OPTIMA 88

Mathematical Optimization Society Newsletter

Philippe L. Toint

MOS Chair's Column

May 1, 2012. When starting to write this column, I must admit that I feel a little bit like cheating, because a substantial part of this very issue is occupied by an overview of work in which I have been personally deeply involved. Fortunately, this coincidence was none of my doing, and the result of independent conversations between Katya Scheinberg, our great Optima editor, and my co-authors. I do hope that you will find the topic interesting … and this also gives me a good excuse to keep the rest of this column short.

The Berlin International Symposium is approaching fast and promises to be a truly exiting event: a record number of talks (over 1700) have already been submitted! Beyond the real dedication of the local organizing committee, several preparatory tasks are also carried out in the various committees of the MOS. The first is the selection of the recipients of the various prizes which will be awarded at the Berlin opening ceremony (don't miss it!). The work of the ad hoc committees is definitely progressing, and I know that some very worthy conclusions have already been reached. I am looking forward to the public announcement of all MOS awards.

The second ongoing discussion is about the location of the next ISMP, beyond our Berlin symposium. Several proposals have been received and are currently being examined by the committee in charge. It is really nice to see that there are various possibilities, and also that colleagues are interested in organizing this important event.

Finally, I could not close this column without mentioning the MOS election process. The Executive Committee is finalizing the ballot which will be sent to all members very soon now, in order to elect a new Chair-elect (who is going to take over from me in due time), a new treasurer and a completely new MOS Council, in line with the Society's tradition. By the time you read these lines, the ballot may already be with you. Please vote for the officers you like best and whom you think would be able to care for the interests and the organization of MOS most effectively. Thank you very much in advance.

Note from the Editors

We are pleased to present the latest Optima issue dedicated to a hot topic in continuous optimization – global complexity bounds for nonlinear optimization methods. Our main article is by Cartis, Gould and Toint and it summarizes a comprehensive and impressive body of work that the three authors have accomplished over the last few years. Their work focuses on nonconvex problems and methods.

The discussion column is by Yurii Nesterov, whose countless contributions to the complexity theory of convex optimization need no introduction in the MOS community (and even outside of it). In fact, Yurii contributed an article to Optima on complexity of first order methods a few years ago. His column in this issue addresses the complexity of obtaining small gradient values in convex optimization and connects gracefully with the nonconvex case.

We hope that you will find the scientific discussion in this issue as enlightening as we do.

Katya Scheinberg, Editor
Sam Burer, Co-Editor
Volker Kaibel, Co-Editor

Coralia Cartis, Nicholas I. M. Gould and Philippe L. Toint

How Much Patience Do You Have? A Worst-Case Perspective on Smooth Nonconvex Optimization

"Though you be swift as the wind, I will beat you in a race",
said the tortoise to the hare.
Aesop

1 Introduction

Nonlinear optimization – the minimization or maximization of an objective function of one or more unknowns which may be restricted by constraints – is a vital component of computational science, engineering and operations research. Application areas such as structural design, weather forecasting, molecular configuration, efficient utility dispatch and optimal portfolio prediction abound. Moreover, nonlinear optimization is an intrinsic part of harder application problems involving integer variables.

When (approximate) first and second derivatives of the objective are available, and no constraints are present, the best known optimization methods are based on the steepest descent [14, 28] and Newton's methods [14, 28]. In the presence of nonconvexity of the objective these techniques may fail to converge from poor,

Contents of Issue 88 / May 2012

or sometimes even good, initial guesses of the solution unless they are carefully safeguarded. State-of-the-art enhancements such as *linesearch* [28] and *trust-region* [13] restrict and/or perturb the local steepest descent or Newton step so as to decrease the objective and ensure (sufficient) progress towards the optimum on each algorithm iteration. Even when convergent, the potential nonconvexity of the objective and the use of derivatives in the calculation of iterative improvement only guarantee local optimality, and most commonly, a point at which the gradient of the objective is (approximately) zero.

Efficient implementations of standard Newton-type methods, with a linesearch or trust-region safeguard, are available in both commercial and free software packages, and are often suitable for solving nonlinear problems with thousands or even hundreds of thousands of unknowns; see GALAHAD, IPOPT, KNITRO, LOQO, PENNON or SNOPT for examples of state of the art software. Often little is known about special properties of a problem under consideration (such as convexity), and so the methods and the software need to be able to cope with a wide spectrum of instances.

Due to this wide range of applicability of generic software, it is essential to provide rigorous guarantees of convergence of the implemented algorithms for large classes of problems under a wide variety of possible algorithm parameters. Much research has been devoted to analysing the local and global convergence properties of standard methods, but what can be said about the rate at which these processes take place? This is significant as a fast rate implies that fewer iterates are generated, saving computational effort and time; the latter is essential for example when the function- and derivative-evaluations required to generate the iterates are computationally expensive to obtain, such as in climate modelling and multi-body simulations.

