A Science Fiction Story in Nonsmooth Optimization Originating at IIASA

Robert Mifflin and Claudia Sagastizábal

2010 Mathematics Subject Classification: 65K05, 49J52, 49M05, 90C30
Keywords and Phrases: Nonsmooth optimization, bundle methods, superlinear convergence

Warning to the reader: despite its title, this story has no otherworldly planets, robots or galactic monsters; just a collection of fading memories confirming that optimization research is a perfect example of human synergy and persistence.

As in a fairy tale, this story starts in a castle, Schloss Laxenburg, one of the residences of the imperial Habsburg family located south of Vienna. In fact, it was one of Maria Theresa's summer houses. Many long years ago (forty plus) there once was a meeting of representatives from the Eastern and Western blocks which begat an international research organization to be located in Laxenburg, Austria. The International Institute for Applied Systems Analysis (IIASA) was thus created, with the purpose of building bridges across the Iron Curtain by means of scientific cooperation. This global, rather than nationalistic, goal was very bold and innovative.

Since its creation, IIASA has pursued the above goal and today it is focused on issues such as energy use and climate change, food and water supplies, poverty and equity, population aging, and sustainable development. The institute's research is independent of political or national interests; and the motto "Science for global insight" appears in its logo. But this is another story; here, we will rather look back, all the way to the IIASA beginnings and somewhat before to 1959, in order to give an answer to the question of whether or not, *superlinear convergence for nonsmooth optimization is science fiction*, as nicely phrased by Claude Lemaréchal in the 1970s.

The Founding Fathers

Before 1975 Claude Lemaréchal and Philip Wolfe independently created bundle methods that minimize a convex function f for which only one subgradient at a point is computable. The work of both authors appears in a 1975 *Mathematical Programming Study.*

Bundle methods are based on and improve on cutting-plane methods due to E. W. Cheney and A. A. Goldstein (1959) and to J. E. Kelley (1960). But this primal interpretation came much later. At first, a dual view was predominant: algorithms were designed to approximate a subdifferential set in such a way as to asymptotically satisfy (the nondifferentiable version of) Fermat's condition, $0 \in \partial f(\bar{x})$ where $\bar{x}$ is a minimizer. Since the new methods seemed to resemble conjugate gradient ones, they were called conjugate subgradient methods by Wolfe. The same algorithms were named extended Davidon methods by Lemaréchal, possibly with the hope for rapid convergence in mind.

Indeed, after W. Davidon (1959) and R. Fletcher and M. Powell (1963) developed superlinearly convergent quasi-Newton methods for smooth minimization, rapid convergence was on everyone's mind. For nonsmooth functions, however, this goal was seen as a wondrous grail, the object of an extended and difficult quest, which would take more than 30 years to achieve.

When Robert Mifflin heard about the new methods, he gave up on an algorithm that moved and shrank an n-dimensional simplex, because bundle methods use previously generated subgradient information in a more efficient manner. He then defined a large class of nonconvex functions, called semismooth, and a dual-type bundle algorithm that achieved convergence to stationary points for such functions. All of the above research provided a way to solve dual formulations of large-scale optimization problems where underlying special structure could be exploited through the future use of parallel computing.

In view of the new advances in the area, Wolfe influenced IIASA to form a nonsmooth optimization (NSO) task-force, including Lemaréchal, Mifflin, and certain Russians and Ukrainians. Among the latter, E. A. Nurminskii was expected at the beginning, but, probably due to the actions of Soviet authorities, could not make it to Laxenburg until one year after the departure of Lemaréchal and Mifflin.

With the support of Michel Balinski (Chairman of the System and Decision Sciences Area at IIASA), the task-force organized at Laxenburg in 1977 a two week long participant-named "First World Conference on Nonsmooth Optimization". From the Soviet side, there were B. T. Polyak and B. N. Pshenichnyi, while the West was represented by R. Fletcher, J. Gauvin, J.-L. Goffin, A. Goldstein, C. Lemaréchal, R. Marsten, R. Mifflin and P. Wolfe. Most of the participants wrote articles published in a 1978 IIASA Proceedings Series book.

At those times when politics mixed with science, researchers were warned that their phones might be tapped and looked for hidden microphones in their table lamps. So this first international workshop was viewed as going beyond mathematics and, in his opening speech, Lemaréchal, feeling the importance of the moment, welcomed the participants with the words, *To begin, let us break the glass.* His emotion made his French (glace) supersede his English (ice)![1]

[1] At a later Cambridge meeting Claude topped that slip of the tongue with the line "Now, I am only speaking in words" rather than the English equivalent "roughly speaking", meaning here, "without mathematical precision".

