A (quick) overview of some complexity issues for nonconvex optimization

Philippe Toint (with Coralia Cartis and Nick Gould)

NAXYS, University of Namur, Belgium

(philippe.toint@fundp.ac.be)

NAXYS Opening Day, Namur, October 2010

The nonlinear unconstrained optimization problem

We consider the unconstrained nonlinear programming problem:

$$\text{minimize} \quad f(x)$$

for $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ smooth, possibly nonconvex

Important special case: the nonlinear least-squares problem

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|F(x)\|^2$$

for $x \in \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ smooth.

Applications: model estimation, nonlinear regression, data assimilation in weather forecasting, geological exploration, image reconstruction, etc., etc.

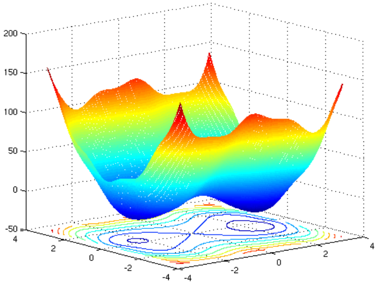
Central to numerical scientific computing !

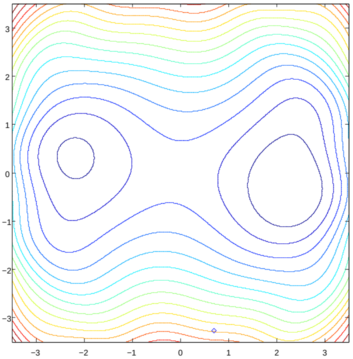
How to solve it?

The main idea: iterative descent methods

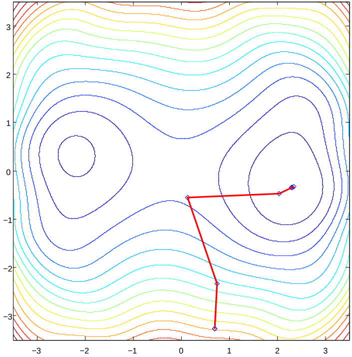
- iterative process generates a sequence of approximate solutions
- each new iterate has a lower value of f than its predecessors
- step based on $\nabla_x f(x_k)$ and (maybe) $\nabla_{xx} f(x_k)$.
- globalization to ensure convergence from arbitrary starting points

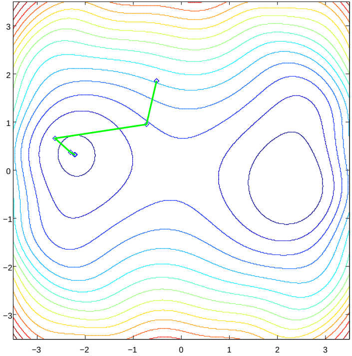
A jungle of *algorithms*!


200
150
100
50
0
-50
4
2
0
-2
-4
-4
-2
0
2
4


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Descent methods: a path towards the lake





When do we stop?

Stop the iteration when

the surface is locally (nearly) flat

i.e. (in maths) when

$$\|\nabla_x f(x_k)\| \le \epsilon$$

($\epsilon \in (0,1)$ is a user-specified accuracy threshold)

A minimization algorithm = a rather complex (discrete) dynamical system moving towards a (possibly very) distant goal

Our question today:

How fast does it get there?

(depends on ϵ, need to count "oracle" calls)

Some notable algorithms (the use of local models)

How to compute the next iterate? Use a local model for f!

- a linear model

$$f(x_k + s) \approx f(x_k) + s^T \nabla_x f(x_k)$$

Cauchy's *steepest descent method*

- a quadratic model

$$f(x_k + s) \approx f(x_k) + s^T \nabla_x f(x_k) + \tfrac{1}{2} s^T \nabla_{xx} f(x_k) s$$

Newton's method

- a quadratic model + bound on the distance

$$f(x_k + s) \approx f(x_k) + s^T \nabla_x f(x_k) + \tfrac{1}{2} s^T \nabla_{xx} f(x_k) s \quad \|s\| \leq \Delta_k$$

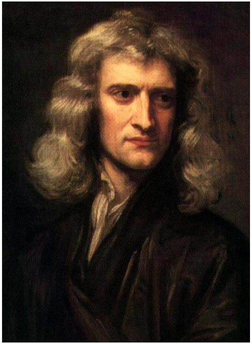
the *trust-region method*

- a quadratic model + cubic penalization of distance

$$f(x_k + s) \approx f(x_k) + s^T \nabla_x f(x_k) + \tfrac{1}{2} s^T \nabla_{xx} f(x_k) s + \tfrac{1}{3} \sigma_k \|s\|^3$$

the *cubic regularization method* (ARC)

Augustin Cauchy (1789-1857) Isaac Newton (1642-1727)



What is known? (1)

How many function evaluations (iterations) (oracle calls) are needed to ensure that $\boxed{\|\nabla_x f(x_k)\| \leq \epsilon?}$

The steepest descent algorithm requires at most

$$\left\lceil \frac{\kappa_C}{\epsilon^2} \right\rceil \text{ function evaluations}$$

(Nesterov)

Newton's method (when convergent) requires at most

??? function evaluations

What is known? (2)

The trust-region algorithm requires at most

??? function evaluations

(Gratton, Sartenaer, T.)