If a "sufficiently good" initial estimate of a well-behaved solution is available, then it is well known (from local convergence results) that Newton-type processes will be fast; they will converge at least super-linearly. However, for general problems (even convex ones), it is impossible or computationally expensive to know a priori the size of this neighbourhood of fast convergence. Frequently, even a good guess is unavailable, and the starting point is far away from the desired solution. Also, optimal points are not always well-behaved, they may be degenerate or lie at infinity, and in such cases, fast convergence may not occur. Therefore, the question of the *global rate of convergence* or *global efficiency* of standard algorithms for general nonconvex sufficiently-smooth problems naturally arises as a much more challenging aspect of algorithm analysis. Until recently, this question has been entirely open for Newton-type methods. Furthermore, due to the wide class of problems being considered, it is more reasonable to attempt to find bounds on this rate, or more precisely upper bounds on the number of iterations the algorithm takes to reach within desired accuracy of a solution. For all algorithms under consideration here, the latter is equivalent to upper bounding the number of function- and/or gradient-evaluations required for approximate optimality, and this count is generally of most interest to users. Hence, we refer to this bound as the *worst-case function-evaluation complexity* of an algorithm. This computational model that counts or bounds the number of calls to the *black-box* or *oracle* generating the objective and/or gradient values is suitably general and appropriate for nonlinear programming due to the diversity of "shapes and sizes" that problems may have. Fundamental contributions and foundational results in this area are presented for instance in [20, 21, 29, 32], where the NP-hardness, -completeness or otherwise of various optimization problem classes and optimization-related questions, such as the calculation of a descent direction, is established.

We begin by mentioning existing complexity results for steepest-descent methods and show that the upper bounds on their global efficiency when applied to sufficiently smooth but potentially non-convex problems are essentially sharp. We then illustrate that, even when convergent, *Newton's method can be – surprisingly – as slow as steepest descent*. Furthermore, all commonly encountered linesearch or trust-region variants turn out to be essentially as inefficient as steepest descent in the worst-case. There is, however, good news: cubic regularization [11, 18, 27] is better than both steepest-descent and Newton's in the worst-case; it is in fact, optimal from the latter point of view within a wide class of methods and problems. We also present bounds on the evaluation complexity of nonconvexly constrained problems, and argue that, for certain carefully devised methods, these can be of the same order as in the unconstrained case, a surprising but reassuring result.

Note that the evaluation complexity of convex optimization problems is beyond the scope of this survey. This topic has been much more thoroughly researched, with a flurry of recent activity in devising and analyzing fast first-order/gradient methods for convex and structured problems; the optimal gradient method [21] has determined or inspired many of the latter developments. Furthermore, the polynomial complexity of interior point methods for convex constrained programming ([26] and others) has changed the landscape of optimization theory and practice for good.

2 Global Efficiency of Standard Methods

2.1 Sharp Bounds for Steepest Descent Methods

Consider the optimization problem

$$\underset{x\in\Re^n}{\text{minimize}}\ f(x),$$

where $f:\Re^n\to\Re$ is smooth but potentially nonconvex. On each iteration, steepest descent methods move along the negative gradient direction so as to decrease the objective $f(x)$; they have the merit of simplicity and theoretical guarantees of convergence under weak conditions when globalized with linesearches or trust-regions [13, 14]. Regarding the evaluation complexity of these methods, suppose that f is globally bounded below by f_{low} and that its gradient g is globally Lipschitz continuous. When applied to minimize $f(x)$, and given a starting point $x_0\in\Re^n$ and an accuracy tolerance $\epsilon>0$, standard steepest descent methods with linesearch or trust-region safeguards have been shown to take at most

$$\left\lceil \frac{\kappa_{\text{sd}}}{\epsilon^2}\right\rceil \qquad (1)$$

function and gradient evaluations to generate an iterate x_k satisfying $\|g(x_k)\|\le\epsilon$ [21, p. 29], [17, Corollary 4.10]. Here κ_{sd} is independent of ϵ, but depends on the initial distance to the optimum $f(x_0)-f_{\text{low}}$, on the Lipschitz constant of the gradient and possibly, on other problem and algorithm parameters. Note that the bound implies at least a sublinear global rate of convergence for the algorithm [21, p. 36]. Despite being the best-known bound for steepest descent methods and even considering the well-known inefficient practical behaviour of gradient-type methods on ill-conditioned problems, (1) may still seem unnecessarily pessimistic. We illustrate however that this bound is essentially sharp as a function of the accuracy ϵ.

Example 1 (Steepest Descent Method). Figure 1a exemplifies a univariate, twice continuously differentiable function with globally Lipschitz continuous gradient on which the steepest descent method with inexact linesearch takes precisely

$$\left\lceil \frac{1}{\epsilon^{2-\tau}}\right\rceil$$

Note that by giving up the Lipschitz continuity requirement on the gradient and Hessian, one can construct functions on which the complexity of Newton's method can be made arbitrarily poor [12, §3.2].

3 Improved Complexity for Cubic Regularization Methods

In a somewhat settled state of affairs (at least for problems without constraints), a new Newton-type approach, based on *cubic regularization*, was proposed independently by Nesterov and Polyak (2006) [27], and Weiser, Deuflhard and Erdmann (2007) [35], and led to the rediscovery of an older (unpublished) fundamental work by Griewank (1981) [18]. Crucially, [27] showed that such a technique requires at most $\mathcal{O}(\epsilon^{-3/2})$ function-evaluations to drive the gradient below ϵ, the first result ever to show that a second-order scheme is better than steepest-descent in the worst-case, when applied to general (nonconvex) functions, a remarkable milestone!

