More on cubic regularization and complexity for nonconvex optimization

Philippe Toint (with Coralia Cartis and Nick Gould)

Department of Mathematics, University of Namur, Belgium

(philippe.toint@fundp.ac.be)

US-Mexico Workshop on Optimization, Oaxaca, January 2011

The 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.

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.

A useful observation

Note the following: if

- f has gradient g and globally Lipschitz continuous Hessian H with constant $2L$

Taylor, Cauchy-Schwarz and Lipschitz imply

$$\begin{aligned} f(x+s) &= f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle \\ &\quad + \int_0^1 (1-\alpha)\langle s, [H(x+\alpha s) - H(x)]s\rangle \, d\alpha \\ &\leq \underbrace{f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle + \tfrac{1}{3}L\|s\|_2^3}_{m(s)} \end{aligned}$$

$\Longrightarrow$ reducing m from $s = 0$ improves f since $m(0) = f(x)$.

The cubic regularization

Change from trust-regions:

$$\min_s \quad f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle \ \text{ s.t. } \ \|s\| \leq \Delta$$

to cubic regularization:

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

σ is the (adaptive) regularization parameter

(ideas from Griewank, Weiser/Deuflhard/Erdmann, Nesterov/Polyak, Cartis/Gould/T)

Cubic regularization highlights

$$f(x+s) \leq m(s) \equiv f(x) + s^T g(x) + \tfrac{1}{2} s^T H(x) s + \tfrac{1}{3} L \|s\|_2^3$$

- Nesterov and Polyak minimize m globally and **exactly**
  - N.B. m may be non-convex!
  - efficient scheme to do so if H has sparse factors
- global (ultimately rapid) convergence to a 2nd-order critical point of f
- better worst-case function-evaluation complexity than previously known

Obvious questions:

- can we avoid the global Lipschitz requirement?
- can we approximately minimize m and retain good worst-case function-evaluation complexity?
- does this work well in practice?

Cubic overestimation

Assume

- $f \in C^2$
- f, g and H at x_k are f_k, g_k and H_k
- symmetric approximation B_k to H_k
- B_k and H_k bounded at points of interest

Use

- cubic overestimating model at x_k

$$m_k(s) \equiv f_k + s^T g_k + \tfrac{1}{2} s^T B_k s + \tfrac{1}{3}\sigma_k \|s\|_2^3$$

  - σ_k is the iteration-dependent regularisation weight
  - easily generalized for regularisation in M_k-norm $\|s\|_{M_k} = \sqrt{s^T M_k s}$ where M_k is uniformly positive definite

Adaptive Regularization with Cubic (ARC)

Algorithm 1.1: The ARC Algorithm

Step 0: Initialization: x_0 and $\sigma_0 > 0$ given. Set $k = 0$

Step 1: Step computation: Compute s_k for which $\boxed{m_k(s_k) \leq m_k(s_k^c)}$

Cauchy point: $s_k^c = -\alpha_k^c g_k$ & $\alpha_k^c = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$

Step 2: Step acceptance: Compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}$

and set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

Step 3: Update the regularization parameter:

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & = \frac{1}{2}\sigma_k \text{ if } \rho_k > 0.9 & \textit{very successful} \\ [\sigma_k, \gamma_1 \sigma_k] & = \sigma_k \text{ if } 0.1 \leq \rho_k \leq 0.9 & \textit{successful} \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & = 2\sigma_k \text{ otherwise} & \textit{unsuccessful} \end{cases}$$

Minimizing the model

$$m(s) \equiv f + s^T g + \tfrac{1}{2} s^T B s + \tfrac{1}{3}\sigma \|s\|_2^3$$

- **Small problems:**

  use Moré-Sorensen-like method with modified secular equation
  (also OK as long as factorization is feasible)

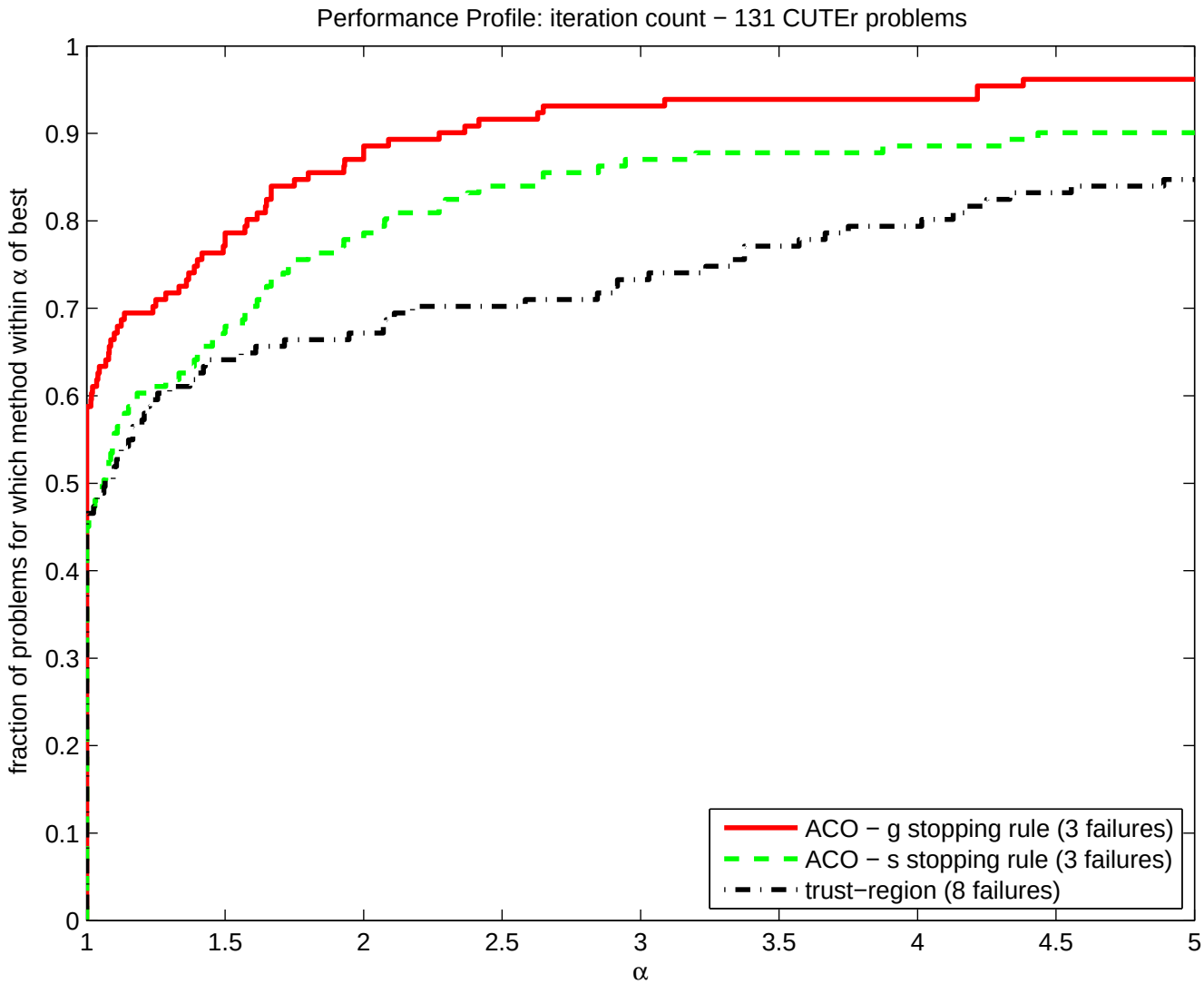
- **Large problems:**

  an iterative Krylov space method

  approximate solution

Numerically sound procedures for computing exact/approximate steps

Numerical experience — small problems using Matlab



Local convergence theory for cubic regularization (1)

