Optimal Newton-type methods for nonconvex smooth optimization

Coralia Cartis (University of Edinburgh, UK)

joint with

Nick Gould (RAL, UK) & **Philippe Toint** (Namur, Belgium)

WID-DOW Seminar Series

University of Wisconsin, Madison

November 14, 2011

Unconstrained optimization — a “mature” area?

Local unconstrained optimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{where} \quad f \in C^1(\mathbb{R}^n) \text{ or } C^2(\mathbb{R}^n).$$

Currently two main competing methodologies:

- Linesearch methods
- Trust-region methods

Much reliable, efficient software for (large-scale) problems.

- Cubic regularization methods ...
- Is there anything more to say?...

Unconstrained optimization — a “mature” area?

Local unconstrained optimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{where} \quad f \in C^1(\mathbb{R}^n) \text{ or } C^2(\mathbb{R}^n).$$

Currently two main competing methodologies:

- Linesearch methods
- Trust-region methods

Much reliable, efficient software for (large-scale) problems.

- Cubic regularization methods ...

- Is there anything more to say?...
  - Global rates of convergence for optimization algorithms
  - Cubic regularization: better than Newton and steepest descent in the worst-case; optimal in second-order class

Global efficiency of standard methods

- number of function & gradient evaluations $\simeq$ iteration complexity.

Steepest descent method (with linesearch or trust-region):

- $f \in \mathcal{C}^2(\mathbb{R}^n)$ with Lipschitz continuous gradient.
- to generate $\|g(x_k)\| \leq \epsilon$, requires at most

[c.f. Nesterov ('04), Gratton et al. ('08)]

$$\left\lceil \frac{\kappa_{sd}}{\epsilon^2} \right\rceil \quad \textbf{function evaluations.}$$

Global efficiency of standard methods

- number of function & gradient evaluations $\simeq$ iteration complexity.

Steepest descent method (with linesearch or trust-region):

- $f \in \mathcal{C}^2(\mathbb{R}^n)$ with Lipschitz continuous gradient.
- to generate $\|g(x_k)\| \leq \epsilon$, requires at most

[c.f. Nesterov ('04), Gratton et al. ('08)]

$$\left\lceil \frac{\kappa_{sd}}{\epsilon^2} \right\rceil \quad \textbf{function evaluations.}$$

Newton's method:

- when convergent, requires at most

$$??? \quad \textbf{function evaluations.}$$

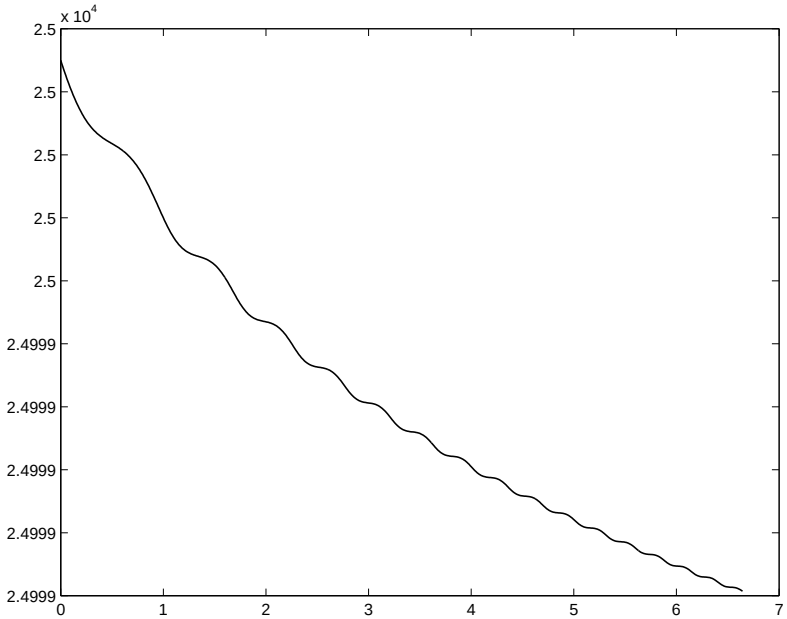
- if Newton step taken within trust-region framework, then steepest descent bound applies.

The worst-case bound is sharp for steepest descent

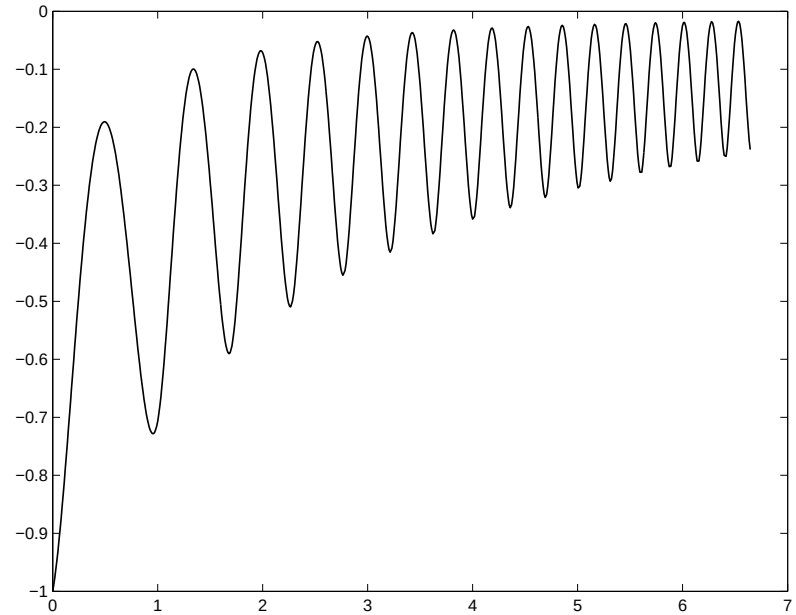
- given x_0, for any $\epsilon > 0$ and $\tau > 0$, steepest descent with (inexact) linesearch applied to f below takes precisely

$$\left\lceil \frac{1}{\epsilon^{2-\tau}} \right\rceil \quad \textbf{function evaluations}$$

to generate $|g(x_k)| \leq \epsilon$.



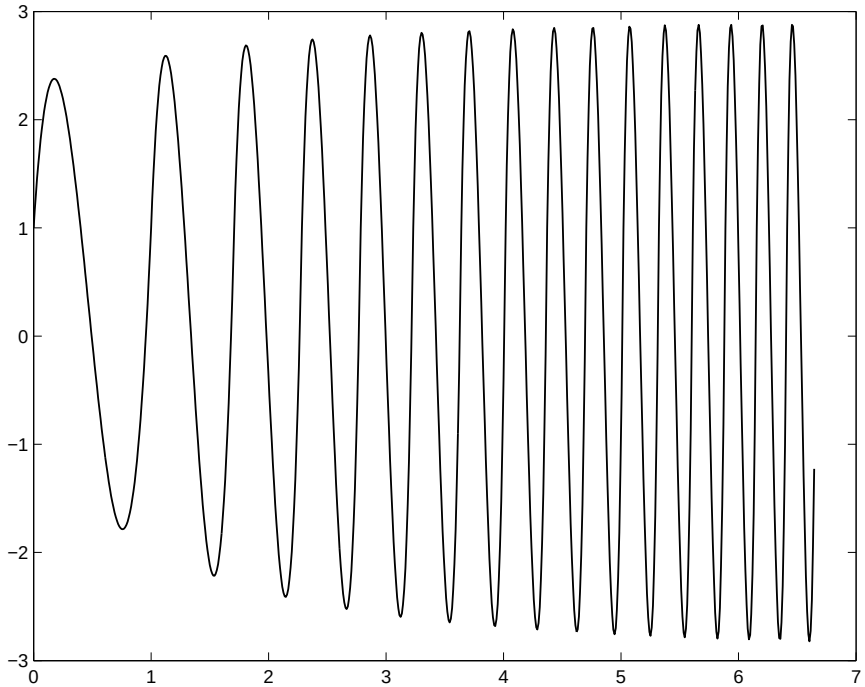
The objective function f.



