Mathematical Programming 33 (1985) 260-279
North-Holland

DESCENT METHODS FOR COMPOSITE NONDIFFERENTIABLE OPTIMIZATION PROBLEMS

James V. BURKE

Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA

Received 7 February 1984
Revised manuscript received 18 April 1985

We present a framework for the development of globally defined descent algorithms for the minimization of non-differentiable objective functions $F := h \circ f$ with h convex. Within our structure the global convergence properties of the Cauchy, Modified Newton, Gauss-Newton, and Variable-Metric methods are easily established along with that of several new approaches. Examples illustrating the calculational techniques are provided.

Key words: Clarke Subdifferential, ε-Subdifferential, Armijo Stepsize, Epi-Convergence, Casting Functions.

1. Introduction

In this paper we present a framework for the development of globally defined descent algorithms that are designed to locate stationary points of functions of the form

$$F = h \circ f, \tag{1.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, and $h: \mathbb{R}^m \to \mathbb{R}$ is convex. This problem and techniques to solve it play a central role in contemporary studies in mathematical programming. For example, the function h may be taken to be the identity, a norm, a penalty function, or the distance function to some convex set. Analyses of such problems where the function h is chosen to have a specific representation abound in the literature, but recently efforts have been made to unify the methodology. The endeavor was initiated by Anderson and Osborne [1], Osborne and Watson [20], Osborne [21], Fletcher [9], and Powell [24]. In [1] Anderson and Osborne provide the first uniform treatment of Gauss-Newton type methods for solving systems of equations via polyhedral norms, then in [20] Osborne and Watson extend this analysis to arbitrary norm structures on $\mathbb{R}^n$ and provide the first indication that these methods could be extended to composite functions of type (1.1). Osborne [21] provides a survey of these results and those of others. In [9] Fletcher coins the term 'composite nondifferentiable optimization', and applies the technique to penalty functions, developing a Trust-Region algorithm that foreshadows the casting function approach of this paper. Powell [24] extends these techniques and analyzes Gauss-Newton, Trust-Region, and Variable-Metric methods for minimizing (1.1).

Further refinements of these methods are developed in the papers of Powell and Yuan [25, 26, 31, 32] where attempts are made to obtain better computational characteristics and faster rates of convergence.

In the present work, we provide a more general theory for the development of algorithms for the minimization of (1.1). Within our framework the global convergence characteristics of all of the standard techniques (e.g. Cauchy, Modified Newton, Gauss–Newton, Variable-Metric) are easily established along with those of several new approaches. We begin in Section 2 with a statement of the general structure of the algorithms to be investigated and prove a rudimentary convergence result. In Section 3 we address the question of the boundedness of search directions, generalize the notion of *casting functions* introduced in Wets [29], and derive several relevant stationarity criteria. In Section 4 we define three general classes of search directions and show that they can be employed within the framework of the model algorithm of Section 1. Finally, in Section 5, we provide a convergence analysis via the notion of *epi-convergence*, and conclude our study in Section 4 with a few examples demonstrating how the necessary underlying computations can be formulated as either *linear* or *quadratic* programs.

The notation that we employ is for the most part the same as that of Rockafellar [27]. A partial list is provided below for the reader's convenience:

$$f'(x; d) := \lim_{\lambda \downarrow 0} \frac{f(x+\lambda d) - f(x)}{\lambda}.$$

Let $f: \mathbb{R}^n \to \mathbb{R}^* := \mathbb{R} \cup \{+\infty\}$ be convex, then

$$\mathrm{Dom}(f) := \{x : f(x) < +\infty\},$$

$$\mathrm{epi}(f) := \{(x, \alpha) : f(x) \leqslant \alpha,\ \alpha \in \mathbb{R}\},$$

$$\partial_\varepsilon f(x) := \{x^* : f(y) \geqslant f(x) + \langle x^*, y - x\rangle - \varepsilon, \text{ for all } y \in \mathrm{Dom}(f)\},$$

$$f^*(x^*) := \sup\{\langle x, x^*\rangle - f(x) : x \in \mathbb{R}^n\}.$$

For $f: \mathbb{R}^n \to \mathbb{R}$ and $C \subset \mathbb{R}^n$,

$$\mathrm{argmin}\{f(x) : x \in C\} = \{\bar{x} \in C : f(\bar{x}) = \min\{f(x) : x \in C\}\}.$$

For $C \subset \mathbb{R}^n$, $\overline{\mathrm{co}}\, C$ is the closed convex hull of C and int C is the interior of C.
For $\|\cdot\|_\nu$ a norm on $\mathbb{R}^n$, $B_\nu := \{x : \|x\|_\nu \leqslant 1\}$.
For C a nonempty convex subset of $\mathbb{R}^n$, we have

$$\psi(x|C) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C, \end{cases}$$

$$\psi^*(x|C) := \sup\{\langle x, x^*\rangle : x^* \in C\},$$

$$\gamma(x|C) := \inf\{\gamma : x \in \gamma C,\ \gamma \geqslant 0\},$$

$$C^0 := \{x^* : \langle x^*, x\rangle \leqslant 1 \text{ for all } x \in C\}.$$

For K a closed convex cone in $\mathbb{R}^n$, we have

$$K^0 = \{x^*: \langle x^*, x\rangle \le 0 \quad \text{for all} \quad x \in K\},$$

$$K^* := -K^0.$$

$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n: x_i \ge 0, i = 1, \ldots, n, \text{ where } x = (x_1, \ldots, x_n)^{\mathrm{T}}\}$, $\mathbb{R}^n_- := -\mathbb{R}^n_+$.
For $x \in \mathbb{R}^n$ we define x_+, x_-, and $|x|$ componentwise as follows: $x = (x_1, \ldots, x_n)^{\mathrm{T}}$,

$$(x_+)_i := \max(0, x_i),$$

$$(x_-)_i := \min(0, x_i),$$

$$(|x|)_i := |x_i|.$$

The vector $e \in \mathbb{R}^k$ is the vector of ones, $e = (1, 1, \ldots, 1)^{\mathrm{T}}$.

2. The model algorithm

The types of algorithms that we concern ourselves with are of the form

$$x_{i+1} := x_i + \lambda_i d_i \tag{2.1}$$

where

$$\lambda_i := \max\{\gamma^k: F(x_i + \gamma^k d_i) - F(x_i) \le c\gamma^k \Delta_i, k = 0, 1, \ldots\},$$

$d_i \in D_i \subset \mathbb{R}^n$, $\Delta_i \le 0$, $c \in (0, 1)$, and $\gamma \in (0, 1)$. In this context it is clear that the choice of the numbers Δ_i and the sets D_i provide the key to the analysis of the algorithm. For our purposes we require that they satisfy the following three conditions:

$$\begin{aligned} &\text{(a)} \quad D_i \ne \emptyset \quad \text{for all} \quad i = 0, 1, 2, \ldots, \\ &\text{(b)} \quad [0 \in D_i] \Leftrightarrow [\Delta_i = 0] \Leftrightarrow [0 \in \partial F(x_i)], \\ &\text{(c)} \quad h(f(x_i) + f'(x_i)d_i) - F(x_i) \le \Delta_i \le 0 \quad \text{for all} \quad i = 0, 1, \ldots. \end{aligned} \tag{2.2}$$

(Here ∂F denotes the *Clarke subgradient* [6].)

Similar model algorithms have been studied in the context of nondifferentiable optimization by several authors [1, 4, 5, 7–12, 14, 20–26, 28, 30, 31]. In particular, the *Armijo type stepsize routine* has been found to be an especially useful tool in the development of very general global optimization strategies. The analysis that we provide for (2.1) is reminiscent of that given in Wolfe [30], as it is our intention to provide a broad framework from which many of the known techniques are easily derived. Moreover, our approach is also somewhat similar to that which may be found in the recent paper by Polak, Mayne and Wardi [23]. In the Polak, Mayne, and Wardi paper, the choice of search direction is an ε-steepest descent direction calculated as the nearest point to the origin of their so-called 'ε-smeared generalized gradient'. Generalized gradients of this type seem to have been first investigated by Goldstein [12], and later by Dixon [7] and Dixon and Gaviano [8], and is defined by the expression

$$\partial_\varepsilon F(\bar{x}) := \overline{\text{co}} \bigcup_{x \in B(\bar{x};\varepsilon)} \partial F(x).$$

In order to assure that their approach is computationally implementable, Polak, Mayne, and Wardi require that the function F to be minimized is *semi-smooth* in the sense of Mifflin [17, 18]. The functions that we study, (1.1), are also semi-smooth (Mifflin [18], Proposition 5) and so the Polak, Mayne, and Wardi algorithm applies. But, as we shall see, due to the special structure of our objective function, $F = h \circ f$, the direction choices that we study have a greater attraction for both theoretical and computational reasons.

