On generalized gradients and optimization *

Erik J. Balder

1 Introduction

There exists a calculus for general nondifferentiable functions that englobes a large part of the familiar subdifferential calculus for convex nondifferentiable functions [1]. This development started with F.H. Clarke, who introduced a *generalized gradient* for functions that are *locally Lipschitz*, but (possibly) nondifferentiable. Generalized gradients turn out to be the subdifferentials, in the sense of convex analysis [1], of generalized directional derivative functions that are canonically associated to the locally Lipschitz functions under consideration. The key point to note is that such generalized directional derivative functions are *automatically convex*, even when the original locally Lipschitz functions are not convex. Two important special cases of functions f that are locally Lipschitz near some point $x_0 \in \mathbb{R}^n$ and their generalized gradients, denoted by $\bar{\partial} f(x_0)$, are as follows: (i) f is continuously differentiable at x_0 and (ii) f is convex on $\mathbb{R}^n$ and $x_0 \in \text{int dom } f$. Then

$$\bar{\partial} f(x_0) = \begin{cases} \{\nabla f(x_0)\} & \text{in case } (i) \\ \partial f(x_0) & \text{in case } (ii) \end{cases}$$

See Propositions 2.7 and 2.8. The corresponding calculus for generalized gradients manages to preserve the relevant parts of the Moreau-Rockafellar and Dubovitskii-Milyutin theorems (see Theorems 2.9 and 2.10 below) that we know in the convex setting from [1]. For problems with only inequality constraints this calculus leads, by means of a pointwise maximum function of Dubovitskii-Milyutin-type (as employed in [1]), to a "not necessarily convex or smooth" version of the Karush-Kuhn-Tucker theorem (Theorem 3.3).[1]

If equality constraints are introduced as well, then a new method is required to obtain such a theorem (Theorem 3.5); this is based on Ekeland's Theorem 3.4. This time, the pointwise maximum function of Dubovitskii-Milyutin-type will also have to involve the equality constraints. This contrasts with [1], but recall that the convex framework could only deal with equality constraints that were linear/affine. In a

*LNMB Ph.D. course "Convex Analysis for Optimization", November, 2008.

[1]However, the set S over which one minimizes is treated by means of the distance function dist_S instead of the characteristic function χ_S; this changes the nature of the obtuse angle property (OAP) somewhat.

certain sense the present approach by means of generalized gradients abridges what was done for the "convex" KKT theorem in [1] and for the "smooth" KKT theorem in Chapter 4 in [2]: Theorem 3.5 comes very close – without doing so precisely, i.e., in all respects – to generalizing both of these results simultaneously. Even if one restricts oneself beforehand to "smooth" optimization problems, generalized gradients can be viewed an alternative, quite useful device to derive important results, such as the "smooth" KKT theorem. Good references for these developments are [3] and [4].

2 Generalized gradients and their calculus

Definition 2.1 (local Lipschitz continuity) A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is *Lipschitz near* a point $x_0 \in \text{int dom } f$ if there exist $K \geq 0$, the so-called Lipschitz constant, such that

$$|f(x) - f(x')| \leq K|x - x'| \text{ for all } x, x' \in B_\delta(x_0),$$

where $\delta > 0$ sufficiently small so as to have $B_\delta(x_0) := \{x \in \mathbb{R}^n : |x - x_0| < \delta\} \subset$ dom f. A *locally Lipschitz* function is a function that is Lipschitz near every point in $\mathbb{R}^n$.

Since all our optimality considerations will be *local* and at points of local Lipschitz continuity that lie in the interior of effective domains of functions, no real purpose is served by retaining functions that can take the value $+\infty$. Therefore, we shall only consider real-valued function from now on, except when dealing with convex functions.

Example 2.2 (*i*) A convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is Lipschitz near every point in int dom f. To see this, let $x_0 \in \text{int dom } f$ be arbitrary. Observe that f is continuous at x_0 by Lemma 2.16 of [1]. Therefore, there exist a and b in $\mathbb{R}$ and $\epsilon > 0$ such that $a \leq f(x) \leq b$ for all $x \in B_\epsilon(x_0)$. Let $0 < \epsilon' < \epsilon$ and $x, x' \in B_{\epsilon'}(x_0)$ be arbitrary. Then $x' = \lambda z + (1 - \lambda)x$, with $z := x' + (\epsilon - \epsilon')(x' - x)/|x' - x|$ and $\lambda := |x' - x|/(|x' - x| + \epsilon - \epsilon')$. Since $|z - x_0| \leq |x' - x_0| + \epsilon - \epsilon' < \epsilon$, we have $z \in B_\epsilon(x_0)$. By convexity of f

$$f(x') - f(x) \leq \lambda(f(z) - f(x)) \leq \lambda(b - a) \leq K|x' - x|,$$

where $K := (b - a)/(\epsilon - \epsilon')$. By reversing the roles of x and x', it then follows that $|f(x') - f(x)| \leq K|x' - x|$ for all $x', x \in B_{\epsilon'}(x_0)$.

(*ii*) A function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz near every point in which f is continuously differentiable.

Exercise 2.1 Prove part (*ii*) of the above Example 2.2 by fixing two points x, x' and applying the usual (scalar) intermediate value theorem. Recall that f is said to be *continuously differentiable* at a point x_0 if all its partial derivatives are continuous functions in some neighborhood of x_0 (in turn, this implies that f is differentiable at x_0).

