

# Gradients on sets

Jan Mankau, Friedemann Schuricht

TU Dresden, Fakultät Mathematik

01062 Dresden, Germany

## Abstract

For a locally Lipschitz continuous function  $f : X \rightarrow \mathbb{R}$  the generalized gradient  $\partial f(x)$  of Clarke is used to develop some (set-valued) gradient on a set  $A \subset X$ . Existence, uniqueness and some approximation are considered for optimal descent directions on set  $A$ . The results serve as basis for nonsmooth numerical descent algorithms that can be found in subsequent papers.

## 1 Introduction

For a smooth function  $f : X \rightarrow \mathbb{R}$  the derivative  $f'(x)$  in particular indicates directions of descent near  $x$ . This fact serves as basis for typical numerical descent algorithms. However such algorithms fail in cases where the direction of descent changes rapidly in a small neighborhood of  $x$ . This typically occurs for functions having large second derivatives and, even worse, for functions that are not differentiable. In such situations it becomes necessary to use more information of  $f$  for the selection of a descent direction. If we consider some  $f$  being the pointwise maximum of two (non-constant) linear functions, we have to realize that also Clarke's set-valued generalized gradient  $\partial f(x)$ , defined for Lipschitz continuous functions  $f$ , does not provide enough information for a stable scheme. Therefore the selection of a robust descent direction is only possible if one uses relevant information of  $f$  from some suitable neighborhood of  $x$ .

We consider locally Lipschitz continuous functions  $f : X \rightarrow \mathbb{R}$  on a Banach space  $X$ . Using the generalized gradients of Clarke we introduce some (set-valued) gradient  $\partial f(A)$  of  $f$  on a set  $A \subset X$  and, with Clarke's generalized directional derivative  $f^0(y; h)$ , we define some directional derivative  $f^0(A; h)$  of  $f$  on  $A$  in direction  $h$ . In Section 2 we verify basic properties for these new quantities where some are quite similar to that in Clarke's calculus. For sequences of sets  $A_k \rightarrow A$  converging in the Hausdorff metric, some general upper semicontinuity is shown. Here the relevance of certain assumptions is illuminated by examples. Moreover we

show that the  $\varepsilon$ -generalized gradient  $\delta_\varepsilon f(x)$  of  $f$  at  $x$  introduced in Goldstein [5] for  $X = \mathbb{R}^n$  (that somehow relies on Rademacher's Theorem for Lipschitz continuous functions) agrees with  $\partial f(\overline{B_\varepsilon(x)})$ . Finally we consider regularity in the sense that  $0 \notin \partial f(A)$  and, in particular, a result from Goldstein [5] is extended to Banach spaces. In Section 3 we define descent directions and optimal descent directions of  $f$  on  $A$ . Then existence and general properties of optimal descent directions are analyzed. An example demonstrates that there might be no optimal descent direction in a non-reflexive Banach space. Uniqueness of an optimal descent direction can be verified for strictly convex Banach spaces. Examples show that the selection of descent directions and optimal descent directions needs much more care in spaces that are not strictly convex. Furthermore we provide some stability and approximation results for optimal descent directions that are very useful for applications in numerics. The advantage of gradients on sets and corresponding descent directions for numerical algorithms is demonstrated by a simple but typical example. Applications of the analytical results to nonsmooth descent algorithms and corresponding numerical simulations can be found in subsequent papers.

*Notation:* By  $X$  we denote a Banach space, by  $X^*$  its dual, and by  $\langle \cdot, \cdot \rangle$  the corresponding duality pairing. We call  $X$  (or  $X^*$ ) strictly or uniformly convex if the norm has that property (cf. [2]). For a set  $M$  we use  $\overline{M}$  for its closure,  $\text{conv } M$  for its convex hull, and  $\overline{\text{conv}}^* M$  for its weak\*-closed convex hull.  $B_\varepsilon(x)$  stands for the open  $\varepsilon$ -neighborhood of point  $x$  and  $B_\varepsilon(M)$  for the open  $\varepsilon$ -neighborhood of set  $M$ . We write  $]x, y[$  and  $[x, y]$  for the open and closed segment (or interval), respectively, generated by the points  $x, y$ . Clarke's generalized directional derivative is denoted by  $f^0(x; h)$  and its generalized gradient by  $\partial f(x) \subset X^*$  (cf. Clarke [3]). Notice that  $\partial f(A)$  denotes the gradient defined in (2.1) and does not mean  $\bigcup_{x \in A} \partial f(x)$ .

