Convergence of the Iterates of Descent Methods for Analytic Cost Functions

P.-A. Absil[†] R. Mahony[‡] B. Andrews[§]

Abstract

In the early eighties Łojasiewicz [Loj84] proved that a bounded solution of a gradient flow for an *analytic* cost function converges to a well-defined limit point. In this paper, we show that the iterates of numerical descent algorithms, for an *analytic* cost function, share this convergence property if they satisfy certain natural descent conditions. The results obtained are applicable to a broad class of optimization schemes and strengthen classical "weak convergence" results for descent methods to "strong limit-point convergence" for a large class of cost functions of practical interest. The result does not require that the cost has isolated critical points, requires no assumptions on the convexity of the cost, nor any non-degeneracy conditions on the Hessian of the cost at critical points.

Key words. gradient flows, descent methods, real analytic functions, Lojasiewicz gradient inequality, single limit-point convergence, line-search, trust-region, Mexican Hat

1 Introduction

Unconstrained numerical optimization schemes can be classified into two principal categories: line-search descent methods and trust-region methods. Seminal work by Goldstein [Gol65] and Wolfe [Wol69] on line-search descent methods introduced easily verifiable bounds on step-size selection that lead to weak convergence results ($\lim \|\nabla\phi(x_k)\| = 0$) for a wide class of inexact line-search descent algorithms; see e.g. [Fle87, theorem 2.5.1] or [NW99, theorem 3.2]. For trust-region methods, classical convergence results guarantee weak convergence ($\lim \|\nabla\phi(x_k)\| = 0$) if the total model decrease is at least a fraction of that obtained at the Cauchy point; see e.g. [NW99, theorem 4.8] or [CGT00, theorem 6.4.6]. Thus, classical convergence results establish that accumulation points of the sequence of iterates are stationary points of the cost function ϕ. Convergence of the whole sequence to a single limit-point is not guaranteed. Curry [Cur44, p. 261] first gave the following intuitive counter-example to the existence of such a result for steepest-descent methods with line minimization:

> Let $G(x, y) = 0$ on the unit circle and $G(x, y) > 0$ elsewhere. Outside the unit circle let the surface have a spiral gully making infinitely many turns about the circle. The path[1] C will evidently follow the gully and have all points of the circle as limit points[2].

[†]School of Computational Science, Florida State University, Tallahassee FL 32306-4120, USA (http://www.csit.fsu.edu/∼absil/). This author's work was supported partially by the School of Computational Science of Florida State University through a postdoctoral fellowship. Part of this work was done while the author was a Research Fellow with the Belgian National Fund for Scientific Research (Aspirant du F.N.R.S.) at the University of Liège.

[‡]Department of Engineering, Australian National University, ACT, 0200, Australia (mahony@ieee.org).

[§]Center for Mathematical Analysis, IAS, Australian National University, ACT, 0200, Australia (Ben.Andrews@anu.edu.au).

[1]The path consists of the sequence of estimates of the numerical method.

[2]A point x is a *limit point* or *accumulation point* of a sequence $\{x_k\}_{k\in\mathbb{N}}$ if there exists a subsequence $\{x_{k_i}\}_{i\in\mathbb{N}}$ that converges to x.

It is possible to prove single limit-point convergence for descent algorithms by exploiting additional properties of the cost that ensure critical points are isolated or impose non-degeneracy conditions on the Hessian of the cost on critical sets [HM94]. Strong convexity of the cost function guarantees that the global minimum is a unique isolated critical point of the cost function and single limit-point convergence is recovered; see e.g. Byrd and Nocedal [BN89] for the BFGS algorithm and Kiwiel and Murty [KM96] for the steepest descent method. Convergence results obtained by Dunn [Dun81, theorem 4.3] require an uniform growth condition of ϕ and uniqueness of the minimizer within a certain subset. For a class of approximate trust-region methods, Moré and Sorensen [MS83, theorem 4.13] show that if the Hessian of ϕ is nonsingular at an accumulation point x^*, then the whole sequence converges to x^*. Conn *et al.* [CGST93] (or see theorem 6.5.2 in [CGT00]) show that the same result holds for a class of trust-region methods that ensure a fraction of Cauchy decrease. The Capture Theorem [Ber95], for a class of line-search methods, shows convergence to a single local minimum x^*, provided x^* is an isolated stationary point of ϕ and the iteration comes sufficiently close to x^*; see also [Dun87].

In this paper, we consider the question of convergence given certain regularity conditions on the cost function considered. The motivation for our study is a result in dynamical systems theory, that has only recently become widely recognized. For a generic smooth cost function, the ω-limit set [Wig90, pg. 42] of a bounded gradient flow is a connected subset of critical points, and not necessarily a single point [HM94, Prop. C.12.1]. If the cost function is real analytic[3], then Łojasiewicz's theorem [Loj84] states that the associated gradient flow converges to a single limit-point; see Section 2 or the introduction of [KMP00] for an overview of Łojasiewicz's argument. A comprehensive treatment of the contintuous-time convergence results with applications in optimization theory is contained in the Diploma thesis [Lag02]. The key to the proof lies in showing that the total length of the solution trajectory to the gradient flow is bounded. The proof utilizes the Łojasiewicz gradient inequality (see Lemma 2.1 on page 3) which gives a lower bound for the norm of the gradient of ϕ in terms of ϕ itself. Due to the importance of this result in the motivation of our work, we provide a review of this result in the early part of the paper, and go on to present an explicit counter example that shows that single limit-point convergence can not be proved in general for C^∞ cost functions.

The main contribution of the paper is to adapt these results to iterates of numerical descent algorithms. We define a pair of descent conditions termed the *strong descent conditions* that characterize the key properties of a sequence of iterates that leads to single limit-point convergence. These conditions are deliberately chosen to be as weak as possible in order to apply to the widest possible class of numerical descent algorithms. For line-search methods, it is sufficient to impose an angle condition and the first Wolfe condition (also known as Armijo condition). For trust-region methods, we give several easily verified conditions involving the Cauchy point that guarantee the strong descent conditions hold. The main theorem uses these conditions to prove that the whole sequence of iterates $\{x_k\}$ of a numerical descent algorithm, applied to an analytic cost function, either escapes to infinity (i.e. $\|x_k\| \to +\infty$) or converges to a single limit-point. An interesting aspect of the development is that the strong descent conditions themselves do not guarantee convergence to a critical point of the cost, $\|\nabla\phi(x_k)\| \not\to 0$. However, combining single limit-point convergence with classical weak convergence results leads to convergence to a single critical point for a wide range of classical numerical descent algorithm applied to analytic cost functions.

Apart from ensuring continuity of ϕ, the only purpose of the analyticity assumption is to guarantee that the Łojasiewicz gradient inequality holds in a neighbourhood of every point. Therefore, the domain of application of our results go beyond the (already large) class of analytic functions to functions that satisfy a simple growth condition (see Eq. 7). If moreover it is known that a point x^* is an accumulation point, then in order to have convergence of the whole sequence to x^* it is sufficient to require that this growth condition holds in a neighbourhood of x^*.

A preliminary version of the results presented in this paper appeared in the proceedings of the 13th MTNS conference [Mah98]. Generalizations to Riemannian manifolds have been considered

[3] A real function is said to be analytic if it possesses derivatives of all orders and agrees with its Taylor series in the neighbourhood of every point.

in [Lag02].

The paper is organized as follows. The continuous-time case is reviewed in Section 2 and the Mexican Hat example is presented. The general convergence theory for descent iterations is developed in Section 3 and applied to line-search and trust-region methods in Section 4. Conclusions are presented in Section 5.