Definition 2.3 (generalized gradient) Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $x_0 \in \mathbb{R}^n$. Then Clarke's *generalized directional derivative* of f at x_0 in the direction $d \in \mathbb{R}^n$ is defined by

$$f^o(x_0; d) := \limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \frac{f(x_0 + h + \lambda d) - f(x_0 + h)}{\lambda} := \lim_{\delta \downarrow 0} \sup_{h \in B_\delta(0), \lambda \in (0,\delta)} \frac{f(x_0 + h + \lambda d) - f(x_0 + h)}{\lambda}.$$

Also, Clarke's *generalized gradient* of f at x_0 is defined by

$$\bar{\partial} f(x_0) := \{\xi \in \mathbb{R}^n : f^o(x_0; d) \geq \xi^t d \text{ for all } d \in \mathbb{R}^n\}.$$

Observe a fundamental difference with the ordinary directional derivative (see Definition 2.13 in [1]): in that definition the "base point" for taking differences was the *fixed* vector x_0, but now it is the *variable* vector $x_0 + h \to x_0$. Precisely this difference is responsible for "automatic convexity" of the generalized directional derivative (see Lemma 2.6) that holds even even when f itself is nonconvex.

Exercise 2.2 Prove that $f^o(x_0; d)$ is the maximum of

$$l((h_k, \lambda_k)_1^\infty) := \limsup_{k \to \infty} \frac{f(x_0 + h_k + \lambda_k d) - f(x_0 + h_k)}{\lambda_k}$$

over all sequences $(h_k, \lambda_k)_1^\infty$ with $h_k \to 0$ and $\lambda_k \downarrow 0$. Here *maximum* means that there actually exists a sequence $(\bar{h}_k, \bar{\lambda}_k)_1^\infty$ for which $f^o(x_0; d) = l((\bar{h}_k, \bar{\lambda}_k)_1^\infty)$.

Example 2.4 (a) Consider $n = 1$ and $f(x) := |x|$. For $x_0 := 0$ we have

$$f^o(0; d) = \limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \left[|\frac{h}{\lambda} + d| - |\frac{h}{\lambda}| \right].$$

By the triangle inequality, this gives $f^o(0; d) \leq |d|$, and it is easily seen that the majorant $|d|$ is also attainable as a limit:

$$|d| = \lim_k \left[|\frac{h_k}{\lambda_k} + d| - |\frac{h_k}{\lambda_k}| \right]$$

for $\lambda_k := k^{-1}$ and $h_k := k^{-2}$. Hence, $f^o(0; d) = |d|$ (see Exercise 2.2), and it follows that $\bar{\partial} f(0) = [-1, +1]$, which coincides with the classical subgradient of f in the sense of convex analysis.

(b) Consider $n = 1$ and $f(x) := -|x|$. Then for $x_0 := 0$ we have

$$f^o(0; d) = \limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \left[|\frac{h}{\lambda}| - |\frac{h}{\lambda} + d| \right],$$

which is again at most $|d|$ by the triangle inequality. Again, but now in a diametrically opposite way, the majorant $|d|$ is attainable as a limit:

$$|d| = \lim_k \left[-|\frac{h_k}{\lambda_k} + d| + |\frac{h_k}{\lambda_k}| \right]$$

by taking $\lambda_k := k^{-2}$ and $h_k := -k^{-1}$ if $d > 0$, and by taking $\lambda_k := k^{-2}$ and $h_k := k^{-1}$ if $d < 0$. Hence, $f^o(0; d) = |d|$, and it follows that $\bar{\partial} f(0) = [-1, +1]$. So the concave function $x \mapsto -|x|$ has the same generalized gradient as the convex function $x \mapsto |x|$ in part (a).

This shows clearly that the theory involving generalized gradients in the sense of Clarke is much less unilateral in nature than convex analysis. As one more indication of this, we observe that, although formally it would be possible to consider extended real-valued functions in Definition 2.3 and further on, the fact that the local Lipschitz property plays such an important role makes such adaptations so obvious that we prefer instead to keep our functions *real-valued* from now on.

Exercise 2.3 Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $x_0 \in \mathbb{R}^n$. Let $\xi \in \bar{\partial} f(x_0)$ and suppose that $f(x_0) \neq 0$. Prove that $\bar{\partial}|f|(x_0) \subset \mathrm{sign}(f(x_0))\bar{\partial} f(x_0)$.

Theorem 2.5 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $x_0 \in \mathbb{R}^n$ with local Lipschitz constant K. Then the following hold:*
(i) $\bar{\partial} f(x_0) \neq \emptyset$.
(ii) $|\xi| \leq K$ *for all* $\xi \in \bar{\partial} f(x_0)$.
(iii) $f^0(x_0; d) = \sup_{\xi \in \bar{\partial} f(x_0)} \xi^t d$ *for every* $d \in \mathbb{R}^n$.

Lemma 2.6 (automatic convexity of generalized directional derivative functions) *In Theorem 2.5 the function $p : d \mapsto f^o(x_0; d)$ is positively homogeneous and subadditive (whence in particular convex). Also, $|p(d)| \leq K|d|$ for all $d \in \mathbb{R}^n$.*

