Convergence of Non-smooth Descent Methods Using the Kurdyka–Łojasiewicz Inequality

Dominikus Noll

Journal of Optimization Theory and Applications

ISSN 0022-3239

J Optim Theory Appl
DOI 10.1007/s10957-013-0391-8

J Optim Theory Appl
DOI 10.1007/s10957-013-0391-8

Convergence of Non-smooth Descent Methods Using the Kurdyka–Łojasiewicz Inequality

Dominikus Noll

Received: 9 October 2012 / Accepted: 2 August 2013

Abstract We investigate the 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 Kurdyka–Łojasiewicz inequality is satisfied. We show, by way of an example, that the Kurdyka–Łojasiewicz inequality alone is not sufficient to prove the convergence to critical points.

Keywords Non-smooth non-convex optimization · Subgradient-oriented descent method · Strict model · Kurdyka–Łojasiewicz inequality · Upper- and lower-C^1

1 Introduction

In smooth optimization, a sequence of descent directions is called gradient-oriented iff angles with negative gradients stay uniformly away from 90°. Convergence of gradient-oriented methods is guaranteed by the Armijo condition in tandem with a safeguard against too small steps [1]. Convergence is *a priori* understood in the sense of subsequences, but the work of Absil et al. [2] and subsequent generalizations [3–6] ensure *a posteriori* convergence to a single critical point if the objective function satisfies the Kurdyka–Łojasiewicz inequality.

In this work we investigate whether the KŁ-inequality allows similar results in non-smooth optimization. We consider non-smooth subgradient-oriented descent, where trial steps are computed by a convex quadratic tangent program, and where step finding uses the Armijo condition in tandem with backtracking. Our approach includes bundling techniques to deal with large size problems.

Communicated by Hedy Attouch.

D. Noll (✉)
Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse, France
e-mail: noll@mip.ups-tlse.fr

Published online: 12 September 2013

It turns out that there is a discrepancy between the smooth and the non-smooth case. In our framework the KŁ-inequality alone is not sufficient to prove convergence to a single critical point. Convergence only occurs under the additional hypothesis that the objective has a strict standard model in the sense of [7]. We demonstrate by way of an example that without this hypothesis convergence to a critical point may even fail for tame convex functions.

Bolte et al. [4] characterize the Kurdyka–Łojasiewicz inequality by the finite length of discrete subgradient trajectories, and by the existence of an approximate Talweg. We show that in the non-smooth case convergence of discrete subgradient trajectories to a single critical point is no longer guaranteed by the KŁ-inequality alone. As before the additional hypothesis of a strict standard model is needed. This is in contrast with [8], where it is shown that under the KŁ-inequality finite length of the continuous subgradient trajectory automatically implies its convergence to a critical point. So yet another discrepancy occurs within the non-smooth framework, now between discrete and continuous subgradient trajectories.

Attouch et al. [9] prove an abstract convergence result under the Kurdyka–Łojasiewicz inequality. We investigate whether the sufficient conditions of [9] can be used in our non-smooth framework, where trial steps are generated by a convex quadratic tangent program.

The paper is organized as follows. Sections 2.1–2.5 present the context and recall the model concept introduced in [7]. Sections 3.1–3.3 prove convergence of subgradient-oriented descent methods for functions with a strict standard model satisfying the KŁ-inequality. Consequences for the Talweg and for discrete gradient trajectories are given in Sect. 4.2. Links with the abstract descent result of [9] are discussed in Sect. 4.3. Limiting examples appear in Sect. 4.4.

2 Preparation

In this section we recall known concepts and discuss technical notions.

2.1 The Kurdyka–Łojasiewicz Inequality

Following [5], a locally Lipschitz function $f:\mathbb{R}^n\to\mathbb{R}$ satisfies the Kurdyka–Łojasiewicz inequality at $x^*\in\mathbb{R}^n$ iff there exist $\eta\in]0,\infty[$, a neighborhood U of x^*, and a concave function $\kappa:[0,\eta]\to[0,\infty[$ with:

(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

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

Here $\partial^L f(x)$ is the limiting subdifferential of f at x. f satisfies the strong KŁ-inequality at x^* iff the same estimate holds for the Clarke subdifferential $\partial f(x)$. For strictly differentiable f this reduces to the standard definition of the KŁ-inequality. Bolte et al. [10, Theorem 11] show that definable functions satisfy the strong KŁ-inequality, which covers a large variety of practical cases.

2.2 Subgradient-Oriented Descent

In a non-smooth framework the angle condition does no longer describe a useful set of search directions. The reason is that directions allowing descent form in general not a half-space, but a cone with opening angle $< 180^\circ$. That means a direction d with $\angle(d, -g_-) < 90^\circ$, g_- the steepest ascent subgradient, does not even need to allow descent. Fortunately, gradient-orientedness has an equivalent definition, which carries over to the non-smooth case.

