THE MINIMIZATION OF DISCONTINUOUS FUNCTIONS: MOLLIFIER SUBGRADIENTS

Yuri M. Ermoliev
International Institute for Applied System Analysis, A-2361 Laxenburg, Austria

Vladimir I. Norkin
Glushkov Institute of Cybernetics, 252207 Kiev, Ukraine

*Roger J-B Wets**
Mathematics, University of California, Davis, CA 95616, U.S.A.

SIAM Journal on Control and Optimization, 33 (1995), 149-167.

Abstract. To minimize discontinuous functions, that arise in the context of systems with jumps for example, we propose a new approach based on approximation via averaged functions (obtained by convolution with mollifiers). The properties of averaged functions are studied, after it is shown that they can be used in an approximation scheme consistent with minimization. A new notion of subgradient is introduced based on approximations generated by mollifiers, and its properties are exploited in the design of minimization procedures.

Keywords: impulse control, discrete events systems, averaged functions, subgradients, subdifferentiability, stochastic quasi-gradients, epi-convergence.

Date: August 30, 1992
Version: May 28, 2009

* Supported in part by a grant from the U.S.-Israel Binational Science Foundation

1. Introduction

It is not unusual to have to deal with optimization problems involving discontinuous functions, for example: optimization problems involving set-up costs or impulse controls (Bensoussan and Lions [**5**]), the control of discrete events systems (Gong and Ho [**14**], Rubinstein [**36**], Ermoliev and Gaivoronski [**9**]), and control problems with pre- and post-accident regimes whose systems' parameters do not evolve continuously. Even a convex optimization problem is sometimes replaced by one involving discontinuous penalties such as indicator or characteristic functions. Problems defined in terms of marginal functions, expressing the dependence of the optimal value of some subproblem (as in stochastic programming problems, for example) on certain parameters are in general discontinuous. In order to deal with such applications, a number of efforts have been made to develop a subdifferential calculus for nonsmooth, and possibly discontinuous, functions. Among the many possibilities let us mention the notions due to Rockafellar [**31**], Aubin [**3**], Clarke [**6**], Ioffe [**18**], Frankowska [**11**], Michel and Penot [**25**] and Mordukhovich [**26**] in the context of variational analysis, to Warga [**43**] for subdifferentials obtained via certain approximating scheme, to Demyanov and Rubinov [**7**] for quasi-differentiable functions, and to Ermoliev [**9**] and Polyak [**30**] in the context of stochastic approximation techniques for optimization problems.

Another approach to the differentiation of "nonclassical" functions, which eventually became known as the *theory of distributions* (in Russia, as the *theory of generalized functions*), was developed in 1930's by Sobolev [**38**] and Schwartz [**37**]. This technique is in wide use in mathematical physics and related engineering problems. Although, one can find in the literature occasional reference to a connection between these two developments, the notion of differentiability in the sense of distributions is not used in variational analysis or in the design of solution procedures for optimization problems involving "nonclassical" functions. Probably, one of the reasons for this, is that in the theory of distributions, (standard) functions defined on $\mathbb{R}^n$ are redefined as functionals on a certain functional space. The same applies to their gradients.

In the development of a subdifferential calculus for (discontinuous) functions, we shall appeal to some of the results of the theory of distributions, but our aim is to bring back the algebraic manipulations to operations that can be carried out in $\mathbb{R}^n$, in particular, by assigning a family of distributions to a point in $\mathbb{R}^n$. More specifically, we associate with a point $x \in \mathbb{R}^n$, a family of mollifiers (density functions) whose support tends toward x and converge to the dirac function $\delta(x - \cdot)$. Given such a family, say $\{\psi_\theta, \theta \in \mathbb{R}_+\}$, a "generalized" function associated with a function $f : \mathbb{R}^n \to \mathbb{R}$ is then defined as the clusters of all possible values generated by the pairings of f with ψ_ν. A set of generalized gradients, called here *mollifier subgradients* is defined in a similar fashion.

From another angle, one can also link this approach to a technique involving "averaged" functions introduced by Steklov [**39**], [**40**] and Sobolev [**38**]. In the case of continuous functions, these averaged functions converge uniformly to f, and is then related to an

approach suggested by Warga [**42-44**], see also Frankowska [**12**].

For the gradients of averaged functions there are simple unbiased stochastic estimators based on finite differences (some will be mentioned later on). This opens up the possibility of minimizing the original (discontinuous) function through the minimization of a sequence of smooth approximating averaged functions. Such an approach, initiated in section 5, relies on the ideas inherent in stochastic quasi-gradient methods and dynamic nonstationary optimization as were used by Ermoliev and Nurminski [**10**], Gaivoronski [**13**], Katkovnik [**19**], Nikolaeva [**27**] in convex nondifferentiable optimization, by Gupal [**15**], Mayne and Polak [**24**] in the Lipschitz continuous case, and by Gupal and Norkin [**17**] in the discontinuous case.

Section 2 introduce a notion of convergence for discontinuous function, and prepares the way to a justification that averaged functions are consistent approximating functions when dealing with the minimization of a discontinuous functions. Section 3 is devoted to the properties of averaged functions, and section 4 introduces the notion of a mollifier subgradient based on the approximation of a discontinuous function by averaged functions. Finally section 5, outlines some potential optimization procedures.

2. eh-Convergence

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper ($f \not\equiv \infty$, $f > -\infty$) extended real-valued function with $\operatorname{dom} f = \{x \in \mathbb{R}^n | f(x) < \infty\}$ the (nonempty) set on which it is finite. Its *epigraphical (or lower semicontinuous) closure* $\operatorname{cl}_e f$ is given by

$$\operatorname{cl}_e f(x) := \liminf_{x' \to x} f(x') = \inf_{x^\nu \to x} \liminf_{\nu \to \infty} f(x^\nu)$$

and its *hypographical (or upper semicontinuous) closure* $\operatorname{cl}_h f$ is

$$\operatorname{cl}_h f(x) := \limsup_{x' \to x} f(x') = \sup_{x^\nu \to x} \limsup_{\nu \to \infty} f(x^\nu);$$

inf and sup are taken over all sequences x^ν converging to x. The function $\operatorname{cl}_e f$ is lower semicontinuous and $\operatorname{cl}_h f$ is upper semicontinuous.

For an arbitrary sequence of functions $\{ f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu \in \mathbb{N} \}$, we denote by $\mathrm{e}{-}\operatorname{li} f^\nu$ its *lower epi-limit*, i.e.,

$$(\mathrm{e}{-}\operatorname{li} f^\nu)(x) := \inf_{x^\nu \to x} \liminf_{\nu \to \infty} f^\nu(x^\nu),$$

and by $\mathrm{h}{-}\operatorname{ls} f^\nu$ its *upper hypo-limit*, i.e.,

$$(\mathrm{h}{-}\operatorname{ls} f^\nu)(x) := \sup_{x^\nu \to x} \limsup_{\nu \to \infty} f^\nu(x^\nu);$$

here also inf and sup are calculated with respect to all sequences converging to x. It is easy to see that $\mathrm{e}{-}\operatorname{li} f^\nu$ is lower semicontinuous and that $\mathrm{h}{-}\operatorname{ls} f^\nu$ is upper semicontinuous, if necessary cf. [**33**] for more details; note that $\mathrm{h}{-}\operatorname{ls} f^\nu = -\,\mathrm{e}{-}\operatorname{li}(-f^\nu)$.

2.1. Definition. *Given a sequence of functions* $\{ f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu \in \mathbb{N} \}$, *a function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is an epi-sublimit of the sequence* $\{f^\nu\}$ *if* $\mathrm{cl}_e\, f \le \mathrm{e{-}li}\, f^\nu$. *It is a hypo-suplimit if* $\mathrm{h{-}ls}\, f^\nu \le \mathrm{cl}_h\, f$. *If* f *is both an epi-sublimit and a hypo-suplimit, we shall say that the sequence* f^ν *eh-converges to* f.

One can view eh-convergence as an *extended graph-convergence*. With $\mathrm{gph}\, f^\nu$, the graph of the function f^ν, eh-convergence means that

$$\mathrm{Lim\,sup}_{\nu\to\infty}\, \mathrm{gph}\, f^\nu \subset \{\, (x,\alpha) \in \mathbb{R}^n \times \overline{\mathbb{R}} \mid \mathrm{cl}_e\, f(x) \le \alpha \le \mathrm{cl}_h\, f(x) \,\}$$

where Lim sup is the *outer (superior) set-limit*; for a sequence of sets C^ν, $\mathrm{Lim\,sup}_\nu\, C^\nu$ consists of the cluster points of all sequences $\{u^\nu\}$ with $u^\nu \in C^\nu$ for ν sufficiently large.

A notion of eh-convergence (for functions with values in a function space) also surfaced in the study of the stability properties of integral functionals with discontinuous integrands, Artstein and Wets [**1**].