PROOF. It is trivial to see that $p(\alpha d) = \alpha p(d)$ for every d and $\alpha \geq 0$ (positive homogeneity). Also, for every pair of directions d and d' it is elementary to see that

$$\limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \frac{f(x_0 + h + \lambda d + \lambda d') - f(x_0 + h)}{\lambda} \leq \limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \frac{f(x_0 + h + \lambda d + \lambda d') - f(x_0 + h + \lambda d')}{\lambda} + $$
$$+ \limsup_{\substack{h \to 0 \\ \lambda \downarrow 0}} \frac{f(x_0 + h + \lambda d') - f(x_0 + h)}{\lambda},$$

whence $p(d+d') \leq p(d)+p(d')$ (subadditivity). Together with the positive homogeneity observed above, it therefore follows that $p(\alpha d+(1-\alpha)d') \leq p(\alpha d)+p((1-\alpha)d') = \alpha p(d) + (1 - \alpha)p(d')$ for every $\alpha \in [0, 1]$ (convexity). Also, for all d, h in $\mathbb{R}^n$ and all $\lambda > 0$ the given local Lipschitz continuity of f gives

$$\left|\frac{f(x_0 + h + \lambda d) - f(x_0 + h)}{\lambda}\right| \leq K|d|,$$

so $p(d) \leq K|d|$ follows immediately. QED

PROOF OF THEOREM 2.5. Denote again $p(d) := f^o(x_0; d)$. By Lemma 2.6 p is a convex real-valued function. So by Lemma 2.16 of [1] it follows that $\partial p(0)$, the subdifferential of p at 0 in the sense of convex analysis, is nonempty, i.e., there exists $\xi \in \mathbb{R}^n$ such that $p(d) \geq p(0) + \xi^t d = \xi^t d$ for all $d \in \mathbb{R}^n$. Hence, $\xi \in \bar{\partial} f(x_0)$. This proves (i). Since $\bar{\partial} f(x_0) = \partial p(0)$, part (ii) follows immediately from Lemma 2.6 and part (iii) follows from Theorem 2.15 in [1], which says that the directional derivative of p at 0 (in the sense of convex analysis) satisfies for every d

$$p'(0; d) = \sup_{\xi \in \partial p(0)} \xi^t d.$$

Here $p'(0; d) = p(d)$, by $p(0) = 0$ and positive homogeneity of p (Lemma 2.6), and $\partial p(0) = \bar{\partial} f(x_0)$. QED

We shall now establish connections between the generalized gradient and two classical concepts: the ordinary gradient and the subdifferential of convex analysis.

Proposition 2.7 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable at $x_0 \in \mathbb{R}^n$. Then*

$$\bar{\partial} f(x_0) = \{\nabla f(x_0)\},$$

i.e., then Clarke's generalized gradient coincides with the gradient of f.

Observe that if f is in addition convex in Proposition 2.7, then we already know its result from Proposition 2.6 of [1], in view of Proposition 2.8 below.

Proof of Proposition 2.7. For every d and h in $\mathbb{R}^n$, with $|h|$ sufficiently small, an application of the mean value theorem to $\lambda \mapsto f(x_0 + h + \lambda d)$ gives for every $\lambda > 0$ the existence of $\lambda' \in (0, \lambda)$ such that

$$\frac{f(x_0 + h + \lambda d) - f(x_0 + h)}{\lambda} = d^t \nabla f(x_0 + h + \lambda' d).$$

Since $x \mapsto \nabla f(x)$ is continuous at x_0, this easily gives $f^o(x_0; d) = d^t \nabla f(x_0)$. So by Theorem 2.5$(iii)$

$$\sup_{\xi \in \bar{\partial} f(x_0)} \xi^t d = d^t \nabla f(x_0) = \sup_{\xi \in \{\nabla f(x_0)\}} \xi^t d,$$

and we can finish as in the proof of Theorem 2.5. QED

Proposition 2.8 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then for every $x_0 \in \mathbb{R}^n$*

$$\bar{\partial} f(x_0) = \partial f(x_0),$$

i.e., then Clarke's generalized gradient coincides with the subdifferential of f.

Proof of Proposition 2.8. Let $d \in \mathbb{R}^n$ be arbitrary. First, observe that by taking $h = 0$ in Definition 2.3 we get trivially $f'(x_0; d) \leq f^o(x_0; d)$. Next, we also prove the converse inequality. Note that by Definition 2.3, for any $c > 0$

$$f^o(x_0; d) = \lim_{\delta \downarrow 0} \sup_{h \in B_{\delta c}(0)} \sup_{\lambda \in (0, \delta)} \frac{f(x_0 + h + \lambda d) - f(x_0 + h)}{\lambda}.$$

But we know that for a convex function the difference quotients are monotone (see Proposition 2.14 in [1]); hence,

$$f^o(x_0; d) \leq \lim_{\delta \downarrow 0} \sup_{h \in B_{\delta c}(0)} \frac{f(x_0 + h + \delta d) - f(x_0 + h)}{\delta}.$$

By Example 2.2, f is Lipschitz near every point, so there exists $K > 0$ such that

$$\left|\frac{f(x_0 + h + \delta d) - f(x_0 + h)}{\delta} - \frac{f(x_0 + \delta d) - f(x_0)}{\delta}\right| \leq \frac{2K|h|}{\delta} \leq 2Kc$$

for all $h \in B_{\delta c}(0)$. Combined with the previous inequality and the triangle inequality this gives

$$f^o(x_0; d) \leq \lim_{\delta \downarrow 0} \frac{f(x_0 + \delta d) - f(x_0)}{\delta} + 2Kc = f'(x_0; d) + 2Kc.$$

Letting $c \downarrow 0$, we thus find $f^o(x_0; d) \leq f'(x_0; d)$. We conclude from the preceding that for every $d \in \mathbb{R}^n$

$$f^o(x_0; d) = f'(x_0; d),$$

which, by Theorem 2.5(*iii*) and the corresponding property of convex functions (i.e., Theorem 2.15 of [1]), can also be written as

$$\sup_{\xi \in \bar{\partial} f(x_0)} \xi^t d = \sup_{\xi \in \partial f(x_0)} \xi^t d. \tag{1}$$

