Nondifferentiable Optimization: Introduction, Applications and Algorithms

Samir Elhedhli
GERAD/Faculty of Management
McGill University, Montreal CA
samir@management.mcgill.ca

Jean-Louis Goffin
GERAD/Faculty of Management
McGill University, Montreal CA
goffin@management.mcgill.ca

Jean-Philippe Vial
Logilab/ Department of Management Studies
Université de Genève, Geneva, Switzerland
jean-philippe.vial@hec.unige.ch

March, 2000

Les Cahiers du GERAD
G–2000–10

Abstract

This article was written for the *Encyclopedia of Optimization*.

Key Words: Nondifferentiable Optimization, Nonsmooth optimization

Résumé

Cet article a été écrit pour: *Encyclopedia of Optimization*.

1 Introduction

Nondifferentiable, also known as *nonsmooth*, optimization (NDO) is concerned with problems where the smoothness assumption on the functions involved is relaxed. Nondifferentiability means that the gradient does not exist, implying that the function may have kinks or corner points. Consequently, the function cannot be approximated locally by a tangent hyperplane, or by a quadratic approximation. In NDO, the smoothness assumption is usually replaced by weaker ones, which at least guarantee the existence of directional derivatives.

NDO problems arise in a variety of contexts, and methods designed for smooth optimization may fail to solve them. This justifies developing a specialized theory and methods that are the object of this short introduction. In the sequel, we will often refer to convex NDO, a subclass of nondifferentiable optimization, in which functions are further assumed to be convex. Due to its global property, convexity allows stronger convergence results and finer analyses. Yet, the difficulties linked with the presence of kinks remain an important aspect, justifying special interest for this class of problems.

In the following section, we give some basic definitions, then discuss examples of nondifferentiable optimization problems and finally, describe a few different solution techniques.

2 Basic Definitions

The basic nondifferentiable optimization problem takes the form

$$[NDP] \quad \min_{x \in \mathcal{R}^n} f(x) \tag{1}$$

where f is a real valued, continuous, nondifferentiable functions. Convexity of f implies that it has at least one *supporting hyperplane* at every point of $\mathcal{R}^n$. The slopes of such hyperplanes form the set of *subgradients*, which is known as the *subdifferential* set or the *generalized gradient* [7]. At differentiable points there is a unique supporting hyperplane whose slope is the gradient. At nondifferentiable points, there is an infinite set of subgradients and, hence, an infinite set of supporting hyperplanes.

A supporting hyperplane to f at a point x_0 is given by

$$y = f(x_0) + \xi_0^T (x - x_0),$$

where ξ_0 is any element of the subdifferential $\partial f(x_0)$ of f at x_0. Recalling the fact that it is a supporting hyperplane leads to the *subgradient inequality*

$$f(x_0) + \xi_0^T (x - x_0) \leq f(x) \tag{2}$$

Subgradients are defined by this inequality.

Determining the whole subdifferential set is generally an extremely difficult, or impossible, task; if the function f is polyhedral, the number of extreme points of the subdifferential may be exponential in the dimension of the underlying space. A complete description of

the subdifferential can be accomplished for simple situations, such as the one when f is the maximum of a finite number of convex differentiable functions: $f(x) = \max_{i \in I} f_i(x)$. The subdifferential $\partial f(x_0)$ is then given by

$$\begin{aligned} \partial f(x_0) = \quad & \{ \sum_{i \in I(x_0)} \alpha_i \nabla f_i(x_0) : \\ & \sum_{i \in I(x_0)} \alpha_i = 1, \alpha_i \geq 0 \} \\ & I(x_0) = \{ i : f_i(x_0) = f(x_0) \} \end{aligned}$$

When f is a Lipschitz function, the subdifferential set can be defined as being the set of cluster points of the gradients $\nabla f(x_i)$ as sequence of differentiable points x_i approaches x [7]. The precise definition of $\partial f(x_0)$ is given by

$$conv \{ \lim \nabla f(x_i) : x_i \to x_0; \quad \nabla f(x_i) \text{ exists} \}$$

In nondifferentiable optimization, the whole subdifferential set is never calculated. Subgradients are calculated when needed and often a single element suffices. It is common practice to isolate the procedures for calculating subgradients into an *oracle*. The number of calls to the oracle can be a basis for comparing different NDO methods.

