Convergence of non-smooth descent methods using the Kurdyka-Łojasiewicz inequality

Dominikus Noll *

Abstract

We investigate convergence of subgradient-oriented descent methods in non-smooth non-convex optimization. We prove convergence in the sense of subsequences for functions with a strict standard model, and we show that convergence to a single critical point may be guaranteed if the strong Kurdyka-Łojasiewicz condition is added. We show by way of an example that the Kurdyka-Łojasiewicz inequality alone is not sufficient to prove convergence to critical points.

Key words: Non-smooth non-convex optimization, subgradient-oriented descent method, strict model, Kurdyka-Łojasiewicz inequality, upper C^1 function, lower C^1 function.

1 Introduction

In smooth optimization a sequence of descent directions d_j at iterates x_j is called gradient-oriented if the angle between d_j and the negative gradient $-\nabla f(x_j)$ stays uniformly away from 90°. Convergence of gradient-oriented methods is guaranteed by the Armijo condition in tandem with a safeguard against too small steps (see [14]). Convergence is *apriori* understood in the sense of subsequences, but the work of Absil *et al.* [4], and subsequent generalizations [7, 11, 8, 21], assures *aposteriori* convergence $x_j \to x^*$ to a single critical point x^* if f satisfies the Kurdyka-Łojasiewicz inequality.

Here we investigate whether similar results may be expected in non-smooth optimization. We are premonished by the well-known fact that for convex functions, the steepest descent method converges only when the stepsizes t_k satisfy $\sum_{k=1}^{\infty} t_k = \infty$, $\sum_{k=1}^{\infty} t_k^2 < \infty$, (cf. [5]), a condition which cannot be verified algorithmically in the non-convex case, where linesearch or related mechanisms are required. It turns out that the non-smooth situation is indeed complicated, and not altogether promising. Convergence of subgradient-oriented methods, even when understood in the sense of subsequences, only occurs when f belongs to the specific class $\mathscr{S}$ of non-smooth functions having a strict standard model [20]. For $f \in \mathscr{S}$, convergence can be forced algorithmically, and in tandem with the Kurdyka-Łojasiewicz property, $f \in \mathscr{S}$ assures convergence to a single critical point. We show by way of an example that convergence to a critical point may fail even for tame convex functions $f \notin \mathscr{S}$.

The work of Bolte *et al.* [11] is of interest to us. These authors characterize the Kurdyka-Łojasiewicz property for convex $C^{1,1}$-functions by length boundedness of piecewise gradient iterates, and by the existence of an approximate talweg. We show that in

*Université de Toulouse, Institut de Mathématiques, Toulouse, France

the non-smooth case convergence of discrete subgradient trajectories, or of the talweg, is no longer linked to the Kurdyka-Łojasiewicz inequality. Something else is needed, namely, a function $f \in \mathscr{S}$. For $f \in \mathscr{S}$ we obtain convergence in the sense of subsequences of a variant of the talweg, and convergence to a single critical points when the strong Kurdyka-Łojasiewicz property is added. These results are in contrast with the continuous case [9], where finite length of the subgradient trajectory automatically implies convergence to a critical point.

Our work is also related to Attouch *et al.* [6], where an abstract convergence result under the Kurdyka-Łojasiewicz inequality is proved. We investigate whether the sufficient conditions of these authors can be assured algorithmically. The results are discussed in section 12.

The structure of the paper is as follows. In section 6 we recall the model concept and prove that upper C^1-functions belong to the class $\mathscr{S}$ of functions with a strict standard model. The central result in sections 7 – 9 proves convergence of subgradient-oriented descent methods for functions $f \in \mathscr{S}$. Consequences for the talweg and for discrete gradient trajectories are given in section 11. The abstract descent result of [6] is discussed in section 12. A limiting example is discussed in section 13.

2 Kurdyka-Łojasiewicz inequality

Following [8], we shall say that a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Kurdyka-Łojasiewicz inequality (for short KL-inequality) at $x^* \in \mathbb{R}^n$ if there exists $0 < \eta < \infty$, a neighborhood U of x^*, and a concave function $\kappa : [0, \eta] \to [0, \infty)$ such that

(i) $\kappa(0) = 0$,

(ii) κ is of class C^1 on $(0, \eta)$,

(iii) $\kappa' > 0$ on $(0, \eta)$,

(iv) For every $x \in U$ with $f(x^*) < f(x) < f(x^*) + \eta$ we have

$$\kappa'(f(x) - f(x^*))\mathrm{dist}(0, \partial^L f(x)) \geq 1.$$

Here $\partial^L(f(x)$ is the limiting subdifferential of f at x. We shall say that f satisfies the strong KL-inequality at x^* if the same estimate holds for the Clarke subdifferential $\partial f(x)$. In Bolte *et al.* [10, Thm. 11] it is shown that definable functions satisfy the strong KL-inequality, and this class is expected to cover a large variety of practical cases.

3 Subgradient-oriented descent

The angle condition does no longer describe a useful set of search directions in a non-smooth setting. The reason is that directions allowing descent form in general not a half-space, but a cone with opening angle $< 180°$. That means a direction d with $\angle(d, -g_-) < 90°$, where g_- is the steepest ascent subgradient, need not even allow descent. Fortunately, gradient-orientedness of d_j could also be defined in the following equivalent way: each d_j is the steepest descent direction at x_j with respect to some euclidian norm $\|x\|_j^2 = x^\top P_j x$ on $\mathbb{R}^n$, such that

$$0 < \lambda \leq \lambda_{\min}(P_j) \leq \lambda_{\max}(P_j) \leq \Lambda < \infty \tag{1}$$

for all $j \in \mathbb{N}$ and certain $0 < \lambda \leq \Lambda < \infty$. The charm of this second definition is that it carries over reasonably to the non-smooth case.

Yet another difference between the smooth and the non-smooth case is that the concept of a descent direction depends on the choice of the subdifferential. We therefore avoid it and simply work with *directions which allow descent*, i.e., search directions d where $f(x + td) < f(x)$ for some $t_0 > 0$ and all $0 < t \leq t_0$. Altogether, this leads to

Definition 1. *A sequence d_j of normalized directions allowing descent at $x_j \in \mathbb{R}^n$ is subgradient-oriented if there exist Clarke subgradients $g_j \in \partial f(x_j)$ such that $d_j = -\frac{P_j g_j}{\|P_j g_j\|}$, with the P_j satisfying (1).*

We examine under what conditions subgradient-oriented method for non-smooth optimization convergence. We are interested in conditions which can be guaranteed algorithmically.

4 Discrete gradient-oriented flow

Bolte *et al.* [11] give a characterization of the Kurdyka-Łojasiewicz condition for convex $C^{1,1}$ functions in terms of finite length of discrete gradient flow trajectories. Here discrete gradient flow means sequences x_j of iterates satisfying the strong descent condition

$$\beta\|\nabla f(x_j)\|\|x_{j+1} - x_j\| \leq f(x_j) - f(x_{j+1}). \tag{2}$$

One can observe that if (2) is to hold for *all* points on the segment $[x_j, x_{j+1}]$, then one obtains the condition $\beta\|\nabla f(x_j)\| \leq -\nabla f(x_j)^\top d_j$, which implies $\cos\angle(-\nabla f(x_j), d_j) \geq \beta > 0$. In other words, the sequence of directions d_j is then gradient-oriented in the usual smooth sense.

Here we analyze the non-smooth and non-convex analogue of this result, using our definition 1. That is, we seek algorithmically verifiable conditions assuring convergence of discrete subgradient trajectories. Our results will be compared to [11, 8] in section 11.

5 Abstract descent method

Attouch *et al.* [6] prove convergence of an abstract non-smooth descent method for functions satisfying the KL-inequality. They require their sequence x_j to satisfy the axiom

$$f(x_j) - f(x_{j+1}) \geq a\|x_j - x_{j+1}\|^2 \tag{3}$$

for some $a > 0$, and the existence of $g_{j+1} \in \partial^L f(x_{j+1})$ satisfying

$$\|g_{j+1}\| \leq b\|x_j - x_{j+1}\| \tag{4}$$

for some $b > 0$. While (3) is built rather along the lines of the usual strong descent condition (2) for subgradient-oriented methods, condition (4) is unexpected, and the immediate question is whether it has a chance to be algorithmically verifiable.

Postponing this question, if one believes that (4) makes sense algorithmically, then it is natural to also consider a similar condition rooted at x_j, i.e., there exists $g_j \in \partial^L f(x_j)$ and $b > 0$ such that

$$\|g_j\| \leq b\|x_j - x_{j+1}\|. \tag{5}$$

In our framework we can explain why both (4), (5) are difficult to force algorithmically. We show in section 13 by way of an example that (4) and (5) both fail for a convex tame function $f \notin \mathscr{S}$.

On the positive side, our approach also shows that for functions $f \in \mathscr{S}$, conditions like (4) and (5) are not even needed. All that is required to prove convergence is a descent condition in the spirit of (3), in tandem with a diligent backtracking strategy (Theorem 2). We will get back to this interesting line in section 12.

6 The model concept

Given a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, we call $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ a *first-order model* of f if $\phi(\cdot, x)$ is convex for every $x \in \mathbb{R}^n$ and the following axioms are satisfied:

(M_1) $\phi(x, x) = f(x)$ and $\partial_1\phi(x, x) \subset \partial f(x)$.

(M_2) For every x and every $\epsilon > 0$ there exists $\delta > 0$ such that $f(y) \leq \phi(y, x) + \epsilon\|y - x\|$ whenever $\|y - x\| \leq \delta$.

(M_3) ϕ is jointly upper semi-continuous, i.e., $(y_j, x_j) \to (y, x)$ implies $\limsup_{j\to\infty} \phi(y_j, x_j) \leq \phi(y, x)$.

One notices the similarity of this concept with the Taylor expansion of differentiable functions, which is corroborated by the fact that every locally Lipschitz function has a first-order model, which we call the standard model,

$$\phi^\sharp(y, x) = f(x) + f^\circ(x, y - x).$$

Here $f^\circ(x, d)$ is the Clarke directional derivative of f at x in direction d. For C^1-functions $\phi^\sharp(y, x) = f(x)+\nabla f(x)(y-x)$ reproduces indeed the Taylor expansion. There is, however, a major difference between the Taylor expansion and the model concept. Taylor expansion wants herself to be unique. The idea of the model concept is the opposite. We wish a given function to have as many models as possible, because every model leads to a different optimization method.

Definition 2. *A first-order model ϕ for the locally Lipschitz function f is called strict at $\bar{x} \in \mathbb{R}^n$ if the following strict version of axiom (M_2) is satisfied:*

($\widehat{M_2}$) *For every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$f(y) \leq \phi(y, x) + \epsilon\|y - x\|$$

for all $y, x \in B(\bar{x}, \delta)$. The model ϕ is called strict if it is strict at every $\bar{x}$.

Definition 3. *A first-order model ϕ for the locally Lipschitz function f is called strong at $\bar{x}$ if the following even stronger version of (M_2) is satisfied*

($\widetilde{M_2}$) *For every $\epsilon > 0$ there exists $\delta > 0$ and $L > 0$ such that*

$$f(y) \leq \phi(y, x) + L\|y - x\|^2$$

for all $x, y \in B(\bar{x}, \delta)$. The model ϕ is called strong if it is strong at every $\bar{x}$.

Remark 1. One notices the resemblance of $(\widehat{M}_2)$ with the Taylor-Young formula, and that of $(\widetilde{M}_2)$ with the Taylor-Lagrange formula.

Remark 2. Notice that if a model is strong at $\bar{x}$, then it is also strong for every $\tilde{x}$ in a neighborhood of $\bar{x}$. Strong models are strict, and strict models are models, but none of these is invertible. For instance, if f is of class $C^{1,1}$, then its Taylor expansion is strong, while it is strict if f is of class C^1. So for $f \in C^1 \setminus C^{1,1}$ the Taylor expansion is strict but not strong. If we consider $f(x) = x^2 \sin(x^{-1})$ with $f(0) = 0$, then $\phi^\sharp$ is a model, which is not strict at $x = 0$.