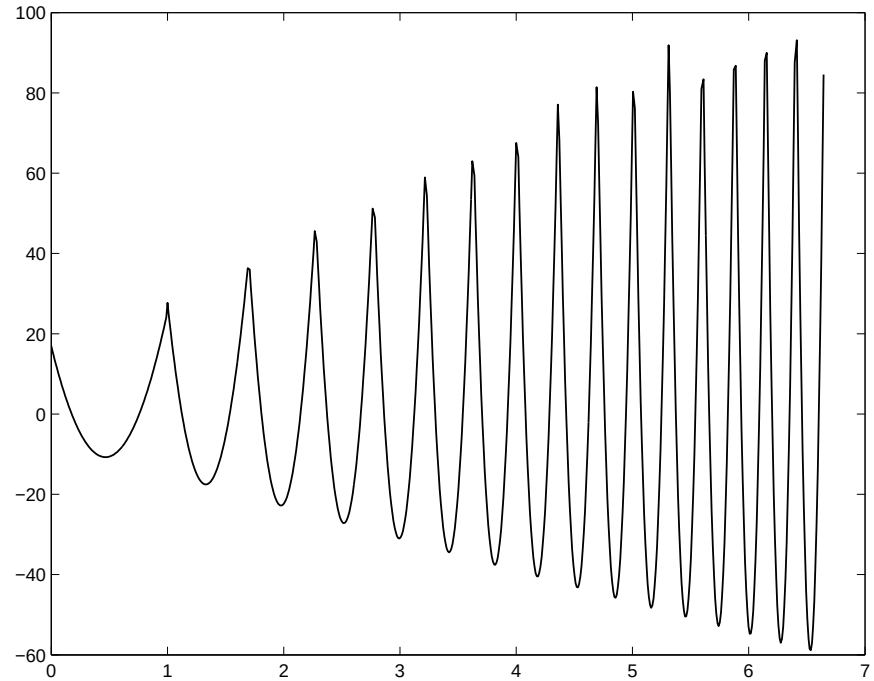
Its gradient g.

The bound is sharp for steepest descent ...

- $f \in \mathcal{C}(\mathbb{R})$ bounded below by zero.
- gradient g is Lipschitz continuous.
- $f(x_k) \to 0 = \inf_{x \in \mathbf{R}} f(x)$, as $k \to \infty$.



The Hessian of f.



The third derivative of f.

The bound is sharp for steepest descent ...

Unidimensional example: $x_0 = 0, \quad 0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} < 2,$

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

where $\eta = \eta(\tau)$ such that $\frac{1}{2} + \eta = \frac{1}{2-\tau}$. Also,

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

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

Use Hermite interpolation on $[x_k, x_{k+1}]$ to construct f s.t.

$$f(x_k) = f_k, \ g(x_k) = g_k \ \text{ and } \ H(x_k) = H_k.$$

$\Longrightarrow \quad k = \left\lceil \frac{1}{\epsilon^{2-\tau}} \right\rceil$ such that $|g_k| \leq \epsilon$.

Newton's method: as slow as steepest descent

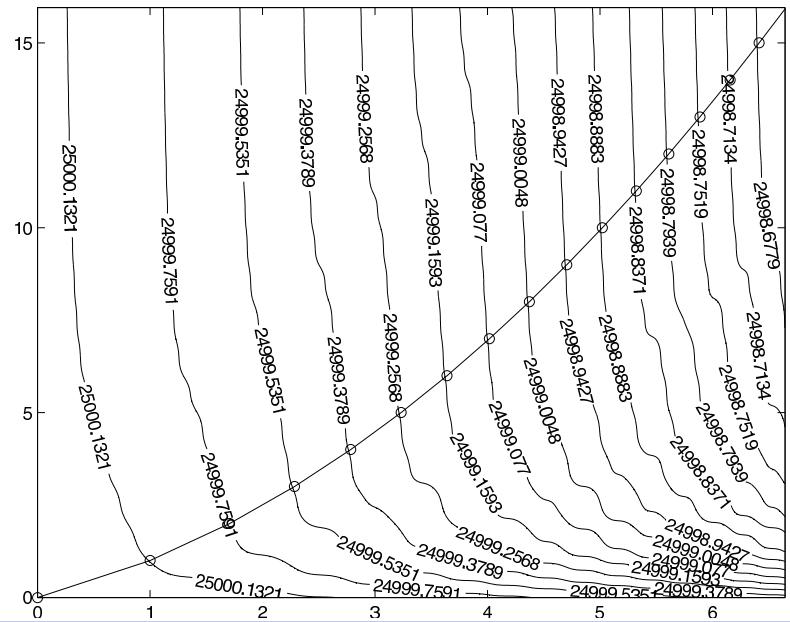
- Newton's method may require as much as

$$\left\lceil \frac{\kappa_{\mathrm{N}}}{\epsilon^{2-\tau}} \right\rceil \quad \textbf{function evaluations}$$

to generate $\|g(x_k)\| \leq \epsilon$, for any ϵ and $\tau > 0$.

- Corollary: trust-region methods' worst-case bound of $\mathcal{O}(\epsilon^{-2})$ is sharp.
- $f \in \mathcal{C}(\mathbb{R}^2)$ with bounded and (segmentwise) Lipschitz continuous Hessian.

Path of iterates and contours of f.



Newton's method: as slow as steepest descent ...

Bidimensional example: $x_0 = (0,0)^T$, $\eta = \eta(\tau)$ s.t. $\frac{1}{2} + \eta = \frac{1}{2-\tau}$;

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

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

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

Use previous example for x^1 (with $\alpha_k = 1$) and Hermite interpolation on $[x_k^2, x_{k+1}^2]$ for $x^2 \Longrightarrow \quad k \geq \left\lceil \frac{1}{\epsilon^{2-\tau}} \right\rceil$ such that $\|g_k\| \leq \epsilon$.

Improved complexity for cubic regularization

- H is globally Lipschitz continuous with Lipschitz constant 2σ:

  Taylor, Cauchy-Schwarz and Lipschitz $\Longrightarrow$

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

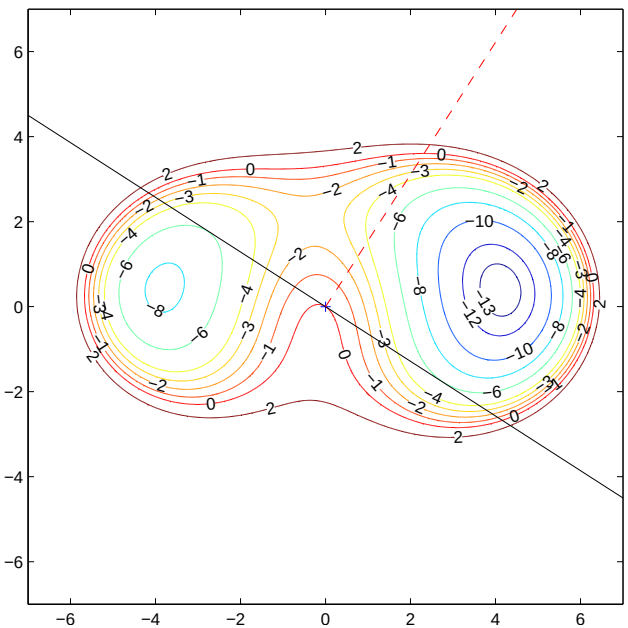
  $\Longrightarrow$ reducing m from $s = 0$ decreases f since $m(0) = f(x)$.