The Cauchy condition:

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{\text{CR}} \|g_k\| \min \left[\frac{\|g_k\|}{1 + \|H_k\|}, \sqrt{\frac{\|g_k\|}{\sigma_k}} \right]$$

The bound on the stepsize:

$$\|s_k\| \leq 3 \max \left[\frac{\|H_k\|}{\sigma_k}, \sqrt{\frac{\|g_k\|}{\sigma_k}} \right]$$

(Cartis/Gould/T)

Local convergence theory for cubic regularization (2)

And therefore...

$$\lim_{k \to \infty} \|g_k\| = 0$$

first-order global convergence

Under stronger assumptions can show that

If s_k minimizes m_k over subspace with orthogonal basis Q_k,

$$\lim_{k \to \infty} Q_k^T H_k Q_k \succeq 0$$

second-order global convergence

Fast convergence

For fast asymptotic convergence $\Longrightarrow$ need to improve on Cauchy point: minimize over Krylov subspaces

- g stopping-rule: $\|\nabla_s m_k(s_k)\| \leq \min(1, \|g_k\|^{\frac{1}{2}})\|g_k\|$
- s stopping-rule: $\|\nabla_s m_k(s_k)\| \leq \min(1, \|s_k\| \quad)\|g_k\|$

If B_k satisfies the Dennis-Moré condition

$$\|(B_k - H_k)s_k\|/\|s_k\| \to 0 \text{ whenever } \|g_k\| \to 0$$

and $x_k \to x_*$ with positive definite $H(x_*)$

$\Longrightarrow$ Q-superlinear convergence of x_k under the g- and s-rules

If additionally $H(x)$ is locally Lipschitz around x_* and

$$\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$$

$\Longrightarrow$ Q-quadratic convergence of x_k under the s-rule

First-order function-evaluation complexity (1)

How many function evaluations (iterations) are needed to ensure that

$$\boxed{\|g_k\| \leq \epsilon?}$$

So long as for very successful iterations $\sigma_{k+1} \leq \gamma_3 \sigma_k$ for $\gamma_3 < 1$

The basic ARC algorithm requires at most

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

for some κ_C independent of ϵ and n

c.f. steepest descent

First-order function-evaluation complexity (2)

How many function evaluations (iterations) are needed to ensure that

$$\|g_k\| \leq \epsilon?$$

If H is globally Lipschitz, the s-rule is applied and additionally s_k is the global (line) minimizer of $m_k(\alpha s_k)$ as a function of α, the ARC algorithm requires at most

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

for some κ_S independent of ϵ and n.

c.f. Nesterov & Polyak

Note: an $O(\epsilon^{-3})$ bound holds for convergence to second-order critical points.

First-order function-evaluation complexity (3)

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

Construct a unidimensional example with

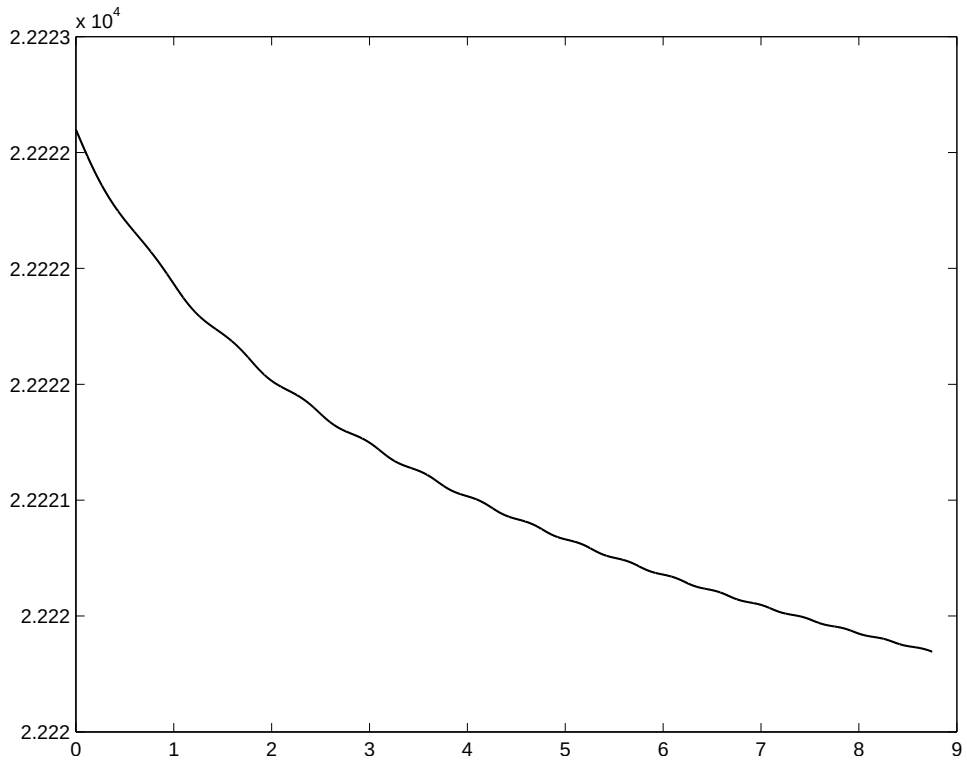
$$x_0 = 0, \quad x_{k+1} = x_k + \left(\frac{1}{k+1}\right)^{\frac{1}{3}+\eta},$$