The ARC algorithm requires at most

$$\left\lceil \frac{\kappa_{\rm C}}{\epsilon^{3/2}} \right\rceil \text{ function evaluations}$$

(Nesterov / Cartis, Gould, T.)

Some new results follow... (Cartis, Gould, T.)

Complexity bound for ARC

Is the bound in $O(\epsilon^{-3/2})$ sharp? YES!!! (under reasonable assumptions)

Construct a unidimensional example with

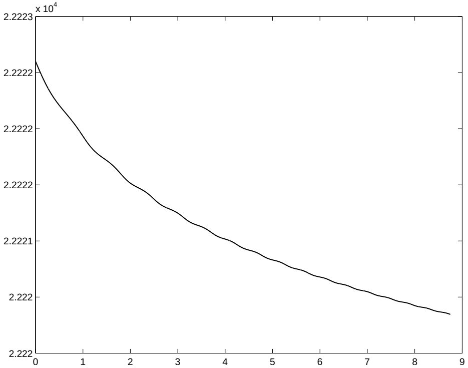
$$x_0 = 0, \quad x_{k+1} = x_k + \left(\frac{1}{k+1}\right)^{\frac{1}{3}+\eta},$$

$$f_0 = \frac{2}{3}\zeta(1+3\eta), \quad f_{k+1} = f_k - \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta},$$

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \quad H_k = 0 \ \text{ and } \ \sigma_k = 1,$$

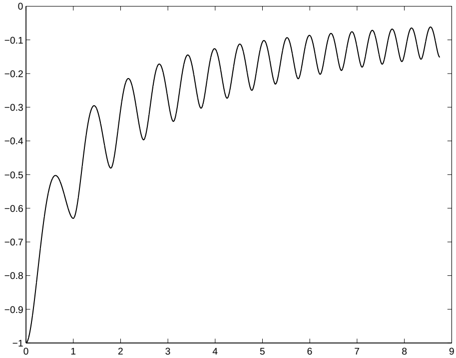
Use Hermite interpolation on $[x_K, x_{k+1}]$.

An example of slow ARC (1)

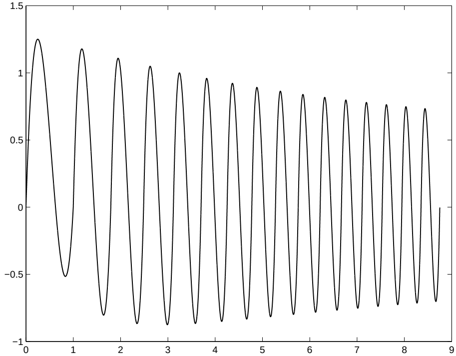


The objective function

An example of slow ARC (2)

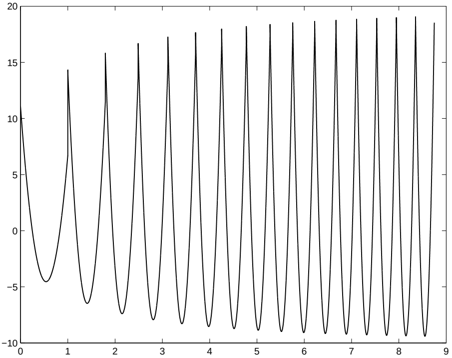


The first derivative



The second derivative

An example of slow ARC (4)



The third derivative

Complexity bound for steepest-descent

Is the bound in $O(\epsilon^{-2})$ sharp? $\boxed{\text{YES!!!}}$ (under reasonable assumptions)

As before, construct a unidimensional example with

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}, \quad \text{and} \quad H_k = 1,$$

$$x_0 = 0, \qquad x_{k+1} = x_k + \alpha_k \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta},$$

