Adaptive Cubic Overestimation for Unconstrained Optimization

Coralia Cartis[1] Nick Gould[2] Philippe Toint[3]

[3]Department of Mathematics,University of Namur, Belgium

(philippe.toint@fundp.ac.be)

[1]University of Edimburgh, UK
[2]Computer Laboratory, University of Oxford, UK

Joint EUROPT-OMS Conference, Prag, July 2007

Outline

1. The new method
2. Convergence theory
3. Practical algorithmics
4. Conclusions

Outline

1. The new method
2. Convergence theory
3. Practical algorithmics
4. Conclusions

Outline

1. The new method
2. Convergence theory
3. Practical algorithmics
4. Conclusions

Outline

1. The new method
2. Convergence theory
3. Practical algorithmics
4. Conclusions

Unconstrained optimization — a "mature" area?

Unconstrained optimization:

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

Currently two main competing (but similar) methodologies

- **Linesearch methods**
  - compute a descent direction s_k from x_k
  - set $x_{k+1} = x_k + \alpha_k s_k$ to improve f
- **Trust-region methods**
  - compute a step s_k from x_k to improve a model m_k of f within the trust-region $\|s\| \leq \Delta$
  - set $x_{k+1} = x_k + s_k$ if m_k and f "agree" at $x_k + s_k$
  - otherwise set $x_{k+1} = x_k$ and reduce the radius Δ

Is there anything more to say?

Recently, Nesterov and Polyak (2006) observed the following: if

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

Taylor, Cauchy-Schwarz and Lipschitz imply

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

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

Nesterov and Polyak highlights

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

- N&P minimize m globally
  - 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 complexity than previously known

Obvious questions:

- can we avoid the global Lipschitz requirement?
- can we approximately minimize m and retain good worst-case 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 Cubic Overestimation (ACO)

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

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

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

c.f. trust-region methods

Convergence to first-order critical points

- $f(x_k) - m_k(s_k) \geq \frac{1}{6\sqrt{2}} \|g_k\| \min \left[\frac{\|g_k\|}{1 + \|B_k\|}, \frac{1}{2}\sqrt{\frac{\|g_k\|}{\sigma_k}} \right]$
- $\|s_k\| \leq \frac{3}{\sigma_k} \max(\|B_k\|, \sqrt{\sigma_k \|g_k\|})$
- if $\|g_k\| \geq \epsilon \quad \forall k \quad \Longrightarrow \quad \exists L \mid \sigma_k \leq \frac{L}{\epsilon} \quad \forall k$
- f bounded below and $g_\ell \neq 0 \quad \forall \ell \quad \Longrightarrow \quad \liminf_{k\to\infty} \|g_k\| = 0$
- f bounded below and $g_\ell \neq 0 \quad \forall \ell \quad \Longrightarrow \quad \boxed{\lim_{k\to\infty} \|g_k\| = 0}$

Under stronger assumptions can show that

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

if s_k minimizes m_k over subspace with orthogonal basis matrix Q_k

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 both 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

Iteration complexity

How many iterations are needed to ensure that $\|g_k\| \leq \epsilon$?

- so long as for very successful iterations $\sigma_{k+1} \leq \gamma_3 \sigma_k$ for $\gamma_3 < 1$
  $\Longrightarrow$ basic ACO algorithm requires at most
  $$\left\lceil \frac{\kappa_{\rm C}}{\epsilon^2} \right\rceil \text{ iterations}$$
  for some $\kappa_{\rm C}$ independent of ϵ c.f. steepest descent
- 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
  $\Longrightarrow$ ACO algorithm requires at most
  $$\left\lceil \frac{\kappa_{\rm S}}{\epsilon^{3/2}} \right\rceil \text{ iterations}$$
  for some $\kappa_{\rm S}$ independent of ϵ c.f. Nesterov & Polyak

Minimizing the model

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

Derivatives:

- $\lambda = \sigma \|s\|_2$
- $\nabla_s m(s) = g + Bs + \lambda s$
- $\nabla_{ss} m(s) = B + \lambda I + \lambda \left(\frac{s}{\|s\|} \right) \left(\frac{s}{\|s\|} \right)^T$