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

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

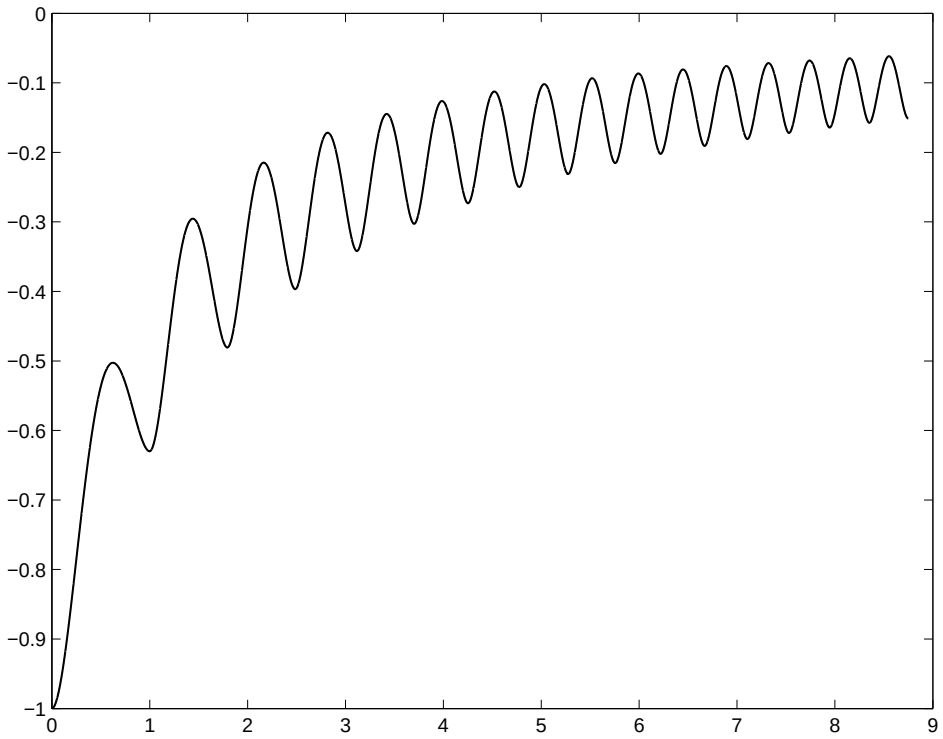
Use Hermite interpolation on $[x_K, x_{k+1}]$.

An example of slow ARC (1)



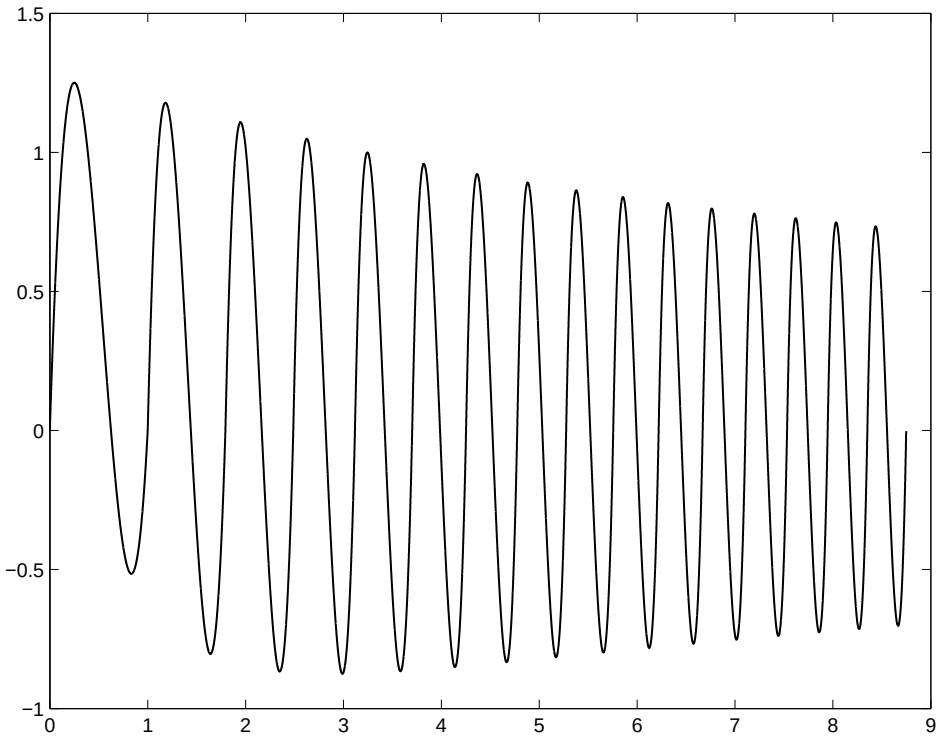
The objective function

An example of slow ARC (2)



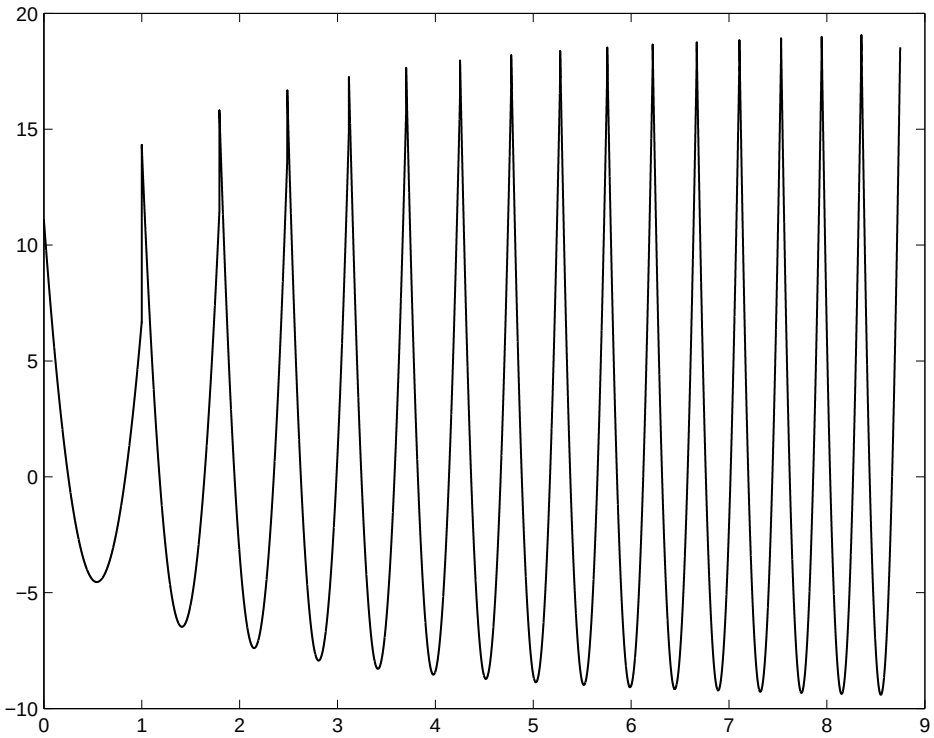
The first derivative

An example of slow ARC (3)



The second derivative

An example of slow ARC (4)



The third derivative

The constrained case

Can we apply regularization to the constrained case?