2 Convergence of Analytic Gradient Descent Flows

In this section, we briefly review Łojasiewicz's argument for the convergence of analytic gradient flows and give an explicit counter-example to show that single limit-point convergence does not hold for certain C^∞ gradient flows. In the past five years, many authors have revisited the original gradient flow convergence results of Łojasiewicz [Loj84]. Our presentation follows the generalization proposed by Lageman [Lag02], where the steepest-descent direction was relaxed to an angle condition. The proof is included to provide motivation for the discrete-time analysis in Section 3. A concise presentation of the standard argument for Łojasiewicz's theorem is contained in [KMP00].

Let $\mathbb{R}^n$ be the linear space of column vectors with n components, endowed with the usual inner product $\langle x, y\rangle = x^T y$. Let $\nabla\phi(x) := (\partial_1\phi(x), \ldots, \partial_n\phi(x))^T$ denote the Euclidean gradient of the differentiable function ϕ. A point x^* where $\nabla\phi(x^*) = 0$ is called a *stationary point* or *critical point* of ϕ.

The proof of Łojasiewicz's theorem is based on the following property of real analytic functions.

Lemma 2.1 (Łojasiewicz gradient inequality) [4] *Let ϕ be a real analytic function on a neighbourhood of x^* in $\mathbb{R}^n$. Then there are constants $c > 0$ and $\mu \in [0, 1)$ such that*

$$\|\nabla\phi(x)\| \geq c|\phi(x) - \phi(x^*)|^\mu \tag{1}$$

in some neighbourhood U of x^*.

Proof. See [Loj65, p. 92], [BM88, prop. 6.8], or the short proof in [KP94]. □

Theorem 2.2 *Let ϕ be a real analytic function and let $x(t)$ be a C^1 curve in $\mathbb{R}^n$, with $\dot{x}(t) = \frac{dx}{dt}(t)$ denoting its time derivative. Assume that there exists a $\delta > 0$ and a real τ such that for $t > \tau$, $x(t)$ satisfies the angle condition*

$$\frac{d\phi(x(t))}{dt} \equiv \langle \nabla\phi(x(t)), \dot{x}(t)\rangle \leq -\delta\|\nabla\phi(x(t))\|\|\dot{x}(t)\| \tag{2}$$

and a weak decrease condition

$$\left[\frac{d}{dt}\phi(x(t)) = 0\right] \Rightarrow [\dot{x}(t) = 0]\,. \tag{3}$$

Then, either $\lim_{t\to+\infty}\|x(t)\| = \infty$, or there exists $x^* \in \mathbb{R}^n$ such that $\lim_{t\to+\infty} = x^*$.

Proof. Assume that $\|x(t)\| \nrightarrow +\infty$ as $t \to +\infty$. Then $x(t)$ has an accumulation point x^* in $\mathbb{R}^n$. It remains to show that $\lim_{t\to+\infty} x(t) = x^*$ and the proof will be complete.

It follows from (2) that $\phi(x(t))$ is nonincreasing. Moreover, since x^* is an accumulation point of $x(t)$, it follows by continuity of ϕ that

$$\phi(x(t)) \downarrow \phi(x^*).$$

[4] The Łojasiewicz gradient inequality is a special instance of a more general Łojasiewicz inequality [Łoj59, Łoj93]. The latter result has been used in the study of error bounds of analytic inequality systems in optimization [LP94, Ded00]. In turn, such error bounds have been used in the convergence analysis of optimization algorithms in the same general spirit as done in the present paper; see, e.g., [FFK00, YDF04]. We thank an anonymous reviewer for pointing this out.

We distinguish two cases.

Case (i): there exists a $t_1 > \tau$ such that $\phi(x(t_1)) = \phi(x^*)$. Since $\phi(x(t))$ is non-increasing then it is straightforward to see that $\phi(x(t)) = \phi(x^*)$ and $\frac{d}{dt}\phi(x(t)) = 0$ for all $t \geq t_1$. From the weak decrease condition (3) this implies that $\dot{x}(t) = 0$ for all $t \geq t_1$ and $x(t) = x(t_1) = x^*$.

Case (ii): $\phi(x(t)) > \phi(x^*)$ for all $t > \tau$. In order to simplify the forthcoming equations we assume without loss of generality that $\phi(x^*) = 0$. It follows from the Łojasiewicz gradient inequality (Lemma 2.1) and from (2) that

$$\frac{d\phi(x(t))}{dt} \leq -\delta\|\nabla\phi(x(t))\|\|\dot{x}(t)\| \leq -\delta c|\phi(x(t))|^{\mu}\|\dot{x}(t)\| \tag{4}$$

holds in a neighbourhood U of x^* for some $\mu \in [0, 1)$. Since we have assumed that $\phi(x(t)) > \phi(x^*) = 0$, it follows from (4)

$$c_1 \frac{d(\phi(x(t)))^{1-\mu}}{dt} \leq -\|\dot{x}(t)\| \tag{5}$$

where $c_1 := [\delta c(1-\mu)]^{-1} > 0$. Given t_1 and t_2 with $\tau < t_1 < t_2$, if $x(t) \in U$ for all $t \in (t_1, t_2)$ then by integration of (5)

$$L_{12} := \int_{t_1}^{t_2} \|\dot{x}(t)\|dt \leq c_1((\phi(x(t_1)))^{1-\mu} - (\phi(x(t_2)))^{1-\mu}) \leq c_1(\phi(x(t_1)))^{1-\mu}. \tag{6}$$

Now let r be such that $B_r(x^*) \subset U$, where

$$B_r(x^*) := \{x \in \mathbb{R}^n : \|x - x^*\| < r\}.$$

We show that $x(t)$ eventually enters and remains in $B_r(x^*)$. Since r is arbitrarily small, it follows that $x(t)$ converges to x^* and the theorem will be proven.

Let t_1 be such that $\|x(t_1) - x^*\| < r/2$ and $c_1\phi^{1-\mu}(x(t_1)) < r/2$. Such a t_1 exists by continuity of ϕ since x^* is an accumulation point of $x(t)$ and $\phi(x^*) = 0$. Then we show that the entire trajectory after t_1 lies in $B_r(x^*)$. By contradiction, suppose not, and let t_2 be the smallest $t > t_1$ such that $\|x(t_2) - x^*\| = r$. Then $x(t)$ lies in U for all $t \in (t_1, t_2)$. Therefore (6) holds and it follows that $L_{12} \leq c_1(\phi(x(t_1)))^{1-\mu} < r/2$. Then $\|x(t_2) - x^*\| \leq \|x(t_2) - x(t_1)\| + \|x(t_1) - x^*\| < L_{12} + r/2 < r$, a contradiction. Thus $x(t)$ remains in $B_r(x^*)$ for all $t \in [t_1, +\infty)$, and the proof is complete. □

The role of the weak decrease condition (3) is to prevent the trajectory $x(t)$ from wandering endlessly in the critical set $\nabla\phi = 0$. It is possible to weaken this condition somewhat to allow the trajectory to spend finite periods of time wandering in this set as long as it eventually either converges or continues to decrease the cost (see [Lag02]).

Considering Theorem 2.2, a natural question to ask is if it is possible to relax the condition of analyticity on the cost function and retain the convergence results. Clearly, analyticity is principally used to invoke the Łojasiewicz gradient inequality. The rationale goes through if ϕ is continuous at an accumulation point x^* of $x(t)$ and a growth condition of the type

$$\|\nabla\phi(x^*)\| \geq \psi(\phi(x(t)) - \phi(x^*)) \tag{7}$$

holds in a neigbourhood of x^*, where $1/\psi$ is positive and integrable on an interval $(0, \epsilon)$. In practice, such a growth condition may be difficult to check. This is especially true when no accumulation point is known a priori so that the condition must be verified on a set.