At the meeting, each participant presented his work during a three hour period in the morning, and the afternoon was devoted to brainstorming. These exchanges increased the participants' awareness of the strong connections between nonlinear programming and nonsmooth optimization. In particular, Roy Marsten explained boxstep methods, and Boris Pshenichnyi's talk suggested a link with Sequential Quadratic Programming, hinting at the possibility of superlinear convergence.

The new conjugate-subgradient-like methods were the subject of many discussions during this first workshop. Their novelty was in that, unlike most subgradient methods that could be thought of as being *forgetful* and also different from smooth algorithms, the new methods kept past basic information in *memory*. Indeed, for progressing from the current iterate to the next one, a direction is defined by solving a quadratic program with data consisting of function and subgradient values from several past points. It is precisely this collection of information generated at previous iterations that is referred to as "the bundle". Actually, the terminology was born during a workshop lunch:

- *bundle* in English;
- *faisceau* in French, a word that raised some concerns among English speaking participants, who wondered if it would connote fascism (it does not); and
- *Schar* in German.

As noted by Wolfe (while chewing Wiener Schnitzel mit Spätzle), the German word sounds close to Shor. In those times, the r-algorithm of N. Z. Shor was the *bête noire* of NSO researchers, because of its reported success in many practical applications. This is, in spite of the method (a combination of steepest descent and conjugate gradients) lacking a general convergence proof. When there is convergence little is known about its rate, except for a recent (2008) work by Jim Burke, Adrian Lewis and Michael Overton, interpreting the r-algorithm as a variable metric method that does not satisfy the secant equation (a partial convergence rate result is given, for a convex quadratic function of two variables). This interpretation could help in unveiling today's mystery behind the excellent performance of the r-algorithm.

The r-algorithm is a *space-dilation* method, a family of (not so amnesic!) subgradient algorithms using information from both a current and a previous iterate, and usually having excellent numerical behavior. This family includes a variant related to the symmetric rank-one quasi-Newton method. It was this type of recurrent finding that kept alive the quest for rapid convergence.

The ε-subdifferential and the road to implementation

A second international workshop took place at IIASA in 1980, with contributions from Y. M. Ermoliev, J.-L. Goffin, C. Lemaréchal, R. Mifflin, E. A.

Nurminskii, R. T. Rockafellar, A. Ruszczynski, and A. P. Wierzbicki. In the conference book, Terry Rockafellar wrote about the important class of lower C^2 functions, calling them particularly amenable to computation;[2] ten years before he had introduced the concept of approximate subgradients, which was extended to nonconvex functions by Al Goldstein in 1977. In 1991, after many years of joint climbing trips in the Dolomites with discussions on this subject, C. Lemaréchal and Jochem Zowe came up with the *eclipsing concept*, aimed at defining a first-order approximation of a multi-valued mapping.

The idea of an approximate subdifferential turned out to be fundamental for nonsmooth optimization. In particular, it is crucial for the effectiveness of bundle methods for large problems, but this is not its only important property. Indeed, on the theoretical side, the incorporation of an "expansion" parameter ε makes the multifunction $\partial_\varepsilon f(x)$ both inner and outer semicontinuous in the variables ε and x. For the exact subdifferential, the latter semicontinuity property holds (the subdifferential has a closed graph).

Inner semicontinuity is of paramount importance, since it guarantees that having sequences $x^k \to \bar{x}$ and $\varepsilon^k \to 0$, and a zero subgradient, $0 \in \partial f(\bar{x})$, there exists an approximate subgradient sequence g^k converging to zero: $\partial_{\varepsilon^k} f(x^k) \ni g^k \to 0$. Since the goal of any sound optimization method is to asymptotically satisfy Fermat's condition, without inner continuity there is no hope. Now, this essential property holds only for approximate subgradients, but the available information is from *exact* subgradients. What to do? Here arises an important algorithmic consequence of the concept, known in the area as a *transportation formula*, introduced by Lemaréchal in his *Thèse d'État* from 1980. This simple, yet powerful, formula for convex functions relates exact subgradients (at one point) to inexact ones (at another point), as follows:

$$g^i \in \partial f(x^i) \implies g^i \in \partial_{\hat{\varepsilon}} f(\hat{x}) \text{ for } \hat{\varepsilon} = f(\hat{x}) - f(x^i) - \left\langle g^i, \hat{x} - x^i \right\rangle .$$

By means of this relation, bundle methods relate past exact subgradient information to a special ε-subgradient at a so-called *serious* point $\hat{x}$, a point which gives significant progress towards the goal of minimizing the objective function (in bundle jargon, non-serious points are called *null*). This special subgradient and its corresponding $\hat{\varepsilon}$ are called the *aggregate* subgradient and error, respectively. Together with a serious subsequence of iterates, these aggregate objects ensure limiting satisfaction of Fermat's condition.