In the lemma that follows, we present the key structural characteristic of functions of the form (1.1) that will allow us to define a variety of descent directions satisfying conditions (2.2).

2.3 Lemma. *If $F = h \circ f$ is such that $f : \mathbb{R}^n \to \mathbb{R}^m$ is Frechet differentiable on $\mathbb{R}^n$, and $h: \mathbb{R}^m \to \mathbb{R}$ is a closed proper convex function on $\mathbb{R}^m$, then*

(a) *$\partial F(x)$, the Clarke subdifferential, has the representation*

$$\partial F(x) := \partial h(f(x)) \circ f'(x) := \{y \in \mathbb{R}^n : y = zf'(x),\ z \in \partial h(f(x))\}$$

for all $x \in \mathbb{R}^n$, and

(b) *$F'(x; d)$ exists for all x and d in $\mathbb{R}^n$, and satisfies*

$$F'(x; d) \leq h(f(x) + f'(x)d) - h(f(x)).$$

Proof. Statement (a) is implicit in the work of both Fletcher [9] and Powell [24], and is easily derived via Clarke [6, Theorem 2.3.10]. Statement (b) is also implicit in the work of Powell [24]. The proof is as follows. By Clarke [6, Theorem 2.3.10], $F'(x; d)$ exists for all x and d in $\mathbb{R}^n$. Choose x and d in $\mathbb{R}^n$ and let K be a local Lipschitz constant for h at x (K exists as h is finite-valued and convex on $\mathbb{R}^n$ [27]). Then for $\lambda \geq 0$ sufficiently small, we have that

$$\begin{aligned} & h(f(x + \lambda d)) - h(f(x)) \\ & \quad = [h(f(x) + \lambda f'(x)d) - h(f(x))] + [h(f(x + \lambda d)) - h(f(x) + \lambda f'(x)d)] \\ & \quad \leq [(1 - \lambda)h(f(x)) + \lambda h(f(x) + f'(x)d) - h(f(x))] \\ & \qquad + K\|f(x + \lambda d) - f(x) - \lambda f'(x)d\| \\ & \quad = \lambda[h(f(x) + f'(x)d) - h(f(x))] + K\text{o}(\lambda) \end{aligned}$$

from which the result follows. □

Thus we see that any direction d for which $h(f(x) + f'(x)d) < h(f(x))$ is a descent direction for F. Therefore condition (2.2), along with the stopping criteria $[0 \in \partial F(x)]$, guarantee that algorithm (2.1) is always well defined.

We now give the fundamental convergence result for algorithms of type (2.1) that satisfy conditions (2.2).

2.4 Theorem. *Let $x_0 \in \mathbb{R}^n$ and let $F = h \circ f$ satisfy the assumptions*

$$\begin{array}{ll}\text{(a)} & \textit{The function } f : \mathbb{R}^n \to \mathbb{R}^m \textit{ is Fréchet differentiable with } f' \\ & \textit{uniformly continuous on } \overline{\text{co}}\{x \colon F(x) \le F(x_0)\}, \textit{ and} \\ \text{(b)} & \textit{the finite-valued convex function } h : \mathbb{R}^m \to \mathbb{R} \textit{ is Lipschitz on} \\ & \overline{\text{co}}\{y \colon h(y) \le F(x_0)\}.\end{array} \tag{2.5}$$

If $\{x_i\}$ is the sequence generated by algorithm (2.1) *with initial point x_0 and stopping criteria $0 \in \partial F(x)$, then provided that condition* (2.2) *is satisfied, one of the following must occur*:

(i) *The algorithm terminates finitely at x_{i_0} with $0 \in \partial F(x_{i_0})$, or* $\lim_{j\in J} \Delta_j = 0$ *for every subsequence J for which the associated subsequence $\{d_j \colon j \in J\}$ is bounded*; *and/or*

(ii) $F(x_i) \downarrow -\infty$; *and/or*

(iii) *the sequence $\{\|d_i\|\}$ diverges to* $+\infty$.

Proof. Suppose to the contrary that none of (i), (ii), and/or (iii) occur. Then there is a subsequence J such that $\sup\{\|d_j\| \colon j \in J\} < \infty$ and $\sup\{\Delta_j \colon j \in J\} \le \beta < 0$ for some $\beta \in \mathbb{R}$. Now $F(x_i) \not\downarrow -\infty$, hence the decreasing sequence $\{F(x_i)\}$ is bounded below, and so has a limit. Therefore $(F(x_{i+1}) - F(x_i)) \to 0$ and thus by the Armijo inequality in (2.1), we get that $\lambda_i \Delta_i \to 0$. Hence we can assume with no loss of generality that $\lim_{j\in J} \lambda_j = 0$ and $\lambda_j < 1$ for all $j \in J$, since $\sup\{\Delta_j \colon j \in J\} \le \beta < 0$. The Armijo inequality now yields the relation

$$c\lambda_j\gamma^{-1}\Delta_j < F(x_j + \lambda_j\gamma^{-1}d_j) - F(x_j)$$

for all $j \in J$. But, as in Lemma 2.3,

$$\begin{aligned} &F(x_j + \lambda_j\gamma^{-1}d_j) - F(x_j) \\ &\quad \le \lambda_j\gamma^{-1}\Delta_j + K\|f(x_j + \lambda_j\gamma^{-1}d_j) - f(x_j) - \lambda_j\gamma^{-1}f'(x_j)d_j\| \\ &\quad \le \lambda_j\gamma^{-1}\Delta_j + K\|\lambda_j\gamma^{-1}\int_0^1 [f'(x_j + t\lambda_j\gamma^{-1}d_j) - f'(x_j)]d_j\,\mathrm{d}t\| \\ &\quad \le \lambda_j\gamma^{-1}[\Delta_j + K\omega(\lambda_j\gamma^{-1}\|d_j\|)\|d_j\|], \end{aligned}$$

for all $j \in J$, where K is a Lipschitz constant for h and the function $\omega : \mathbb{R} \to \mathbb{R}_+$ is the modulus of continuity for f'. Therefore

$$0 < (1-c)\Delta_j + K\omega(\lambda_j\gamma^{-1}\|d_j\|)\|d_j\| \le (1-c)\beta + K\omega(\lambda_j\gamma^{-1}\|d_j\|)\|d_j\|$$

for all $j \in J$. Now taking the limit as $j \to \infty$, $j \in J$, and employing the boundedness of the subsequence $\{d_j \colon j \in J\}$, we obtain the contradiction

$$0 \le (1-c)\beta < 0,$$

yielding the result. □

Thus far, the discussion has been placed in a very general setting and in fact, without the refinements to be introduced later, the preceding results would be of little consequence. The primary benefit of these results and especially Theorem 2.4, though, is that they isolate the potential structual defects of algorithm (2.1) that must be compensated for in the designation of the sets D_i and the numbers Δ_i. Specifically, we need a device to induce the boundedness of the search direction choices d_i. Thus we are led to the notion of *casting functions.*

3. Casting functions

3.1 Definition. A mapping $\rho:\mathbb{R}^n\to\mathbb{R}^*:=\mathbb{R}\cup\{+\infty\}$, is called a casting function if it satisfies the following four conditions:

(1) ρ is a closed proper convex function,

(2) $0\in\operatorname{int}(\operatorname{Dom}(\rho))$,

(3) $0=\rho(0)=\min\{\rho(d)\colon d\in\mathbb{R}^n\}$,

(4) ρ is inf-compact, or equivalently, $\lim_{\|x\|\to\infty}\rho(x)=+\infty$.

We denote by $\mathscr{C}$ the class of all casting functions, and by $\mathscr{C}'$ those that are Fréchet differentiable at the origin.

Remark. The definition of casting function that we present here is a generalization of that which is given by Wets [29]. In his definition, it is required that the functions be symmetric. In fact, it is these symmetric casting functions that form the most important subclass of casting functions.