Definition 2.1 A sequence d_j of normalized directions allowing descent of f at x_j is called subgradient-oriented iff there exist Clarke subgradients $g_j \in \partial f(x_j)$ and symmetric matrices P_j satisfying

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

for $0 < \lambda < \Lambda < \infty$ and all $j \in \mathbb{N}$, such that $d_j = -\frac{P_j g_j}{\|P_j g_j\|}$. In other words, the d_j are steepest descent directions at x_j with respect to the uniformly equivalent Euclidean norms $\|x\|_j^2 = x^\top P_j x$.

2.3 Discrete Subgradient-Oriented Flow

Bolte et al. [4] characterize the KŁ-inequality for convex $C^{1,1}$ functions by finite length of discrete gradient flow. This refers to sequences x_j satisfying the strong descent condition

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

If (2) is to hold for *all* points on $[x_j, x_{j+1}]$, then $\beta \|\nabla f(x_j)\| \leq -\nabla f(x_j)^\top d_j$, hence $\cos \angle(-\nabla f(x_j), d_j) \geq \beta > 0$, so d_j is gradient-oriented in the usual sense. Here we analyze the non-smooth and non-convex analogue of this result, using Definition 2.1. We seek algorithmic conditions ensuring convergence of discrete subgradient trajectories. Our results will be compared to [4, 5] in Sect. 4.2.

2.4 Abstract Descent Method

Attouch et al. [9] prove convergence of an abstract non-smooth descent method under the KŁ-inequality. Their sequence x_j has 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 the strong descent condition (2) and can be forced by backtracking, it is less obvious how condition (4) can be forced algorithmically. The challenge is to find a finite process at each iteration j which ensures (4) for the entire sequence x_j.

If it is possible to force (4) algorithmically, then it is natural to also consider a similar condition rooted at x_j, i.e., there exist $g_j \in \partial^L f(x_j)$ and $b > 0$ such that

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

because in tandem with the KŁ-inequality and tangent program (6), condition (5) is also sufficient to imply convergence to a single critical point.

We shall explain why (4) and (5) are difficult to force algorithmically for nonsmooth programs if computable local models are used in the tangent program. We will get back to this line in Sect. 4.3.

2.5 The Model Concept

Definition 2.2 (Compare [7, 13]) $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a *first-order model* of the locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ iff $\phi(\cdot, x)$ is convex for every $x \in \mathbb{R}^n$ and satisfies the following axioms:

(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, that is, $(y_j, x_j) \to (y, x)$ implies

$$\limsup_{j\to\infty} \phi(y_j, x_j) \leq \phi(y, x).$$

The first-order model ϕ is called strict at $\bar{x} \in \mathbb{R}^n$ iff the following strict version of axiom (M_2) is satisfied:

$(\widehat{M}_2)$ For every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in B(\bar{x}, \delta)$,

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

The model ϕ is called strict iff it is strict at every $\bar{x}$.

The first-order model ϕ is called strong at $\bar{x}$ iff the following even stronger version of (M_2) is satisfied:

$(\widetilde{M}_2)$ There exist $\delta > 0$ and $L > 0$ such that for all $x, y \in B(\bar{x}, \delta)$

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

ϕ is called strong iff it is strong at every $\bar{x}$.

Remark 2.1 One notes the resemblance with the Taylor expansion. Every locally Lipschitz function has a first-order model, which we call the standard model:

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

where $f^\circ(x, d)$ is the Clarke directional derivative of f. For C^1-functions $\phi^\sharp(y, x) = f(x) + \nabla f(x)(y - x)$ *is* the Taylor expansion. Note, however, that the Taylor expansion is unique, while we wish f to have as many models as possible, because each leads to a new optimization method.

Following [11, 12] a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ is lower C^k at x_0 iff there exist a compact K and a continuous $F : B(x_0, \delta) \times K \to \mathbb{R}$ with all partial derivatives of order $\leq k$ with respect to x 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 iff it is lower C^k at every x. Following [12] lower C^2 functions are already lower C^k for every $k \geq 2$, but the class lower C^1 is substantially larger than the class lower C^2. Finally, we call f upper C^k iff $-f$ is lower C^k.

Proposition 2.1 *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 $\epsilon > 0$. By Daniilidis and Georgiev [14] 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

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

Taking the limsup $t \to 0^+$ 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. □

The following definition is useful for the analysis of the subsequent sections.

Definition 2.3 A locally Lipschitz function f belongs to the class $\mathcal{S}$ iff its standard model $\phi^\sharp$ is strict.

Remark 2.2 The property $f \in \mathcal{S}$ seems weaker than upper C^1. Indeed, from [11] 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 $g \in \partial f(x)$ 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: 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

$$t^{-1}\big(f(x + td) - f(x)\big) \leq f^\circ(x, -d) + \epsilon,$$

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

$$t^{-1}\big(f(x + td) - f(x)\big) \leq f^\circ(x, d) + \epsilon.$$

3 Convergence

In this central section we develop our algorithm and prove convergence to a single critical point under the KŁ-inequality.

