following fact:

Lemma 2. *Given numbers $a > b > 0$ and $\mu \in (0, 1)$,*

$$\frac{a - b}{a^{1-\mu}} \leq \frac{a^\mu - b^\mu}{\mu}.$$

Proof. Define the function $t > 0 \mapsto z(t) = t^\mu$, $\mu \in (0, 1)$. Then $z'(t) = \mu t^{\mu-1}$ decreases for $t > 0$, z is concave and hence

$$\begin{aligned} z(b) &\leq z(a) + (b - a)z'(a) \\ b^\mu &\leq a^\mu - (a - b)\mu a^{\mu-1} \\ a^\mu - b^\mu &\geq \frac{a - b}{a^{1-\mu}}\mu, \end{aligned}$$

completing the proof. □

Theorem 1. *Let x^k be a sequence generated by a strong descent algorithm, and assume that (H2) holds. Then $\sum_{k=0}^{\infty} \left\|d^k\right\| < \infty$, and hence the sequence (x^k) converges.*

Proof. Assume for simplicity that $f^* = 0$. At an iteration k, (H2) gives

$$f(x^k)^{1-\mu} \leq MF'(x^k)$$

for some $M > 0$ and $\mu \in (0, 1)$. Using the strong descent property,

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq \sigma F'(x^k) \left\| d^k \right\| \\ &\geq \frac{\sigma}{M} f(x^k)^{1-\mu} \left\| d^k \right\| \\ \frac{\sigma}{M} \left\| d^k \right\| &\leq \frac{f(x^k) - f(x^{k+1})}{f(x^k)^{1-\mu}}. \end{aligned}$$

Using Lemma 2,

$$\frac{\sigma}{M} \left\| d^k \right\| \leq \frac{1}{\mu}(f(x^k) - f(x^{k+1})).$$

Summing this expression we get

$$\frac{\sigma}{M} \sum_{i=0}^{k-1} \left\| d^i \right\| \leq \frac{1}{\mu}(f(x^0) - f(x^k)) \leq \frac{1}{\mu} f(x^0) \leq +\infty,$$

completing the proof.

2.2 Analysis based on (H1)

Simplification. We assume for simplicity that $f^* = 0$, and then use the function $x \in \mathbb{R}^n \mapsto g(x) = \beta d(x, X^*)^p$.

Two level sets are associated with each point $z \in S$:
$S(z) = \{y \in \mathbb{R}^n \mid f(y) \leq f(z)\}$,
$B(z) = \{y \in \mathbb{R}^n \mid g(y) \leq f(z)\} = \{y \in \mathbb{R}^n \mid d(y, X^*) \leq f(z)^{1/p}/\beta\}$.

It follows from the hypothesis on g that for all $s \in S$, $S(x) \subset B(x)$. The set $B(z)$ may be seen as a neighborhood of the set X^*, with a radius $\mathcal{R}(z) = \sqrt[p]{f(z)}/\beta$. Let us fix a point $x \in S$ and study a strong descent step $x^+ = x + d$ from x. We shall prove that the reduction in the radius of the neighborhoods, $\mathcal{R}(x) - \mathcal{R}(x^+)$ is bounded below by a multiple of $\|d\|$: this implies that the sequence $\left(\left\| d^k \right\|\right)$ has finite sum. The treatment uses elementary geometrical properties represented in Fig 1.

By definition, $f(x^+) \leq f(x) + -\alpha F'(x)\|d\|$. Define also the projection $x^* = P_{X^*}(x)$ and set $h = (x - x^*)/\|x - x^*\|$.

We study the restrictions of f and g to the line $\{x^* + \lambda h \mid \lambda \geq 0\}$. See Fig.1 for the geometry. Note that for $\lambda \geq 0$, $g(x^* + \lambda h) = \beta\lambda^p$, because $d(x^* + \lambda h, X^*) = \lambda$.

Our first lemma relates the descent of f along the step d and along the direction $-h$.