Observe that both $\bar{\partial} f(x_0)$ and $\partial f(x_0)$ are closed (and in fact compact) convex sets. Let ξ_0 be arbitrary in $\bar{\partial} f(x_0)$. Suppose that ξ_0 did not belong to the compact convex set $\partial f(x_0)$. Then we could separate strictly by Theorem A.2 in [1]: there would be $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that

$$\xi_0^t d > \alpha \geq \sup_{\xi \in \partial f(x_0)} \xi^t d = f^o(x_0; d),$$

where the identity on the right holds by Theorem 2.5(*iii*). But by that same theorem, $\xi_0^t d \leq f^o(x_0; d)$, so we would get a contradiction to (1). Hence, it follows that ξ_0 belongs to $\partial f(x_0)$, and we conclude that $\bar{\partial} f(x_0) \subset \partial f(x_0)$. One can prove the opposite inclusion in precisely the same way. QED

Exercise 2.4 Give an alternative proof of the last step in the proof by means of conjugate functions and the Fenchel-Moreau theorem. *Hint:* Note that (1) expresses the equality of two conjugate functions.

Some basic calculus rules for generalized gradients will now be stated. They are rather similar to rules developed for the subgradients/subdifferentials of convex analysis in [1]. The following rule retains the "interesting" inclusion of the Moreau-Rockafellar theorem:

Theorem 2.9 *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $x_0 \in \mathbb{R}^n$ and let $\alpha > 0$. Then*

$$\bar{\partial}(\alpha f)(x_0) = \alpha \bar{\partial} f(x_0),$$

$$\bar{\partial}(f+g)(x_0) \subset \bar{\partial} f(x_0) + \bar{\partial} g(x_0).$$

PROOF. For every $d \in \mathbb{R}^n$ the identity $(\alpha f)^o(x_0; d) = \alpha f^o(x_0; d)$ is obvious, and then the first result immediately follows. Also, for every $d \in \mathbb{R}^n$ the inequality

$$(f+g)^o(x_0; d) \leq f^o(x_0; d) + g^o(x_0; d)$$

follows by elementary subadditivity properties of the limes superior. Now let ξ_0 be arbitrary in $\bar{\partial}(f+g)(x_0)$. Suppose that ξ_0 did not belong to the compact convex set $\bar{\partial} f(x_0) + \bar{\partial} g(x_0)$. Then the strict separation Theorem A.2 of [1] applies: there would exist $d \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that

$$\xi_0^t d > \beta \geq \sup_{\xi \in \bar{\partial} f(x_0), \xi' \in \bar{\partial} g(x_0)} (\xi + \xi')^t d.$$

Since

$$\sup_{\xi \in \bar{\partial} f(x_0), \xi' \in \bar{\partial} g(x_0)} (\xi + \xi')^t d = \sup_{\xi \in \bar{\partial} f(x_0)} \xi^t d + \sup_{\xi' \in \bar{\partial} g(x_0)} \xi'^t d = f^o(x_0; d) + g^o(x_0; d)$$

(here the last identity holds by Theorem 2.5(*iii*)), we get $\xi_0^t d > f^o(x_0; d) + g^o(x_0; d)$. But by that same result also $\xi_0^t d \leq (f+g)^o(x_0; d)$, and so we have a contradiction with the above inquality. It therefore follows that ξ_0 belongs to $\bar{\partial} f(x_0) + \bar{\partial} g(x_0)$, and, since ξ_0 was arbitrary in $\bar{\partial}(f+g)(x_0)$, the desired inclusion has been proven. QED

Exercise 2.5 Show by means of a counterexample that in the second part of Theorem 2.9 it is in general not possible to have the opposite inclusion "$\supset$" (note: this is in contrast to the Moreau-Rockafellar theorem).

As the second calculus rule, we consider the following analogue for generalized gradients of the "interesting" inclusion in the Dubovitskii-Milyutin theorem:

Theorem 2.10 *Let $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $x_0 \in \mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by*

$$f(x) := \max_{1 \leq i \leq m} f_i(x)$$

and let $I(x_0)$ be the (nonempty) set of all $i \in \{1, \ldots, m\}$ for which $f_i(x_0) = f(x_0)$. Then

$$\bar{\partial} f(x_0) \subset \mathrm{co}\left(\cup_{i \in I(x_0)} \bar{\partial} f_i(x_0)\right).$$

PROOF. First, we have that for $|h|$ sufficiently small $I(x_0 + h) \subset I(x_0)$. Indeed, if $j \notin I(x_0)$, then continuity of f_j and f implies that there exists $\delta_j > 0$ such that $f_j(x_0 + h) < f(x_0 + h)$ for $|h| < \delta_j$. Hence, $j \notin I(x_0 + h)$. So $I(x_0 + h) \subset I(x_0)$ for $|h| < \min_{j \notin I(x_0)} \delta_j$. It follows that for every $d \in \mathbb{R}^n$

$$f(x_0+h+\lambda d) - f(x_0+h) \leq \max_{i \in I(x_0+h+\lambda d)} f_i(x_0+h+\lambda d) - f_i(x_0+h) \leq \max_{i \in I(x_0)} f_i(x_0+h+\lambda d) - f_i(x_0+h),$$

where the last inequality holds by the above inclusion $I(x_0+h+\lambda d) \subset I(x_0)$, provided that $|h|$ and λ are sufficiently small. After division by λ and taking the limes superior in the usual way, we find

$$f^o(x_0; d) \leq \max_{i \in I(x_0)} f_i^o(x_0; d) \text{ for every } d \in \mathbb{R}^n. \tag{2}$$