3.1 Descent Step Finding

We start with the question how to compute a subgradient-oriented descent step. The difficulty is that if $g \in \partial f(x)$, then due to non-smoothness, $-g$ will not necessarily allow descent. Directions allowing descent form a cone, not a half-space. It is therefore harder to find one. As we focus on subgradient-oriented methods, we will work with $\phi^\sharp$, even though some of the results hold for more general ϕ.

A function $\phi_k^\sharp : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a *first-order working model* iff $\phi_k^\sharp(\cdot, x)$ is convex, $\phi_k^\sharp \leq \phi^\sharp$, $\phi_k^\sharp(x,x) = \phi^\sharp(x,x) = f(x)$, 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 $\mathcal{G}_k \subset \partial f(x)$. Note 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

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

where $\|x\|_P^2 = x^\top P x$ is an Euclidean 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

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

or, equivalently,

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

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

Algorithm 1 Descent step finding by backtracking

```
Input: Current serious iterate $x$. Output: New 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)   $\min_{y \in \mathbb{R}^n} \phi_k^\sharp(y, x) + \frac{1}{2t_k}\|y - x\|_P^2$.
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 $x^+$.
   Otherwise, if $\rho_k < \gamma$, go to 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 $\mathcal{G}_{k+1}$ 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.
```

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 iff $\rho_k \geq \gamma$. 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 continues, 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 3.1 Note that our test $\rho_k{>}\gamma$ replaces the descent conditions (2) and (3).

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

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

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

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

with $g_i \in \mathcal{G}_k$ and $\lambda_i > 0$. Including g_k^* in the set $\mathcal{G}_{k+1}$ allows 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 has now 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 Euclidean norm $\| \cdot \|_P$. In other words, this case covers all non-smooth subgradient-oriented descent method with backtracking linesearch in the sense of Definition 2.1.

Theorem 3.1 *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 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. In both cases we will have to achieve a contradiction with the hypothesis $0 \notin \partial f(x)$. The first case may now be settled along the lines of [7, Lemma 4], while the second case uses the method of proof in [13, Lemma 6] or [7, Lemma 5]. □

Remark 3.2 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 (see e.g. Sect. 8.2 of [7, 16]).

3.2 Algorithm

In this section we state the main algorithm and comment 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, $t_{k+1} \in [\theta t_k, \Theta t_k]$. This means that in the inner loop (Algorithm 1) the stepsize is never