These cubic techniques can be described by a well-known overestimation property. Assume that our objective $f(x)$ is twice continuously differentiable with globally Lipschitz continuous Hessian H of Lipschitz constant $2L_H$. Then the latter property, a second-order Taylor expansion and the Cauchy-Schwarz inequality imply that at any given x_k,

$$f(x_k+s) \leq f(x_k) + g(x_k)^T s + \frac{1}{2}s^T H(x_k)s + \frac{L_H}{3}\|s\|^3 \stackrel{\text{def}}{=} m_{k,L}(x_k+s), \text{ for any } s \in \Re^n, \quad (3)$$

where $\|\cdot\|$ is the usual Euclidean norm [27, Lemma 1], [11, (1.1)]. Thus if we consider x_k to be the current best guess of a (local) minimizer of $f(x)$, then the right-hand side of (3) provides a local cubic model $m_{k,L}(x_k+s)$, $s \in \Re^n$, such that $f(x_k) = m_{k,L}(x_k)$. Further, if $x_k + s_k$ is the global minimizer of the (possibly nonconvex but bounded below) model $m_{k,L}$, then due to (3), f can be shown to decrease by a significant amount at the new point $x_k + s_k$ from its value at x_k [27, Lemmas 4, 5]. Although theoretically ideal, using $m_{k,L}$ is impractical and unrealistic as L is unknown in general, may be expensive to compute exactly and may not even exist for a general smooth function. Thus, in the algorithmic framework Adaptive Regularization with Cubics (ARC) [11, Algorithm 2.1], we propose to employ instead the local cubic model

$$m_k(x_k+s) \stackrel{\text{def}}{=} f(x_k)+g(x_k)^T s+\frac{1}{2}s^T B_k s+\frac{\sigma_k}{3}\|s\|^3, \quad s \in \Re^n, \quad (4)$$

where B_k is an approximation to the Hessian of f at x_k; the latter is also a practical feature, essential when the Hessian is unavailable or expensive to compute. Even more importantly, $\sigma_k > 0$ is a regularization parameter that ARC adjusts automatically and is no longer conditioned on the computation or even existence of a (global) Hessian Lipschitz constant. In particular, σ_k is increased by say, a constant multiple factor until *approximate* function decrease [11, (2.4)] – rather than the more stringent overestimation property – is achieved; on such iterations, the current iterate is left unchanged as no progress has been made. When sufficient objective decrease has been obtained (relative to the model decrease), we update the iterate by $x_{k+1} = x_k + s_k$ and may even allow σ_k to decrease in order to prevent the algorithm from taking unnecessarily short steps. Global convergence of ARC can be shown under very mild assumptions on f and approximate model minimization conditions on the step s_k [11, Corollary 2.6]. Adaptive σ_k updates and cubic models have also been proposed in [18, 27, 35] but these proposals still rely on ensuring overestimation at each step and on the existence of global Hessian Lipschitz constants, while the ARC approach shows that local constant estimation is sufficient.

Exact Model Minimization. Essential to ARC's and any cubic regularization method's fast local and global rates of convergence is that minimizing $m_k(s)$ over $s \in \Re^n$, despite being a nonconvex problem (as Figure 2a illustrates), can be solved efficiently – in polynomial time – to find the global minimizer s_*, a rare instance in the nonconvex optimization literature! In particular, any global minimizer s_* of (4) satisfies the system

$$(B_k + \lambda_* I)s_* = -g(x_k), \quad \text{where } B_k + \lambda_* I \text{ is positive semidefinite and } \lambda_* = \sigma_k\|s_*\|. \quad (5)$$

See [18], [27, §5.1], [11, Theorem 3.1] for a proof. The first and last set of equations in (5) express that the gradient of the model $m_k(x_k+s)$ is zero at s_*, which are first-order necessary optimality conditions that hold at any local or global minimizer of the model. The global optimality of s_* is captured in the eigenvalue condition $\lambda_k \geq \max\{-\lambda_1, 0\}$, where λ_1 is the left-most eigenvalue of B_k, and which is more stringent than local second-order optimality conditions for the model.

The characterization (5) can be used to compute s_* as follows [11, §6.1]. Express $s_* = s(\lambda)$ as a function of λ from the first set of equations in (5) and then replace it in the third condition $\|s(\lambda)\| = \lambda/\sigma_k$ which is now a univariate nonlinear equation in λ. We can apply Newton's method for finding the root of the latter equation in the interval $(\max\{-\lambda_1, 0\}, \infty)$, as represented in Figure 2b. Applying Newton's method in this context requires repeated factorizations of diagonally perturbed B_k matrices, and so this approach is only suitable when B_k is sparse or not too large.