Consider the constrained nonlinear programming problem:

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

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
- define using the generalized Cauchy point idea
- prove global convergence + function-evaluation complexity

Constrained step computation

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

subject to

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

σ is the (adaptive) regularization parameter

Criticality measure: (as before)

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

The generalized Cauchy point for ARC

Cauchy step: Goldstein-like piecewise linear seach on m_k along the gradient path projected onto $\mathcal{F}$

Find

$$x_k^{\mathrm{GC}} = P_{\mathcal{F}}[x_k - t_k^{\mathrm{GC}} g_k] \stackrel{\mathrm{def}}{=} x_k + s_k^{\mathrm{GC}} \quad (t_k^{\mathrm{GC}} > 0)$$

such that

$$m_k(x_k^{\mathrm{GC}}) \leq f(x_k) + \kappa_{\mathrm{ubs}} \langle g_k, s_k^{\mathrm{GC}} \rangle \qquad \text{(below linear approximation)}$$

and either

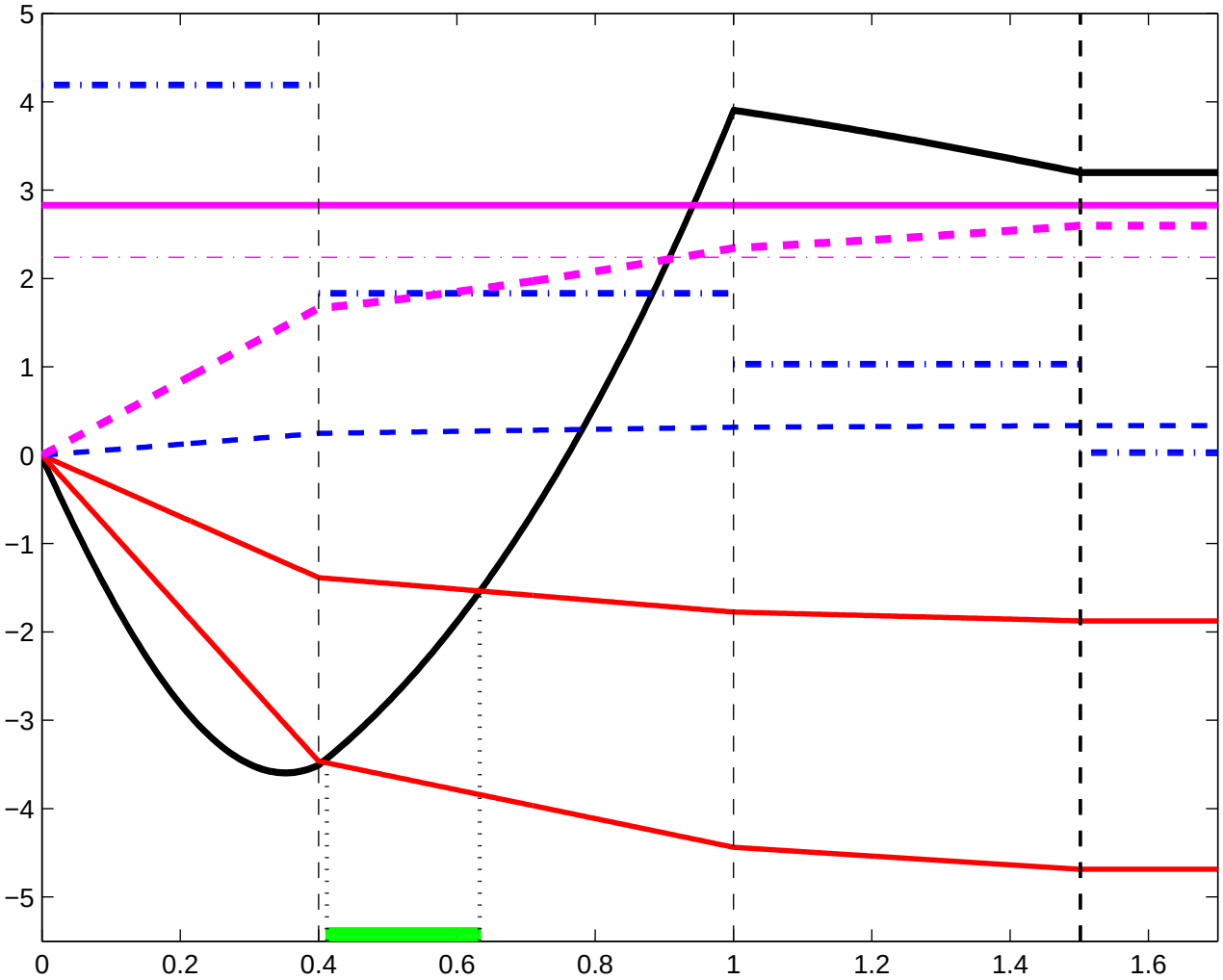
$$m_k(x_k^{\mathrm{GC}}) \geq f(x_k) + \kappa_{\mathrm{lbs}} \langle g_k, s_k^{\mathrm{GC}} \rangle \qquad \text{(above linear approximation)}$$

or

$$\|P_{T(x_k^{\mathrm{GC}})}[-g_k]\| \leq \kappa_{\mathrm{epp}} |\langle g_k, s_k^{\mathrm{GC}} \rangle| \qquad \text{(close to path's end)}$$

no trust-region condition!

Searching for the ARC-GCP



$$m_k(0+s) = -3.57s_1 - 1.5s_2 - s_3 + s_1s_2 + 3s_2^2 + s_2s_3 - 2s_3^2 + \tfrac{1}{3}\|s\|^3 \quad \text{such that} \quad s \leq 1.5$$

A constrained regularized algorithm

Algorithm 2.1: ARC for Convex Constraints (COCARC)

Step 0: Initialization. $x_0 \in \mathcal{F}$, σ_0 given. Compute $f(x_0)$, set $k = 0$.

Step 1: Generalized Cauchy point. If x_k not critical, find the generalized Cauchy point x_k^{GC} by piecewise linear search on the regularized cubic model.

Step 2: Step calculation. Compute s_k and $x_k^+ \stackrel{\rm def}{=} x_k + s_k \in \mathcal{F}$ such that $m_k(x_k^+) \leq m_k(x_k^{GC})$.

Step 3: Acceptance of the trial point. Compute $f(x_k^+)$ and ρ_k. If $\rho_k \geq \eta_1$, then $x_{k+1} = x_k + s_k$; otherwise $x_{k+1} = x_k$.

Step 4: Regularisation parameter update. Set

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases}$$

Local convergence theory for COCARC

The Cauchy condition:

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{\text{CR}} \chi_k \min \left[\frac{\chi_k}{1 + \|H_k\|}, \sqrt{\frac{\chi_k}{\sigma_k}}, 1 \right]$$

The bound on the stepsize:

$$\|s_k\| \leq 3 \max \left[\frac{\|H_k\|}{\sigma_k}, \left(\frac{\chi_k}{\sigma_k} \right)^{\frac{1}{2}}, \left(\frac{\chi_k}{\sigma_k} \right)^{\frac{1}{3}} \right]$$

And therefore...

$$\lim_{k \to \infty} \chi_k = 0$$

(Cartis/Gould/T)

Function-Evaluation Complexity for COCARC (1)

But

What about first-order function-evaluation complexity?

If, for very successful iterations, $\sigma_{k+1} \leq \gamma_3 \sigma_k$ for $\gamma_3 < 1$, the COCARC algorithm requires at most

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

(for some κ_C independent of ϵ and n) to achieve $\chi_k \leq \epsilon$

c.f. steepest descent

Do the nicer bounds for unconstrained optimization extend to the constrained case?