A natural solution method in nonsmooth analysis is an iterative method, where a search is done following descent directions. A decent direction is one along which a small movement of f leads to a strict improvement. In other words

$$f'(x_0; d) = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}$$

should be strictly negative. $f'(x_0; d)$ is called the *directional derivative* and it is related to the subgradient through the following relation

$$f'(x_0; d) = \sup \{ \xi^T d : \xi \in \partial f(x_0) \} \tag{3}$$

This relation implies that for d to be a descent direction, $-d$ has to make an acute angle with every subgradient of f at x_0.

3 Sources of NDO Problems

Nonsmooth problems are encountered in many disciplines. In some instances, they occur naturally and in others they result from mathematical transformations.

In statistics for example, rectilinear data fitting, which was long discovered to be superior to the Euclidean one —it has the advantage of overcoming the effect of outliers, [27]— results directly in an NDO problem. Similarly, functions involving ℓ_1 or ℓ_∞ norms, *Euclidean* or *Chebychev* distances, a maximum of convex functions are typical NDO problems. As an example, the ℓ_∞ solution of an over-determined linear system is found by solving the nondifferentiable convex function:

$$\min_{x \in \mathcal{R}^n} \quad ||Ax - b||_\infty \quad = \quad \min_{x \in \mathcal{R}^n} \max_{i=1 \ldots m} \left| a_i^T x - b_i \right|, \tag{4}$$

where $x \in \mathcal{R}^n$, $b \in \mathcal{R}^m$ and $A \in \mathcal{R}^{m \times n}$ with rows a_i^T. This problem can be traced back to the Russian mathematician Chebychev who studied it in the 1850's [27].

Among the mathematical transformations that lead to NDO problems is the technique that changes constrained problems into unconstrained ones through the use of *exact penalty functions* [11] . Equality constraints, $\phi(x) = 0$ and inequality constraints $\varphi(x) \leq 0$ are placed in the objective using penalty parameters and nondifferentiable functions $|\phi(x)|$ and $\max\{0, \varphi(x)\}$ respectively. In other word, a solution to the constrained problem

$$\begin{array}{ll} \min & f(x) \\ s.t. & \phi(x) = 0, \\ & \varphi(x) \leq 0, \end{array} \tag{5}$$

is determined by solving

$$\min f(x) + t_1 |\phi(x)| + t_2 \max\{0, \varphi(x)\}$$

for large enough values of t_1 and t_2.

Still, the major source of optimization problems are master problems resulting from the application of relaxation/restriction techniques such as *Lagrangean relaxation* [13], [12], *Benders decomposition* [4], [14] and *Dantzig-Wolfe decomposition* [10], [9].

These different approaches are conceptually similar, at least in the linear case, and end up solving the same NDO problem. To show that, let us consider the linear program

$$\begin{array}{lll} [LP]: & \min & c^T x \\ & \text{s.t.} & Ax \geq b, \\ & & Dx \geq d, \\ & & x \geq 0. \end{array}$$

Where, we assume for the ease of exposition that $\{x : Ax \geq b; x \geq 0\}$ is a bounded, nonempty polytope. The dual of $[LP]$ is

$$\begin{array}{lll} [LD]: & \max & b^T u + d^T v \\ & \text{s.t.} & A^T u + D^T v \leq c, \\ & & u, v \geq 0. \end{array}$$

Applying Lagrangean relaxation to $[LP]$ is equivalent to relaxing $Dx \geq d$ using positive dual multipliers v, leading to

$$\max_{v \geq 0} \left\{ \min_{x \geq 0} c^T x + v^T (d - Dx) : Ax \geq b \right\}. \tag{6}$$

Benders decomposition applied to $[LD]$ results in

$$\max_{v \geq 0} \left\{ \max_{u \geq 0} b^T u + d^T v : A^T u \leq c - D^T v \right\},$$

where v is assumed to be the complicating variable. Replacing the inside problem by its dual leads to

$$\max_{v\geq 0}\left\{\min_{x\geq 0} c^T x + v^T(d - Dx) : Ax \geq b\right\}. \tag{7}$$