Theorem 2.2 does not hold for the general class of smooth cost functions $\phi \in C^{\infty}$. It is instructive to provide an explicit counter example. The following function $f \in C^{\infty}$ (cf. Figure 1) is a smooth example of a 'Mexican Hat' cost function. Let

$$f(r, \theta) := \begin{cases} e^{-\frac{1}{1-r^2}}\left[1 - \frac{4r^4}{4r^4 + (1-r^2)^4}\sin\left(\theta - \frac{1}{1-r^2}\right)\right] & \text{if } r < 1, \\ 0 & \text{if } r \geq 1, \end{cases} \tag{8}$$

where (r, θ) denote polar coordinates in $\mathbb{R}^2$.

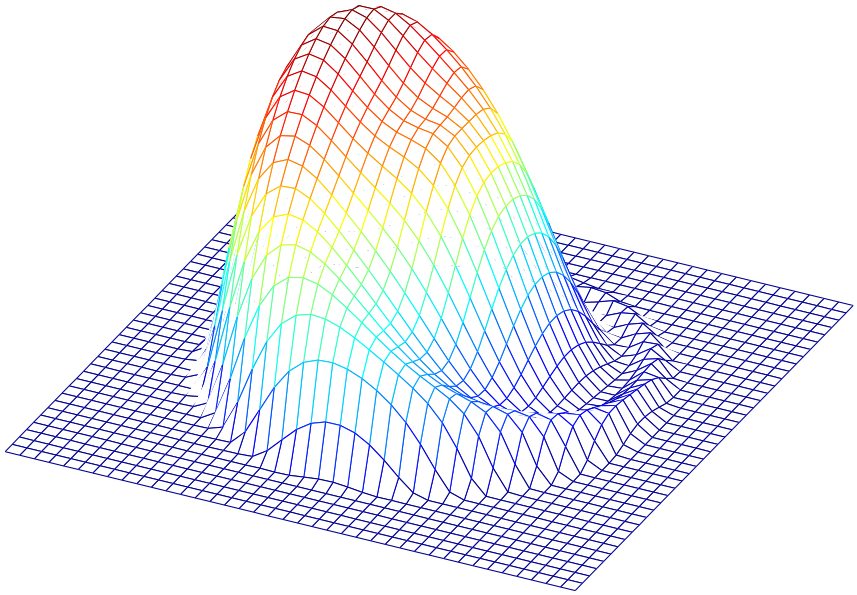

Figure 1: A plot of the smooth 'Mexican Hat' function defined in (8).

Since $0 \leq \frac{4r^4}{4r^4+(1-r^2)^4} < 1$ for all $r < 1$, it follows that $f(r,\theta) > 0$ for all $r < 1$. The exponential factor in f ensures that all derivatives at $r = 1$ are well defined (and equal to zero) and it follows that $f \in C^\infty$. The example has been constructed such that, for initial conditions (r_0, θ_0) with $\theta_0(1 - r_0^2) = 1$ and $0 < r_0 < 1$, the solution $(r(t), \theta(t))$ of the gradient descent flow (expressed in polar coordinates) satisfies

$$\theta(t) = \frac{1}{1 - r(t)^2}. \tag{9}$$

By inspection, the ω-limit set of the trajectory Eq. 9 considered is the entire circle $\{(r,\theta) \,|\, r = 1\}$.

The origin of the colloquial name 'Mexican Hat' function for a counter-example of this form is not clear. Certainly, the structure of the counter-example was known by the time of Curry [Cur44]. Prior examples of Mexican Hats were proposed in [Zou76] (mentioned in [Ber95, Exercise 2.18]) and [PdM82, example 3 p. 13]. The merit of the cost function (8) is to provide a closed-form trajectory (9) and render the convergence analysis trivial.

3 Convergence of analytic descent iterations

In this section, a discrete-time analogue of Theorem 2.2 (Łojasiewicz's theorem with an angle condition) is obtained. We propose a pair of 'strong descent conditions' that encapsulate the key properties of the iterates of a numerical descent algorithm that lead to single limit-point convergence for an analytic cost function. In later sections we show that the strong descent conditions are satisfied naturally by most numerical descent algorithm iterates.

3.1 Main result

In the discrete-time case, a solution trajectory is a sequence $\{x_k\}$ instead of a curve $x(t)$. The key to extending the results of Section 2 to this case is to adapt the conditions (2) and (3) to the discrete-time case. For (2) we propose a *primary descent condition*:

$$\phi(x_k) - \phi(x_{k+1}) \geq \sigma \|\nabla\phi(x_k)\| \|x_{k+1} - x_k\| \tag{10}$$

for all k and for some $\sigma > 0$. Condition (10) is satisfied under Armijo's condition (22) along with an angle condition (20). This fact will be exploited in Section 4.1 in the context of line-search methods. Moreover (10) is sufficiently general to accomodate the framework of trust-region methods; see Section 4.2.

Condition (10) itself does not preclude $\{x_k\}$ from endlessly wandering in a critical set of ϕ. To overcome this, we introduce a *complementary descent condition*:

$$[\phi(x_{k+1}) = \phi(x_k)] \Rightarrow [x_{k+1} = x_k] \tag{11}$$

This condition simply requires that any nonvanishing update, $x_{k+1} \neq x_k$, produce a change in the cost function. Condition (11) adds information to (10) only when x_k is a critical point (i.e. $\nabla\phi(x_k) = 0$). Note that conditions (10) and (11) allow the sequence $\{x_k\}$ to stagnate for arbitrarily many iterations, a behaviour observed e.g. in trust-region methods when the model estimate turns out to be so poor that the proposed update is rejected (see Section 4.2). Together, we term conditions (10) and (11) the *strong descent conditions*.

Definition 3.1 (strong descent conditions) *We say that a sequence $\{x_k\}$ in $\mathbb{R}^n$ satisfies the* strong descent conditions *if* (10)-(11) *hold for some $\sigma > 0$ and for all k larger than some K.*

The main result (Theorem 3.2 below) shows that if the iterates $\{x_k\}$ of a numerical descent algorithm satisfy the strong descent conditions (Definition 3.1) and the cost function ϕ is analytic then $\{x_k\}$ converges to a single point or diverges to infinity. Note that we do not claim that the limit-point is a stationary point of ϕ; indeed, the assumptions are not strong enough (in particular, they do not preclude stagnation). For classical descent algorithms, convergence to a stationary point can be obtained by invoking classical weak convergence results ($\nabla\phi \to 0$) in combination with Theorem 3.2.

Theorem 3.2 (main result) *Let $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ be an analytic cost function. Let the sequence $\{x_k\}_{k=1,2,\ldots}$ satisfy the strong descent conditions (Definition 3.1). Then, either $\lim_{k\to\infty} \|x_k\| = +\infty$, or there exists a single point $x^* \in \mathbb{R}^n$ such that*

$$\lim_{k\to\infty} x_k = x^*.$$

Proof. Without loss of generality, discard all iterates up to the K iterate and relabel the sequence, such that (10) and (11) hold explicitly on the new sequence. Assume moreover that $\|x_k\| \nrightarrow \infty$, i.e., $\{x_k\}$ has at least one accumulation point x^* in $\mathbb{R}^n$. It is sufficient to show that $\lim_{k\to+\infty} x_k = x^*$ to complete the proof.