The notion of an approximate subdifferential was algorithmically exploited for the first time by Dimitri Bertsekas and S. Mitter, early on in 1971. In 1974 Rockafellar visited Kiev and gave a talk on the subject which was translated into Russian by Pshenichnyi. This made it possible for Evgenii Nurminskii to learn about the subject. He then started to study the semicontinuity properties of this new set-valued operator and, after some joint work with

[2]These functions had been introduced in 1974 by Robert Janin in his University of Paris IX PhD dissertation *Sur la dualité et la sensibilité dans les problèmes de programmation mathématique.*

Lemaréchal, eventually established its continuity. A comprehensive set of useful ε-subdifferential calculus rules was developed by Jean-Baptiste Hiriart-Urruty.

An interesting application of the ε-subdifferential, significant for numerical performance, is that past bundle information can be "compressed" into the aggregate subgradients and errors, without loss of global convergence. The compression mechanism allows for discarding bundle information, keeping only enough to construct the last bundle subproblem solution, for example, only the solution itself. This makes the next direction defining subproblem easier to solve, a feature that is not present in the original cutting-plane method, which has to keep all of the past information for the sake of convergence. For this reason cutting-plane methods often suffer from a slow tailing-off convergence effect.

Thanks to their potential for practical implementation, bundle methods were considered in several variants in the early 1990s. Trust region bundle methods and zig-zag searches were developed for convex and nonconvex functions by Zowe and his PhD student H. Schramm. Level variants were brought from Moscow to Paris by Arkadi Nemirovski and Yuri Nesterov, who wrote a paper with Lemaréchal on this subject. The development of technical tools for showing convergence of bundle methods and incorporating a compression mechanism in the algorithmic process is due to Krzysztof Kiwiel. He also developed a very efficient quadratic programming solver for the bundle direction subproblems, and systematically extended the methodology to different cases such as nonconvex and constrained ones.

The first VU and the primal view

The issue of increasing convergence speed of NSO methods was a recurrent obsession.

For single variable problems, a superlinearly convergent method was devised by Lemaréchal and Mifflin in 1982. It has a very simple rule for deciding if, near a serious point, the function's graph looks V-shaped (nonsmooth piecewise linear), or U-shaped (smooth quadratic). In the former case, a V-model, made from two cutting planes, is used to approximate the function. In the latter case, the difference of two "serious-side" derivatives is used to give second-order information for creating a quadratic U-model. Since cutting-plane methods are known to have finite termination for piecewise affine functions, these cases are solved efficiently with V-model minimizers. The same holds for smooth cases, because they are handled well via quasi-Newton moves.

Nevertheless, this fast algorithm had the handicap of not extending directly to functions of several variables. The difficulty with extending VU-concepts to multidimensional problems was eventually solved, but it took almost 20 years to find the right objects, after a detour involving work descending from that of J.-J. Moreau and K. Yosida.

The challenge was to find a generalization for the notion of a Hessian which is adequate for a black-box setting, that is, one that could be constructed from bundle information consisting of function and single subgradient values at each computed point. At this stage, the primal interpretation of bundle methods became handy, since when considered as a stabilized cutting-plane method, there is a direct link between certain bundle iterates and the proximal point theory initiated by B. Martinet in 1970. After the seminal work on this subject by Terry Rockafellar in 1976, theoretical proximal results blossomed during the 1980s and 90s. An important step towards practical implementation was taken by Masao Fukushima and Alfred Auslender, who independently showed that by not stopping bundling with a serious point one produced a sequence converging to a proximal point. Ending null steps with a serious step leads to an approximation of a proximal point.

In 1993 Claude Lemaréchal and Claudia Sagastizábal interpreted the bundle direction as coming from a preconditioned gradient direction for minimizing the Moreau-Yosida regularization function associated with the proximal points. This interpretation led to a BFGS proximal approach opening the way to *variable prox-metric* bundle methods, which made quasi-Newton updates for a Moreau-Yosida regularization that was not fixed (the proximal parameter varies with the iterations). So the approach looked, in fact, like a dog chasing its tail.

