数理解析研究所講究録
第 680 巻 1989 年 107-116

NOTES ON NONSMOOTH OPTIMIZATION

Masao FUKUSHIMA (Kyoto University)
福島雅夫（京都大学）

1. Introduction

Nonsmooth Optimization (NSO) or Nondifferentiable Optimization (NDO) deals with optimization problems whose objective and constraint functions are not necessarily differentiable. The nonsmooth functions which are encountered in practice are often defined as the max function

$$f(x) = \max\{ F(x,y) \mid y \in Y \}, \qquad (1.1)$$

where F is assumed to be smooth with respect to x. Some examples of nonsmooth functions are as follows.

Example 1. Multifacility location problem: Francis and White (1974).

Let m facilities be situated at points $a_1, a_2, \ldots, a_m$ on a plane. Suppose that n new facilities should be located and that cost should be imposed on transportation of goods between the new facilities and the existing facilities and between the new facilities themselves. Then the problem of efficiently locating the new facilities may be formulated as

$$\text{minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ji}(\|x_j-a_i\|) + \sum_{j=1}^{n-1}\sum_{k=j+1}^{n} g_{jk}(\|x_j-x_k\|), \qquad (1.2)$$

where $x_1, x_2, \ldots, x_n$ are the locations of the new facilities to be found, the functions f_{ji} and g_{jk} are nondecreasing, and $\|\cdot\|$ is a norm in R^2. Since $\|\cdot\|$ has discontinous first derivatives at the origin, the objective function of problem (1.2) is usually nonsmooth.

Example 2. Exact penalty function: Fletcher (1981).

Consider the nonlinear programming problem

$$\begin{array}{lll} \text{minimize} & f(x) & \qquad (1.3) \\ \text{subject to} & c_i(x) = 0, & i=1,2,\ldots,m' \\ & c_i(x) \leqq 0, & i=m'+1,\ldots,m, \end{array}$$

where f and c_i are assumed to be smooth. Then it is well known that, under some regularity conditions, a solution of problem (1.3) may be obtained by solving the unconstrained problem

$$\text{minimize} \quad F_r(x), \tag{1.4}$$

where F_r is defined by

$$F_r(x) = f(x) + r[\sum_{i=1}^{m'} |c_i(x)| + \sum_{i=m'+1}^{m} \max(0, c_i(x))]$$

and r is a sufficiently large positive constant. The function F_r is clearly nonsmooth.

<u>Example 3</u>. Decomposition of large problems: Fukushima (1987).

Consider the nonlinear programming problem

$$\begin{aligned} &\text{minimize} \quad && F(x) + \langle q,y\rangle \\ &\text{subject to} \quad && Ax + By \leqq b, \\ & && c_i(x) = 0, \quad i=1,2,\ldots,m, \end{aligned} \tag{1.5}$$

where F and c_i are smooth, A, B, q and b are matrices or vectors of appropriate dimension, and $\langle\cdot,\cdot\rangle$ denotes the inner product. Then, as in Benders' decomposition method, we may rewrite (1.5) as

$$\begin{aligned} &\text{minimize} \quad && F(x) + f(x) \\ &\text{subject to} \quad && c_i(x) = 0, \quad i=1,2,\ldots,m, \end{aligned} \tag{1.6}$$

where f is given by

$$\begin{aligned} f(x) &= \min\{ \langle q,y\rangle \mid By \leqq b - Ax \} \\ &= \max\{ \langle Ax-b,w\rangle \mid B^T w = -q,\ w \geqq 0\}. \end{aligned} \tag{1.7}$$

(The last equality follows from the duality theory in linear programming.) The function f is a (polyhedral) convex function which is generally nonsmooth.

<u>Example 4</u>. Multicommodity flow problem: Fukushima (1984).