Lemma 3. *Assume that $x^+ = x + d$ is a strong descent step from $x \in S$, and define $y = x^* + b\,h$ such that $f(y) = f(x^+)$. Then $\|x^+ - x\| \leq \|y - x\| /\alpha$.*

Proof. From the definition of strong descent and from the convexity of f, we have

$$\begin{aligned} f(x^+) - f(x) &\leq -\alpha F'(x) \left\| x^+ - x \right\| \\ f(y) - f(x) &\geq f'(x, -h) \|y - x\| \geq -F'(x) \|y - x\|. \end{aligned}$$

As $f(x^+) = f(y)$, merging the expressions above we obtain $\alpha F'(x) \left\| x^+ - x \right\| \leq F'(x) \|y - x\|$. Dividing both terms by $F'(x) > 0$, we complete the proof.

□

Now we study the behavior of $f(x^* + \lambda h)$ for $\lambda \in \mathbb{R}_+$. Set $\chi = \|x\|$ and define the linear upper approximation of f between x^* and x: $\lambda \in [0, \chi] \mapsto \bar{f}(\lambda) = \lambda f(x)/\chi$. Clearly $\bar{f}(\lambda) \geq f(\lambda h)$ for $\lambda \in [0, \chi]$.

Define the following points in R_+: a, b, χ, s, t such that:
$f(x) = f(x^* + \chi h) = g(x^* + t\,h)$,

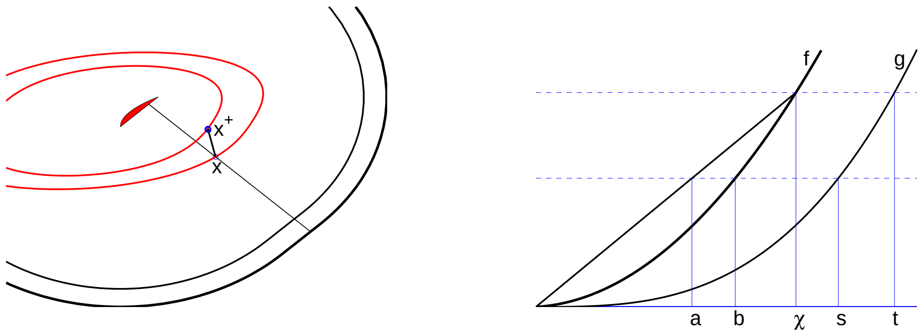

Figure 1: The geometry of a descent step.

$f(x^+) = \bar{f}(a\,h) = f(x^* + b\,h) = g(x^* + s\,h)$.
Note that $t = \mathcal{R}(x)$, $s = \mathcal{R}(x^+)$, $\chi = \|x - x*\|$, and of course the projection of all points on the line $\{\lambda\,h\,|\,h \geq 0\}$ onto X^* is x^*.

Lemma 4. *Consider the construction above. Then*

$$\chi - b \leq \chi - a \leq p(t - s) = p(\mathcal{R}(x) - \mathcal{R}(x^+)).$$

Proof. The first inequality stems immediately from the fact that $\bar{f}(\lambda) \geq f(x^* + \lambda\,h)$ on $[0, \chi]$ and $\bar{f}(a) = f(x^* + b\,h)$. Let us prove the second inequality.

$$g'(x^* + t\,h, h) = p\beta t^{p-1} = p\,\frac{g(x^* + t\,h)}{t} = p\,\frac{f(x)}{t}.$$

By convexity of g, $g(x^* + s\,h) - g(x^* + t\,h) \geq -g'(x^* + t\,h, h)(t - s)$, and hence using the definitions of s and t,

$$f(x^+) - f(x) \leq -p\,\frac{f(x)}{t}(t - s) \leq -p\,\frac{f(x)}{\chi}(t - s), \tag{3}$$

where the last inequality holds because the (H1) implies $s \geq \chi$. On the other hand, from the definitions of $\bar{f}$, a and χ,

$$f(x) - f(x^+) = (\chi - a)\frac{f(x)}{\chi}.$$

Substituting this into (3) we obtain immediately the result, completing the proof. □

We are ready to prove the main result.