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 Euclidean 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 go to inner loop.
3: Inner loop. Given $x^j$ and the Euclidean 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 go to 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 $t_{j+1}^\sharp = t_k$. Go to 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 new $P_{j+1}$ such that $c\|\cdot\| \leq \|\cdot\|_{j+1} \leq C\|\cdot\|$. Then go to step 2.
```

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 is medium $\gamma \leq \rho_k < \Gamma$, then we memorize the last stepsize used.

Algorithm 2 contains the steepest descent method, and all subgradient-oriented descent methods in the sense of Definition 2.1, as special cases. On the other hand, it is more general because it allows to approximate these methods by numerically implementable iterative technique. This is beneficial in practical situations, where the full subdifferential $\partial f(x)$ is inaccessible to direct computation.

Step 5 is void if $\underline{t} = 0$, and the same for step 6 when $\overline{t} = \infty$. This is indicated by the term *optional*. We wish to avoid these rules in the proofs, even though they are acceptable in practice. For instance, linesearch methods tempting second-order steps always put $\underline{t} = 1$. Note that if $\underline{t} = 0$ and $\overline{t} = \infty$, then the step length is fully memorized between serious steps, an important option for large scale programs.

The idea to fully memorize the steplength was analyzed in [6], 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 3.2 below as non-smooth extensions of Theorems 1, 2 in [6], and of the results in [2].

3.3 Convergence

In this section we prove subsequence convergence of Algorithm 2. Convergence to a single critical point follows under the strong KŁ-inequality.

Theorem 3.2 *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., $f \in \mathcal{S}$, 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 is not needed: $\underline{t} = 0$).*
4. *If the standard model $\phi^\sharp$ is strict and f satisfies the strong Kurdyka–Łojasiewicz inequality, then x^j converges to a single critical point (and the small stepsize safeguard rule is not needed: $\underline{t} = 0$).*

In all these cases the large stepsize safeguard rule is not needed, i.e., $\overline{t} = \infty$.

Proof By Theorem 3.1, at serious iterate x^j Algorithm 1 finds a new serious iterate $x^{j+1} = y^{k_j}$ passing the acceptance test at the k_jth step of the inner loop. From acceptance $\rho_{k_j} \geq \gamma$ and optimality (7) we obtain

$$t_{k_j}^{-1} \|x^j - x^{j+1}\|^2 \leq \gamma^{-1}\bigl(f\bigl(x^j\bigr) - f\bigl(x^{j+1}\bigr)\bigr). \tag{9}$$

During the rest we will concentrate on the case where the sequence x^j is infinite. Now statement 3 may be obtained from the proof of Theorem 1 in [7]. Statement 2 can be dealt with using the proof of Theorem 2 of [7], because the only type of subsequences excluded in that proof is ruled out by the small stepsize safeguard rule. In the case of statement 1 we only need one subsequence with an accumulation point, and that is again essentially covered by Theorem 2 in [7]. See also Theorem 2 in [13].

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^*. It follows from statement 1 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 a constant value on the set L.

By assumption, for every $x \in L$, there exist 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

$$\kappa_x'\bigl(f\bigl(x'\bigr) - f(x)\bigr)\mathrm{dist}\bigl(0, \partial f\bigl(x'\bigr)\bigr) \geq 1$$

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 $V := \{x \in \mathbb{R}^n : \mathrm{dist}(x, L) < \epsilon\} \subset \bigcup_{i=1}^r U(x_i)$. 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 $x' \in V = \{x' : \mathrm{dist}(x', L) < \epsilon\}$ with $f(x) < f(x') < f(x) + \eta$ we have

$$\kappa'\bigl(f\bigl(x'\bigr) - f(x)\bigr)\, \mathrm{dist}\bigl(0, \partial f\bigl(x'\bigr)\bigr) \geq 1. \tag{10}$$

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

$$\begin{aligned}1 &\le \kappa'_{x_i}\big(f\big(x'\big) - f(x_i)\big)\operatorname{dist}\big(0, \partial f\big(x'\big)\big)\\ &\le \kappa'\big(f\big(x'\big) - f(x)\big)\operatorname{dist}\big(0, \partial f\big(x'\big)\big),\end{aligned}$$

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

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

$$\kappa\big(f\big(x^j\big)\big) - \kappa\big(f\big(x^{j+1}\big)\big) \ge \kappa'\big(f\big(x^j\big)\big)\big(f\big(x^j\big) - f\big(x^{j+1}\big)\big)$$

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

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

By the strong KŁ-inequality (10), and using $f(x) = 0$, we have $\kappa'(f(x^j)) \ge \|g\|^{-1}$ for every Clarke subgradient $g \in \partial f(x^j)$. Therefore $\kappa'(f(x^j)) \ge \|g_j^*\|^{-1}$ for the aggregate subgradient, which due to the specific form of the Clarke model $\phi^\sharp$ belongs to $\partial f(x^j)$. We deduce

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

for some constant c' independent of j. 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. □

4 Applications

In this section we present several consequences and applications of the main Theorem 3.2. Then discrete subgradient trajectories, the approximate Talweg, and abstract descent are discussed.

4.1 Consequences of the Main Theorem

We start with proving convergence of the steepest descent method, as promised. Recall that we wanted algorithmically verifiable criteria for convergence, as opposed to conditions like [19]. The price to pay for this is that we require $f \in \mathcal{S}$.

Corollary 4.1 *Suppose f is upper C^1 and satisfies the strong KŁ-inequality. Let $\{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$ be bounded and let x^j be generated by a subgradient-oriented descent method, where the stepsize is fully memorized. Then x^j converges to a critical point of f.*

Proof By Proposition 2.1 we have $f \in \mathcal{S}$. Therefore Algorithm 2 converges for the special case, where step finding uses Algorithm 1 with $\phi_k^\sharp = \phi^\sharp$. Note that we are in the case $\underline{t} = 0$ and $\overline{t} = \infty$, 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 KŁ-inequality is not available:

Corollary 4.2 *Let $f \in \mathcal{S}$ and suppose Algorithm 2 is operated 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 3.2 is critical, a contradiction. □

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

Corollary 4.3 *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. As the proof of Theorem 2 in [13] shows, all other subsequences, have an accumulation point which is critical, and that proves the result. □

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

Corollary 4.4 *Suppose the set K of critical points of $f \in \mathcal{S}$ is* a priori *known to be totally disconnected. If Algorithm 2 is operated with both safeguard rules*, i.e., $0 < \underline{t} < \overline{t} < \infty$, *then the sequence x^j converges to a single critical point x^*.*

Proof From the proof of Theorem 3.2 we know that $t_{k_j}^{-1}\|x^j - x^{j+1}\| \to 0$. Hypothesis $\overline{t} < \infty$ yields $t_{k_j} \leq \overline{t} < \infty$, so we deduce $x^j - x^{j+1} \to 0$, $j \to \infty$. As a consequence, by Ostrowski's Theorem [20], the set L of accumulation points of the sequence x^j is either a singleton or a nontrivial compact continuum.

Secondly, the hypothesis $\underline{t} > 0$ ensures 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 $\overline{t} < \infty$ if we knew for other reasons that $x^j - x^{j+1} \to 0$.

4.2 Talweg and the Unskilled Skier's Descent

In [8] the Kurdyka–Łojasiewicz inequality was used to prove finite length of subgradient trajectories $\dot{x}(t) \in -\partial f(x(t))$. In the continuous case this automatically implies convergence to a critical point. Now subgradient-oriented descent is a discrete form of the subgradient trajectory. This point of view is taken in [4], where the authors use the finite lengths of such trajectories to characterize the KŁ-property. However, as we shall see in Sect. 4.4, in the non-smooth case, the finite length of a discrete subgradient trajectory does *not* imply its convergence to a critical point. In order to ensure convergence, we need again strictness of $\phi^\sharp$, i.e., $f \in \mathcal{S}$.

In [4] the authors use yet another discrete construction related to the KŁ-property, which they call the Talweg. Again, finite length (but not convergence to a critical point) of the Talweg characterizes the KŁ-inequality. Here we consider the following:

```
**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.)

The step from x to $\widehat{x}$ is identical with the serious step of Algorithm 2. In [5] sequences with jumps like $\widehat{x} \to x^+$ are called piecewise subgradient trajectories.

Theorem 4.1 *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, i.e., $f \in \mathcal{S}$, and f satisfies the strong KŁ-inequality, then $x^j, \widehat{x}^j$ converge to a single critical point x^*.*

Proof Concentrating on item 4., the argument of Theorem 3.2 shows that $\sum_j \|x^j - \widehat{x}^j\| < \infty$. By step 3 of Algorithm 3, $\|x^{j+1} - \widehat{x}^j\| \leq K\|\widehat{x}^j - x^j\|$, so that $\sum_j \|x^j - x^{j+1}\|$ converges. □

Remark 4.1 If f has the strong KŁ-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 Sect. 4.4.

4.3 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 [9], and that of the alternative condition (5).

As we see from part (1) of the proof of Theorem 3.2, 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 [9], and coincides with it when the $t_{k_j}^{-1}$ are bounded below. We could force this by the large stepsize safeguard rule, $\overline{t} < \infty$, but we only do this in Corollary 4.4, because in all other cases it represents an unnecessary limitation. Nonetheless, in the light of our result, condition (3) used in [9] may be considered sub-optimal but reasonable, because in practice we expect t_{k_j} to be bounded above most of the time. More importantly, our analysis shows how (3) can be *forced.*

Let us now focus on condition (4), which is coined on the proximal point algorithm, where a sequence x_j is generated iteratively as

$$x_{j+1} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x) + \frac{1}{2t_j}\|x - x_j\|^2. \tag{11}$$

Here (3) is satisfied if stepsizes stay bounded away from zero, $t_j \geq \underline{t} > 0$, and (4) is then also satisfied with $b = \underline{t}^{-1}$ and $g_{j+1} = t_j^{-1}(x_j - x_{j+1}) \in \partial f(x_{j+1})$. Reference [9] presents other cases where (4) is satisfied. In our terminology, (11) corresponds to using f *as its own model* in the tangent program (6).

Remark 4.2 The situation becomes more delicate if only a local model $\phi(\cdot, x_j)$ of f is available in the tangent program at x_j:

$$\min_{y \in \mathbb{R}^n} \phi(y, x_j) + \frac{1}{2t_j}\|y - x_j\|^2.$$

Here the optimality condition $t_j^{-1}(x_j - x_{j+1}) \in \partial_1\phi(x_{j+1}, x_j)$ gives only information about $\partial_1\phi(x_{j+1}, x_j)$, which may be difficult, or even impossible, to relate to the information about $\partial f(x_{j+1})$ required in (4).

The dilemma is that if a model $\phi(\cdot, x_j)$ of f is used, the Armijo condition (3) is not satisfied for prior chosen stepsizes t_j, due to the discrepancy between $\phi(y^k, x_j)$ and $f(y^k)$. To force (3) we then have to take shorter steps t_j. This, however, is in conflict with condition (4), which can *not* be forced by backtracking, as it requires steps to be *not too small*. So as soon as f is replaced by a local model $\phi(\cdot, x_j)$, one has to prove that there exists a nonempty set of stepsizes t for which both conditions (3) and (4) are true, and one has to provide an algorithm which *finds* a step t in this set in finite time. That this may fail is shown in Examples 4.1, 4.2.

Let us next focus on condition (5), which we expect to be somewhat weaker than (4), and therefore easier to verify. This is corroborated by the following.

Proposition 4.1 *Suppose x_j is bounded and subgradient-oriented with respect to the norms $\|x\|_j^2 = x^\top P_j x$. Let f be upper C^2 at every accumulation point x^* of x_j. If the sequence x_j satisfies condition* (4), *then it also satisfies* (5), *possibly with a different constant $b' > 0$.*

Proof By (1) there exists $c > 0$ such that $\|x\|_j^2 = x^\top P_j x \geq c\|x\|\|P_j x\|$ for all x and all j.

