Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

1133

Krzysztof C. Kiwiel

Methods of Descent for Nondifferentiable Optimization

Springer-Verlag
Berlin Heidelberg New York Tokyo

Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

1133

Krzysztof C. Kiwiel

Methods of Descent for Nondifferentiable Optimization

Springer-Verlag
Berlin Heidelberg New York Tokyo

Author

Krzysztof C. Kiwiel
Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland

Mathematics Subject Classification: 49-02, 49D37, 65-02, 65K05, 90-02, 90C30

ISBN 3-540-15642-9 Springer-Verlag Berlin Heidelberg New York Tokyo
ISBN 0-387-15642-9 Springer-Verlag New York Heidelberg Berlin Tokyo

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2146/3140-543210

Lecture Notes in Mathematics

For information about Vols. 1–925, please contact your bookseller or Springer-Verlag.

Vol. 926: Geometric Techniques in Gauge Theories. Proceedings, 1981. Edited by R. Martini and E.M.de Jager. IX, 219 pages. 1982.

Vol. 927: Y. Z. Flicker, The Trace Formula and Base Change for GL (3). XII, 204 pages. 1982.

Vol. 928: Probability Measures on Groups. Proceedings 1981. Edited by H. Heyer. X, 477 pages. 1982.

Vol. 929: Ecole d'Eté de Probabilités de Saint-Flour X – 1980. Proceedings, 1980. Edited by P.L. Hennequin. X, 313 pages. 1982.

Vol. 930: P. Berthelot, L. Breen, et W. Messing, Théorie de Dieudonné Cristalline II. XI, 261 pages. 1982.

Vol. 931: D.M. Arnold, Finite Rank Torsion Free Abelian Groups and Rings. VII, 191 pages. 1982.

Vol. 932: Analytic Theory of Continued Fractions. Proceedings, 1981. Edited by W.B. Jones, W.J. Thron, and H. Waadeland. VI, 240 pages. 1982.

Vol. 933: Lie Algebras and Related Topics. Proceedings, 1981. Edited by D. Winter. VI, 236 pages. 1982.

Vol. 934: M. Sakai, Quadrature Domains. IV, 133 pages. 1982.

Vol. 935: R. Sot, Simple Morphisms in Algebraic Geometry. IV, 146 pages. 1982.

Vol. 936: S.M. Khaleelulla, Counterexamples in Topological Vector Spaces. XXI, 179 pages. 1982.

Vol. 937: E. Combet, Intégrales Exponentielles. VIII, 114 pages. 1982.

Vol. 938: Number Theory. Proceedings, 1981. Edited by K. Alladi. IX, 177 pages. 1982.

Vol. 939: Martingale Theory in Harmonic Analysis and Banach Spaces. Proceedings, 1981. Edited by J.-A. Chao and W.A. Woyczyński. VIII, 225 pages. 1982.

Vol. 940: S. Shelah, Proper Forcing. XXIX, 496 pages. 1982.

Vol. 941: A. Legrand, Homotopie des Espaces de Sections. VII, 132 pages. 1982.

Vol. 942: Theory and Applications of Singular Perturbations. Proceedings, 1981. Edited by W. Eckhaus and E.M. de Jager. V, 363 pages. 1982.

Vol. 943: V. Ancona, G. Tomassini, Modifications Analytiques. IV, 120 pages. 1982.

Vol. 944: Representations of Algebras. Workshop Proceedings, 1980. Edited by M. Auslander and E. Lluis. V, 258 pages. 1982.

Vol. 945: Measure Theory. Oberwolfach 1981, Proceedings. Edited by D. Kölzow and D. Maharam-Stone. XV, 431 pages. 1982.

Vol. 946: N. Spaltenstein, Classes Unipotentes et Sous-groupes de Borel. IX, 259 pages. 1982.

Vol. 947: Algebraic Threefolds. Proceedings, 1981. Edited by A. Conte. VII, 315 pages. 1982.

Vol. 948: Functional Analysis. Proceedings, 1981. Edited by D. Butković, H. Kraljević, and S. Kurepa. X, 239 pages. 1982.

Vol. 949: Harmonic Maps. Proceedings, 1980. Edited by R.J. Knill, M. Kalka and H.C.J. Sealey. V, 158 pages. 1982.

Vol. 950: Complex Analysis. Proceedings, 1980. Edited by J. Eells. IV, 428 pages. 1982.