It should be clear that both of the classes $\mathscr{C}$ and $\mathscr{C}'$ are closed under addition, multiplication by positive scalars, and by squaring. Moreover, the pointwise supremum of any finite subset of $\mathscr{C}$ is also in $\mathscr{C}$. Important examples of casting functions are the support, gauge, and convex indicator functionals of closed convex sets containing the origin in their interior. Other important examples are generated by symmetric positive definite bilinear forms.

We now employ casting functions in defining our primary analytic tool for the development of techniques intended to minimize (1.1), that is, the class of convex functions

$$d\mapsto\phi(d\,;x,\rho):\mathbb{R}^n\to\mathbb{R}$$

defined by the relation

$$\phi(d\,;x,\rho):=h(f(x)+f'(x)d)+\rho(d) \tag{3.2}$$

for every $x\in\mathbb{R}^n$ and $\rho\in\mathscr{C}$. In conjunction with these functions the following sets will play an important role in our analysis: Let $\rho_0 \in \mathscr{C}$ and $x_0 \in \mathbb{R}^n$, then

$$\mathscr{C}(\rho_0) := \{\rho \in \mathscr{C}\colon \rho_0(x) \leq \rho(x) \forall x \in \mathbb{R}^n\}, \tag{3.3a}$$

$$\mathscr{C}'(\rho_0) := \mathscr{C}(\rho_0) \cap \mathscr{C}', \tag{3.3b}$$

$$L(x_0) := \{x\colon F(x) \leq F(x_0)\}, \tag{3.3c}$$

$$S(\rho_0, x_0) := \{d\colon \phi(d; x_0, \rho_0) \leq F(x_0)\}, \tag{3.3d}$$

$$\Phi(\rho_0, x_0) := \bigcup_{\rho \in \mathscr{C}(\rho_0)} \bigcup_{x \in L(x_0)} S(\rho, x), \tag{3.3e}$$

$$\Phi(\rho_0, *) := \{d \in \mathbb{R}^n\colon \phi(d; x, \rho) < \infty \text{ for some } x \in \mathbb{R}^n \text{ and } \rho \in \mathscr{C}(\rho_0)\}, \tag{3.3f}$$

and observe that $S(\rho, x) \subset \Phi(\rho, x) \subset \Phi(\rho, *)$, for all $x \in \mathbb{R}^n$ and $\rho \in \mathscr{C}$. We now have the following basic results concerning these sets.

3.4 Lemma. *Let $x_0 \in \mathbb{R}^n$, let $F = h \circ f$ satisfy condition* (2.5a) *of Theorem* 2.4, *and suppose that f is Fréchet differentiable on $\mathbb{R}^n$.*

(a) *Let $S \subset \mathbb{R}^n$ and let $\rho \in \mathscr{C}$. If ρ is bounded on S, then S is bounded.*

(b) *If h is bounded below, then for any $\rho_0 \in \mathscr{C}$, the set $\Phi(\rho_0, x_0)$ is bounded.*

(c) *Let $\|\cdot\|_\nu$ be a norm on $\mathbb{R}^n$. If f' is bounded on the set $L(x_0)$, then the set $\Phi(\|\cdot\|_\nu^\alpha, x_0)$ is bounded for all $\alpha > 1$.*

(d) *Let $\delta > 0$ and let $\|\cdot\|_\nu$ be a norm on $\mathbb{R}^n$ with unit ball B_ν. Then the set $\Phi(\psi(\cdot|\delta B_\nu), *)$ is bounded by δ.*

(e) *Suppose $x_0 \in \mathbb{R}^n$ and $\rho \in \mathscr{C}$ are such that the set $S(\rho, x)$ is bounded for all $x \in L(x_0)$. Then for every $x \in L(x_0)$ there exists a $d_x \in \mathbb{R}^n$, not necessarily unique, such that*

$$\phi(d_x; x, \rho) = \min\{\phi(d; x, \rho)\colon d \in \mathbb{R}^n\}.$$

Proof. (a) This follows immediately from condition 4 of Definition 3.1.

(b) Let M be a lower bound for h, and choose $\rho_0 \in \mathscr{C}$. Then for every $d \in \Phi(\rho_0, x_0)$ we have the inequality $\rho_0(d) \leq F(x_0) - M$. Hence $\Phi(\rho_0, x_0)$ is bounded by (a).

(c) Let $\|\cdot\|_\nu$ be a norm on $\mathbb{R}^n$, and let $\|\cdot\|_{\nu'}$ be a norm on $\mathbb{R}^m$ consistent with $\|\cdot\|_\nu$. Let K_1 be h's Lipschitz constant with respect to $\|\cdot\|_{\nu'}$, and let K_2 be a bound on f' with respect to $\|\cdot\|_\nu$ and $\|\cdot\|_{\nu'}$. Choose $\alpha > 1$. Then for $d \in \Phi(\|\cdot\|_\nu^\alpha, x_0)$, there is some $x \in L(x_0)$ such that

$$\|d\|_\nu^\alpha \leq F(x) - h(f(x) + f'(x)d) \leq K_1\|f(x) - f(x) - f'(x)d\|_{\nu'} \leq K_1K_2\|d\|_\nu.$$

Hence $\|d\|_\nu^{\alpha-1} \leq K_1K_2$ and so $\Phi(\|\cdot\|_\nu^\alpha, x_0)$ is bounded.

(d) This follows immediately from Definition (3.3f).

(e) Since both h and ρ are closed proper convex functions, and $f'(x)$ is linear in d for every $x \in L(x_0)$, we know that for each $x \in L(x_0)$, ϕ is continuous on $S(\rho, x)$ and that $S(\rho, x)$ is closed and bounded. Hence ϕ attains its infimum on $S(\rho, x)$. Consequently, ϕ attains its global infimum. □

Part (e) of the above lemma provides a foreshadowing of things to come as it indicates one of the procedures by which our search directions will be calculated. The boundedness of such directions is guaranteed, under mild assumptions, by the other parts of the lemma. But even with boundedness, condition (2.2) still must be satisfied. In order to appropriately deal with this question, we extend the definition of casting functions so that we can consider them as functions of both d and x. To this end, we will employ the following notation:

We denote by $\mathscr{C}^*$ the set of all functions $\rho:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ that satisfy the conditions (3.5)
(a) $\rho(\cdot,x)\in\mathscr{C}$ for all $x\in\mathbb{R}^n$, and
(b) $[0\in\partial F(x)+\partial\rho(0,x)] \Leftrightarrow [0\in\partial F(x)]$.

Remark. Note that condition (3.5b) is superfluous if $\rho(\cdot,x)\in\mathscr{C}'$ for all $x\in\mathbb{R}^n$. An example of such a function that is not necessarily differentiable at the origin for all $x\in\mathbb{R}^n$ is as follows: Let $\rho_0\in\mathscr{C}$ be such that $\partial\rho_0(0)$ is contained in the unit ball, B_ν, of some norm $\|\cdot\|_\nu$. Define $\rho(d,x)$ by the relation

$$\rho(d,x):=\begin{cases}\frac{1}{2}\operatorname{dist}_\nu(0,\partial F(x))\rho_0(d) & \text{if } 0\notin\partial F(x),\\ \rho_0(d) & \text{if } 0\in\partial F(x),\end{cases}$$

for all $x\in\mathbb{R}^n$. (Here $\operatorname{dist}_\nu(x,C):=\inf\{\|x-y\|_\nu: y\in C\}$.) Then $\rho\in\mathscr{C}^*$.

The following theorem is the starting point for the analysis required to verify condition (2.2).

