Subgradient Optimization in Nonsmooth Optimization (including the Soviet Revolution)

Jean-Louis Goffin

2010 Mathematics Subject Classification: 26A27, 46N10
Keywords and Phrases: Nondifferentiable optimization, nonsmooth optimization, subgradient optimization, relaxation method, Soviet revolution

1 Introduction

Convex nondifferentiable, also known as convex nonsmooth, optimization (NDO) looks at problems where the functions involved are not continuously differentiable. The gradient does not exist, implying that the function may have kinks or corner points, and thus cannot be approximated locally by a tangent hyperplane, or by a quadratic approximation. Directional derivatives still exist because of the convexity property.

NDO problems are widespread, often resulting from reformulations of smooth, or linear problems, that are formulated in a space with much smaller number of variables than in the original problem. Examples of this are the reformulation implicit in Dantzig-Wolfe decomposition or column generation [4] and [5], which are equivalent by duality to Cheney's cutting plane method [20]. These methods do not work well if an aggregated formulation is used. Shor's subgradient method [35, 36] provided a superior alternative, leading to a true Soviet revolution. His work was expanded both in theory and in practice by numerous authors. Held and Karp [17], unaware of the work of Shor, developed a method for the traveling salesman problem that uses subgradient optimization to compute a bound in a Lagrangean relaxation scheme. This seminal contribution also led to a huge following; see for instance Fisher [11].

2 Basic definitions

The basic nondifferentiable optimization problem takes the form

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

where f is a real-valued, continuous, convex, and nondifferentiable function. Sometimes there is a restriction that $x \in X$, a closed convex set, for which a projection map is available:

$$x^*(x) = \Pi_X(x) = \big\{\bar{x} : \|\bar{x} - x\| \le \|y - x\|,\ \forall y \in X\big\};$$

and the problem becomes:

$$[NDOc] \quad \min_{x \in X} f(x).$$

The convexity of f implies that it has at least one supporting hyperplane at every point of $\mathcal{R}^n$. The subdifferential is the set of such slopes, i.e.,

$$\partial f(x) = \big\{\xi : f(x) + \langle \xi, (y - x)\rangle \le f(y),\ \forall y \in \mathcal{R}^n\big\}.$$

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.

The derivative in the direction d is given by:

$$f'(x; d) = \sup\big\{\xi^T d : \xi \in \partial f(x)\big\}$$

and the direction of steepest descent is given by d^*:

$$\min_{\|d\|=1} f'(x; d) = f'(x; d^*);$$

it can be shown that if $0 \notin \partial f(x)$ and $\hat{d}$ is the element of minimum norm in the subdifferential $\partial f(x)$, then

$$d^* = -\frac{\hat{d}}{\|\hat{d}\|}.$$

The use of the steepest descent method *with exact line searches* is not recommended as:

1. The steepest descent method with exact line searches may converge to a nonoptimum point, see Wolfe [43];
2. In the frequent case where $f(x) = \max_{i \in I}\{\langle a_i, x\rangle + b_i\}$, and the set I is computed by an oracle or subroutine, an LP or an IP, the cardinality of I may be exponential, and the subdifferential is given by:

$$\begin{aligned}\partial f(x) = &\Big\{ \sum_{i \in I(x)} \alpha_i a_i : \\ &\sum_{i \in I(x)} \alpha_i = 1, \alpha_i \ge 0\Big\}, \\ &I(x) = \big\{i : \langle a_i, x\rangle + b_i = f(x)\big\};\end{aligned}$$

so that it is unrealistic to expect that the full subdifferential will be available.

In NDO, one assumes that the function f is given by an oracle which for every value of x returns the value of f, i.e., $f(x)$, and one arbitrary subgradient $\xi(x) \in \partial f(x)$.

3 Subgradient methods: The Soviet revolution

Subgradient methods were developed by Shor [35] and [36] in the 1960's.