The equilibrium traffic assignment problem and the optimal routing problem in a packet switched communication network can be formulated as a convex cost multicommodity flow problem. Consider a directed network with the set N of nodes and the set A of arcs. Let K denote the set of commodities to be transported through the network. Assume that each commodity $k \in K$ has a single origin-destination (OD) pair (s_k, t_k) and let P_k be the set of paths between the OD pair (s_k, t_k). Then the minimum cost multicommodity flow problem may be stated as

$$\begin{aligned} &\text{minimize} \quad \sum_{a \in A} f_a\Big(\sum_{k \in K} \sum_{p \in P_k} \delta_{ap} y_p\Big) && (1.8) \\ &\text{subject to} \quad \sum_{p \in P_k} y_p = D_k, \quad \forall k \in K, \\ &\qquad\qquad y_p \geq 0, \quad \forall p \in P_k, \quad \forall k \in K, \end{aligned}$$

where f_a is the cost function for arc a, x_a is the total flow on arc a, y_p is the flow (of commodity k) along path p, D_k is the given nonnegative flow requirement for commodity k, and δ_{ap} are elements of the arc-path incidence matrix associated with the given network. If we assume that the functions f_a are convex, then the dual of problem (1.8) may be given as

$$\text{minimize} \quad \sum_{a \in A} f_a^*(u_a) + \sum_{k \in K} D_k h_k(u), \qquad (1.9)$$

where f_a^*, the conjugate function of f_a, is defined by

$$f_a^*(u_a) = \sup\{x_a u_a - f_a(x_a) \mid x_a \in R\}, \qquad (1.10)$$

and h_k is defined by

$$h_k(u) = -\min\{\sum_{a \in A} u_a \delta_{ap} \mid p \in P_k\}. \qquad (1.11)$$

Note that both f_a^* and h_k are convex. In particular, since f_a^* is a function of a single variable, it is often possible to obtain an explicit representation of f_a^* from (1.10). Moreover, (1.11) indicates that, for each k, $h_k(u)$ can be calculated by finding a shortest path between the OD pair (s_k, t_k) in the network with arc lengths given by u_a, $a \in A$.

2. Algorithms

The key tool of dealing with nonsmooth functions is the notion of subgradient (Rockafellar, 1970). Let f be a convex function on R^n. Then, a vector g is called a subgradient of f at x if

$$f(x') - f(x) \geq \langle g, x'-x \rangle \quad \text{for all } x'. \qquad (2.1)$$

The set of subgradients of f at x is denoted by $\partial f(x)$. The notion of subgradient has been generalized to various classes of nonconvex functions. (See, e.g., Clarke (1983).)

Roughly speaking, we may classify two different types of nonsmooth functions, depending upon the information available. Womersley and Fletcher (1986) use the terminology, "composite nonsmooth problems" and "basic nonsmooth problems". For the former, there is sufficient information available at each x to calculate the set $\partial f(x)$ of subgradients completely, while for the latter, the set of subgradients must be approximated using information evaluated at various point around x. When the nonsmooth function is defined as a max function of the form (1.1), the former correspond to the case where the set Y is finite and all component functions $F(\cdot,y)$ are easily enumerated, and the latter correspond to the case where the number of elements in Y is either infinite or too many to enumerate individually. Examples 1 and 2 in the previous section belong to the class of composite nonsmooth problems, and Examples 3 and 4 typify the basic nonsmooth problems.

2.1. Algorithms for Composite Nonsmooth Problems

Typical composite nonsmooth problems are represented as

$$\text{minimize} \quad \phi(x) \triangleq F(x) + h(c(x)), \tag{2.2}$$

where $F: R^n \to R$ and $c: R^n \to R^m$ are smooth and $h: R^m \to R$ is a polyhedral convex function. Fletcher (1981) presents an algorithm which constructs a quadratic approximation of problem (2.2) at each trial point and uses a trust region technique to guarantee the validity of the quadtaric model as a good approximation to the original problem. Specifically, the algorithm of Fletcher (1981) solves at each iteration the following subproblem:

$$\begin{aligned} &\text{minimize} \quad q^k(d) + h(\ell^k(d)) && (2.3)\\ &\text{subject to} \quad \|d\|_\infty \leq \Delta^k \end{aligned}$$

where

$$q^k(d) = F(x^k) + \nabla F(x^k)^T d + \frac{1}{2} d^T W^k d,$$
$$\ell^k(d) = c(x^k) + \nabla c(x^k)^T d,$$
$$W^k = \nabla^2 F(x^k) + \sum_{i=1}^{m} \lambda_i^k \nabla^2 c_i(x^k),$$

λ_i^k are estimates of the optimal Lagrange multipliers, and Δ^k is the radius of the trust region. Note that problem (2.3) can be transformed into a quadratic programming problem.

