A Note About The Complexity Of Minimizing Nesterov's Smooth Chebyshev-Rosenbrock Function

by C. Cartis, N. I. M. Gould and Ph. L. Toint

Report NAXYS-20-2011 15 July 2011

University of Namur
61, rue de Bruxelles, B5000 Namur (Belgium)

http://www.fundp.ac.be/sciences/naxys

A Note About The Complexity Of Minimizing Nesterov's Smooth Chebyshev-Rosenbrock Function

C. Cartis*, N. I. M. Gould† and Ph. L. Toint‡

15 July 2011

Abstract

This short note considers and resolves the apparent contradiction between known worst-case complexity results for first and second-order methods for solving unconstrained smooth nonconvex optimization problems and a recent note by Jarre (2011) implying a very large lower bound on the number of iterations required to reach the solution's neighbourhood for a specific problem with variable dimension.

Keywords: evaluation complexity, worst-case analysis, nonconvex optimization.

1 Introduction

The worst-case complexity of algorithms for unconstrained nonconvex smooth optimization has recently been intensively studied by several authors. In particular, we refer the reader to Vavasis (1993), Nesterov (2004) and Cartis, Gould and Toint (2010*b*) for an analysis of steepest descent, to Gratton, Sartenaer and Toint (2008) and Cartis, Gould and Toint (2011*c*) for trust-regions algorithms, to Cartis et al. (2010*b*) for Newton's method, to Nesterov and Polyak (2006), Cartis, Gould and Toint (2010*a*, 2011*b*, 2011*c*) for regularized variants, or to Vicente (2010) and Cartis, Gould and Toint (2010*c*) for finite-difference and/or derivative-free schemes. The common feature of all these contributions is that they discuss upper (and sometimes lower) bounds on the number of function evaluations that are necessary for the algorithm under consideration to produce an approximate first-order critical point, that is an iterate at which the Euclidean norm of the objective function's gradient is below some user-prescribed tolerance ϵ. Remarkably, these results show that such bounds have the form

$$\left\lceil \frac{\kappa}{\epsilon^{\alpha}} \right\rceil \tag{1.1}$$

where κ is a problem-dependent constant and α is an algorithm-dependent constant ranging between 3/2 and 2. These bounds are often sharp (Cartis et al., 2010*b*) and are optimal for some regularization methods (Cartis et al., 2011*b*). It is important for our purposes to note that κ typically depends, possibly exponentially, on problem dimension via the relevant gradient and perhaps Hessian global Lipschitz constants (which are assumed to exist). We also note that all the algoritms considered in these results are descent methods, in the sense that they generate a sequence of iterates with non-increasing objective function values.

An interesting development occured when F. Jarre recently published a report (Jarre, 2011) where he pointed out that, on a specific problem with variable dimension, any descent algorithm would require a number of iterations (and hence of function evaluations) which is exponential in problem dimension to reach the (unique) critical point. Since ϵ and α in (1.1) are independent of dimension, this behaviour

*School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland, UK. Email: coralia.cartis@ed.ac.uk

†Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK. Email: nick.gould@sftc.ac.uk

‡Namur Center for Complex Systems (naXys) and Department of Mathematics, FUNDP-University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@fundp.ac.be

could easily be made compatible with the results mentioned above if the problem's Lipschitz constants depended exponentially on dimension on the domain of interest. However, it turns out that, for the considered example, both these constants are *independent* of problem size, implying that the bound (1.1) is also independent of problem dimension. It is the purpose of this short note to resolve this apparent contradiction.

2 Some details

We first need to elaborate on the details of the context. In what follows, we consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ x \in \mathbb{R}^n & \end{array}$$

where f is a twice continuously differentiable possibly nonconvex function from $\mathbb{R}^n$ to $\mathbb{R}$, which is assumed to be bounded below (by some value $f_{\rm low}$). To solve this problem, we may then apply the ARC algorithm, which can be outlined as follows. At iteration k, a step s_k from the current iterate x_k is computed, which (approximately) minimizes the cubic model