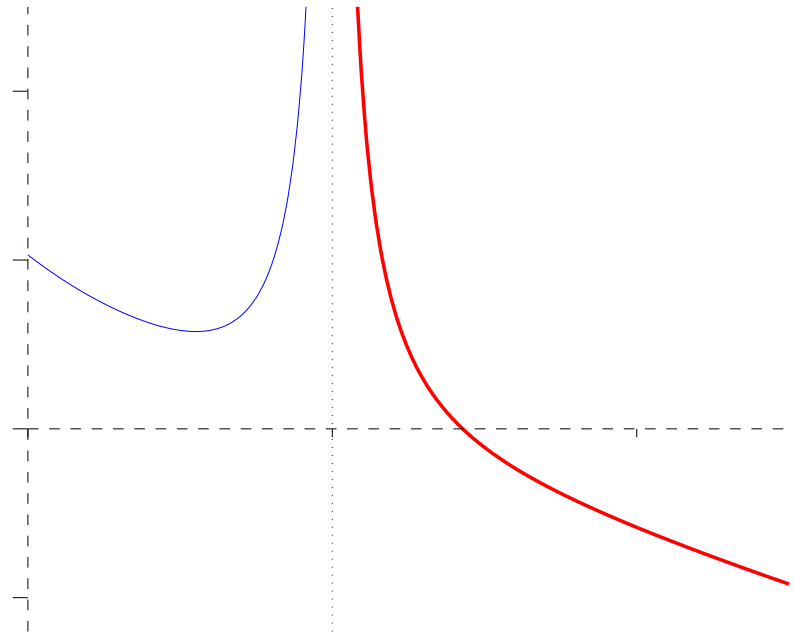
- compute $s_k \longrightarrow \mathbf{min}_s \, m_k(s)$.
  - A. Griewank (1981, technical report).
  - Y. Nesterov & B. Polyak (2006).
  - M. Weiser, P. Deuflhard & B. Erdmann (2007).
  - C. C., N. I. M. Gould & Ph. L. Toint (2009, 2010).

Minimizing the cubic model

- f nonconvex $\longrightarrow$ $m_k(s)$ may be nonconvex!

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





Necessary and sufficient optimality: any **global** minimizer s_* of m satisfies $(H + \lambda_* I)s_* = -g$ and $\lambda_* = \sigma\|s_*\|_2$,

- $H + \lambda_* I$ is positive semidefinite

[cf. Nesterov et al. (2006), CGT (2009)]

Adaptive cubic regularization

Assume

- $f \in C^1(\mathbb{R}^n)$ (maybe $C^2(\mathbb{R}^n)$)
- f, g (and H) at x_k are f_k, g_k (and H_k)
- symmetric approximation B_k (to H_k) ◁
- B_k bounded above independently of k

Use

- **cubic regularization** 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 **regularization weight** ◁

Adaptive Regularization with Cubics (ARC)

Given x_0, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until convergence,

- compute a step s_k for which $m_k(s_k) \leq m_k(s_k^C)$
  - **Cauchy point**: $s_k^C = -\alpha_k^C g_k$ & $\alpha_k^C = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$

Adaptive Regularization with Cubics (ARC)

Given x_0, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until convergence,

- compute a step s_k for which $m_k(s_k) \leq m_k(s_k^C)$
  - **Cauchy point**: $s_k^C = -\alpha_k^C g_k$ & $\alpha_k^C = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}$

Adaptive Regularization with Cubics (ARC)

Given x_0, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until convergence,

- compute a step s_k for which $m_k(s_k) \leq m_k(s_k^C)$
  - **Cauchy point**: $s_k^C = -\alpha_k^C g_k$ & $\alpha_k^C = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}$
- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

Adaptive Regularization with Cubics (ARC)

Given x_0, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until convergence,

- compute a step s_k for which $m_k(s_k) \leq m_k(s_k^C)$
  - **Cauchy point**: $s_k^C = -\alpha_k^C g_k$ & $\alpha_k^C = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}$
- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$
- given $\gamma_2 \geq \gamma_1 > 1$, set

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

Adaptive Regularization with Cubics (ARC)

Given x_0, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until convergence,

- compute a step s_k for which $m_k(s_k) \leq m_k(s_k^C)$
  - **Cauchy point**: $s_k^C = -\alpha_k^C g_k$ & $\alpha_k^C = \arg\min_{\alpha \in \mathbf{R}_+} m_k(-\alpha g_k)$
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}$
- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$
- given $\gamma_2 \geq \gamma_1 > 1$, set

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

[cf. trust-region methods]

Basic ARC: global convergence and complexity

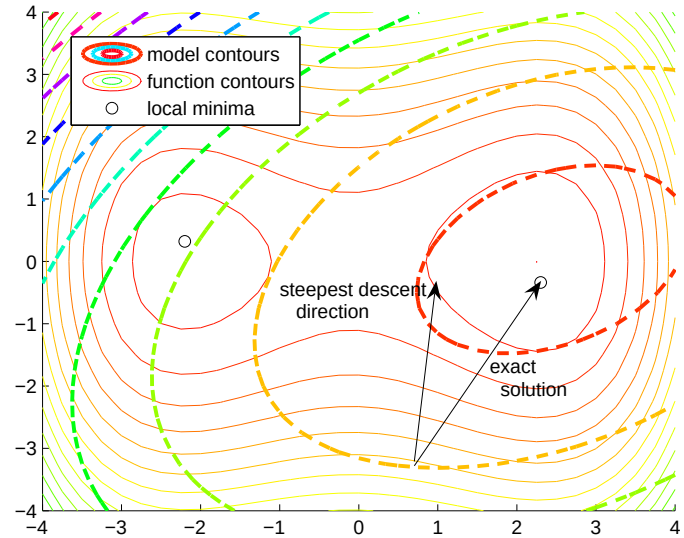
- f bounded below $\implies$ $\liminf_{k \to \infty} \|g_k\| = 0.$
- additionally, if g is uniformly continuous $\implies$ $\lim_{k \to \infty} \|g_k\| = 0.$
- additionally, if g Lipschitz continuous $\implies$ $\mathcal{O}(\epsilon^{-2})$ function/gradient-evaluations for $\|g_k\| \leq \epsilon$.

[cf. steepest-descent-like methods]

Second-order ARC: beyond the Cauchy point

$m_k(s_k) \leq m_k(s_k^C)$ achieved if:

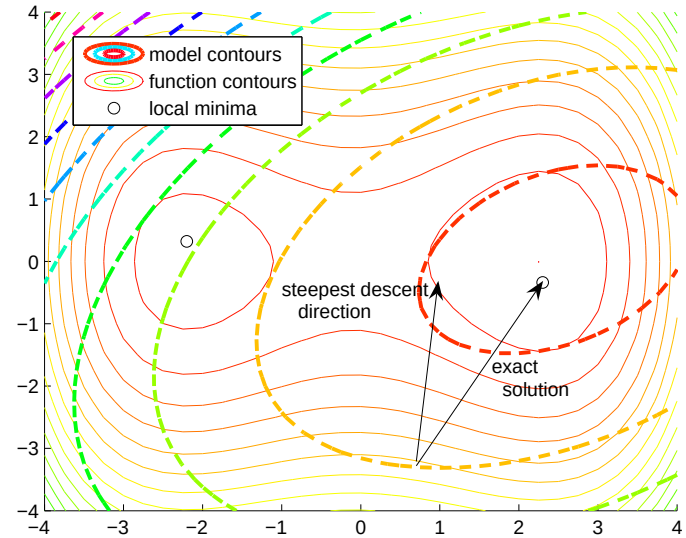
- $s_k = s_k^C \quad \longrightarrow$ inefficient.
- $s_k =$global argmin$_{s \in \mathbf{R}^n}\ m_k(s)$
  $\longrightarrow$ expensive (large-scale).