## 2 Gradients on sets

Let  $X$  be a Banach space and let  $f : X \rightarrow \mathbb{R}$  be a locally Lipschitz continuous function. We denote the generalized gradient at  $x$  by  $\partial f(x)$  and the generalized directional derivative at  $x$  in direction  $y$  by  $f^0(x; y)$  (cf. Clarke [3]). While these quantities somehow express the behavior of  $f$  at the point  $x$ , we are interested in information that expresses the behavior of  $f$  on a whole set. Therefore we introduce some set-valued gradient of  $f$  on a set  $A \subset X$  by using Clarke's pointwise quantities. Later sets  $A = \overline{B_\varepsilon(x)}$  with  $\varepsilon > 0$  will be of particular interest.

For  $A \subset X$  ( $\neq \emptyset$ ) we define the *gradient* of  $f$  on  $A$  by

$$\partial f(A) := \overline{\text{conv}}^* \bigcup_{y \in A} \partial f(y) \quad (2.1)$$

(where  $\overline{\text{conv}}^*$  denotes the weak\* closure of the convex hull) and the *directional derivative* of  $f$  at  $A$  in direction  $h \in X$  by

$$f^0(A; h) := \sup_{y \in A} f^0(y; h). \quad (2.2)$$

Clearly  $\partial f(x) = \partial f(\{x\})$  and  $f^0(x; h) = f^0(\{x\}; h)$ . Let us start with some basic properties.

**Proposition 2.3.** *Let  $A \subset X$  be nonempty and let  $f : X \rightarrow \mathbb{R}$  be Lipschitz continuous of rank  $L$  on a neighborhood of  $A$ . Then:*

- (1)  $\partial f(A)$  is nonempty, convex, weak\*-compact and bounded by  $L$ .
- (2)  $f^0(A; \cdot)$  is finite, positively homogeneous, subadditive, and Lipschitz continuous of rank  $L$ . Moreover it is the support function of  $\partial f(A)$  with

$$f^0(A; h) = \max_{a \in \partial f(A)} \langle a, h \rangle \quad \text{for all } h \in X. \quad (2.4)$$

- (3) We have

$$\partial f(A) = \{a \in X^* \mid \langle a, h \rangle \leq f^0(A; h) \text{ for all } h \in X\}. \quad (2.5)$$

- (4) Let  $h \in X$  with  $f^0(A; h) < 0$ , let  $x \in A$ , and let  $t > 0$  with  $]x, x + th[ \subset A$ . Then

$$f(x + th) \leq f(x) + tf^0(A; h) < f(x).$$

PROOF. For (1) we recall that  $\partial f(y)$  is nonempty and bounded by  $L$  for all  $y \in A$  (cf. [3, Prop. 2.1.2]). Thus the stated properties follow easily from the definition of  $\partial f(A)$  and the Banach Alaoglu Theorem.

For (2) we first notice that  $f^0(y; \cdot)$  is the support function of  $\partial f(y)$  (cf. [3, Prop. 2.1.2]). Therefore we obtain for the support function of  $\partial f(A)$  at  $h \in X$

$$\begin{aligned} \sup_{a \in \partial f(A)} \langle a, h \rangle &= \sup \left\{ \langle a, h \rangle \mid a \in \overline{\text{conv}}^* \left( \bigcup_{y \in A} \partial f(y) \right) \right\} \\ &= \sup \left\{ \langle a, h \rangle \mid a \in \text{conv} \left( \bigcup_{y \in A} \partial f(y) \right) \right\} \\ &= \sup \left\{ \langle a, h \rangle \mid a \in \left( \bigcup_{y \in A} \partial f(y) \right) \right\} \\ &= \sup_{y \in A} \sup_{a \in \partial f(y)} \langle a, h \rangle = \sup_{y \in A} f^0(y; h) = f^0(A; h). \end{aligned}$$

Since  $\partial f(A)$  is weak\*-compact, the supremum is attained and (2.4) follows. The remaining properties are now easy consequences.

For (3) we notice that characterization (2.5) is as general property of support functions (cf. [3, Prop. 2.1.4]).

For (4) we use Lebourg's mean value theorem (cf. [3, Prop. 2.3.7]) to get some  $z \in ]x, x + th[$  and some  $a \in \partial f(z) \subset \partial f(A)$  such that

$$f(x + th) - f(x) = \langle a, th \rangle \stackrel{(2.4)}{\leq} tf^0(A; h) < 0,$$

which directly implies the assertion. ◇

**Proposition 2.6** (upper semicontinuity). *Let  $f : X \rightarrow \mathbb{R}$  be locally Lipschitz continuous, let  $h_k \rightarrow h$  in  $X$ , and let  $A_k, A \subset X$  with  $A$  compact and  $A_k \rightarrow A$  in the Hausdorff metric, i.e.*

$$d(A_k, A) := \inf \{ \delta > 0 \mid A \subset B_\delta(A_k) \text{ and } A_k \subset B_\delta(A) \} \xrightarrow{k \rightarrow \infty} 0.$$

Then

$$\limsup_{k \rightarrow \infty} f^0(A_k; h_k) \leq f^0(A; h) \quad (2.7)$$

$$\{a \in X^* \mid a_k \xrightarrow{*} a \text{ for } a_k \in \partial f(A_k)\} \subset \partial f(A). \quad (2.8)$$

If  $A \subset A_k$  for all  $k \in \mathbb{N}$ , then we have equality in (2.8) and

$$\lim_{k \rightarrow \infty} f^0(A_k; h_k) = f^0(A; h) \quad (2.9)$$

With  $A = \{x\}$  we directly derive the following statement.