Let d^k be a solution of (2.3). If d^k gives a sufficient reduction in the objective function ϕ of the original problem (2.2), then the next iterate x^{k+1} is chosen to be x^k+d^k. Otherwise, the radius of the trust region has to be controlled to enforce the reduction of the function ϕ. Fletcher (1981) shows that, under suitable conditions, such a step restriction strategy generates a sequence whose accumulation point satisfies the optimality conditions for (2.2).

Concerning the rate of convergence, it is known that the trust region algorithm of Fletcher (1981) suffers from the so called Maratos effect which may hinder superlinear convergence. To overcome this difficulty, Fletcher (1982) proposes to solve another subproblem of the following form, after the subproblem (2.3) is solved:

$$\begin{array}{ll} \text{minimize} & q^k(d) + h(\bar{c}^k + \nabla c(x^k)^T d) \qquad (2.4) \\ \text{subject to} & \|d\|_\infty \leq \Delta^k, \end{array}$$

where $\bar{c}^k = c(x^k+d^k) - \nabla c(x^k)^T d^k$ and d^k is the solution of (2.3) . Fletcher (1981) establishes a superlinear convergence result for this modified algorithm.

More recently, Yamakawa, Fukushima and Ibaraki (1989) have proposed another remedy of overcoming the Maratos effect, using the idea given by Fukushima (1986b) for the SQP methods in nonlinear programming. Specifically, this algorithm solves, instead of (2.4), the following subproblem:

$$\begin{array}{ll} \text{minimize} & \hat{q}^k(d) + h(\hat{\ell}^k(d)) \qquad (2.5) \\ \text{subject to} & \|d\|_\infty \leq \Delta^k, \end{array}$$

where

$$\hat{q}^k(d) = F(x^k) + (p^k)^T d + \tfrac{1}{2} d^T W^k d,$$
$$\hat{\ell}_i^k(d) = c_i(x^k) + (a_i^k)^T d,$$
$$p^k = \nabla F(x^k) - \sum_{i=1}^{m} \lambda_i^k r_i^k,$$
$$a_i^k = \nabla c_i(x^k) + r_i^k, \quad i=1,2,\dots,m,$$
$$r_i^k = \tfrac{1}{2}\nabla^2 c_i(x^k) d^k.$$

Note that the calculation of the above vectors requires no extra work of evaluating the second derivatives, because they have already been used to evaluate the matrix W^k in subproblem (2.3). In Yamakawa, Fukushima and Ibaraki (1989), it is shown that the Maratos effect can be avoided by using the solutions of subproblems (2.5) and very promising numerical results are reported for some ill-conditioned test problems.

2.2 Algorithms for Basic Nonsmooth Problems

For a function involved in basic nonsmooth problems, it is usually assumed that only one subgradient is available at each point. Early numerical methods that can deal with nonsmooth functions are subgradient methods and cutting plane methods. In particular, the subgradient method is one of the antecedents of the ellipsoid method for linear programming. The subgradient method and the ellipsoid method are fully described by Shor (1985) and Bland, Goldfarb and Todd (1981), respectively. The cutting plane method, which was independently developed by Cheney and Goldstein and Kelley about thirty years ago, has played an important role in nonsmooth optimization. The cutting plane method may be viewed as a predecessor of various descent methods developed by a number of authors inluding Lemarechal, Wolfe, Mifflin, Fukushima, Kiwiel and Auslender (see, e.g., Kiwiel (1985) and the references therein.) The primitive cutting plane method has the drawback of accumulating cutting planes infinitely. An attempt to overcome this difficulty is proposed by Fukushima (1983) for nonsmooth convex programs and this idea has been extended further to variational inequalities by Fukushima (1986a).

In the rest of this section, we briefly describe a class of descent methods which may be considered a modification of the proximal method studied by Rockafellar (1976).