The smoothing effect of the Moreau-Yosida operator led to the belief that the key to defining an appropriate Hessian was to find proper proximal parameters (as in the BFGS proximal approach). This was a false track; in 1997 Lemaréchal and Sagastizábal showed that for the Moreau-Yosida regularization to have a Hessian everywhere, the (nonsmooth!) function f needed to be sufficiently smooth and have a Hessian itself ... once again, the elusive rapid convergence seemed out of reach.

Moving fast is possible, if in the right subspace

In their negative results from 1997, when studying the Moreau-Yosida Hessian, Lemaréchal and Sagastizábal noticed that a nonsmooth function f exhibits some kind of second order behavior *when restricted to a special subspace.* More precisely, the function has kinks on (a translation of) the tangent cone to $\partial f(\bar{x})$ at the zero subgradient and appears smooth or "U-shaped" on (a translation of) the normal cone. Under reasonable assumptions related to the minimizer $\bar{x}$ being nondegenerate, the cones above are in fact complementary subspaces, called V and U, because they concentrate, respectively, all of the nonsmoothness and smoothness of f near $\bar{x}$. In the same work it was noticed that a Newton step based on the Hessian of the Moreau-Yosida regularization has no V-subspace component.

The seed of just dropping off the regularization began to germinate.

In the period 1984–96 Mifflin came up with similar concepts and conclusions in a different manner based on the bundle algorithm itself. The algebra associated with the bundle method subproblem solution naturally breaks it into local V and U components with all the active subgradients having the same U-component, which suggests that U is the space of differentiability. Associated with this he also developed the idea of an algorithm step being the sum of a bundle serious step and a U-Newton step.

The U-Lagrangian from 2000, defined by Lemaréchal, François Oustry, and Sagastizábal, proved useful as a theoretical tool to extract implicitly second order information from a nonsmooth function without resorting to the Moreau-Yosida regularization. Its associated U-Hessian turns out to be the correct second order object for NSO, akin to the projected Hessian in smooth nonlinear programming. In some favorable cases (involving strong minimizers) a conceptual VU-Newton step, constructed from the sum of a V-step and a U-step depending on the result of the V-step, can produce a superlinearly convergent sequence of iterates. Paraphrasing Lemaréchal words: with the U-Lagrangian came the realization that, when moving along a V-shaped valley of nondifferentiability which is tangent to the U-subspace at the minimizer, a Newton-like method could drive the algorithm convergence with the desired speed.

The jackpot had been finally hit!

Or not yet? In a manner similar to the proximal point algorithm, the U-Lagrangian superlinear scheme was highly conceptual, as it depended on information at the minimizer being sought, i.e. assuming the dog had already caught its tail.

It would take some more years of hard work to produce implementable VU-versions. The process was started by Oustry, who produced a rapidly convergent VU-algorithm with dual steps for the special case of a max-eigenvalue function. Two quadratic programming problems needed to be solved per iteration, instead of only one, as in classical bundle algorithms. Unfortunately, the method, tailored for eigenvalue optimization, used *rich* black-boxes that computed more than one subgradient at each point.

Mifflin and Sagastizábal developed VU-theory further, defining a class of functions structured enough to generate certain primal and dual tracks (the class includes the max-eigenvalue case). In the meantime, the importance of structure producing nonsmoothness was noticed by Lewis, whose *partly smooth* functions formalize, in a general nonconvex setting, VU structure. This was followed by works by Aris Daniilidis, Warren Hare, Jerôme Malick and others. A nice connection between U-Lagrangian methods and Sequential Quadratic Programming was given by Scott Miller and J. Malick.

By relating primal and dual tracks to U-Lagrangians and proximal points, Mifflin and Sagastizábal succeeded in creating a superlinearly convergent VU algorithm for very general convex functions. The method also sequentially solves pairs of quadratic programs, corresponding to finding approximations in both the primal and dual tracks. This culminated over 30 years of effort by many researchers, not limited to the ones mentioned here, and brought us to

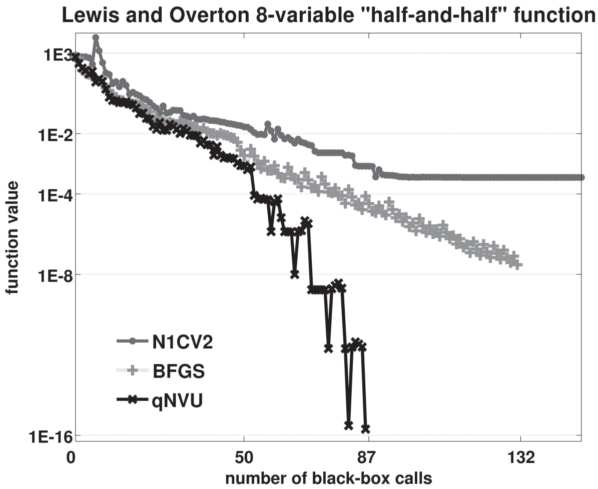

Figure 1: Sublinear, linear, and supernatural convergence

our current realization of science fiction: Figure 1 shows rapid convergence of a quasi-Newton version of the VU-algorithm.