Function-evaluation complexity for COCARC (2)

As for unconstrained, impose a termination rule on the subproblem solution:

- Do not terminate solving $\min_{x_k+s\in\mathcal{F}} m_k(x_k+s)$ before

$$\chi_k^{\mathsf{m}}(x_k^+) \leq \min(\kappa_{\text{stop}}, \|s_k\|)\,\chi_k$$

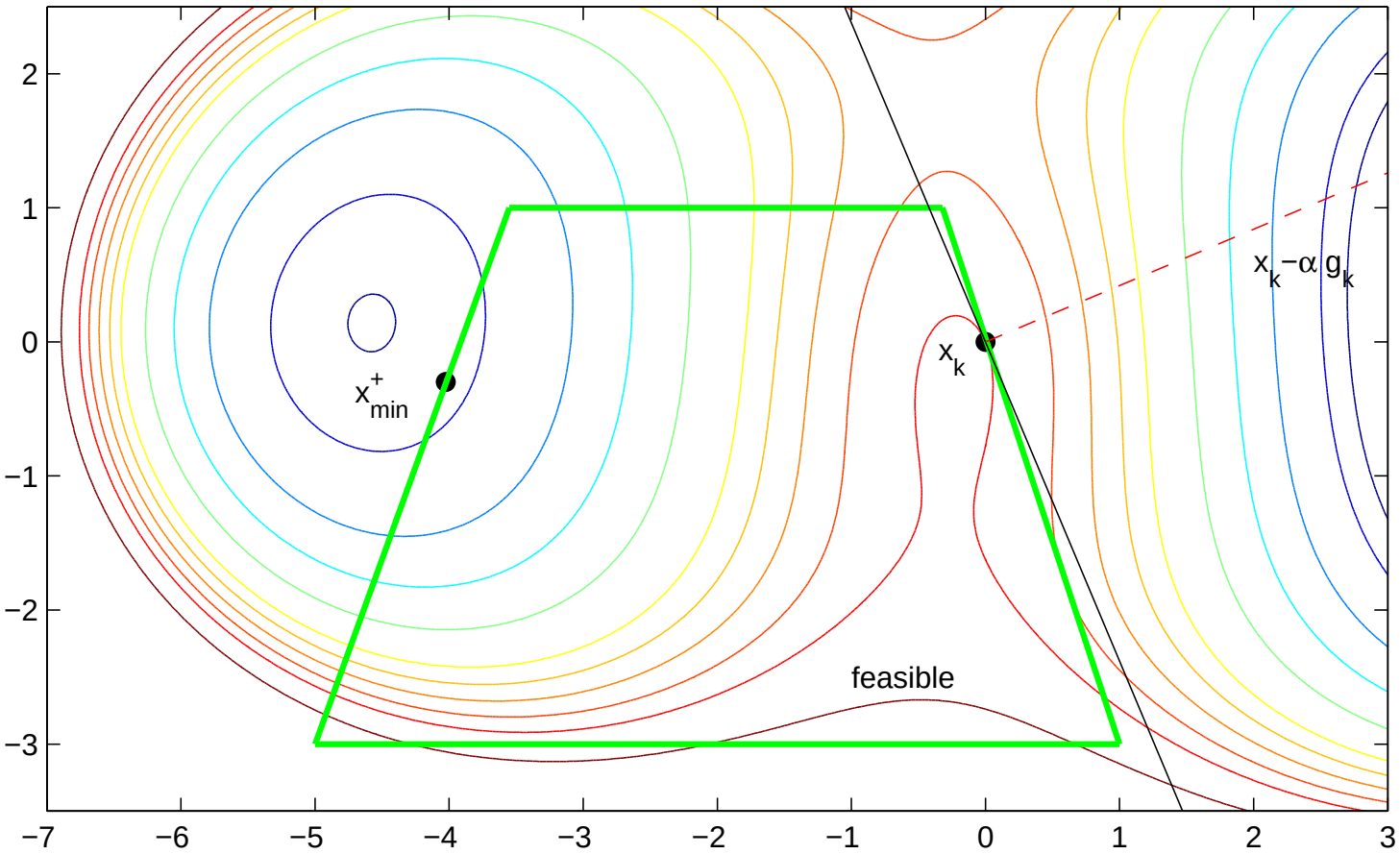
where

$$\chi_k^{\mathsf{m}}(x) \stackrel{\text{def}}{=} \left| \min_{x+d\in\mathcal{F},\|d\|\leq 1} \langle \nabla_x m_k(x), d\rangle \right|$$

c.f. the "s-rule" for unconstrained

Note: OK at local constrained model minimizers

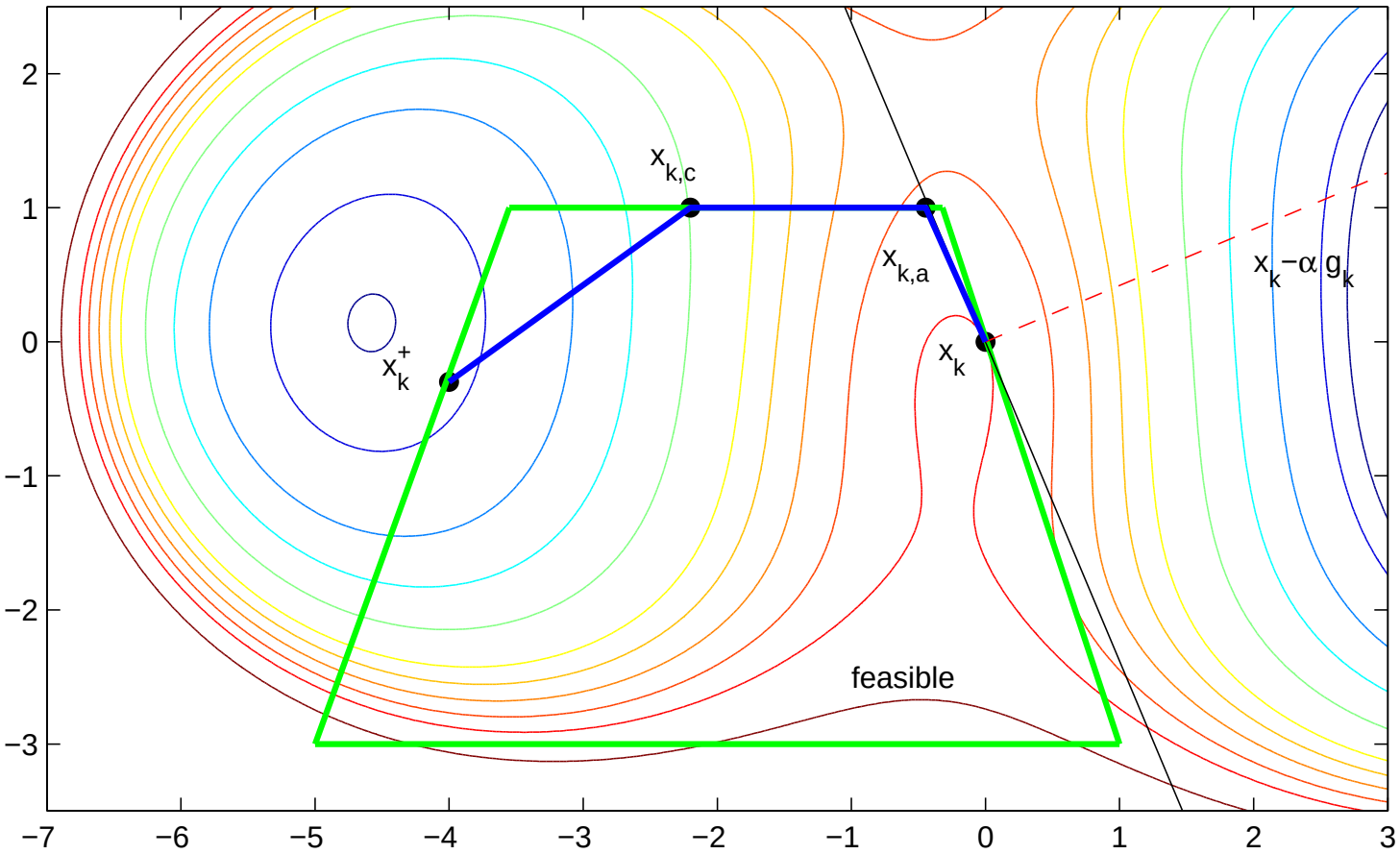
Walking through the pass...