We recall from Spingarn [24] and Rockafellar and Wets [23] that a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ is lower C^k at x_0 if there exists a compact space K and a continuous function $F : B(x_0, \delta) \times K \to \mathbb{R}$ for which all partial derivatives of order $\leq k$ are continuous, such that

$$f(x) = \max_{y \in K} F(x, y), \qquad x \in B(x_0, \delta).$$

f is called lower C^k if it is lower C^k at every x. Recall that lower C^2 functions are already lower C^k for every $k \geq 2$ (cf. [23]), but the class of lower C^1 functions is strictly larger than the class of lower C^2 functions. Finally, we call f upper C^k if $-f$ is lower C^k.

Proposition 1. *(Cf. [18]). Let f be locally Lipschitz. If f is upper C^1, then its standard model $\phi^\sharp$ is strict, and if f is upper C^2, then $\phi^\sharp$ is strong.*

Proof: 1) Let f be upper C^1 at $\bar{x}$. Let $\epsilon > 0$. According to Daniilidis and Georgiev [12] there exists $\delta > 0$ such that $-f(tx + (1-t)y)) \leq -tf(y) - (1-t)f(x) + \epsilon t(1-t)\|x - y\|$ for all $x, y \in B(\bar{x}, \delta)$ and $0 \leq t \leq 1$. This can be re-arranged as

$$f(y) \leq f(x) + t^{-1}\left(f(x + t(y - x)) - f(x)\right) + \epsilon(1-t)\|x - y\|.$$

Taking the limit superior $t \to 0^+$ readily implies $f(y) \leq f(x) + f^\circ(x, y - x) + \epsilon\|x - y\| = \phi^\sharp(y, x) + \epsilon\|x - y\|$, hence strictness of $\phi^\sharp$ at $\bar{x}$.

2) The proof of the upper C^2 case is similar. □

Remark 3. Having a strict standard model $\phi^\sharp$ seems a weaker property than upper C^1. Indeed, from Spingarn [24] we know that upper C^1 at $\bar{x}$ is equivalent to the following: For every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in B(\bar{x}, \delta)$ and *every* $g \in \partial f(x)$ one has $-f(y) + f(x) \geq g^\top(y - x) - \epsilon\|y - x\|$. In contrast, for strictness of $\phi^\sharp$ it suffices that this be true for *some* $g \in \partial f(x)$. We may represent this in a more compact form as follows: f is upper C^1 at $\bar{x}$ iff for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, x + td \in B(\bar{x}, \delta)$, $\|d\| = 1$, $t > 0$, we have

$$\frac{f(x + td) - f(x)}{t} \leq -f^\circ(x, -d) + \epsilon,$$

whereas strictness of the standard model replaces this by the formally weaker

$$\frac{f(x + td) - f(x)}{t} \leq f^\circ(x, d) + \epsilon.$$

Remark 4. If f is convex then $\phi(\cdot, x) = f$ is a strong model in the sense of Definition 3. We say that a convex function is its own strong model. Since f has also a standard model $\phi^\sharp$, we see that a function f will in general have several models.

Remark 5. Every convex composite function $f = g \circ F$ with g convex and F of class C^1 has the natural strict model $\phi(\cdot, x) = g(F(x) + F'(x)(\cdot - x))$. Convergence theory for natural models was developed in [2, 3, 20].

In the same vein, if f is lower C^2, then given a bounded convex set B, we can find $\mu > 0$ such that for every $x \in B$, $f + \mu\| \cdot -x\|^2$ is convex on B. In consequence, $\phi(\cdot, x) = f+\mu\|\cdot-x\|^2$ is a strong first-order model of f on B. Notice that descent directions based on the natural model or the lower C^2 model are in general not subgradient-oriented, and it is open whether the KL-property applies in this setting.

The model concept extends even to lower C^1-function:

Proposition 2. *Let f be locally Lipschitz and lower C^1. Then f has a strict first-order model on every bounded set B.*

Proof: Let $B = B(0, M)$ be a bounded disk. We construct a strict model $\phi(\cdot, x)$ for the $x \in B$. For $y \in B$ and $g \in \partial f(y)$ let $t_{y,g}(\cdot) = f(y) + g^\top(\cdot - y)$ be a tangent to f at y. We define the downshift of $t_{y,g}(\cdot)$ with respect to x as

$$s = s(x, y, g) = [t_{y,g}(x) - f(x)]_+ + c\|y - x\|^2,$$

where $c > 0$ is a fixed constant. Then we put

$$m_{y,g}(\cdot, x) = t_{y,g}(\cdot) - s(x, y, g),$$

which we call the downshifted tangent. Now we define

$$\phi(\cdot, x) = \max\{m_{y,g}(\cdot, x) : y \in B, g \in \partial f(y)\},$$

and one can check that ϕ is indeed a strict first-order model of f at every $x \in B$. For a more detailed proof see [18]. □

We have applied bundling techniques to lower C^1 functions quite successfully in the context of automatic control. We refer to [20, 18, 1, 19, 13] for theoretical material supporting this branch of the theory. One may observe that for lower C^1 functions the standard model $\phi^\sharp$ is not the best choice, as more natural strict models ϕ or oracles in the sense of [18] are available.

7 Descent step finding

In this section we discuss the question how to compute a subgradient-oriented descent step. The difficulty may be condensed to the observation that if $g \in \partial f(x)$, then due to non-smoothness, $-g$ may not necessarily allow descent. The directions allowing descent form a cone, not a half-space, and it is therefore harder to find one pointing into this cone. As the general theme of our work is the analysis of gradient-oriented methods, we will during the following work with the standard model $\phi^\sharp$, even though some of the results hold for more general models.

A function $\phi_k^\sharp : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a *first-order working model* if $\phi_k^\sharp(\cdot, x)$ is convex, $\phi_k^\sharp(\cdot, x) \leq \phi^\sharp(\cdot, x)$, $\phi_k^\sharp(x, x) = \phi^\sharp(x, x) = f(x)$, and $\partial_1\phi_k^\sharp(x, x) \subset \partial_1\phi^\sharp(x, x) = \partial f(x)$. Working models are maintained and updated iteratively during the inner loop (algorithm 1) with counter k by adding cutting planes. Here cutting planes means tangents to $\phi^\sharp(\cdot, x)$

at the various null steps y^k. In other words, due to the specific structure of $\phi^\sharp$, each $\phi_k^\sharp$ has the form

$$\phi_k^\sharp(\cdot, x) = \sup_{g \in \mathcal{G}_k} f(x) + g^\top(\cdot - x)$$

for a suitable $\mathcal{G}_k \subset \partial f(x)$. Notice that the standard model itself has the same structure with $\mathcal{G} = \partial f(x)$, i.e.,

$$\phi^\sharp(\cdot, x) = \sup_{g \in \partial f(x)} f(x) + g^\top(\cdot - x),$$

which guarantees $\phi_k^\sharp \leq \phi^\sharp$ and $\partial_1 \phi_k^\sharp(x, x) \subset \partial_1 \phi^\sharp(x, x)$. As we shall see, the management of the sets $\mathcal{G}_k$ during the inner loop has to respect two basic rules, referred to as cutting planes and aggregation, which we proceed to explain.

Given the current working model $\phi_k^\sharp(\cdot, x)$ at x, the step-finding algorithm computes the solution y^k of the tangent program

$$\min_{y \in \mathbb{R}^n} \sup_{g \in \mathcal{G}_k} f(x) + g^\top(y - x) + \tfrac{1}{2t_k}\|y - x\|_P, \tag{6}$$

where $\|x\|_P^2 = x^\top P x$ is an euclidian norm fixed during the inner loop at x. The solution y^k of (6) is called the trial step, t_k is called the stepsize, while $t_k^{-1} > 0$ is sometimes referred to as the proximity control parameter. The necessary optimality condition for (6) implies

$$0 \in \partial_1 \phi_k^\sharp(y^k, x) + t_k^{-1} P(y^k - x),$$

or what is the same,

$$g_k^* := t_k^{-1} P(x - y^k) \in \partial_1 \phi_k^\sharp(y^k, x). \tag{7}$$

We call g_k^* the aggregate subgradient and $f(x) + g_k^{*\top}(\cdot - x)$ the aggregate plane at y^k. Notice that $g_k^* \in \partial f(x)$ due to the specific structure of $\phi^\sharp$.

Having computed a trial step y^k by solving (6), we test acceptance by computing the test parameter

$$\rho_k = \frac{f(x) - f(y^k)}{f(x) - \phi_k^\sharp(y^k, x)}.$$

We say that y^k satisfies the descent condition if $\rho_k \geq \gamma$. If this is the case, we accept $x^+ = y^k$ as the new serious iterate, and the step finding algorithm terminates successfully. On the other hand, if $\rho_k < \gamma$, then we call y^k a null step. In this case the inner loop has to continue, and this requires improving the working model by modifying the set $\mathcal{G}_k$, and possibly by shortening the stepsize t_k. In the case of a null step y^k, we compute $g_k \in \partial f(x)$ such that $\phi^\sharp(y^k, x) = f(x) + g_k^\top(y^k - x)$ and include g_k in the new set $\mathcal{G}_{k+1}$. Moreover, we also include the aggregate subgradient g_k^* in the set $\mathcal{G}_{k+1}$.

Remark 6. Notice that our test $\rho_k > \gamma$ replaces the descent conditions (2) and (3).

We mention two specific ways of constructing the working model $\phi_k^\sharp$. The first option is to maintain a finite set $\mathcal{G}_k = \{g_0, \ldots, g_{k-1}\}$, where at each step k the new cutting plane g_k is added. In this case the tangent program has the simple form

$$\min_{y \in \mathbb{R}^n} \max_{i=0,\ldots,k-1} f(x) + g_i^\top(y - x) + \tfrac{1}{2t_k}\|y - x\|_P^2, \tag{8}$$

Algorithm 1. Descent step-finding by backtracking.

Input: Current iterate x. **Output:** Serious iterate x^+.
Parameters: $0 < \gamma < \widetilde{\gamma} < 1$, $0 < \theta < \Theta < 1$.

1: **Initialize.** Put counter $k = 1$, fix $t_1 > 0$ and $g_0 \in \partial f(x)$. Put $\mathcal{G}_1 = \{g_0\}$.

2: **Tangent program.** Given $t_k > 0$, the current $\mathcal{G}_k \subset \partial f(x)$ and working model $\phi_k^\sharp(\cdot, x) = f(x) + \max_{g \in \mathcal{G}_k} g^\top(\cdot - x)$, compute solution y^k of the tangent program

$$(TP) \quad \min_{y \in \mathbb{R}^n} \phi_k^\sharp(y, x) + \tfrac{1}{2t_k}\|y - x\|_P^2.$$

3: **Acceptance test.** Compute

$$\rho_k = \frac{f(x) - f(y^k)}{f(x) - \phi_k^\sharp(y^k, x)}.$$

If $\rho_k \geq \gamma$, then put $x^+ = y^k$ and quit successfully with new serious step. Otherwise, if $\rho_k < \gamma$, goto step 4.

4: **Cutting plane.** Pick a subgradient $g_k \in \partial f(x)$ such that $f(x) + g_k^\top(y^k - x) = \phi^\sharp(y^k, x)$, or equivalently, $f^\circ(x, y^k - x) = g_k^\top(y^k - x)$. Include g_k into the new set $\mathcal{G}_{k+1}$ for the next sweep.

5: **Aggregate plane**. Include the aggregate subgradient g_k^* in the new set $\mathcal{G}_{k+1}$, and allow the inclusion of additional subgradients from $\partial f(x)$.