For simplicity, we assume without loss of generality that $\phi(x^*) = 0$. If the sequence $\{x_k\}$ is eventually constant (i.e. there exists a K such that $x_k = x_K$ for all $k > K$), then the result follows directly. For the remaining case we remove from the sequence all the x_k's such that $x_{k+1} = x_k$ and we renumber the sequence accordingly. It follows that the new sequence is infinite, never stagnates, and admits the same limit-set as the original sequence. By continuity of ϕ, since x^* is an accumulation point of $\{x_k\}$ and $\phi(x_k)$ is strictly decreasing as a consequence of (11), it follows that

$$\phi(x_0) > \phi(x_1) > \ldots > 0. \tag{12}$$

(Note that this is the only place in this proof where (11) is utilized.)

It then follows from the Łojasiewicz gradient inequality (Lemma 2.1) and the primary descent condition (10) that, in some neighbourhood U of x^*,

$$\phi(x_k) - \phi(x_{k+1}) \geq \sigma \|\nabla\phi(x_k)\| \|x_{k+1} - x_k\| \geq \sigma c |\phi(x_k)|^\mu \|x_{k+1} - x_k\|.$$

That is, since we have shown that $\phi(x_k) > 0$ for all k,

$$\|x_{k+1} - x_k\| \leq \frac{\phi(x_k) - \phi(x_{k+1})}{\sigma c \, (\phi(x_k))^\mu}, \tag{13}$$

provided x_k belongs to U.

Since $\mu \in [0, 1)$, it follows from (12) that $\frac{1}{(\phi(x_k))^\mu} \leq \frac{1}{\phi^\mu}$ for all ϕ in $[\phi(x_{k+1}), \phi(x_k)]$, and therefore

$$\frac{\phi(x_k) - \phi(x_{k+1})}{(\phi(x_k))^\mu} = \int_{\phi(x_{k+1})}^{\phi(x_k)} \frac{1}{(\phi(x_k))^\mu} d\phi \leq \int_{\phi(x_{k+1})}^{\phi(x_k)} \frac{1}{\phi^\mu} d\phi = \frac{1}{1-\mu} \left((\phi(x_k))^{1-\mu} - (\phi(x_{k+1}))^{1-\mu} \right). \tag{14}$$

Substituting (14) into (13) yields

$$\|x_{k+1} - x_k\| \leq \frac{1}{\sigma c (1-\mu)} \left((\phi(x_k))^{1-\mu} - (\phi(x_{k+1}))^{1-\mu} \right). \tag{15}$$

This bound plays a role similar to the bound (5) on the exact derivative obtained in the continuous-time case.

Given $k_2 > k_1$ such that the iterates x_{k_1} up to x_{k_2-1} belong to U, we have

$$\sum_{k=k_1}^{k_2-1} \|x_{k+1} - x_k\| \leq c_1 \left((\phi(x_{k_1}))^{1-\mu} - (\phi(x_{k_2}))^{1-\mu} \right) \leq c_1 (\phi(x_{k_1}))^{1-\mu} \tag{16}$$

where $c_1 = [\sigma c (1-\mu)]^{-1}$. This bound plays the same role as (6).

Now the conclusion comes much as in the proof of Theorem 2.2. Let $r > 0$ be such that $B_r(x^*) \subset U$, where

$$B_r(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| < r\}$$

is the open ball of radius r centered at x^*. Let k_1 be such that $\|x_{k_1} - x^*\| < r/2$ and $c_1 (\phi(x_{k_1}))^{1-\mu} < r/2$. Such a k_1 exists since x^* is an accumulation point and $\phi(x^*) = 0$. Then we show that $x_{k_2} \in B_r(x^*)$ for all $k_2 > k_1$. By contradiction, suppose not, and let K be the smallest $k > k_1$ such that $\|x_K - x^*\| \geq r$. Then x_k remains in U for $k_1 \leq k < K$, so it follows from (16) that $\sum_{k=k_1}^{K-1} \|x_{k+1} - x_k\| \leq c_1 (\phi(x_{k_1}))^{1-\mu} < r/2$. It then follows that $\|x_K - x^*\| \leq \|x_K - x_{k_1}\| + \|x_{k_1} - x^*\| \leq \sum_{k=k_1}^{K-1} \|x_{k+1} - x_k\| + \|x_{k_1} - x^*\| < \frac{r}{2} + \frac{r}{2} \leq r$. But we have supposed that $\|x_K - x^*\| \geq r$, a contradiction.

We have thus shown that, given r sufficiently small, there exists k_1 such that $\|x_{k_2} - x^*\| < r$ for all $k_2 > k_1$. Since $r > 0$ is arbitrary (subject to $B_r(x^*) \subset U$), this means that the whole sequence $\{x_k\}$ converges to x^*, and the proof is complete. (The same conclusion comes by noting that the "length" $\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|$ is finite.) □

3.2 Discussion

We now comment on Theorem 3.2 and propose a few variations and extensions to this result.

3.2.1 C^∞ is not sufficient to guarantee single limit-point convergence

Similar to the continous-time case, it is natural to wonder whether the analyticity assumption on ϕ can be relaxed to indefinite differentiability ($\phi \in C^\infty$). The answer is again negative: as we now show, there exist a sequence $\{x_k\}$ in $\mathbb{R}^n$ and a function ϕ in C^∞ such that $\{x_k\}$ satisfies the strong

descent conditions (Definition 3.1) and nevertheless the limit-set of $\{x_k\}$ contains more than one point of $\mathbb{R}^n$.

Consider the Mexican Hat function (8) and let $x_k = (r_k \cos\theta_k, r_k \sin\theta_k)^T$ with $\theta_k = k\omega$ and $r_k = \sqrt{(\theta_k - 1)/\theta_k}$, so that x_k belongs to the trajectory given by (9). Choose $\omega > 0$ such that ω/π is an irrational number. Then the limit-set of $\{x_k\}$ is the unit circle in $\mathbb{R}^2$. However, f is C^∞ and the primary descent condition (10) is satisfied for $\sigma = \frac{1-e^{-\omega}}{4}$. Indeed, simple manipulations yield

$$\partial_r f(r_k, \theta_k) = -e^{-\frac{1}{1-r_k^2}} \frac{2r_k(1-r_k^2)^2}{4r_k^4+(1-r_k^2)^4}$$
$$\frac{1}{r}\partial_\theta f(r_k, \theta_k) = -e^{-\frac{1}{1-r_k^2}} \frac{4r_k^3}{4r_k^4+(1-r_k^2)^4}$$
$$\|\nabla_x f(x_k)\| = e^{-\theta_k} \frac{2r_k}{4r_k^4+(1-r_k^2)^4} \sqrt{(1-r_k^2)^4 + 4r_k^4}.$$

Thus $\|\nabla_x f(x_k)\| \leq 2e^{-k\omega}$ when r_k is sufficiently close to 1, i.e. when k is sufficiently large. Thus

$$f(x_k) - f(x_{k+1}) = e^{-k\omega} - e^{-(k+1)\omega} = e^{-k\omega}(1-e^{-\omega}) \geq \frac{1-e^{-\omega}}{4} 2e^{-k\omega}\, 2 \geq \frac{1-e^{-\omega}}{4} \|\nabla f(x_k)\| \|x_k - x_{k+1}\|.$$

3.2.2 Ruling out escape to infinity

There are several ways to rule out the case $\lim_{k\to+\infty} \|x_k\| = \infty$ in Theorem 3.2. Convergence results often assume that ϕ has compact sublevel sets, in which case $\{x_k\}$ is bounded. Note also that $\lim_{k\to\infty} \|x_k\| = +\infty$ occurs if and only if $\{x_k\}$ has no accumulation point in $\mathbb{R}^n$.