Vol. 951: Advances in Non-Commutative Ring Theory. Proceedings, 1981. Edited by P.J. Fleury. V, 142 pages. 1982.

Vol. 952: Combinatorial Mathematics IX. Proceedings, 1981. Edited by E. Billington, S. Oates-Williams, and A.P. Street. XI, 443 pages. 1982.

Vol. 953: Iterative Solution of Nonlinear Systems of Equations. Proceedings, 1982. Edited by R. Ansorge, Th. Meis, and W. Törnig. VII, 202 pages. 1982.

Vol. 954: S.G. Pandit, S.G. Deo, Differential Systems Involving Impulses. VII, 102 pages. 1982.

Vol. 955: G. Gierz, Bundles of Topological Vector Spaces and Their Duality. IV, 296 pages. 1982.

Vol. 956: Group Actions and Vector Fields. Proceedings, 1981. Edited by J.B. Carrell. V, 144 pages. 1982.

Vol. 957: Differential Equations. Proceedings, 1981. Edited by D.G. de Figueiredo. VIII, 301 pages. 1982.

Vol. 958: F.R. Beyl, J. Tappe, Group Extensions, Representations, and the Schur Multiplicator. IV, 278 pages. 1982.

Vol. 959: Géométrie Algébrique Réelle et Formes Quadratiques, Proceedings, 1981. Edité par J.-L. Colliot-Thélène, M. Coste, L. Mahé, et M.-F. Roy. X, 458 pages. 1982.

Vol. 960: Multigrid Methods. Proceedings, 1981. Edited by W. Hackbusch and U. Trottenberg. VII, 652 pages. 1982.

Vol. 961: Algebraic Geometry. Proceedings, 1981. Edited by J.M. Aroca, R. Buchweitz, M. Giusti, and M. Merle. X, 500 pages. 1982.

Vol. 962: Category Theory. Proceedings, 1981. Edited by K.H. Kamps D. Pumplün, and W. Tholen, XV, 322 pages. 1982.

Vol. 963: R. Nottrot, Optimal Processes on Manifolds. VI, 124 pages. 1982.

Vol. 964: Ordinary and Partial Differential Equations. Proceedings, 1982. Edited by W.N. Everitt and B.D. Sleeman. XVIII, 726 pages. 1982.

Vol. 965: Topics in Numerical Analysis. Proceedings, 1981. Edited by P.R. Turner. IX, 202 pages. 1982.

Vol. 966: Algebraic K-Theory. Proceedings, 1980, Part I. Edited by R.K. Dennis. VIII, 407 pages. 1982.

Vol. 967: Algebraic K-Theory. Proceedings, 1980. Part II. VIII, 409 pages. 1982.

Vol. 968: Numerical Integration of Differential Equations and Large Linear Systems. Proceedings, 1980. Edited by J. Hinze. VI, 412 pages. 1982.

Vol. 969: Combinatorial Theory. Proceedings, 1982. Edited by D. Jungnickel and K. Vedder. V, 326 pages. 1982.

Vol. 970: Twistor Geometry and Non-Linear Systems. Proceedings, 1980. Edited by H.-D. Doebner and T.D. Palev. V, 216 pages. 1982.

Vol. 971: Kleinian Groups and Related Topics. Proceedings, 1981. Edited by D.M. Gallo and R.M. Porter. V, 117 pages. 1983.

Vol. 972: Nonlinear Filtering and Stochastic Control. Proceedings, 1981. Edited by S.K. Mitter and A. Moro. VIII, 297 pages. 1983.

Vol. 973: Matrix Pencils. Proceedings, 1982. Edited by B. Kågström and A. Ruhe. XI, 293 pages. 1983.

Vol. 974: A. Draux, Polynômes Orthogonaux Formels – Applications. VI, 625 pages. 1983.

Vol. 975: Radical Banach Algebras and Automatic Continuity. Proceedings, 1981. Edited by J.M. Bachar, W.G. Bade, P.C. Curtis Jr., H.G. Dales and M.P. Thomas. VIII, 470 pages. 1983.

Vol. 976: X. Fernique, P.W. Millar, D.W. Stroock, M. Weber, Ecole d'Eté de Probabilités de Saint-Flour XI – 1981. Edited by P.L. Hennequin. XI, 465 pages. 1983.

Vol. 977: T. Parthasarathy, On Global Univalence Theorems. VIII, 106 pages. 1983.