Let ξ_0 be arbitrary in $\bar{\partial} f(x_0)$. Suppose that ξ_0 did not belong to the compact convex set $C := \text{co } \cup_{i \in I(x_0)} \bar{\partial} f_i(x_0)$. Then we could separate strictly by Theorem A.2 in [1]: there would exist $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that

$$\xi_0^t d > \alpha \geq \sup_{\xi \in C} \xi^t d = \max_{i \in I(x_0)} \sup_{\xi \in \bar{\partial} f_i(x_0)} \xi^t d = \max_{i \in I(x_0)} f_i^o(x_0; d),$$

where the identity on the right holds by Theorem 2.5(*iii*). But by that same theorem, $\xi_0^t d \leq f^o(x_0; d)$, so we would get a contradiction to (2). Hence we conclude that ξ_0 belongs to co $\cup_{i \in I(x_0)} \bar{\partial} f_i(x_0)$. QED

Exercise 2.6 Show by means of a counterexample that in Theorem 2.10 it is in general not possible to have the opposite inclusion "$\supset$" (note: this is in contrast to the Dubovitskii-Milyutin theorem).

3 Generalized gradients and the KKT theorem

It is not surprising that Theorem 2.10 *à la Dubovitskii-Milyutin* will again serve us well in connection with the inequality constraints in proving the Lagrangain inclusion and complementary slackness relationship. But, in connection with the equality constraints, the counterpart of the obtuse angle property can *no longer* be expected to come from gradients connected with indicator functions, since indicator functions completely lack the local Lipschitz property that Clarke's generalized gradient calculus requires. First, we discuss the analogue of the "small" KKT Theorem 2.10 in [1]. Recall that this is the "set constraints only" version of the KKT theorem. Below we write $\text{dist}_S(x) := \inf_{x' \in S} |x - x'|$.

Theorem 3.1 ("small" KKT) *Let S be a closed subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\bar{x} \in S$. Consider the optimization problem*

$$(P) \quad \inf_{x \in S} f(x).$$

If $\bar{x}$ is a local optimal solution of (P), then there exists $\bar{\eta} \in \mathbb{R}^n$ such that

$$0 \in \bar{\partial} f(\bar{x}) + \bar{\eta}$$

and

$$\bar{\eta} \in \cup_{t>0} t\, \bar{\partial} \text{dist}_S(\bar{x}) \quad \textbf{(normal cone property)}.$$

A notable difference with its counterpart in [1] is that this result has *only* necessary conditions for (local) optimality. In contrast, Theorem 2.10 in [1] contained also sufficient conditions for (global) optimality, which was possible because of its convexity conditions. This pattern will be reproduced in the KKT theorems that follow below: only necessary conditions for (local) optimality can be stated within the present locally Lipschitz framework.

PROOF. By hypothesis, there exists $\epsilon > 0$ such that $f(\bar{x} \leq f(x)$ for all $x \in S \cap B_\epsilon(\bar{x})$ and such that f has the local Lipschitz property on $B_\epsilon(\bar{x})$ with Lipschitz constant $K > 0$.

Case 1: $S = \mathbb{R}^n$. In this case we simply take $h = 0$ in Definition 2.3 and get

$$f^o(\bar{x}; d) \geq \lim_{\delta \downarrow 0} \sup_{\lambda \in (0,\delta)} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \limsup_{\lambda \downarrow 0} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} \geq 0 = 0^t d$$

for every d, since $f(\bar{x} + \lambda d) \geq f(\bar{x})$ for $\lambda < \epsilon/|d|$. Hence,

$$0 \in \bar{\partial} f(\bar{x}).$$

Since $\text{dist}_{\mathbb{R}^n}(x) = 0$ for all x, we finish by setting $\bar{\eta} := 0$.

Case 2: general case. We reduce this to the unconstrained situation of Case 1 as follows: We claim that $\bar{x}$ is also a local optimal solution of the unconstrained auxiliary problem

$$(P') \quad \inf_{x \in B_{\epsilon/2}(\bar{x})} \left[f(x) + K \text{dist}_S(x) \right].$$

For suppose that there existed an $x \in B_{\epsilon/2}(\bar{x})$ (note already: this implies $\text{dist}_S(x) < \epsilon/2$) for which $f(x) + K\text{dist}_S(x) < f(\bar{x})$; then, by [2, Theorem 2.4.1] there would exist $x' \in S$ such that $|x - x'| = \text{dist}_S(x) < \epsilon/2$. Notice that then $|x' - \bar{x}| < \epsilon$ by the triangle inequality. Yet the local Lipschitz property gives $f(x') \leq f(x) + K|x' - x| = f(x) + K\text{dist}_S(x) < f(\bar{x})$, which contradicts the given local optimality property of $\bar{x}$. This proves our claim about (P'). By Case 1 and Theorem 2.9 this implies

$$0 \in \bar{\partial}(f + K\text{dist}_S)(\bar{x}) \subset \bar{\partial} f(\bar{x}) + K\bar{\partial}\text{dist}_S(\bar{x}).$$

Hence, the result follows. QED

Remark 3.2 (return of the (OAP)) The set $\cup_{t>0} t\,\bar{\partial}\text{dist}_S(\bar{x})$ is called the *normal cone* to S at $\bar{x}$. If S is additionally *convex*, then this normal cone consists precisely of all vectors having the obtuse angle property that we used in [1]. For then dist_S is obviously convex, so $\bar{\partial}\text{dist}_S(\bar{x}) = \partial\text{dist}_S(\bar{x})$ (Proposition 2.8), and from this the above fact follows easily.

Exercise 3.1 Give a complete proof of the fact mentioned in the above remark.