A "beyond the pass" constrained problem with

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

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)

Assume also

- $x_k \leftarrow x_k^+$ in a bounded number of feasible descent substeps
- $\|H_k - \nabla_{xx} f(x_k)\| \leq \kappa \|s_k\|^2$
- $\nabla_{xx} f(\cdot)$ is globally Lipschitz continuous
- $\{x_k\}$ bounded

The COCARC algorithm requires at most

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

(for some κ_C independent of ϵ and n) to achieve $\chi_k \leq \epsilon$

Caveat: cost of solving the subproblem c.f. unconstrained case!!!

Without regularization ?

What is known for unregularized (standard) methods?

The steepest descent method requires at most

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

for obtaining $\|g_k\| \leq \epsilon$.

Sharp???

Newton's method (when convergent) requires at most

??? function evaluations

for obtaining $\|g_k\| \leq \epsilon$.

Slow steepest descent (1)

For steepest descent, the bound of

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

is sharp on functions with Lipschitz continuous gradients.

As before, construct a unidimensional example with

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

for some steplength $\alpha_k > 0$ such that

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

giving the step

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

Slow steepest descent (1)

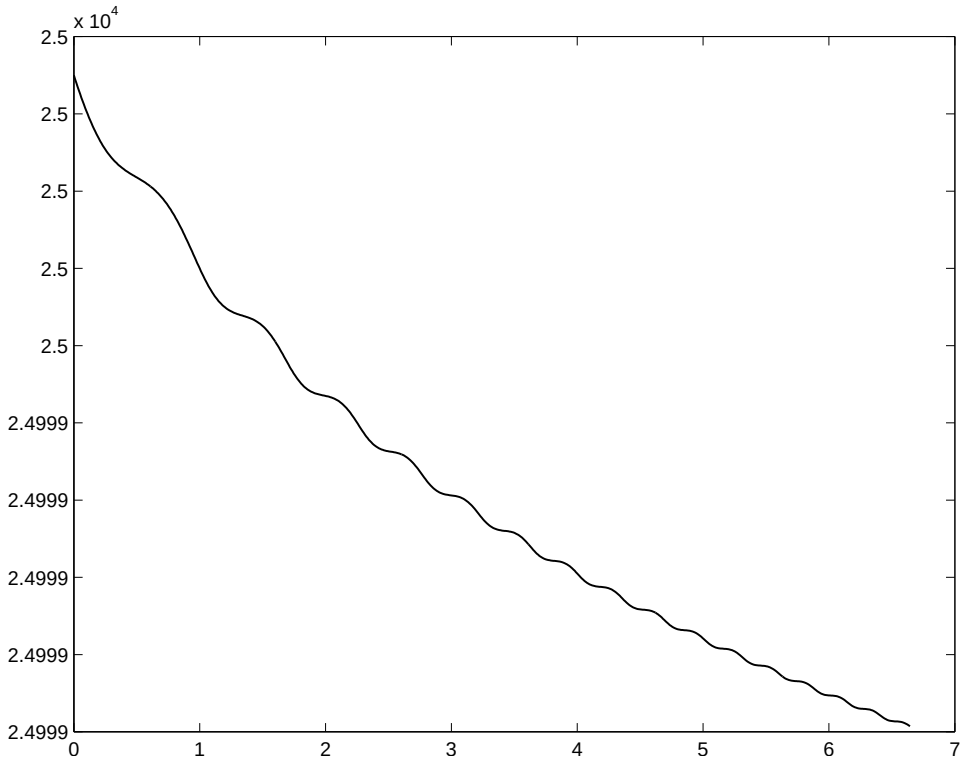
Also set

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

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

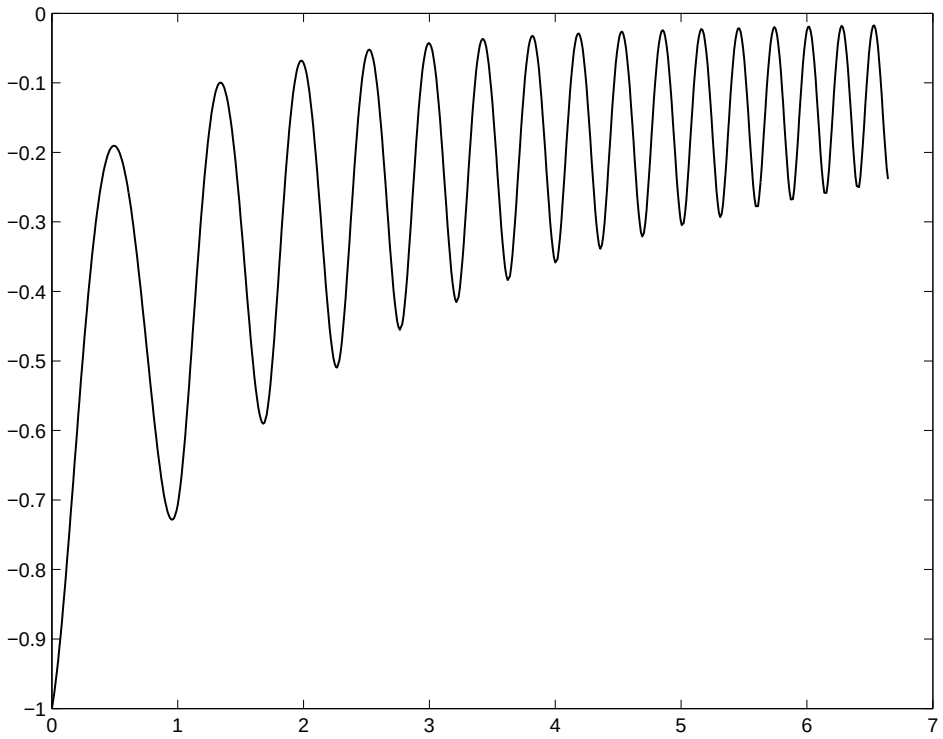
Use Hermite interpolation on $[x_K, x_{k+1}]$.