3. Averaged functions

Averaged functions will be defined relative of a specific family of mollifiers; our usage of the term mollifier differs somewhat from the standard one in that we do not require that mollifiers be necessarily analytic.

3.1. Definition. *Given a locally integrable function* $f : \mathbb{R}^n \to \mathbb{R}$ *and a family of mollifiers* $\{\, \psi_\theta : \mathbb{R}^n \to \mathbb{R}_+, \theta \in \mathbb{R}_+ \,\}$ *that by definition satisfy*

$$\int_{\mathbb{R}^n} \psi_\theta(z)\, dz = 1, \quad \mathrm{supp}\, \psi_\theta := \{\, z \in \mathbb{R}^n \mid \psi_\theta(z) > 0 \,\} \subset \rho_\theta\, \mathbb{B} \quad \textit{with } \rho_\theta \downarrow 0 \textit{ as } \theta \downarrow 0,$$

the associated family $\{\, f_\theta, \theta \in \mathbb{R}_+ \,\}$ *of averaged functions is defined by*

$$f_\theta(x) := \int_{\mathbb{R}^n} f(x-z)\psi_\theta(z)\, dz = \int_{\mathbb{R}^n} f(z)\psi_\theta(x-z)\, dz.$$

For example, the family of mollifiers could be of the following type: let ψ be a density function with $\mathrm{supp}\, \psi$ bounded, $\alpha_\theta \downarrow 0$ as $\theta \downarrow 0$, and

$$\psi_\theta(z) := \frac{\psi(z/\alpha_\theta)}{(\alpha_\theta)^n}\,.$$

A mollifier is thus a probability density function defined on $\mathbb{R}^n$ but the family $\{\psi_\theta\}$ must possess some specific properties. One can also express f_θ as a convolution

$$f_\theta = f \star \psi_\theta.$$

Sobolev [**38**] introduced "averaged functions" in his study of generalized functions (distributions) that could serve as solutions of certain equations in mathematical physics; he also required that the mollifiers ψ_θ be analytic (of class C^∞). In terms of the theory of distributions, $f_\theta(x)$ is the value of the distribution f at $\psi_\theta(x - \cdot)$, x playing the role of a parameter.

3.2. Theorem. *Let $\{ f_\theta, \theta \in \mathbb{R}_+ \}$ be a family of averaged functions associated with a locally integrable function $f : \mathbb{R}^n \to \mathbb{R}$, and suppose that $x^\theta \to x$ as $\theta \downarrow 0$. Then*

$$\mathrm{cl}_e\, f(x) \le \liminf_{\theta \downarrow 0} f_\theta(x^\theta) \le \limsup_{\theta \downarrow 0} f_\theta(x^\theta) \le cl_h f(x).$$

Consequently, the averaged functions f_θ eh-converge to f.

Proof. It will suffice to prove the first inequality, the second one is evident and the proof of the last one is similar to that of the first. eh-convergence is an immediate consequence of this string of inequalities.

By definition of lower semicontinuity, for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists V, a neighborhood of 0, such that $f(x - z) \ge \mathrm{cl}_e\, f(x) - \varepsilon$ for all $z \in V$. For θ sufficiently small, $\operatorname{supp} \psi_\theta \subset V$ and then

$$\begin{aligned} f_\theta(x^\theta) = \int_{\mathbb{R}^n} f(x^\theta - z)\psi_\theta(z)\, dz &= \int_V f(x^\theta - z)\psi_\theta(z)\, dz \\ &\ge \int_V \mathrm{cl}\, f(x^\theta - z)\psi_\theta(z)\, dz \ge (\mathrm{cl}_e\, f(x^\theta) - \varepsilon) \int \psi_\theta(z)\, dz. \end{aligned}$$

Hence, $\liminf_{\theta \downarrow 0} f_\theta(x^\theta) \ge \mathrm{cl}_e\, f(x) - \varepsilon$. The proof is completed by letting $\varepsilon \downarrow 0$. □

3.3. Corollary. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous, and $\{f_\theta, \theta \in \mathbb{R}_+\}$ an associated family of averaged functions. Then, the averaged functions f_θ converge continuously to f, i.e., $f_\theta(x^\theta) \to f(x)$ for all $x^\theta \to x$. In fact, the averaged functions f_θ converge uniformly to f on every bounded subset of $\mathbb{R}^n$.*

Proof. Evident. □

When the function f is not continuous, one cannot expect to have continuous convergence of the averaged functions to f. But that is also more than what is required. For our purposes, we only need to establish that the averaged functions converge to f is a sense that will guarantee the convergence of minimizers and infima. This is precisely what is accomplished by epi-convergence.

3.4. Definition (Aubin and Frankowska, [**4**], Rockafellar and Wets [**33**])**.** *A sequence of functions $\{ f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu \in \mathbb{N} \}$ epi-converges to $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at x if*

(i) $\liminf_{\nu \to \infty} f^\nu(x^\nu) \ge f(x)$ *for all* $x^\nu \to x$*;*

(ii) $\lim_{\nu \to \infty} f^\nu(x^\nu) = f(x)$ *for some sequence* $x^\nu \to x$*.*

The sequence $\{f^\nu\}_{\nu \in \mathbb{N}}$ epi-converges to f if this holds for all $x \in \mathbb{R}^n$, in which case we write $f = \mathrm{e}\text{-}\lim f^\nu$.

Clearly, if f is the epi-limit of some sequence, then f is necessarily lower semicontinuous. Moreover, if the f^ν converge continuously, and a fortiori uniformly, to f, they also epi-converge to f.

For example, if $(x,y) \mapsto g(x,y) : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is (jointly) lsc at $(\bar{x},\bar{y})$ and is continuous in y at $\bar{y}$, then for any sequence $y^\nu \to \bar{y}$, the corresponding sequence of functions $\{\, f^\nu(\cdot,y^\nu), \nu \in \mathbb{N}\,\}$ epi-converges to $f(\cdot,\bar{y})$ at $\bar{x}$.

3.5. Theorem (Attouch and Wets [**2**]). *If the sequence of functions $\{\, f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu \in \mathbb{N}\,\}$ epi-converges to $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at all $x \in D \subset \mathbb{R}^n$, then for any compact set $K \subset D$, one has*

$$\inf_K f^\nu \;\longrightarrow\; \inf_K f,$$

and

$$\forall\, x^\nu \to x : \quad [\, f^\nu(x^\nu) \le \inf f^\nu + \varepsilon_\nu, \quad \varepsilon_\nu \downarrow 0, \,] \quad \Longrightarrow \quad x \in \operatorname{argmin} f.$$

Epi-convergence of the averaged functions f_θ to f will be guaranteed by the following property of f:

3.6. Definition. *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strongly lower semicontinuous at x, if it is lower semicontinuous at x and there exists a sequence $x^\nu \to x$ with f continuous at x^ν (for all ν) such that $f(x^\nu) \to f(x)$. The function f is strongly lower semicontinuous if this holds at all x.*

Strong lower semi-continuity excludes the possibility of discontinuities that are localized on lower dimensional subsets of $\mathbb{R}^n$. If we think of $(x, f(x))$ as the state of a system, strong lower semicontinuity means that this state can always be reached by following a path along which the evolution of the system is continuous (with no jumps). If x is "time-dependent", then although we may expect sudden changes from one state to another, either before or after the jump, the evolution will be continuous, one doesn't expect instantaneous jumps followed by an immediate return to normal regime.

3.7. Theorem. *For any strongly lower semicontinuous, locally integrable function $f : \mathbb{R}^n \to \mathbb{R}$, and any associated family $\{\, f_\theta,\, \theta \in \mathbb{R}_+\,\}$ of averaged functions, one has that $f = \mathrm{e}{-}\lim f_\theta$, i.e., for any sequence $\theta^\nu \downarrow 0$, $f = \mathrm{e}{-}\lim f_{\theta^\nu}$.*

Proof. Pick any x. We are going to show that the f_θ epi-converge to f at x. The strong lower semicontinuity of f at x provides us with a sequence $x^\nu \to x$ such that $f(x^\nu) \to f(x)$ with f continuous at x^ν. From corollary 3.3, it follows that for all ν, $f_\theta(x^\nu) \to f(x^\nu)$, and consequently a standard diagonalization process will yield (for any sequence $\theta^k \to 0$ as $k \to \infty$) a sequence x^k such that $f_{\theta^k}(x^k) \to f(x)$. This yields condition (ii) in definition 3.4. For condition (i) of definition 3.4, we simply appeal to proposition 3.2. □

Theorem 3.7 tells us that if one has to minimize the function f, the averaged functions f_θ could be used in a consistent approximation scheme, i.e., that implies the convergence of the minimizers. However, before we follow this route, we would have to make sure that their properties makes them amenable to minimization by existing —or possibly, modified— algorithmic procedures. The remainder of this section is devoted to the continuity and

differentiability properties of averaged functions, in particular for the class of Steklov (averaged) functions.