6: **Step management.** Compute the test quotient

$$\widetilde{\rho}_k = \frac{f(x) - \phi^\sharp(y^k, x)}{f(x) - \phi_k^\sharp(y^k, x)}.$$

If $\widetilde{\rho}_k \geq \widetilde{\gamma}$, then select $t_{k+1} \in [\theta t_k, \Theta t_k]$. On the other hand, if $\widetilde{\rho}_k < \widetilde{\gamma}$ then keep $t_{k+1} = t_k$. Increment counter k and go back to step 2.

which can be converted to a convex quadratic program. Here the aggregate subgradient has the form

$$g_k^* = \sum_{i=0}^{k-1} \lambda_i g_i, \quad \lambda_i \geq 0, \quad \sum_{i=0}^{k-1} \lambda_i = 1,$$

and the $g_i \in \mathcal{G}_k$ with $\lambda_i > 0$ are said to be called by the aggregate g_k^*. Including g_k^* in the set $\mathcal{G}_{k+1}$ allows us to drop older elements of $\mathcal{G}_k$, so that the size of $\mathcal{G}_k$ can be limited.

The second case of interest is when $\phi_k^\sharp = \phi^\sharp$ for all k. Here the test quotient $\widetilde{\rho}_k$ has constant value 1, so we always reduce the stepsize in case of a null step. Adding cutting planes and aggregate planes has no effect, because they are already included in $\mathcal{G} = \partial f(x)$. The only action taken by the algorithm is backtracking. The solution y^k of the tangent program now has the specific form $y^k = x - t_k P g_-$, where $g_- \in \partial f(x)$ is the projection of 0 onto $\partial f(x)$ with respect to the euclidian norm $\|\cdot\|_P$. In other words, this case covers all non-smooth subgradient-oriented descent method with backtracking linesearch in the sense of definition 1.

Theorem 1. *Let f be locally Lipschitz. Suppose $0 \notin \partial f(x)$. Then after a finite number of trials k the descent step-finding algorithm locates $g_k \in \partial f(x)$ and a stepsize $t_k > 0$ such that $x^+ = x - t_k P^{-1} g_k$ satisfies the descent condition $\rho_k \geq \gamma$.*

Proof: 1) We assume, contrary to what is claimed, that the algorithm turns infinitely, generating a sequence y^k of trial points which all fail the acceptance test. That means $\rho_k < \gamma$ for all $k \in \mathbb{N}$. According to step 5 of the algorithm the step size t_k is either kept invariant, or reduced by a factor $\theta < 1$, but it is never increased. We have therefore two cases. Case 1 is when $t_k \to 0$, case 2 is when t_k is bounded away from 0. The latter means $t_k = t_{k_0}$ for some $k_0 \in \mathbb{N}$ and all $k \geq k_0$. In both cases we will have to achieve a contradiction with the hypothesis $0 \notin \partial f(x)$. Since the norm $\|\cdot\|_P$ is fixed during the entire proof, we may without loss assume $P = I$ during the following in order to simplify the notations.

2) Let us start with the case where $t_k \to 0$. By step 5 of the algorithm the step size t_k is only reduced to $t_{k+1} \leq \Theta t_k < t_k$ when $\widetilde{\rho}_k \geq \widetilde{\gamma}$. We deduce that there exists an infinite subset $\mathcal{K}$ of $\mathbb{N}$ such that $\widetilde{\rho}_k \geq \widetilde{\gamma}$ for every $k \in \mathcal{K}$.

By the optimality condition (7) we have $\|y^k - x\| = t_k \|g_k^*\|$, where g_k^* is the k^{th} aggregate subgradient. Since $\|g_k^*\| \leq \max\{\|g\| : g \in \partial f(x)\} < \infty$, we must have $y^k \to x$. Using the fact that g_k^* is a subgradient of $\phi^\sharp(\cdot, x)$ at y^k, the subgradient inequality implies

$$g_k^{*\top}(x - y^k) \leq \phi_k^\sharp(x, x) - \phi_k^\sharp(y^k, x),$$

which in view of $\phi_k^\sharp(x, x) = f(x)$ and $g_k^{*\top}(x - y^k) = \|g_k^*\|\|x - y^k\|$ gives

$$\|g_k^*\|\|x - y^k\| \leq f(x) - \phi_k^\sharp(y^k, x) \leq g_0^\top(x - y^k),$$

the latter because $f(x) + g_0^\top(y - x) \leq \phi_k^\sharp(y, x)$ for all k. We deduce $\phi_k^\sharp(y^k, x) \to f(x)$. Moreover, since $0 \notin \partial f(x)$, we have $\|g_k^*\| \geq \eta > 0$ for some $\eta > 0$ and all k. That shows

$$f(x) - \phi_k^\sharp(y^k, x) \geq \eta \|x - y^k\| \tag{9}$$

for all k.

Next observe that $\phi_{k+1}^\sharp(y^k, x) = \phi^\sharp(y^k, x)$ by step 4 of the algorithm. Therefore, expanding the control parameter $\widetilde{\rho}_k$, gives

$$\begin{aligned} \widetilde{\rho}_k &= \rho_k + \frac{f(y^k) - \phi^\sharp(y^k, x)}{f(x) - \phi_k^\sharp(y^k, x)} \\ &\leq \rho_k + \frac{\epsilon_k \|x - y^k\|}{\eta \|x - y^k\|} = \rho_k + \epsilon_k/\eta \end{aligned}$$

for a sequence $\epsilon_k \to 0^+$. Here we use (9) and the fact that $y^k \to x$ and $f(y) - \phi^\sharp(y, x) \leq o(\|y - x\|)$ as $y \to x$ by the definition of the Clarke subdifferential. But $\rho_k < \gamma$ for all k and $\epsilon_k/\eta \to 0$, hence $\widetilde{\rho}_k \leq \gamma + \epsilon_k/\eta < \widetilde{\gamma}$ for all k large enough. This contradicts $\widetilde{\rho}_k \geq \widetilde{\gamma}$ for the infinitely many $k \in \mathcal{K}$, and settles the case where $t_k \to 0$.

3) Let us now consider the case where the step size t_k is frozen from some counter k_0 onwards, i.e., $t_k =: t > 0$ for all $k \geq k_0$. That means $\widetilde{\rho}_k < \widetilde{\gamma}$ for all $k \geq k_0$. As in part 2) of the proof we wish to show $y^k \to x$, but with the step size frozen this turns out more complicated to verify.

Let us introduce the objective function $\psi_k(\cdot, x) = \phi_k^\sharp(\cdot, x) + \frac{1}{2t}\|\cdot - x\|^2$ of program (6) for $k \geq k_0$. We know that $\phi_k^\sharp(y^k, x) = f(x) + g_k^{*\top}(y^k - x)$ by step 4 of the algorithm. Therefore

$$\psi_k(y^k, x) = f(x) + g_k^{*\top}(y^k - x) + \tfrac{1}{2t}\|y^k - x\|^2.$$

We define the quadratic function

$$\psi_k^*(\cdot, x) = f(x) + g_k^{*\top}(\cdot - x) + \tfrac{1}{2t}\|\cdot - x\|^2$$

then

$$\psi_k(y^k, x) = \psi_k^*(y^k, x) \tag{10}$$

by what we have just seen. By the definition of the aggregate subgradient we have $f(x) + g_k^{*\top}(\cdot - x) \leq \phi_{k+1}^\sharp(\cdot, x)$, so that

$$\psi_k^*(\cdot, x) \leq \psi_{k+1}(\cdot, x). \tag{11}$$

Notice that $\nabla\psi_k(y, x) = g_k^* + t^{-1}(y - x)$, so that $\nabla\psi_k^*(y^k, x) = g_k^* + t^{-1}(y^k - x) = 0$ by (7). This proves the representation

$$\psi_k^*(y, x) = \psi_k^*(y^k, x) + \tfrac{1}{2t}\|y - y^k\|^2. \tag{12}$$

Now we have

$$\begin{array}{rcll} \psi_k(y^k, x) & \leq & \psi_k^*(y^k, x) + \frac{1}{2t}\|y^k - y^{k+1}\|^2 & \text{(using (10))} \\ & = & \psi_k^*(y^{k+1}, x) & \text{(using (12))} \\ & \leq & \psi_{k+1}(y^{k+1}, x) & \text{(using (11))} \\ & \leq & \psi_{k+1}(x, x) & (y^{k+1} \text{ minimizer of } \psi_{k+1}) \\ & = & \phi_{k+1}^\sharp(x, x) = f(x). & \end{array} \tag{13}$$

Therefore the sequence $\psi_k(y^k, x)$ is monotonically increasing and bounded above by $f(x)$, and converges to a value $\psi^* \leq f(x)$. This shows that $\frac{1}{2t}\|y^k - y^{k+1}\|^2$ is sandwiched in between two terms with the same limit, ψ^*, hence $\|y^k - y^{k+1}\| \to 0$. As $\|\cdot\|$ is euclidian and the sequence y^k is bounded, we deduce

$$\|y^k - x\|^2 - \|y^{k+1} - x\|^2 \to 0. \tag{14}$$

Now using both convergence results (14) and $\psi_k(y^k, x) \to \psi^*$, we deduce

$$\phi_{k+1}^\sharp(y^{k+1}, x) - \phi_k^\sharp(y^k, x) = \tag{15}$$

$$\psi_{k+1}(y^{k+1}, x) - \psi_k(y^k, x) - \tfrac{1}{2t}\|y^{k+1} - x\|^2 + \tfrac{1}{2t}\|y^k - x\|^2 \to 0. \tag{16}$$

Now recall that $f(x) + g_k^\top(\cdot - x)$ is an affine support function of $\phi_{k+1}^\sharp(\cdot, x)$ at y^k. By the subgradient inequality we obtain

$$g_k^\top(y - y^k) \leq \phi_{k+1}^\sharp(y, x) - \phi_{k+1}^\sharp(y^k, x).$$

Since $\phi_{k+1}^\sharp(y^k, x) = \phi^\sharp(y^k, x)$, we have

$$\phi^\sharp(y^k, x) + g_k^\top(y - y^k) \leq \phi_{k+1}^\sharp(y, x). \tag{17}$$

Now we expand

$$\begin{array}{rcl} 0 & \leq & \phi^\sharp(y^k, x) - \phi_k^\sharp(y^k, x) \\ & = & \phi^\sharp(y^k, x) + g_k^\top(y^{k+1} - y^k) - \phi_k^\sharp(y^k, x) - g_k^\top(y^{k+1} - y^k) \\ & \leq & \phi_{k+1}^\sharp(y^{k+1}, x) - \phi_k^\sharp(y^k, x) + \|g_k\|\|y^{k+1} - y^k\| \qquad \text{(using (17)).} \end{array}$$

But the last term converges to 0 as a consequence of (15), boundedness of the g_k, and $y^{k+1} - y^k \to 0$. This proves

$$\phi^\sharp(y^k, x) - \phi_k^\sharp(y^k, x) \to 0. \tag{18}$$

Let us argue that $\phi_k(y^k,x)\to f(x)$. If this is not the case, then $\liminf_{k\to\infty}\phi_k^\sharp(y^k,x)=f(x)-\eta$ for some $\eta>0$. Choose $\delta>0$ such that $\delta<(1-\widetilde{\gamma})\eta$. From (18) we know that $\phi^\sharp(y^k,x)-\delta\le\phi_k^\sharp(y^k,x)$ for all $k\ge k_1$ and some $k_1\ge k_0$. Using $\widetilde{\rho}_k\le\widetilde{\gamma}$ for all $k\ge k_0$ in tandem with $f(x)>\phi_k^\sharp(y^k,x)$ gives $f(x)-\phi_{k+1}^\sharp(y^k,x)\le\widetilde{\gamma}\left(f(x)-\phi_k^\sharp(y^k,x)\right)$. Therefore

$$\widetilde{\gamma}\left(\phi_k^\sharp(y^k,x)-f(x)\right)\le\phi_{k+1}^\sharp(y^k,x)-f(x)=\phi^\sharp(y^k,x)-f(x)\le\phi_k^\sharp(y^k,x)+\delta-f(x).$$

Passing to the limit for a subsequence realizing the limit inferior gives

$$-\widetilde{\gamma}\eta\le-\eta+\delta,$$

contradicting the choice of δ. This shows $\eta=0$ and proves $\phi_k^\sharp(y^k,x)\to f(x)$. From (18) we immediately deduce $\phi^\sharp(y^k,x)\to f(x)$. And this is from where we now deduce $y^k\to x$. Indeed, from estimate (13) we get

$$\psi_k(y^k,x)=\phi_k(y^k,x)+\tfrac{1}{2t}\|y^k-x\|^2\le\psi^*\le f(x),$$

so that $\phi_k(y^k,x)\to f(x)$ shows $\frac{1}{2t}\|y^k-x\|^2\to 0$ and also $\psi^*=f(x)$.