$$m(x_k + s) = \langle g(x_k), s\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle + \tfrac{1}{6}\sigma_k\|s\|^2,$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are the Euclidean inner product and norm, respectively, $g(x) \stackrel{\rm def}{=} \nabla_x f(x)$, $H(x) = \nabla_{xx} f(x)$ and $\sigma_k \geq \sigma_{\min} > 0$ is an adaptive regularization parameter whose value is recurred inside the algorithm. The step s_k may be successful (it $f(x_k + s_k) \leq f(x_k) + \eta\, m(x_k + s_k)$ for some $\eta \in (0, 1)$), in which case it is accepted as the next iterate, or unsuccessful, in which case it is rejected and the regularization parameter suitably increased. Further details of the algoritm are irrelevant here. Crucially for our purposes, it has been proved (see Nesterov and Polyak, 2006, Cartis et al., 2010*a*) that, if we assume that $H(x)$ is Lipshitz continuous (with constant L) on each of the segments $[x_k, x_k + s_k]$ and if we define an ϵ-approximate critical iterate as an iterate x_k such that

$$\|g(x_k)\| \leq \epsilon, \tag{2.1}$$

where $\epsilon \in (0, 1)$ is a user-specified accuracy, then the ARC algorithm started from the initial point x_0 will produce such an iterate in at most

$$\left\lceil (f(x_0) - f_{\rm low})\frac{\kappa_{\rm ARC}}{\epsilon^{3/2}} \right\rceil \tag{2.2}$$

iterations. The constant $\kappa_{\rm ARC}$ only depends (sublinearly) on L and an upper bound on $\|H(x)\|$ on the segments $[x_k, x_k + s_k]$, as well as on fixed, dimension independent, algorithmic parameters (such as η and $\sigma_{\min}$). We will also make use of a property of the ARC algorithm, namely that, for all $k \geq 0$,

$$\|s_k\| \leq 3\max\left[\frac{\|H(x_k)\|}{\sigma_k}, \sqrt{\frac{\|g(x_k)\|}{\sigma_k}}\right], \tag{2.3}$$

(see Lemma 1.1 of Cartis, Gould and Toint, 2011*a*).

Jarre's example of slow minimization uses the Chebyshev-Rosenbrock function attributed to Nesterov in Gurbuzbalaban and Overton (2011), which is defined, for some $\rho > 0$ and $n \geq 2$, by

$$f(x) = \tfrac{1}{4}(x_1 - 1)^2 + \rho\sum_{i=1}^{n-1}(x_{i+1} - 2x_i^2 + 1)^2 \stackrel{\rm def}{=} \tfrac{1}{4}(x_1 - 1)^2 + \rho\sum_{i=1}^{n-1} v_i(x)^2, \tag{2.4}$$

and whose gradient is given by

$$g_1(x) = \tfrac{1}{2}(x_1 - 1) - 8\rho x_1 v_1(x) \tag{2.5}$$

$$g_i(x) = 2\rho(v_{i-1}(x) - 4\,x_i\,v_i(x)), \quad (i = 2, \ldots, n-1), \tag{2.6}$$

and

$$g_n(x) = 2\rho v_{n-1}. \tag{2.7}$$

The nonzero entries of its Hessian are given (up to symmetry) by

$$H_{1,1}(x) = \tfrac{1}{2} - 8\rho v_1(x) + 32\rho x_1^2, \quad H_{1,2}(x) = -4\rho x_1, \tag{2.8}$$

$$H_{i,i}(x) = 2\rho(1 - 4v_i(x) + 16x_i^2), \quad H_{i,i+1}(x) = -8\rho x_i, \quad (i = 2, \ldots, n-1) \tag{2.9}$$

and

$$H_{n,n}(x) = 2\rho, \tag{2.10}$$

while those of its third derivative tensor $T(x)$ are given by

$$T_{1,1,1}(x) = -32\rho x_1, \quad T_{1,1,2}(x) = -8\rho, \quad T_{1,2,1} = -4\rho, \tag{2.11}$$

$$T_{i,i,i}(x) = -16\rho x_i, \quad T_{i,i,i+1}(x) = -8\rho, \quad (i = 2, \ldots, n-1). \tag{2.12}$$

The level contours for this function are shown in Figure 2.1, the leftmost graph showing the levels in the (x_1, x_2) plane and the rightmost the levels in the (x_i, x_{i+1}) plane, for any i between 2 and $n-1$. The unique first- (and second-)order critical point is $x_* = (1, 1, \ldots, 1)^T$, which is marked on the upper right of each graph.