To quote from a paper by B. T. Polyak [33] delivered at the Task Force on nondifferentiable optimization organized at IIASA by Lemaréchal and Mifflin, (this paper also includes an excellent bibliography of work done in the USSR before 1977):

> The subgradient method was developed in 1962 by N.Z. Shor and used by him for solving large-scale transportation problems of linear programming [35]. Altough published in a low-circulation publication, this pioneering work became widely known to experts in the optimization area in the USSR. Also of great importance for the propagation of nondifferentiable concepts were the reports by the same author presented in a number of conferences in 1962–1966.
>
> Publication of papers by Ermoliev [9], Polyak [30] and Ermoliev and Shor [10] giving a precise statement of the method and its convergence theorems may be regarded as the culmination of the first stage in developing subgradient techniques.

All of their massive contributions to the field are well reported in their two books Shor[40] and Polyak[32], as well as in the second book by Shor[41]; see also the book by Nesterov [27].

So subgradient optimization simply moves the current iterate in the direction of a scaled subgradient by a stepsize that is decided a prori:

$$x_{k+1} = \Pi_X \left(x_k - t_k \frac{\xi_k}{\|\xi_k\|} \right),$$

where x_k is the current point, $\xi_k \in \partial f(x_k)$ is an arbitrary subgradient of f at x_k, t_k is a stepsize and Π_X is the projection map on the constraint set X. It is assumed that the projection map is easily computed, such as if X is a sphere, a box or a simplex. A subgradient is not a direction of descent for the function f but it is one for the distance to the optimal set.

Shor [35] states that a constant stepsize $t_k = t$ does not converge, as the example of $f(x) = |x|$ clearly shows. He also shows that the iterates eventually reach an $O(t)$ neighborhood of the optimum.

This follows from an equivalent proof, extended to the case of a constraint set:

Theorem 3.1 (Nesterov [27]). *Let f be Lipschitz continuous on $B_2(x^*, R)$ with constant M and $x_0 \in B_2(x^*, R)$. Then*

$$f_k^* - f^* \leq M \frac{R^2 + \sum_{i=0}^k h_i^2}{2 \sum_{i=0}^k h_i}. \tag{1}$$

In this statement $f_k^* = \min_{i=0}^k f(x_k)$ and $f^* = \min_{x \in X} f(x)$.

It follows that if the sequence t_k is chosen as $t_k = R\epsilon, \forall k = 1, \ldots, N$, and $N = \lceil \frac{1}{\epsilon^2} \rceil$ then: $f_N^* - f^* \leq MR\epsilon$; see also Shor [40] pp. 23–24.

This means that subgradient optimization is an optimal algorithm, uniformly in the dimension of the problem, see Nemirovski and Yudin [25]. Almost quoting from Polyak again [33]:

> Reference [35] has described the following way of stepsize regulation resting upon this result, although it is not entirely formalized. A certain ϵ is chosen and the computation proceeds with $t_k = R\epsilon$ until the values of $f(x_k)$ start to oscillate about a certain limit. After this ϵ is halved and the process is repeated.

This leads readily to the divergent series of stepsizes, suggested by Polyak [30] and Ermoliev[9], and studied in Shor and Ermoliev [10]:

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

Theorem 3.2. *Theorem 3.1 shows that f_k^* converges to f^*.*

An often used stepsize is $t_k = \frac{R}{\sqrt{k+1}}$, which guarantees convergence in $O^*(\frac{1}{\sqrt{k+1}})$ steps [27], where O^* means the term of higher order, ignoring lower order terms; the proof of this can be improved, see Nemirovski [26], who shows that $\varepsilon_N \leq O(1)\frac{RM}{\sqrt{N}}$, where $\varepsilon_N = f_N^* - f^*$.

Unfortunately, the divergent stepsize rule can and is extremely slow. So the question arose, as to whether geometric convergence can be obtained.