Theorem 2. *Assume that the algorithm generates an infinite sequence (x^k) such that for $k \in \mathbb{N}$, $x^{k+1} = x^k + d^k$ is a strong descent step with constant $\alpha > 0$, and assume that (H1) holds. Then*

(i) $\mathcal{R}(x^k) - \mathcal{R}(x^{k+1}) \geq \frac{\alpha}{p}\left\|d^k\right\|$.

(ii) $\sum_{k=0}^{\infty} \left\|d^k\right\| \le \frac{p}{\alpha}\mathcal{R}(x^0) < \infty$,
and hence the sequence (x^k) *converges.*

Proof. Let $x = x^k$, for a given $k \in \mathbb{N}$, define $x^+ = x^{k+1}$. With the construction made above, we obtain from Lemma 3 and then Lemma 4,

$$\left\|x^k - x^{k+1}\right\| \le \frac{\chi - b}{\alpha} \le \frac{p}{\alpha}(\mathcal{R}(x^k) - \mathcal{R}(x^{k+1})),$$

proving (i). Summing (i) for $i = 0, 1, \dots, k$, we obtain

$$\mathcal{R}(x^0) - \mathcal{R}(x^k) \ge \frac{\alpha}{p}\sum_{i=0}^{k}\left\|d^i\right\|$$

But $\mathcal{R}(x^k) > 0$ for $k \in \mathbb{N}$, which results in (ii), completing the proof. □

This proves the convergence of the iterates. Of course, if the steps are taken deliberately short, the sequence may converge to a non-stationary point, even for smooth functions. For non-smooth problems, strong descent in this sense is not enough to prove convergence to an optimal solution.

3 Strong descent in constrained convex problems

Consider now the constrained problem

$$\underset{x \in C}{\text{minimize}} \quad f(x)$$

where $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a lower semi-continuous convex function, C is closed and convex, and assume that $X^* = \operatorname{argmin}_{x \in C} f(x) \ne \emptyset$. Assume also that an initial point $x^0 \in C$ is given, $f(x^0) < +\infty$, and define the set $S = \left\{x \in C \mid f(x) \le f(x^0)\right\}$.
Denote $f^* = f(x^*)$, for an arbitrary $x^* \in X^*$.

In the unconstrained case, the slope is defined by minimizing $f'(x, h)$ for h in a ball, whose radius does not influence the outcome. Now we must consider feasible directions, and the radius will matter. Let us define the slope associated with a point $x \in C$ and a radius $\Delta > 0$:

$$F'_{\Delta}(x) = \inf\left\{f'(x, d) \mid x + d \in C, \|d\| \le \Delta\right\}.$$

We call a step $x^+ = x + d \in C$ strong if

$$f(x^+) \le f(x) + \alpha F'_{\|d\|}(x). \tag{4}$$

Now we use the same construction as in the previous section and prove a lemma similar to Lemma 3, defining $x^* = P_{X^*}(x)$, $h = (x - x^*)/\|x - x^*\|$.

Lemma 5. *Assume that* $x^+ = x + d$ *satisfies* (4) *and define* $y = x^* + b\,h$ *such that* $f(y) = f(x^+)$*. Then* $\|x^+ - x\| \le \|y - x\|/\alpha$.

Proof. If $\|y - x\| \le \|d\|$, the result is trivial for $\alpha \in (0, 1)$. Assume then that $\|y - x\| > \|d\|$, and $y = x - \|d\|\,h \in C$. From the definitions of slope and strong descent, and using the convexity of f, we have

$$\begin{aligned} f'(x, -\|d\|\,h) &\ge F'_{\|d\|}(x), \\ f'(x, -h) &\ge F'_{\|d\|}(x)/\|d\|, \\ f(x^+) - f(x) &\le \alpha F'_{\|d\|}(x)\left\|x^+ - x\right\|, \\ f(y) - f(x) &\ge f'(x, -h)\,\|y - x\| \ge F'_{\|d\|}(x)\,\|y - x\|/\|d\|. \end{aligned}$$

As $f(x^+) = f(y)$, merging the two last expressions above we complete the proof. □

From now on, the treatment is the same as in the previous section, since Lemma 4 is not influenced by the presence of the constraint. The conclusion of Theorem 2 applies to the constrained case.