Approximate Model Minimization. In the large-scale case, we have proposed [11, §3.2, §6.2] to set s_k to be only an approximate global minimizer of $m_k(x_k+s)$ that can be computed using Krylov-type methods, thus requiring only matrix-vector products. In particular, for each k, successive trial steps $s_{k,j}$ are computed as global minimizers of the cubic model $m_k(x_k+s)$ over increasing subspaces $s \in \mathcal{L}_j$[1] until the inner model minimization termination condition

$$\|\nabla_s m_k(x_k+s_{k,j})\| \leq \kappa_\theta \min\left\{1, \|s_{k,j}\|\right\}\|g(x_k)\| \quad (6)$$

is satisfied for some $\kappa_\theta \in (0,1)$. We then set $s_k = s_{k,j}$ where j is the final inner iteration. Since $\nabla_s m_k(x_k) = g(x_k)$, this termination criterion is a relative error condition, which is clearly satisfied at any stationary point of the model m_k. Generally, one hopes that the inner minimization will be terminated before this inevitable outcome. To be specific, one may employ a Lanczos-based approach that generates the Krylov subspace $\left\{g(x_k), B_k g(x_k), B_k^2 g(x_k)\ldots\right\}$. Then the Hessian of the reduced cubic model in the subspace is tridiagonal, and hence inexpensive to factorize when solving the characterization (5) in the subspace.

We have shown that ARC with approximate model minimization inherits the fast local convergence [11, §4.2] and complexity of cubic regularization with exact model minimization[2] [27]. We recall the bound on the worst-case performance of ARC.

Theorem 1. *[10, Corollary 5.3] Assume that f is bounded below by f_{low}, and that its gradient g and Hessian H are globally Lipschitz continuous on the path of the iterates[3]. Assume that ARC with exact or approximate model minimization is applied to minimizing f starting from some $x_0 \in \Re^n$, with $\sigma_k \geq \sigma_{\min} > 0$ and the approximate Hessian B_k satisfying $\|[B_k - H(x_k)]s_k\| = \mathcal{O}(\|s_k\|^2)$.[4] Then ARC takes at most*

$$\left\lceil \frac{\kappa_{\text{arc}}}{\epsilon^{\frac{3}{2}}} \right\rceil \quad (7)$$

5.1 Detour I: Minimizing a Nonsmooth Composite Function

Consider the unconstrained problem

$$\underset{x\in\Re^n}{\text{minimize}}\ h(r(x)), \tag{8}$$

where $r : \Re^n \to \Re^p$ is smooth but potentially nonconvex and $h : \Re^p \to \Re$ is convex but potentially nonsmooth; we may think of h as a norm. First-order methods have been devised for this problem [9, 23, 24] that satisfy the same evaluation complexity bound $\mathcal{O}(\epsilon^{-2})$ as in the unconstrained smooth case, despite the non-smoothness of h.

The *quadratic regularization* approach in [9] computes the trial step s_k from the current iterate x_k by solving the convex problem that linearizes the smooth parts of the composite objective but leaves the non-smooth parts unchanged, namely,

$$\underset{s\in\Re^n}{\text{minimize}}\ \underbrace{h(r(x_k) + A(x_k)s)}_{l(x_k,s)} + \frac{\sigma_k}{2}\|s\|^2,$$

where $A(x)$ denotes the Jacobian of $r(x)$ and $\sigma_k > 0$ is a regularization weight.[7] There is an underlying assumption that h is simple enough to make the above subproblem inexpensive to solve, as in the case of polyhedral norms. The parameter σ_k is adjusted in a similar way as for ARC to ensure sufficient objective decrease.

Assuming that h and A are globally Lipschitz continuous and the composite function is bounded below, the quadratic regularization framework can be shown to take at most

$$\left\lceil \frac{\kappa_{\text{qr}}}{\epsilon^2} \right\rceil \tag{9}$$

residual evaluations to achieve

$$\Psi(x_k) = l(x_k, 0) - \min_{\|s\|\le 1} l(x_k, s) \le \epsilon, \tag{10}$$

where $\Psi(x_k)$ is a first-order criticality measure [9, Theorem 2.7].

5.2 A First-Order Algorithm for Equality and Inequality Constrained Problems

Now consider the smooth nonconvex equality constrained problem

$$\underset{x\in\Re^n}{\text{minimize}}\ f(x) \text{ subject to } c(x) = 0. \tag{11}$$

As illustrated in Figure 4a, the Short-Step Steepest-Descent (ShS-SD) algorithm relies on two phases, one for feasibility and a second for optimality [3]. In **Phase 1**, ShS-SD attempts to generate a feasible iterate (if possible), by minimizing $\|c(x)\|$. This nonsmooth objective is of the form (8) with $h = \|\cdot\|$ and

$$r(x) \overset{\text{def}}{=} c(x), \tag{12}$$

and can thus be solved by the quadratic regularization approach for (8). If an iterate satisfying $\|c(x_1)\| \le \epsilon$ is found at the end of Phase 1, then **Phase 2** is entered, where we iteratively and approximately track the trajectory

$$\mathcal{T} = \{x \in \Re^n : c(x) = 0 \text{ and } f(x) = t\}$$

for decreasing values of t from some initial t_1 corresponding to the initial feasible iterate x_1. Namely, for the current target t_k, we do one quadratic regularization iteration from the current iterate x_k aimed at minimizing the merit function

$$\Phi(x, t_k) \overset{\text{def}}{=} \|c(x)\| + |f(x) - t_k|,$$

which again is of the form (8) with $r(x) \overset{\text{def}}{=} r(x, t_k)$ and

$$r(x, t) \overset{\text{def}}{=} \begin{pmatrix} c(x) \\ f(x) - t_k \end{pmatrix}. \tag{13}$$

If $\Phi(x_{k+1}, t_k)$ has not decreased sufficiently compared to $\Phi(x_k, t_k)$, we keep t_k unchanged and repeat; otherwise, we update t_k to t_{k+1} so as to ensure $\Phi(x_{k+1}, t_{k+1}) = \epsilon$. The latter implies that $\|c(x_{k+1})\| \le \epsilon$ and so we remain approximately feasible at the new iterate. Phase 2 terminates when (10) corresponding to $\Phi(x_k, t_k)$ holds.