3.6 Theorem. *Let $F=h\circ f$ where $h:\mathbb{R}^m\to\mathbb{R}$ is a finite-valued convex function on $\mathbb{R}^m$ and $f:\mathbb{R}^n\to\mathbb{R}$ is Fréchet differentiable on $\mathbb{R}^n$.*
(1) *If $0\in\partial F(x)$, then $0\in\partial\phi(0;x,\rho)$ for all $\rho\in\mathscr{C}$.*
(2) *The following statements are equivalent:*
(a) $0\in\partial F(x)$;
(b) $0\in\partial\phi(0;x,\rho)$ *for some or all $\rho\in\mathscr{C}^*$;*
(c) $F(x)=\min\{\phi(d;x,\rho): d\in\mathbb{R}^n\}$ *for some or all $\rho\in\mathscr{C}^*$;*
(d) $0\in\operatorname{argmin}\{\phi(d;x,\rho): d\in\mathbb{R}^n\}$ *for some or all $\rho\in\mathscr{C}^*$.*

Proof. (1) From the definition of $\mathscr{C}$, we know that $0\in\partial\rho(0)$ for every $\rho\in\mathscr{C}$. Hence the result follows immediately from the fact that

$$\partial\phi(0;x,\rho)=\partial h(f(x))\circ f'(x)+\partial\rho(0,x)=\partial F(x)+\partial\rho(0,x).$$

(2) (a)$\Leftrightarrow$(b): The implication (a)$\Rightarrow$(b) follows from part (1). Conversely, if $0\in\partial\phi(0;x,\rho)=\partial F(x)+\partial\rho(0,x)$, then $0\in\partial F(x)$, since $\rho\in\mathscr{C}^*$.

(a)$\Rightarrow$(c): Let $0\in\partial F(x)$ and define $\phi(\cdot\,;x,0):\mathbb{R}^n\to\mathbb{R}$ by

$$\phi(d;x,0):=h(f(x)+f'(x)d).$$

Then $0 \in \partial\phi(0; x, 0) = \partial F(x)$, hence $\phi(\cdot\,; x, 0)$ is minimized at $d = 0$ due to its convexity. The implication now follows from the inequality

$$F(x) \geqslant \min\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\} \geqslant \min\{\phi(d; x, 0) \colon d \in \mathbb{R}^n\}$$

where ρ is any element of $\mathscr{C}$.

(d)⇒(a): Suppose there exists $\rho \in \mathscr{C}^*$ for which $0 \in \operatorname{argmin}\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\}$. Then $0 \in \partial\phi(0; x, \rho)$, and so $0 \in \partial F(x)$ by the implication (b)⇒(a).

(c)⇒(d): Suppose there exists $\rho_0 \in \mathscr{C}^*$ for which $F(x) = \min\{\phi(d; x, \rho_0) \colon d \in \mathbb{R}^n\}$. Then clearly $0 \in \operatorname{argmin}\{\phi(d; x, \rho_0) \colon d \in \mathbb{R}^n\}$. But then by the string of implications (d)⇒(a)⇒(c), we have that $F(x) = \min\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\}$ for every $\rho \in \mathscr{C}^*$, and so $0 \in \operatorname{argmin}\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\}$ for all $\rho \in \mathscr{C}^*$. □

Before leaving this section, we state one more result concerning the stationary characteristics of the functions (3.2). Since the result is a straightforward application of Propositions 1 and 2 in Bertsekas and Mitter [3], we omit its proof.

3.7 Theorem. *Let the assumptions of Theorem* 3.6 *hold and let the functions* $\phi(\cdot\,; x, \rho)$ *be as in* (3.2) *with* $\rho \in \mathscr{C}^*$. *Then*

(a) $[0 \leqslant F(x) - \inf\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\} \leqslant \varepsilon] \Leftrightarrow [0 \in \partial_\varepsilon\phi(0; x, \rho)]$ *where* ∂_ε *represents the usual ε-subgradient operator of convex analysis* [27], *and*

(b) *if* $0 \notin \partial_\varepsilon\phi(0; x, \rho)$ *and* $d \in \mathbb{R}^n$ *is any vector such that* $\psi^*(d \mid \partial_\varepsilon\phi(0; x, \rho)) < 0$, *then*

$$F(x) - \inf_{\lambda \geqslant 0} \phi(\lambda d; x, \rho) > \varepsilon.$$

4. Search directions

For $x \in \mathbb{R}^n$ and $\rho \in \mathscr{C}^*$ we define the function $d \mapsto \Delta(d; x, \rho) \colon \mathbb{R}^m \to \mathbb{R}$ by the relation

$$\Delta(d; x, \rho) := \phi(d; x, \rho) - F(x) \tag{4.1}$$

and note that $F'(x; d) \leqslant \Delta(d; x, \rho)$ for all x and d in $\mathbb{R}^n$, and $\rho \in \mathscr{C}$, by Lemma 2.3. Given $\rho \in \mathscr{C}^*$, we define the following three classes of search directions.

4.2. Set

$$\Delta_1(x, \rho) := \inf\{\Delta(d; x, \rho) \colon d \in \mathbb{R}^n\}$$

and

$$D_1(x, \rho, r) := \{d \colon \Delta(d; x, \rho) \leqslant r\Delta_1(x, \rho)\}$$

where $r \in (0, 1]$ is a relaxation parameter.

4.3. Let $\sigma \colon \mathbb{R}^n \to \mathbb{R}^n$ be a selection from $\operatorname{argmin}\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\}$, i.e.

$$\sigma(x) \in \operatorname{argmin}\{\phi(d; x, \rho) \colon d \in \mathbb{R}^n\} \text{ for all } x \in \mathbb{R}^n,$$

and set

$$\Delta_2(x, \rho) := h(f(x) + f'(x)\sigma(x)) - F(x)$$

and

$$D_2(x, \rho, r) := \{d \colon h(f(x) + f'(x)d) - F(x) \leqslant r\Delta_2(x, \rho)\}$$

where $r \in (0, 1]$ is a relaxation parameter. (In order to simplify the presentation, we have chosen to suppress the choice of selection, $\sigma(x)$, from the notation.)

4.4. Choose $r \in (0, 1)$ and define

$$\varepsilon(x) := \begin{cases} 0 & \text{if } 0 \in \partial\phi(0; x, \rho), \\ \max\{r^p \colon 0 \notin \partial_{r^p}\phi(0; x, \rho),\ p = 0, 1, 2, \ldots\}, & \text{otherwise,} \end{cases}$$

then set

$$\Delta_3(x, \rho) := -\varepsilon(x)$$

and

$$D_3(x, \rho, r) := \begin{cases} \{0\}, & \text{if } \varepsilon(x) = 0, \\ \{d \colon \psi^*(d \mid \partial_{\varepsilon(x)}\phi(0; x, \rho)) < 0, \text{ and } \Delta(d; x, \rho) \leqslant \Delta_3(x, \rho)\}, & \\ & \text{otherwise.} \end{cases}$$

(Here we have suppressed the parameter r in the notation for Δ_3 for the sake of simplicity.)

If $0 \notin \partial\phi(0; x, \rho)$ the statement that the set $D_3(x, \rho, r)$ is nonempty is easily seen to be equivalent to Theorem 3.7, part b. In fact, this choice of search direction is simply a generalization of that which is employed in the Bertsekas–Mitter ε-subgradient algorithm for convex functions [3]. For $k \in \{1, 2\}$, we establish the nonemptiness of $D_k(x, \rho, r)$ in various situations by employing the results of Lemma 3.4. Clearly, $D_k(x, \rho, r) \supset D_k(x, \rho, 1)$ for every $r \in (0, 1]$. Hence we need only establish the nonemptiness of $D_k(x, \rho, 1)$. But, by Lemma 3.4e, $D_k(x, \rho, 1)$ is nonempty if $S(\rho, x)$ is bounded. Finally, very general conditions for obtaining the boundedness of $S(\rho, x)$ are established in parts b, c, and d of Lemma 3.4.

Given the nonemptiness of the sets $D_k(x, \rho, r)$, the following lemma confirms that these search direction choices do indeed satisfy the requirements of condition (2.2).

4.5 Lemma. *Let the assumptions of Theorem* 3.6 *hold. Choose* $x_0 \in \mathbb{R}^n$, $\rho \in \mathscr{C}^*$, *and* $k \in \{1, 2, 3\}$, *then select* $r \in (0, 1]$ *if* $k \in \{1, 2\}$; *otherwise,* $r \in (0, 1)$. *If the set* $S(\rho, x)$ *is bounded for all* $x \in L(x_0)$, *then*

(a) $D_k(x, \rho, r) \neq \emptyset$ *for all* $x \in L(x_0)$,
(b) $[0 \in D_k(x, \rho, r)] \Leftrightarrow [\Delta_k(x, \rho) \eqslantless 0] \Leftrightarrow [0 \in \partial F(x)]$, *and*
(c) $h(f(x) + f'(x)d) - F(x) \leqslant \Delta_k(x, \rho)$ *for all*
$x \in L(x_0)$ *whenever* $d \in D_k(x, \rho, r)$.