First let us consider the unconstrained problem

$$\text{minimize} \quad f(x), \tag{2.6}$$

where f is a convex function. Given the current iterate x, the proximal method (approximately) solves the subproblem

$$\text{minimize} \quad \frac{\lambda}{2}\|p\|^2 + f(x+p), \tag{2.7}$$

where λ is a positive parameter. Once the solution p of (2.7) is obtained, the next iterate x_+ is then determined by $x_+ = x + p$.

Fukushima (1984a) observes that if $\bar{p}$ solves (2.7), then the vector $\bar{g} = \frac{1}{\lambda}\bar{p}$ is the minimum norm element of $\partial_\varepsilon f(x)$, the set of ε-subgradients of f at x, defined by

$$\partial_\varepsilon f(x) = \{\, g \mid f(x')-f(x) \geqq \langle g, x'-x\rangle - \varepsilon,\ \forall x'\}, \tag{2.8}$$

where $\varepsilon \geqq 0$. Thus $\bar{p}$ is a descent direction of f at x, and the next iterate x_+ may be determined as

$$x_+ = x + \alpha\,\bar{p}, \tag{2.9}$$

where the step-size $\alpha > 0$ is chosen in such a way that a sufficient reduction in the objective value is obtained. Another important fact is that solving (2.7) only approximately suffices to guarantee convergence of the whole algorithm. To solve (2.7), Fukushima (1984a) uses a cutting plane technique in which cutting planes are generated only for the convex term $f(x+p)$ and hence search direction p is obtained by solving a sequence of problems of the form

$$\text{minimize} \quad \frac{\lambda}{2}\|p\|^2 + \hat{f}(p), \tag{2.10}$$

where $\hat{f}$ is a polyhedral (outer) approximation of f. Note that (2.10) can easily be transformed into a quadratic programming problem. Recently, Auslender (1987) extended this idea to get a more general method.

The proximal method has been generalized to the nonconvex problem

$$\text{minimize} \quad F(x) + f(x), \tag{2.11}$$

where F is smooth but not necessarily convex, and f is convex but not necessarily smooth. Fukushima and Mine (1981) propose to determine a search direction at x by solving the "convex" subproblem

$$\text{minimize} \quad \frac{\lambda}{2}\|p\|^2 + \langle \nabla F(x), p\rangle + f(x+p), \tag{2.12}$$

where λ is a positive parameter. It can be shown that the solution of (2.12) is a descent direction of the objective function of (2.11), so that the next iterate may be found by line search. The algorithm presented by Fukushima and Mine (1981) is conceptual in the sense that it requires the exact solution of (2.12) which is usually impractical. Recently Kiwiel (1986) has improved it to obtain an implementable algorithm.

The proximal method can further be extended to a class of constrained nonsmooth nonconvex optimization problems (Fukushima, 1987). Let us consider the problem

$$\begin{aligned} &\text{minimize} && F(x) + f(x) && (2.13)\\ &\text{subject to} && c_i(x) = 0,\ i=1,2,\ldots,m, && \end{aligned}$$

where F and f are the same as in (2.11), and c_i are assumed to be smooth. We note here that the algorithm to be described can deal with smooth inequality constraints by slight modification. Given the current iterate x, the direction finding subproblem can be defiend by

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\langle Bp,p\rangle + \langle \nabla F(x),p\rangle + f(x+p) && (2.14)\\ &\text{subject to} && c(x) + A(x)p = 0, && \end{aligned}$$

where B is symmetric and positive definite, $c(x) = (c_1(x),\ldots,c_m(x))^T$ and $A(x) = \nabla c(x)^T$. Note that the matrix λI used in (2.7) and (2.12) has been replaced by a more general matrix B in (2.14). Problem (2.14) is a linearly constrained "convex" programming problem. Again we can apply the cutting plane technique to find an approximate solution of (2.14). That is, we solve a sequence of problems of the form