The particular updating rule for t_{k+1} [3, (2.11)] also provides that the decrease in t_k is at least as much as that in the objective $\Phi(\cdot, t_k)$, namely,

$$t_k - t_{k+1} \ge \Phi(x_k, t_k) - \Phi(x_{k+1}, t_k) \ge \kappa \cdot \epsilon^2 \tag{14}$$

for some problem constant κ. The second inequality in (14) follows from the guaranteed function decrease on successful quadratic regularization iterations prior to termination [9, (2.38)]. Figure 4b illustrates the ℓ_1-neighbourhoods $\Phi(x, t) \le \epsilon$ in the two-dimensional plane $(\|c\|, f)$ and the inequalities (14) with $(x_k, t_k) = (x, t)$ and $(x_{k+1}, t_{k+1}) = (x_+, t_+)$. The main complexity result follows.

Theorem 3. *[3, Theorem 3.6] Assume that $c \in C^1(\Re^n)$ with globally Lipschitz continuous Jacobian J, and f is bounded below by f_{low} and above by f_{up} and has Lipschitz continuous gradient g in a small neighbourhood of the feasibility manifold. Then, for some problem constant κ_{sh}, the ShS-SD algorithm takes at most*

$$\left\lceil \left(\|c(x_0)\| + f_{\text{up}} - f_{\text{low}}\right) \frac{\kappa_{\text{sh}}}{\epsilon^2} \right\rceil$$

problem evaluations[8] to find an iterate x_k that is either an infeasible critical point of the feasibility measure $\|c(x)\|$ – namely, $\|c(x_k)\| > \epsilon$ and $\|J(x_k)^T z\| \le \epsilon$ for some z – or an approximate KKT point of (11), namely, $\|c(x_k)\| \le \epsilon$ and $\|g(x_k) + J(x_k)^T y\| \le \epsilon$ for some multiplier y.

Sketch of proof. Clearly, the total evaluation complexity is the sum of the complexity of each Phase. Phase 1's complexity follows directly from (9) and (12). In Phase 2, the target t_k remains unchanged for only a problem-constant number of 'unsuccessful' steps, and then it is decreased by at least ϵ^2 due to (14). The targets t_k are bounded below due to $f(x_k)$ being bounded and close to the targets, and so Phase 2 must terminate at the latest when t_k has reached its lower bound.

Crucially, (10) corresponding to $\Phi(x_k, t_k)$ implies that $\|g(x_k) + J(x_k)^T y\| \le \epsilon$ for some y [9, Theorem 3.1]; letting $g = 0$ gives the criticality condition for $\|c\|$, with the remark that if $\|z\| < 1$, we are guaranteed to have $\|c(x_k)\| \le \epsilon$ [9, (3.11)]. □

Note that no constraint qualification is required to guarantee termination and complexity of ShS-SD.

This approach also applies to inequality-constrained problems, by replacing $\|c(x)\|$ with $\|\min\{c(x), 0\}\|$ throughout.

6 Improved Complexity for Constrained Problems

It is natural to ask, as before, if there is an algorithm for constrained problems that has better worst-case complexity than $\mathcal{O}(\epsilon^{-2})$. For this, cubic regularization-based methods are the obvious candidates since their complexity in the unconstrained case is the best we know. The question thus becomes, can we extend cubic regularization methods for constrained problems while retaining their good complexity? We attempt to answer this question for the remainder of this survey. Again, we begin by taking a detour.

References

[1] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. ERGO Technical Report 12-001, School of Mathematics, University of Edinburgh, 2012.

[2] C. Cartis, N. I. M. Gould and Ph. L. Toint. A note about the complexity of minimizing Nesterov's smooth Chebyshev-Rosenbrock function. ERGO Technical Report 11-013, School of Mathematics, University of Edinburgh, 2011.

[3] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the complexity of finding first-order critical points in constrained nonlinear programming. ERGO Technical Report 11-005, School of Mathematics, University of Edinburgh, 2011.

[4] C. Cartis, N. I. M. Gould and Ph. L. Toint. Optimal Newton-type methods for nonconvex smooth optimization problems. ERGO Technical Report 11-009, School of Mathematics, University of Edinburgh, 2011.

[5] C. Cartis, N. I. M. Gould and Ph. L. Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optimization Methods and Software*, DOI:10.1080/10556788.2011.602076, 2011.

[6] C. Cartis, N. I. M. Gould and Ph. L. Toint. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, doi: 10.1093/imanum/drr035, 2011.