3.8. Definition. *Given a locally bounded function $f : \mathbb{R}^n \to \mathbb{R}$, the Steklov (averaged) functions are defined as follows: for $\alpha > 0$*

$$f_\alpha(x) = \int_{\mathbb{R}^n} f(x-z)\psi_\alpha(z)\,dz$$

where

$$\psi_\alpha(z) = \begin{cases} 1/\alpha^n, & \text{if } \max_{1,\dots,n} |z_i| \le \alpha/2; \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently,

$$f_\alpha(x) = \frac{1}{\alpha^n} \int_{x_1-\alpha/2}^{x_1+\alpha/2} dy_1 \dots \int_{x_n-\alpha/2}^{x_n+\alpha/2} dy_n\, f(y).$$

This class of averaged functions were introduced by Steklov [**39**] in 1907, and used by Kolmogorov and Fréchet for compactness tests in $\mathcal{L}^p$. In the context of smooth optimization, they were used by Katkovnik [**19**], Nikolaeva [**27**], Gupal [**15-16**] and Mayne and Polak [**24**].

The next proposition records the well-know fact that Steklov functions are locally Lipschitz continuous.

3.9. Proposition. *For locally bounded functions $f : \mathbb{R}^n \to \mathbb{R}$, the associated Steklov functions f_α are locally Lipschitz continuous, i.e., on each compact set $K \subset \mathbb{R}^n$, the function f_α is Lipschitz continuous on K with Lipschitz constant κ,*

$$\kappa = (2n/\alpha) \sup_{x \in K_\alpha} f(x), \quad \text{where} \quad K_\alpha := \{\, x + z \,|\, x \in K,\ \max_{i=1,\dots,n} |z_i| \le \alpha/2 \,\}.$$

Differentiability of average functions, however, cannot be guaranteed in general, unless the mollifiers ψ_θ are sufficiently smooth or if f itself has a sufficient level of continuity.

3.10. Proposition (Sobolev [**38**], Schwartz [**37**])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally integrable. Whenever the mollifiers ψ_θ are smooth (of class C^1), so are the associated averaged functions f_θ with gradient*

$$\nabla f_\theta(x) = \int_{\mathbb{R}^n} f(y) \nabla \psi_\theta(x-y)\,dy$$

3.11. Proposition (Gupal [**15**])**.** *For $f : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, the Steklov (averaged) functions f_α are continuously differentiable, and their gradients are given by*

$$\nabla f_\alpha(x) = \sum_{i=1}^{n} e_i \frac{1}{\alpha^n - 1} \int_{x_1-\alpha/2}^{x_1+\alpha/2} dy_1 \cdots \int_{x_{i-1}-\alpha/2}^{x_{i-1}+\alpha/2} dy_{i-1} \int_{x_{i+1}-\alpha/2}^{x_{i+1}+\alpha/2} dy_{i+1} \cdots \int_{x_n-\alpha/2}^{x_n+\alpha/2} dy_n$$
$$\frac{1}{\alpha}\left[f(y_1, \dots, y_{i-1}, x_i - \tfrac{1}{2}\alpha, y_{i+1}, \dots y_n) - f(y_1, \dots, y_{i-1}, x_i + \tfrac{1}{2}\alpha, y_{i+1}, \dots y_n) \right]$$

where e_i is the i-th unit coordinate vector.

This gradient can also be expressed as

$$\nabla f_\alpha(x) = \sum_{i=1}^{n} e_i \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_{i-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_{i+1} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_n \, \lambda_\alpha(x,\xi)$$

where

$$\begin{aligned}\lambda_\alpha(x,\xi) = \frac{1}{\alpha}\,[&f(x_1+\alpha\xi_1,\ldots,x_{i-1}+\alpha\xi_{i-1},x_i+\tfrac{1}{2}\alpha,x_{i+1}+\alpha\xi_{i+1},\ldots,x_n+\alpha\xi_n)\\ &- f(x_1+\alpha\xi_1,\ldots,x_{i-1}+\alpha\xi_{i-1},x_i-\tfrac{1}{2}\alpha,x_{i+1}+\alpha\xi_{i+1},\ldots,x_n+\alpha\xi_n)].\end{aligned}$$

This means that $\nabla f_\theta(x)$ is the expectation of the random vector $\boldsymbol{\lambda}_\alpha(x,\boldsymbol{\xi})$ where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\ldots\boldsymbol{\xi}_n)$ is a random vector, whose elements are independent and uniformly distributed on $[\,-1/2,1/2\,]$. In other words, $\boldsymbol{\lambda}_\alpha(x,\boldsymbol{\xi})$ is an unbiased estimator of the gradient of f_α at x.

3.12. Remark. Although, in the case of discontinuous functions f, we cannot "reach" differentiability for Steklov functions, it is always possible to do so, if the averaging process is repeated a second time. This follows immediately from propositions 3.9 and 3.11. Given a locally integrable function $f : \mathbb{R}^n \to \mathbb{R}$, let

$$\begin{aligned}f_{\alpha\beta}(x) &:= \int_{\mathbb{R}^n} f_\alpha(x-z)\psi_\beta(z)\,dz\\ &= \int_{\mathbb{R}^n} dy \int_{\mathbb{R}^n} dz\, f(x-y-z)\psi_\alpha(y)\psi_\beta(z)\end{aligned}$$

with the densities ψ_α and ψ_β as in definition 3.8. We can also express this as an expectation,

$$f_{\alpha\beta}(x) = E\{f(x-\alpha\boldsymbol{\xi}-\beta\boldsymbol{\eta})\}$$

with $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ random vectors whose elements are independent and uniformly distributed on $[\,-1/2,1/2\,]$. The gradient can be calculated from proposition 3.11. One has

$$\nabla f_{\alpha\beta}(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_n \Big(\int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_{i-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_{i+1} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_n \, \lambda_{\alpha\beta}(x,\xi,\eta)\Big)$$

where, with $z_i^{\alpha\beta}(\xi,\eta) := x_i - \alpha\xi_i - \beta\eta_i$,

$$\begin{aligned}\lambda_{\alpha\beta}(x,\xi,\eta) := \sum_{i=1}^{n} e_i \big[\, &f(z_1^{\alpha\beta}(\xi,\eta),\ldots,z_{i-1}^{\alpha\beta}(\xi,\eta),x_i+\alpha\xi_i+\frac{\beta}{2},z_{i+1}^{\alpha\beta}(\xi,\eta),\ldots,z_n^{\alpha\beta}(\xi,\eta))\\ &- f(z_1^{\alpha\beta}(\xi,\eta),\ldots,z_{i-1}^{\alpha\beta}(\xi,\eta),x_i+\alpha\xi_i-\frac{\beta}{2},z^{\alpha\beta}(\xi,\eta),\ldots,z_n^{\alpha\beta}(\xi,\eta))\,\big]\beta^{-1}.\end{aligned}$$

Again, $\boldsymbol{\lambda}_{\alpha\beta}(x,\boldsymbol{\xi},\boldsymbol{\eta})$ is an unbiased estimate of the gradient $\nabla f_{\alpha\beta}(x)$ with $\boldsymbol{\xi},\boldsymbol{\eta}$ random vectors whose elements are independent and uniformly distributed on $[\,-1/2,1/2\,]$. □

3.13. Remark. Let us also record an important relationship between the estimates of the gradients of averaged functions and stochastic gradients. We consider the following averaged functions:

$$f_\theta(x) = \frac{1}{\theta^n}\int_{\mathbb{R}^n} f(z)\psi(\,\frac{x-z}{\theta}\,)\,dz = \int_{\mathbb{R}^n} f(x-\theta z)\psi(z)\,dz,$$

with f locally integrable, ψ is a density function with compact support and such that $\nabla\psi$ is Lipschitz continuous. Then, the gradient of f_θ,

$$\nabla f_\theta = \frac{1}{\theta^{n+1}}\int_{\mathbb{R}^n} f(z)\nabla\psi(\,\frac{x-z}{\theta}\,)\,dz$$

is locally Lipschitz with constants proportional to $1/\theta^2$. The following random vector (cf. Gupal [**16**])

$$\boldsymbol{\lambda}_{\theta,\triangle}(x,\boldsymbol{\xi},\boldsymbol{\eta}) = \frac{1}{\triangle}\,[\,f(x-\theta\boldsymbol{\xi}+\triangle\boldsymbol{\eta}) - f(x-\theta\boldsymbol{\xi})\,]\boldsymbol{\eta}$$

is a stochastic quasi-gradient of f_θ at x (Ermoliev [**9**]), where $\boldsymbol{\xi}$ is distributed in accordance with the density function ψ, and $\boldsymbol{\eta}$ is a random vector whose elements are independent and uniformly distributed on $[\,-1,1\,]$. To see this, note that

$$\begin{aligned} E^{\xi,\eta}\{\boldsymbol{\lambda}_{\theta,\triangle}(x,\boldsymbol{\xi},\boldsymbol{\eta})\} &= E^\eta\,\frac{1}{\triangle}\,[\,f_\theta(x+\triangle\boldsymbol{\eta}) - f_\theta(x)\,]\boldsymbol{\eta} \\ &= \frac{2}{3}\,\nabla f_\theta(x) + \frac{\triangle}{2}\,O(x,\theta,\triangle) \end{aligned}$$

where $O(x,\theta,\triangle)$ is locally bounded.