The answer is given in the following theorem, proved only in the unconstrained case:

Theorem 3.3 (Shor [40] pp. 30–31). *Let f be a convex function defined on $\mathcal{R}^n$. Assume that for some φ satisfying $0 \leq \varphi < \pi/2$, and for all $x \in \mathcal{R}^n$ the following inequality holds:*

$$\langle \xi(x), x - x^*(x) \rangle \geq \cos\varphi \|\xi(x)\| \, \|x - x^*(x)\|, \tag{2}$$

where $\xi(x) \in \partial f(x)$, and $x^*(x)$ is the point in the set of minima that is nearest to x. If for a given x_0 we choose a stepsize t_1 satisfying:

$$t_1 \geq \begin{cases} \|x^*(x_0) - x_0\| \cos\varphi & \text{for} \quad \pi/4 \leq \varphi < \pi/2 \\ \|x^*(x_0) - x_0\|/(2\cos\varphi) & \text{for} \quad 0 \leq \varphi < \pi/4, \end{cases}$$

define $\{t_k\}_{k=1}^{\infty}$ by

$$t_{k+1} = t_k r(\varphi), \quad k+1, \ldots, \infty$$

where

$$r(\varphi) = \left\{ \begin{array}{ll} \sin\varphi & \text{for} \quad \pi/4 \leq \varphi < \pi/2 \\ 1/(2\cos\varphi) & \text{for} \quad 0 \leq \varphi < \pi/4 \end{array} \right\},$$

and generate $\{x_k\}_{k=0}^{\infty}$ *according to the formula*

$$x_{k+1} = x_k - t_{k+1} \frac{\xi(x_k)}{\|\xi(x_k)\|}.$$

Then either $\xi(x_k^*) = 0$ *for some* k^*, *i.e.,* x_k^* *is a minimum point, or for all* $k = 1, \ldots, \infty$ *the following inequality holds*

$$\|x_k - x^*(x_k)\| \begin{cases} t_{k+1}/\cos\varphi & \text{for } \pi/4 \leq \varphi < \pi/2 \\ 2t_{k+1}\cos\varphi & \text{for } 0 \leq \varphi < \pi/4 \end{cases}$$

This theorem was first proved in this form by Shor and Gamburd [38] and by Shor [39]. An earlier version that used the asphericity σ of the level set of f instead of $\cos\varphi$ was proved by Shor [37]. This is a slightly weaker result as $\cos\varphi \geq 1/\sigma$.

In practice, a most widely used stepsize is $t_k = \lambda(f(x_k) - \bar{f})/||\xi_k||$ where $\lambda \in (0, 2)$ and $\bar{f}$ is expected to be a good estimate of the optimal value $f(x^*)$. It can be either the exact optimum f^*, an overestimate $\hat{f} > f^*$, or an underestimate $\check{f} < f^*$. This was suggested and studied by Polyak, see for instance [32].

The most general theorem is due to Nemirovski [26], under the assumption that $\bar{f} = f^*$:

$$\varepsilon_N \leq M\|x_0 - x^*\|N^{-1/2}.$$

Polyak [31], see also Shor [40] shows that if in addition to the Lipschitz condition on f one has a lower bound on the variation of f such as

$$f(x) \geq md(x, X^*)^\alpha$$

where $d(x, X^*)$ is the distance to the optimal set X^* and $\alpha = 1$ or 2 then:

$$\|x_k - x^*\| \leq q^k\|x_0 - x^*\|,$$

where $q = \sqrt{1 - \lambda(2-\lambda)\frac{m^2}{M^2}}$.

The more practical case of $\bar{f} < f^*$, as an underestimate of f^*, can be computed by getting a feasible dual solution, was studied by Eremin [6, 7, 8] who studied the Chebyshev solution to an infeasible system of linear inequalities:

$$P = \big\{x : \langle a_i, x\rangle + b_i \leq 0, \quad \forall i \in I\big\}.$$

This is equivalent to minimizing the function $f(x) = \max_{i\in I}\{< a_i, x > +b_i\}$, where $f^* > 0$, and taking the stepsize $t_k = \lambda_k f(x_k)/\|\xi_k\|$. He shows convergence of $(x_k)_{k=1,\ldots,\infty}$ to a point in X^* if $(\lambda_k)\ _{k=0,\ldots,\infty} > 0$ is a divergent series that converges to 0.

From a practical point of view subgradient optimization has solved quite successfully a wide range of problems. This means that many problems are quite surprisingly well conditioned. Subgradient optimization fails miserably on ill conditioned problems such as highly nonlinear multicommodity flow problems.

4 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.