Dantzig-Wolfe decomposition replaces $[LP]$ by its convex representation in terms of the convex points of $\{x : Ax \geq b; x \geq 0\}$ that is indexed by $\mathcal{E}$, to get

$$\begin{array}{ll} \min & \sum\limits_{h\in\mathcal{E}} \alpha_h(c^T x^h) \\ \text{s.t.} & \sum\limits_{h\in\mathcal{E}} \alpha_h(Dx^h) \geq d, \\ & \sum\limits_{h\in\mathcal{E}} \alpha_h = 1, \\ & \alpha_h \geq 0,\ h \in \mathcal{E}. \end{array}$$

Taking the dual results in

$$\begin{array}{ll} \max\limits_{v\geq 0, v_0} & v_0 + d^T v \\ \text{s.t.} & c^T x^h + v^T D x^h \geq v_0 \quad \forall h \in \mathcal{E}, \end{array}$$

which is equivalent to

$$\max_{v\geq 0}\left\{\min_{h\in\mathcal{E}} c^T x^h + v^T(d - Dx^h)\right\} \tag{8}$$

The equivalence between (6) and (7) is obvious. Using the fact that there is always an extreme point solution to a linear program, the equivalence between (6), (7) and (8) is established. Therefore, Lagrangean relaxation applied to the primal is exactly Benders decomposition applied to the dual, and is equivalent by duality to Dantzig-Wolfe decomposition. Furthermore, all three solve (8), which is the maximum of a concave piecewise linear function that is nondifferentiable at intersection points.

Clarke [7] and [8] discusses further examples from physics, engineering, economics and optimal control. Other mathematical problems leading to NDO optimization include semi-infinite programming, eigenvalue optimization and variational inequalities [15].

4 Solution Approaches

Due to the existence of successful solution methods for differentiable optimization, an other solution approach tries to transform nonsmooth problem into smooth ones. As an example, the absolute value function $|x|$, which is nondifferentiable at zero can be approximated by

$$\left\{\begin{array}{cc} -x & x \geq t, \\ \frac{x^2}{t} & -t \leq x \leq t, \\ x & x \geq t, \end{array}\right.$$

for small values of the parameter t. For these transformations to be successful, the right transformation should be found and the the nondifferentiable points should be known. A solution approach based on this transformation is discussed in [22].

Other solution approaches that try to eliminate nondifferentiability, do so by transforming an unconstrained nonsmooth problem into a constrained smooth one. This approach is highly efficient for problems that can be transformed into easily solvable constrained problems such as linear programs. The ℓ_∞ optimization problem described in (4) is equivalent to the linear program

$$\begin{array}{ll} \min & y \\ s.t. & Ax - ye \leq b, \\ & Ax + ye \geq b, \end{array}$$

where e is the appropriate dimension vector whose entries are all ones. Being a linear program with a special type of matrix, most linear programming techniques were modified to solve (4). This includes the simplex-like algorithm of Barrodale and Philips [3] and the interior point algorithms of Ruzinsky and Olsen [24] and Zhang [29].

4.1 Subgradient Methods

The first methods for nondifferentiable optimization tried to extend the gradient-based methods that were successful for smooth optimization. The transition from gradients to subgradients is not straightforward as some subgradient-based search direction are not necessarily improving directions. Wolfe [28] gives an example where the extension of the steepest descent method fails. To overcome that, some designed methods [18], [20] will only take a serious step only when the next iterate is a better one.

Subgradient methods were developed by Shor [25] in 1960's. They are basically an iterative technique where iterates are updated using a current subgradient and a carefully-chosen step size. Applied to (1), iterates are given by

$$x_{k+1} = x_k + t_k \xi_k,$$

where x_k is the current point, ξ_k is a subgradient of f at x_k and t_k is a step size. Shor [25] states that a constant step size does not converge, even for the simple function $|x|$. He proposes the use of a step size that satisfies

$$\sum_{k=0}^{\infty} t_k = \infty, \quad t_k \to 0.$$

In practice, the most widely used step-size is $\theta[f(x_k) - f^*]/||\xi_k||$ where $\theta \in (0, 2]$ and f^* is the best estimate of the optimal value $f(x^*)$.