[7] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22(1):66–86, 2012.

[8] C. Cartis, N. I. M. Gould and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[9] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[10] C. Cartis, N. I. M. Gould and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, 2011.

[11] C. Cartis, N. I. M. Gould and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[12] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[13] A. R. Conn, N. I. M. Gould and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, USA, 2000.

[14] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1983. Reprinted as *Classics in Applied Mathematics 16*, SIAM, Philadelphia, USA, 1996.

[15] R. Garmanjani and L. N. Vicente. Smoothing and worst case complexity for direct-search methods in non-smooth optimization. Preprint 12-02, Department of Mathematics, University of Coimbra, Portugal, 2012.

[16] S. M. Goldfeld, R. E. Quandt and H. F. Trotter. Maximization by quadratic hill-climbing. *Econometrica*, 34:541–551, 1966.

[17] S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1), 414–444, 2008.

[18] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981.

[19] F. Jarre. On Nesterov's smooth Chebyshev-Rosenbrock function. Technical report, University of Düsseldorf, Düsseldorf, Germany, May 2011.

[20] A. S. Nemirovskii and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley and Sons, Chichester, England, 1983.

[21] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[22] Yu. Nesterov. Cubic regularization of Newton's method for convex problems with constraints. CORE Discussion Paper 2006/9, Université Catholique de Louvain, Belgium, 2006.

[23] Yu. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Paper 2007/76, Université Catholique de Louvain, Belgium, 2007.

[24] Yu. Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimization Methods and Software*, 22(3):469–483, 2007.

[25] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, Series A, 112(1):159–181, 2008.

[26] Yu. Nesterov and A. S. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming* SIAM, Philadelphia, USA, 1994.

[27] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[28] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 1999. Second edition, 2006.

[29] K. A. Sikorski. *Optimal Solutions of Nonlinear Equations*. Oxford University Press, Oxford, England, 2001.

[30] K. Ueda and N. Yamashita. Regularized Newton method without line search for unconstrained optimization. Technical Report 2009-007, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Japan, 2009.

[31] K. Ueda and N. Yamashita. On a global complexity bound of the Levenberg-Marquardt method. *Journal of Optimization Theory and Applications*, 147:443–453, 2010.

[32] S. A. Vavasis. *Nonlinear Optimization: Complexity Issues*. International Series of Monographs on Computer Science. Oxford University Press, Oxford, England, 1992.

[33] S. A. Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.

[34] L. N. Vicente. Worst case complexity of direct search. Preprint 10-17, Department of Mathematics, University of Coimbra, Portugal, 2010.

[35] M. Weiser, P. Deuflhard and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, 22(3):413–431, 2007.

Discussion Column

Yurii Nesterov

How to Make the Gradients Small

In many situations, the points with small gradients perfectly fit our final goals. Consider for example, the dual approach for solving the problem $f^* = \min_{x \in Q}\{f(x) : Ax = b\}$ with convex Q and strongly convex objective. Then the dual problem is

$$\max_y \left\{\phi(y) = \min_{x \in Q}\left[f(x) + \langle y, b - Ax\rangle\right]\right\} = f^*.$$

Let $x(y) \in Q$ be the unique solution of the internal problem. Then $\phi'(y) = b - Ax(y)$. Therefore

$$f(x(y)) - \phi(y) = -\langle y, \phi'(y)\rangle \le \|y\| \cdot \|\phi'(y)\|.$$

Thus, the value $\|\phi'(y)\|$ serves as the measure of feasibility and optimality of the primal solution.

In Convex Optimization, the traditional theoretical target is the fast convergence of the objective to f^*. The rate of convergence for the gradients is addressed very rarely. Let us present here the main available results. All supporting inequalities can be found in [1], [2], and [3].

1. For a problem of unconstrained smooth convex minimization, each iteration of the *Gradient Method* decreases the objective as follows:

$$f(x_k) - f(x_{k+1}) \ge \frac{1}{2L}\|f'(x_k)\|^2, \qquad (1)$$

where L is the Lipschitz constant of the gradient. On the other hand, we have $f(x_k) - f^* \le \frac{2LR^2}{k+4}$, where $R = \|x_0 - x^*\|$. Summing up (1)

for $k = m+1, \dots, N$, with $N = 2m$, we get

$$\frac{2LR^2}{m+4} \geq f(x_m) - f^* \geq f(x_{N+1}) - f^* + \frac{1}{2L} \sum_{k=m+1}^{N} \|f'(x_k)\|^2 \quad (2)$$
$$\geq \frac{m}{2L} \cdot \min_{0 \leq k \leq N} \|f'(x_k)\|^2,$$

Thus, we can find a point $\bar{x}$ with $\|f'(\bar{x})\| \leq \epsilon$ in $\frac{4LR}{\epsilon}$ iterations.

2. For the same problem, the *Fast Gradient Methods* (FGM) converge as $f(x_k) - f^* \leq \frac{4LR^2}{(k+2)^2}$. Let us introduce in these schemes an additional gradient step ensuring the decrease (1) between the best point of the previous iteration and the starting point of the next one. Then we can apply the above reasoning and obtain a chain of inequalities (2) with the new left-hand side $\frac{4LR^2}{(m+2)^2}$. Thus, we obtain $\|f'(\bar{x})\| \leq \epsilon$ in $O\left((\frac{LR}{\epsilon})^{2/3}\right)$ iterations of FGM.