The most complete reference on NDO problems is Chapter 5 of Shor's book [40]. In Shor original work [35], he mentions solving the transportation problem using subgradient optimization.

A standard transportation problem is a special case of an NDO that occurs when optimizing the Lagrangean dual of a constrained optimization problem:

$$\begin{array}{ll} \min & \langle c, y\rangle \\ s.t. & Ay \geq b \\ & By \geq d \end{array}$$

Dualizing the first set of constraints, with dual variables x, one gets the partial dual:

$$f(x) = \max_{x\geq 0}(\langle b, x\rangle + \min_{y\in Y}\langle c - A^T x, y\rangle),$$

where $Y = \{y : By \geq d\}$ is a polyhedron, assumed to be compact, and with a set of extreme points given by$\{y^i : i \in I\}$.

One subgradient is thus any $b - Ay^{i(x)}$ where $y^{i(x)}$ is a minimizer of $\min_{y\in Y} < c - A^T x, y >$. The formulation with an objective variable:

$$\begin{array}{ll} \min & \langle b, x\rangle + w \\ s.t. & w \leq \langle c - A^T x, y^i\rangle \forall i \in I \end{array}$$

is the dual of the extended form of the Dantzig-Wolfe decomposition reformulation.

5 Other contributions

The seminal contribution by Held and Karp [17] on the traveling salesman problem introduced Lagrangean relaxation and the solution of the partial Lagrangean dual by subgradient optimization. They were not aware at that time of the Soviet revolution in this field, so they developed subgradient optimization from scratch. The symmetric traveling-salesman problem seeks to find a minimum cost tour in a complete undirected graph. A minimum tour k^* can be shown to be a 1-tour k with the added constraint that every node has degree 2. A 1-tree consists of a tree on the vertex set $\{2, 3, \ldots, n\}$, together with two distinct edges at vertex 1. Therefore a formulation of the TSP is:

$$\begin{array}{ll} \min_k & c_k \\ s.t.: & d_{i,k} = 2 \end{array}$$

and $d_{i,k}$ is the degree of vertex i in the k^{th} 1-tree, and c_k is the cost of the 1-tree. *Dualizing* the degree constraints with *multipliers* π_k leads to:

$$f(\pi) = \min_k \left\{ c_k + \sum_{i=1}^{n} (d_{i,k} - 2)\pi_i \right\}$$

The cost of a minimum cost tour C^* is greater than or equal to $\max_\pi f(\pi)$, which provides a lower bound on C^*. The computation of $f(\pi)$ and a subgradient ξ involves the computation of a minimum cost 1-tree which can be done in $O(n)$ steps. This formulation can be solved by the dual of Dantzig-Wolfe decomposition; this method shows the long tail typical of DW when no disaggregation is available, as seems the case here. Held and Karp [17] suggested the use of subgradient optimization, i.e.,

$$\pi^{m+1} = \pi^m + t_m \xi^m,$$

and proved a result analogous to Shor's [35], with a constant $t_m = \bar{t}$ and convergence to within $O(\bar{t})$ of the optimum is achieved. The solution of the TSP by branch and bound, using the bound computed here, was extremely successful, and led the authors to claim that:

> In fact, this experience with the traveling-salesman problem indicates that some form of the relaxation method may be superior to the simplex method for linear programs including a very large number of inequalities.

The authors sought the wisdom of Alan Hoffman, who advised them that the method they just developed was closely related to the relaxation method for linear inequalities due to Agmon [1], and Motzkin and Schoengerg [23]. The relaxation method attempts to solve a system of linear inequalities $\{x : \langle a_i, x\rangle + b_i \leq 0 : i \in I\}$ by projecting, in the case of Agmon, or reflecting in the case of Motzkin and Schoenberg on the most distant inequality. This amounts to minimizing the convex function

$$f(x) = \max\left\{0, \max_{i\in I}\left\{\frac{< a_i, x > +b_i}{\|a_i\|}\right\}\right\},$$

by using what became known as subgradient optimization with a stepsize that uses the information that $f^* = 0$. The algorithm is thus $x_{k+1} = x_k + \lambda_k \xi_k$, where

$$\xi_k = \frac{a_{\bar{i}}}{\|a_{\bar{i}}\|},$$

with $\bar{i}$ one of the indices that satisfies $\frac{\langle a_i, x\rangle + b_i}{\|a_i\|} = f(x)$.