**Corollary 2.10.** *Let  $x \in X$  and  $\varepsilon_k \rightarrow 0$  such that  $x \in A_k \subset B_{\varepsilon_k}(x)$  and let  $h \in X$ . Then*

$$\lim_{k \rightarrow \infty} f^0(A_k; h) = f^0(x; h) \quad \text{and} \quad \bigcap_{k \in \mathbb{N}} \partial f(A_k) = \partial f(x).$$

PROOF of Proposition 2.6. By definition and assumption there exist  $x_k \in A_k$  and  $z_k \in A$  with

$$f^0(A_k; h_k) \geq f^0(x_k; h_k) \geq f^0(A_k; h_k) - \frac{1}{k} \quad \text{and} \quad \|x_k - z_k\| \rightarrow 0.$$

By compactness of  $A$  we get, possibly for a subsequence,

$$f^0(x_k; h_k) \rightarrow \limsup_{k \rightarrow \infty} f^0(A_k; h_k) \quad \text{and} \quad z_k \rightarrow z \in A.$$

Consequently  $x_k \rightarrow z$ . Since  $f^0(\cdot; \cdot)$  is upper semicontinuous (cf. [3, Prop. 2.1.1]),

$$f^0(A; h) \geq f^0(z; h) \geq \lim_{k \rightarrow \infty} f^0(x_k; h_k) = \limsup_{k \rightarrow \infty} f^0(A_k; h_k)$$

and we have (2.7).

Let now  $a_k \in \partial f(A_k)$  with  $a_k \xrightarrow{*} a$ . Hence

$$f^0(A; h) \geq \limsup_{k \rightarrow \infty} f^0(A_k; h) \geq \lim_{k \rightarrow \infty} \langle a_k, h \rangle = \langle a, h \rangle \quad \text{for all } h \in X.$$

Thus  $a \in \partial f(A)$  by (2.5). If  $A \subset A_k$ , then  $f^0(A; \cdot) \leq f^0(A_k; \cdot)$  and  $\partial f(A) \subset \partial f(A_k)$  by definition. Hence, equality in (2.8) follows in the case that  $A \subset A_k$  for all  $k \in \mathbb{N}$ . Furthermore

$$f^0(A; h) \leq \liminf_{k \rightarrow \infty} f^0(A_k; h_k) \leq \limsup_{k \rightarrow \infty} f^0(A_k; h_k) \stackrel{(2.7)}{\leq} f^0(A; h)$$

and (2.9) follows. ◇

**Example 2.11.** We present some examples showing the necessity of central assumptions in Proposition 2.6.

(1) Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be given by  $f(x) = |x|$  and let

$$h_k = h = -1, \quad A_k := ] -\frac{1}{k}, 1[, \quad A := ]0, 1[.$$

Obviously  $d(A_k, A) \rightarrow 0$ , but  $A$  is not compact. We have  $\partial f(A) = \{1\}$  and

$$f^0(A_k; h_k) \geq f^0(0; h) = 1 > -1 \stackrel{(2.4)}{=} f^0(A; h).$$

Hence (2.7) is not satisfied.

(2) Let again  $f : \mathbb{R} \rightarrow \mathbb{R}$  be given by  $f(x) = |x|$  and let

$$h_k = h = -1, \quad A_k := [\frac{1}{k}, 1], \quad A := [0, 1].$$

Here  $A$  is compact and  $d(A_k, A) \rightarrow 0$ , but  $A \not\subset A_k$ . We have  $\partial f(A_k) = \{1\}$  and

$$f^0(A; h) \geq f^0(0; h) = 1 > -1 \stackrel{(2.4)}{=} f^0(A_k; h_k).$$

Therefore (2.7) is satisfied, but without equality as in (2.9).