Second-order ARC: beyond the Cauchy point

$m_k(s_k) \leq m_k(s_k^C)$ achieved if:

- $s_k = s_k^C \longrightarrow$ inefficient.
- $s_k =$global argmin$_{s \in \mathbb{R}^n}\ m_k(s)$ $\longrightarrow$ expensive (large-scale).



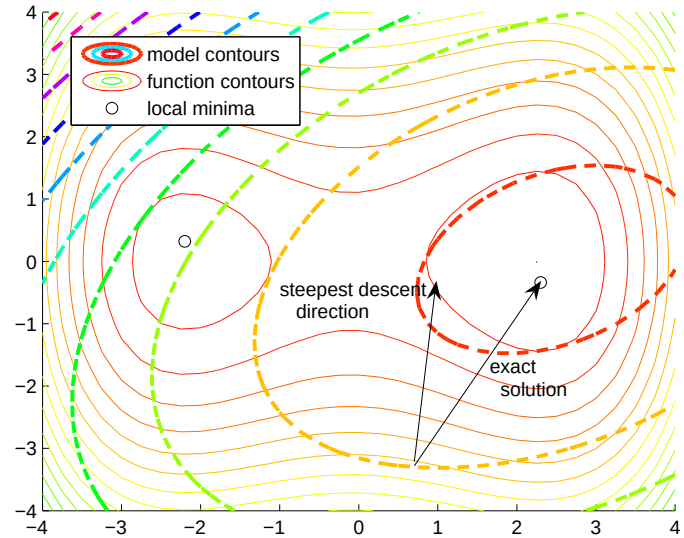
- $\mathbf{ARC_S}$: $s_k =$global min of $m_k(s)$ over $s \in \mathcal{S} \leq \mathbb{R}^n$, where $g \in \mathcal{S}$ $\longrightarrow$ increase subspaces to satisfy termination criteria:

$$\|\nabla_s m_k(s_k)\| \leq \min(1, \|s_k\|)\|g_k\|.$$

Second-order ARC: beyond the Cauchy point

$m_k(s_k) \leq m_k(s_k^C)$ achieved if:

- $s_k = s_k^C \quad \longrightarrow$ inefficient.
- $s_k =$global argmin$_{s \in \mathbb{R}^n}\ m_k(s)$
  $\longrightarrow$ expensive (large-scale).



- $\mathbf{ARC_S}$: $s_k =$global min of $m_k(s)$ over $s \in \mathcal{S} \leq \mathbb{R}^n$, where $g \in \mathcal{S}$
  $\longrightarrow$ increase subspaces to satisfy termination criteria:
  $$\|\nabla_s m_k(s_k)\| \leq \min(1, \|s_k\|)\|g_k\|.$$
  - superlinear local rate with Dennis-Moré on B_k;
    Q-quadratic if H locally Lipschitz continuous at x_*.
  - if H globally Lipschitz continuous: global convergence to 2nd order critical points in the subspaces.

Minimizing the cubic model over subspaces

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

Seek global minimizer of $m(s)$ in a j-dimensional ($j \ll n$) subspace $\mathcal{S} \subseteq \mathbb{R}^n$ with $g \in \mathcal{S}$

- Q orthogonal basis for $\mathcal{S} \Longrightarrow s = Qu$ where

  $$u = \operatorname*{arg\,min}_{u \in \mathbb{R}^j} f + u^T (Q^T g) + \tfrac{1}{2} u^T (Q^T B Q) u + \tfrac{1}{3}\sigma_k \|u\|_2^3$$

  $\Longrightarrow$ use secular equation to find u

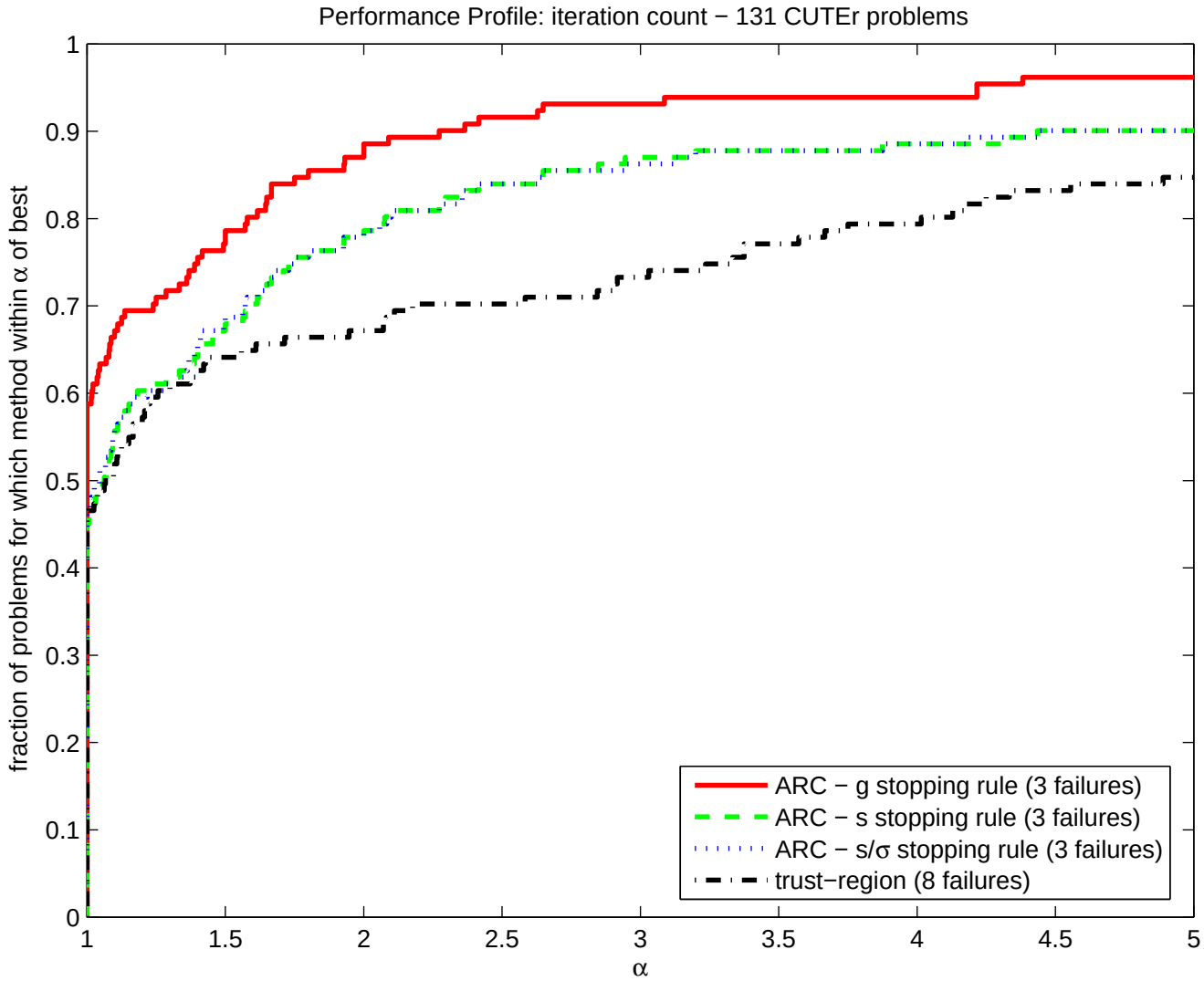
- if $\mathcal{S}$ is the Krylov space generated by $\{B^i g\}_{i=0}^{j-1}$

  $\Longrightarrow$ $Q^T B Q = T$, tridiagonal

  $\Longrightarrow$ can factor $T + \lambda I$ to solve sec. eq. even if j is large

Average-case performance of ARC variants

Preliminary numerical experience — using Matlab



Worst-case performance of ARC$_S$

How many function- & gradient-evaluations are needed to ensure that $\|g_k\| \leq \epsilon$?