Agmon [1] showed that for $\lambda = 1$ the convergence to a feasible point $x^* \in P = \{x : f(x) = 0\}$ is geometric at a rate $\sqrt{1 - \mu^{*2}}$, unless finite convergence occurs. Motzkin and Schoenberg [23] showed that if P is full-dimensional, finite convergence occurs if $\lambda = 2$. It was shown by the author [14] that Agmon's definition of μ^* can be written as $\mu^* = \inf_{x \notin P} f(x)/d(x, P)$, where $d(x, P)$ is the distance from x to P. It can also be shown [14] that $\mu^* = \cos\varphi$ as defined by Shor and Gamburd in Theorem 3.3.

The works by Agmon and Motzkin and Schoenberg may be viewed as a precursors to the Soviet revolution.

The successful solution of the traveling-salesman problem by computing bounds using subgradient optimization led to a true explosion of works in Lagrangean relaxation in the West; for example Fisher [11] and the many references therein.

Karp, who was my thesis adviser, asked me to read the Held and Karp [17] paper as well as the ones by Agmon [1] and Motzkin and Schoenberg [23], and apply subgradient optimization to the transportation problem, and see if something could be done to explain the success of subgradient optimization. He also mentioned that the simplex method when applied to a "normallly" formulated system of equalities converges in a number of iterations which is a small multiple of the number of constraints, but that in the case where the number of variables is exponential, as in Dantzig-Wolfe decomposition, this estimate does not hold, thus requiring another solution technique. I engaged in a thorough review of the Soviet literature, and found the works of Eremin and Polyak, but missed the huge contributions by Shor.

My 1971 thesis, published later as Goffin [12], has the following result, extending Motzkin and Schoenberg: the relaxation method converges finitely to a point $x^* \in P$, where P is assumed to be full dimensional, if

$$\lambda \in [1, 2] \quad \text{if } P \text{ is obtuse}$$
$$\lambda \in \Big[\frac{2}{1 + 2\nu(P)\sqrt{1 - \nu^2(P)}}, 2\Big], \quad \text{if } \nu(P) < \sqrt{2}/2,$$

where the condition number $\nu(P)$ equals the minimum over all tangent cones to P of the sine of the half aperture of the largest spherical cone included in a tangent cone. It is easy to show that $\mu^* \geq \nu(P)$, and that if the constraints defining every tangent cone are linearly independent then $\mu^* = \nu(P)$.

Unfortunately, both $\nu(P)$ and μ^* are not polynomial, showing that the relaxation method is not a polynomial algorithm; see, for instance, Todd [42]. An unpublished result by the author shows that if $\{a_i : i \in I\}$ forms a totally unimodular matrix, then $\nu(P) \geq 1/n$.

The author then extended this convergence theory to subgradient optimization [13], and at the IIASA meeting in 1977, B. T. Polyak mentioned the work by Shor and Gamburd [38], and helped translate it, showing that this author's results were essentially identical to that work. A very nice extension of the geometric convergence to the case of functional constraints has been published by Rosenberg [34], extending also results by Polyak [30].

A thorough study of subgradient optimization and its applications was performed by Held, Wolfe and Crowder [18]. They cite Polyak [30, 31] and

Shor [36]. As stepsize they use an underestimate $\bar{f}$ of the function minimum $f^* = \min_{x\in X} f(x)$ and use the Agmon relaxation step for an infeasible system:

$$x_{k+1} = \Pi_X\Big(x_k - \lambda_k \frac{f(x_k) - \bar{f}}{\|\xi_k\|^2}\xi_k\Big) \tag{3}$$

where $\xi_k \in \partial f(x_k)$. Paraphrasing from the Held et al. [18] paper on the "Validation of Subgradient Optimization": We observed that the results did not seem to depend critically on the exact value of $\bar{f}$. Of course it is necessary that the stepsize converges to 0, which we will not accomplish, with an underestimate $\bar{f}$, unless we choose a sequence λ_k which tends to zero. Generally (but not always) a good rule is to set $\lambda = 2$ for $2n$ iterations (where n is a measure of the problem size), and then successively halve both the value of λ and the number of iterations until the number of iterations reaches some threshold z. λ is then halved every z iterations until the resulting λ_k is sufficiently small. It is thus possible to converge to a point not in the optimal set, altough in our work that almost never happened. We would particularly point out choice of stepsize as an area which is imperfectly understood.