(3) For  $X = \ell^2$  (sequences  $x = (\xi_i)_{i \in \mathbb{N}}$  in  $\mathbb{R}$  with  $\|x\|^2 = \sum_{k \in \mathbb{N}} |\xi_k|^2 < \infty$ ) we consider

$$A := \{0\}, \quad A_k := \overline{B_1(0)} \cap \{(\xi_i) \in \ell^2 \mid \xi_j = 0 \text{ for } j < k\}.$$

Clearly  $A$  is compact and  $A = \bigcap_{k \in \mathbb{N}} A_k$ . But  $d(A_k, A) \not\rightarrow 0$ , since there are  $x_k \in A_k$  with  $\|x_k\| = 1$ . With fixed  $z \in \ell^2 \setminus \{0\}$  and  $\phi \in C^\infty(\mathbb{R}, \mathbb{R})$  satisfying

$$\phi(\alpha) = 1 \text{ for } \alpha < \frac{1}{4}, \quad \phi(\alpha) = 0 \text{ for } \alpha > \frac{3}{4},$$

we define  $f : \ell^2 \rightarrow \mathbb{R}$  by

$$f(x) := \langle z, x \rangle \phi(\|x\|).$$

Obviously  $f$  is locally Lipschitz continuous with

$$f^0(x; -z) = 0 \text{ if } \|x\| = 1 \quad \text{and} \quad f^0(0; -z) = -\|z\|^2 \neq 0.$$

For  $h_k = h = -z$  we obtain

$$f^0(A_k; h_k) \geq f^0(x_k; -z) = 0 > -\|z\|^2 = f^0(0; -z) = f^0(A; h)$$

and, again, (2.7) is violated.

For  $X = \mathbb{R}^n$  and  $\varepsilon \geq 0$  the  $\varepsilon$ -generalized gradient of  $f$  at  $x \in X$  is given according to Goldstein [5] by

$$\delta_\varepsilon f(x) := \text{conv} \bigcap_{k=1}^{\infty} \overline{\left\{ f'(y) \mid y \in \overline{B_{\varepsilon + \frac{1}{k}}(x)}, f'(y) \text{ exists} \right\}} \quad (2.12)$$

(where  $f'(x)$  denotes the usual derivative).

**Corollary 2.13.** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be locally Lipschitz continuous. Then*

$$\delta_\varepsilon f(x) = \partial f(\overline{B_\varepsilon(x)}) \quad \text{for all } x \in \mathbb{R}^n, \varepsilon \geq 0.$$

PROOF. Using the characterization of  $\partial f(x)$  in  $\mathbb{R}^n$  (cf. [3, Theorem 2.5.1]), we get

$$\partial f(\overline{B_\varepsilon(x)}) \subset \delta_\varepsilon f(x) \stackrel{(2.12)}{\subset} \bigcap_{k \in \mathbb{N}} \partial f(\overline{B_{\varepsilon+\frac{1}{k}}(x)}) \stackrel{(2.8)}{\subset} \partial f(\overline{B_\varepsilon(x)})$$

(most right inclusion is already an equality by  $\overline{B_\varepsilon(x)} \subset \overline{B_{\varepsilon+\frac{1}{k}}(x)}$  for all  $k \in \mathbb{N}$ ).  $\diamond$

The following statement somehow generalizes Goldstein [5, Propostion 2.8] from  $X = \mathbb{R}^n$  to a general Banach space  $X$ .

**Proposition 2.14.** *Let  $f : X \rightarrow \mathbb{R}$  be locally Lipschitz continuous and let  $A \subset X$  be compact such that  $0 \notin \partial f(x)$  for all  $x \in A$ . Then there exists  $\varepsilon > 0$  and  $\sigma > 0$  such that*

$$\min \{ \|a\| \mid a \in \partial f(\overline{B_\varepsilon(x)}) \} \geq \sigma \quad \text{for all } x \in A. \quad (2.15)$$

PROOF. Notice that there is a minimum in (2.15), since the norm  $\|\cdot\|$  in  $X^*$  is weak\* lower semicontinuous and  $\partial f(\overline{B_\varepsilon(x)})$  is weak\* compact. If the statement would be wrong, then there are  $x_k \in A$  with

$$\min \{ \|a\| \mid a \in \partial f(\overline{B_{\frac{1}{k}}(x_k)}) \} < \frac{1}{k} \quad \text{for all } k \in \mathbb{N}.$$

By compactness of  $A$  we can assume that  $x_k \rightarrow x \in A$ . Moreover we find  $a_k \in \partial f(\overline{B_{\frac{1}{k}}(x_k)})$  with  $a_k \rightarrow 0$ . Since  $d(\overline{B_{\frac{1}{k}}(x_k)}, \{x\}) \rightarrow 0$ , Proposition 2.6 gives the contradiction  $0 \in \partial f(x)$ .  $\diamond$

Let us finally show that  $0 \notin \partial f(x)$  implies some regularity also in a small neighborhood of  $x$ .