Let us now single out an accumulation point x^* and consider a subsequence x_j converging to x^*. For simplicity call this subsequence x_j again. By assumption there exist $\delta > 0$ and $K > 0$ such that $f - K\|\cdot - x\|^2$ is concave for all $x \in B(x^*, \delta)$. Since $x_j \in B(x^*, \delta)$ for $j \geq j_0$, we have

$$-g_{j+1}^\top (x_{j+1} - x_j) + g_j^\top (x_{j+1} - x_j) \geq -K\|x_{j+1} - x_j\|^2$$

for $j \geq j_0$ by monotonicity of $-f + K\|\cdot - x_j\|^2$. That shows

$$\begin{aligned} -g_j^\top (x_{j+1} - x_j) &\leq K\|x_{j+1} - x_j\|^2 - g_{j+1}^\top (x_{j+1} - x_j) \\ &\leq K\|x_{j+1} - x_j\|^2 + \|g_{j+1}\|\|x_{j+1} - x_j\|. \end{aligned}$$

Since $x_{j+1} - x_j = t_j d_j$, with $d_j = -P_j g_j / \|P_j g_j\|$ subgradient-oriented, we have $-g_j^\top (x_{j+1} - x_j) \geq c\|g_j\|\|x_{j+1} - x_j\|$. Moreover, since (4) holds with b, we have $\|g_{j+1}\| \leq b\|x_{j+1} - x_j\|$, hence

$$c\|g_j\|\|x_{j+1} - x_j\| \leq (K + b)\|x_{j+1} - x_j\|^2,$$

which shows that (5) is true with $b' = (K + b)/c$ for $j \geq j_0$. □

Remark 4.3 One consequence of Proposition 4.1 is that it is reasonable to look for descent methods satisfying (3) and (5). Not only is there overlap with methods based on (3) and (4), which work in situations discussed in [9]. The proof of Theorem 3.1 also shows that (3) and (5) give convergence under the KŁ-inequality in their own right, a fact which was first presented in [21]. However, the limitations of both approaches become evident in practical situations, when computable local models $\phi(\cdot, x_j)$ have to be used in the tangent program.

4.4 Examples

In this section we consider several limiting examples.

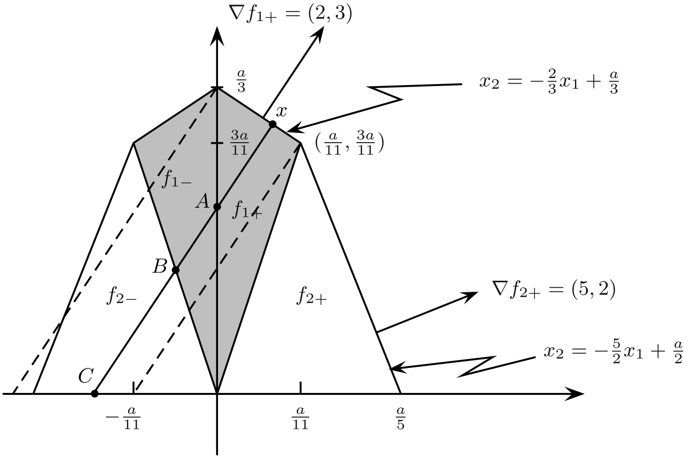
Example 4.1 The following example adapted from [22] can be used to show the difficulties with non-smooth subgradient-oriented descent. We define a convex piecewise affine function $f:\mathbb{R}^2\to\mathbb{R}$ as

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

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 $[f\le a]$, which lies inside the polygon, and above the x_1-axis.

We decompose the lower level set $[f\le a]\cap[x_2\ge 0]$ into four regions where different branches of f are active, i.e., $[f=f_{1+}]$, $[f=f_{1-}]$, indicated 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 $[f_{1-}=f_{1+}]\cap[x_2\ge 0]$ is the positive x_2-axis.

Let R_a denote the rhombus $(0,0)$, $(\frac{a}{11},\frac{3a}{11})$, $(0,\frac{a}{3})$, $(-\frac{a}{11},\frac{3a}{11})$, shaded gray in the plot. Consider the current x on the upper right part of level curve $[f=a]$. Then $x_1=\frac{\tau a}{11}$ for some $0<\tau\le 1$, and $x_2=-\frac{2}{3}x_1+\frac{a}{3}$. The steepest descent direction at x is $-\nabla f_{1+}=(-2,-3)$. This is the line parting at x through A,B,C. (The two limiting positions for x are the parallel dashed lines.) We now construct an instance of steepest descent, where steps from x which are accepted by the test $\rho\ge\gamma$ lie before B. With the exception of A, this means 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.