Observe also that if $\boldsymbol{\xi}$ is distributed in accordance with the density function ψ_θ and $\boldsymbol{\eta}$ is a random vector whose elements are independent and uniformly distributed on $[\,-1,1\,]$, then

$$\boldsymbol{\lambda}_{\theta,\triangle}(x,\boldsymbol{\xi},\boldsymbol{\eta}) = \frac{1}{\triangle}\,[\,f(x-\boldsymbol{\xi}+\triangle\boldsymbol{\eta}) - f(x-\boldsymbol{\xi})\,]\boldsymbol{\eta}$$

is a quasi-gradient for the averaged function f_θ, i.e., it provides a, possibly biased, estimate of the gradient of f_θ as calculated in proposition 3.10. □

3.14. Remark. To complete this analysis of averaged functions, let us point out that the class of averaged functions that we have introduced is based on convolutions with mollifiers that are of the same nature as those used in theory of distributions. One could however have worked with a more general class and still obtain a convergence result similar to that of theorem 3.2; in fact, not just eh-convergence, but most of the results in this section. Let $\{\,\varphi_\theta : \mathbb{R}^n \to \mathbb{R}_+, \theta \in \mathbb{R}_+\,\}$ be a class of integrable functions such that $\int \varphi_\theta(z)\,dz = 1$.

Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ and the $\{\varphi_\theta\}$ are such that $f_\theta = f \star \varphi_\theta$ is well-defined (on $\mathbb{R}^n$) and that for all $\delta > 0$:

$$\lim_{\theta \downarrow 0} \int_{|z|>\delta} |f(z)| \varphi_\theta(x - z)\, dz = 0, \text{ uniformly in } x, \lim_{\theta \downarrow 0} \int_{|z| \le \delta} \varphi_\theta(z)\, dz = 1;$$

To see that the functions f_θ still eh-converge to f, note for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists V, a neighborhood of 0, such that $f(x - z) \ge \mathrm{cl}_e\, f(x) - \varepsilon$ for all $z \in V$ and that for $x^\theta \to x$ as $\theta \downarrow 0$, for all $\delta > 0$ and θ sufficiently small,

$$\begin{aligned} f_\theta(x^\theta) &= \int_{|z| \le \delta} f(z) \varphi_\theta(x^\theta - z)\, dz + \int_{|z|>\delta} f(z) \varphi_\theta(x^\theta - z)\, dz \\ &\ge (\mathrm{cl}_e\, f(x) - \varepsilon) \int_{|z|>\delta} \varphi_\theta(z)\, dz + \frac{\varepsilon}{2} \end{aligned}$$

and hence $\liminf_{\theta \downarrow 0} f_\theta(x^\theta) \ge \mathrm{cl}_e\, f(x)$ (after letting $\varepsilon \downarrow 0$). For example, let φ be the gaussian density function, i.e.,

$$\varphi(y) = \frac{1}{(2\pi)^{n/2}}\, e^{-|y|^2}.$$

Consider the following family of functions

$$f_\theta(x) = \frac{1}{\theta^n} \int_{\mathbb{R}^n} f(y) \varphi\left(\frac{x - y}{\theta}\right) dy, \quad \theta > 0.$$

Suppose that $|f(x)| \le \gamma_1 + \gamma_2 |x|^{\gamma_3}$ with γ_1, γ_2, γ_3 positive constants. Then, the functions f_θ eh-converge to f as $\theta \downarrow 0$ and each functions f_θ is analytic. One has

$$\nabla f_\theta(x) = \frac{1}{\theta^{n+2}} \int_{\mathbb{R}^n} f(x - y) \varphi\left(\frac{y}{\theta}\right) dy = \frac{1}{\theta} \int_{\mathbb{R}^n} [\, f(x - \theta z) - f(x)\,] z \varphi(z)\, dz;$$

passing differentiation under the integral sign is justified by the theory of tempered distributions, cf. Schwartz [**37**]. Thus the random vector $\boldsymbol{\lambda}_\theta(x, \boldsymbol{\xi})$, defined by

$$\lambda_\theta(x, \xi) = \frac{1}{\theta} [\, f(x - \theta\xi) - f(x)\,] \xi$$

with $\boldsymbol{\xi}$ a gaussian random variable (density φ), is an unbiased statistical estimator of $\nabla f_\theta(x)$. □

4. Mollifier subgradients

We are going to exploit the fact that averaged functions determine an epi-convergent family of approximating functions, and that rather explicit expressions can be obtained for their gradients, to define a new notion of subgradient based on a family of mollifiers. In the next section, these subgradients are used to design minimization procedures aimed, in particular, at the minimization of discontinuous functions.

4.1. Definition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally integrable and $\{f^\nu := f_{\theta^\nu}, \nu \in \mathbb{N}\}$ a sequence of averaged functions obtained from f by convolution with the sequence of mollifiers $\{\psi^\nu := \psi_{\theta^\nu} : \mathbb{R}^n \to \mathbb{R}, \nu \in \mathbb{N}\}$ where $\theta^\nu \downarrow 0$ as $\nu \to \infty$. Assume that the mollifiers are such that the averaged functions f^ν are smooth (of class C^1), as would be the case if the mollifiers ψ^ν are smooth. The set of the ψ-mollifier subgradients of f at x is by definition*

$$\partial_\psi f(x) := \operatorname{Lim\,sup}_{\nu\to\infty}\{\nabla f^\nu(x^\nu) \,|\, x^\nu \to x\},$$

i.e., the cluster points of all possible sequences $\{\nabla f^\nu(x^\nu)\}$ such that $x^\nu \to x$. The full (Ψ-) mollifier subgradient set is

$$\partial_\Psi f(x) := \bigcup\nolimits_\psi \partial_\psi f(x)$$

where ψ ranges over all possible sequences of mollifiers that generate smooth averaged functions.

The set $\partial_\psi f(x)$ of ψ- mollifier subgradients is closed, and clearly depends on the choice of the sequence $\{\psi^\nu\}$ that is used in its construction. The full mollifier subgradient set $\partial_\Psi f(x)$ is also convex and clearly does not depend on any particular choice of mollifiers. The sets $\partial_\psi f(x)$ and $\partial_\Psi f(x)$ are always nonempty if the function f is almost everywhere smooth and its gradient is locally bounded on the set where it exists (as in corollary 3.3 but applied here to ∇f).

4.2. Definition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally integrable and $\{f^\nu := f_{\theta^\nu}, \nu \in \mathbb{N}\}$ a sequence of averaged functions obtained from f by convolution with the sequence of mollifiers $\{\psi^\nu := \psi_{\theta^\nu} : \mathbb{R}^n \to \mathbb{R}, \nu \in \mathbb{N}\}$ where $\theta^\nu \downarrow 0$ as $\nu \to \infty$. Assume that the mollifiers are such that the averaged functions f^ν are smooth (of class C^1), as would be the case if the mollifiers ψ^ν are smooth ψ^ν are smooth (of class C^1). The ψ-mollifier subderivative of f at x in direction u is*

$$f'_\psi(x;u) := \mathrm{h}{-}\mathrm{ls}\,(f^\nu)'(x;u) = \sup_{\{x^\nu\to x\}} \limsup_{\nu\to\infty} (f^\nu)'(x^\nu;u)$$

where $(f^\nu)'(x;u)$ is the derivative of f^ν at x in direction u; sup *is taking with respect to all sequences $x^\nu \to x$. The full (Ψ-)mollifier subderivative of f at x in direction u is*

$$f'_\Psi(x;u) := \sup_\psi f'_\psi(x;u)$$

where ψ ranges over all possible sequences of mollifiers generating smooth averaged functions.

Henceforth, when referring to f we always assume that it is locally integrable and that $\{f^\nu\}$ is a sequence of smooth averaged functions obtained from f by convolution with a sequence of mollifiers $\{\,\psi^\nu,\, \nu \in \mathbb{N}\,\}$.