Proof. As was noted in the discussion preceding the lemma, $D_k(x, \rho, r) \neq \emptyset$ whenever $S(\rho, x)$ is bounded for $k=3$. Moreover, (c) follows from the construction of $\Delta_k(x, \rho)$ and $D_k(x, \rho, r)$. Thus we need only establish (*b*). For $k=3$, (b) again follows by construction, and for $k=1$, (b) is an immediate consequence of Theorem 3.6(2). For $k=2$, (b) would also follow from Theorem 3.6(2) if we knew that

$$[0 \in D_2(x, \rho, 1)] \Leftrightarrow [\Delta_2(x, \rho)=0] \Leftrightarrow [0 \in \partial F(x)].$$

In order to see that this is indeed the case, recall that

$$[0 \in \partial F(x)] \Leftrightarrow [F'(x; d) \geq 0 \text{ for all } d \in \mathbb{R}^n] \tag{4.6}$$

[5, Proposition 2.3.2]. Next let $\sigma(x)$ be the selection from $\operatorname{argmin}\{\phi(d; x, \rho)\colon d \in \mathbb{R}^n\}$ used in defining $D_2(x, \rho, 1)$. Then, by Lemma 2.3b, we have the inequality

$$F'(x; \sigma(x)) \leq \Delta_2(x, \rho) \leq \Delta_1(x, \rho) \leq 0.$$

Hence if $0 \in \partial F(x)$, statement (4.6) implies that $\Delta_2(x, \rho)=0$. Conversely, if $\Delta_2(x, \rho)=0)=0$, then $\Delta_1(x, \rho)=0$, yielding $0 \in \partial F(x)$ via Theorem 3.6(2). Finally, the equivalence of $0 \in D_2(x, \rho, 1)$ and $\Delta_2(x, \rho)=0$ is apparent from their definitions. □

5. Convergence

In this section we determine conditions under which accumulation points of sequences generated by algorithm (2.1) are also stationary points of *F*. Theorem 2.3 eschews this issue and only speaks of the convergence of functional values. In fact, without the imposition of further requirements on the choice of casting functions, the efficacy of algorithm (2.2) is in serious doubt, as is illustrated by the following example.

5.1 Example. Choose $x_0 \in \mathbb{R}^n$, $\delta > 0$, and $\rho_0 \in \mathscr{C}$. Define $\rho \in \mathscr{C}^*$ as follows:

$$\rho(d, x) := \begin{cases} \rho_0(d) & \text{if } \|x - x_0\| \geq \delta, \\ \psi\left(d \,\middle|\, \left(\dfrac{\delta - \|x - x_0\|}{2}\right) B\right) & \text{otherwise,} \end{cases}$$

where $B := \{x\colon \|x\| \leq 1\}$. Then no matter what the function *F*, the iterates generated by algorithm (2.2) with initial point x_0 and search direction choice $D_k(x, \rho, r)$ for $k \in \{1, 2\}$, cannot escape the δ-ball about x_0.

Thus we see that further restrictions on the choice of casting function $\rho \in \mathscr{C}^*$ are required in order to obtain meaningful convergence results. We begin by observing that implicit in the usage of either of the stepsize choices $D_1(x, \rho, r)$ or $D_2(x, \rho, r)$ is the minimization of the convex functions $\phi(d; x, \rho)$ at every iteration. Thus we may view the algorithm as successively minimizing a sequence of convex functions that are themselves local approximations to the function in which our real interest lies. The natural and appropriate technique by which such optimization schemes are analyzed is via the notion of *epi-convergence.* The basic properties of epi-convergent sequences of convex functions as applied to optimiztion problems, are developed in, for example, the works of Attouch and Wets [2, 29].

5.2 Definition. *Let $\{f_i\}_{i=0}^{\infty}$ be a sequence of closed convex functions with domain in $\mathbb{R}^n$ and range $\mathbb{R}^* := \mathbb{R} \cup \{+\infty\}$. We say that $\{f_i\}$ converges pointwise to the closed convex function $f: \mathbb{R}^n \to \mathbb{R}^*$ and write $f_i \to^{\mathrm{p}} f$ if $\lim_i f_i(x) = f(x)$ for all $x \in \mathbb{R}^n$. We say that $\{f_i\}$ epi-converges to f and write $f_i \to^{\mathrm{e}} f$ if the epi-graphs of the f_i converge to the epi-graph of f, that is*

$$\operatorname{limsup} \operatorname{epi}(f_i) = \operatorname{epi}(f) = \operatorname{liminf} \operatorname{epi}(f_i),$$

where the epi-graph of a convex function $g: \mathbb{R}^n \to \mathbb{R}^*$ is the set $\operatorname{epi}(g) := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}: g(x) \leq \alpha, x \in \operatorname{Dom}(g)\}$.

The central result of this section is as follows.

5.3 Theorem. *Let $x_0 \in \mathbb{R}^n$ and let $F = h \circ f$ satisfy the hypothesis* (2.5). *Choose $\rho \in \mathscr{C}^*$ such that the set*

$$S^*(\rho, x_0) := \bigcup_{x \in L(x_0)} S(\rho, x)$$

is bounded. Let $k \in \{1, 2, 3\}$, and if $k \in \{1, 2\}$ select $r \in (0, 1]$; otherwise, select $r \in (0, 1)$. Suppose that $\{x_i\}$ is the sequence generated by algorithm (2.1) *with initial point x_0, stopping criteria $0 \in \partial F(x)$, and the designations*

$$D_i := D_k(x_i, \rho, r) \quad \text{and} \quad \Delta_i := \begin{cases} r\Delta_k(x_i, \rho) & \text{if } k \in \{1, 2\}, \\ \Delta_3(x_i, \rho) & \text{otherwise,} \end{cases}$$

for all $i = 0, 1, 2, \ldots$. If x^* is an accumulation point of $\{x_i\}$ with $y_j \to x^*$ and $\rho(\cdot, y_j) \to^{\mathrm{p}} \rho(\cdot, x^*)$ for some subsequence $\{y_j\}$ of $\{x_i\}$, then

$$\lim_j (\min\{\phi(d; y_j, \rho): d \in \mathbb{R}^n\}) = \min\{\phi(d; x^*, \rho): d \in \mathbb{R}^n\} = F(x^*), \tag{5.4}$$

$$\lim\sup[\operatorname{argmin}\{\phi(d; y_j, \rho): d \in \mathbb{R}^n\}] \subset \operatorname{argmin}\{\phi(d; x^*, \rho): d \in \mathbb{R}^n\}, \tag{5.5}$$

$F(x_i) \downarrow F(x^*)$, and $0 \in \partial F(x^*)$.

Proof. Let $\{y_j\}$ be as in the hypothesis with $\rho(\cdot, y_j) \to^{\mathrm{p}} \rho(\cdot, x^*)$. Then $\phi(\cdot\,; y_j, \rho) \to^{\mathrm{p}} \phi(\cdot\,; x^*, \rho)$ with $\operatorname{int}(\operatorname{Dom}(\phi(\cdot\,; x^*, \rho))) \neq \emptyset$. Hence by [29, Corollary 4], $\phi(\cdot\,; y_j, \rho) \to^{\mathrm{e}} \phi(\cdot\,; x^*, \rho)$. Thus, by [29, Theorem 7], we have the first half of (5.4), and by [29, Theorem 9], (5.5) also holds. Furthermore, $F(x_i) \downarrow F(x^*)$, since $\{F(x_i)\}$ is a decreasing sequence, and the sequence $\{\|d_i\|\}$ is uniformly bounded, since $\{d_i\} \subset S^*(\rho, x_0)$. Hence $\Delta_i \to 0$ by Theorem 2.4.

Case 1: $k=1$. Since $\Delta_i \to 0$ we have that

$$\lim_i \,[F(x_i) - \min\{\phi(d\,; x_i, \rho): d \in \mathbb{R}^n\}] = 0.$$

Therefore, by the first half of (5.4), we have that

$$F(x^*) = \lim_j F(y_j) = \lim_j \,[\min\{\phi(d\,; y_j, \rho): d \in \mathbb{R}^n\}]$$
$$= \min\{\phi(d\,; x^*, \rho): d \in \mathbb{R}^n\},$$

and so $0 \in \partial F(x^*)$ by Theorem 3.6(2).