Vol. 978: J. Ławrynowicz, J. Krzyż, Quasiconformal Mappings in the Plane. VI, 177 pages. 1983.

Vol. 979: Mathematical Theories of Optimization. Proceedings, 1981. Edited by J.P. Cecconi and T. Zolezzi. V, 268 pages. 1983.

PREFACE

This book is about numerical methods for problems of finding the largest or smallest values which can be attained by functions of several real variables subject to several inequality constraints. If such problems involve continuously differentiable functions, they can be solved by a variety of methods well documented in the literature. We are concerned with more general problems in which the functions are locally Lipschitz continuous, but not necessarily differentiable or convex. More succintly, this book is about numerical methods for nondifferentiable optimization.

Nondifferentiable optimization, also called nonsmooth optimization, has many actual and potential applications in industry and science. For this reason, a great deal of effort has been devoted to it during the last decade. Most research has gone into the theory of nonsmooth optimization, while surprisingly few algorithms have been proposed, these mainly by C.Lemaréchal, R.Mifflin and P.Wolfe. Frequently such algorithms are conceptual, since their storage and work per iteration grow infinitely in the course of calculations. Also their convergence properties are usually weaker than those of classical methods for smooth optimization problems.

This book gives a complete state-of-the-art in general-purpose methods of descent for nonsmooth minimization. The methods use piecewise linear approximations to the problem functions constructed from several subgradients evaluated at certain trial points. At each iteration, a search direction is found by solving a quadratic programming subproblem and then a line search produces both the next improved approximation to a solution and a new trial point so as to detect gradient discontinuities. The algorithms converge to points satisfying necessary optimality conditions. Also they are widely applicable, since they require only a weak semismoothness hypothesis on the problem functions which is likely to hold in most applications.

A unifying theme of this book is the use of subgradient selection and aggregation techniques in the construction of methods for nondifferentiable optimization. It is shown that these techniques give rise in a totally systematic manner to new implementable and globally convergent modifications and extensions of all the most promising algorithms which have been recently proposed. In effect, this book should give the reader a feeling for the way in which the subject has developed and is developing, even though it mainly reflects the author's research.

This book does not discuss methods without a monotonic descent (or ascent) property, which have been developed in the Soviet Union.

The reason is that the subject of their effective implementations is still a mystery. Moreover, these subgradient methods are well described in the monograph of Shor (1979). We refer the reader to Shor's excellent book (its English translation was published by Springer-Verlag in 1985) for an extensive discussion of specific nondifferentiable optimization problems that arise in applications. Due to space limitations, such applications will not be treated in this book.

In order to make the contents of this book accessible to as wide a range of readers as possible, our analysis of algorithms will use only a few results from nonsmooth optimization theory. These, as well as certain other results that may help the reader in applications, are briefly reviewed in the introductory chapter, which also contains a review of representative existing algorithms. The reader who has basic familiarity with nonsmooth functions may skip this chapter and start reading from Chapter 2, where methods for unconstrained convex minimization are described in detail. The basic constructions of Chapter 2 are extended to the unconstrained nonconvex case in two fundamentally different ways in Chapters 3 and 4, giving rise to competitive methods. Algorithms for constrained convex problems are treated in Chapter 5, and their extensions to the nonconvex case are described in Chapter 6. Chapter 7 presents new versions of the bundle method of Lemaréchal and its extensions to constrained and nonconvex problems. Chapter 8 contains a few numerical results.

The book should enable research workers in various branches of science and engineering to use methods for nondifferentiable optimization more efficiently. Although no computer codes are given in the text, the methods are described unambiguously, so computer programs may readily be written.

The author would like to thank Claude Lemaréchal and Dr. A.Ruszczyński for introducing him to the field of nonsmooth optimization, and Prof. K.Malanowski for suggesting the idea of the book. Without A.Ruszczyński's continuing help and encouragement this book would not have been written. Part of the results of this book were obtained when the author worked for his doctoral dissertation under the supervision of Prof. A.P.Wierzbicki at the Institute of Automatic Control of the Technical University of Warsaw. The help of Prof. R.Kulikowski and Prof. J.Hołubiec from the Systems Research Institute of the Polish Academy of Sciences, where this book was written, is gratefully acknowledged. Finally, the author wishes to thank Mrs. I.Forowicz and Mrs. E.Grudzińska for patiently typing the manuscript.