To finish the proof, let us now achieve a contradiction by showing $0\in\partial f(x)$. Namely, by the subgradient inequality,

$$\begin{aligned}t^{-1}(x-y^k)^\top(y-y^k)&\le\phi_k^\sharp(y,x)-\phi_k^\sharp(y^k,x)\\&\le\phi^\sharp(y,x)-\phi_k^\sharp(y^k,x)\qquad(\text{using }\phi_k^\sharp\le\phi^\sharp)\end{aligned}$$

Passing to the limit gives

$$0\le\phi^\sharp(y,x)-f(x)=\phi^\sharp(y,x)-\phi^\sharp(x,x).$$

As this is true for every y, we obtain $0\in\partial_1\phi^\sharp(x,x)\subset\partial f(x)$, the desired contradiction. That settles the proof of case 2. □

Remark 7. The result tells us that if we are not able to find a step which allows descent at x, then $0\in\partial f(x)$. However, the converse is not true. Even when $0\in\partial f(x)$, we may still be able to find a descent step. Take for instance $f(x)=-|x|$ at $x=0$. Then $0\in\partial f(x)$. Yet, if we initialize the step finding algorithm with $g_0=1$, then we will immediately get a descent step which passes the acceptance test $\rho_k{\ge}\gamma$ in step 3. It is therefore recommended to initialize the algorithm with a limiting subgradient $g_0\in\partial^L f(x)$. When $0\notin\partial^L f(x)$, we increase our chances of finding a step allowing descent.

Remark 8. The above algorithm requires a method to compute $g\in\partial f(x)$ where the maximum $g^\top d=f^\circ(x,d)=\max\{h^\top d:h\in\partial f(x)\}$ is attained for a given d. The existence of such an oracle is a realistic hypothesis.

As an illustration consider eigenvalue optimization, where $f(x)=\lambda_1\left(F(x)\right)$ with $F:\mathbb{R}^n\to\mathbb{S}^m$ of class C^1 and $\lambda_1:\mathbb{S}^m\to\mathbb{R}$ the maximum eigenvalue function on the space $\mathbb{S}^m$ of $m\times m$ symmetric matrices. Then $\partial f(x)=F'(x)^*\partial\lambda_1\left(F(x)\right)$ is computable as soon as $F'(x)^*$ is, because computation of $\partial\lambda_1(X)$ is well-known. More precisely,

$$f^\circ(x,d)=\lambda_1'(X,D)=\lambda_1\left(Q^\top DQ\right),$$

where $X=F(x)\in\mathbb{S}^m$, $D=F'(x)d\in\mathbb{S}^m$, and where the columns of Q form an orthonormal basis of the maximum eigenspace of X. Then $G=QQ^\top\in\partial\lambda_1(X)$ attains the maximum $\lambda_1\left(Q^\top DQ\right)=G\bullet D$, and $g=F'(x)^*G$ therefore attains $f^\circ(x,d)=g^\top d$.

8 Algorithm

In this section we state the main algorithm formally, and give a few comments on its rationale. Recall first that the step finding algorithm 1 combines successive improvement of the working model, achieved by adding cutting planes, with occasional backtracking steps, $t_{k+1} = t_k$ or $t_{k+1} \in [\theta t_k, \Theta t_k]$. This means that in the inner loop (algorithm 1) the stepsize is never increased. Therefore, in the outer loop, we allow the stepsize $t_{j+1}^\sharp = \theta^{-1} t_k$ to increase if acceptance gives a good ratio $\rho_k \geq \Gamma$. If acceptance gives a ratio $\gamma \leq \rho_k < \Gamma$, then me memorize the last stepsize used.

We stress that the following algorithm contains the steepest descent method, and all subgradient-oriented descent methods in the sense of definition 1, as special cases. On the other hand it is more general because it allows to approximate these methods by an iterative technique. This is beneficial in practical situations, where the full subdifferential $\partial f(x)$ is inaccessible to direct computation.

Algorithm 2. Subgradient-oriented descent method.

Parameters: $0 < \gamma < \widetilde{\gamma} < 1$, $0 < \gamma < \Gamma < 1$, $0 < \theta < \Theta < 1$, $0 < c < C < \infty$, $0 \leq \underline{t} < \overline{t} \leq \infty$.

1: **Initialize.** Put counter $j = 1$, choose initial guess x^1, and fix $t_1^\sharp > 0$. Choose an euclidian norm $\|x\|_1^2 = x^\top P_1 x$ such that $c\|\cdot\| \leq \|\cdot\|_1 \leq C\|\cdot\|$.

2: **Stopping**. At counter j, stop if $0 \in \partial f(x^j)$. Otherwise goto inner loop.

3: **Inner loop.** Given x^j and the euclidian norm $\|\cdot\|_j$ satisfying $c\|\cdot\| \leq \|\cdot\|_j \leq C\|\cdot\|$, use the step-finding algorithm with proximity control (algorithm 1) started at stepsize $t_j^\sharp$ to find a stepsize $t_k > 0$ such that the k^{th} trial point y^k satisfies $\rho_k \geq \gamma$. Put $x^{j+1} = y^k$ and goto step 4.

4: **Updating stepsize**. Check whether $\rho_k \geq \Gamma$ at acceptance $x^{j+1} = y^k$. If this is the case, put $t_{j+1}^\sharp = \theta^{-1} t_k$, otherwise put $t_{j+1}^\sharp = t_k$. Goto step 5.

5: **Small stepsize safeguard rule** (Optional). Replace $t_{j+1}^\sharp$ by $\max\{t_{j+1}^\sharp, \underline{t}\}$.

6: **Large stepsize safeguard rule** (Optional). Replace $t_{j+1}^\sharp$ by $\min\{\overline{t}, t_{j+1}^\sharp\}$.

7: **Updating norm**. Choose a new P_{j+1} such that $c\|\cdot\| \leq \|\cdot\|_{j+1} \leq C\|\cdot\|$. Then goto step 2.

Notice that step 5 is void if $\underline{t} = 0$, and the same for step 6 when $\overline{t} = \infty$. This is what we indicate by the term optional. In fact, we wish to avoid these rules in the convergence proofs, even though they are certainly acceptable in practice. For instance, linesearch methods tempting second-order steps always put $\underline{t} = 1$. Notice that if $\underline{t} = 0$ and $\overline{t} = \infty$, then the step length is fully memorized between serious steps.

In the smooth case, the idea of fully memorizing the steplength has been analyzed in [21], with the outcome that stepsize *may* be fully memorized for $C^{1,1}$-functions, whereas this is *not* possible for C^1 functions. Here the linesearch has to be started at $t_1 \geq \underline{t}$ for a threshold $\underline{t} > 0$. Since C^1 functions are upper C^1, and $C^{1,1}$-functions are upper C^2, we can consider items 2 and 3 of Theorem 2 below as non-smooth extensions of Theorems 1, 2 in [21], and of the results in [4].

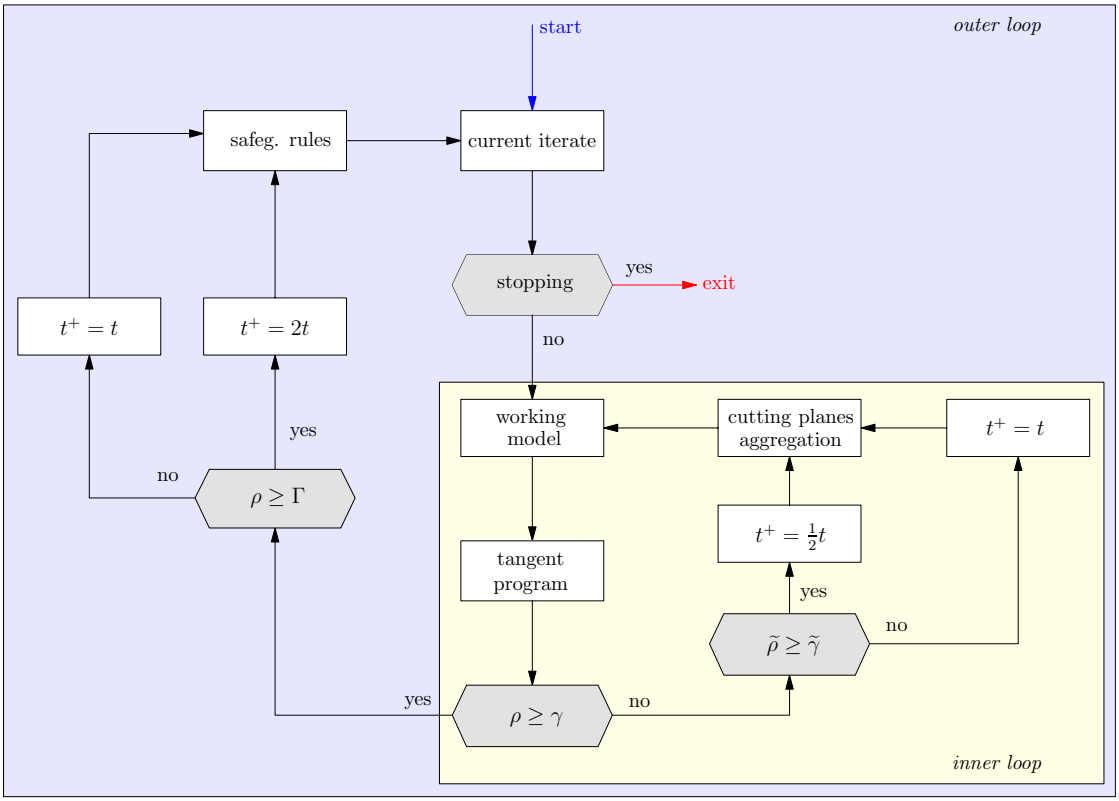


Figure 1: Flowchart of Algorithm 2

9 Convergence

In this section we prove convergence in the sense of subsequences of algorithm 2. Convergence to a single critical point will follow if the strong KL-property is satisfied.

Theorem 2. *Suppose f is locally Lipschitz and $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Let x^j be the sequence generated by algorithm 2. Then the following are satisfied:*

1. *If the standard model $\phi^\sharp$ of f is strict, i.e., $f \in \mathscr{S}$, then x^j has at least one accumulation point which is critical.*

2. *If the standard model is strict and algorithm 2 is operated with the small stepsize safeguard rule $\underline{t} > 0$, then* every *accumulation point of x^j is critical.*

3. *If the standard model is strong, then every accumulation point of the x^j is critical (and the small step safeguard rule may be dispensed with: $\underline{t} = 0$).*

4. *If the standard model $\phi^\sharp$ is strict and f satisfies the strong Kurdyka-Łojasiewicz property, then x^j converges to a single critical point (and the small stepsize safeguard rule may be dispensed with: $\underline{t} = 0$).*

In all these cases the large stepsize safeguard rule may be dispensed with, i. e., $\bar{t} = \infty$.