Worst-case performance of ARC$_S$

How many function- & gradient-evaluations are needed to ensure that $\|g_k\| \leq \epsilon$?

- if H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$
  $\Longrightarrow$ ARC$_S$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \text{function evaluations.}$$

Worst-case performance of ARC$_S$

How many function- & gradient-evaluations are needed to ensure that $\|g_k\| \leq \epsilon$?

- if H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$
  $\Longrightarrow$ ARC$_S$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

To also ensure that for the same k as above

$$-\lambda_{\min}(Q_k^T B_k Q_k) \leq \epsilon,$$

where Q_k orthogonal basis matrix of minimization subspace
$\Longrightarrow$ ARC$_S$ algorithm requires at most

$$\left\lceil \kappa_{curv} \cdot \epsilon^{-3} \right\rceil \quad \textbf{function evaluations.}$$

[cf. Nesterov & P.]

$\mathbf{ARC_S}$: first-order worst-case bound

If H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$, then $\mathrm{ARC_S}$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

$\mathbf{ARC_S}$: first-order worst-case bound

If H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$, then $\mathrm{ARC_S}$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

For all k successful,

- $f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(s_k)] \geq \frac{\eta_1}{6}\sigma_k\|s_k\|^3$

$\mathbf{ARC_S}$: first-order worst-case bound

If H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$, then $\mathrm{ARC_S}$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

For all k successful,

- $f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(s_k)] \geq \frac{\eta_1}{6}\sigma_k\|s_k\|^3$
- $\|s_k\| \geq C\|g_{k+1}\|^{\frac{1}{2}}$ and $\sigma_k \geq \sigma_{\min} > 0$

$\mathbf{ARC_S}$: first-order worst-case bound

If H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$, then $\mathrm{ARC_S}$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

For all k successful,

- $f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(s_k)] \geq \frac{\eta_1}{6}\sigma_k\|s_k\|^3$
- $\|s_k\| \geq C\|g_{k+1}\|^{\frac{1}{2}}$ and $\sigma_k \geq \sigma_{\min} > 0$

$\Longrightarrow$

$$f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{j}[f(x_k) - f(x_{k+1})] \geq \frac{\eta_1\sigma_{\min}}{6}\sum_{k=0}^{j}\|g_k\|^{\frac{3}{2}}$$

$\mathbf{ARC_S}$: first-order worst-case bound

If H Lipschitz continuous on iterates' path and $\|(B_k - H_k)s_k\| = O(\|s_k\|^2)$, then $\mathrm{ARC_S}$ requires at most

$$\left\lceil \kappa_S \cdot \epsilon^{-3/2} \right\rceil \quad \textbf{function evaluations.}$$

For all k successful,

- $f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(s_k)] \geq \frac{\eta_1}{6}\sigma_k\|s_k\|^3$
- $\|s_k\| \geq C\|g_{k+1}\|^{\frac{1}{2}}$ and $\sigma_k \geq \sigma_{\min} > 0$

$\Longrightarrow$

$$f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{j}[f(x_k) - f(x_{k+1})] \geq \frac{\eta_1\sigma_{\min}}{6}\sum_{k=0}^{j}\|g_k\|^{\frac{3}{2}}$$

While $\|g_k\| \geq \epsilon$,

$$\Longrightarrow\ f(x_0) - f_{\text{low}} \geq |\mathcal{S}_j|\frac{\eta_1\sigma_{\min}}{6}\epsilon^{\frac{3}{2}} \Longrightarrow\ |\mathcal{S}_j| \leq \frac{6(f(x_0)-f_{\text{low}})}{\eta_1\sigma_{\min}}\epsilon^{-\frac{3}{2}}.$$

ARC$_S$: the worst-case first-order bound is sharp

Unidimensional example: $x_0 = 0$, $\eta = \eta(\tau)$ s.t. $\frac{1}{3} + \eta = \frac{1}{3-2\tau}$;

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

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

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

Use Hermite interpolation on $[x_k, x_{k+1}]$ to construct f s.t.

$$f_k = f(x_k), \; g_k = g(x_k) \; \text{ and } \; H_k = H(x_k).$$

$\Longrightarrow \quad k = \left\lceil \epsilon^{-\frac{3}{2}+\tau} \right\rceil$ such that $|g_k| \leq \epsilon$.

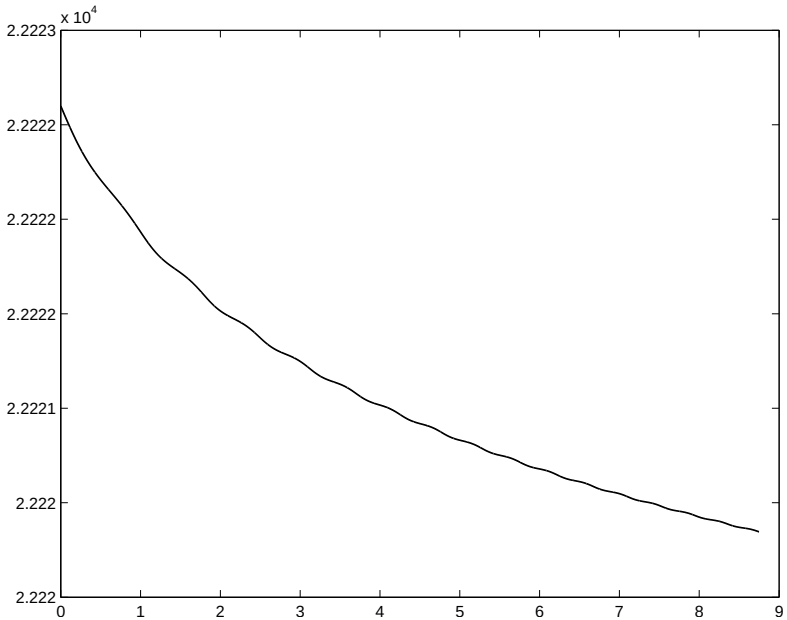
- here, s_k = global minimizer of cubic model $m_k(s)$, $s \in \mathbb{R}^n$.

$\mathbf{ARC_S}$: the first-order bound is sharp ...