5 Steepest descent and ϵ-subgradient methods

Subgradient methods are not monotonic, as they do not guarantee to improve the value of the minimized function. Descent methods are designed to overcome this drawback. As

an example we discuss the steepest descent method which chooses its search direction by solving

$$\min_{||d|| \leq 1} f'(x; d).$$

Using relation (3), the steepest descent direction, at a point x_k, is given by

$$d_k = \frac{\xi_k}{||\xi_k||}; \xi_k = \arg \max_{\xi \in \partial f(x_0)} ||\xi||.$$

The method proceeds iteratively, updating the iterates by

$$x_{k+1} = x_k + \alpha_k \xi_k$$

and choosing the step length α_k so that $f(x_{k+1}) < f(x_k)$.

The main difficulty with the steepest descent resides in the calculation of the direction d_k which necessitates the knowledge of the whole differential set $\partial f(x)$. To overcome that, ϵ-*subgradient* methods prefer to calculate approximate steepest descent direction by searching through subgradients of neighboring points through the use of the ϵ-*subdifferential* set

$$\partial_\epsilon f(x) = \{\xi : f(x_0) + \xi(x - x_0) + \epsilon \leq f(x), \forall x\}.$$

Details of the method can be found in [5].

5.1 Cutting Plane Methods

Kelley [17] and Cheney and Goldstein [6] were the first to realize the potential of such methods for convex programming. Applied to (1), cutting plane algorithms use the subgradient inequality to approximate f by

$$f(x) \quad \cong \quad \max_{i \in I} f(x_i) + \xi_i^T (x - x_i).$$

Where $\xi_i^f, i \in I$ are subgradients of f at $x_i, i \in I$. Thus, $[NDP]$ is replaced by,

$$\min_x \left\{ \max_{i \in I} f(x_i) + \xi_i^T (x - x_i) \right\},$$

which is equivalent to,

$$\begin{array}{ll} \min & v \\ s.t. & f(x_i) + \xi_i^T (x - x_i) \leq v \quad \forall i \in I \end{array} \tag{9}$$

Problem [9] is a linear program that is easier to deal with than the original problem. It is to note, however, that this is only an approximation of $[NDP]$, which gets better as more constraints are added. Let us denote by $[MP_k]$ the relaxed master problem [9] with index set I_k.

By transforming $[NDP]$ to [9], a nondifferentiable problem is replaced by a constrained problem having a large number of constraints. Cutting plane methods use only a subset of these constraints and generate the rest as needed. In fact, they would solve a series of relaxed master problems $[MP_k]$ and stop when an optimal (satisfactory) solution to $[NDP]$ is reached.

Various cutting plane methods were proposed over the years. Each variant generates cuts at a different point called the *query point*. *Kelley's classical cutting plane* [17] method chooses the minimum of the relaxed master problem $[MP_k]$ as a query point. Although, it may work well for some problems, this method suffers from slow convergence [23]. The analytic centre cutting plane method (*ACCPM*) [16], [15], on the other hand, chooses the *analytic centre* as its query point. Its calculation makes use of interior point concepts and has shown promising results for a number of applications [1], [2]. *Bundle methods* [19], [20], choose the query point by solving a quadratic program that contains a small number of cutting planes. The information (bundle of cutting planes) is updated regularly and kept moderately small.

6 Conclusion

Nondifferentiable optimization tackles a class of problems that are intractable to classical optimization methods. Most of the theory is based on the notion of subgradients and most of the work is done for the convex case. It has an abundance of applications in real life, because the nondifferentiability aspect captures some of the inherent complexity in real-life problems. Like all disciplines, favoring an easily implementable and understood method will not necessarily lead to a good solution method. This corresponds to the subgradient method in NDO. Although it is easily implementable, it has slow convergence. More sophisticated methods, such as Bundle or ACCPM are more promising from a computational point of view but require more know-how of the method and of the numerical linear algebra.

References

[1] O. Bahn, J. L. Goffin, J. P. Vial, and O. Du Merle. "Implementation and behavior of an interior point cutting plane algorithm for convex programming : An application to geometric programming." *Discrete Applied Mathematics***49** (1994), 3–23.

[2] O. Bahn, O. Du Merle J. L. Goffin and J. P. Vial. "A cutting plane method from analytic centers for stochastic programming." *Mathematical Programming, Series B,* **69** (1995) 45–73 editors J.-L. Goffin and J.-P. Vial.

[3] I. Barrodale and C. Phillips "An improved algorithm for discrete Chebychev linear approximation". *In Proceedings of the Fourth Manitoba conference on Numerical Mathematics, University of Manitoba, Winnipeg, Canada* (1974).