Case 2: $k=2$. Since $\Delta_i \to 0$, we know that $\Delta_2(x_i, \rho) \to 0$ and so $\Delta_1(x_i, \rho) \to 0$, since $\Delta_2(x_i, \rho) \leqslant \Delta_1(x_i, \rho) \leqslant 0$. Hence we are back in Case 1, and so $0 \in \partial F(x^*)$ with $F(x^*) = \min\{\phi(d\,; x^*, \rho): d \in \mathbb{R}^n\}$, again by Theorem 3.6(2).

Case 3: $k=3$. From Theorem 3.7a we know that

$$0 \leqslant F(x_i) - \min\{\phi(d\,; x_i, \rho): d \in \mathbb{R}^n\} \leqslant -r^{-1}\Delta_3(x_i, \rho)$$

for all i sufficiently large, since $\Delta_i \to 0$. But then $\Delta_1(x_i, \rho) \to 0$, and so we are back in Case 1, thereby establishing the result. □

Due to the generality of the above result, there exist a multitude of corollaries and refinements that can be derived by, for example, identifying the specific type of localizing function one is interested in considering. Such analyses, although of great importance, are best left to papers wherein the local properties of these algorithms are also considered. In lieu of this analysis, however, we do provide a short list of examples implicated by our study and briefly indicate how the necessary computations can be performed by using only linear or quadratic programming subroutines.

6. Examples

In this section we indicate how one can compute the entities $\Delta_i(x, \rho)$ and $D_i(x, \rho, r)$ of Section 4 for two standard classes of problems. The first class of problems is that of unconstrained minimization of differentiable functions. In this case the convex function h is simply the identity map, and so, as one would expect, we simply recover many of the classical techniques. The second class of problems is that of constrained optimization via exact penalty methods. Indeed, it is for this class of problems that our approach was originally intended, and so, as we shall see, requires a good deal more care and effort.

6.1. Techniques for unconstrained minimization

In this section the convex function h is taken to be the identity map.

6.1.1. Cauchy methods

Define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \psi(d|B_2)$ for all d and x in $\mathbb{R}^n$, where B_2 is the unit ball for the l_2-norm. Then $\Delta_i(x, \rho) = -\|\nabla f(x)\|_2$ and $D_i(x, \rho, 1) = \{-\nabla f(x)\|\nabla f(x)\|_2^{-1}\}$ for $i = 1, 2$. For the ε-subdifferential approach, we need to choose $r \in (0, 1)$, in which case

$$\varepsilon(x) = \max\{r^p\colon r^p < \|\nabla f(x)\|_2, p = 0, 1, 2, \ldots\} = -\Delta_3(x, \rho),$$

and

$$D_3(x, \rho, r) = \{d \in B_2\colon \varepsilon(x)\|d\|_2 < -\nabla f(x)d, \text{ and } \varepsilon(x) \leqslant -\nabla f(x)d\}.$$

In particular, $-\nabla f(x)\|\nabla f(x)\|_2^{-1} \in D_3(x, \rho, r)$.

6.1.2. Variable metric methods

Let $\mathscr{H}$ be a set of real positive definite symmetric matrices, all of whose eigenvalues lie in a compact set, and suppose that to each $x \in \mathbb{R}^n$ there is associated some $H_x \in \mathscr{H}$. Define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \frac{1}{2}d^{\mathrm{T}}H_x d$. Then

$$2\Delta_1(x, \rho) = \Delta_2(x, \rho) = -\nabla f(x)^{\mathrm{T}}H_x^{-1}\nabla f(x)$$

and

$$2D_1(x, \rho, 1) = D_2(x, \rho, 1) = \{-H_x^{-1}\nabla f(x)\}.$$

For the ε-subdifferential approach choose $r \in (0, 1)$, then

$$\varepsilon(x) = \max\{r^p\colon 2r^p < \|\nabla f(x)\|_{H_x^{-1}}^2, p = 0, 1, 2, \ldots\} = -\Delta_3(x, \rho),$$

$$\partial_{\varepsilon(x)}\phi(0; x, \rho) = \{d^*\colon \|d^* - \nabla f(x)\|_{H_x^{-1}}^2 \leqslant 2\varepsilon(x)\},$$

and

$$D_3(x, \rho, r) = \{d\colon \psi^*(d|\partial_{\varepsilon(x)}\phi(0; x, \rho)) < 0, \nabla f(x)^{\mathrm{T}}d + \tfrac{1}{2}d^{\mathrm{T}}H_x d \leqslant -\varepsilon(x)\},$$

where $\|\cdot\|_{H_x^{-1}}$ is defined by the relation

$$\|z\|_{H_x^{-1}} := (z^{\mathrm{T}}H_x^{-1}z)^{1/2}$$

for all $z \in \mathbb{R}^n$. In particular, it is a simple matter to verify that $-H_x^{-1}\nabla f(x) \in D_3(x, \rho, r)$.

6.2. Exact penalty methods

Let $h\colon \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ be the function

$$h(x, y) := x + \alpha \operatorname{dist}_\nu(y|-K), \tag{6.2.1}$$

where $\alpha > 0$, $K = \mathbb{R}^m_+$, and $\operatorname{dist}_\nu(y|-K) := \inf\{\|y + k\|_\nu\colon k \in K\}$. Let $f\colon \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^m$ be defined by

$$f(y) := \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix}, \tag{6.2.2}$$

where $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}^m$ are both Fréchet differentiable, and set $F := h \circ f$. We now present a list of lemmas that serve to dissect the structure of the function h. The proofs we provide are in fact only sketches of proofs as we omit the explication of several details in the computations. For further information regarding these types of computations, the reader is referred to the following excellent references: [3, 13, 19, 27].

6.2.3 Lemma. *Let K be a closed convex cone in the real normed linear space Y. Then*

$$\operatorname{dist}_\nu(y|-K) = \gamma(y|B_\nu - K) = \psi^*(y|B_\nu^0 \cap K^*),$$

where $B_\nu := \{y : \|y\|_\nu \leq 1\}$.

(For the proof, see [4, 5].)

6.2.4 Definition. A norm $\|\cdot\|_\nu$ on $\mathbb{R}^m$ is said to be monotone if

$$|x_1| \leq |x_2| \;\Rightarrow\; \|x_1\|_\nu \leq \|x_2\|_\nu.$$

(Note: The l_p-norms are monotone for $1 \leq p \leq \infty$.)

6.2.5 Lemma. *If $\|\cdot\|_\nu$ is a monotone norm on $\mathbb{R}^m$, then*

$$\operatorname{dist}_\nu[y|\mathbb{R}^m_-] = \|y_+\|_\nu.$$

Proof. Note that if $z \in \mathbb{R}^m_+$, then $0 \leq x_+ \leq (x+z)_+$. Hence

$$\begin{aligned} \|x_+\|_\nu &\geq \inf_{z \in \mathbb{R}^m_+} \|x+z\|_\nu = \inf_{z \in \mathbb{R}^m_+} \|(x+z)_+ - (x-z)_-\|_\nu \\ &\geq \inf_{z \in \mathbb{R}^m_+} \|(x+z)_+\|_\nu \geq \|x_+\|_\nu. \qquad \square \end{aligned}$$

6.2.6 Lemma. *Let h be as in* (6.2.1), *with $\|\cdot\|_\nu$ monotone, then*

$$h^*(x^*, y^*) = \psi(x^*|\{1\}) + \psi(y^*|\alpha(B_\nu^0 \cap K^*))$$

and

$$\partial_\varepsilon h(x, y) = \{1\} \times \alpha\{y^* : y^* \in B_\nu^0 \cap K^*, \|y_+\|_\nu - \varepsilon\alpha^{-1} \leq \langle y^*, y\rangle\}.$$