3. A better complexity bound can be obtained by the regularization technique. Consider the function $f_\delta(x) = f(x) + \frac{\delta}{2}\|x - x_0\|^2$. It is strongly convex with parameter δ. Therefore, FGM can find $\bar{x}$ with $\|f'_\delta(\bar{x})\| \leq \frac{\epsilon}{2}$ in $O(\sqrt{\frac{L}{\delta}} \ln \frac{LR}{\epsilon})$ iterations. For $\delta = \frac{\epsilon}{2R}$, we get $\|f'(\bar{x})\| \leq \frac{\epsilon}{2} + \delta\|\bar{x} - x_0\| \leq \epsilon$. Thus, we need $O\left((\frac{LR}{\epsilon})^{1/2} \ln \frac{LR}{\epsilon}\right)$ iterations. Up to a logarithmic factor, this is an optimal complexity bound. There are no known direct methods, i.e., methods not using some form of regularization, with this efficiency estimate.

4. Let us look now at the efficiency estimates for the second-order schemes. Assume that the Hessian $f''(x)$ is Lipschitz continuous with constant M. Then, the cubic regularization of the Newton Method [2] decreases the functional gap with the rate $f(x_k) - f^* \leq \frac{27MR^3}{2(k+1)^2}$. It can be accelerated by the technique of estimate functions [3] up to the rate $f(x_k) - f^* \leq \frac{14MR^3}{k(k+1)(k+2)}$. Let us apply it to the regularized function $F_\delta(x) = f(x) + \frac{\delta}{3}\|x - x^0\|^3$. We introduce in this method a regular restart after m iterations. Since F_δ is uniformly convex of degree three,

$$\frac{\delta}{3}\|x_m - x^*_\delta\|^3 \leq F_\delta(x_m) - F_\delta(x^*_\delta)$$
$$\leq \frac{14M}{m(m+1)(m+2)}\|x_0 - x^*_\delta\|^3.$$

Thus, if $m = O((\frac{M}{\delta})^{1/3})$, then the value $\|x_m - x^*_\delta\|^3$ can be made at most half of $\|x_0 - x^*_\delta\|^3$. Let us repeat these series of m steps. Denote the last point of the k-th series by y_k with $y_0 = x_0$. After each series we compute a point u_k by taking one Cubic Newton Step from the point y_k. This point is taken as a starting point for the next series. In this case,

$$\left(\frac{1}{2}\right)^k \frac{M}{3}R^3 \geq F_\delta(y_k) - F_\delta(x^*_\delta) \geq \frac{1}{12M^{1/2}}\|F'_\delta(u_k)\|^{3/2}.$$

Therefore, in order to get $\|F'_\delta(\bar{x})\| \leq \frac{\epsilon}{2}$, we need $K = O(\ln \frac{MR^2}{\epsilon})$ series. After the last one, we have $\|f'(u_K)\| \leq \frac{\epsilon}{2} + \delta R^2$. Thus, we need $\delta = \frac{\epsilon}{2R^2}$. Hence, we perform at most $O\left((\frac{MR^2}{\epsilon})^{1/3} \ln \frac{MR^2}{\epsilon}\right)$ iterations in order to obtain the norm of the gradient smaller than ϵ. For such a goal, this is the best dependence in ϵ achieved so far in Convex Optimization. The lower complexity bounds for these settings are not known.

5. Let us discuss now the complexity bounds of the gradient norm minimization in nonconvex case. The main article in this issue by Cartis, Gould and Toint, provides us with very interesting arguments, which show that the lower complexity bound for our problem is $O(\frac{f_0 - f^*}{\epsilon^{3/2}})$. Moreover, this bound is achieved by the Cubic Newton Method (see [2]). Let us show that a minor change in the initial conditions dramatically changes our conclusions. Consider the following situation.

Problem class. Nonconvex functions with Lipschitz continuous Hessian. There exists at least one point x^* such that $f'(x^*) = 0$ and $\|x^*\|_\infty \leq R$.

Goal. Find a point $\bar{x}$ such that $\|f'(\bar{x})\|_\infty < \epsilon$ and $\|\bar{x}\|_\infty \leq R$.

Theorem. *The lower complexity bound for our problem class is* $\left(\frac{MR^2}{4\epsilon}\right)^{n/2}$. *It is implemented by the Uniform Grid Method.*

Idea of the proof. Let us fix an integer $p \geq 1$. We apply the following, so-called, resisting oracle: at each test point x generated by the method, it answers that $f'(x) = \epsilon 1_n$, (where 1_n is the n-dimensional vector of 1s) and $f''(x) = 0$. Assume that the number of questions N of our method is smaller than p^n. Then there exists a box $B \stackrel{\text{def}}{=} \{x \in R^n : \bar{x} \leq x \leq \bar{x} + \frac{R}{p}1_n\}$ where there were no questions. We define $f'(x) \equiv \epsilon 1_n$ for $x \notin B$. Inside the box, for each coordinate $f'_i(x)$ we smoothly connect the level ϵ at the points $\bar{x}^{(i)}$ and $\bar{x}^{(i)} + \frac{1}{p}$ with the zero level attained in the center of the interval. A simple computation shows that for declaring that our goal is not reached it is enough to choose $\epsilon = 2\frac{M}{2}\left(\frac{R}{2p}\right)^2$. This contradiction shows that $N \geq p^n$.