TABLE OF CONTENTS

CHAPTER 1

Fundamentals

1. Introduction

The nonlinear programming problem, also known as the mathematical programming problem, can be taken to have the form

$$P: \quad \text{minimize} \quad f(x), \quad \text{subject to} \quad F_i(x) \le 0 \quad \text{for} \quad i=1,\dots,m,$$

where the objective function f and the constraint functions F_i are real-valued functions defined on the N-dimensional Euclidean space R^N. The value of $m \ge 0$ is finite; when $m=0$ the problem is unconstrained. Often the optimization problem P is smooth: the problem functions f and F_i are continuously differentiable, i.e. they have continuous gradients ∇f and ∇F_i, $i=1,\dots,m$. But in many applications this is not true. Nonsmooth problems are the subject of nonsmooth optimization, also called nondifferentiable optimization.

Owing to actual and potential applications in industry and science, recently much research has been conducted in the area of nonsmooth optimization both in the East (see the excellent monographs by Gupal (1979), Nurminski (1979) and Shor (1979)) and in the West (see the comprehensive bibliographies of Gwinner (1981) and Nurminski (1982)).

Nonsmooth problems that arise in applications have certain common features. They are more complex and have poorer analytical properties than standard mathematical programming problems, cf. (Bazaraa and Shetty, 1979; Pshenichny and Danilin, 1975). A single evaluation of the problem functions usually requires solutions of auxiliary optimization subproblems. In particular, it is very common to encounter a nondifferentiable function which is the pointwise supremum of a collection of functions that may themselves be differentiable - a max function.

Functions with discontinuous gradients, such as max functions, cannot be minimized by classical nonlinear programming algorithms. This observation applies both to gradient-type algorithms (the method of steepest descent, conjugate direction methods, quasi-Newton methods) and to direct search methods which do not require calculation of derivatives (the method of Nelder and Mead, the method of Powell, etc.), see (Lemarechal, 1978 and 1982; Wolfe, 1975).

This work is concerned with numerical methods for finding (approximate) solutions to problem P when the problem functions are locally Lipschitzian, i.e. Lipschitz continuous on each bounded subset of R^N, but not necessarily differentiable.

The advent of F.H.Clarke's (1975) analysis of locally Lipschitzian functions provided a unified approach to both nondifferentiable and nonconvex problems (Clarke, 1976). Clarke's subdifferential analysis, the pertinent part of which is briefly reviewed in the following section, suffices for establishing properties of a vast class of optimization problems that arise in applications (Pshenichny, 1980; Rockafellar, 1978).

2. Basic Results of Nondifferentiable Optimization Theory

In this section we describe general properties of nondifferentiable optimization problems that are the subject of this work. Basic familiarity is, however, assumed. Source material may be found in (Clarke, 1975; Clarke, 1976; Rockafellar, 1970; Rockafellar, 1978; Rockafellar, 1981).

The section is organized as follows. First, we review concepts of differentiability and elementary properties of the Clarke subdifferential. The proofs are omitted, because only simple results, such as Lemma 2.2, will be used in subsequent chapters. Other results, in particular the calculus of subgradients, should help the reader who is mainly interested in applications. Secondly, we study convex first order approximations to nondifferentiable functions. Such approximations are then used for deriving necessary conditions of optimality for nondifferentiable problems. Our approach is elementary and may appear artificial. However, it yields useful interpretations of the algorithms described in subsequent chapters.

The following notation is used. We denote by $\langle\cdot,\cdot\rangle$ and $|\cdot|$, respectively, the usual inner product and norm in finite-dimensional, real Euclidean space. R^N denotes Euclidean space of dimension $N<\infty$. We use x_i to denote the i-th component of the vector x. Thus $\langle x,y\rangle = \sum_{i=1}^{N} x_i y_i$ and $|x| = \langle x,x\rangle^{1/2}$ for $x,y \in R^N$. Superscripts are used to denote different vectors, e.g. x^1 and x^2. All vectors are column vectors. However, for convenience a column vector in R^{N+n} is sometimes denoted by (x,y) even though x and y are column vectors in R^N and R^n, respectively. $[x,y]$ denotes the line segment joining x and y in R^N, i.e. $[x,y]=\{z \in R^N: z=\lambda x+(1-\lambda)y$ for some λ satisfying $0 \le \lambda \le 1\}$.