Proof

$$\begin{aligned} h^*(x^*, y^*) &= \sup_{(x,y)} \{\langle (x, y), (x^*, y^*)\rangle - h(x, y)\} \\ &= \sup_x \{(x^* - 1)x\} + \alpha \sup_y \{\langle y, \alpha^{-1}y^*\rangle - \psi^*(y|B_\nu^0 \cap K^*)\} \\ &= \psi(x^*|\{1\}) + \psi(y^*|\alpha(B_\nu^0 \cap K^*)) \end{aligned}$$

$$\begin{aligned}\partial_\varepsilon h(x,y) &= \{(x^*,y^*)\colon h(x,y)+h^*(x^*,y^*)\leq\langle(x,y),(x^*,y^*)\rangle+\varepsilon\}\\ &= \{(1,y^*)\colon y^*\in\alpha(B^0_\nu\cap K^*),\ \alpha\psi^*(y|B^0_\nu\cap K^*)-\varepsilon\leq\langle y,y^*\rangle\}\\ &= \{(1,\alpha y^*)\colon y^*\in B^0_\nu\cap K^*,\ \psi^*(y|B^0_\nu\cap K^*)-\varepsilon\alpha^{-1}\leq\langle y^*,y\rangle\}\\ &= \{1\}\times\alpha\{y^*\colon y^*\in B^0_\nu\cap K^*,\ \|y_+\|_\nu-\varepsilon\alpha^{-1}\leq\langle y^*,y\rangle\}.\qquad\square\end{aligned}$$

6.2.7 Lemma. *Let f_1 and f_2 be two closed proper convex functions mapping $\mathbb{R}^n$ into $\mathbb{R}$ with* $\operatorname{int}(\operatorname{Dom}(f_1))\cap\operatorname{int}(\operatorname{Dom}(f_2))\neq\emptyset$. *Then*

$$\partial_\varepsilon(f_1+f_2)(x_0)=\bigcup_{\substack{\varepsilon_1\geq 0,\varepsilon_2\geq 0\\ \varepsilon_1+\varepsilon_2=\varepsilon}}\{\partial_{\varepsilon_1}f_1(x_0)+\partial_{\varepsilon_2}f_2(x_0)\}$$

for all $x_0\in\operatorname{Dom}(f_1)\cap\operatorname{Dom}(f_2)$.

Proof. This is just a special case of Theorem 2.1 in [13]. □

6.2.8 Lemma. *Let h and f be as in* (6.2.1) *and* (6.2.2), *respectively, and set $F=h\circ f$. For $\varepsilon>0$, define*

$$\partial_\varepsilon F(x):=\partial_\varepsilon h(f(x))\circ f'(x)\quad\textit{and}\quad f'_\varepsilon(x;d):=\psi^*(d|\partial_\varepsilon F(x))$$

for all $x\in\mathbb{R}^n$. Then given $x\in\mathbb{R}^n$ and a norm $\|\cdot\|_\nu$ on $\mathbb{R}^n$, we have that

$$-\operatorname{dist}_\nu[0,\partial_\varepsilon F(x)]=\min_{d\in B^0_\nu}f'_\varepsilon(x;d)$$

where there exist $\bar d\in B^0$ and $x^*\in\partial_\varepsilon F(x)$ satisfying

$$\min_{d\in B^0_\nu}f'_\varepsilon(x;d)=f'_\varepsilon(x;\bar d)=\langle x^*,\bar d\rangle=-\|x^*\|_\nu=-\operatorname{dist}_\nu[0,\partial_\varepsilon F(x)].$$

Proof. The result is a straightforward application of the Minimum Norm Duality Theorem for Convex Sets which can be found in [15, Theorem 1, p. 136]. (For a generalization to semi-norms, see [5].) □

Remark. We shall call the direction $\bar d$ obtained in the above lemma the *ε-steepest descent direction* for F at x.

The above lemmas provide the theoretical tools required to compute the examples that follow. We omit the derivations as they are quite lengthy although straightforward.

In the first set of examples, we indicate how the search direction choices $D_1(x,\rho,r)$ and $D_2(x,\rho,r)$ can be characterized by showing how to solve $\min\{\phi(d;x,\rho)\colon d\in\mathbb{R}^n\}$.

6.2.9(a) Variable-metric techniques. Let $\mathcal{H}$ be a set of real positive definite symmetric matrices, all of whose eigenvalues lie in a compact set, and suppose that to each

$x \in \mathbb{R}^n$, there is associated some $H_x \in \mathcal{H}$. Define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \frac{1}{2}d^{\mathrm{T}}H_x d$, and set $\|\cdot\|_\nu := \|\cdot\|_1$ in the definition of h in 6.2.1. Then

$$\min_{d \in \mathbb{R}^n} \phi(d; x, \rho) = \min_{(d,\gamma)} f_1'(x)d + \alpha\gamma + \tfrac{1}{2}d^{\mathrm{T}}H_x d$$

$$\text{subject to} \quad f_2(x) + f_2'(x)\, d \leq \gamma e$$

$$0 \leq \gamma.$$

6.2.9(b) Variable-metric with a trust region. Let $\mathcal{H}$ be as defined in 6.9(a), let T be a compact set of positive real numbers, and suppose that to each $x \in \mathbb{R}^n$ there is associated some $H_x \in \mathcal{H}$ and some $\beta_x \in T$. Define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \frac{1}{2}d^{\mathrm{T}}H_x d + \psi(d|\beta_x B_\infty)$, and set $\|\cdot\|_\nu := \|\cdot\|_1$ in the definition of h in 6.2.1. Then

$$\min_{d \in \mathbb{R}^n} \phi(d; x, \rho) = \min_{(d,z)} f_1'(x)d + \alpha e^{\mathrm{T}}z + \tfrac{1}{2}d^{\mathrm{T}}H_x d$$

$$\text{subject to} \quad f_2(x) + f_2'(x)\, d \leq z,$$

$$0 \leq z,$$

$$-\beta_x e \leq d \leq \beta_x e.$$

6.2.9(c) Sequential linear programming. Let T be as in 6.9(b) and define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \psi(d|\beta_x B_\infty)$, and set $\|\cdot\|_\nu := \|\cdot\|_\infty$ in the definition of h in 6.2.1. Then

$$\min_{d \in \mathbb{R}^n} \phi(d; x, \rho) = \min_{(d,\gamma)} f_1'(x)d + \alpha\gamma$$

$$\text{subject to} \quad f_2(x) + f_2'(x)d \leq \gamma e,$$

$$0 \leq \gamma,$$

$$-\beta_x e \leq d \leq \beta_x e.$$

In the second set of examples, we indicate how a search direction $d \in D_3(x, \rho, r)$ can be determined. Our approach is to use the directions of ε-steepest descent, thereby simultaneously determining whether or not $0 \in \partial_\varepsilon\phi(0; x, \rho)$. For these examples, however, one should note that we restrict ourselves to cases where ρ is representable as a linear combination of polyhedral norms and indicator functions so that $\partial_\varepsilon\phi(0; x, \rho)$ can be represented by linear systems of equations and inequalities.

6.2.10(a). Let $\beta: \mathbb{R}^n \to \mathbb{R}_+$ be such that

$$\xi \operatorname{dist}_1[0, \partial F(x)] \leq \beta(x) \leq \xi_2 \operatorname{dist}_1[0, \partial F(x)]$$

for all $x \in \mathbb{R}^n$, where $0 < \xi_1 \leq \xi_2 < 1$, and define $\rho \in \mathscr{C}^*$ as in the remark after definition (3.5) by

$$\rho(d, x) := \beta(x)\|d\|_1 \quad \text{for all } x \in \mathbb{R}^n.$$

Finally, let $\|\cdot\|_\nu := \|\cdot\|_1$ in the definition of h in 6.2.1. Then

$$\begin{aligned}\operatorname{dist}_2[0, \partial_\varepsilon\phi(0; x, \rho)] = \min_{(z_1,z_2)} \ & \tfrac{1}{2}\|f_1'(x)+\alpha f_2'(x)^{\mathrm{T}}z_1+\beta(x)z_2\|_2^2 \\ \text{subject to} \ & 0 \leqslant z_1 \leqslant e, \\ & -e \leqslant z_2 \leqslant e, \\ & \|f_2(x)_+\|_1 - \varepsilon\alpha^{-1} \leqslant \langle z_1, f_2(x)\rangle.\end{aligned}$$