Note that each component of the constructed vector field is a function of one variable. Therefore this field has a potential. □

It is interesting to compare our results with the bound $O(\frac{f_0 - f^*}{\epsilon^{3/2}})$ for $n = 1$. For this case, we have the bound $\left(\frac{MR^2}{4\epsilon}\right)^{1/2}$. The difference seems to be very big. However, the apparent contradiction is resolved by the fact that in our example $f_0 - f^* = O(\epsilon)$.

Yurii Nesterov, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 34 voie du Roman Pays, 1348, Louvain-la-Neuve, Belgium. nesterov@core.ucl.ac.be

References

[1] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer, Boston, 2004.

[2] Yu. Nesterov, B. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, **108**(1), 177–205 (2006).

[3] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, **112**(1), 159–181 (2008).

Announcements

MIP 2012

You are cordially invited to participate in the upcoming workshop in Mixed Integer Programming (MIP 2012). The 2012 Mixed Integer Programming workshop will be the ninth in a series of annual

workshops held in North America designed to bring the integer programming community together to discuss very recent developments in the field. The workshop is especially focused on providing opportunities for junior researchers to present their most recent work. The workshop series consists of a single track of invited talks.

Confirmed Speakers

Shabbir Ahmed (Georgia Tech), Gennadiy Averkov (Otto-von-Guericke-Universität Magdeburg), Sam Burer (The University of Iowa), Philipp Christophel (SAS), Jesús A. De Loera (University of California, Davis), Alberto Del Pia (ETH Zurich), Friedrich Eisenbrand (EPFL), Ricardo Fukasawa (University of Waterloo), Vineet Goyal (Columbia University), Marcos Goycoolea (Universidad Adolfo Ibañez), Yongpei Guan (University of Florida), Volker Kaibel (Otto-von-Guericke-Universität Magdeburg), Kiavash Kianfar (Texas A&M University), Mustafa Kilinc (University of Pittsburgh), Fatma Kilinc-Karzan (Carnegie Mellon University), David Morton (The University of Texas at Austin), Ted Ralphs (Lehigh University), Edward Rothberg (Gurobi Optimization), Siqian Shen (University of Michigan), Renata Sotirov (Tilburg University), Dan Steffy (ZIB and Oakland University), Alejandro Toriello (University of Southern California), Christian Wagner (ETH Zurich),

The workshop is designed to provide ample time for discussion and interaction between the participants, as one of its aims is to facilitate research collaboration. Thanks to the generous support by our sponsors, registration is free.

MIP 2012 Organizing Committee
mip2012@math.ucdavis.edu

Further information: http://www.math.ucdavis.edu/mip2012/

ICCOPT 2013

The Fourth International Conference on Continuous Optimization, will take place in Lisbon, Portugal, from July 27 to August 1, 2013. ICCOPT is a recognized forum of discussion and exchange of ideas for researchers and practitioners in continuous optimization, and one of the flagship conferences of the Mathematical Optimization Society.

ICCOPT 2013 is organized by the Department of Mathematics of FCT, Universidade Nova de Lisboa, in its Campus de Caparica, located near a long beach, 15 minutes away by car (and 30 by public transportation) from the center of Lisbon, on the opposite side of the river Tagus.

The three previous versions of ICCOPT were organized respectively in 2004 at Rensselaer Polytechnic Institute (Troy, NY, USA), in 2007 at McMaster University (Hamilton, Ontario, Canada), and in 2010 at University of Chile (FCFM, Santiago, Chile).

ICCOPT 2013 includes a Conference and a Summer School. The Conference (July 29 to August 1) comprehends a series of invited lectures, organized and contributed sessions, and poster sessions. The Summer School (July 27–28) is directed to PhD students and young researchers in the field of continuous optimization.

The meeting is chaired by Luis Nunes Vicente (Organizing Committee) and Katya Scheinberg (Program Committee) and locally coordinated by Paula Amaral (Local Organizing Committee).

Further information: http://eventos.fct.unl.pt/iccopt2013

IMPRINT
Editor: Katya Scheinberg, Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA. katyas@lehigh.edu ■ **Co-Editors:** Samuel Burer, Department of Management Sciences, The University of Iowa, Iowa City, IA 52242-1994, USA. samuel-burer@uiowa.edu ■ Volker Kaibel, Institut für Mathematische Optimierung, Otto-von-Guericke Universität Magdeburg, Universitätsplatz 2, 39108 Magdeburg, Germany. kaibel@ovgu.de ■ **Founding Editor:** Donald W. Hearn ■ Published by the Mathematical Optimization Society ■ Design and typesetting by Christoph Eyrich, Berlin, Germany. optima@0x45.de ■ Printed by Oktoberdruck AG, Berlin, Germany.