4.3. Proposition. *The ψ-mollifier subgradient mapping $x \mapsto \partial_\psi f(x)$ is outer semicontinuous (closed graph) and $f'_\psi(x;\cdot)$ is upper semicontinuous. Also*

$$f'_\psi(x;u) \geq \sup\{\,\langle g,u\rangle \,|\, g \in \partial_\psi f(x)\,\},$$
$$f'_\Psi(x;u) \geq \sup\{\,\langle g,u\rangle \,|\, g \in \partial_\Psi f(x)\,\}.$$

Proof. Follows directly from the definitions; $f f'_\psi(x;\cdot)$ is a hypo-limit. □

4.4. Proposition. *The function $u \mapsto f'_\psi(x;u)$ is sublinear, i.e., $f'_\psi(x;\cdot)$ is convex and positively homogeneous. The set-valued mapping*

$$x \mapsto G_\psi(x) := \{\, g \in \mathbb{R}^n \,|\, \langle g,u\rangle \leq f'_\psi(x;u),\, \forall\, u \in \mathbb{R}^n\,\}$$

is closed-, convex-valued.

Proof. Since the functions f^ν are smooth, one has

$$(f^\nu)'(x^\nu;u_1+u_2) = (f^\nu)'(x^\nu;u_1) + (f^\nu)'(x^\nu;u_2).$$

Taking *limsup* on both sides over all sequences $x^\nu \to x$ yields

$$f'_\psi(x;u_1+u_2) \leq f'_\psi(x;u_1) + f'_\psi(x;u_2).$$

Similarly, the positive homogeneity of $f'_\psi(x;\cdot)$ follows from the linearity of the derivatives of the functions $(f^\nu)'(x;\cdot)$. The assertions about the set-valued mapping G_ψ follow directly from the sublinearity of $f'_\psi(x;\cdot)$. □

4.5. Proposition. *One always has*

$$\operatorname{con}\partial_\psi f(x) \subset G_\psi(x) := \{\, g \in \mathbb{R}^n \,|\, \langle g,u\rangle \leq f'_\psi(x;u),\, \forall\, u \in \mathbb{R}^n\,\}$$

where con *denotes the convex hull. If $\partial_\psi f(x)$ is bounded then $\operatorname{con}\partial_\psi f(x) = G_\psi(x)$.*

Proof. We begin with the inclusion. To any $g \in \partial_\psi f(x)$, there corresponding a subsequence $\{\nu_k\} \subset \{\nu\}$ and $x^k \to x$ such that $\nabla f^{\nu_k}(x^k) \to g$. Since $(f^{\nu_k})'(x^k;u) = \langle \nabla f^{\nu_k}(x^k), u\rangle$, it follows that

$$\langle g,u\rangle = \lim_{k\to\infty}\langle \nabla f^{\nu_k}(x^k),u\rangle = \lim_{k\to\infty}(f^{\nu_k}(x^k;u) \leq f'_\psi(x;u).$$

Thus $\partial_\psi f(x) \subset G_\psi(x)$ and the convexity of $G_\psi(x)$ then yields $\operatorname{con} \partial_\psi f(x) \subset G_\psi(x)$.

Suppose now that $\partial_\psi f(x)$ is bounded. If $h \in G_\psi(x) \setminus \operatorname{con} \partial_\psi f(x)$, i.e., $G_\psi(x) \not\subset \operatorname{con} \partial_\psi f(x)$, then by the separation theorem for convex sets, there exists $\bar{u}$ such that $\langle h, \bar{u}\rangle > \langle g, \bar{u}\rangle$ for all $g \in \operatorname{con} \partial_\psi f(x)$. But $f'_\psi(x;\bar{u}) \geq \langle h, \bar{u}\rangle$ and, passing to a subsequence whenever necessary, there exists $x^\nu \to x$ so that

$$\nabla f^\nu(x^\nu) \;\longrightarrow\; g \in \partial_\psi f(x)$$

and

$$(f^\nu)'(x^\nu;\bar{u}) = \langle \nabla f^\nu(x^\nu), \bar{u}\rangle \;\longrightarrow\; f'_\psi(x;\bar{u}).$$

Thus, we would have that

$$f'_\psi(x;\bar{u}) = \langle g, \bar{u}\rangle \geq \langle h, \bar{u}\rangle > \langle g, \bar{u}\rangle,$$

clearly contradicting the existence of such a h. □

4.6. Remark. The approach laid out here could be used to define subdifferentials of higher order. For example, if the mollifiers ψ_{θ^ν} are of class C^2, then the resulting averaged function f^ν are also twice continuously differentiable. With $\nabla^2 f^\nu(x)$ the hessian of f^ν at x, we could define the second order ψ-mollifier subhessian of f at x as

$$\partial^2_\psi f(x) := \operatorname{Lim\,sup}_{\nu\to\infty}\{\, \nabla^2 f^\nu(x^\nu) \,|\, x^\nu \to x \,\},$$

i.e., the cluster points of all possible sequences $\{\nabla^2 f^\nu(x^\nu)\}$ of matrices with $x^\nu \to x$. The function

$$f''_\psi(x;H) := \limsup_{x^\nu\to x}\langle \nabla^2 f^\nu(x^\nu), H\rangle = \limsup_{x^\nu\to x} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i \partial x_j} f^\nu(x^\nu) h_{ij}$$

could be called the second order ψ-mollifier subderivative of f in direction H. The mapping $x \mapsto \partial^2_\psi f(x)$ is closed, the function $f''_\psi(x;\cdot)$ is upper semicontinuous and one has

$$\operatorname{con} \partial^2_\psi f(x) = \{\, H \in \mathbb{R}^{n^2} \,|\, Hu \leq f''_\psi(x;U),\ \forall\, U \in \mathbb{R}^{n^2} \,\}.$$

The next theorem justifies a minimization approach based on mollifier subgradients.

4.7. Theorem. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is strongly lower semicontinuous and locally integrable. Then, for any sequence $\{\psi^\nu\}$ of smooth mollifiers, one has*

$$0 \in \partial_\psi f(x) \;\textit{ whenever } x \textit{ is a local minimizer of } f.$$

Proof. Let x be a local minimizer of f. For V a compact neighborhood of x sufficiently small, define

$$\varphi : V \to \mathbb{R} \text{ with } \varphi(z) = f(z) + |z - x|^2.$$

The function φ achieves its global minimum (on V) at x. Consider also the averaged functions

$$\varphi^\nu(z) = \int_{\mathbb{R}^n} \varphi(y-z)\psi^\nu(y)\,dy = f^\nu(z) + \beta^\nu(x,z)$$

where $\beta^\nu(x,z) = \int |y-z-x|^2\psi^\nu(y)\,dy$. From theorem 3.10, it follows that the function φ^ν are continuously differentiable and theorem 3.7 implies that they epi-converge to φ on V. Suppose φ^ν achieves its minimum at some point $z^\nu \in V$. It follows from theorem 3.5 that $z^\nu \to x$, and thus

$$\nabla\varphi^\nu(z^\nu) = \nabla f^\nu(z^\nu) + \nabla\beta^\nu(x,z^\nu) = 0.$$

Hence

$$\nabla f^\nu(z^\nu) = -\nabla\beta^\nu(x,z^\nu) \;\longrightarrow\; 0 \text{ as } \nu \to \infty,$$

and consequently $0 \in \partial_\psi f(x)$. □

In the remainder of this section we explore the relationship between mollifier subgradient and some other subgradients notions.

For function $f : \mathbb{R}^n \to \mathbb{R}$ continuous on a neighborhood V of x, Warga [**42-44**] defines subgradients of f at x as follows: Let $\{f^k, k \in \mathbb{N}\}$ be a sequence of smooth functions converging uniformly to f on V, we shall refer to

$$\partial_W f(x) = \bigcap_{j=1}^{\infty} \bigcap_{\delta>0} \mathrm{cl}\,[\bigcup_{k\geq j, |x-y|\leq\delta} \nabla f^k(y)\,]$$

as the set of Warga-subgradients of f at x (cl denotes closure).

4.8. Proposition. *For $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on V a neighborhood of x, and $\{f^k, k \in \mathbb{N}\}$ a sequence of smooth functions converging uniformly to f on V, then*

$$\partial_W f(x) = \mathrm{Lim\,sup}_{k\to\infty}\{\,\nabla f^k(x^k) \,|\, \forall x^k \to x\,\}.$$

Consequently, when f is continuous, $\partial_W f(x)$ coincides with $\partial_\psi f(x)$ if in the construction of $\partial_W f(x)$ the f^k are averaged functions generated by the sequence of smooth mollifiers $\{\psi^k\}$.

Proof. Let

$$D(x) = \mathrm{Lim\,sup}_{k\to\infty}\{\,\nabla f^k(x^k) \,|\, \forall x^k \to x\,\}.$$

Let us first show that $D(x) \subset \partial_W f(x)$. Let $g \in D(x)$ be such that, passing to a subsequence if necessary, $g = \lim_k \nabla f^k(x^k)$ for some specific sequence $x^k \to x$. We have to show that for all j and $\delta > 0$,

$$g \in G_{j,\delta}(x) := \mathrm{cl}\,[\bigcup_{k\geq j, |x-y|\leq\delta} \nabla f^k(y)\,].$$

Obviously, if $k \geq j$ and $|x^k - x| \leq \delta$, then

$$\nabla f^k(x^k) \in G_{j,\delta}(x).$$

Since $G_{j,\delta}(x)$ is closed, each cluster point of the sequence $\{\nabla f^k(x^k)\}$ belongs to $G_{j,\delta}(x)$. Hence, $g \in \partial_W f(x)$ and $D(x) \subset \partial_W f(x)$.