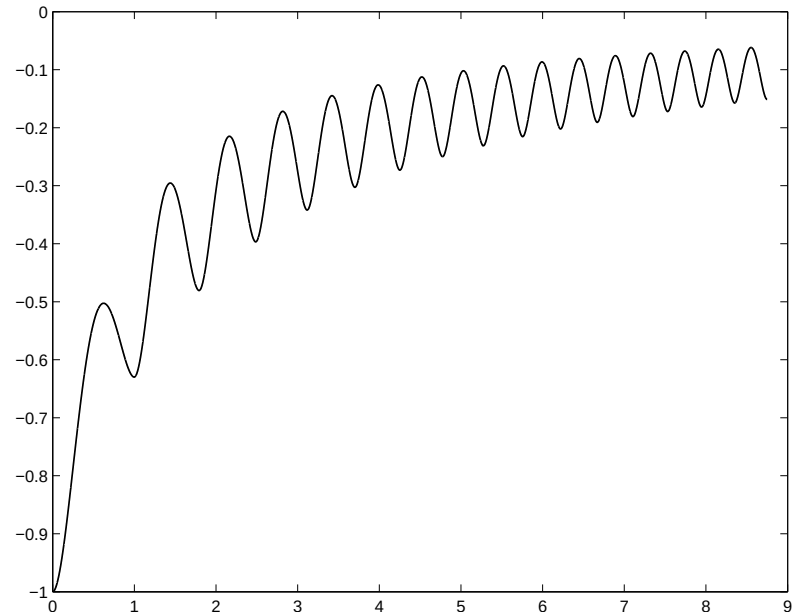
- given x_0, for any $\epsilon > 0$ and $\tau > 0$, $\mathrm{ARC_S}$ applied to f below takes precisely

$$\left\lceil \epsilon^{-\frac{3}{2}+\tau} \right\rceil \quad \textbf{function evaluations}$$

to generate $|g(x_k)| \leq \epsilon$.



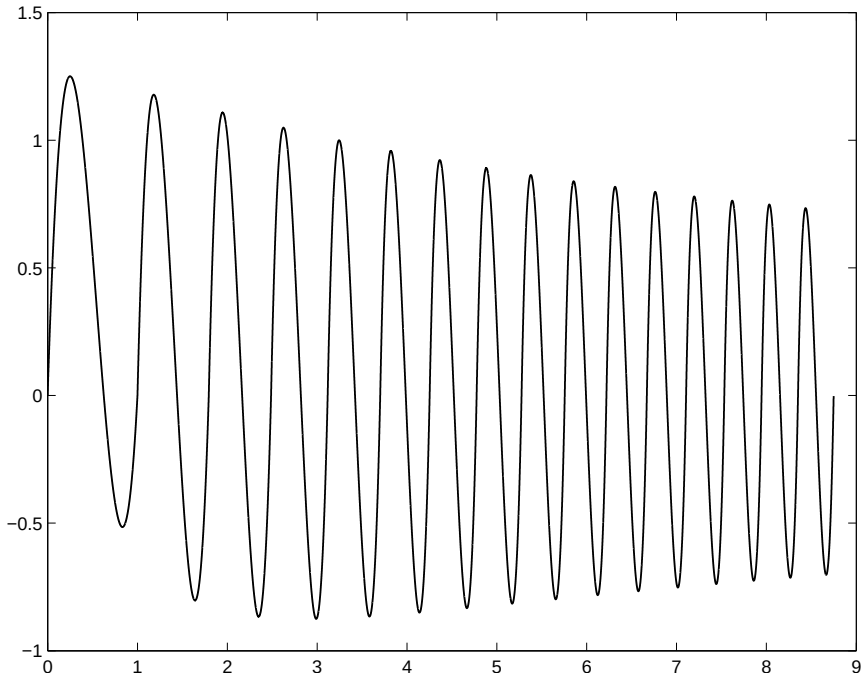
The objective function f.



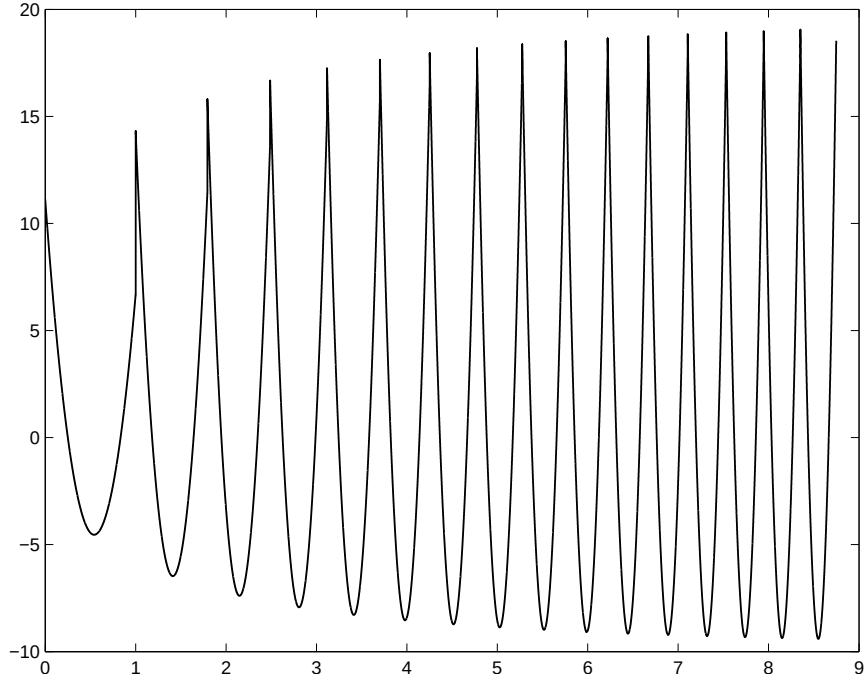
Its gradient g.

ARC$_S$: the first-order bound is sharp ...

- $f \in \mathcal{C}(\mathbb{R})$ bounded below by zero.
- Hessian H is bounded above and Lipschitz continuous on the path of iterates.
- $f(x_k) \to 0 = \inf_{x \in \mathbb{R}} f(x)$, as $k \to \infty$.



The Hessian of f.



The third derivative of f.

A general class of methods and objectives

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$, for some $\alpha \in [0, 1]$.

A general class of methods and objectives

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$, for some $\alpha \in [0, 1]$.

Class of objectives A.α: $f \in \mathcal{C}^2$ bounded below;
g globally Lipschitz continuous and H α-Hölder continuous on the path of the iterates.

- $\alpha \in [0, 1]$; $\alpha = 1$: H Lipschitz-continuous. A.1 $\subset$ A.α.

A general class of methods and objectives

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$, for some $\alpha \in [0, 1]$.

Class of objectives A.α: $f \in \mathcal{C}^2$ bounded below;
g globally Lipschitz continuous and H α-Hölder continuous on the path of the iterates.

- $\alpha \in [0, 1]$; $\alpha = 1$: H Lipschitz-continuous. A.1 $\subset$ A.α.

Properties of class M.α:

- $f \in \mathrm{A}.\alpha$ and $\mathcal{M} \in \mathrm{M}.\alpha \Longrightarrow \|s_k\| \geq C\|g_{k+1}\|^{\frac{1}{1+\alpha}}$.
- $\|g_{k+1}\| \leq c\|g_k\|^{1+\alpha} \Longrightarrow$ lower bound on the step.

Examples of methods in M.α

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$.

Examples of methods in M.α

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$.

Step calculation and sufficient decrease: make use of model

$$m_k(s) = f_k + g_k^T s + \tfrac{1}{2} s^T (H_k + \beta_k I)s,$$

with $\beta_k = \beta_k(s) \geq 0$ and $\beta_k \leq \lambda_k$.

Examples of methods in M.α

Class of methods M.α: $x_{k+1} = x_k + s_k, \quad k \geq 0$;

- $(H_k + \lambda_k I)s_k = -g_k$ with $\lambda_k \geq 0$ and $H_k + \lambda_k I \succeq 0$
- $\|s_k\| \leq \kappa_s$ and $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha$.

Step calculation and sufficient decrease: make use of model

$$m_k(s) = f_k + g_k^T s + \tfrac{1}{2} s^T (H_k + \beta_k I) s,$$

with $\beta_k = \beta_k(s) \geq 0$ and $\beta_k \leq \lambda_k$.