A set $S \subset R^N$ is called <u>convex</u> if $[x,y] \subset S$ for all x and y belonging to S. A linear combination $\sum_{j=1}^{k} \lambda_j x^j$ is called a <u>convex combination</u> of points $x^1,\dots,x^k$ in R^N if each $\lambda_j \ge 0$ and $\sum_{j=1}^{k} \lambda_j = 1$. The <u>convex hull</u> of a set $S \subset R^N$, denoted conv S, is the set of all convex combinations of points in S. conv S is the smallest convex set containing S, and S is convex if and only if S=conv S. An important property of convex hulls is described in

Lemma 2.1 (Caratheodory's theorem; see Theorem 17.1 in (Rockafellar, 1970)).
If $S \subset R^N$ then $x \in \text{conv } S$ if and only if x is expressible as a convex combination of $N+1$ (not necessarily different) points of S.

Any nonzero vector $g \in R^N$ and number γ define a hyperplane

$$H = \{x \in R^N : \langle g,x \rangle = \gamma\},$$

which is a translation of the (N-1)-dimensional subspace $\{x \in R^N : \langle g,x \rangle = 0\}$ of R^N. H divides R^N into two closed half-spaces $\{x \in R^N : \langle g,x \rangle \le \gamma\}$ and $\{x \in R^N : \langle g,x \rangle \ge \gamma\}$, respectively. We say that H is a supporting hyperplane to a set $S \subset R^N$ at $\bar{x} \in S$ if $\langle g,\bar{x} \rangle = \gamma$ and $\langle g,x \rangle \le \gamma$ for all $x \in S$. Any closed convex set S can be described as an intersection of all the closed half-spaces that contain S.

We use the set notation

$$S^1+S^2 = \{z^1+z^2 : z^1 \in S^1, z^2 \in S^2\},$$

$$\text{conv}\{S^i : i=1,2\} = \text{conv}\{z : z \in S^1 \cup S^2\}$$

for any subsets S^1 and S^2 of R^N.

A function $f: R^N \longrightarrow R$ is called convex if

$$f(\lambda x^1+(1-\lambda)x^2) \le \lambda f(x^1)+(1-\lambda)f(x^2) \quad \text{for all } \lambda \in [0,1] \text{ and } x^1,x^2 \in R.$$

This is equivalent to the epigraph of f

$$\text{epi } f = \{(x,\beta) \in R^{N+1} : \beta \ge f(x)\}$$

being a convex subset of R^{N+1}. A function $f: R^N \longrightarrow R^1$ is called concave if the function $(-f)(x) = -f(x)$ is convex. If $F_i: R^N \longrightarrow R$ is convex and $\lambda_i \ge 0$ for each $i=1,\dots,k$, then the functions

$$\phi_1(x) = \sum_{i=1}^{k} \lambda_i f_i(x),$$
$$\phi_2(x) = \max\{f_i(x) : i=1,\dots,k\} \tag{2.1}$$

are convex.

A function $f: R^N \longrightarrow R$ is strictly convex if $f(\lambda x^1+(1-\lambda)x^2) < \lambda f(x^1)+(1-\lambda)f(x^2)$ for all $\lambda \in (0,1)$ and $x^1 \ne x^2$. For instance, the function $|.|^2$ is strictly convex.

A function $f : R^N \longrightarrow R$ is said to be locally Lipschitzian if for each bounded subset B of R^N there exists a Lipschitz constant $L = L(B) < \infty$ such that

$$|f(x^1)-f(x^2)| \le L|x^1-x^2| \quad \text{for all} \quad x^1,x^2 \in B. \tag{2.2}$$

Then in particular f is continuous. Examples of locally Lipschitzian functions include continuously differentiable functions, convex functions, concave functions and any linear combination or pointwise maximum of a finite collection of such functions, cf. (2.1).

Following (Rockafellar, 1978), we shall now describe differentiability properties of locally Lipschitzian functions. Henceforth let f denote a function satisfying (2.2) and let x be an interior point of B, i.e. $x \in \text{int } B$.