**Proposition 2.16.** *Let  $f : X \rightarrow \mathbb{R}$  be locally Lipschitz continuous and let  $0 \notin \partial f(x)$  for some  $x \in X$ . Then there exist  $\varepsilon > 0$  and  $h \in X$  with  $\|h\| = 1$  such that*

$$- \|a\| \leq \langle a, h \rangle \stackrel{(2.4)}{\leq} f^0(A; h) < 0 \quad \text{for all } A \subset B_\varepsilon(x), a \in \partial f(A). \quad (2.17)$$

PROOF. By  $0 \notin \partial f(x)$ , property (2.5) with  $A = \{x\}$  provides the existence of some  $h \in X$  with  $\|h\| = 1$  and  $f^0(x; h) < 0$ . Proposition 2.6 implies

$$\lim_{k \rightarrow \infty} f^0(B_{\frac{1}{k}}(x); h) = f^0(x; h) < 0.$$

Hence we get the most right inequality in (2.17) for some  $\varepsilon > 0$  sufficiently small. With (2.4) we obtain for any  $a \in \partial f(A)$

$$- \|a\| \leq \langle a, h \rangle \leq f^0(A; h)$$

which verifies the assertion.  $\diamond$

### 3 Optimal descent directions

Motivated by Proposition 2.3 (4) we say that  $h \in X$  is a *descent direction* of  $f$  on  $A$  if  $f^0(A; h) < 0$  (cf. also Clarke [4, Ex. 10.7]). We call  $\tilde{h} \in X$  *steepest* or *optimal* descent direction of  $f$  on  $A$  with respect to  $\|\cdot\|$  if

$$\|\tilde{h}\| = 1 \quad \text{and} \quad f^0(A; \tilde{h}) = \min_{\|h\| \leq 1} f^0(A; h) < 0. \quad (3.1)$$

For reflexive Banach spaces the existence of optimal descent directions follows from duality theory.

**Proposition 3.2** (existence of optimal descent directions). *Let  $A \subset X$  be nonempty and let  $f : X \rightarrow \mathbb{R}$  be Lipschitz continuous on a neighborhood of  $A$ . Then:*

(1) *There is some  $\tilde{a} \in \partial f(A)$  such that*

$$\inf_{\|h\| \leq 1} f^0(A; h) = - \min_{a \in \partial f(A)} \|a\| = -\|\tilde{a}\|. \quad (3.3)$$

(2) *For every pair  $(\tilde{a}, \tilde{h}) \in \partial f(A) \times \overline{B_1(0)}$  with*

$$\|\tilde{a}\| = \min_{a \in \partial f(A)} \|a\| \quad \text{and} \quad f^0(A; \tilde{h}) = \min_{\|h\| \leq 1} f^0(A; h) \quad (3.4)$$

*we have*

$$-\|\tilde{a}\| = \langle \tilde{a}, \tilde{h} \rangle = f^0(A; \tilde{h}). \quad (3.5)$$

(3) *If  $X$  is reflexive, then there exists a pair  $(\tilde{a}, \tilde{h}) \in \partial f(A) \times \overline{B_1(0)}$  satisfying (3.4).*

Before providing the proof we still formulate a simple consequence.

**Corollary 3.6.** *Let  $A \subset X$  be nonempty and let  $f : X \rightarrow \mathbb{R}$  be Lipschitz continuous on a neighborhood of  $A$ . Then*

$$\inf_{\|h\| \leq 1} f^0(A; h) < 0 \quad \iff \quad 0 \notin \partial f(A). \quad (3.7)$$

*Moreover, if  $0 \notin \partial f(A)$  and  $(\tilde{a}, \tilde{h}) \in \partial f(A) \times \overline{B_1(0)}$  satisfies (3.4), then  $\tilde{h}$  is an optimal descent direction of  $f$  on  $A$ .*

**PROOF** of Propostion 3.2. For (1) we readily see that we have a minimizer  $\tilde{a} \in \partial f(A)$  and we use (2.4) to get

$$\inf_{\|h\| \leq 1} f^0(A; h) = \inf_{\|h\| \leq 1} \max_{a \in \partial f(A)} \langle a, h \rangle.$$

Since  $\partial f(A)$  is weak\* compact, we can exchange inf and max by Aubin's lopsided minimax theorem (cf. [1, Theorem 6.2.7]) and obtain

$$\inf_{\|h\| \leq 1} f^0(A; h) = \max_{a \in \partial f(A)} \inf_{\|h\| \leq 1} \langle a, h \rangle = \max_{a \in \partial f(A)} -\|a\| = - \min_{a \in \partial f(A)} \|a\|. \quad (3.8)$$

For (2) let  $(\tilde{a}, \tilde{h}) \in \partial f(A) \times \overline{B_1(0)}$  satisfy (3.4). Then

$$-\|\tilde{a}\| \stackrel{(3.3)}{=} f^0(A; \tilde{h}) \stackrel{(2.4)}{=} \max_{a \in \partial f(A)} \langle a, \tilde{h} \rangle \geq \langle \tilde{a}, \tilde{h} \rangle \geq \inf_{\|h\| \leq 1} \langle \tilde{a}, h \rangle = -\|\tilde{a}\|$$

which readily gives (3.5).