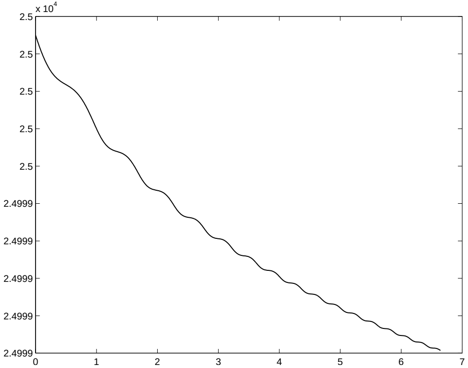
for some steplength $\alpha_k > 0$ such that

$$0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} < 2,$$

giving the step

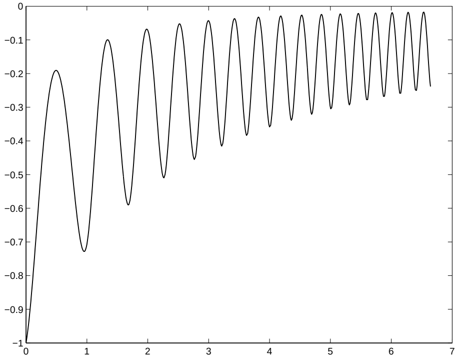
$$s_k \stackrel{\text{def}}{=} x_{k+1} - x_k = \alpha_k \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}.$$

An example of slow steepest descent (1)



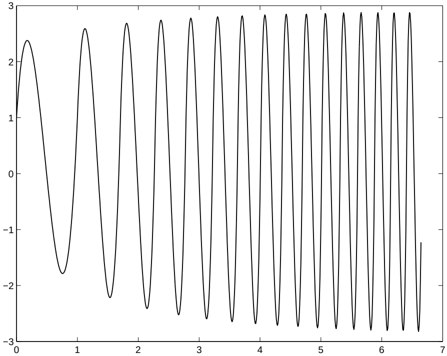
The objective function

An example of slow steepest-descent (2)



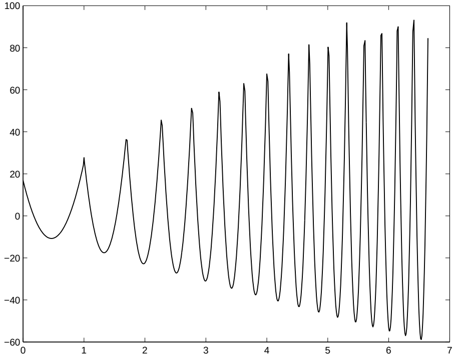
The first derivative

An example of slow steepest-descent (3)



The second derivative

An example of slow steepest descent (4)



The third derivative

Slow Newton

A big surprise:

Newton's method may require as much as

$$\left\lceil \frac{\kappa_C}{\epsilon^2} \right\rceil \text{ function evaluations}$$

to obtain $\|g_k\| \leq \epsilon$ (under reasonable assumptions)

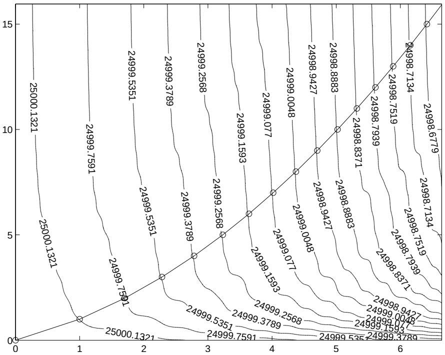
The trsut-region method may require as much as

$$\left\lceil \frac{\kappa_C}{\epsilon^2} \right\rceil \text{ function evaluations}$$

to obtain $\|g_k\| \leq \epsilon$ (under reasonable assumptions)

Example (for both) now bi-dimensional

An example of slow Newton



The path of iterates on the objective's contours

More general second-order methods (work in progress)

Assume that, for $\beta \in (0, 1]$, the step is computed by

$$(H_k + \lambda_k I)s_k = -g_k \;\text{ and }\; 0 \le \lambda_k \le \kappa_s \|s_k\|^\beta$$

(ex: Newton, ARC, (TR), ...)

The corresponding method may require as much as

$$\left\lceil \frac{\kappa_C}{\epsilon^{-(\beta+2)/(\beta+1)}} \right\rceil \text{ function evaluations}$$

to obtain $\|g_k\| \le \epsilon$ on functions with bounded and (segment-wise) β-Hölder continuous Hessians.