The Clarke generalized directional derivative of f at x in a direction d

$$f^o(x;d) = \limsup_{y \to x, t \downarrow 0} [f(y+td)-f(y)]/t \tag{2.3}$$

is a finite, convex function of d and $f^o(x;d) \le L|d|$. The Dini upper directional derivative of f at x in a direction d

$$f^D(x;d) = \limsup_{t \downarrow 0} [f(x+td)-f(x)]/t \tag{2.4}$$

exists for each $d \in R^N$ and satisfies

$$f(x+td) \le f(x)+tf^D(x;d)+o(t), \tag{2.5}$$

where $o(t)/t \to 0$ as $t \downarrow 0$. The limit

$$f'(x;d) = \lim_{t \downarrow 0} [f(x+td)-f(x)]/t \tag{2.6}$$

is called the (one-sided) directional derivative of f at x with respect to d, if it exists. The two-sided derivative (the Gateaux derivative) corresponds to the case $f'(x;-d)=-f'(x;d)$. Clearly,

$$\begin{aligned} f^D(x;d) &\le f^o(x;d), \\ f'(x;d) &\le f^D(x;d), \end{aligned} \tag{2.7}$$

whenever $f'(x;d)$ exists.

If $f'(x;d)$ is linear in d (Gateaux differentiable at x)

$$f'(x;d) = \langle g_f, d \rangle \quad \text{for all} \quad d \in R^N, \tag{2.8}$$

then the vector g_f is called the gradient of f at x and denoted by $\nabla f(x)$. The components of $\nabla f(x)=(\frac{\partial f}{\partial x_1}(x),\ldots,\frac{\partial f}{\partial x_N}(x))$ are the coordinate-wise two-sided partial derivatives of f at x. The function f is (Frechet) differentiable at x if

$$f(x+d)=f(x)+\langle\nabla f(x),d\rangle+o(|d|) \quad \text{for all } d\in R^N, \tag{2.9}$$

where $o(t)/t\to 0$ as $t\downarrow 0$. The above relation is equivalent to

$$\lim_{d'\to d,t\downarrow 0} [f(x+td')-f(x)]/t=\langle\nabla f(x),d\rangle \quad \text{for all } d\in R^N. \tag{2.10}$$

If

$$\lim_{y\to x,t\downarrow 0} [f(y+td)-f(y)]/t = \langle\nabla f(x), d\rangle \quad \text{for all } d \text{ in } R^N, \tag{2.11}$$

then f is called strictly differentiable at x. In this case f is differentiable at x and the gradient $\nabla f:R^N\to R^N$ is continuous at x relative to its domain

$$\text{dom } \nabla f = \{y\in R^N: f \text{ is differentiable at } y\}$$

It is known that a locally Lipschitzian function $f:R^N\to R$ is differentiable at almost all points $x\in R^N$, and moreover that the gradient mapping ∇f is locally bounded on its domain. Suppose that (2.2) holds for some neighborhood B of a point $x\in R^N$. Then

$$\langle\nabla f(y),d\rangle = f'(y;d) = \lim_{t\downarrow 0}[f(y+td)-f(y)]/t \le L|d|$$

for all $y\in B\cap \text{dom } \nabla f$ and $d\in R^N$, and this implies

$$|\nabla f(y)|\le L \text{ for all } y\in B\cap \text{dom } \nabla f. \tag{2.12}$$

Since dom ∇f is dense in B, there exist sequences $\{y^j\}$ such that f is differentiable at y^j and $y^j\to x$. The corresponding sequence of gradients $\{\nabla f(y^j)\}$ is bounded and has accumulation points (each being the limit of some convergent subsequence). It follows that the set

$$M_f(x) = \{z\in R^N: \nabla f(y^j)\to z \text{ for some sequence } y^j\to z \text{ with f differentiable at } y^j\} \tag{2.13a}$$

is nonempty, bounded and closed. The set

$$\partial f(x) = \text{conv } M_f(x) \tag{2.13b}$$

is called the subdifferential of f at x (called the generalized gradient by Clarke (1975)). Each element $g_f \in \partial f(x)$ is called a subgradient of f at x. Thus

$$\partial f(x)=\text{conv}\{\lim \nabla f(y^j): y^j \to x, \quad f \text{ differentiable at } y^j\}. \tag{2.14}$$

In particular therefore, $\partial f(x)=-\partial(-f)(x)$. Three immediate consequences of the definition are listed in

Lemma 2.2. (i) $\partial f(x)$ is a nonempty convex compact set.

(ii) The point-to-set mapping $\partial f(\cdot)$ is locally bounded (bounded on bounded subsets of R^N), i.e. if $B \subset R^N$ is bounded then the set $\{g_f \in \partial f(y): y \in B\}$ is bounded.