[4] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems", *Numerische Mathematik* **4** (1962), 238-252.

[5] D. P. Bertsikas, "Nonlinear Programming" *Athena Scientific* **1995.**

[6] W. Cheney and A.A. Goldstein "Newton's method for convex programming and Chebyshev approximation", *Numerische Mathematik* **1-5** (1959), 253-268.

[7] F. H. Clarke, "Optimization and Nonsmooth Analysis ", *Les publications CRM, Montreal* (1989).

[8] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern and P. R. Wolenski, "Nonsmooth analysis and control theory" *Graduate texts in mathematics* Springer-Verlag (1998).

[9] G. B. Dantzig and P. Wolfe, "The decomposition algorithm for linear programming", *Econometrica* **29** (4), (1961), 767-778.

[10] G. B. Dantzig and P. Wolfe, "Decomposition principle for linear programs", *Operations Research*, **8**, (1960) 101-111.

[11] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming : Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, New York, 1968. Reprint : Volume 4 of *SIAM Classics in Applied Mathematics*, SIAM Publications, Philadelphia, PA 19104–2688, USA, 1990.

[12] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problems", *Management Science* **27** (1981) 1-18.

[13] A. M. Geoffrion, "Lagrangean relaxation for integer programming", *Mathematical Programming Study*, **2** (1974) 82-114.

[14] A. M. Geoffrion, "Generalized Benders decomposition", *Journal of Optimization Theory and Applications* **10** (1972) 237-260.

[15] J.-L. Goffin and J.-P. Vial, "Interior point methods for nondifferentiable optimization", Operations Research Proceedings 1997 P. Kishka, H.W. Lorenz, U. Derigs, W. Domschke, P. Kleinschmidt, R. Moehring, editors, Springer-Verlag, Berlin, (1998), 35–49.

[16] J.-L. Goffin, A. Haurie, and J.-P. Vial. "Decomposition and nondifferentiable optimization with the projective algorithm." *Management Science*, **38(2)**,(1992), 284-302.

[17] J. E. Kelley, "The cutting plane method for solving convex programs", *Journal of the SIAM* **8** (1960), 703-712.

[18] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, lecture Notes in Mathematics 1133, Springer-Verlag, 1985.

[19] K. C. Kiwiel, "Proximity control in Bundle methods for convex nondifferentiable optimization", *Mathematical Programming* **46** (1990), 105-122.

[20] C. Lemaréchal, "Bundle Methods in Nonsmooth Optimization", in *Nonsmooth Optimization*, Proceedings of the IIASA Workshop March 28 - April 8,1977, Lemaréchal, C., and Mifflin, R., eds. Pergamon Press, (1978).

[21] C. Lemaréchal, A. Nemirovskii and Y. Nesterov, "New variants of bundle methods " , in Nonsmooth Optimization, *Mathematical Programming* **69** (1995), 111–147.

[22] K. Madsen, H. B. Nielsen and M.Ç. Pınar, " New characterizations of ℓ_1 solutions of over-determined linear systems" *Operations Research Letters*, **16** (1993).

[23] A. S. Nemirovskii and D. B. Yudin, *Problem complexity and method efficiency in optimization*, John Wiley, Chichester (1983).

[24] S. A. Ruzinsky and E. T. Olsen, " ℓ_1 and ℓ_∞ minimization via a variant of Karmarkar's algorithm" *IEEE Transactions on Acoustics Speech and Signal Processing*, **37** (1989)

[25] N. Z. Shor, *subgradient methods: A survey of soviet research Nonsmooth optimization: Proceedings of the IIASA workshop March 28–April 8,1977* C. Lemaréchal and R. Mifflin eds. Pergamon Press 1978.

[26] N. Z. Shor, *Minimization Methods for Non-differentiable Functions* (in Russian), Naukova Dumka, Kiev, 1979 [English translation: Springer, Berlin, 1985].

[27] G. A. Watson *Approximation theory and numerical methods*, John Wiley, New York **1980**.

[28] P. Wolfe. "A method of conjugate subgradients for minimizing nondifferentiable functions," *Mathematical programming study*, **3** (1975) 145–173.

[29] Y. Zhang, "Primal-dual interior point approach for computing the ℓ_1-solutions and ℓ_∞-solutions of over-determined linear systems" *Journal of Optimization Theory and Applications*, **77:2** (1993).