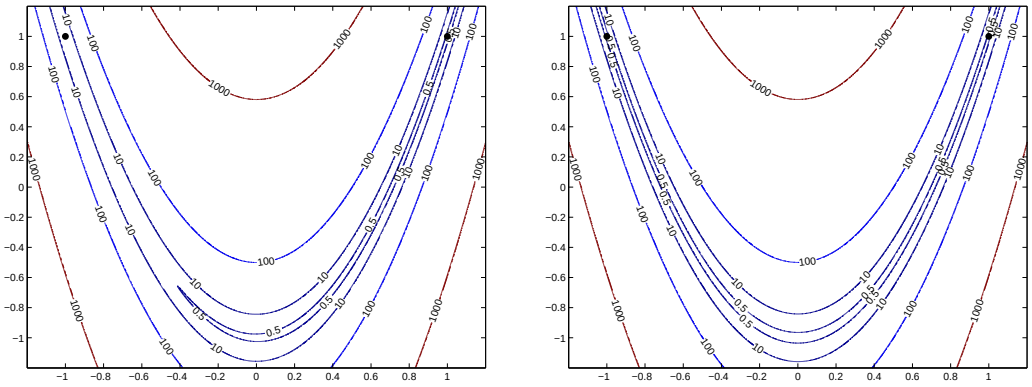

Figure 2.1: Contours of $f(x)$ in the (x_1, x_2) plane (left) and in (x_i, x_{i+1}) plane (for any $2 \leq i \leq n-1$) (right)

The unconstrained minimization of this function is started from $x_0 = (-1, 1, 1, \ldots, 1)^T$ (also marked in the upper left part of the graphs of Figure 2.1), at which $f(x_0) = 1$ and $\|g(x_0))\| = 1$, and for which it is easy to verify that

$$\mathcal{L}_0 \stackrel{\rm def}{=} \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\} \subset [-1, 1]^n.$$

Thus any descent algorithm will remain in $[-1, 1]^n$ and we may therefore derive from (2.5)-(2.7) and the sparse nature of (2.8)-(2.10) that there exist constants $\kappa_g > 0$ and $\kappa_H > 0$ independent of n such that, for iterates generated by the ARC algorithm,

$$\|g(x_k)\| \leq \kappa_g \quad \text{and} \quad \|H(x_k)\| \leq \kappa_H \tag{2.13}$$

for all $k \geq 0$. Moreover, (2.3) then implies that steps generated by the ARC algorithm satisfy the inequality

$$\|s_k\| \leq 3 \max\left[\frac{\kappa_H}{\sigma_{\min}}, \sqrt{\frac{\kappa_g}{\sigma_{\min}}}\right] \stackrel{\rm def}{=} \kappa_s.$$

As a consequence, we obtain that, for all $k \geq 0$,

$$[x_k, x_k + s_k] \subset [-1 - \kappa_s, 1 + \kappa_s]^n \stackrel{\rm def}{=} \mathcal{L}$$

and therefore, using the mean-value theorem, that $H(x)$ is Lipschitz continuous in $\mathcal{L}$ with constant $\max_{x \in \mathcal{L}} \|T(x)\|$, which is itself independent of n because of the sparsity of T (see (2.11)-(2.12)). As

a consequence, the value of κ_{ARC} in (2.2) is independent of n, and, because we may obviously choose $f_{\rm low} = 0$ since $f(x)$ is the sum of squared terms, the upper bound on the maximum number of iterations necessary to achieve (2.1) starting from x_0 is also fixed for given ϵ.

On the other hand, Jarre's observation is that any descent algorithm (including ARC) must take

$$\text{at least} \quad 1.44 \times 1.618^n \quad \text{iterations} \tag{2.14}$$

to move from x_0 to x_*, at which $f(x_*) = 0 = f_{\rm low}$. Moreover, at least half that number of iterations is required to obtain an iterate with $x_1 \geq 0$, which ensures that (2.2) cannot be interpreted as an upper bound on the number of iterations needed to reach an ϵ-dependent neighbouhood of x_*.

3 Resolving the apparent contradiction

We first notice that (2.2) and (2.14) are obviously compatible if

$$\epsilon \leq \left(\frac{\kappa_{\rm ARC}}{1.44 \times 1.618^n} \right)^{2/3} \stackrel{\rm def}{=} \theta(n), \tag{3.1}$$

as in this case the accuracy requirement is tight enough to allow for the number of steps indicated by Jarre's bound. But what happens if (3.1) is violated is not clear. Using the famous Sherlock Holmes adage that "When you have eliminated the impossible, whatever remains, however improbable, must be the truth" (Conan Doyle, 1890), we must conclude in this case that, if an ϵ-approximate first-order critical point can be reached in a dimension independent number of iterations, but that this point cannot be x_*, then it must be that $f(x)$ admits other *approximate* critical points in $\mathcal{L}_0$ within a fixed distance from x_0. And indeed this happens to be the case. The leftmost graph of Figure 3.2 shows (as a continuous line) the evolution with n of

$$\tau(n) = \min_{x \in \{x_1, \ldots, x_{50}\}} \|g(x)\|,$$

where the x_k are the iterates generated by the ARC algorithm applied to minimize $f(x)$ (with dimension n), starting from x_0. The dashed line in the same graph corresponds to the parallel evolution of $\theta(n)$, the right-hand side of (3.1). The distance

$$\delta(n) = \left\| x_0 - \arg \min_{x \in \{x_1, \ldots, x_{50}\}} \|g(x)\| \right\|$$

is shown in the rightmost graph.