Note that our algorithm would also accept A on the positive x_2 axis, and this is indeed the only escape point from the rhombus. Once an iterate 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 is not algorithmically feasible, even more so as we have not specified any condition which distinguishes A from the other points 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 convex curve with two kinks, at the points $A = (0, \frac{a}{3} - \frac{13}{6}x_1)$ and $B = (\frac{13}{27}x_1 - \frac{2}{27}a; -\frac{13}{9}x_1 + \frac{2}{9}a)$, the lowest value being at B. 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 $f(A) = f_{1+}(0, \frac{a}{3} - \frac{13}{6}x_1) = a - \frac{13}{2}x_1$, $f(B) = f_{1-}(\frac{13}{27}x_1 - \frac{2}{27}a, -\frac{13}{9}x_1 + \frac{2}{9}a) = -\frac{26}{27}x_1 + \frac{4}{27}a - \frac{13}{3}x_1 + \frac{2}{3}a = \frac{22}{27}a - \frac{143}{27}x_1$ $f_{1+}(B) = -\frac{91}{27}x_1 + \frac{14}{27}a$ and $f(C) = f_{2-}(\frac{13}{9}x_1 - \frac{2}{9}a, 0) = -\frac{13 \cdot 5}{9}x_1 + \frac{10}{9}a$. 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 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. Note that this interval of acceptance corresponds also to the interval of points accepted by condition (3). That means that the new serious iterate x^+ will have exactly the same properties as discussed for x, now in the rhombus R_{a^+}.

The question is how criteria (4) and (5) from Sect. 2.4 behave. Can we find $x^+ \in]x, B[$ where $\|\partial f(x^+)\|_- \leq b\|x - x^+\|$? Since $x - x^+ \to 0$ and the gradient is constant on $[f = f_{1-}]$ and $[f = f_{1+}]$, the only candidate to be accepted by (4) is A. The Clarke subgradients are $g_t = t(2, 3) + (1 - t)(-2, 3) = (4t - 2, 3)$, $0 \leq t \leq 1$. Unfortunately, $\|g_t\| \geq 3$, so A does not work. There is *no* point on the entire segment $[x, B]$ which is accepted by (4). One would at least have hoped that A were 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 A, but a linesearch trying 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 (5) fails.

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 $x^* = (0, 0)$, a fact which can be verified directly.

Remark 4.4 The ideal subgradient trajectory $\dot{x}(t) \in -\partial f(x(t))$ switches to the escape line at point A, allowing to leave 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.

Example 4.2 Consider Nesterov's 2008 non-smooth variant [23] of the Rosenbrock function $\hat{f}(x_1,x_2)=\frac{1}{4}(x_1-1)^2+|x_2-2x_1^2+1|$, a contour plot of which can be seen in Fig. 1 (left) of [24]. The unique global minimizer is $x^*=[1,1]$, $\hat{f}$ is non-smooth at points on the manifold $\mathcal{M}=\{x\in\mathbb{R}^2:x_2^2-2x_1+1=0\}$. For $x\notin\mathcal{M}$ we have $\|\nabla\hat{f}(x)\|\geq 1$, so that $\|g_{j+1}\|\leq b\|x_{j+1}-x_j\|$ for $g_{j+1}\in\partial\hat{f}(x_{j+1})$ can only be arranged if $\|x_{j+1}-x_j\|\geq 1/b>0$ for all j where $x_{j+1}\notin\mathcal{M}$. That means that *every* iterative descent method which tries to satisfy (3) in tandem with either (4) or (5) has only two choices: It must either find the minimum x^* in a finite number of steps, or it must stay precisely on the manifold $\mathcal{M}$, i.e., $x_j\in\mathcal{M}$ for all $j\geq j_0$. Since $\hat{f}$ is not sharp at x^*, finite convergence will not occur, not even for the proximal point method, and if the manifold $\mathcal{M}$ is not explicitly known to the step-finding routine, identifying iterates $x_{j+1}\in\mathcal{M}$ may not be a numerically feasible. For instance, in [24] the authors have tested a non-smooth version of the BFGS-method, which zig-zags around the manifold $\mathcal{M}$, but does not find iterates on $\mathcal{M}$ even though it converges to $x^*\in\mathcal{M}$.

Example 4.3 One may argue that tangent program (11) is not realistic, because it is just as difficult to solve as the original problem $\min_{x\in\mathbb{R}^n}f(x)$. While this is true as a rule, we might make a concession for non-convex f: Here the use of (11) might be justified if the tangent program is convex, hence easier. This requires search for a convexifying $(1/2t_j)\|\cdot-x_j\|^2$ at each step j. However, this will spoil (4) if f is not lower C^2. As an example on the real line, define $f'(x)=-1$ for $x\leq 0$, and $f'(x)=1$ for $x\in[2^{-k},\alpha_k]$, $f'(x)=-k$ for $x\in(\alpha_k,2^{-k+1}]$, where $\alpha_k=2^{-k}(5/4+2k)/(1+k)$, $k=1,2,\dots$, then the primitive f of f' has a unique minimum at $x=0$, but in order to have convexity for f at iterates $x_j\in\bigcup_{j=1}^{\infty}(\alpha_j,2^{-k+1}]$, one needs smaller and smaller t_j in (11) as $x_j\to 0$. Then conditions (4) and (5) fail, while the backtracking strategy of Algorithm 2 is still functional.