We now can prove a first version of the KKT theorem by means of the "small" KKT Theorem 3.1 and the Dubovitskii-Milyutin-like Theorem 2.10. This development is very similar to the proof of Theorem 3.1 in [1], which is the convex counterpart of the following result:

Theorem 3.3 (KKT – no equality constraints) *Let $S \subset \mathbb{R}^n$ be closed and let $f, g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\bar{x} \in S$. Consider the* nonlinear programming problem

$$(P) \ \inf_{x \in S} \{f(x) : g_1(x) \leq 0, \ldots, g_m(x) \leq 0\}.$$

Denote by $I(\bar{x})$ the set of all $i \in \{1, \ldots, m\}$ for which $g_i(\bar{x}) = 0$.

If $\bar{x}$ is a local optimal solution of (P), then there exist $\bar{u}_0 \in \{0, 1\}$ and $\bar{u} := (\bar{u}_1, \ldots, \bar{u}_m) \in \mathbb{R}^m_+$, $(\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_m) \neq (0, 0, \ldots, 0)$, and $\bar{\eta} \in \mathbb{R}^n$ such that

$$\bar{u}_i g_i(\bar{x}) = 0 \text{ for } i = 1, \ldots, m \quad \textbf{(complementary slackness)},$$

$$0 \in \bar{u}_0 \bar{\partial} f(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{u}_i \bar{\partial} g_i(\bar{x}) + \bar{\eta} \quad \textbf{(Lagrange inclusion)},$$

$$\bar{\eta} \in \cup_{t>0} t \, \bar{\partial} \mathrm{dist}_S(\bar{x}) \quad \textbf{(normal cone property)}.$$

Proof. Let us write $I := I(\bar{x})$. Consider the auxiliary optimization problem

$$(P') \ \inf_{x \in S} \phi(x),$$

where $\phi(x) := \max[f(x) - f(\bar{x}), \max_{1 \leq i \leq m} g_i(x)]$. Since $\bar{x}$ is a local optimal solution of (P), it is not hard to see that $\bar{x}$ is also local optimal for (P') (observe that $\phi(\bar{x}) = 0$ and that $x \in S$ is feasible if and only if $\max_{1 \leq i \leq m} g_i(x) \leq 0$). By Theorem 3.1 there exists $\bar{\eta}$ in $\mathbb{R}^n$ such that $\bar{\eta}$ has the normal cone property and $-\bar{\eta} \in \bar{\partial}\phi(\bar{x})$. By Theorem 2.10 this gives

$$-\bar{\eta} \in \bar{\partial}\phi(\bar{x}) = \mathrm{co}(\bar{\partial} f(\bar{x}) \cup \cup_{i \in I} \bar{\partial} g_i(\bar{x})).$$

Since generalized gradients form convex sets, we get the existence of $(u_0, \xi_0) \in \mathbb{R}_+ \times \bar{\partial} f(\bar{x})$ and $(u_i, \xi_i) \in \mathbb{R}_+ \times \bar{\partial} g_i(\bar{x})$, $i \in I$, such that $\sum_{i \in \{0\} \cup I} u_i = 1$ and

$$-\bar{\eta} = \sum_{i \in \{0\} \cup I} u_i \xi_i.$$

In case $u_0 = 0$, we are done by setting $\bar{u}_i := u_i$ for $i \in \{0\} \cup I$ and $\bar{u}_i := 0$ otherwise. Observe that in this case $(\bar{u}_1, \ldots, \bar{u}_m) \neq (0, \ldots, 0)$ by $\sum_{i \in I} u_i = 1$. In case $u_0 \neq 0$, we know that $u_0 > 0$, so we can set $\bar{u}_i := u_i / u_0$ for $i \in \{0\} \cup I$ and $\bar{u}_i := 0$ otherwise. QED

From Chapter 4 in [2] we already know that the inclusion of equality constraints in the KKT theorem requires considerable additional analytical efforts.

Recall that in Chapter 4 of [2] these efforts involved a certain version of the implicit function theorem. Recall also that this theorem is proven by appealing to completeness of the underlying (Euclidean) space. The situation is not different for the treatment by means of generalized gradients: We shall need extra analytical results, and these are based on the completeness of the underlying (Euclidean) space.

Theorem 3.4 (Ekeland) *Let $S \subset \mathbb{R}^n$ be closed and let $F : S \to \mathbb{R}$ be lower semi-continuous on and bounded from below. Let $x_0 \in S$ be such that $F(x_0) \leq \inf_S F + \epsilon$ for $\epsilon > 0$. Then there exists $\tilde{x} \in S$ such that $|x_0 - \tilde{x}| \leq \sqrt{\epsilon}$ and*

$$F(\tilde{x}) \leq F(x) + \sqrt{\epsilon}|x - \tilde{x}| \text{ for all } x \in S.$$

This so-called variational principle states, rather surprisingly, that an almost-mimimizer of F is, in a certain sense, very close to an exact minimizer of an "almost-F" function. A proof can be found in [4, pp. 266-268].