It is interesting to consider what happens in the close vicinity of a critical point. Proposition 3.3 guarantees that if the iteration starts close enough to a local minimum x^* of ϕ, and if the complementary descent condition (11) is replaced by a termination condition, then the sequence of iterates stays in a neighbourhood of x^*. Strengthening the weak descent condition to condition (18) is required since it is not possible to centre the analysis at an accumulation point as was done in the proof of Theorem 3.2.

Proposition 3.3 (Lyapunov stability of minima) *Let x^* be a (possibly nonstrict) local minimum of the analytic cost-function ϕ. Let*

$$x_{k+1} = F(x_k) \tag{17}$$

be a discrete-time dynamical system satisfying the primary descent condition (10) *and the termination condition*

$$\nabla\phi(x_k) = 0 \Rightarrow \textit{terminate}. \tag{18}$$

Then x^* is Lyapunov-stable for (17)*. That is, given $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\|x_0 - x^*\| \leq \delta \Rightarrow \|x_k - x^*\| \leq \epsilon \textit{ for all } k.$$

Proof. Without loss of generality, we again assume that $\phi(x^*) = 0$. Let U_m be a neighbourhood of x^* such that $\phi(x) \geq \phi(x^*)$ for all $x \in U_m$. Let $U_{\text{Ł}}$ be a neighbourhood of x^* where the Łojasiewicz inequality (Lemma 2.1) holds. Let ϵ be such that $B_\epsilon(x^*) \subset U_m \cap U_{\text{Ł}}$. Let $\delta < \epsilon/2$ be such that $c_1(\phi(x))^{1-\mu} < \epsilon/2$ for all $x \in B_\delta(x^*)$, where $c_1 = [\sigma c(1-\mu)]^{-1}$ and c, μ and σ are the constants appearing in the Łojasiewicz inequality and the primary descent condition (10). Then we show that x_k belongs to $B_\epsilon(x^*)$ and the proof is complete. By contradiction, suppose that x_k eventually leaves $B_\epsilon(x^*)$. Let K be the smallest k such that x_k is not in $B_\epsilon(x^*)$. We dismiss the trivial case where the algorithm terminates. Thus $\nabla\phi(x_k) \neq 0$ for all $k < K$. It follows that $\phi(x_k) > 0$ for all $k < K$, otherwise the assumption on U_m would not hold. The rationale given in the proof of Theorem 3.2 yields that

$$\|x_k - x_0\| \leq c_1(\phi(x_0))^{1-\mu} < \epsilon/2$$

for all $k \leq K$, and it comes from the triangle inequality that $\|x_K - x^*\| < \epsilon$, a contradiction. □

Note that Proposition 3.3 is a Lyapunov stability result and one must prove local attractivity of x^* in addition to Proposition 3.3 to prove asymptotic stability of x^*. It would be sufficient to require additionally, weak convergence of the iterates (i.e. $\nabla\phi(x_k) \to 0$) and that x^* is an isolated stationary point of ϕ. A similar result is given by the Capture Theorem [Ber95].

3.2.3 A stronger result

The proof of Theorem 3.2 does not use the analyticity of ϕ to its full extent. Instead, the proof only requires that ϕ be continuous at the accumulation point x^* and that the Łojasiewicz gradient inequality hold in a neighbourhood of x^*.

There is a large class of functions that are not real analytic but nevertheless satisfy the Łojasiewicz gradient inequality; see, e.g., [Kur98, BDL04]. As an illustration, consider $\phi(x) = f(g(x))$ where g is real analytic and f is C^1. Assume for simplicity that $g(x^*) = 0$ and $f(0) = 0$, so $\phi(x^*) = 0$. Assume moreover that $f'(0) = c_1 > 0$, where f' denotes the first derivative of f. Since $f' \circ g$ is continuous, it follows that there exists a neighbourhood U of x^* such that $f'(g(x)) > \frac{c_1}{2}$ for all $x \in U$. Shrinking U if necessary, it follows from the Łojasiewicz gradient inequality on g that there are constants $c > 0$ and $\mu \in [0, 1)$ such that $\|\nabla\phi(x)\| = |f'(g(x))| \cdot \|\nabla g(x)\| \geq \frac{c_1}{2}\|\nabla g(x)\| \geq \frac{c_1}{2}c|g(x)|^\mu$ for all $x \in U$. Shrinking U further if necessary, since $f'(0) = c_1 > 0$ and $f(0) = 0$ and $f \in C^1$, we have $|g(x)| \geq |f(g(x))|/(2c_1)$ for all $x \in U$. Consequently, $\|\nabla\phi(x)\| \geq \frac{c_1 c}{2(2c_1)^\mu}|\phi(x)|^\mu$ for all $x \in U$, and this is a Lojasiewicz inequality.

It is interesting to consider what would be the weakest general condition on the cost function that would ensure single limit-point convergence of a descent iteration under the strong descent conditions (Definition 3.1). In general, this question is difficult to answer, however, the following result provides the weakest condition on the cost function such that the proof given for Theorem 3.2 applies. Note that the class of functions covered is larger again than those satisfying the Łojasiewicz gradient inequality, shown earlier to be a superset of analytic functions.

Theorem 3.4 *Let x^* be a point of $\mathbb{R}^n$ and let ϕ be a cost function on $\mathbb{R}^n$ continuous at x^*. Assume that there exists a neighbourhood U of x^*, an $\epsilon > 0$ and a nondecreasing strictly positive function $\psi : (0, \epsilon) \to \mathbb{R}$ such that $1/\psi$ is integrable over $(0, \epsilon)$ and*

$$\|\nabla\phi(x)\| \geq \psi(\phi(x) - \phi(x^*))$$

for all x in $\{x \in U : 0 < \phi(x) - \phi(x^*) < \epsilon\}$. Consider a sequence $\{x_k\}$ satisfying the strong descent conditions (Definition 3.1) and assume that x^* is an accumulation point of $\{x_k\}$. Then $\lim_{k\to\infty} x_k = x^*$.

4 Application to classical optimization schemes

In this section, we show that the strong descent conditions (Definition 3.1) hold for a wide variety of numerical optimization methods. Consequently, these methods have single limit-point convergence when the cost function is analytic, or more generally when the conditions of Lemma 3.4 are satisfied. We will successively consider methods of the line-search type and of the trust-region type. References on numerical optimization include [DS83, Fle87, Ber95, NW99, CGT00].

4.1 Convergence of Line-Search Methods

Any line-search method proceeds in two steps. First, the algorithm chooses a search direction p_k from the current iterate x_k. Then the algorithm searches along this direction for a new iterate

$$x_{k+1} = x_k + \alpha_k p_k \tag{19}$$

satisfying some criteria.

We first consider the choice of the search direction p_k. An obvious choice is the steepest-descent direction $p_k = -\nabla\phi(x_k)$, which is often relaxed to a direction p_k satisfying an angle condition

$$\frac{\langle p_k, \nabla\phi(x_k)\rangle}{\|p_k\|\|\nabla\phi(x_k)\|} = \cos\theta_k \leq -\delta < 0, \tag{20}$$

i.e., the angle between p_k and $-\nabla\phi(x_k)$ is bounded away from 90°. A wide variety of optimization schemes obtain the search direction by solving an equation of the form

$$B_k p_k = -\nabla\phi(x_k). \tag{21}$$

In particular, the choice $B_k = \nabla^2\phi(x_k)$, the Hessian of ϕ at x_k, yields the Newton direction. When some approximation of the Hessian is used, p_k is called a quasi-Newton search direction. From (20) and (21), standard manipulations (see e.g. [NW99, p. 45]) yield $\cos\theta_k \geq 1/\kappa(B_k)$ where $\kappa(B_k) = \|B_k\|\,\|B_k^{-1}\|$ is the condition number of B_k. Therefore, for the angle condition (20) to hold true with (21), it is sufficient that the condition number of B_k be bounded.