The answers provided to that question did not appear in the works of Shor [40] or Polyak [31], who prove rather weak results. The following result which extends [12] for Part 1 and Eremin [6, 7] for Part 2 appears in Allen et al. [2]:

Theorem 5.1. *In algorithm (3),*
1. *given* $\delta > 0$ *and* $0 < \lambda_k = \lambda < 2$, *there is some* K *such that*

$$f(x_K) \leq f^* + \big(\lambda/(2-\lambda)\big)(f^* - \bar{f}) + \delta;$$

2. *if* $\sum_{k=1}^{\infty} \lambda_k = \infty$, *and* $\lambda_k \to 0$, *then* $f_K^* = \min_{k=1}^{K} f(x_k)$ *converges to* f^*.

This shows that the strategy of using $\lambda_k \to 0$ is the correct one. The stepsize chosen by Held et al. [18] was, towards the end of the sequence, a halving of λ at each five iterations. This is equivalent to $r(\varphi) = (\frac{1}{2})^{1/5} \cong .85$, where $r(\varphi)$ is defined in Shor's theorem (3.3), assuming that Shor's result of (3.3) applies in this case, which nobody has proven, but which seems quite likely to be provable.

Held et al. [18] experimented with great success on a variety of problems, including the assignment problem, the multicommodity flow problems and the TSP, concluding:

> Briefly, we think that subgradient optimization holds promise for alleviating some of the computational difficulties of large-scale optimization. It is no panacea, though, and needs careful work to make it effective, but its basic simplicity and its wide range of applicability indicate that it deserves to be more widely studied.

Further developments include:

1. An updating procedure for the target $\bar{f}$ which can be either an overestimate $\bar{f} > f^*$ or an underestimate $\bar{f} < f^*$, which now becomes a variable $\bar{f}_k$ to be adjusted depending on the behaviour of the sequence $f(x_k)$. Both Ahn et al. [21] and [15] show an updating rule for $\bar{f}_k$ that guarantees that $f_\infty = \inf_k f(x_k) = f^*$.
2. The computation of the primal variables y in section 4 can be done in the limit. This was shown by Shor [40] pp. 117–118 and improved by Anstreicher and Wolsey [3] and Nesterov [28]. Define the subgradient optimization by the recursive relation:

$$x_{k+1} = \Pi_X(x_k - t_k \xi_k),$$

and the convex combination

$$\bar{t}_i^k = \frac{t_i}{\sum_{j=1}^k t_j}.$$

Then the sequence defined by

$$\bar{y}_k = \sum_{i=1}^k \bar{t}_i^k y^i$$

has the following properties
Theorem 5.2. *Let the sequence x_k in the problem of section 4 be generated according to the formulae above, and*

$$t_i \to 0, \quad \sum_{i=1}^\infty t_i = \infty, \text{ and } \sum_{i=1}^\infty t_i^2 < \infty.$$

Then $x_k \to x^* \in X^*$, and any accumulation point of $\bar{y}_k$ is in the optimal set Y^*.
3. Nedic and Berstsekas [24] showed how to use the disaggregation structure, often available in problems obtained from Dantzig-Wolfe decomposition, by introducing an incremental subgradient method that cycles between the subgradients of the individual functions.
4. A recent paper by Nesterov [29] shows how to use subgradient optimization successfully on huge-scale problems, by using sparse updates of the subgradient, leading to excellent computational results.