Proof: 1) From the analysis of section 7 we know that after a finite number of trials k the descent step finding algorithm 1 at serious iterate x^j locates a new iterate x^{j+1} satisfying the acceptance test $\rho_k \geq \gamma$, unless we have finite termination due to $0 \in \partial f(x^j)$. Excluding

this case, let x^j be the infinite sequence of steps generated by algorithm 2. Suppose x^{j+1} is accepted at inner loop counter k_j, i.e., $x^{j+1} = y^{k_j}$ passes the acceptance test, while the y^k with $k < k_j$ are null steps. This means

$$f(x^j) - f(x^{j+1}) \geq \gamma \left(f(x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j \right). \tag{19}$$

Now from optimality (7) we know that $g_j^* = t_{k_j}^{-1} P_j(x^j - x^{j+1}) \in \partial_1 \phi^\sharp_{k_j}(x^{j+1}, x^j)$, hence the subgradient inequality gives

$$(x^j - x^{j+1})^\top t_{k_j}^{-1} P_j (x^j - x^{j+1}) \leq \phi^\sharp_{k_j}(x^j, x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j) = f(x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j).$$

In combination with (19), and using $\|u\|_j^2 = u^\top P_j u$, this gives the estimate

$$t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2 \leq \gamma^{-1} \left(f(x^j) - f(x^{j+1}) \right). \tag{20}$$

Summing (20) over $j = 1, \ldots, J-1$ on both sides gives

$$\sum_{j=1}^{J-1} t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2 \leq \gamma^{-1} \left(f(x^1) - f(x^J) \right),$$

and since the algorithm is of descent type and the set of iterates x^j is bounded, the right hand side is bounded above, which implies summability of the series $\sum_j t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2$. In particular, this implies $t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2 \to 0$, and since the norms $\|\cdot\|_j$ are uniformly equivalent, $t_{k_j} \|x^j - x^{j+1}\|^2 \to 0$.

2) Let us consider an infinite subsequence $j \in \mathcal{N}$ of $\mathbb{N}$ where $g_j^* \to 0$, $j \in \mathcal{N}$. We will show that every accumulation point of the x^j, $j \in \mathcal{N}$, is critical. Let x^* be such an accumulation point, and passing to a subsequence if necessary, assume $x^j \to x^*$, $j \in \mathcal{N}$.

Since g_j^* is a subgradient of $\phi^\sharp_{k_j}(\cdot, x^j)$ at $x^{j+1} = y^{k_j}$, the subgradient inequality gives for every test vector h:

$$g_j^{*\top} h \leq \phi^\sharp_{k_j}(x^{j+1} + h, x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j) \leq \phi^\sharp(x^{j+1} + h, x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j).$$

As $x^{j+1} = y^{k_j}$ was accepted, we have $f(x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j) \leq \gamma^{-1} \left(f(x^j) - f(x^{j+1}) \right)$. Substituting this above gives

$$\begin{aligned} g_j^{*\top} h &\leq \phi^\sharp(x^{j+1} + h, x^j) - f(x^j) + f(x^j) - \phi^\sharp_{k_j}(x^{j+1}, x^j) \\ &\leq \phi^\sharp(x^{j+1} + h, x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1}) \right). \end{aligned}$$

Now putting $h = x^j - x^{j+1} + h'$, we obtain

$$g_j^{*\top}(x^j - x^{j+1}) + g_j^{*\top} h' \leq \phi^\sharp(x^j + h', x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1}) \right).$$

Passing to the limit $j \in \mathcal{N}$, using $x^j \to x^*$, boundedness of $x^j - x^{j+1}$, $g_j^* \to 0$ and $f(x^j) - f(x^{j+1}) \to 0$, we obtain

$$0 \leq \phi^\sharp(x^* + h', x^*) - f(x^*) = \phi^\sharp(x^* + h', x^*) - \phi^\sharp(x^*, x^*).$$

Since h' was arbitrary, this implies $0 \in \partial_1 \phi^\sharp(x^*, x^*) \subset \partial f(x^*)$, which proves what was claimed.

3) We shall now have to deal with the more complicated case of infinite subsequences $\mathcal{J}$ of $\mathbb{N}$ satisfying $\|g_j^*\| \geq \eta > 0$ for all $j \in \mathcal{J}$. We first claim that under this assumption, $t_{k_j} \to 0$, $j \in \mathcal{J}$. Indeed, if $t_{k_j} \geq \tau > 0$ for all $j \in \mathcal{J}$, then on passing to a subsequence $\mathcal{J}' \subset \mathcal{J}$, we may assume $P_j \to P$, $x^j - x^{j+1} \to \delta x$, $t_{k_j}^{-1} \to t^{-1} \leq \tau^{-1}$, and such that $t^{-1}\|\delta x\|_P \geq \eta$. But at the same time $(x^j - x^{j+1})^\top t_{k_j}^{-1} P_j (x^j - x^{j+1}) \to 0$ implies $t^{-1}\|\delta x\|_P^2 = 0$, a contradiction. This shows $t_{k_j} \to 0$ for the subsequence $j \in \mathcal{J}$.

4) Let us for convenience call infinite sequences $\mathcal{J}$ with $\|g_j^*\| \geq \eta > 0$ for all $j \in \mathcal{J}$ problematic. We distinguish two types of problematic sequences. The first type are those $\mathcal{J}_1$ where for every $j \in \mathcal{J}_1$ the backtracking rule (step 6 of algorithm 1) was applied at least once during the j^{th} inner loop. The second type are those $\mathcal{J}_2$ where the backtracking rule was never applied during the j^{th} inner loop for any one of the $j \in \mathcal{J}_2$. More formally,

$$\mathcal{J}_1 \subset \{j \in \mathbb{N} : t_j^\sharp > t_{k_j}\}, \quad \text{and} \quad \mathcal{J}_2 \subset \{j \in \mathbb{N} : t_j^\sharp = t_{k_j}\}.$$

Notice that every problematic subsequence $\mathcal{J}$ has either subsequences $\mathcal{J}_1$ of type 1, or $\mathcal{J}_2$ of type 2, or both.

5) Consider a problematic subsequence $j \in \mathcal{J}_1$ of the first type. Despite the fact that $t_{k_j} \to 0$, it is conceivable that $t_j^\sharp \to \infty$ on a subsequence. Fixing a small threshold $\vartheta > 0$, we consider the set $B = \{x^j, y^k : k \leq k_j, j \in \mathcal{J}_1, t_k \leq \vartheta\}$. On the other hand, for a problematic subsequence $j \in \mathcal{J}_2$ we simply choose $B = \{x^j, y^k : k \leq k_j, j \in \mathcal{J}_2\}$. We show that in both cases B is bounded.

Observe that the set of all serious iterates is bounded, so the question hinges on whether the null step y^k visited during the inner loop at x^j remain uniformly bounded. Now let $g_{0j} \in \partial f(x^j)$ be the first subgradient picked in the inner loop, which stays in the set $\mathcal{G}_k$ at all counters k of the j^{th} inner loop. Then $\{g_{0j} : j \in \mathcal{J}\}$ is bounded by the local boundedness of the Clarke subdifferential. By the subgradient inequality,

$$g_k^{*\top}(x^j - y^k) \leq \phi_k^\sharp(x^j, x^j) - \phi_k^\sharp(y^k, x^j) \leq g_{0j}^\top(x^j - y^k) \leq \|g_{0j}\|\|x^j - y^k\|,$$

and moreover, by the specific structure of the aggregate (see part 2) of the proof of Theorem 1), $g_k^{*\top}(x^j - y^k) = \|g_k^*\|\|x^j - y^k\|$. We deduce $\|g_k^*\| \leq \|g_{0j}\| \leq M < \infty$ for all j and all $1 \leq k \leq k_j$. This shows $t_k^{-1}\|x^j - y^k\| \leq M$ for all k, j, and implies boundedness of those y^k where t_k^{-1} is bounded away from 0. This applies to those k where $t_k \leq \vartheta$, and proves our claim.

6) Let us consider a problematic subsequence $j \in \mathcal{J}_1$ of type 1. Let $\hat{x}$ be an accumulation point of $\mathcal{J}_1$. We will show that $\hat{x}$ is critical. Passing to a subsequence if necessary, we may assume $x^j \to \hat{x}$, $j \in \mathcal{J}_1$.

Suppose that for $j \in \mathcal{J}_1$ the backtracking rule was applied for the last time at stage $k_j - \nu_j$ with $\nu_j \geq 1$. In other words,

$$t_{k_j} = t_{k_j - 1} = \cdots = t_{k_j - \nu_j + 1} < t_{k_j - \nu_j}, \tag{21}$$

with $t_{k_j} = \theta_{k_j - \nu_j} t_{k_j - \nu_j}$ for some $0 < \theta \leq \theta_{k_j - \nu_j} \leq \Theta < 1$. From the inner loop (algorithm 1) we know that backtracking occurs solely when $\rho < \gamma$, $\widetilde{\rho} \geq \widetilde{\gamma}$. In consequence, we have

$$\rho_{k_j - \nu_j} = \frac{f(x^j) - f(y^{k_j - \nu_j})}{f(x^j) - \phi_{k_j - \nu_j}^\sharp(y^{k_j - \nu_j}, x^j)} < \gamma, \quad \widetilde{\rho}_{k_j - \nu_j} = \frac{f(x^j) - \phi^\sharp(y^{k_j - \nu_j}, x^j)}{f(x^j) - \phi_{k_j - \nu_j}^\sharp(y^{k_j - \nu_j}, x^j)} \geq \widetilde{\gamma}.$$

From (21) we know that $\widetilde{g}_j := \theta_{k_j - \nu_j}^{-1} t_{k_j}^{-1} P_j(x^j - y^{k_j - \nu_j}) \in \partial_1 \phi_{k_j - \nu_j}^\sharp(y^{k_j - \nu_j}, x^j)$. We will show that $\widetilde{g}_j \to 0$, and subsequently, that this implies $0 \in \partial f(\hat{x})$.

By the subgradient inequality we have

$$\begin{aligned}(x^j - y^{k_j-\nu_j})^\top \theta_{k_j-\nu_j}^{-1} t_{k_j}^{-1} P_j (x^j - y^{k_j-\nu_j}) &\leq \phi^\sharp_{k_j-\nu_j}(x^j, x^j) - \phi^\sharp_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \\ &= f(x^j) - \phi^\sharp_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j). \end{aligned} \tag{22}$$

Now as $t_{k_j}^{-1} \to \infty$ for problematic subsequences, and by boundedness of the $\theta_{k_j-\nu_j}$, and boundedness of the set y^k of trial points and serious iterates shown in part 5), we must have $y^{kj-\nu_j} - x^j \to 0$. Since $x^j \to \hat{x}$, $j \in \mathcal{J}_1$, we have $y^{k_j-\nu_j} \to \hat{x}$, too.

7) We claim that the $\widetilde{g}_j = \theta_{k_j-\nu_j}^{-1} t_{k_j}^{-1} P_j(x^j - y^{k_j-\nu_j})$, $j \in \mathcal{J}_1$, are bounded. This can be seen from (22). Indeed, the left hand side behave asymptotically like $c\|\widetilde{g}_j\|\|x^j - y^{k_j-\nu_j}\|$. The right hand side of (22) is majorized by $f(x^j) - m_0(y^{k_j-\nu_j}, x^j)$, where $m_0(\cdot, x^j)$ is the exactness plane of the j^{th} inner loop. Therefore, (22) may be transformed into

$$c\|\widetilde{g}_j\|\|x^j - y^{k_j-\nu_j}\| \leq \|g_{0j}\|\|x^j - y^{k_j-\nu_j}\|.$$

But the $g_{0j} \in \partial f(x^j)$ are bounded due to boundedness of the x^j and local boundedness of the Clarke subdifferential operator ∂f, hence the claim.