Now consider the choice of α_k in (19). A very usual condition on α is the first Wolfe condition, also known as the *Armijo condition* (see e.g. [NW99]):

$$\phi(x_k) - \phi(x_{k+1}) \geq -c_1\langle\nabla\phi(x_k), x_{k+1} - x_k\rangle \tag{22}$$

where $c_1 \in (0,1)$ is a constant. The Armijo condition is satisfied for all sufficiently small values of α_k. Therefore, in order to ensure that the algorithm makes sufficient progress, it is usual to require moreover that, for some constant $c_2 \in (c_1, 1)$,

$$\langle\nabla\phi(x_{k+1}), x_{k+1} - x_k\rangle \geq c_2\langle\nabla\phi(x_k), x_{k+1} - x_k\rangle, \tag{23}$$

known as the *curvature condition*. Conditions (22) and (23) are known collectively as the *Wolfe conditions*. Several schemes exist that compute an α_k such that the Wolfe conditions hold; see e.g. [Ber95, NW99].

Theorem 4.1 *(i) Consider the line-search descent algorithm given by* (19)*. Let the algorithm terminate if $\nabla\phi(x_k) = 0$. Assume that the search direction p_k satisfies the angle condition* (20)*. Let the step-size be selected such that the Armijo condition* (22) *holds. Then the strong descent conditions (Definition 3.1) hold.*
(ii) Assume moreover that the cost function ϕ is analytic.
Then either $\lim_{k\to\infty}\|x_k\| = +\infty$, or there exists a single point $x^* \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} x_k = x^*.$$

(iii) In the latter case, if moreover the curvature condition (23) *holds, then x^* is a stationary point of ϕ, i.e.*

$$\nabla\phi(x^*) = 0.$$

Proof. (i) Combining the angle condition (20) and the Armijo condition (22) yields $\phi(x_k) - \phi(x_{k+1}) \geq c_1\delta\|\nabla\phi(x_k)\|\|x_{k+1} - x_k\|$, i.e. the primary descent condition (10) with $\sigma = c_1\delta$. The complementary descent condition (11) is also satisfied: if $\nabla\phi(x_k) = 0$ then the algorithm terminates and if $\nabla\phi(x_k) \neq 0$ then (11) follows from (10).
(ii) Direct from (i) and Theorem 3.2.
(iii) Direct consequence of (ii) and a classical convergence result (proven e.g. in [Fle87, theorem 2.5.1] and [NW99, section 3.2]). □

4.2 Convergence of Trust-Region Methods

Most trust-region methods compute the trust-region step such that the model decrease is at least a fraction of that obtained at the so-called Cauchy point. This condition alone is not sufficient to

guarantee that the primary descent condition (10) hold. However, we show in this section that the strong descent conditions (Definition 3.1) hold under a mild modification of the Cauchy decrease condition.

Before stating the results in Theorem 4.4, we briefly review the underlying principles of trust-region methods. The vast majority of trust-region methods proceed along the following lines. At each iterate x_k, a *model* $m_k(p)$ is built that agrees with $\phi(x_k + p)$ to the first order, that is

$$m_k(p) = \phi(x_k) + \nabla\phi(x_k)^T p + \frac{1}{2} p^T B_k p \tag{24}$$

where B_k is some symmetric matrix. Then the problem

$$\min_{p \in \mathbb{R}^n} m_k(p) \quad \text{s.t.} \quad \|p\| \leq \Delta_k \tag{25}$$

where $\Delta_k > 0$ is the trust-region radius, is solved within some approximation, yielding an update vector p_k. Finally the actual decrease of ϕ is compared with the decrease predicted by m_k in the ratio

$$\rho_k = \frac{\phi(x_k) - \phi(x_k + p_k)}{m_k(0) - m_k(p_k)}. \tag{26}$$

If ρ is exceedingly small, then the model is very bad: the step must be rejected and the trust-region radius must be reduced. If ρ is small but less dramatically so, then the step is accepted but the trust-region radius is reduced. If ρ is close to 1, then there is a good agreement between the model and the function over the step, and the trust-region can be expanded. This can be formalized into the following algorithm (similar formulations are given e.g. in [MS83, CGT00]).

Algorithm 4.2 (Trust Region, see e.g. [NW99]) *Given* $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$, *and* $\eta \in (0, \frac{1}{4})$*:*
for $k = 0, 1, 2, \ldots$
  Obtain p_k, $\|p_k\| < \Delta_k$, *by (approximately) solving* (25)*;*
  Evaluate ρ_k *from* (26)*;*
  if $\rho_k < \frac{1}{4}$
    $\Delta_{k+1} = \frac{1}{4}\|p_k\|$
  else if $\rho_k > \frac{3}{4}$ *and* $\|p_k\| = \Delta_k$
    $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$
  else
    $\Delta_{k+1} = \Delta_k$*;*
  if $\rho_k > \eta$
    $x_{k+1} = x_k + p_k$
  else
    $x_{k+1} = x_k$*;*
end (for).

Trust-region methods essentially differ in the way they approximately solve the trust-region subproblem (25). Most of the algorithms compute a step such that the model decrease is at least a fraction of that obtained at the Cauchy point. By definition, the Cauchy point is the solution p_k^C of the one-dimensional problem

$$p_k^C = \arg\min\{m_k(p) : p = \alpha\nabla\phi(x_k),\ \|p\| \leq \Delta_k\}. \tag{27}$$

The class of methods that ensure a fraction of the Cauchy decrease includes the dogleg method of Powell [Pow70], the double-dogleg method of Dennis and Mei [DM79], the truncated conjugate-gradient method of Steihaug [Ste83] and Toint [Toi81], and the two-dimensional subspace minimization strategy of Byrd *et al.* [BSS88]. These methods have weak convergence properties ($\|\nabla\phi(x_k)\| \to 0$) in general; see e.g. [NW99, theorem 4.8]. Other methods, including the one of Moré and Sorensen [MS83], do even better as they attempt to find a nearly exact solution of the trust-region subproblem (25). In this case a strong limit-point convergence result is available [MS83, theorem 4.13] under some additional hypotheses, including nonsingularity of the Hessian of ϕ at an accumulation point.

Assuming that the cost function ϕ is analytic, we have to check that the strong descent conditions (Definition 3.1) hold in order to apply our main result (Theorem 3.2) and conclude to single limit-point convergence.

The following technical lemma will prove to be useful.