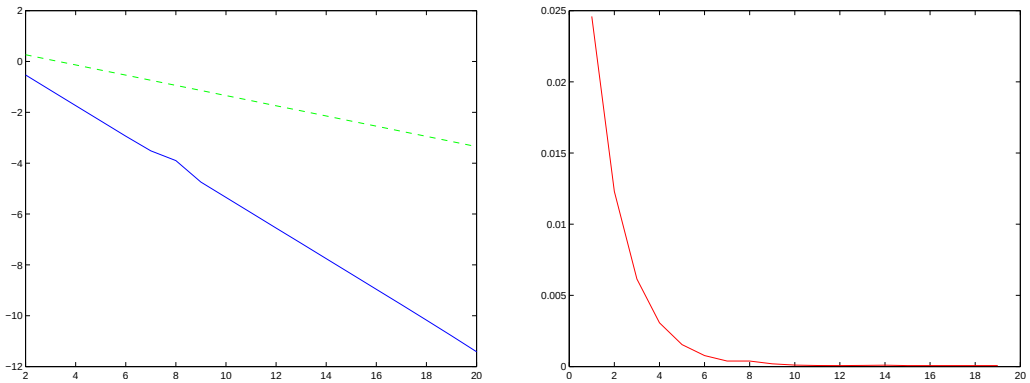

Figure 3.2: Evalution of $\tau(n)$ and $\theta(n)$ (dashed) (left, in $\log_{10}$ scale) and $\delta(n)$ (right) as functions of n

We may conclude from this figure that, for ϵ above the threshold given by (3.1), suitable approximate first-order critical points of $f(x)$ exist close to x_0 (and can be found relatively easily by standard optimization methods). A further investigation of these approximate critical points is possible, using the

analytical expression of $f(x)$. Without entering into too much detail, we may simply say that the gradient norm at such points is dominated by the magnitude of g_n, which is proportional to $|v_{n-1}|$ because of (2.7). As it turns out, (2.6) and the fact that all g_i ($i = 2, \ldots, n-1$) must also be small impose that the $|v_i|$ decrease as an approximate geometric progression. The freedom left for each $|g_i|$ to be small (of the order of $|g_n|$) is enough to counterbalance the effect of x_1 in g_1 given by (2.5). However, this explanation remains problem specific, which considerably limits its interest and applicability.

It remains remarkable that our analysis shows the existence of (potentially many) approximate first-order critical points for a dimension-dependent family of smooth functions for which the gradient and Hessian Lipschitz constants are dimension independent, at a level of approximation which improves exponentially with problem size. It is the authors' view that the implications of this observation (for instance on the geometry of smooth infinite dimensional maps) deserves more study.

Acknowledgements

The work of the second author is funded by EPSRC Grant EP/E053351/1. All three authors are grateful to the Royal Society for its support through the International Joint Project 14265.

References

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 2010*a*. DOI: 10.1007/s10107-009-0337-y.

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*b*.

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. Technical Report naXys-03-2010, Namur Centre for Complex Systems (naXys), FUNDP-University of Namur, Namur, Belgium, 2010*c*.

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

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, **(to appear)**, 2011*b*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Optimal Newton-type methods for nonconvex optimization. Technical Report naXys-17-2011, Namur Centre for Complex Systems (naXys), FUNDP-University of Namur, Namur, Belgium, 2011*c*.

A. Conan Doyle. The sign of the four. *Lippincott's Monthly Magazine*, February 1890.

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.

M. Gurbuzbalaban and M. L. Overton. On Nesterov's nonsmooth Chebyshev-Rosenbrock functions. Optimization Online, 2011. http://www.optimization-online.org/DB_FILE/2011/2923.pdf.

F. Jarre. On Nesterovs smooth Chebyshev-Rosenbrock function. Technical report, University of Düsseldorf, Düsseldorf, Germany, May 2011.

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

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, **108**(1), 177–205, 2006.

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

L. N. Vicente. Worst case complexity of direct search. Technical report, Department of Mathematics, University of Coimbra, Coimbra, Portugal, May 2010.