8) We will now show that $\widetilde{g}_j \to 0$, $j \in \mathcal{J}_1$. Assume contrary to what is claimed that $\|\widetilde{g}_j\| \geq \theta > 0$ for all $j \in \mathcal{J}_1$. This shows

$$f(x^j) - \phi^\sharp_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \geq \theta\|x^j - y^{k_-\nu_j}\| \tag{23}$$

for all $j \in \mathcal{J}_1$. Indeed, from the subgradient inequality,

$$\widetilde{g}_j^\top (x^j - y^{k_j-\nu_j}) \leq \phi^\sharp_{k_j-\nu_j}(x^j, x^j) - \phi^\sharp_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) = f(x^j) - \phi^\sharp_{kj-\nu_j}(y^{k_j-\nu_j}, x^j).$$

But the left hand side behaves asymptotically like $\|\widetilde{g}_j\|\|x^j - y^{k_j-\nu_j}\|$, hence the claim (23).

9) This is now the moment where we apply the fact that f has a strict standard model $\phi^\sharp$. We apply axiom $(\widehat{M}_2)$ to the sequences $x^j \to \hat{x}$ and $y^{k_j-\nu_j} \to \hat{x}$. That means there exist $\epsilon_j \to 0$ such that

$$f(y^{k_j-\nu_j}) - \phi^\sharp(y^{k_j-\nu_j}, x^j) \leq \epsilon_j\|x^j - y^{k_j-\nu_j}\|. \tag{24}$$

Now we expand

$$\begin{aligned}\widetilde{\rho}_{k_j-\nu_j} &= \rho_{k_j-\nu_j} + \frac{f(y^{k_j-\nu_j}) - \phi^\sharp(y^{k_j-\nu_j}, x^j)}{f(x^j) - \phi^\sharp_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j)} \\ &\leq \rho_{k_j-\nu_j} + \frac{\epsilon_j\|x^j - y^{k_j-\nu_j}\|}{\theta\|x^j - y^{k_j-\nu_j}\|} = \rho_{k_j-\nu_j} + \epsilon_j/\theta. \end{aligned}$$

Since $\epsilon_j \to 0$, this shows $\limsup \widetilde{\rho}_{k_j-\nu_j} \leq \limsup \rho_{k_j-\nu_j} \leq \gamma < \widetilde{\gamma}$, contradicting the fact that $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma}$ for every $j \in \mathcal{J}_1$. This proves $\widetilde{g}_j \to 0$. Using the argument employed in part 2), we now deduce $0 \in \partial f(\hat{x})$.

10) It remains to deal with problematic subsequences of type 2. This can only be dealt with if the model is strong, or if the small stepsize safeguard rule is applied (that is, $\underline{t} > 0$). Namely, in both cases, the existence of problematic subsequences of type 2 can simply be ruled out. For $\underline{t} > 0$ this is clear, because then $t_{j+1}^\sharp$ is bounded below, and cannot go to zero. For $\phi^\sharp$ strong this could be obtained from [20].

11) Let us now assume that f satisfies the Kurdyka-Łojasiewicz inequality. We have to show that x^j converges to a single critical point x^*.

We have shown that the sequence x^j has at least one accumulation point x^* which is critical. Moreover, the set of accumulation points L of x^j is closed, as can be proved by a diagonal argument. Since $f(x^j)$ is decreasing, we conclude that f has constant value on the set L.

By assumption, for every $x \in L$, there exists an open neighborhood $U(x)$ of x and a continuous concave function $\kappa_x : [0, \eta_x] \to [0, \infty)$ of class C^1 on $(0, \eta_x)$ with $\kappa_x(0) = 0$, $\kappa_x' > 0$ on $(0, \eta_x)$, such that

$$\kappa_x' \left(f(x') - f(x)\right) \operatorname{dist}\left(0, \partial f(x')\right) \geq 1$$

whenever $x' \in U(x)$ satisfies $f(x) < f(x') < f(x) + \eta_x$. Using compactness of L, we find finitely many points $x_1, \ldots, x_r \in L$ such that the $U(x_1), \ldots, U(x_r)$ cover L. Choose $\epsilon > 0$ such that $V := \{x \in \mathbb{R}^n : \operatorname{dist}(x, L) < \epsilon\} \subset \bigcup_{i=1}^r U(x_i)$. Put $\eta = \min_{i=1,\ldots,r} \eta_{x_i}$, and define the function $\kappa'(t) = \max_{i=1,\ldots,r} \kappa_{x_i}'(t)$, then κ' is continuous and decreasing because all the κ_{x_i}' are. Putting $\kappa(t) = \int_0^t \kappa'(\tau)\, d\tau$ therefore defines a concave class C^1 function on $[0, \eta]$ with $\kappa(0) = 0$ and $\kappa' > 0$ on $(0, \eta)$. In addition, κ has the following property: For every $x \in L$ and every $x' \in V = \{x' : \operatorname{dist}(x', L) < \epsilon\}$ with $f(x) < f(x') < f(x) + \eta$ there holds

$$\kappa' \left(f(x') - f(x)\right) \operatorname{dist}\left(0, \partial f(x')\right) \geq 1. \tag{25}$$

Indeed, to see this let x, x' as above. Find x_i such that $x' \in U(x_i)$. Then

$$\begin{aligned} 1 &\leq \kappa_{x_i}' \left(f(x') - f(x_i)\right) \operatorname{dist}(0, \partial f(x')) \\ &\leq \kappa' \left(f(x') - f(x)\right) \operatorname{dist}(0, \partial f(x')), \end{aligned}$$

using $\kappa_{x_i}' \leq \kappa'$ and $f(x_i) = f(x)$. That proves our claim.

Let us for the following assume without loss that $f \equiv 0$ on L. Recall that the aggregate subgradient $g_j^* = t_{k_j}^{-1} P_j(x^j - x^{j+1})$ satisfies $t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2 \leq \gamma^{-1} \left(f(x^j) - f(x^{j+1})\right)$ by acceptance $\rho \geq \gamma$. Concavity of κ gives the estimate

$$\kappa\left(f(x^j)\right) - \kappa\left(f(x^{j+1})\right) \geq \kappa'\left(f(x^j)\right)\left(f(x^j) - f(x^{j+1})\right)$$

whenever $0 < f(x^j) < \eta$, $0 < f(x^{j+1}) < \eta$. Combining these twain gives

$$\kappa\left(f(x^j)\right) - \kappa\left(f(x^{j+1})\right) \geq \kappa'\left(f(x^j)\right) \gamma t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2.$$

By the strong KL-inequality (25), and using $f(x) = 0$, we have $\kappa'\left(f(x^j)\right) \geq \|g\|^{-1}$ for every Clarke subgradient $g \in \partial f(x^j)$. Therefore in particular $\kappa'\left(f(x^j)\right) \geq \|g_j^*\|^{-1}$ for the aggregate subgradient, which due to the specific form of the Clarke model $\phi^\sharp$ belongs to $\partial f(x^j)$. We deduce

$$\kappa\left(f(x^j)\right) - \kappa\left(f(x^{j+1})\right) \geq \gamma \frac{t_{k_j}^{-1} \|x^j - x^{j+1}\|_j^2}{t_{k_j}^{-1} \|P_j(x^j - x^{j+1})\|} \geq c' \|x^j - x^{j+1}\|$$

for some constant c'. That proves summability of $\|x^j - x^{j+1}\|$, hence x^j is a Cauchy sequence, which converges to x^* and $L = \{x^*\}$. Since L was shown to contain at least one critical point of f, we conclude that x^* is critical. That completes the proof of the Theorem.

□

10 Consequences and comments

In this section we present several consequences of the main Theorem 2, and give some comments. The following result is the expected convergence of the steepest descent method. Notice that we obtain algorithmically verifiable criteria for convergence, as opposed to conditions like [5]. The price to pay for this is that f has to belong to the class $\mathscr{S}$.

Corollary 1. *Suppose f is upper C^1 and satisfies the strong KL-inequality. Let $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ be bounded and let x^j be generated by a subgradient-oriented descent method, where the stepsize may be fully memorized. Then x^j converges to a critical point of f.*

Proof: By Proposition 1 we have $f \in \mathscr{S}$. Therefore algorithm 2 converges for the special case, where step finding uses algorithm 1 with $\phi_k^\sharp = \phi^\sharp$. Notice that we are in the case $\underline{t} = 0$ and $\bar{t} = \infty$, so *no restriction at all* is made on the stepsize, which means it is fully memorized. □

The next result describes a situation where the use of the small stepsize safeguard rule $\underline{t} > 0$ may be beneficial. Namely, it gives a satisfactory answer for stopping even when the KL-inequality is not available:

Corollary 2. *Let $f \in \mathscr{S}$ and suppose algorithm 2 is run with $\underline{t} > 0$. Suppose $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Then for every $\epsilon > 0$ there exists $j_0 \in \mathbb{N}$ such that all iterates x^j, $j \geq j_0$, are within distance ϵ of some critical point of f.*

Proof: Suppose there exist $\bar{\epsilon} > 0$ and infinitely many x^j, $j \in \mathcal{J}$, which have no critical point of f within $\bar{\epsilon}$ reach. Due to $\underline{t} > 0$, this sequence x^j, $j \in \mathcal{J}$, has an accumulation point, which by Theorem 2 is critical, a contradiction. □

The small stepsize safeguard rule is not needed if f has a strong model.

Corollary 3. *Suppose f is upper C^2 and $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Then for every $\epsilon > 0$ there exists $j_0 \in \mathbb{N}$ such that every iterate x^j, $j \geq j_0$, is within distance ϵ of some critical point of f.*

Proof: Since f has a strong standard model, infinite subsequences x^j, $j \in \mathcal{J}_2$, where $t_j^\sharp = t_{k_j} \to 0$ can be excluded. Those were named problematic subsequences of type 2 in the proof of Theorem 2. As the proof of Theorem 2 shows, all other subsequences (unproblematic, or problematic of type 1), have an accumulation point which is critical, and that proves the result. □

So far we have never needed the large stepsize safeguard rule $\bar{t} < \infty$. There is one specific situation, where this rule is beneficial, because it gives an additional option to converge to a single critical point without the KL-inequality.

Corollary 4. *Suppose the set K of critical points of $f \in \mathscr{S}$ is* apriori *known to be totally disconnected. If algorithm 2 is operated with both safeguard rules, i.e., $0 < \underline{t} < \bar{t} < \infty$, then the sequence x^j converges to a single critical point x^*.*

Proof: From the proof of Theorem 2 we know that $t_{k_j}^{-1}\|x^j - x^{j+1}\| \to 0$. Hypothesis $\bar{t} < \infty$ yields $t_{k_j} \leq \bar{t} < \infty$, so we deduce $x^j - x^{j+1} \to 0$, $j \to \infty$. As a consequence, by

Ostrowski's theorem [22], the set L of accumulation points of the sequence x^j is either a singleton or a nontrivial compact continuum.

Secondly, hypothesis $\underline{t} > 0$ assures that every accumulation point of the sequence x^j of serious iterates is critical, so that $L \subset K$. Since by hypothesis the only connected components of K are the singletons, L must be singleton, hence x^j converges to a single critical point x^*. □

Once again we could dispense with $\underline{t} > 0$ if f was upper C^2, respectively, if model $\phi^\sharp$ was strong, and we could dispense with $\bar{t} < \infty$ if we knew from other reasons that $x^j - x^{j+1} \to 0$.

11 Talweg and the unskilled skier's descent

The original motivation for the KL-property was to prove finite length of subgradient trajectories $\dot{x}(t) \in -\partial f(x(t))$. In the continuous case this immediately implies convergence to a critical point [9]. Subgradient-oriented descent may be understood as a discrete form of subgradient trajectories, and in [11] the authors use indeed finite length of such trajectories to characterize the KL-property. However, as we shall see in section 13, in the discrete case finite length of the trajectory does *not* imply convergence to a critical point. In order to assure convergence to a critical point, we need strictness of $\phi^\sharp$, i.e., $f \in \mathscr{S}$.