6 Conclusions

From my doctoral thesis:

"To simplex, to relax: This thesis' question
Whether 'tis faster on P to iterate
On the narrowing edge slung between vertices

Or to take the normal against a sea of planes
And by opposing it, to leap to end today."[1]

Silly and somewhat arrogantly optimistic. But as we have seen in this journey, subgradient optimization outperforms the simplex method in many instances. When it is good it's very good, but when it is bad it is very bad, as is the case of ill-conditioned problems, or in the terminolgy of Shor, gully shaped functions. This has given rise to a set of more complex methods that deal well with ill conditioned problems. Among them are:

1. The r-algorithm due to Shor [40], which introduces a variable metric on top of the subgradient; it worked quite well with a heuristic choice of pararmeters, until a theoretically selected choice of the parameters by Yudin and Nemirovksi [25] led to the ellipsoid method and its deep theoretical significance
2. The mirror descent method of Yudin and Nemirovski [25]
3. The bundle method developed by Lemaréchal and Kiwiel and many others, about which a chapter appears in this book by Mifflin and Sagastizabal [22]
4. The analytic center cutting plane method by Goffin and Vial [16]

Acknowledgments. The author's research has been funded by the Natural Research Council in Science and Engineering of Canada for 39 years. I sincerely apologize to the many friends whose work I could not cite.

References

[1] S. Agmon, "The Relaxation Method for Linear Inequalities", *Canadian Journal of Mathematics*, 6, 1954, 382–392.

[2] E. Allen, R. Helgason and J. Kennigton, " A Generalization of Polyak's Convergence Result for Subgradient Optimization", *Mathematical Programming*, 37, 1987, 309–317.

[3] K.M Anstreicher and L.A. Wolsey, "Two 'Well-Known' properties of Subgradient Optimization", *Mathematical Programming*, Ser. B 2009 120:213–220.

[4] G. B. Dantzig and P. Wolfe, "The Decomposition Algorithm for Linear Programming", *Econometrica* 29 (4), (1961), 767–778.

[5] G. B. Dantzig and P. Wolfe, "Decomposition Principle for Llinear Programs", *Operations Research*, 8, (1960) 101–111.

[6] I.I. Eremin, "Incompatible Solutions of Linear Inequalities", *Soviet Mathematics Doklady*, 2, 1961, 821–824.

[1]The simplex method referred here is the one applied to a problem with an exponential number of hyperplanes. On normally formulated linear programs, A. Hoffman et al. [19] showed that the simplex method is vastly superior to the relaxation method.

[7] I.I. Eremin, “An Iterative Method for Cebysev Approximation of Incompatible Solutions of Linear Inequalities”, *Soviet Mathematics Doklady*, 3, 1962, 570–572.

[8] I.I. Eremin, “A Generalization of the Motzkin-Agmon Relaxation Method”, *Uspekhi Mathematicheski Nauk*, 20, 1965, 183–187.

[9] Yu.M Ermoliev: M. “Methods of Solutions of Nonlinear Extremal Problems”, *Cybernetics* 2,4, 1–16.

[10] Yu.M. Ermoliev and N.Z. Shor, “On the Minimization of Nondifferentable Functions”, *Cybernetics*, 3, 1, 72.

[11] M. L. Fisher, “The Lagrangian relaxation method for solving integer programming problems”, *Management Science* 27 (1981) 1–18.

[12] J.L. Goffin: “The Relaxation Method for Solving Systems of Linear Inequalities”, *Mathematics of Operations Research*, 5,3 1980, 388–414.

[13] J.L. Goffin, “On Convergence Rates of Subgradient Optimization Methods”, *Mathematical Programming*, 13, 1977, 329–347.

[14] J.L. Goffin, “Nondifferentiable Optimization and the Relaxation Method”, *Nonsmooth optimization: Proceedings of the IIASA workshop March 28–April 8, 1977* C. Lemaréchal and R. Mifflin eds. Pergamon Press 1978, 31–50.

[15] J.L. Goffin and K.C. Kiwiel, “ Convergence of a Simple Subgradient method”, *Mathematical Programming*, 85, 1999, 207–211.

[16] J.L. Goffin and J.P. Vial, “Convex Nondifferentiable Optimization: a Survey Focused on the Analytic Center cutting Plane Method”, *Optimization Methods and Software*, 17, 2002, 805–867.