Theorem 3.5 (KKT – general case) *Let $S \subset \mathbb{R}^n$ be closed and let $f, g_1, \ldots, g_m, h_1, \ldots, h_p : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\bar{x} \in S$. Consider the* nonlinear programming problem

$$(P) \quad \inf_{x \in S} \{f(x) : g_1(x) \leq 0, \ldots, g_m(x) \leq 0, h_1(x) = \cdots = h_p(x) = 0\}.$$

Denote by $I(\bar{x})$ the set of all $i \in \{1, \ldots, m\}$ for which $g_i(\bar{x}) = 0$. If $\bar{x}$ is a local optimal solution of (P), then there exist multipliers $\bar{u}_0 \in \{0, 1\}$, $\bar{u} \in \mathbb{R}^m_+$, $\bar{v} \in \mathbb{R}^p$, $(\bar{u}_0, \bar{u}, \bar{v}) \neq (0, 0, 0)$, and $\bar{\eta} \in \mathbb{R}^n$ such that the complementary slackness relationship and normal cone property hold as in Theorem 3.3, as well as:

$$0 \in \bar{u}_0 \bar{\partial} f(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{u}_i \bar{\partial} g_i(\bar{x}) + \sum_{j=1}^{p} \bar{v}_j \bar{\partial} h_j(\bar{x}) + \bar{\eta} \quad \textbf{(Lagrange inclusion)}.$$

Lemma 3.6 (closure property) *Let $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $x_0 \in \mathbb{R}^n$. Suppose that $(x_k) \subset \mathbb{R}^n$ converges to x_0 and that (ξ_k), with $\xi_k \in \bar{\partial} F(x_k)$ for each k, converges to ξ_0. Then $\xi_0 \in \bar{\partial} F(x_0)$.*

PROOF. Let $d \in \mathbb{R}^n$ be arbitrary. For each k we have $F^o(x_k; d) \geq \xi_k^t d$. Hence, there exist $h_k \in \mathbb{R}^n$ and $\lambda_k > 0$ such that $|h_k|, \lambda_k < 1/k$ and

$$\frac{F(x_k + h_k + \lambda_k d) - F(x_k + h_k)}{\lambda_k} > \xi_k^t d - 1/k.$$

Since $x_k + h_k = x_0 + h'_k$, where $h'_k := x_k - x_0 + h_k \to 0$, Definition 2.3 gives immediately $F^o(x_0; d) \geq \xi_0^t d$. Since d was arbitrary, we get $\xi_0 \in \bar{\partial} F(x_0)$. QED

PROOF OF THEOREM 3.5. *Step 1: application of Ekeland's theorem and Theorem 3.1.* We form the following function $F_\epsilon : \mathbb{R}^n \to \mathbb{R}$; here $\epsilon > 0$:

$$F_\epsilon(x) := \max \left[f(x) - f(\bar{x}) + \epsilon, \max_{1 \leq i \leq m} g_i(x), \max_{1 \leq j \leq p} |h_j(x)| \right].$$

Clearly, $F_\epsilon \geq 0$ and we have $F_\epsilon(\bar{x}) = \epsilon$. So by Ekeland's theorem there exists a $\tilde{x}_\epsilon$, $|\tilde{x}_\epsilon - \bar{x}| < \sqrt{\epsilon}$, that is a local optimal solution of the following problem:

$$(P_\epsilon) \quad \inf_{x \in S} \left[F_\epsilon(x) + \sqrt{\epsilon}|x - \tilde{x}_\epsilon| \right].$$

Note already that $F_\epsilon(\tilde{x}_\epsilon) > 0$ for sufficiently small ϵ (or else we would have an obvious contradiction to the local optimality of $\bar{x}$ in (P)). We now repeat the part of the proof of Theorem 3.3 involving the auxiliary problem (P') and the successive application of Theorems 3.1 and 2.9, to get

$$0 \in \bar{\partial} F_\epsilon(\tilde{x}_\epsilon) + \sqrt{\epsilon}\bar{\partial} b_\epsilon(\tilde{x}_\epsilon) + K\bar{\partial}\mathrm{dist}_S(\tilde{x}_\epsilon),$$

where $b_\epsilon(x) := |x - \tilde{x}_\epsilon|$.

Step 2: extraction of convergent subsequences. Observe that $\bar{\partial} b_\epsilon(\tilde{x}_\epsilon) = \mathrm{cl}\ B_1(0)$, by an easy adaptation of Example 2.4. So the result of step 1 can be phrased as follows: there exist $\xi_\epsilon \in \bar{\partial} F_\epsilon(\tilde{x}_\epsilon)$ and $\eta_\epsilon \in K\bar{\partial}\mathrm{dist}_S(\tilde{x}_\epsilon)$ such that $|\xi_\epsilon + \eta_\epsilon| \leq \sqrt{\epsilon}$. Thus, since $\bar{\partial}\mathrm{dist}_S(\tilde{x}_\epsilon) \subset \mathrm{cl}\ B_1(0)$ (apply Theorem 2.5(ii), observing that dist_S has Lipschitz constant 1), we may invoke the Bolzano-Weierstrass theorem and conclude that a certain subsequence $(\xi_{\epsilon'}, \eta_{\epsilon'})$ converges to some (ξ, η) with $|\xi + \eta| = 0$ (i.e., $\xi = -\eta$).

Step 3: proof of (NCP). Since $|\tilde{x}_{\epsilon'} - \bar{x}| \leq \sqrt{\epsilon'} \to 0$, it follows already from Lemma 3.6 that $\eta \in K\bar{\partial}\mathrm{dist}_S(\bar{x})$. Hence, η belongs to the normal cone to S at $\bar{x}$.