This is directly applicable to the method of projected gradients: let f be differentiable, $x \in C$ and $x^+ = P_C(x - \lambda \nabla f(x))$. Then it is possible to prove by direct application of optimality conditions that $x^+ = x + d$ where $d = \text{argmin}\left\{\nabla f(x)^T h \mid \|h\| \leq \|d\|, x + d \in C\right\}$, and hence $F'_\Delta(x) = \nabla f(x)^T d$.

4 Examples

Let us describe some applications of the theory.

• Proximal point method: a proximal iteration [12] from a point x solves $\bar{d} = \text{argmin}\{f(x + d) + c\|d\|^2\}$, with $c > 0$. If follows trivially that $f(x + \bar{d}) = \min\{f(x + d) \mid \|d\| \leq \|\bar{d}\|\}$, and in our construction $\chi - b \geq \|\bar{d}\|$, because $f(x^* + bh) = f(x + \bar{d})$. So the result of Lemma 3 is true and the complete analysis holds.

• Quasi-Newton methods with Armijo line search: a typical quasi-Newton method for a function of class C1 computes at an iteration from a point x

$$d = \underset{s \in \mathbb{R}^n}{\text{argmin}} \left\{\nabla f(x)^T s + s^T A s/2\right\}, \tag{5}$$

where A is a positive definite matrix with bounded condition number $C(A) \leq \bar{C}$. The next iterate is then computed by an Armijo line search along d. As we commented above, this will be a strong descent step if $\nabla f(x)^T d/(\|\nabla f(x)\| \|d\|)$ is bounded away from zero.

From (5), $\nabla f(x) + Ad = 0$, or $d = -A^{-1}\nabla f(x)$. Hence

$$\begin{aligned} \|d\| &= \left\|A^{-1}\nabla f(x)\right\| \leq \|\nabla f(x)\| / \mu_{min}, \\ \nabla f(x)^T d &= -\nabla f(x)^T A^{-1} \nabla f(x) \leq -\|\nabla f(x)\|^2 / \mu_{max}, \end{aligned}$$

where μ_{min} and μ_{max} are respectively the smallest and largest eigenvalues of A. If follows by substitution that

$$\frac{\nabla f(x)^T d}{\|\nabla f(x)\| \|d\|} \leq -\frac{\mu_{min}}{\mu_{max}} \leq -\frac{1}{\bar{C}},$$

completing the proof.

• Trust region, Steinhaug and dog-leg methods: the strong descent condition for these methods applied to functions of class C1 is proved in the textbook by Nocedal and Wright [13, Sec.4.2].

Convergence to an optimizer. A strong descent algorithm may converge to a non-optimal point if the steps are too short. Let us define the step length of a descent iteration $x^{k+1} = x^k + d^k$ as $\lambda_k = \|d^k\| / F'(x^k)$ (this agrees with the notation for the steepest descent $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$).

Lemma 6. *Assume that a strong descent algorithm in our general setting generates an infinite sequence of iterates. If* $\sum_{k=0}^{\infty} \lambda_k = \infty$ *then* (x^k) *converges to a minimizer of* f.

Proof. Here we use the following properties of convex functions: the point to set map $x \in \mathbb{R}^n \mapsto \partial f(x)$ is upper semi-continuous (see for instance [8]), which implies that the function $x \in \mathbb{R}^n \mapsto F'(x)$ is upper semi-continuous.