An example of slow steepest descent (1)



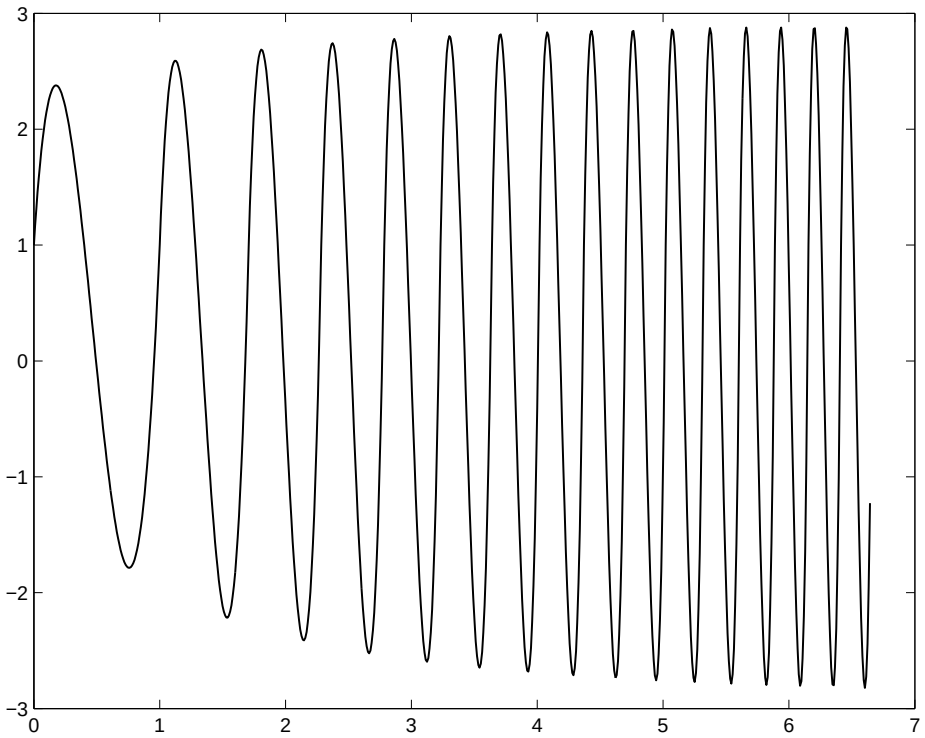
The objective function

An example of slow steepest-descent (2)



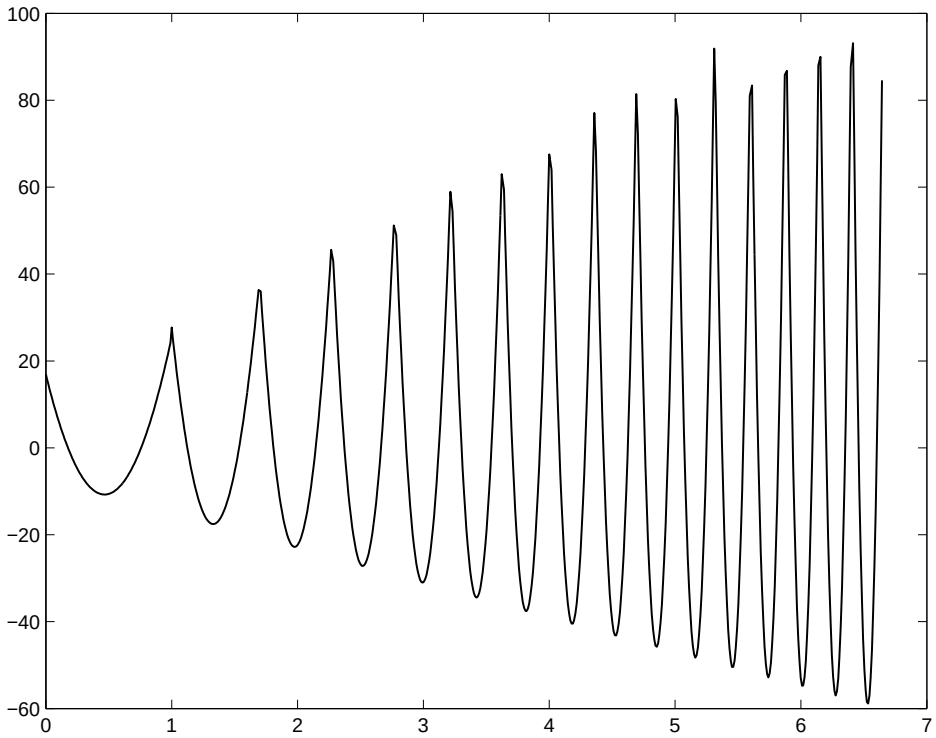
The first derivative

An example of slow steepest-descent (3)



The second derivative

An example of slow steepest descent (4)



The third derivative

Slow Newton (1)

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 $\|g_k\| \leq \epsilon$ on functions with bounded and (segment-wise) Lipschitz continuous Hessians.

Example now bi-dimensional

Slow Newton (2)

The conditions are now:

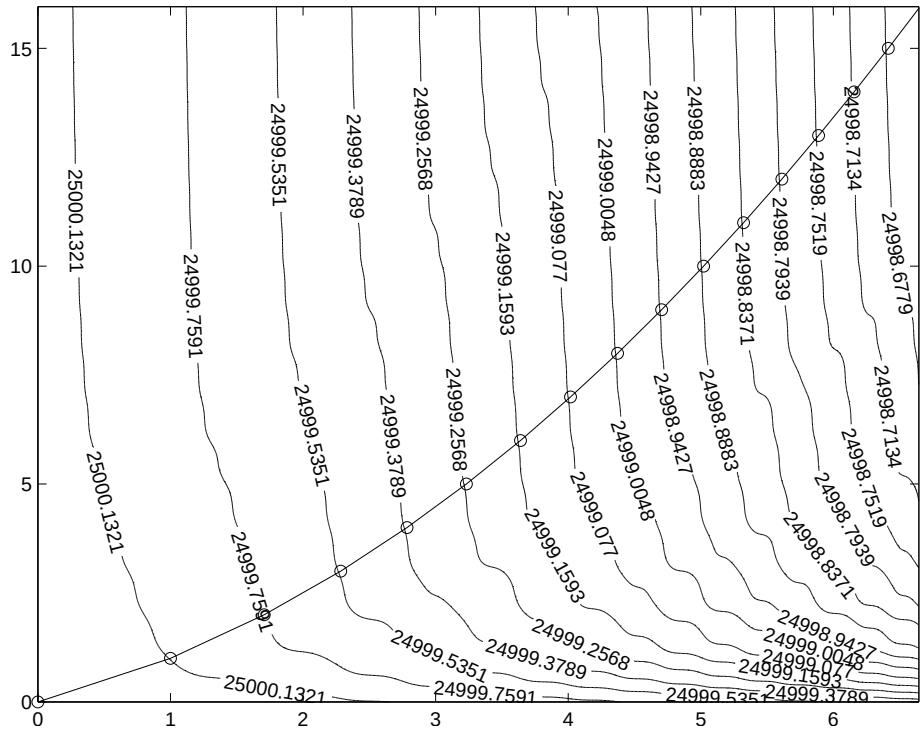
$$x_0 = (0,0)^T, \quad x_{k+1} = x_k + \begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 1 \end{pmatrix},$$