[17] M. Held and R.M. Karp, “The Traveling-Salesman Problem and Minimum Spanning Trees:Part II”, *Mathematical Programming* 1, 1971, 6–25.

[18] M. Held, P. Wolfe and H.P. Crowder, “Validation of Subgradient Optimization”, *Mathematical Programming*, 6, 1974, 62–88.

[19] A. Hoffman, M. Mannos, D. Sokolovsky and N. Wiegmann, “Computational Experience in Solving Linear Programs”, *Journal of the SIAM*, Vol. 1, No. 1 Sep., 1953.

[20] J. E. Kelley, “The cutting plane method for solving convex programs”, *Journal of the SIAM* 8 (1960), 703–712.

[21] S. Kim, H. Ahn and S-C. Cho, “ Variable Target Value Subgradient Method”, *Mathematical Programming*, 49, 1991, 359–369

[22] R. Mifflin and C. Sagastizabal, "A Scince Fiction Story in Nonsmooth Optimization Originating at IIASA", this volume.

[23] T. Motzkin and I.J. Schoenberg, "The Relaxation Method for Linear Inequalities", *Canadian Journal of Mathematics*, 6, 1954, 393–404.

[24] A. Nedic and D.P. Bertsekas, "Incremental Subgradient Methods for Nondifferentiable Optimization", *SIAM J. OPTIM.*, Vol. 12, No. 1., 2001.

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

[26] A.S. Nemirovski, "Efficient Methods in Convex Programming". *Lecture Notes, Technion-Faculty of Industrial Engineering & Management*, Fall Semester 1994/1995.

[27] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, Doordrecht, London, 2004

[28] Yu. Nesterov, "Primal-Dual Subgradient Methods for Convex Problems", *Mathematical Programming*, Ser B. 2009, 120:221–259.

[29] Yu. Nesterov, "Subgradient Methods for Huge-Scale Optimizations Problems", *CORE Discusion Papers*, 2012/2.

[30] B.T. Polyak, " A General Method of Solving Extremal Problems", *Soviet Math. Doklady*, 8, 593–597 , 1967.

[31] B.T. Polyak, " Minimization of Unsmooth Functionals", *U.S.S.R. Computational Mathematics and Mathematical Physics*, 9, 509–521, 1969.

[32] B.T. Polyak, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.

[33] B.T. Polyak, "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, 5–30.

[34] E. Rosenberg, "A Geometrically Convergent Subgradient Optimization Method for Nonlnearly Conctrained Convex Programs", *Mathematics of Operations Research*, 13, 3, 1988.

[35] N.Z Shor: "An application of the method of gradient descent to the solution of the network transportation problem". In: *Materialy Naucnovo Seminara po Teoret i Priklad. Voprosam Kibernet. i Issted. Operacii, Nucnyi Sov. po Kibernet*, Akad. Nauk Ukrain. SSSR, vyp. 1, pp. 9–17, Kiev 1962.

[36] N.Z Shor: "On the structure of algorithms for numerical solution of problems of optimal planning and design". Diss. Doctor Philos. Kiev 1964

[37] N.Z. Shor, "On the Rate of Convergence of the Generalized Gradient Method", *Kibernetika*, 4, 3, 1968.

[38] N.Z. Shor and P.R. Gamburd, "Some Questions Concerning the Convergence of the Generalized Gradient Method", *Kibernetika*, 7,6, 1971.

[39] N.Z. Shor, "Generalizations of Gradient Methods for Nonsmooth Functions and their Applications to Mathematical Programming", *Economic and Mathematical Methods*, Vo. 12, No. 2 pp. 337–356 (in Russian) 1976.

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

[41] N. Z. Shor, *Nondifferentiable Optimization and Polynomial Problems*, Kluwer Academic Publishers, Boston, Doordrecht, London 1998.

[42] M.J. Todd, "Some Remarks on the Relaxation Method for Linear Inequalities", Technical Report No. 468, SORIE, Cornell University, Ithaca, New York, 1980.

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

Jean-Louis Goffin
Professor emeritus in
Management Science
Desautels Faculty
of Management
McGill University
Montreal, Quebec
Canada
jean-louis.goffin@mcgill.ca