To prove the converse inclusion, one needs to show that for each point g in $\partial_W f(x)$ one can find a sequence $x^k \to x$ such that $\nabla f^k(x^k) \to g$. By definition of ∂_W for all j and $\delta > 0$, $g \in G_{j,\delta}(x)$. Let us choose a sequence $\delta_j \downarrow 0$ as $j \to \infty$. Since $g \in G_{j,\delta_j}(x)$ for all j,

$$g \in \mathrm{cl}\{\, \nabla f^k(y) \;:\; k \geq j, |y - x| \leq \delta_j \,\}.$$

Thus in this set, there exists an element $g^j = \nabla f^{k_j}(y^j)$ such that $|g^j - g| < 1/j$. Clearly $y^j \to x$, $k_j \to \infty$ and $g^j \to g$, so that $g \in D(x)$ and $\partial_W f(x) \subset D(x)$.

The equality between the Warga- and the ψ-mollifier subgradient sets then follow from the formula we just proved, and the definition of ψ-mollifier subgradients. □

In variational analysis, the *Clarke subderivative* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$(d_C f)(x; u) = \limsup_{y \to x, \lambda \downarrow 0} \frac{1}{\lambda} [\, f(y + \lambda u) - f(y) \,]$$

with the lim sup calculated with respect to all sequences $y \to x$, $\lambda \downarrow 0$. The set of *Clarke subgradients* is

$$\partial_C f(x) = \{\, g \in \mathbb{R}^n \mid \langle g, u \rangle \leq d_C f(x; u),\ \forall\, u \in \mathbb{R}^n \,\}.$$

This notion was proposed by Clarke [**6**] for locally Lipschitz continuous functions; for just lower semicontinuous functions this notion needs further adjustments, consult Rockafellar [**31**].

4.9. Proposition. *For $f : \mathbb{R}^n \to \mathbb{R}$ locally integrable, one has $f'_\psi(x; \cdot) \leq d_C f(x; \cdot)$. If f is also continuous, then $f'_\psi(x; \cdot) = d_C f(x; \cdot)$.*

Proof. By definition of $d_C f(x; u)$ it follows that for an arbitrary $\varepsilon > 0$, there exist δ_1, δ_2 such that whenever $|y - x| < \delta_1$, and $\lambda \in (0, \delta_2)$,

$$\frac{1}{\lambda} [\, f(y + \lambda u) - f(y) \,] < d_C f(x; u) + \varepsilon.$$

Let f^ν be the averaged function obtained as the convolution of f and the mollifier ψ^ν. Consider the finite differences

$$\triangle_\nu(y, u, \lambda) := \frac{1}{\lambda} [\, f^\nu(y + \lambda u) - f^\nu(y) \,] = \int_{\mathbb{R}^n} \frac{1}{\lambda} [\, f(y - z + \lambda u) - f(y - z) \,] \psi^\nu(z)\, dz.$$

If $|y-x| < \delta_1/2$, $\lambda < \delta_2/2$ and $|z| \le \delta_1/2$, then

$$\triangle_\nu(y,u,\lambda) \le (d_C f(x;u) + \varepsilon) \int_{|z|\le\delta_1/2} \psi^\nu(z)\, dz.$$

Thus for y close enough to x,

$$(f^\nu)'(y;u) = \lim_{\lambda \downarrow 0} \triangle_\nu(y,u,\lambda) \le (d_C f(x;u) + \varepsilon) \int_{|z|\le\delta_1/2} \psi^\nu(z)\, dz$$

from which, after letting $\varepsilon \downarrow 0$, it follows that $f'_\psi(x;u) \le d_C f(x;u)$.

We next set out to prove the reverse inequality, assuming that f is continuous. Let $x^\nu \to x$ and $\lambda_\nu \downarrow 0$ be such that

$$d_C f(x;u) = \lim_{\nu\to\infty} \frac{1}{\lambda_\nu} \left[f(x^\nu + \lambda_\nu u) - f(x^\nu) \right].$$

From corollary 3.3, we know that when f is continuous, the averaged functions f^ν converge uniformly to f on some neighborhood, say V, of x. Thus, with $\varepsilon_\nu = \lambda_\nu/\nu$, one can always find k_ν such that

$$\sup_{y\in V} |f(y) - f^{k_\nu}(y)| < \varepsilon_\nu.$$

Now from the mean value theorem follows the existence of $y^\nu := x^\nu + \tau_\nu u$, $\tau_\nu \in [\,0, \lambda_\nu\,]$ such that

$$\frac{1}{\lambda_\nu} \left[f^{k_\nu}(x^\nu + \lambda_\nu u) - f^{k_\nu}(x^\nu) \right] = (f^{k_\nu})'(y^\nu; u).$$

Thus for ν sufficiently large, with $x^\nu \in V$ and $x^\nu + \lambda_\mu u \in V$, one has

$$\begin{aligned}
&[\, f(x^\nu + \lambda_\nu u) - f(x^\nu)\,] \\
&\quad = [\, f^{k_\nu}(x^\nu + \lambda_\nu u) - f^{k_\nu}(x^\nu)\,] + [\, f(x^\nu + \lambda_\nu u) - f^{k_\nu}(x^\nu + \lambda_\nu u)\,] - [\, f(x^\nu) - f^{k_\nu}(x^\nu)\,] \\
&\quad \le \lambda_\nu((f^{k_\nu})'(y^\nu; u) + 2/\nu).
\end{aligned}$$

Taking lim sup with respect to ν yields

$$d_C f(x;u) \le \limsup_{\nu\to\infty} (f^{k_\nu})'(y^\nu;u) \le f'_\psi(x;u),$$

which completes the proof. □

4.10. Theorem. *If $f : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous and locally integrable, then*

$$\operatorname{con} \partial_\psi f(x) \subset \partial_\Psi f(x) \subset \partial_C f(x).$$

If, in addition f is locally Lipschitz continuous, then

$$\operatorname{con} \partial_\psi f(x) = \partial_\Psi f(x) = \partial_C f(x).$$

Proof. The first inclusion follows from the relationship between $\partial_\psi f(x)$ and $\partial_\Psi f(x)$ (with this last set convex), and the second inclusion follows from the preceding proposition. If f is locally Lipschitz, then also the averaged functions f^ν are locally Lipschitz and $\partial_\psi f(x)$ is bounded. Equality then follows from propositions 4.5 and 4.9. □

4.11. Corollary (Gupal [**15**]). *If f is locally Lipschitz continuous, then for all $\alpha_\nu \downarrow 0$ and $x^\nu \to x$, all clusters points of the sequences $\{\nabla f_{\alpha_\nu}(x^\nu)$ belong to $\partial_C f(x)$.*

4.12. Remark. For the sake of completeness, let us also record the fact that for convex functions, $f : \mathbb{R}^n \to \mathbb{R}$, we actually have that

$$\partial_\psi f(x) = \partial_C f(x) = \{\, g \in \mathbb{R}^n \mid f(z) \geq f(x) + \langle g, z - x\rangle,\ \forall\, z \in \mathbb{R}^n \,\}.$$

For convex functions, as is well known, the set of gradients can be characterized in terms of the expression on the right, cf. [**32**], for example. In view of the preceding theorem, it will thus be sufficient to show that if $g \in \partial_C f(x)$, then g is also included in $\partial_\psi f(x)$. Let us consider the function

$$\varphi(y) = f(y) + |y - x|^2 - f(x) - \langle g, y - z\rangle.$$

The function $\varphi \geq 0$ and attains its minimum ($= 0$) at x; due to the strict convexity of φ, x is a unique minimizer of φ. Let

$$\begin{aligned}\varphi^\nu(y) &= \int_{\mathbb{R}^n} \varphi(y - z)\psi^\nu(z)\, dz \\ &= f^\nu(y) + \beta^\nu(x, y) - f(x) - \langle g, y - x\rangle - \int_{\mathbb{R}^n} \langle g, z\rangle \psi^\nu(z)\, dz\end{aligned}$$

be the averaged functions associated with φ by convolution with the ψ^ν; here $\beta^\nu(x, y) = \int |y - z - x|^2 \psi^\nu(z)\, dz$. The averaged functions φ^ν uniformly converge of φ on some neighborhood V of x (corollary 3.3). Due to the strict convexity of φ, for ν sufficiently large, the averaged functions ψ^ν have a (global) minimizer on V, say y^ν. Moreover, $y^\nu \to x$, since x is a unique minimizer of $\varphi = \mathrm{e}{-}\mathrm{lm}\,\varphi^\nu$ (theorem 3.7). The averaged functions φ^ν, f^ν and $\beta^\nu(x, \cdot)$ are smooth (theorem 3.10), and thus