In [11] the authors use yet another discrete construction related to the KL-property, which they call a talweg. Again, finite length of the talweg may be used to characterize the KL-property. Here we consider the following slight modification of the talweg:

Algorithm 3. Unskilled skier's descent into the valley

Parameters: $0 < \gamma < \widetilde{\gamma} < 1$, $0 < \gamma < \Gamma < 1$, $0 < \theta < \Theta < 1$, $0 < c < C < \infty$, $K > 0$.

1: Given the current serious iterate x, stop if $0 \in \partial f(x)$. Otherwise use the step-finding algorithm 1 to find $\widehat{x}$ satisfying the acceptance test $\rho \geq \gamma$.
2: Manage the stepsize $t^\sharp$ as in algorithm 2.
3: Given the intermediate iterate $\widehat{x}$, find the new serious iterate x^+ on the same level curve, i.e., $f(x^+) = f(\widehat{x})$, such that $\|x^+ - \widehat{x}\| \leq K\|\widehat{x} - x\|$. Then go back to step 1.

The interpretation is as follows. The novice skier, lacking control, starts steepest descent (schuss) downhill from his current position x. Not being able to wedel, this leads him straight to $\widehat{x}$, with sufficient decrease $\rho \geq \gamma$ achieved quickly. Stopping at $\widehat{x}$ is arranged by sitting down on the bottom. In need of some rest, the clumsy skier now puts his skis in parallel with the level line to be stable for a while and then walks some distance along the level curve from $\widehat{x}$ to x^+. From here the procedure loops on by another pair of schuss-walk steps. The obvious question is whether the unskilled skier ever reaches the valley, i.e., whether the method converges to a critical point. (Finite length of the trajectory without convergence to a critical point is no consolation for the novice skier, because the ski lodge is at the bottom of the valley at a critical point. Convergence to a non-critical point means St. Bernhard dogs will have to pick him up on the slope a few days later).

We notice that the step from x to $\widehat{x}$ is identical with the serious step of algorithm 2. In [8] sequences with jumps like $\widehat{x} \to x^+$ are called piecewise gradient trajectories.

Theorem 3. *Suppose f is locally Lipschitz and $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Let x^1, $\widehat{x}^1$, x^2, $\widehat{x}^2, \ldots$ be the sequence generated by the unskilled skier's descent method. Then the following are satisfied:*

1. *If the standard model $\phi^\sharp$ is strict, then x^j, $\widehat{x}^j$ have at least one common accumulation point x^* which is critical.*
2. *If the standard model is strong, then* every *accumulation point of x^j, $\widehat{x}^j$ is critical.*
3. *If the standard model is strict and the small step safeguard rule ($\underline{t} > 0$) is used, then* every *accumulation point of x^j, $\widehat{x}^j$ is critical.*
4. *If the standard model is strict and f satisfies the strong KL-inequality, then $x^j, \widehat{x}^j$ converge to a single critical point x^*.*

Proof: The argument of Theorem 2 shows that $\sum_j \|x^j - \widehat{x}^j\| < \infty$. But $\|x^{j+1} - \widehat{x}^j\| \leq K\|\widehat{x}^j - x^j\|$, so that $\sum_j \|x^j - x^{j+1}\|$ converges, too. □

Remark 9. If f has the strong KL-property, but the standard model of f fails to be strict at x^*, then $x^j, \widehat{x}^j$ still converge to x^*, but x^* may fail to be critical. An example of this behavior is given in section 13.

12 Links with abstract convergence

We are now in the position to discuss the role of the sufficient conditions (3) and (4) in the convergence result of [6], and that of the alternative condition (5).

As we see from part 1) of the proof of Theorem 2, our acceptance test $\rho \geq \gamma$ forces the descent condition $f(x^j) - f(x^{j+1}) \geq \gamma t_{k_j}^{-1} \|x^j - x^{j+1}\|^2$, which is weaker than (3) in [6], and coincides with it when the $t_{k_j}^{-1}$ are bounded below. We could force the latter by the large stepsize safeguard rule, i.e., by choosing $\bar{t} < \infty$, but we only do this in the situation of Corollary 4, because in all other cases it represents an unnecessary limitation. Nonetheless, in the light of our result, condition (3) may be considered reasonable, because in practice we expect t_{k_j} to be bounded above most of the time, and more importantly, our analysis shows how (3) can be *forced* algorithmically.

It is more difficult to understand condition (4), because in our approach subgradient information at trial points y^k generated in the inner loop is only evaluated and registered at null steps, while we accept on the basis of (3) only. Therefore, if condition (4) is to be forced algorithmically, one has to *add* it to the acceptance test of the inner loop (step 3 of algorithm 1). However, we had convinced ourselves a long time ago that this foils finiteness of the inner loop. In fact, the case which poses problems is the one analyzed in part 3) of the proof of Theorem 1. We do not see how (4) could be forced algorithmically, and our example in section 13 shows that there is very little margin to succeed. We believe that condition (4) is not realistic for non-smooth descent methods.

Concerning conditions (5), observe that the aggregate Clarke subgradient g_j^* in part 2) of the proof is an element of $\partial f(x^j)$ due to the specific structure of the standard model $\phi^\sharp$, and it could therefore be a candidate for condition (5). However, our construction gives $\|g_j^*\| \sim t_{k_j}^{-1} \|x^j - x^{j+1}\|$, and this could only be bounded above by $a\|x^j - x^{j+1}\|$ if the t_{k_j} are bounded away from 0, an unlikely case. Since g_j^* is the most natural candidate to

converge to $0 \in \partial f(x^*)$, we conclude that (5) is again too strong to be realistic. This is corroborated by the example in section 13.

Remark 10. Conditions (4) and (5) are typical examples for the fact that criteria which work in the smooth case may not be mechanically transferred to the non-smooth setting in replacing gradients by subgradients. In fact, estimates (4), (5) go the wrong way, as our proof shows. It is the aggregate subgradient g_j^* which arranges convergence $g_j^* \to 0$, but the aggregate satisfies $\|g_j^*\| \leq M\|\partial f(x^j)\|$, and since $g_j^* \sim t_{k_j}^{-1}(x^j - x^{j+1})$, the estimate which *is* true reads $t_{k_j}^{-1}\|x^j - x^{j+1}\| \leq M\|\partial f(x^j)\|$.

Remark 11. The fact that the aggregate subgradient $g_j^* \to 0$ is a convex combination of model subgradients at null steps leads to the conclusion that the correct subdifferential to be used in non-convex bundling is the Clarke subdifferential. This is why our present theory needs the strong KL-inequality.

13 Failure of subgradient-oriented descent

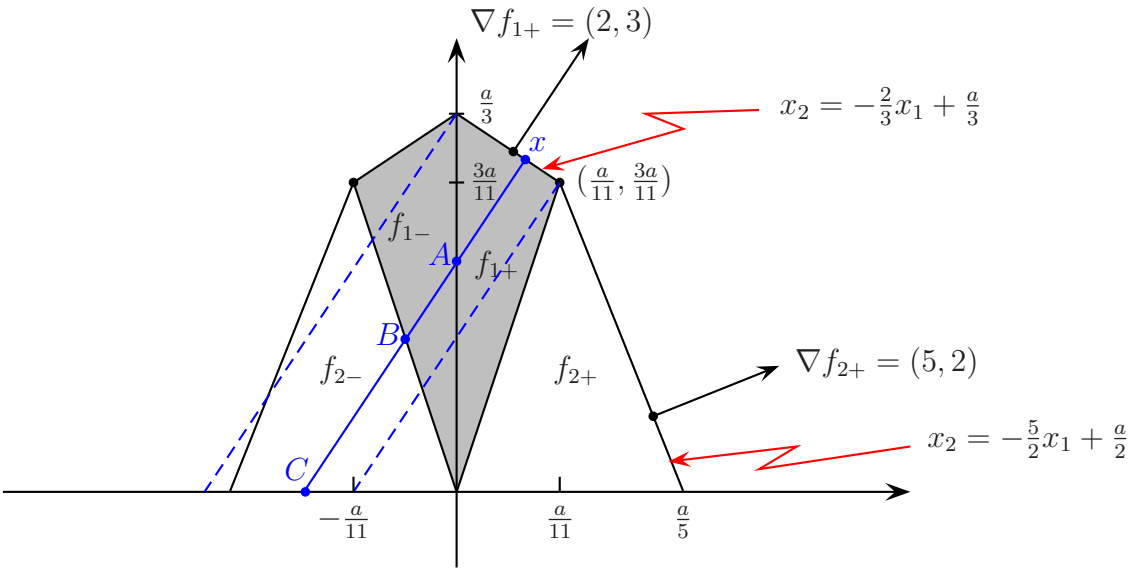
The following example adapted from [15] can be used to show the difficulties with non-smooth subgradient-oriented descent. We define a convex piecewise affine function $f : \mathbb{R}^2 \to \mathbb{R}$ as

$$f(x) = \max\{f_0(x), f_{\pm 1}(x), f_{\pm 2}(x)\}$$

where

$$f_0(x) = -100, f_{\pm 1}(x) = \pm 2x_1 + 3x_2, f_{\pm 2}(x) = \pm 5x_1 + 2x_2.$$

The following plot shows that part of the level curve $[f = a]$ which lies in the upper half plane $x_2 > 0$. It consists of the polygon connecting the five points $(-\frac{a}{5}, 0)$, $(-\frac{a}{11}, \frac{3a}{11})$, $(0, \frac{a}{3})$, $(\frac{a}{11}, \frac{3a}{11})$, $(\frac{a}{5}, 0)$. We are interested in the lower level set $[f \leq a]$ which lies inside the polygon, and above the x_1-axis.

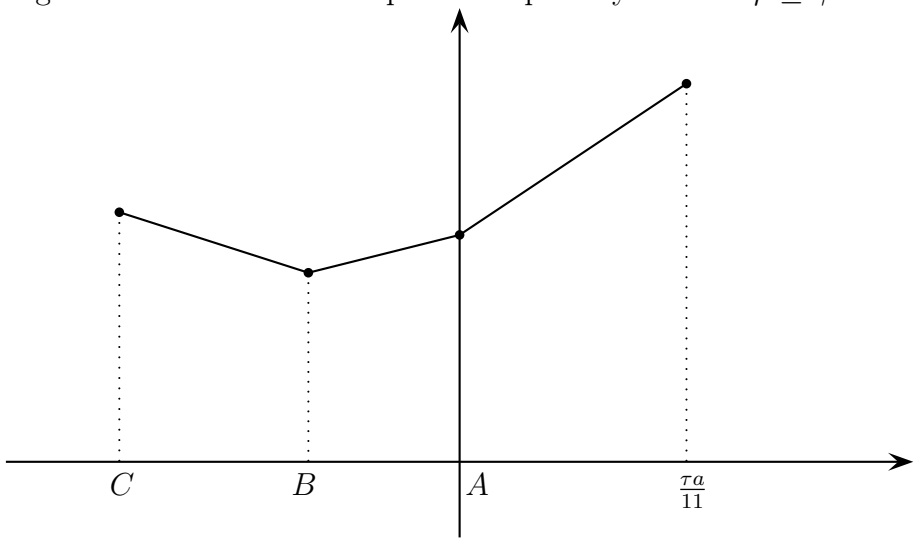


We decompose the lower level set $[f \leq a] \cap [x_2 \geq 0]$ into 4 regions where the 4 different branches of f are active, i.e., $[f = f_{1+}]$, $[f = f_{1-}]$, etc. In the plot we indicate these by the symbols f_{1+}, f_{1-}, etc. The lines $[f_{1+} = f_{2+}]$ and $[f_{1-} = f_{2-}]$ connect the origin to the points $(\pm\frac{a}{11}, \frac{3a}{11})$, while $[f_{1-} = f_{1+}] \cap [x_2 \geq 0]$ is the positive x_2-axis.