Note: ranges form ϵ^{-2} to $\epsilon^{-3/2}$

ARC is optimal within this class

The constrained case

Can we apply the same ideas to the constrained case?

$$\begin{array}{ll} \text{minimize} & f(x) \\ & x \in \mathcal{F} \end{array}$$

for $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ smooth, and where

$$\mathcal{F} \text{ is convex.}$$

Main ideas:

- exploit (cheap) projections on convex sets
- use the same conceptual approach

Constrained step computation (1)

$$\min_s \quad f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle + \tfrac{1}{3}\sigma\|s\|^3$$

subject to

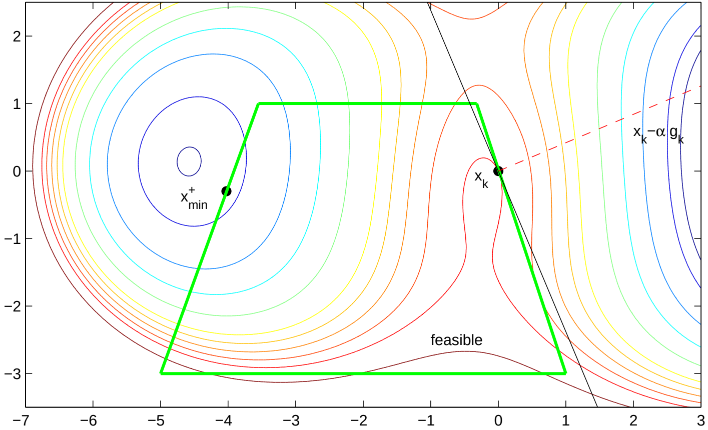
$$x + s \in \mathcal{F}$$

σ is the (adaptive) regularization parameter

Criticality measure: (replaces $\|\nabla_x f(x_k)\| \le \epsilon$)

$$\chi(x_k) \stackrel{\text{def}}{=} \left| \min_{x+d\in\mathcal{F},\|d\|\le 1} \langle \nabla_x f(x_k), d\rangle \right|,$$

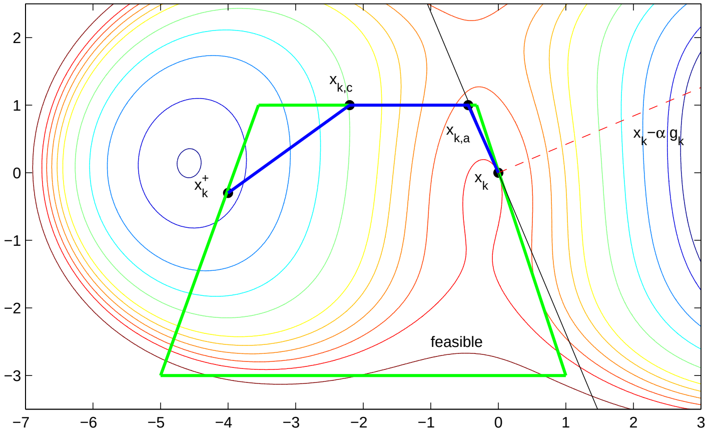
Walking through the pass...



A "beyond the pass" constrained problem with

$$m(x,y) = -x - \frac{42}{100}y - \frac{3}{10}x^2 - \frac{1}{10}y^3 + \frac{1}{3}[x^2+y^2]^{\frac{3}{2}}$$

Walking through the pass...with a sherpa



A piecewise descent path from x_k to x_k^+ on

$$m(x, y) = -x - \tfrac{42}{100} y - \tfrac{3}{10} x^2 - \tfrac{1}{10} y^3 + \tfrac{1}{3}[x^2 + y^2]^{\frac{3}{2}}$$

Function-Evaluation Complexity for COCARC (2)

The COCARC algorithm requires at most

$$\left\lceil \frac{\kappa_{\rm C}}{\epsilon^{3/2}} \right\rceil \text{ function evaluations}$$

(for some $\kappa_{\rm C}$ independent of ϵ) to achieve $\chi(x_k) \leq \epsilon$

Caveat: cost of solving the subproblem

c.f. unconstrained case!!!

Conclusions

- More known on 1rst-order and DFO methods
  (Vicente / Cartis, Gould, T.)
- Many open questions . . . but very interesting
- Algorithm design profits from complexity analysis
- Many issues regarding regularizations still unresolved
- ARC is optimal amongst second-order methods

Many thanks for your attention!

Some references

1. A. S. Nemirovsky and B. B. Yudin,
Problem Complexity and Method Efficiency in Optimization,
Wiley, 1983.
2. Y. Nesterov,
Introductory Lectures on Convex Optimization,
Kluwer, 2004
3. Y. Nesterov and B. T. Polyak,
Cubic regularization of Newton method and its global performance,
Mathematical Programming, 108(1), pp. 177-205, 2006.
4. S. Gratton, A. Sartenaer and Ph. L. Toint,
Recursive Trust-Region Methods for Multiscale Nonlinear Optimization,
SIAM Journal on Optimization, 19(1), pp. 414-444, 2008.
5. 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, to appear, 2010.
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,
FUNDP Techreport, 2009.
7. 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), pp. 2833-2852, 2008.

. . . and more if you are interested.