The half-and-half function $f(x) = \sqrt{x^T Ax} + x^T Bx$ was created by Lewis and Overton to analyze BFGS behavior when minimizing a nonsmooth function. The 8-variable example in the figure has a matrix A with all elements zero, except for ones on the diagonal at odd numbered locations ($A(i,i) = 1$ for $i = 1, 3, 5, 7$). The matrix B is diagonal with elements $B(i,i) = 1/i^2$ for $i = 1, \ldots, 8$. The minimizer of this partly smooth convex function is at $\bar{x} = 0$, where the V and U subspaces both have dimension 4; hence, the name half-and-half.

Each graph in the figure shows function values from all points generated by its corresponding algorithm starting from the point having all components equal to 20.08. The top curve was obtained with a proximal bundle method, implemented in the code N1CV2 by Lemaréchal and Sagastizábal. The middle curve corresponds to the BFGS implementation by Overton, who adapted the method for nonsmooth functions via a suitable line search developed with Lewis. They argue that the linear convergence of "vanilla BFGS" as exhibited by this example is surprisingly typical for nonsmooth optimization. However, so far this has been proved only for a two variable example with the use of exact line searches, i.e., by exploiting nonsmoothness. It pays to exploit nonsmoothness, even in more than one dimension, and it can be done implicitly as shown by the (supernatural) curve at the bottom of the figure. This one results from the quasi-Newton VU algorithm that uses a BFGS update formula to approximate U-Hessians. Only its serious point subsequence has proven Qsuperlinear convergence.[3] The tops of the ending "humps" in this graph are due to "clumps" of null steps.

In the bundle business, null steps remain the hard cookies to digest. Null points can be thought of as intermediate unavoidable steps, needed to make the bundle "sufficiently rich", until enough progress is achieved and an iterate can be declared serious. This fact was also commented on by Stephen Robinson, who in 1999 proved R-linear convergence of ε-subgradient descent methods (including the serious subsequence of proximal bundle algorithms), for functions satisfying a certain inverse growth condition. The feature of eliminating unnecessary null steps is yet to be found in NSO, because it is not known what unnecessary means. An empirical observation of how the algorithmic process drives the aggregate gradient and error to zero shows that, in general, the aggregate error goes to zero fast, while it takes long time (including many null steps) for the aggregate gradient to attain a small norm. This phenomenon suggests there is a minimal threshold, which cannot be avoided, for the number of null steps between two serious iterates. But except for complexity results (referring to a worst case that is rare in practice), there is not yet a clear understanding of how to determine a realistic value for the threshold. Maybe in another 30 or 40 years the answer will be spoken in words in a future ISMP Optimization History book. In the meantime the quest continues with the search for rapid convergence to local minimizers for nonconvex functions.

Concluding remarks

The astute reader probably noticed that IIASA was not directly involved in VU theory and algorithm developments. The reason is that the institution discontinued support for nonsmooth optimization when its last man standing, Vladimir Demyanov, left IIASA in 1985. He had organized the last IIASA Workshop on Nondifferential Optimization, held in Sopron, Hungary in 1984, and was a very early contributor to the field with a minimax paper in 1968.

The same reader of this article will notice a lack of references as the authors are "only speaking in words" to minimize the level of technicality. This choice was made to avoid the embarrassment of missed citations.

[3] However, one can envision a smooth outer envelope function, starting at about evaluation number 37, which touches some points, is strictly concave and has an ending slope looking very close to minus infinity. It empirically shows R-superlinear convergence of the qNVU algorithm.

Acknowledgements. The authors are grateful to C. Lemaréchal and E. A. Nurminskii for sharing various NSO memories, from IIASA and elsewhere.

They also thank AFOSR, CNPq, Faperj, INRIA and NSF for many years of research support.

Robert Mifflin
Neill 103
Washington State University
Pullman WA 99164-3113
USA
mifflin@math.wsu.edu

Claudia Sagastizábal
IMPA
Estrada Dona Castorina 110
22460-320 Jardim Botanico
Rio de Janeiro – RJ
Brazil
sagastiz@impa.br