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\langle Bp,p\rangle + \langle \nabla F(x),p\rangle + \hat{f}(p) && (2.15)\\ &\text{subject to} && c(x) + A(x)p = 0, && \end{aligned}$$

where f is a polyhedral (outer) approximation of f, defined by the cutting planes generated so far. Problem (2.15) can be transformed into a convex quadratic programming problem. Once an approximate solution of (2.14) satisfying some criteria is found, then the next iterate x_+ is determined by line serach using the exact penalty function

$$F_r(x) = F(x) + f(x) + r\sum_{i=1}^{m} |c_i(x)|. \qquad (2.16)$$

The penalty parameter $r > 0$ must be large enough to guarantee global convergence to a solution of (2.13), but we do not know a priori how large it should be. Therefore we have to adjust r to a suitable value automatically in the course of solving QP subproblems (2.15). This can be done using information on the optimal Lagrange multipliers of QP subproblems (2.15). We can show that, under some standard assumptions, the penalty parameter remains constant after a finite number of iterations and, moreover, an approximate optimal solution of any desirable accuracy can be obtained for the original problem (2.13) after finitely many steps.

To summarize, the proximal method can be extended in conjunction with descent methods to solve various nonsmooth optimization problems. In particular, the method presented by Fukushima (1987) may be viewed as a natural generalization of the globally convergent successive quadratic programming (SQP) methods for smooth nonlinear programming problems.

References

Auslender, A. (1987), Numerical methods for nondifferentiable convex optimization, Math. Programming Study **30**, pp. 102-126.

Bland, R.G., Goldfarb, D. and Todd, M.J. (1981), The ellipsoid method: A survey, Operations Res. **29**, pp. 1039-1091.

Clarke, F.H. (1983), Optimization and Nonsmooth Analysis, Wiley.

Fletcher, R. (1981), Practical Methods of Optimization: Vol. 2, Wiley.

Fletcher, R. (1982), Second order corrections for nondifferentiable optimization, Numerical Analysis, Dundee 1981, G.A. Watson, ed., Springer, pp. 85-114.

Francis, R.L. and White, J.A. (1974), Facility Layout and Location, Prentice-Hall.

Fukushima, M. (1983), An outer approximation algorithm for solving general convex programs, Operations Res. **31**, pp. 101-113.

Fukushima, M. (1984a), A descent algorithm for nonsmooth convex optimiza tion, *Math. Programming* **30**, pp. 163-175.

Fukushima, M. (1984b), A nonsmooth optimization approach to nonlinear multicommodity network flow problems, *J. Oper. Res. Soc. Japan* **27**, pp. 151-176.

Fukushima, M. (1986a), A relaxed projection method for variational inequalities, *Math. Programming* **35**, pp. 58-70.

Fukushima, M. (1986b), A successive quadratic programming algorithm with global and superlinear convergence properties, *Math. Programming* **35**, pp. 253-264.

Fukushima, M. (1987), A successive quadratic programming method for a class of constrained nonsmooth optimization problems, Technical Report #87024, Dept. of Appl. Math. and Phys., Kyoto University.

Fukushima, M. and Mine, H. (1981), A generalized proximal point algorithm for certain nonconvex minimization problems, *Int. J. Syst. Sci.* **12**, pp. 989-1000.

Kiwiel, K.C. (1985), *Methods of Descent for Nondifferentiable Optimization*, Springer.

Kiwiel, K.C. (1986), A method for minimizing the sum of a convex function and a continuously differentiable function, *J. Opt. Theory Appl.* **48**, pp. 437-449.

Rockafellar, R.T. (1970), *Convex Analysis*, Princeton University Press.

Rockafellar, R.T. (1976), Monotone operators and the proximal point algorithm, *SIAM J. Control. Opt*. **14**, pp. 877-898.

Shor, N.Z. (1985), *Minimization Methods for Nondifferentiable Functions*, Springer.

Womersley, R.S. and Fletcher, R. (1986), An algorithm for composite nonsmooth optimization problems, *J. Opt. Theory Appl.* **48**, pp. 493-521.

Yamakawa, E., Fukushima, M. and Ibaraki, T. (1989), An efficient trust region algorithm for minimizing nondifferentiable composite functions, *SIAM J. Sci. Stat. Comput.* **10**.