Step 4: proof of (CS). In step 2 we found $\xi_{\epsilon'} \in \bar{\partial} F_{\epsilon'}(\tilde{x}_{\epsilon'})$, so by Theorem 2.10 there exist $(u_{0,\epsilon'}, \xi_{0,\epsilon'}) \in \mathbb{R}_+ \times \bar{\partial} f(\tilde{x}_{\epsilon'})$, $(u_{i,\epsilon'}, \xi_{i,\epsilon'}) \in \mathbb{R}_+ \times \bar{\partial} g_i(\tilde{x}_{\epsilon'})$, $i = 1, \ldots, m$ and $(w_{j,\epsilon'}, \xi'_{j,\epsilon'}) \in \mathbb{R}_+ \times \bar{\partial}|h_j|(\tilde{x}_{\epsilon'})$, $j = 1, \ldots, p$, such that

$$\xi_{\epsilon'} = \sum_{i=0}^{m} u_{i,\epsilon'}\xi_{i,\epsilon'} + \sum_{j=1}^{p} w_{j,\epsilon'}\xi'_{j,\epsilon'}, \tag{3}$$

and $\sum_{i=0}^{m} u_{i,\epsilon'} + \sum_{j=1}^{p} w_{j,\epsilon'} = 1$. This allows us to suppose without loss of generality (rather than extracting a suitable subsequence by the Bolzano-Weierstrass theorem) that $(u_{i,\epsilon'})$ converges to $u_i \in \mathbb{R}_+$ for each i and that $(w_{j,\epsilon'})$ converges to $w_j \in \mathbb{R}_+$ for each j. Observe that in the limit we retain the convex combination property:

$$\sum_{i=0}^{m} u_i + \sum_{j=1}^{p} w_j = 1. \tag{4}$$

Still by Theorem 2.10, we also have for $i = 1, \ldots, m$ that if $g_i(\tilde{x}_{\epsilon'}) < F_{\epsilon'}(\tilde{x}_{\epsilon'})$, then $u_{i,\epsilon'} = 0$ and for $j = 1, \ldots, p$ that if $|h_j|(\tilde{x}_{\epsilon'}) < F_{\epsilon'}(\tilde{x}_{\epsilon'})$, then $w_{j,\epsilon'} = 0$. Now if $i \notin I(\bar{x})$, then $g_i(\bar{x}) < 0$, whence $g_i(\tilde{x}_{\epsilon'}) < 0$ for ϵ' sufficiently small. But this means $u_{i,\epsilon'} = 0$ by the above (recall from step 1 that $F_\epsilon(\tilde{x}_{\epsilon'}) > 0$). This proves $u_i = 0$ for $i \notin I(\bar{x})$. In other words, (CS) holds.

Step 5: Application of Exercise 2.3. We claim for each j the existence of $v_{j,\epsilon'} \in \mathbb{R}$, $|v_{j,\epsilon'}| = w_{j,\epsilon'}$ and $\xi''_{j,\epsilon} \in \bar{\partial} h_j(\tilde{x}_{\epsilon'})$ such that

$$w_{j,\epsilon'}\xi'_{j,\epsilon'} = v_{j,\epsilon'}\xi''_{j,\epsilon'}. \tag{5}$$

Indeed, if $h_j(\tilde{x}_{\epsilon'}) = 0$, then $0 = |h_j|(\tilde{x}_{\epsilon'}) < F_{\epsilon'}(\tilde{x}_{\epsilon'})$, so $w_{j,\epsilon'} = 0$ by step 4. Of course, then we can match by taking $v_{j,\epsilon'} := 0$. On the other hand, if $h_j(\tilde{x}_{\epsilon'}) \neq 0$, then the result follows from Exercise 2.3: we then set $v_{j,\epsilon'} := \mathrm{sign} h_j(\tilde{x}_{\epsilon'})w_{j,\epsilon'}$. It is clear that, without loss of generality, we may suppose that for each j the sequence $v_{j,\epsilon'}$ converges to some $v_j \in \mathbb{R}$.

Step 6: proof of (LI). In steps 4, 5 we found $\xi_{0,\epsilon'} \in \bar{\partial} f(\tilde{x}_{\epsilon'})$, $\xi_{i,\epsilon'} \in \bar{\partial} g_i(\tilde{x}_{\epsilon'})$, and $\xi''_{j,\epsilon'} \in \bar{\partial} h_j(\tilde{x}_{\epsilon'})$. Applying Theorem 2.5($ii$) we may suppose without loss of generality that $(\xi_{i,\epsilon'})$ converges to some $\xi_i \in \mathbb{R}^n$ for each i and that $(\xi''_{j,\epsilon})$ converges to some $\xi''_j \in \mathbb{R}^n$ for each j; application of Lemma 3.6 then gives

$$\xi_0 \in \bar{\partial}(\bar{x}), \xi_i \in \bar{\partial} g_i(\bar{x}), \xi''_j \in \bar{\partial} h_j(\bar{x}).$$

Taking now the limit in (3), while taking into consideration (5), gives

$$-\eta = \xi = \sum_{i=0}^{m} u_i \xi_i + \sum_{j=1}^{p} v_j \xi''_j.$$

In case $u_0 = 0$, we set $\bar{u}_i := u_i$, $\bar{v}_j := v_j$ for all i, j and we set $\bar{\eta} := \eta$. Observe that then $(\bar{u}, \bar{v}) \neq (0, 0)$ by (4). In case $u_0 > 0$, we normalize: set $\bar{u}_i := u_i/u_0$, $\bar{v}_j := v_j/u_0$ for all i, j and we set $\bar{\eta} := \eta/u_0$. QED

References

[1] E.J. Balder, *On subdifferential calculus.* handout September, 2008.

[2] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms.* Wiley, New York, 1993.

[3] J. Borwein and A.S. Lewis, *Convex Analysis and Nonlinear Optimization*, CMS Books in Mathematics, Springer-Verlag, New York, 2000.

[4] F.H. Clarke, *Optimization and Nonsmooth Analysis.* SIAM Classics in Applied Mathematics No. 5, 1990.

[5] J. van Tiel, *Convex Analysis: An Introductory Text.* Wiley, 1984.