For (3) we first observe that there is a minimizer  $\tilde{a} \in \partial f(A)$  satisfying the left part of (3.4) (cf. also (1)). For the right part we use that  $f^0(A; \cdot)$  is convex and continuous and, thus, weakly lower semicontinuous. Since  $X$  is reflexive, there is a minimizer  $\tilde{h}$  on the bounded set  $\overline{B_1(0)}$  by the Weierstraß Theorem.  $\diamond$

The following example shows that there might not be an optimal descent direction in a non-reflexive Banach space  $X$ .

**Example 3.9.** For  $X = c_0$  (sequences  $x = (\xi_i)_{i \in \mathbb{N}}$  in  $\mathbb{R}$  with  $\xi_i \rightarrow 0$  and  $\|x\| = \max_{i \in \mathbb{N}} |\xi_i|$ ) the dual is  $X^* = \ell^1$  (sequences  $x = (\xi_i)_{i \in \mathbb{N}}$  in  $\mathbb{R}$  with  $\|x\| = \sum_{k \in \mathbb{N}} |\xi_k| < \infty$ , cf. [6, Satz II.2.3]). Then  $f = (\frac{1}{2^{i+1}})_{i \in \mathbb{N}} \in X^*$  is a Lipschitz continuous function on  $c_0$  with

$$f(x) = \langle f, x \rangle = \sum_{i \in \mathbb{N}} \frac{\xi_i}{2^{i+1}} \quad \text{and} \quad \|f\| = 1.$$

By linearity,  $\partial f(x) = \partial f(A) = \{f\}$  for all  $x \in c_0$  and all nonempty  $A \subset c_0$ . Hence  $\tilde{a} = f$  always satisfies (3.4) and we have

$$\inf_{\|h\| \leq 1} f^0(A; h) = \inf_{\|h\| \leq 1} \langle f, h \rangle = -1 \quad \text{for all nonempty } A \subset c_0.$$

But there is no  $\tilde{h} \in c_0$  with  $\|\tilde{h}\| \leq 1$  such that  $f^0(A; \tilde{h}) = -1$ , i.e. there is no optimal descent direction. We merely find arbitrarily good approximations as e.g.  $h_k = (\xi_i^k)_{i \in \mathbb{N}} \in c_0$  with

$$\xi_i^k = -1 \text{ for } i \leq k, \quad \xi_i^k = 0 \text{ for } i > k.$$

Obviously  $\|h_k\| = 1$  and, using (2.4), we readily get  $f^0(A; h_k) = \langle f, h_k \rangle \rightarrow -1$ .

**Theorem 3.10** (uniqueness of optimal descent direction). *Let  $X$  be reflexive and let  $X, X^*$  be strictly convex, let  $A \subset X$  be nonempty, and let  $f : X \rightarrow \mathbb{R}$  be Lipschitz continuous on a neighborhood of  $A$ . Then there is a unique  $\tilde{a} \in \partial f(A)$  with*

$$\|\tilde{a}\| = \min_{a \in \partial f(A)} \|a\|. \tag{3.11}$$

Moreover, if  $0 \notin \partial f(A)$ , then there exists a unique optimal descent direction  $\tilde{h}$  of  $f$  which is characterized by

$$\langle \tilde{a}, \tilde{h} \rangle = -\|\tilde{a}\| \quad \text{with} \quad \|\tilde{h}\| = 1. \tag{3.12}$$

**PROOF.** By Propostion 3.2 there are  $\tilde{a}$  and  $\tilde{h}$  satisfying (3.11), (3.12). Since  $X^*$  is strictly convex and  $\partial f(A)$  convex,  $\tilde{a}$  in (3.11) is unique. Since  $X$  is strictly convex,  $\tilde{h}$  in (3.12) is also unique.  $\diamond$



**Remark 3.13.** Notice that for every reflexive Banach space  $X$  there exists an equivalent norm such that  $X$  and  $X^*$  are strictly convex (cf. [2, Theorem III.2.9]). However, since the optimal descent direction  $\tilde{h}$  depends on the norm in general,  $\tilde{h}$  might change by a change of norm. In particular, the derivative  $f'(x)$  of a smooth function  $f$  is independent of an equivalent norm, but the optimal descent direction  $\tilde{h}$  on  $A = \{x\}$  might be different for an equivalent norm.

**Example 3.14.** We consider  $X := \mathbb{R}^2$  with the non strictly convex norms  $\|x\|_1$  (1-norm) and  $\|x\|_\infty$  (maximum norm). We will demonstrate that the selection of a descent direction needs more care in a reflexive but not strictly convex space where (3.12) is not sufficient for the selection.