Lemma 4.3 *If p_k^C is the Cauchy point defined in* (27)*, then*

$$m_k(0) - m_k(p_k^C) \geq \frac{1}{2}\|\nabla\phi(x_k)\|\|p_k^C\|.$$

Proof. The Cauchy point p_k^C is given explicitly by (see, e.g., [NW99, eq. (4.8)])

$$p_k^C = -\tau_k \frac{\Delta_k}{\|\nabla\phi(x_k)\|}\nabla\phi(x_k), \tag{28a}$$

where

$$\tau_k = \begin{cases} 1 & \text{if } \nabla\phi(x_k)^T B_k \nabla\phi(x_k) \leq 0; \\ \min(\frac{\|\nabla\phi(x_k)\|^3}{\Delta_k \nabla\phi(x_k)^T B_k \nabla\phi(x_k)}, 1) & \text{otherwise.} \end{cases} \tag{28b}$$

We have

$$m_k(0) - m_k(p_k^C) - \frac{1}{2}\|\nabla\phi(x_k)\|\,\|p_k^C\| = \beta_k \left(1 - \frac{\tau_k \Delta_k}{\|\nabla\phi(x_k)\|^3}\nabla\phi(x_k)^T B_k \nabla\phi(x_k)\right)$$

with $\beta_k := \frac{1}{2}\tau_k\Delta_k\|\nabla\phi(x_k)\|$, thus the claim is equivalent to

$$1 - \frac{\tau_k \Delta_k}{\|\nabla\phi(x_k)\|^3}\nabla\phi(x_k)^T B_k \nabla\phi(x_k) \geq 0,$$

which follows from the definition of τ_k. □

Due to the variety of trust-region methods and the flexibility in the choice of the update direction it is not possible to prove a generic convergence result of the nature of Theorem 4.1. Instead, Theorem 4.4 provides several easily verified conditions for the iterates of Algorithm 4.2 in order that its iterates satisfy the strong descent conditions (Definition 3.1). Once this is verified then the results of Theorem 3.2 apply. Convergence to a critical point again depends on additional weak convergence results for the algorithm considered.

The conditions given in Theorem 4.4 are progressively more restrictive on the iterates of Algorithm 4.2. Condition (B) imposes condition (10) on the model m_k. We show that this in turn implies condition (10) on the cost function ϕ. Condition (C) imposes a fraction of the Cauchy decrease that becomes more restrictive as the ratio $\|p_k\|/\|p_k^C\|$ grows. Condition (D) simply states that the model decrease is at least a fraction of that obtained at the Cauchy point. This condition holds for most of the standard trust region algorithms. However, (D) alone is not sufficient to guarantee single limit-point convergence for analytic ϕ. To see this, consider for example the function in $\mathbb{R}^3$ given by

$$f(x) = (\sqrt{x_1^2 + x_2^2} - 1)^2 + x_3^2$$

which has a symmetry of revolution around the third axis. If B_k is chosen to be singular along the θ direction, then a sequence $\{x_k\}$ can be constructed that satisfies (D) but nevertheless loops endlessly towards the set $\{x : x_1^2 + x_2^2 = 1,\ x_3 = 0\}$. Condition (D) becomes sufficient with complementary conditions, like (E) which imposes a bound on $\|p_k\|/\|p_k^C\|$ or like (F) which imposes that B_k remains positive definite and does not become ill-conditioned.

Theorem 4.4 *Let $\{x_k\}$, $\{\Delta_k\}$, $\{p_k\}$, $\{\phi(x_k)\}$, $\{\nabla\phi(x_k)\}$ and $\{B_k\}$ be infinite sequences generated by Algorithm 4.2 (Trust-Region). Let m_k, ρ_k and p_k^C be defined as in* (24)*,* (26) *and* (27)*, respectively. Consider the following conditions:*

(A) The strong descent conditions (Definition 3.1) hold.

(B) There exists $\sigma_1 > 0$ such that for all k with $\nabla\phi(x_k) \neq 0$

$$m_k(0) - m_k(p_k) \geq \sigma_1 \|\nabla\phi(x_k)\| \|p_k\|. \tag{29}$$

(C) There exists $\sigma_2 > 0$ such that for all k with $\nabla\phi(x_k) \neq 0$

$$\frac{m_k(0) - m_k(p_k)}{m_k(0) - m_k(p_k^C)} \geq \sigma_2 \frac{\|p_k\|}{\|p_k^C\|}. \tag{30}$$

(D) There exists $c_2 > 0$ such that for all k with $\nabla\phi(x_k) \neq 0$

$$m_k(0) - m_k(p_k) \geq c_2(m_k(0) - m_k(p_k^C)). \tag{31}$$

(E) There exists $\kappa_1 > 0$ such that for all k with $\nabla\phi(x_k) \neq 0$

$$\|p_k\| \leq \kappa_1 \|p_k^C\|. \tag{32}$$

(F) B_k is positive definite for all k and there exists a $\kappa_2 \geq 1$ such that $\mathrm{cond}(B_k) := \|B_k\|\|B_k^{-1}\| \leq \kappa_2$ for all k (where the matrix norms are 2-norms).

Then (D and F) $\Rightarrow$ (D and E) $\Rightarrow$ C $\Rightarrow$ B $\Rightarrow$ A. Furthermore, if (A) holds and the cost function ϕ is analytic, then either $\lim_{k\to\infty} \|x_k\| = +\infty$, or there exists a single point $x^* \in \mathbb{R}^n$ such that $\lim_{k\to\infty} x_k = x^*$.

Proof. First note that the condition $\nabla\phi(x_k) \neq 0$ guarantees that $p_k^C \neq 0$ and $m_k(0) - m_k(p_k^C) > 0$.

(D and F) $\Rightarrow$ (D and E). If $\|p_k^C\| = \Delta_k$ then (E) holds with $\kappa_1 = 1$. Assume then that $\|p_k^C\| < \Delta_k$. Let $\lambda_{\max}(B_k)$, resp. $\lambda_{\min}(B_k)$, denote the largest, resp. smallest, eigenvalue of the positive definite matrix B_k. Then

$$\frac{\|\nabla\phi(x_k)\|}{\lambda_{\max}(B_k)} \leq \frac{\|\nabla\phi(x_k)\|^3}{\nabla\phi(x_k)^T B_k \nabla\phi(x_k)} = \|p_k^C\|, \tag{33}$$

where the equality follows from (28) and $\|p_k^C\| < \Delta_k$. In view of (D), one has $m_k(0) - m_k(p_k) \geq 0$ and thus $-\nabla\phi(x_k)^T p_k - \frac{1}{2} p_k^T B_k p_k \geq 0$. Therefore

$$\frac{1}{2}\lambda_{\min}(B_k)\|p_k\|^2 \leq \frac{1}{2} p_k^T B_k p_k \leq -\nabla\phi(x_k)^T p_k \leq \|\nabla\phi(x_k)\| \, \|p_k\|. \tag{34}$$

It follows from (33) and (34) that

$$\|p_k\| \leq 2\frac{\|\nabla\phi(x_k)\|}{\lambda_{\min}(B_k)} \leq 2\frac{\lambda_{\max}(B_k)}{\lambda_{\min}(B_k)}\|p_k^C\| = 2\mathrm{cond}(B_k)\|p_k^C\| \leq 2\kappa_2\|p_k^C\|,$$

i.e., (E) holds with $\kappa_1 := 2\kappa_2$.

(D and E) $\Rightarrow$ C. Direct, with $\sigma_2 = c_2/\kappa_1$.

C $\Rightarrow$ B. Directly follows from Lemma 4.3, with $\sigma_1 = \sigma_2/2$.

B $\Rightarrow$ A. If $x_{k+1} = x_k$, then the strong descent conditions trivially hold. Assume then that $x_{k+1} \neq x_k$, in which case the complementary descent condition (11) holds by the definition of Algorithm 4.2. If $\nabla\phi(x_k) = 0$, then the primary descent condition (10) trivially holds. On the other hand, if $\nabla\phi(x_k) \neq 0$, then it follows from (B) that (10) holds with $\sigma = \eta\sigma_1$ where η is defined in Algorithm 4.2.