$$\nabla\varphi^\nu(y^\nu) = \nabla f^\nu(y^\nu) + \nabla_y \beta^\nu(x, y^\nu) - g,$$

$$\nabla_y \beta^\nu(x, y) = \int_{\mathbb{R}^n} \nabla_y |y - z - x|^2 \psi^\nu(z)\, dz = 2(y - x) - 2\gamma^\nu,$$

$$\gamma^\nu = \int_{\mathbb{R}^n} z\psi^\nu(z)\, dz.$$

From the conditions imposed on the mollifiers ψ^ν, it follows that $\gamma^\nu \to 0$, and hence $\nabla_y \beta^\nu(x, y^\nu) \to 0$, and

$$\nabla f^\nu(y^\nu) = g - \nabla_y \beta^\nu(x, y^\nu) \ \longrightarrow\ g \ \text{ as } \nu \to \infty$$

which means that $g \in \partial_\psi f(x)$, as claimed. □

5. Numerical procedures

Let us consider the problem of minimizing a strongly lower semicontinuous φ on X, a compact subset of $\mathbb{R}^n$. Let

$$\mathbb{1}_X(x) = \begin{cases} 1 & \text{if } x \in X; \\ 0 & \text{if } x \notin X. \end{cases}$$

Then, instead of the original problem, one could work with one of the following unconstrained problems involving discontinuous penalty functions:

$$\text{minimize } f(x) := \varphi(x)\mathbb{1}_X(x) + \gamma(1 - \mathbb{1}_X(x))$$

or

$$\text{minimize } f(x) := \varphi(x)\mathbb{1}_X(x) + \gamma(1 - \mathbb{1}_X(x))d(x, X)$$

where $d(x, X) = \min\{\, |x - y| : y \in X \,\}$ and γ is sufficiently large.

If the function φ is bounded on X and $\gamma > \sup\{|\varphi(x)| : x \in X\}$, all local minima of φ on X are also locally minima of the function f.

Assuming that f is also strongly lower semicontinuous, in view of theorems 3.7 and 3.10, on can always find a sequence of smooth averaged functions f^ν (generated by mollifiers $\{\psi^\nu\}$) that epi-converge to f, and by theorem 4.7, the condition $0 \in \partial_\psi f(x^*)$ is necessary for a point x^* to be a local minimizer of f

Let us now consider some optimization procedures for f making use of the approximating averaged function f^ν.

5.1. Method. Suppose a sequence $\{x^\nu\}$ of global minimizers of f^ν can be calculated. Then, according to theorem 3.5 any cluster point of such a sequence is a (global) minimizer of f. □

However finding global minimizers of the f^ν could be quite complicated. This leads us to consider the next method.

5.2. Method. Here a sequence of approximating solutions $\{x^\nu\}$ is build in accordance with the following rule. Each function f^ν is minimized —initiating the procedure at $x^{\nu-1}$— until a point x^ν is found such that $|\nabla f^\nu(x^\nu)| \le \varepsilon_\nu$ where $\varepsilon_\nu \downarrow 0$; the starting point x^0 is chosen arbitrarily. In this method, if $\bar{x}$ is a cluster point of the sequence $\{x^\nu\}$, then by definition of $\partial_\psi f(\bar{x})$, passing to a subsequence if necessary,

$$\lim_{\nu\to\infty} \nabla f^\nu(x^\nu) = 0 \in \partial_\psi f(\bar{x}).$$

Moreover, this would also mean that $0 \in \partial_C f(\bar{x})$ (theorem 4.10), i.e., $d_C f(x; u) \ge 0$ for all $u \in \mathbb{R}^n$. □

This approach requires estimates of $|\nabla f^\nu(x^\nu)|$ during the iteration process. In general, this could be computationally expensive involving the calculation of multidimensional integrals. One can however, produce these estimates in parallel with the optimization process by a well-known averaging procedure (cf. Ermoliev [**8**]): let

x^0, z^0 be chosen arbitrarily in $\mathbb{R}^n$;
$x^{k+1} = x^k - \rho_k z^k, \quad k = 0, 1, \ldots;$
$z^{k+1} = z^k - \tau_k(z^k - \lambda_k(x^k)), \quad k = 0, 1, \ldots;$

where x^k approximates $\operatorname{argmin} f^\nu$, z^k are averaged estimates of $\nabla f^\nu(x^k)$, $\lambda_\nu(x^k)$ are stochastic (finite-difference unbiased) estimates for $\nabla f^\nu(x^k)$ such that their mathematical expectation $E\{\lambda_\nu(x^k)\} = \nabla f^\nu(x^k)$ (see the observations that follow proposition 3.11), $\rho_k \geq 0$ and $\tau_k > 0$ are sequences such that

$$\sum_{k=0}^{\infty} \rho_k = \infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < \infty, \quad \lim_{k\to\infty} \rho_k/\tau_k = 0.$$

5.3. Proposition (Ermoliev [**8**, theorem V.8])**.** *If the sequences* $\{x^k\}$, $\{z^k\}$ *are almost surely bounded, then almost surely*

$$\lim_{k\to\infty} |z^k - \nabla f^\nu(x^k)| = 0, \text{ and } x^k \longrightarrow \{x \,|\, \nabla f^\nu(x) = 0\}.$$

Thus in method 5.2, we can proceed with the minimization of each f^ν until the estimate z^k of the gradient of $\nabla f^\nu(x^k)$ satisfy the condition $|z^k| \leq \varepsilon_\nu$.

5.4. Method. A sequence of approximate solutions x^ν is generated by the following rule

$x^0 \in \mathbb{R}^n$ is chosen arbitrarily;
$x^{\nu+1} = x^\nu - \rho_\nu \lambda_\nu(x^\nu), \quad \nu = 0, 1. \ldots$

where $\boldsymbol{\lambda}_\nu(x^\nu)$ is a stochastic (finite-difference unbiased) estimator for $\nabla f^\nu(x^\nu)$ with expectation $E\{\boldsymbol{\lambda}_\nu(x^\nu)\} = \nabla f^\nu(x^\nu)$ (see the observations following proposition 3.11 and remark 3.12), $\rho_\nu \geq 0$ is a deterministic sequence of multipliers. □

This method combines ideas from the method of stochastic quasi-gradients with those of dynamic nonstationary optimization techniques, see Ermoliev and Nurminski [**10**] and Gaivoronski [**13**]. The following theorem is an example of the possible convergence results.

5.5. Theorem (Gupal and Norkin [**17**])**.** *Suppose the gradient estimates are those in example 3.12, i.e.,* $\lambda_\nu(x) = \lambda_{\alpha_\nu \alpha_\nu}(x, \xi, \eta)$, *the sequence* $\{x^\nu\}$ *belongs to some compact set and* $\rho_\nu \geq 0$, α_ν *satisfy the conditions*

$$\sum_{\nu=1}^{\infty} \rho_\nu = \infty, \quad \sum_{\nu=1}^{\infty} \left(\frac{\rho_\nu}{\alpha_\nu^2}\right)^2 < \infty, \quad \lim_{\nu\to\infty} \alpha_\nu = \lim_{\nu\to\infty} \frac{\alpha_\nu - \alpha_{\nu+1}}{\alpha_\nu \rho_\nu} = 0.$$

Then, almost surely, the sequence $\{x^\nu\}$ *admits a cluster point* x^* *such that* $0 \in \partial_\psi f(x^*)$.

5.6. Example. Let us consider the minimization of a probability function:

$$f(x) = \mathbf{P}\,[\, g(x, \omega) \geq 0\,].$$

We can express f as a mathematical expectation

$$f(x) = \int_{\Omega} \mathbb{1}_{\{g(x,\omega)\geq 0\}}(\omega)\, P(d\omega).$$

Since the function $\mathbb{1}_{\{.\}}$ is discontinuous, the function f will in general, not be differentiable. To estimate $f(x)$ and its "gradient," Tamm [**41**] and Lepp [**21**] proposed the use of Parzen-Rosenblatt kernel-type estimates [**29**], [**35**]:

$$f_{\varepsilon}(x) = \frac{1}{\varepsilon} \int_{\Omega} P(d\omega) \int_{-\infty}^{0} d\tau\, \psi(\frac{\tau + g(x,\omega)}{\varepsilon}),$$

$$\nabla f_{\varepsilon}(x) = \frac{1}{\varepsilon} \int_{\Omega} \psi(\frac{g(x,\omega)}{\varepsilon}) \nabla_x g(x,\omega)\, P(d\omega)$$

where ψ is some symmetric density function on $[-\infty, \infty]$; more recently Marti [**23**] has suggested a similar approach to deal with reliability constraints in structural optimization. The funcion f_{ε} can also be written as

$$f_{\varepsilon}(x) = \int_{\Omega} \psi_{\varepsilon}(g(x,\omega)\, P(d\omega),$$

where

$$\begin{aligned}\psi_{\varepsilon}(t) &= \frac{1}{\varepsilon} \int_{-\infty}^{t} \psi(\frac{\tau}{\varepsilon})\, d\tau \\ &= \frac{1}{\varepsilon} \int_{-\infty}^{\infty} \mathbb{1}_{\{t-\tau\geq 0\}}(\tau)\psi(\frac{\tau}{\varepsilon})\, d\tau = \frac{1}{\varepsilon} \int_{-\infty}^{\infty} \mathbb{1}_{\{t+\tau\geq 0\}}(\tau)\psi(-\frac{\tau}{\varepsilon})\, d\tau\end{aligned}$$