Assume that $x^k \to \bar{x}$. Then $\liminf_{k\to\infty} F'(x^k) \geq F'(\bar{x})$. Suppose by contradiction that $F'(\bar{x}) > \epsilon > 0$ and that $\sum_{k=0}^{\infty} \lambda_k = \infty$. For k sufficiently large, $F'(x^k) > \epsilon$, and using the definition of strong descent

$$f(x^k) - f(x^{k+1}) \geq \alpha F'(x^k) \left\|d^k\right\| = \alpha \lambda_k F'(x^k)^2 > \alpha \lambda_k \epsilon^2.$$

Summing this expression we obtain $f(x^0) - f(\bar{x}) \geq \alpha\epsilon^2 \sum_{k=0}^{\infty} \lambda_k = \infty$. This contradicts the fact that $f(x^k) \to f(\bar{x})$, completing the proof. □

An interesting fact, not usually known, is proved in [4], and we repeat it here for completeness. It says that when the step lengths have a finite sum the limit point cannot be stationary. For this fact the function does not need to be convex.

Lemma 7. *Assume that f is differentiable and consider an infinite sequence of non-stationary points*

$$x^{k+1} = x^k + d^k, \qquad \left\|d^k\right\| = \lambda_k \left\|\nabla f(x^k)\right\|.$$

Assume that $\{x_k\}$ is contained in a bounded set $\Omega \in \mathbb{R}^n$ where $\nabla f(\cdot)$ is Lipschitz continuous. If $\sum_{k=0}^{\infty} \lambda_k < +\infty$ then (x^k) converges to a non-stationary point.

Proof. Let us denote $g^k = \nabla f(x^k)$. We first prove that (x^k) converges (this is already known in the case considered above in this paper).

Assume that If $\sum_{k=0}^{\infty} \lambda_k < +\infty$. If (x^k) is bounded, so is (g^k). Then the length of the sequence (x^k) is

$$\sum_{k=0}^{\infty} \left\|d^k\right\| < M \sum_{k=0}^{\infty} \lambda_k < \infty,$$

where M is an upper bound for $\left\|g^k\right\|, k \in \mathbb{N}$. Hence (x^k) converges.

We have $\left\|x^{k+1} - x^k\right\| = \lambda_k \left\|g^k\right\|$ and if L is a Lipschitz constant for ∇f in Ω,

$$\left\|g^{k+1} - g^k\right\| \leq L \left\|x^{k+1} - x^k\right\| = L\lambda_k \left\|g^k\right\|.$$

Assume by contradiction that $\left\|g^k\right\| \to 0$. We may choose K such that $\left\|g^K\right\| \geq \left\|g^k\right\|$ for $k > K$, and also $\sum_{k=K}^{\infty} < 1/(2L)$. Then for $k > K$,

$$\left\|g^k - g^K\right\| \leq \sum_{j=K}^{k-1} \left\|g^{j+1} - g^j\right\| \leq \sum_{j=K}^{k-1} L\lambda_j \left\|g^j\right\| < L \left\|g^K\right\| /(2L) = \left\|g^K\right\| /2.$$

Hence $\left\|g^k\right\| > \left\|g^K\right\| /2$ for $k > K$, which establishes a contradiction, completing the proof.

□

References

[1] P.-A Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[2] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forwardbackward splitting, and regularized gaussseidel methods. *Math. Programming, Ser.A*, 137:91–129, 2013.

[3] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2006.

[4] J. F. Bonnans and C. C. Gonzaga. Convergence of interior point algorithms for the monotone linear complementarity problem. *Mathematics of Operations Research*, 21(1):1–25, 1996.

[5] R. Burachik, L. Drummond, A. Iusem, and B. Svaiter. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 32:137–146, 1995.

[6] R. Correa and C. Lemaréchal. Convergence of some algorithms for convex minimization. *Math. Program.*, 62(2):261–275, 1993.

[7] C. C. Gonzaga. Two facts on the convergence of the Cauchy algorithm. *Journal of Optimization Theory and Applications*, 107(3):591–600, 2000.

[8] J-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, New York, 1993.

[9] A. N. Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational and Applied Mathematics*, 22(1):37–52, 2003.

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

[11] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48:769–783, 1998.

[12] B. Lemaire. The proximal algorithm. *International Series of Numerical Mathematics, 87*, pages 73–87, 1989.

[13] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

[14] S. Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Ann. Inst. Fourier*, 43:1575–1595, 1993.

[15] S. M. Robinson. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. *Math. Program*, 86:41–50, 1996.