(iii) $\partial f(\cdot)$ is upper semicontinuous, i.e. if a sequence $\{y^j\}$ converges to x and $g_f^j \in \partial f(y^j)$ for each j then each accumulation point g_f of $\{g_f^j\}$ satisfies $g_f \in \partial f(x)$.

In general, $\partial f(x)$ does not reduce to $\nabla f(x)$ when the gradient ∇f is discontinuous at x.

Lemma 2.3. The following are equivalent:

(i) $\partial f(x)$ consists of a single vector;

(ii) $\nabla f(x)$ exists and ∇f is continuous at x relative to dom ∇f;

(iii) f is strictly differentiable at x.

Moreover, when these properties hold one has $\partial f(x)=\{\nabla f(x)\}$.

Frequently $\partial f(x)$ is a singleton for almost every x. A locally Lipschitzian function $f: R^N \to R$ is subdifferentially regular at $x \in R^N$ if for every $d \in R^N$ the ordinary directional derivative (2.6) exists and coincides with the generalized one in (2.3):

$$f'(x;d) = f^o(x;d) \quad \text{for all } d. \tag{2.15}$$

If (2.15) holds at each $x \in R^N$ then $\partial f(x)$ is actually single-valued at almost every x. Below we give two important examples of subdifferentially regular functions.

Lemma 2.4. If f is convex then f is subdifferentially regular and

$$f'(x;d) = \max\{\langle g_f,d\rangle : g_f \in \partial f(x)\} \quad \text{for all } x,d. \tag{2.16}$$

Lemma 2.5. Suppose that

$$f(x) = \max\{f_u(x): u\in U\} \qquad \text{for all } x \in R^N, \tag{2.17}$$

where the index set U is a compact topological space (e.g. a finite set in the discrete topology), each f_u is locally Lipschitzian, uniformly for u in U, and the mappings $f_u(x)$ and $\partial f_u(x)$ are upper semicontinuous in (x,u) (e.g. each f_u is a differentiable function such that $f_u(x)$ and $\nabla f_u(x)$ depend continuously on (x,u)). Let

$$U(x) = \{u\in U: f_u(x) = f(x)\}. \tag{2.18}$$

Then f is locally Lipschitzian and

$$\partial f(x) \subset \text{conv}\ \{\partial f_u(x) : u\in U(x)\}. \tag{2.19}$$

If each f_u is subdifferentially regular at x, then so is f, equality holds in (2.19), and

$$f'(x;d) = \max\{\langle g_u,d\rangle: g_u\in\partial f_u(x), u\in U(x)\} \quad \text{for all } d. \tag{2.20}$$

Corollary 2.6. Suppose that

$$f(x) = \max\{f_i(x) : i \in I\} \qquad \text{for all } x \text{ in } R^N , \tag{2.21}$$

where the index set I is finite, and let $I(x)=\{i\in I:f_i(x)=f(x)\}$.
(i) If each f_i is continuously differentiable then

$$\left.\begin{aligned} f'(x;d) &= \max\{\langle\nabla f_i(x),d\rangle : i\in I(x)\} \quad \text{for all } d, \\ \partial f(x) &= \text{conv}\{\nabla f_i(x): i\in I(x)\}. \end{aligned}\right. \tag{2.22}$$

(ii) If each f_i is convex then

$$\left.\begin{aligned} f'(x;d) &= \max\{\langle g_{f_i},d\rangle: g_{f_i}\in\partial f_i(x),\ i\in I(x)\} \quad \text{for all } d, \\ \partial f(x) &= \text{conv}\{g_{f_i} \in \partial f_i(x): i\in I(x)\}. \end{aligned}\right. \tag{2.23}$$

When f is smooth, there exists an apparatus for computing ∇f in terms of the derivatives of other functions from which f is composed. The calculus of subgradients, which generalizes rules like $\nabla(f_1+f_2)(x) = \nabla f_1(x) + \nabla f_2(x)$, is based on the following results.