Examples of methods in M.α (applied to functions in A.α):

- Newton's method: $\lambda_k = 0$, $\beta_k = 0$ and $\alpha \in [0, 1]$.
- Regularization methods: $\lambda_k = \sigma_k \|s_k\|^\alpha$, $\beta_k = \frac{2}{2+\alpha} \sigma_k \|s_k\|^{2+\alpha}$

  $\Longrightarrow \alpha = 1$: cubic regularization.

Examples of methods in M.α ...

Examples of methods in M.α (applied to functions in A.α):

- Goldfeld-Quandt-Trotter: $\beta_k = 0$;

$$\lambda_k = \begin{cases} 0, & \text{when } \lambda_{\min}(H_k) \geq R_k \|g_k\|^{\frac{\alpha}{1+\alpha}}; \\ -\lambda_{\min}(H_k) + R_k \|g_k\|^{\frac{\alpha}{1+\alpha}}, & \text{otherwise,} \end{cases}$$

with $R_k > 0$.

- Trust-region methods, when λ_k is at least uniformly bounded above; $\beta_k = 0$.

Examples of methods in M.α ...

Examples of methods in M.α (applied to functions in A.α):

- Goldfeld-Quandt-Trotter: $\beta_k = 0$;

$$\lambda_k = \begin{cases} 0, & \text{when } \lambda_{\min}(H_k) \geq R_k \|g_k\|^{\frac{\alpha}{1+\alpha}}; \\ -\lambda_{\min}(H_k) + R_k \|g_k\|^{\frac{\alpha}{1+\alpha}}, & \text{otherwise,} \end{cases}$$

with $R_k > 0$.

- Trust-region methods, when λ_k is at least uniformly bounded above; $\beta_k = 0$.

Remarks:

- $\lambda_k \leq \kappa_\lambda \|s_k\|^\alpha \Longrightarrow \lambda_k + \lambda_{\min}(H_k) \leq \kappa \max\{|\lambda_{\min}(H_k)|, \|g_k\|^{\frac{\alpha}{1+\alpha}}\}$
- choose $\sigma_k \geq \sigma_{\min} > 0$ and $R_k \geq R_{\min} > 0$.

(Order) optimality of regularization methods

Theorem: Let $\mathcal{M} \in \mathbf{M}.\alpha$. Then there exists a function $f^{\mathcal{M}} \in \mathbf{A}.\alpha$ such that $\mathcal{M}$ takes (at least)

$$\epsilon^{-\frac{2+\alpha}{1+\alpha}+\tau}$$

iterations/function-evaluations to generate $\|g_k\| \leq \epsilon$, for any $\tau > 0$ arbitrarily small.

$\Longrightarrow$ $(2+\alpha)$-regularization method is optimal for the class $\mathbf{M}.\alpha$ when applied to functions in $\mathbf{A}.\alpha$, as its complexity upper bound coincides in order to the lower bound.

- Extension to examples with finite minimizers: possible.

Construction of the function $f^{\mathcal{M}}$

Assume $\{f_k\}$ given;

$$g_k = -\left(\frac{1}{k+1}\right)^t, \; k \geq 0, \text{ for some } t \in (0,1];$$

$$\overline{\kappa}_h |g_k|^{\frac{\alpha}{1+\alpha}} \geq H_k \geq -\kappa_h |g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0;$$

$$x_0 = 0, \quad x_{k+1} - x_k = s_k = -\frac{g_k}{H_k + \lambda_k},$$

with $H_k + \lambda_k \succ 0$ and $\lambda_k \geq 0$ satisfying M.α properties.

Construction of the function $f^{\mathcal{M}}$

Assume $\{f_k\}$ given;

$$g_k = -\left(\frac{1}{k+1}\right)^t, \; k \geq 0, \text{ for some } t \in (0,1];$$

$$\overline{\kappa}_h |g_k|^{\frac{\alpha}{1+\alpha}} \geq H_k \geq -\kappa_h |g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0;$$

$$x_0 = 0, \quad x_{k+1} - x_k = s_k = -\frac{g_k}{H_k + \lambda_k},$$

with $H_k + \lambda_k \succ 0$ and $\lambda_k \geq 0$ satisfying M.α properties.

Use Hermite interpolation and let

$$f^{\mathcal{M}}(x) = p_k(x - x_k) + f_{k+1}, \quad \text{for } x \in [x_k, x_{k+1}] \text{ and } k \geq 0,$$

where p_k is a 5th degree polynomial satisfying
$p_k(0) = f_k - f_{k+1}$ and $p_k(s_k) = 0$; $p_k'(0) = g_k$ and $p_k'(s_k) = g_{k+1}$;
$p_k''(0) = H_k$ and $p_k''(s_k) = H_{k+1}$. A lot more details....

ARC: improved complexity for structured problems

- f has bounded level sets and at local min x_*, $H(x_*) \succ 0$

$\Longrightarrow$ ARC Q-quadratic local rate when sufficiently close to x_*

$\Longrightarrow$ $\log|\log\epsilon|$ iteration complexity asymptotically, near x_*.

Further, can estimate δ such that $\|g_{k+1}\| \leq \delta\|g_k\|^2$

$\Longrightarrow$ $\mathcal{N}(x_*) \cap \{x : \|g(x)\| \leq \frac{1}{\delta}\} = \mathcal{N}$ quadratic convergence.

ARC: improved complexity for structured problems

- f has bounded level sets and at local min x_*, $H(x_*) \succ 0$

  $\Longrightarrow$ ARC Q-quadratic local rate when sufficiently close to x_*

  $\Longrightarrow$ $\log|\log\epsilon|$ iteration complexity asymptotically, near x_*.

  Further, can estimate δ such that $\|g_{k+1}\| \leq \delta\|g_k\|^2$

  $\Longrightarrow$ $\mathcal{N}(x_*) \cap \{x : \|g(x)\| \leq \frac{1}{\delta}\} = \mathcal{N}$ quadratic convergence.

- If there is no $x \notin \mathcal{N}$ s.t. $\epsilon \leq \|g(x)\| \leq 1/\delta$, then $\|g_k\| \leq \epsilon$ requires at most

$$\left\lceil \kappa_1 \cdot \delta^{3/2} + \kappa_2 \cdot \log|\log\epsilon| \right\rceil \quad \textbf{function evaluations}$$

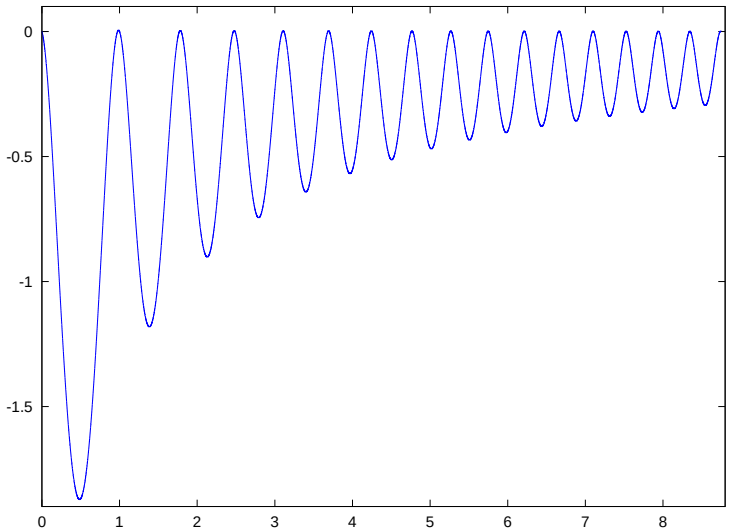
  Example: Rosenbrock function.