We denote the rhombus described by the four points $(0,0)$, $(\frac{a}{11},\frac{3a}{11})$, $(0,\frac{a}{3})$, $(-\frac{a}{11},\frac{3a}{11})$ by R_a. In the plot this area is shaded gray.

Now we consider a current point x on the upper right part of level curve $[f=a]$, that is, a point with $x_1=\frac{\tau a}{11}$ for some $0<\tau\leq 1$, and with $x_2=-\frac{2}{3}x_1+\frac{a}{3}$. In other words, x is on the right upper boundary of the rhombus R_a. The steepest descent direction at x is $-\nabla f_{1+}=(-2,-3)$. This is indicated by the blue line parting at x and passing through the points A,B,C. (The two limiting positions for x are the parallel dashed blue lines). Informally, what we now do is the following. We construct an instance of the steepest descent method, where steepest descent steps away from x along the blue line which are accepted by the test $\rho\geq\gamma$ lie before the point B. With the exception of the point A, which is also accepted, this means that we will stop at a point x^+ which is again on the upper part of a rhombus R_{a^+}, where $a^+=f(x^+)<f(x)$, possibly on the other side of the x_2-axis. Proceeding in this fashion, we will generate a sequence $x,x^+,x^{++},\ldots$ which will never escape from the rhombi $R_{f(x)}$, $R_{f(x^+)}$, $R_{f(x^{++})}$, and will converge to the origin, which is not a critical point of f.

Notice that our algorithm would also accept the point A on the positive x_2 axis, and this is indeed the only escape point on the blue line. Once an iterate of the form A is found, the steepest descent direction switches to $(0,-3)$ and we leave the rhombi through the origin. Our argument is that finding the only escape point A on the blue line is not algorithmically feasible, even more so as we have not specified any condition which distinguishes A from the other point accepted by the test $\rho\geq\gamma$.



If we plot the function $t\mapsto f(x+td)$, where d is the steepest descent direction $d=(-2,-3)$ at x with $0<x_1\leq\frac{a}{11}$, then we get a piecewise linear curve with two kinks corresponding to the points $A=(0,\frac{a}{3}-\frac{13}{6}x_1)$ and $B=(\frac{13}{27}x_1-\frac{2}{27}a;-\frac{13}{9}x_1+\frac{2}{9}a)$. Finally, the line hits the x_1-axis at $C=(\frac{13}{9}x_1-\frac{2}{9}a,0)$.

The function values at these points are

$$f(A)=f_{1+}\left(0,\frac{a}{3}-\frac{13}{6}x_1\right)=a-\frac{13}{2}x_1,$$

$$f(B)=f_{1-}\left(\frac{13}{27}x_1-\frac{2}{27}a,-\frac{13}{9}x_1+\frac{2}{9}a\right)=-\frac{26}{27}x_1+\frac{4}{27}a-\frac{13}{3}x_1+\frac{2}{3}a=\frac{22}{27}a-\frac{143}{27}x_1$$

$$f_{1+}(B)=-\frac{91}{27}x_1+\frac{14}{27}a$$

and

$$f(C) = f_{2-}\left(\frac{13}{9}x_1 - \frac{2}{9}a, 0\right) = -\frac{13\cdot 5}{9}x_1 + \frac{10}{9}a.$$

We obtain

$$\rho = \frac{a - f(B)}{a - f_{1+}(B)} = \frac{a - \frac{22}{27}a + \frac{143}{27}x_1}{a - \frac{14}{27}a + \frac{91}{27}x_1} = \frac{5a + 143x_1}{13a + 91x_1}.$$

If we put $x_1 = \tau\frac{a}{11}$ with $0 < \tau \leq 1$, then

$$\rho = \frac{5 + \frac{143\tau}{11}}{13 + \frac{91\tau}{11}} = \frac{55 + 143\tau}{143 + 91\tau}.$$

This quotient is independent of a and has its largest value at $\tau = 1$, namely, $\rho = \frac{198}{242}$. Therefore, if we put $1 > \gamma > \frac{198}{242}$, then none of the points B is accepted by the test $\rho \geq \gamma$, meaning that the interval of acceptance $(x, x^+] \subset (x, B)$ lies before B, and contains A in its interior. Notice that this interval of acceptance corresponds also to the interval of points accepted by condition (3). That means, the new serious iterate x^+ will have exactly the same properties as discussed for x, now in the rhombus R_{a^+}.

The question is now how convergence criteria (4) and (5) from section 5 behave in this situation. Can we find a point x^+ on the segment (x, B) where $\|\partial f(x^+)\|_- \leq b\|x - x^+\|$? Since $x - x^+ \to 0$ and the gradient is constant on the parts $[f = f_{1-}]$ and $[f = f_{1+}]$, the only candidate to be accepted by (4) is A, because here we get a convex combination of two gradients. The Clarke subgradients are $t(2,3)+(1-t)(-2,3) = (4t-2,3)$, $0 \leq t \leq 1$. Unfortunately, those are norm bounded below by 3, so A does not work. There is *no* point on the entire segment $[x, B]$ which is accepted by condition (4). This is bad in two aspects. Firstly, it is not good to have a hypothesis which is void. Secondly, one would at least have hoped that the point A could be accepted, since from A onward the steepest descent direction will pick another track and escape from the rhombus. In fact, the escape line *is* the positive x_2-axis. (Recall that our own method *does* accept the point A, but a linesearch which tries to locate a single point could not claim to work in practice). In contrast, (4) rejects even the escape point A. The same argument shows that condition (5) fails badly.

We still have to explain why convergence to a critical point fails here. According to our main theorem, this is due to the fact that the Clarke model is not strict at $x^* = (0,0)$. In order to verify this directly, consider points $x = (\xi, \eta)$, $y = (\xi', \eta')$ in the rhombus R_a, but with $\xi > 0$, $\xi' < 0$. If $\phi^\sharp(\cdot, (0,0))$ was to be strict, we would have to have $f(y) \leq f(x) + f^0(x, y - x) + o(\|x - y\|)$ for $(x, y) \to (0,0)$. But $f(y) = -2\xi' + 3\eta'$, $f(x) = 2\xi + 3\eta$, hence strictness requires

$$-2\xi' + 3\eta' \overset{!}{\leq} 2\xi + 3\eta + \langle(2,3), (\xi' - \xi, \eta' - \eta)\rangle + \epsilon\|x - y\|,$$

where $\epsilon \to 0$ as $\xi' \to 0$, $\xi \to 0$, $\eta' \to 0$, $\eta \to 0$. This gives

$$-2\xi' \overset{!}{\leq} 2\xi' + \epsilon\left(|\xi' - \xi| + |\eta' - \eta|\right)$$

if we use for simplicity the 1-norm on the right. This must obviously also hold when $\eta' = \eta \to 0$, so

$$-4\xi' \overset{!}{\leq} \epsilon|\xi' - \xi| = \epsilon(\xi - \xi'),$$

as $\xi' \to 0$, $\xi \to 0$, the latter because $\xi > 0$ and $\xi' < 0$. Suppose $\xi = -3\xi' \to 0$, then we should have $-4\xi' \leq -4\epsilon\xi'$, which requires $\epsilon \geq 1$. Therefore $\phi^\sharp$ is not strict at $(0, 0)$.

Let us finally observe that the ideal subgradient trajectory, where $\dot{x}(t) \in -\partial f(x(t))$ at almost all times t, will switch to the escape line as soon as it crosses it at point A, allowing an escape from the rhombus. This leads to the observation that in non smooth optimization, and this is in stark contrast with smooth optimization, looking at the continuous trajectory associated with a class of descent methods is useless, because it tells us nothing about the discrete method.

14 Conclusion

We have shown that convergence of subgradient-oriented non-smooth descent methods to critical points relies on two pillars. The Kurdyka-Łojasiewicz condition is sufficient to guarantee summability of $\sum_j \|x^j - x^{j+1}\| < \infty$ and therefore finite length of the discrete trajectory. Strictness of the standard model assures convergence to critical points in the sense of subsequences. When combined, these two assure convergence to a single critical point.

References

[1] D. Alazard, D. Noll, M. Gabarrou. Design of a flight control architecture using a non-convex bundle method. Mathematics of Control, Signals and Systems.

[2] P. Apkarian, D. Noll, O. Prot. A trust region spectral bundle method for nonconvex eigenvalue optimization. SIAM J. Optim. 10(1):281-306,2008.

[3] P. Apkarian, D. Noll, O. Prot. A proximity control algorithm to minimize non-smooth and non-convex semi-infinite maximum eigenvalue functions. J. Convex Anal. 16:641-666,2009.

[4] P.A. Absil, R. Mahony, B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.*, 16(2):531-547, 2005.

[5] Y.I. Alber, A.N. Iusem, M.V. Solodov. On the projected subgradient method for non-smooth convex optimization in a Hilbert space. *Math. Programming*, 81:23-35,1998.

[6] H. Attouch, J. Bolte, B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Programming, Ser. A

[7] H. Attouch, J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Programming, 116(1-2, Ser. B):5-16,2009.

[8] H. Attouch, J. Bolte, P. Redont, A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. 35(2):438-457,2010.

[9] J. Bolte, A. Daniilidis, A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Opt. 17(4):1205-1223,2007.

[10] J. Bolte, A. Daniilidis, A. Lewis, M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, 18(2):556-572,2007.

[11] J. Bolte, A. Daniilidis, O. Ley, L. Mazet. Characterizations of Łojesiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319-3363,2010.

[12] A. Daniilidis, P. Georgiev. Approximate convexity and submonotonicity. J. math. Anal. Appl. 291:117-144,2004.

[13] N. M. Dao, D. Noll. Minimizing the memory of a system. Submitted.

[14] J. E. Dennis jr., R. Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall Series in Computational Mathematics, 1983.

[15] J.-B. Hiriart-Urruty, C. Lemaréchal. Convex Analysis and Minimization Algorithms, vol. I and II: Advanced Theory and Bundle Methods, vol. 306 of Grundlehren der mathematischen Wissenschaften, Springer Verlag, New York, heidelberg, Berlin, 1993.

[16] K. Kurdyka. On gradients of functions definable in o-minimal structured. *Ann. Inst. Fourier*, 48(3):769-783,1998.

[17] S. Łojasiewicz. Sur les ensembles semi-analytiques. In *Actes du Congès International des Mathématiques (Nice, 1970), Tome 2, pages 237-241.* Gauthier-Villars, Paris, 1971.

[18] D. Noll. Cutting plane oracles to minimize non-smooth non-convex functions, Journal of Set-Valued and Variational Analysis, 18(3-4):531-568,2010.

[19] D. Noll. Bundle methods for non-convex minimization with inexact subgradient and function values. Computational and Analytical Mathematics. Springer proceedings in Mathematics, 2012.

[20] D. Noll, O. Prot, A. Rondepierre. A proximity control algorithm to minimize nonsmooth non-convex functions. Pacific J. Optim. 4(3):2008,569-602.

[21] D Noll, A. Rondepierre. Convergence of linesearch and trust-region methods using the Kurdyka-Łojasiewicz inequality. *Computational and Analytical Mathematics. Springer Proceedings in Mathematics.* 2012.

[22] A. M. Ostrowski. Solution of Equations in Euclidean and Banach Spaces. Academic Press, New York, 1973.

[23] R. T. Rockafellar, R. J-B. Wets. Variational Analysis. Springer Verlag, 2004.

[24] J. E. Spingarn. Submonotone subdifferentials of Lipschitz functions. Trans. Amer. math. Soc., 264:77-89,1981.