- (1) Let  $X = (\mathbb{R}^2, \|\cdot\|_1)$  and, thus, its dual  $X^* = (\mathbb{R}^2, \|\cdot\|_\infty)$ . We define  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  by

$$f(x_1, x_2) = x_1 + |x_2|.$$

With  $x = (1, 0)$  and  $A = \{x\}$  we get

$$\partial f(A) = \partial f(x) = \{(1, \lambda) \mid \lambda \in [-1, 1]\}.$$

Obviously any  $\tilde{a} \in \partial f(A)$  satisfies (3.11) and, with Proposition 3.2,

$$-1 = -\|\tilde{a}\|_\infty = -\min_{a \in \partial f(A)} \|a\|_\infty = \min_{\|\tilde{h}\|_1 \leq 1} f^0(A; \tilde{h}).$$

Taking  $\tilde{a} = (1, 1) \in \partial f(A)$  we obtain (3.12) e.g. for  $\tilde{h} = (0, -1)$ . However  $f$  is strictly increasing in the directions  $\pm \tilde{h}$  and  $f^0(A; \tilde{h}) = 1$ . Hence  $\tilde{h}$  is not a descent direction and (3.12) is not sufficient for their selection. Obviously  $\tilde{h} = (-1, 0)$  is an optimal descent direction on  $A$  and satisfies (3.12) for every  $\tilde{a} \in \partial f(A)$ .

- (2) Let  $X = (\mathbb{R}^2, \|\cdot\|_\infty)$  and, thus, its dual  $X^* = (\mathbb{R}^2, \|\cdot\|_1)$ . We define  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  by

$$f(x, y) = \frac{1}{2}(x + y + |x - y|).$$

For  $x = (0, 0)$  and  $A = \{x\}$  we have

$$\partial f(A) = \partial f(x) = \{(\lambda, 1 - \lambda) \mid \lambda \in [0, 1]\}.$$

Again any  $\tilde{a} \in \partial f(A)$  satisfies (3.11) and, with Proposition 3.2,

$$-1 = -\|\tilde{a}\|_1 = -\min_{a \in \partial f(A)} \|a\|_1 = \min_{\|\tilde{h}\|_\infty \leq 1} f^0(A; \tilde{h}).$$

With  $\tilde{a} = (1, 0) \in \partial f(A)$  and  $\tilde{h} = (-1, 1)$  we have (3.12), but in both directions  $\pm \tilde{h}$  function  $f$  is strictly increasing and  $f^0(A; \tilde{h}) = 1$ . Hence  $\tilde{h}$  is not a descent direction and also here (3.12) is not sufficient for their selection. We readily verify that  $\tilde{h} = -(1, 1)$  is an optimal descent direction on  $A$  and satisfies (3.12) for every  $\tilde{a} \in \partial f(A)$ .

As a consequence of Theorem 3.10 we obtain that descent directions are stable.

**Corollary 3.15** (stability of descent directions). *Let the assumptions of Theorem 3.10 with  $0 \notin \partial f(A)$  be satisfied, let  $\tilde{a}, \tilde{h}$  be as there, and let  $L$  be the Lipschitz constant of  $f$  on a neighborhood of  $A$ . Then every  $h \in X$  with  $\|h - \tilde{h}\| < \frac{\|\tilde{a}\|}{L}$  is a descent direction on  $A$ .*

PROOF. Let  $h \in X$  be as in the statement. By (2.4) there is  $a \in \partial f(A)$  such that

$$\begin{aligned} f^0(A; h) &= \langle a, h \rangle = \langle a, h - \tilde{h} \rangle + \langle a, \tilde{h} \rangle \stackrel{\text{Prop. 2.3}}{\leq} L\|h - \tilde{h}\| + f^0(A; \tilde{h}) \\ &< \|\tilde{a}\| + f^0(A; \tilde{h}) \stackrel{(3.5)}{=} 0. \end{aligned}$$

Hence  $h$  is a descent direction.  $\diamond$

The stability of descent directions allows to work with approximations of an optimal descent direction.

**Corollary 3.16** (approximation of an optimal descent direction). *Let  $X$  be uniformly convex (or finite dimensional and strictly convex) and let  $X^*$  be strictly convex. Moreover let  $A \subset X$  be nonempty, let  $f : X \rightarrow \mathbb{R}$  be Lipschitz continuous on a neighborhood of  $A$  with  $0 \notin \partial f(A)$ , and let  $\tilde{a} \in \partial f(A)$  be as in Theorem 3.10. Then for any  $\delta \in ]0, 1[$  there is some  $\tau > 0$  such that for every  $a' \in \partial f(A)$  with*

$$\|a'\| \leq \min_{a \in \partial f(A)} \|a\| + \tau \quad (= \|\tilde{a}\| + \tau)$$

the unique  $h' \in X$  satisfying

$$\langle a', h' \rangle = -\|a'\| \quad \text{with} \quad \|h'\| = 1 \quad (3.17)$$

is a descent direction on  $A$  with

$$\left( \max_{a \in \partial f(A)} \langle a, h' \rangle = \right) \quad f^0(A; h') < -\delta \|\tilde{a}\|.$$

Recall that uniformly convex Banach spaces are reflexive (cf. [2, Theorem II.2.9]) and, thus, the results of Theorem 3.10 are available in the corollary.