The final claim follows directly from Theorem 3.2. □

Convergence of the iterates of Algorithm 4.2 to a critical point depends on additional weak convergence ($\|\nabla\phi(x_k)\| \to 0$) results for the particular algorithm considered. For example, if assumptions (D) and (E) hold, ϕ is analytic, and $\|B_k\| \leq \beta$ for some constant β then either $\lim_{k\to\infty} \|x_k\| = +\infty$, or there exists a single point $x^* \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} x_k = x^* \text{ and } \nabla\phi(x^*) = 0.$$

This follows from the above result along with a classical convergence result for trust-region methods (see [NW99, theorem 4.8]).

5 Conclusion

We have shown strong limit-point convergence results that do not rely on the usual requirement that critical points are isolated. Instead, we require two conditions: the Łojasiewicz gradient inequality (1), i.e. a lower bound on the norm of the gradient of the cost function in terms of the cost function itself, and some 'strong descent conditions' stated in Definition 3.1. The Łojasiewicz gradient inequality is satisfied in particular for analytic cost functions. The strong descent conditions are satisfied for a wide variety of optimization schemes; they include line-search methods with an angle condition on the search direction and Armijo's condition on the step length, and trust-region methods under the condition that the length of the update vector is bounded by a multiple of the length of the Cauchy update.

Acknowledgements

The authors would like to thank all reviewers who have provided comments on this paper during its various revisions. We would particularly like to thank both J.-C. Gilbert and E. W. Sachs for handling the paper as associate editors and for providing excellent suggestions and recommendations to improve the material. In addition, the authors wish to thank K. Kurdyka and A. L. Tits for useful discussions related to this topic.

References

[BDL04] J. Bolte, A. Daniilidis, and A. S. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems generalized eigenvalue problems*, submitted to Transactions of the American Mathematical Society, http://www.orie.cornell.edu/∼aslewis/publications/2004.html, 2004.

[Ber95] D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, Massachusetts, 1995.

[BM88] E. Bierstone and P. D. Milman, *Semianalytic and subanalytic sets*, Inst. Hautes tudes Sci. Publ. Math **67** (1988), 5–42.

[BN89] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal. **26** (1989), no. 3, 727–739.

[BSS88] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *Approximate solution of the trust region problem by minimization over two-dimensional subspaces*, Math. Programming **40** (1988), no. 3, (Ser. A), 247–263.

[CGST93] A. R. Conn, N. Gould, A. Sartenaer, and Ph. L. Toint, *Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints*, SIAM J. Optim. **3** (1993), no. 1, 164–221.

[CGT00] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-region methods*, MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, and Mathematical Programming Society (MPS), Philadelphia, PA, 2000.

[Cur44] H. B. Curry, *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math. **2** (1944), 258–261.

[Ded00] Jean-Pierre Dedieu, *Approximate solutions of analytic inequality systems*, SIAM J. Optim. **11** (2000), no. 2, 411–425 (electronic). MR 2002b:90085

[DM79] J. E. Dennis, Jr. and H. H. W. Mei, *Two new unconstrained optimization algorithms which use function and gradient values*, J. Optim. Theory Appl. **28** (1979), no. 4, 453–482.

[DS83] J. E. Dennis and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice Hall Series in Computational Mathematics, Prentice Hall, Englewood Cliffs, NJ, 1983.

[Dun81] J. C. Dunn, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim. **19** (1981), no. 3, 368–400.

[Dun87] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl. **55** (1987), no. 2, 203–216.

[FFK00] Francisco Facchinei, Andreas Fischer, and Christian Kanzow, *On the identification of zero variables in an interior-point framework*, SIAM J. Optim. **10** (2000), no. 4, 1058–1078 (electronic). MR 2001f:90069

[Fle87] R. Fletcher, *Practical methods of optimization, second edition*, John Wiley & Sons, Chichester, 1987.

[Gol65] A. A. Goldstein, *On steepest descent*, J. Soc. Indust. Appl. Math. Ser. A Control **3** (1965), 147–151.

[HM94] U. Helmke and J. B. Moore, *Optimization and dynamical systems*, Springer, 1994.

[KM96] K. C. Kiwiel and K. Murty, *Convergence of the steepest descent method for minimizing quasiconvex functions*, J. Optim. Theory Appl. **89** (1996), no. 1, 221–226.

[KMP00] K. Kurdyka, T. Mostowski, and A. Parusiński, *Proof of the gradient conjecture of R. Thom*, Ann. of Math. (2) **152** (2000), no. 3, 763–792.

[KP94] Krzysztof Kurdyka and Adam Parusiński, $\mathbf{w}_f$*-stratification of subanalytic functions and the Łojasiewicz inequality*, C. R. Acad. Sci. Paris Sér. I Math. **318** (1994), no. 2, 129–133. MR MR1260324 (95d:32012)

[Kur98] Krzysztof Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble) **48** (1998), no. 3, 769–783. MR 2000b:03139

[Lag02] C. Lageman, *Konvergenz reell-analytischer gradientenähnlicher Systeme*, Diplomarbeit im Fach Mathematik, Mathematisches Institut, Universität Würzburg (Betreuer: Prof. Dr. Uwe Helmke), Januar 2002.

[Łoj59] S. Łojasiewicz, *Sur le problème de la division*, Studia Math. **18** (1959), 87–136. MR 21 #5893

[Loj65] S. Lojasiewicz, *Ensembles semi-analytiques*, Inst. Hautes Études Sci., Bures-sur-Yvette, 1965.

[Loj84] ———, *Sur les trajectoires du gradient d'une fonction analytique*, Seminari di Geometria 1982-1983 (Università di Bologna, Istituto di Geometria, Dipartimento di Matematica), 1984, pp. 115–117.

[Łoj93] Stanislas Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Ann. Inst. Fourier (Grenoble) **43** (1993), no. 5, 1575–1595. MR 96c:32007

[LP94] Zhi-Quan Luo and Jong-Shi Pang, *Error bounds for analytic systems and their applications*, Math. Programming **67** (1994), no. 1, Ser. A, 1–28. MR 96f:90110

[Mah98] R. E. Mahony, *Convergence of gradient flows and gradient descent algorithms for analytic cost functions*, proceedings of the thirteenth International Symposium on the Mathematical Theory of Networks and Systems (MTNS98), Padova, Italy, 1998, pp. 653–656.

[MS83] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput. **4** (1983), 553–572.

[NW99] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research, Springer-Verlag, New York, 1999.

[PdM82] J. Palis, Jr. and W. de Melo, *Geometric theory of dynamical systems*, Springer-Verlag, New York, 1982, An introduction, Translated from the Portuguese by A. K. Manning.

[Pow70] M. J. D. Powell, *A new algorithm for unconstrained optimization*, Nonlinear Programming (Proc. Sympos., Univ. of Wisconsin, Madison, Wis., 1970), Academic Press, New York, 1970, pp. 31–65.

[Ste83] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal. **20** (1983), 626–637.

[Toi81] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, Sparse Matrices and Their Uses (I. S. Duff, ed.), Academic Press, London, 1981, pp. 57–88.

[Wig90] S. Wiggins, *Introduction to applied nonlinear dynamical systems and chaos*, Texts in Applied Mathematics, no. 2, Springer-Verlag, New York, 1990.

[Wol69] Philip Wolfe, *Convergence conditions for ascent methods*, SIAM Rev. **11** (1969), 226–235. MR MR0250453 (40 #3690)

[YDF04] Nobuo Yamashita, Hiroshige Dan, and Masao Fukushima, *On the identification of degenerate indices in the nonlinear complementarity problem with the proximal point algorithm*, Math. Program. **99** (2004), no. 2, Ser. A, 377–397. MR 2 039 046

[Zou76] G. Zoutendijk, *Mathematical programming methods*, North-Holland Publishing Co., Amsterdam, 1976. MR 56 #4718