Lemma 2.7. Let $g:R^n \to R$ and $h_i:R^N \to R$, $i=1,\dots,n$, be locally Lipschitzian. Let $h(x)=(h_1(x),\dots,h_n(x))$ and $(g\circ h)(x)=g(h(x))$ for all $x \in E^N$. Then $g\circ h$ is locally Lipschitzian and

$$\partial(g\circ h)(x) \subset \text{conv}\ \{ \sum_{i=1}^{n} u_i \partial h_i(x) : (u_1,\dots,u_n) \in \partial g(h(x)) \}. \tag{2.24}$$

Moreover, equality holds in (2.24) if one of the following is satisfied:
(i) g is subdifferentially regular at h(x), each h_i is subdifferentially regular at x and $\partial g(h(x)) \subset R^n_+$ ($R^n_+ = \{z \in R^n: z_i \geq 0$ for all $i\}$);
(ii) g is subdifferentially regular at h(x) and each h_i is continuously differentiable at x;
(iii) Each h_i is continuously differentiable at x and either g (or - g) is subdifferentially regular at h(x) or the Jacobian matrix of h at x is surjective;
(iv) n=1, g is continuously differentiable at h(x) or g (or - g) is subdifferentially regular at h(x) and h is continuously differentiable at x. In cases (ii) - (iv) the symbol "conv" is superfluous in (2.24). If (ii) holds then $g\circ h$ is subdifferentially regular at x.

Corollary 2.8. Suppose that f_1 and f_2 are locally Lipschitzian on R^N. For each $x \in R^N$ let $(f_1+f_2)(x)=f_1(x)+f_2(x)$, $(f_1f_2)(x)=f_1(x)f_2(x)$ and $(f_1/f_2)(x)=f_1(x)/f_2(x)$ if $f_2(x)\neq 0$. Then

$$\partial(f_1+f_2)(x) \subset \partial f_1(x)+\partial f_2(x), \tag{2.25a}$$

$$\partial(f_1f_2)(x) \subset f_2(x)\ \partial f_1(x)+f_1(x)\partial f_2(x), \tag{2.25b}$$

$$\partial(f_1/f_2)(x) \subset \frac{1}{(f_2(x))^2}\ [f_2(x)\partial f_1(x)-f_1(x)\partial f_2(x)]. \tag{2.25c}$$

Equality holds in (2.25a) if each f_i is subdifferentially regular at x, and in (2.25b) if in addition $f_i(x) \geq 0$.

Clarke (1975) established the following crucial relations between the subdifferential and the generalized directional derivatives of a locally Lipschitzian function f defined on R^N

$$f^o(x;d) = \max\{<g_f,d>: g_f \in \partial f(x)\} \text{ for all } x,d, \tag{2.26}$$

$$\partial f(x) = \{g_f \in R^N: <g_f,d> \leq f^o(x;d) \text{ for all } d\} \text{ for all } x. \tag{2.27}$$

We shall now interpret these relations in geometric terms. In what follows let x be a fixed point in R^N.

First, suppose that f is continuously differentiable at x. From Lemma 2.3, (2.26) and (2.8) we have

$$\partial f(x) = \{\nabla f(x)\}, \tag{2.28a}$$

$$f^o(x;d) = f'(x;d) = <\nabla f(x),d> \quad \text{for all } d. \tag{2.28b}$$

Suppose that $\nabla f(x) \neq 0$. Then $\nabla f(x)$ corresponds to the hyperplane

$$H_{\nabla f} = \{(z,\beta) \in R^{N+1}: \beta = f(x) + <\nabla f(x),z-x>\}$$

being tangent to the <u>graph</u> of f

$$\text{graph } f = \{(z,\beta) \in R^{N+1}: \beta = f(z)\}$$

at the point (x,f(x)). Here β denotes the "vertical" coordinate of a point $(x,\beta) \in R^{N+1}$. Moreover, the hyperplane

$$H_C = \{z \in R^N : <\nabla f(x),z-x> = 0\}$$

is tangent at x to the <u>contour</u> of f at x

$$C = \{z \in R^N: f(z) = f(x)\}.$$

$\nabla f(x)$ is perpendicular to C at x and is the direction of steepest ascent for f at x. Define the following <u>linearization</u> of f at x

$$\bar{f}(z) = f(x) + <\nabla f(x),z-x> \quad \text{for all } z \text{ in } R^N \tag{2.29}$$

and observe that $\nabla\bar{f}(z) = \nabla f(x)$ for all z (x is fixed). Therefore this linearization has the same differentiability properties as f at x in the sense that

$$\partial\bar{f}(x) = \partial f(x), \tag{2.30a}$$

$$\bar{f}^o(x;d) = \bar{f}'(x;d) = f^o(x;d) \text{ for all } d, \tag{2.30b}$$

cf. (2.28). In particular, by (2.28a), (2.9) and (2.30b), for any $d \in R^N$ we have