6.2.10(b). Let T be as in 6.2.9(b) and define $\rho \in \mathscr{C}^*$ by $\rho(d, x) := \psi(d|\beta_x B_1)$ for all $x \in \mathbb{R}^n$, and set $\|\cdot\|_\nu := \|\cdot\|_\infty$ in the definition of h in 6.2.1, then

$$\begin{aligned}\operatorname{dist}_\infty[0, \partial_\varepsilon\phi(0; x, \rho)] = \min_{(\varepsilon_1,\varepsilon_2,z_1,z_2,\gamma)} \ & \gamma \\ \text{subject to} \ & \varepsilon = \varepsilon_1+\varepsilon_2, \ 0 \leqslant \varepsilon_1, \ 0 \leqslant \varepsilon_2. \\ & 0 \leqslant z_1, \ e^{\mathrm{T}}z_1 \leqslant 1, \\ & \|f_2(x)_+\|_\infty - \varepsilon_1\alpha^{-1} \leqslant \langle z_1, f_2(x)\rangle, \\ & -\varepsilon_2 e \leqslant z_2 \leqslant \varepsilon_2 e, \\ & -\gamma e \leqslant f_1'(x)+\alpha f_2'(x)^{\mathrm{T}}z_1+\beta_x z_2 \leqslant \gamma e.\end{aligned}$$

Once $x^* := \binom{z_1}{z_2}$, solving $\operatorname{dist}[0, \partial_\varepsilon\phi(0; x, \rho)]$, is obtained, one employs the alignment condition of Lemma 6.2.8 to compute a direction of ε-steepest descent, $\bar{d}$. (Alternatively, one could solve the dual of the above programs to get $\bar{d}$ directly.) Next, perform the one-dimensional minimization

$$\min_{\lambda \geqslant 0} \phi(\lambda\bar{d}; x, \rho), \tag{6.2.11}$$

which is also a linear or quadratic program, to obtain $\bar{\lambda}$. More specifically, the line search (6.2.11) can be performed by a recent algorithm of Lemarechal and Mifflin [16], as their method is finitely convergent for the examples that we have considered. Finally, by Theorem 3.7, $\bar{\lambda}\bar{d} \in D_3(x, \rho, r)$. A more detailed exposition of computations of this type can be found in Bertsekas and Mitter [3].

Remark. In examples 6.2.9(b) and (c), and 6.2.10(b), some care must be taken with respect to the set T and the association $\beta_x \in T$ to each $x \in \mathbb{R}^n$ in order to obtain the existence of a subsequence $\{y_j\} \subset \{x_i\}$ for which $\rho(\cdot, y_j) \to^{\mathrm{P}} \rho(\cdot, x^*)$, as is required for the conclusion of Theorem 5.3 to hold. The existence of such a subsequence $\{y_j\}$ is only guaranteed if there is a subsequence $\{\beta_{x_j}: j \in J\}$ of the sequence $\{\beta_{x_i}\}$ that is nonincreasing and for which $\lim_{j\in J} x_j = x^*$.

Acknowledgement

The careful reading and detailed comments of two referee's is gratefully acknowledged.

References

[1] D.H. Anderson and M.R. Osborne, "Discrete, nonlinear approximation problems in polyhedral norms", *Numerische Mathematik* 28 (1977) 143–156.

[2] H. Attouch and R.J.-B. Wets, "Approximation and convergence in nonlinear optimization", in: O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, eds., *Nonlinear programming* 4 (Academic Press, New York, 1981) pp. 367–394.

[3] D.P. Bertsekas and S. Mitter, "A descent numerical method for optimization problems with nondifferentiable cost functionals", *SIAM Journal on Control and Optimization* 11 (1973) 637–652.

[4] J.V. Burke and S.-P. Han, "A Gauss–Newton approach to solving generalized inequalities", Preprint, Department of Mathematics, University of Kentucky (Lexington, Kentucky, 1984).

[5] J.V. Burke, *Methods for solving generalized inequalities with applications to nonlinear programming*, Ph.D. Thesis, Department of Mathematics, University of Illinois at Urbana-Champaign (1983).

[6] F.H. Clarke, *Optimization and nonsmooth analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts (John Wiley & Sons, NY, 1983).

[7] L.C.W. Dixon, "Reflections on nondifferentiable optimization, Part I, Ball gradient", *Journal of Optimization Theory and Applications* 32 (1980) 123–133.

[8] L.C.W. Dixon and M. Gaviano, "Reflections on nondifferentiable optimization, Part 2, Convergences", *Journal of Optimization Theory and Applications* 32 (1980) 259–275.

[9] R. Fletcher, "A model algorithm for composite nondifferentiable optimization problems", *Mathematical Programming Study* 17 (1982) 67–76.

[10] U.M. Garcia-Palomares and A. Restuccia, "A global quadratic algorithm for solving a system of mixed equalities and inequalities", *Mathematical Programming* 21 (1981) 290–300.

[11] U.M. Garcia-Palomares and A. Restuccia, "Application of the Armijo stepsize rule to the solution of a nonlinear system of equalities and inequalities", *Journal of Optimization Theory and Applications* 41 (1983) 405–415.

[12] A.A. Goldstein, "Optimization of Lipschitz continuous functions", *Mathematical Programming* 13 (1977) 14–22.

[13] J.-B. Hiriart-Urruty, "ε-Subdifferential calculus", Proceedings of the Colloquium "Convex Analysis and Optimization", Imperial College (London, 1980).

[14] C.L. Lawson and R. Hanson, *Solving least squares problems* (Prentice-Hall, Englewood Cliffs, NJ, 1974).

[15] D.G. Luenberger, *Optimization by vector space methods* (John Wiley & Sons, NY, 1969).

[16] C. Lemarechal and R. Mifflin, "Global and superlinear convergence of an algorithm for one-dimensional minimization of convex functions", *Mathematical Programming* 24 (1982) 241–256.

[17] R. Mifflin, "Semi-smooth and semi-convex functions in constrained optimization", *SIAM Journal on Control and Optimization* 15 (1977) 959–972.

[18] R. Mifflin, "An algorithm for constrained optimization with semi-smooth functions", *Mathematics of Operations Research* 2 (1977) 191–207.

[19] J.-J. Moreau, "Fonctionelles convex", Lecture Notes, Séminaire "Equations aux derivées partielles", Collège de France, 1966.

[20] M.R. Osborne and G.A. Watson, "Nonlinear approximation problems in vector norms", in: G.A. Watson, ed., *Numerical analysis, Dundee 1977*, Lecture Notes in Mathematics 630 (Springer-Verlag, Berlin, 1978) pp. 115–132.

[21] M.R. Osborne, "Algorithms for nonlinear approximation", in: C.T.H. Baker and C. Phillips, eds., *The numerical solution of nonlinear problems* (Clarendon Press, Oxford, 1981) pp. 270–286.

[22] M.R. Osborne, "Strong uniqueness in nonlinear approximation", in: C.T.H. Baker and C. Phillips, eds., *The Numerical Solution of Nonlinear Problems* (Clarendon Press, Oxford, 1981) pp. 287-304.

[23] E. Polak, D.Q. Mayne, and Y. Wardi, "On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems", *SIAM Journal on Control and Optimization* 21 (1983) 179-203.

[24] M.J.D. Powell, "General algorithms for discrete nonlinear approximation calculations", in: C.K. Chui, L.L. Schumaker, and J.D. Ward, eds., *Approximation theory IV* (Academic Press, NY, 1983) pp. 187-218.

[25] M.J.D. Powell, "On the global convergence of trust-region algorithms for unconstrained minimization", *Mathematical Programming* 29 (1984) 297-303.

[26] M.J.D. Powell and Y. Yuan, "Conditions for superlinear convergence in l_1 and l_∞ solutions of overdetermined nonlinear equations", *IMA Journal of Numerical Analysis* 4 (1984) 241-251.

[27] R.T. Rockafellar, *Convex analysis* (Princeton University Press, Princeton, NJ, 1970).

[28] K. Schittkowski, "Numerical solution of systems of nonlinear inequalities", in: *Optimization and operations research*, Lecture Notes in Economics and Mathematical Systems 117 (Springer-Verlag, Berlin, 1975) pp. 259-272.

[29] R.J.-B. Wets, "Convergence of convex functions, variational inequalities, and convex optimization problems", in: R.W. Cottle, F. Giannessi, and J.-L. Lions, eds., *Variational inequalities and complementarity problems* (John Wiley & Sons, NY, 1980) pp. 375-403.

[30] P. Wolfe, "Convergence conditions for ascent methods", *SIAM Review* 11 (1969) 226-235.

[31] Y. Yuan, "Global convergence of trust region algorithms for nonsmooth optimization", Report DAMTP 1983, Cambridge University (Cambridge, England, 1983).

[32] Y. Yuan, "Some properties of trust region algorithms for non-smooth optimization", Report DAMTP 1983, Cambridge University (Cambridge, England, 1983).