- improved efficiency bounds for nonconvex problems based on gradient's "phases" $\longrightarrow$ additive complexity bounds.

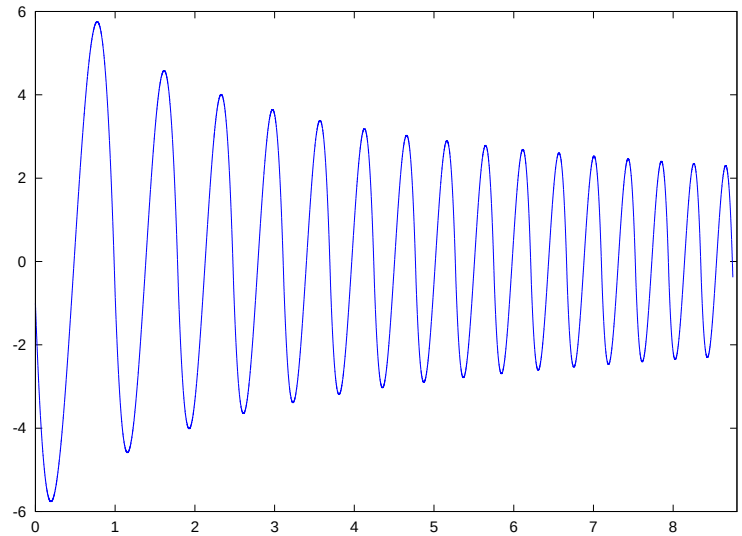
Second-order optimality complexity bounds

$\longrightarrow$ are also tight for ARC and trust-region methods.

- $\mathcal{O}(\epsilon^{-3})$ evaluations for ARC and trust-region to ensure both $\|g_k\| \leq \epsilon$ and $\lambda_{\min}(H_k) \geq -\epsilon$.
- this bound is tight for each method.



The gradient g.



The Hessian H.

First-order finite-difference and derivative-free ARC

- forward gradient-differences to construct approximate Hessian $B_k \longrightarrow \mathcal{O}\left(n(\epsilon^{-3/2} + |\log \epsilon|)\right)$ gradient- and $\mathcal{O}\left(\epsilon^{-3/2}\right)$ function-evaluations for $\|g_k\| \leq \epsilon$.

- central difference scheme for approximating g_k + function finite-differences for B_k with same stepsize (attention to termination criterion) $\longrightarrow \mathcal{O}\left(n^2(\epsilon^{-3/2} + |\log \epsilon|)\right)$ function-evaluations for $\|g_k\| \leq \epsilon$.

Evaluation complexity of constrained problems

Consider the general constrained nonlinear programming:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c_E(x) = 0; \; c_I(x) \geq 0,$$

where f and $c_{E,I} : \mathbb{R}^n \to \mathbb{R}^{m,p}$ are smooth and nonconvex.

Complexity of computing an (approximate) KKT point?

(Question not restricted to cubic regularization algorithms)

A detour: minimizing non-smooth composite functions

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + h(c(x)),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ smooth and nonconvex, and $h : \mathbb{R}^m \to \mathbb{R}$ nonsmooth but convex. $(h = \|\cdot\|)$

[considered by Nesterov (2006, 2007) and CGT (2011)]

Minimizing a non-smooth composite function

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad f(x) + h(c(x))$$

First-order method: compute a step s_k by solving the (convex) problem

$$\underset{\|s\| \leq \Delta_k}{\text{minimize}} \quad l(x_k, s) = f(x_k) + g(x_k)^T s + h(c(x_k) + J(x_k)s),$$

for some trust-region radius Δ_k (also possible using quadratic regularization).

Main result: Assume f, c, h are globally Lipschitz continuous. Then the "algorithm" takes at most $\mathcal{O}(\epsilon^{-2})$ problem-evaluations to achieve $\Psi(x_k) \leq \epsilon$, where $\Psi(x_k)$ is a first-order criticality measure

$$\Psi(x_k) = l(x_k, 0) - \min_{\|s\| \leq 1} l(x_k, s)$$

A first-order algorithm for EC-NLO

Consider now

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0.$$

Idea for a first-order algorithm:

- get feasible (if possible) by minimizing $\|c(x)\|$.
- track the trajectory

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

for decreasing values of t from some t_0 (corresponding to the first feasible iterate).

A first-order algorithm for EC-NLO ...

A Short-Step Steepest-Descent (SSSD) algorithm:

feasibility: apply nonsmooth composite minimization to

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \|c(x)\|.$$

$\Longrightarrow$ at most $\mathcal{O}(\epsilon^{-2})$ function evaluations.

tracking: successively

- apply one (successful) step of nonsmooth composite minimization to

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \Phi(x) = \|c(x)\| + |f(x) - t|.$$

- decrease t (proportionally to the decrease in $\Phi(x)$)

$\Longrightarrow$ at most $\mathcal{O}(\epsilon^{-2})$ problem evaluations.

A complexity result for EC-NLO

Assume that f, its gradient g, constraints c and Jacobian J are globally Lipschitz continuous; f bounded below and above in an $\epsilon-$neighbourhood of feasibility. Then the SSSD algorithm takes at most

$$\mathcal{O}\left(\epsilon^{-2}\right) \text{ problem evaluations}$$

to find an iterate x_k with either

$$\|c(x_k)\| \leq \epsilon \text{ and } \|J(x_k)^T y + g(x_k)\| \leq \epsilon,$$

for some y, or

$$\|c(x_k)\| > \kappa_f \epsilon \text{ and } \|J(x_k)^T z\| \leq \epsilon,$$

for some z and for some user-defined $\kappa_f > 0$.

- also applies to inequality-constrained problems: replace $\|c(x)\|$ by $\|\min(c(x), 0)\|$.

Evaluation complexity of constrained problems...

A (first-order) exact penalty method for EC-NLO:

To generate an approximate KKT point (within ϵ), requires at most:

- $\mathcal{O}(\epsilon^{-2})$ problem-evaluations when the penalty parameter is bounded
- $\mathcal{O}(\epsilon^{-4})$ problem-evaluations, otherwise.

$\longrightarrow$ Use first-order trust-region or quadratic-regularization for minimizing the composite nonsmooth penalty function

$$\Phi_\rho(x) = f(x) + \rho\|c(x)\|.$$

- also applies to inequality-constrained problems: replace $\|c(x)\|$ by $\|\min(c(x), 0)\|$.

Evaluation complexity of constrained problems...

Second-order methods for nonlinear (nonconvex) equality- and inequality-constrained smooth problems

$\longrightarrow$ ARC variants for problems with convex constraints and nonconvex objective considered: at most $\mathcal{O}(\epsilon^{-3/2})$ problem-evaluations required for approximate first-order.

$\longrightarrow$ general (nonconvex) constraints? (work in progress)

Remark:

Often, subproblem solution does not require additional problem-evaluations $\longrightarrow$ complexity bounds ignore the cost of solving the subproblem. Then, evaluation cost of constrained optimization $\equiv$ unconstrained optimization.

Conclusions and work in progress

Algorithm design profits from complexity analysis.

- Problem dimension-dependence of complexity bounds $\longrightarrow$ Jarre's example.
- Cubic regularization: the next generation of optimization software?
- Function-evaluation complexity of second-order methods for nonconvex constrained problems.

Some references:

- CGT, On the complexity of steepest-descent, Newton's and regularized Newton's method for unconstrained optimization. SIAM Journal on Optimization, 2010.
- CGT, Optimal Newton-type methods for nonconvex smooth optimization, 2011 (Optimization Online).
- CGT, Adaptive cubic regularization methods for unconstrained optimization. Part I and II. Math Programming 2010, 2011.