5 Conclusions

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

The present approach, which is particularly useful for large size problems due to the bundling mechanism, can be adapted to deal with constraints e.g. by using the progress function approach of [17, 25], which was developed in [26] for smooth problems, or the improvement function of [27]. It is also possible to use weighted maxima as in [28]. Extensions to inexact values can be found in [18].

It remains open whether the Kurdyka–Łojasiewicz property can also be brought to work for bundle methods based on more general and more practical oracles, like downshifted tangents [17], tilted tangents [27], or to nonstandard models used e.g. in eigenvalue optimization [15]. Substantially new techniques of proof will have to be brought forward to settle these cases. Fortunately, results like Corollary 4.2 *can* be extended to these cases and give a satisfactory convergence theory for practical purposes even without the KŁ-inequality.

Acknowledgements The author was supported by the EADS Foundation grant *Technicom.*

References

1. Dennis, J.E. Jr., Schnabel, R.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall Series in Computational Mathematics. Prentice Hall, New York (1983)
2. Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. **16**(2), 531–547 (2005)
3. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program., Ser. B **116**(1–2),, 5–16 (2009)
4. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Łojesiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
5. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: 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)
6. Noll, D., Rondepierre, A.: Convergence of linesearch and trust-region methods using the Kurdyka–Łojasiewicz inequality. In: Bailey, D.H., Bauschke, H.H., Borwein, P., Garvan, F., Théra, M., Vanderwerff, J., Wolkowicz, H. (eds.) Computational and Analytical Mathematics. Proceedings in Mathematics and Statistics, vol. 50 (2013). In Honor of Jonathan Borwein's 60th Birthday
7. Noll, D., Prot, O., Rondepierre, A.: A proximity control algorithm to minimize non-smooth nonconvex functions. Pac. J. Optim. **4**(3), 569–602 (2008)
8. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
9. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program., Ser. A **137**(1), 91–129 (2013)
10. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
11. Spingarn, J.E.: Submonotone subdifferentials of Lipschitz functions. Trans. Am. Math. Soc. **264**, 77–89 (1981)
12. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (2004)
13. Noll, D.: Cutting plane oracles to minimize non-smooth non-convex functions. Set-Valued Var. Anal. **18**(3–4), 531–568 (2010)
14. Daniilidis, A., Georgiev, P.: Approximate convexity and submonotonicity. J. Math. Anal. Appl. **291**, 117–144 (2004)
15. Apkarian, P., Noll, D., Prot, O.: A trust region spectral bundle method for nonconvex eigenvalue optimization. SIAM J. Optim. **10**(1), 281–306 (2008)
16. Apkarian, P., Noll, D., Prot, O.: A proximity control algorithm to minimize non-smooth and nonconvex semi-infinite maximum eigenvalue functions. J. Convex Anal. **16**, 641–666 (2009)
17. Gabarrou, M., Noll, D., Alazard, D.: Design of a flight control architecture using a non-convex bundle method. Math. Control Signals Syst. **25**(2), 257–290 (2013)
18. Noll, D.: Bundle methods for non-convex minimization with inexact subgradient and function values. In: Bailey, D.H., Bauschke, H.H., Borwein, P., Garvan, F., Théra, M., Vanderwerff, J., Wolkowicz, H. (eds.) Computational and Analytical Mathematics. Springer Proceedings in Mathematics and Statistics, vol. 50 (2013). In Honor of Jonathan Borwein's 60th Birthday
19. Alber, Y.I., Iusem, A.N., Solodov, M.V.: On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. Math. Program. **81**, 23–35 (1998)
20. Ostrowski, A.M.: Solution of Equations in Euclidean and Banach Spaces. Academic Press, New York (1973)
21. Noll, D.: A bundle method for non-smooth and non-convex optimization. Talk at the 2009 ISMP, Chicago

22. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms, vol. I and II: Advanced Theory and Bundle Methods. Grundlehren der Mathematischen Wissenschaften, vol. 306. Springer, New York (1993)
23. Nesterov, Y.:. Private communication (2013)
24. Gürbüzbalaban, M., Overton, M.L.: On Nesterov's nonsmooth Chebyshev-Rosenbrock functions. Nonlinear Anal. **75**, 1282–1289 (2012)
25. Apkarian, P., Noll, D., Rondepierre, A.: Mixed H_2/H_∞ control via nonsmooth optimization. SIAM J. Control Optim. **47**(3), 1516–1546 (2008)
26. Polak, E.: Optimization: Algorithms, and Consistent Approximation. Springer, Berlin (1997)
27. Sagastizábal, C., Solodov, M.: An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or filter. SIAM J. Optim. **16**(1), 146–169 (2005)
28. Apkarian, P., Noll, D.: Nonsmooth optimization for multiband frequency domain control design. Automatica **43**, 724–731 (2007)