Thus ψ_{ε} is an averaged function (with base function $\mathbb{1}_{\{.\geq 0\}}$). Instead of the original function f, we have a sequence of approximating function f_{ε} constructed (indirectly) by means of averaged functions. Tamm [**41**] in the differentiable case, and Norkin [**28**] in the nondifferentiable (but continuous) case, provided conditions under which f_{ε} converges uniformly to f, and they proposed methods, similar to method 5.2., to minimize f making use of the approximating functions f_{ε}. Lepp [**22**] and Roenko [**34**] analyzed stochastic iterative methods, like method 5.4, for the minimization f when it is differentiable, using statistical estimates for $\nabla f_{\varepsilon}(x)$. □

References

[1] Z. Artstein and R. J-B. Wets, Stability of stochastic programs with possibly discontinuous objective functions, Manuascript, Weizmann Institut, Rehovot, 1992.

[2] H. Attouch and R. J-B. Wets, Approximation and convergence in nonlinear optimization, in Nonlinear Programming 4, O. Mangasarian, R. Meyer and S. Robinson, eds., Academic Press, New York, 1981, pp. 367–394.

[3] J-P. Aubin, Lipschitz behavior of solutions to convex minimization problems, Mathematics of Operations Research, 8 (1984), pp. 87–111.

[4] J-P. Aubin and H. Frankowska, Set-Valued Analysis, Birkhaüser, Basel, 1990.

[5] A. Bensoussan and J-L. Lions, Control Impulsional et Inequations Quasi-Variationelles, Bordas, Paris, 1982.

[6] F. H. Clarke, Optimization and Nonsmooth Analysis, J. Wiley, New York, 1983.

[7] V. F. Demyanov and A. M. Rubinov, Foundations of Nonsmooth Analysis and Quasi-Differential Calculus, Nauka, Moscow, 1990, (in Russian).

[8] Y. M. Ermoliev, Methods of Stochastic Programming, Nauka, Moscow, 1976, (in Russian).

[9] Y. M. Ermoliev and A. A. Gaivoronski, On optimization of discontinuous systems, Working paper WP-91-41, IIASA, Laxenburg, 1991.

[10] Y. M. Ermoliev and E. A. Nurminski, Limit extremal problems, Kibernetika, 1973.

[11] H. Frankowska, Inclusions adjointes associées aux trajectoires minimales d'inclusions différentielles, Comptes Rendus de l'Académie des Sciences de Paris, 297 (1983), pp. 461–464.

[12] ———, The first order necessary conditions for nonsmooth variational and control problems, SIAM J. on Control and Optimization, 22 (1984), pp. 1–12.

[13] A. A. Gaivoronski, On nonstationaty stochastic optimization problems, Kibernetika, 1978, (English translation: Cybernetics, vol. 14, no. 4).

[14] W. B. Gong and Y. C. Ho, Smoothed (conditional) perturbations analysis of discrete event dynamic systems, IEEE Transactions on Automatic Control, 32 (1987), pp. 856–866.

[15] A. M. Gupal, On a method for the minimization of almost differentiable functions, Kibernetika,, 1977, (English translation: Cybernetics, vol. 13, no. 1).

[16] ———, Stochastic Methods for Solving Stochastic Extremal Problems, Naukova Dumka, Kiev, 1979, (in Russian).

[17] A. M. Gupal and V. I. Norkin, An algorrithm for the minimization of discontinuous functions, Kibernetika, 1977, (English translation: Cybernetics, vol. 13, no. 2).

[18] A. D. Ioffe, Nonsmooth analysis: differential calculus of nondifferentiable mappings, Transactions of the American Mathematical Society, 266 (1981), pp. 1–56.

[19] V. Y. Katkovnik, Linear Estimators and Stochastic Optimization Problems, Nauka, Moscow, 1976, (in Russian).

[20] A. N. Kolmogorov, Selected Works, Mathematics and Mechanics, Nauka, Moscow, 1985, (in Russian).

[21] R. Lepp, The maximization of a probability function over simple sets, Izvestia Akademii Nauk Estonskoy SSR. Physics and Mathematics, 28 (1979), pp. no. 4, 303–309, (in Russian).

[22] ———, Stochastic approximation type algorithm for the maximization of a probability function, Izvestia Akademii Nauk Estonskoy SSR. Physics and Mathematics, 32 (1983), pp. no. 2, 150–156, (in Russian).

[23] K. Marti, Stochastic optimization methods in structural mechanics, Z. für Angewandte Mathematik und Mechanik, 70 (1990), pp. T742–T745.

[24] D. Q. Mayne and E. Polak, Nondifferential optimization via adaptive smoothing, J. of Optimization Theory and Applications, 43 (1984), pp. 19–30.

[25] P. Michel and J-P. Penot, Calcul sous-differentiel pour des fonctions lipschitziennes et non lipschitziennes, Comptes Rendus de l'Académie des Sciences de Paris, 298 (1984), pp. 269–272..

[26] B. S. Mordukhovich, Approximation Methods in Problems of Optimization and Control, Nauka, Moscow, 1988, (in Russian).

[27] N. D. Nikolaeva, On an algorithm for solving convex programming problems, Ekonomika i Matematicheskie Metody, 10 (1974), pp. 941–946, (in Russian).

[28] V. Norkin, Optimization of probabilities, Prepint 89-9, Glushkow Institut of Cybernetics, Kiev, 1989.

[29] E. Parzen, On estimation of a probability density function and the mode, Annals of Mathematical Statistics, 33 (1962), pp. 1065–1076.

[30] B. Polyak, Nonlinear Programming methods in the presence of noise, Mathematical Programming, 14 (1978), pp. 87–97.

[31] R. T. Rockafellar, The Theory of Subgradients and its Application to Problems of Optimization: Convex and Nonconvex Functions, Helderman Verlag, Berlin, 1981.

[32] ———, Generalized subgradients in mathematical programming, in Mathematical Programming: The State of the Art 1982, A. Bachem, M. Grötschel and B. Korte, eds., Springer Verlag, Berlin, 1983, pp. 368–380.

[33] R. T. Rockafellar and R. J-B. Wets, Variational systems, an introduction., in Multifunctions and Integrands., G. Salinetti, ed., Springer-Verlag Lecture Notes in Mathematics 1091, Berlin, 1984, pp. 1–54.

[34] N. V. Roenko, Stochastic programming problems with integral functionals from multivalued mappings, Abstract of Ph.D. thesis, Glushkov Institute of Cybernetics, Kiev, 1983.

[35] M. Rosenblatt, Remarks on some nonparametric estimates of a density functions, Annals of Mathematical Statistics, 27 (1966), pp. 832–835.

[36] R. Rubinstein, How to optimize discrete-event systems from a single path by the score function method, Annals of Operations Research, 27 (1991), pp. 175–212.

[37] L. Schwartz, Théorie des Distributions, Hermann, Paris, 1966.

[38] S. L. Sobolev, Some Applications of Functional Analysis in Mathematical Physics, Nauka, Moscow, 1988, (3rd edition, in Russian).

[39] V. A. Steklov, Sur les expressions asymptotiques de certaines fonctions définies par les équations différentielles du second ordre et leurs applications au problème du dévelopement d'une fonction arbitraire en séries procédant suivant les diverses fonctions, Communication of Charkov Mathematical Society, Serie 2, 10 (1907), pp. 97–199, (in Russian).

[40] ———, Main Problems of Mathematical Physics, Nauka, Moscow, 1983, (in Russian, 1st edition: Petrograd, 1922).

[41] E. Tamm, On a probability function optimization, Izvestia Akademii Nauk Estonskoy SSR. Physics and Mathematics, 28 (1979), pp. no. 1, 17–24., (in Russian).

[42] J. Warga, Necessary conditions without differentiability assumptions in optimal control, J. Differential Equations, 15 (1975), pp. 41–61.

[43] ———, Derivative containers , inverse functions and controllability, in Calculus of Variations and Control Theory, D. Russell, ed., Academic Press, New York, 1976, pp. 13–46.

[44] ———, Fat homeomorphisms and unbounde derivative containers, J. Mathematical Analysis and Applications, 81 (1981), pp. 545–560.