PROOF . The usual dual mapping  $j : X^* \setminus \{0\} \rightarrow \{x \in X \mid \|x\| = 1\}$  is given by

$$\langle a, j(a) \rangle = \|a\|.$$

Hence (3.17) just means  $h' = -j(a')$  and (3.12) gives  $\tilde{h} = -j(\tilde{a})$  (notice that  $0 \notin \partial f(A)$ ). If the assertion would be false, then there are  $\delta > 0$  and  $a'_k \in \partial f(A)$  such that

$$\|a'_k\| \leq \|\tilde{a}\| + \frac{1}{k} \quad \text{and} \quad f^0(A; -j(a'_k)) \geq -\delta \|\tilde{a}\| \quad \text{for all } k \in \mathbb{N}. \quad (3.18)$$

By (2.4) there are  $a_k \in \partial f(A)$  with  $f^0(A; -j(a'_k)) = \langle a_k, -j(a'_k) \rangle$ . Since  $X$  is reflexive and  $\partial f(A)$  weak\*-compact, we have up to a subsequence that

$$a'_k \xrightarrow{*} : a' \in \partial f(A) \quad \text{and} \quad a_k \xrightarrow{*} : a \in \partial f(A).$$

With (3.11) we obtain

$$\|\tilde{a}\| \leq \|a'\| \leq \liminf_{k \rightarrow \infty} \|a'_k\| \leq \limsup_{k \rightarrow \infty} \|a'_k\| \stackrel{(3.18)}{\leq} \|\tilde{a}\|.$$

Since  $\tilde{a}$  is uniquely determined by (3.11), we get  $a' = \tilde{a}$  and  $\|a'_k\| \rightarrow \|\tilde{a}\|$ . Uniform convexity (or finite dimension) of  $X$  implies  $a'_k \rightarrow \tilde{a}$ . Reflexivity of  $X$  and strict convexity of  $X$  and  $X^*$  imply continuity of  $j$  (cf. [2, Prop. II.5.5]) and, thus,

$$-\delta\|\tilde{a}\| \stackrel{(3.18)}{\leq} \liminf_{k \rightarrow \infty} f^0(A; -j(a'_k)) = \lim_{k \rightarrow \infty} \langle a_k, -j(a'_k) \rangle = \langle a, \tilde{h} \rangle \stackrel{(2.4)}{\leq} f^0(A; \tilde{h}) \stackrel{(3.5)}{=} -\|\tilde{a}\|.$$

But this is a contradiction and the assertion follows.  $\diamond$

Let us finally demonstrate with a simple but typical example how the introduced optimal descent direction can improve numerical descent methods.

**Example 3.19.** For  $X = \mathbb{R}^2$  equipped with the Euclidean norm we consider

$$f(x_1, x_2) := |x_1| + \alpha|x_2| \quad \text{with} \quad 0 < \alpha \ll 1$$

Here steepest descent methods starting from  $(x_1, x_2)$  with  $x_2 \gg |x_1|$  easily approach (but usually do not reach) the axis  $\{x_1 = 0\}$  after a few steps. Then they highly oscillate around that axis, since the gradients switch between  $(\pm 1, \alpha)$ . But with a nonsmooth strategy we would choose a suitable ball  $A = B_\varepsilon(x)$  at an iteration point  $x$  near  $\{x_1 = 0\}$ . If  $0 \in B_\varepsilon(x)$ , then  $0 \in \partial f(B_\varepsilon(x))$  and we either stop the algorithm or we decrease “step size”  $\varepsilon$ . If otherwise  $0 \notin B_\varepsilon(x)$ , then

$$\partial f(B_\varepsilon(x)) = \{(\lambda, \alpha) \mid \lambda \in [-1, 1]\}.$$

Obviously  $\tilde{a} = (0, \alpha)$  has the smallest norm in  $\partial f(B_\varepsilon(x))$  and the corresponding optimal descent direction on  $B_\varepsilon(x)$  according to Theorem 3.10 is  $\tilde{h} = (0, -1)$ . Now a descent step or a line-search in direction  $\tilde{h}$  goes quite directly to the minimizer  $(0, 0)$ .

## References

- [1] J. Aubin, I. Ekeland. *Applied Nonlinear Analysis*. John Wiley & Sons, New York 1984.
- [2] I. Cioranescu. *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*. Kluwer Academic Publishers, Dordrecht 1990.
- [3] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York 1983.
- [4] F. Clarke: *Functional Analysis, Calculus of Variations and Optimal Control*. Springer, London 2013.
- [5] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Math. Program.* 13 (1977) 1422.
- [6] D. Werner. *Funktionalanalysis*. 5. Aufl., Springer, Berlin 2005.