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

$$g_k = -\begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \left(\frac{1}{k+1}\right)^2 \end{pmatrix}, \text{ and } H_k = \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{k+1}\right)^2 \end{pmatrix}$$

Use previous example for x_1 and Hermite interpolation on $[x_K, x_{k+1}]$ for x_2.

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

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

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

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

ARC is optimal within the class corresponding to $\beta = 1$

Finding weak unconstrained minimizers (1)

Now consider finding unconstrained second-order approximate minimizers

$$\|g_k\| \leq \epsilon_g \text{ and } \lambda_{\min}(H_k) \geq -\epsilon_H$$

Require both Cauchy decrease and eigen-point decrease

$$m_k(s_k) \leq \min\left[m_k(s_k^C),\ m_k(s_k^E) \right],$$

whenever $\lambda_{\min}(H_k) < 0$, where

$$s_k^E = \alpha_k^E u_k \quad \text{and} \quad \alpha_k^E = \arg\min_{\alpha} m_k(\alpha u_k), \tag{4.1}$$

with

$$\langle g_k, u_k \rangle \leq 0 \quad \text{and} \quad \langle u_k, B_k u_k \rangle \leq \kappa_{\text{snc}} \lambda_{\min}(H_k) \|u_k\|^2 \tag{4.2}$$

Note: no multidimensional global minimization!

Finding weak unconstrained minimizers (2)

The ARC algorithm requires at most

$$\left\lceil \kappa_{\mathrm{C}} \max \left[\epsilon_g^{-3/2}, \epsilon_H^{-3} \right] \right\rceil \text{ function evaluations}$$

(for some κ_{C} independent of ϵ and n) to achieve

$$\|g_k\| \leq \epsilon \text{ and } \lambda_{\min}(H_k) \geq -\epsilon_H,$$

More surprisingly...

The trust-region algorithm requires at most

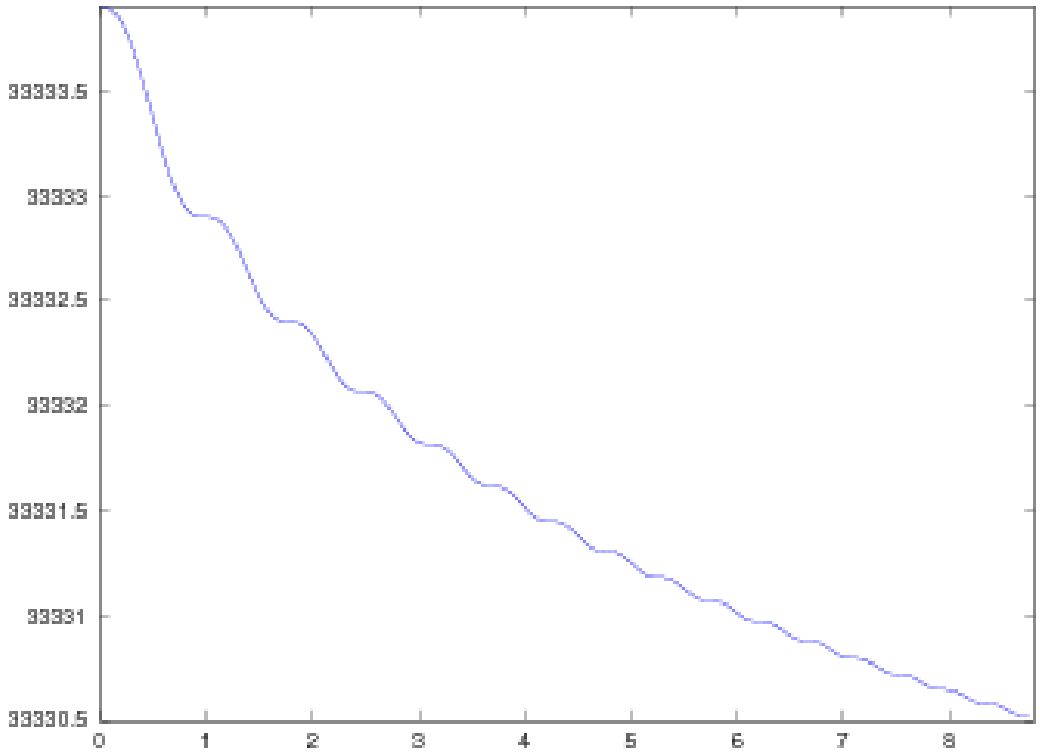
$$\left\lceil \kappa_{\mathrm{TR}}^{\mathrm{2nd}} \max \left[\epsilon_g^{-2} \epsilon_H^{-1}, \epsilon_H^{-3} \right] \right\rceil \text{ function evaluations}$$

(for some κ_{C} independent of ϵ and n) to achieve

$$\|g_k\| \leq \epsilon \text{ and } \lambda_{\min}(H_k) \geq -\epsilon_H$$

An example of slow ARC (2nd-order)

Use Hermite interpolation again with $g_k = 0$ and $\lambda_{\min}(H_k) = -\left(\frac{1}{k+1}\right)^{\frac{1}{3}+\delta}$



The objective function

(applies for both ARC and trust-region)

Finding weak unconstrained minimizers (3)

	ARC	Trust-region
$\epsilon_H \leq \epsilon_g$	$O(\epsilon_H^{-3})$ sharp	$O(\epsilon_H^{-3})$ sharp
$\epsilon_g < \epsilon_H < \sqrt{\epsilon_g}$	$O(\epsilon_H^{-3})$ sharp	$O(\epsilon_H^{-\{3,5\}})$
$\sqrt{\epsilon_g} \leq \epsilon^H$	$O(\epsilon_g^{-3/2})$ sharp	$O(\epsilon_g^{-[2,5/2]})$ "sharp"

Practically sensible: $\epsilon_H \approx \sqrt{\epsilon_g}$

In brief...

Not covered here

- Extension of some of these results to finite-differences (multiplied by n for FDG, by n^2 for DFO)
- Stronger first-order results for the convex case ($O(\log\log\epsilon)$)

Conclusions

- Many open questions ... and very interesting
- ARC is an optimal second-order method
- Algorithm design profits from complexity analysis

Many thanks for your attention!