Optimality: any global minimizer s_* of m satisfies

$$(B + \lambda_* I) s_* = -g$$

- $\lambda_* = \sigma \|s_*\|_2$
- $B + \lambda_* I$ is positive semi-definite

The (adapted) secular equation

Require

$$(B + \lambda I)s = -g \quad \text{and} \quad \lambda = \sigma \|s\|_2$$

Define $s(\lambda)$:

$$(B + \lambda I)s(\lambda) = -g$$

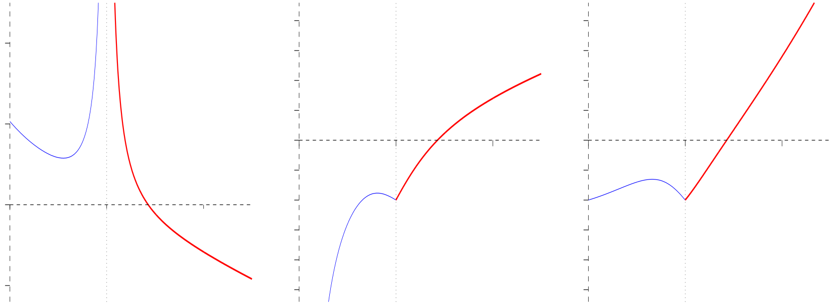
and find scalar λ as the root of secular equations

$$\|s(\lambda)\|_2 - \frac{\lambda}{\sigma} = 0 \quad \text{or} \quad \frac{1}{\|s(\lambda)\|_2} - \frac{\sigma}{\lambda} = 0 \quad \text{or} \quad \frac{\lambda}{\|s(\lambda)\|_2} - \sigma = 0$$

- values and derivatives of $s(\lambda)$ satisfy linear systems with symmetric positive definite $B + \lambda I$
- need to be able to factorize $B + \lambda I$

Plots of secular functions against λ

Example: $g = (0.25 \;\; 1)^T$, $H = \text{diag}(-1 \;\; 1)$ and $\sigma = 2$



$$\|s(\lambda)\|_2 - \frac{\lambda}{\sigma} = 0 \qquad \frac{1}{\|s(\lambda)\|_2} - \frac{\sigma}{\lambda} = 0 \qquad \frac{\lambda}{\|s(\lambda)\|_2} - \sigma = 0$$

Large problems — approximate solutions

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

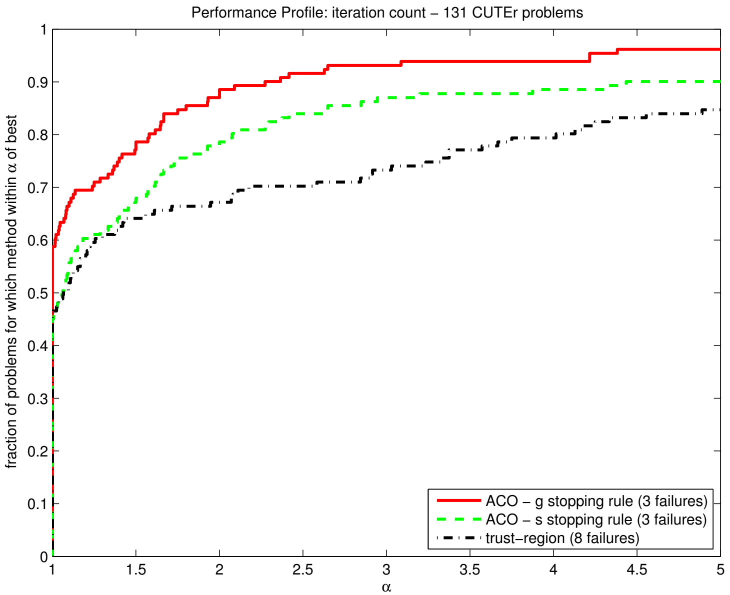
- $g \in \mathcal{S} \Longrightarrow$ ACO algorithm globally convergent
- Q orthogonal basis for $\mathcal{S} \quad \Longrightarrow \quad s = Qu$ where

$$u = \arg\min_{u \in \mathbb{R}^j} \; f + u^T(Q^T g) + \tfrac{1}{2} u^T (Q^T B Q) u + \tfrac{1}{3} \|u\|_2^3$$

$\Longrightarrow$ 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 secular equation even if j is large
- using g- or s-stopping rule $\Longrightarrow$ fast asymptotic convergence for ACO
- using s-stopping rule $\Longrightarrow$ good iteration complexity for ACO

Numerical experience — small problems using Matlab



Conclusions (1)

Encouraging so far!

- promising alternative to linesearch and trust-region methods for unconstrained optimization
- globally convergent to first- (and weak second-) order critical points
- fast asymptotic rate possible
- achieves best-known worst-case iteration complexity bound
- suitable for large-scale problems
- sophisticated implementation as part of GALAHAD underway

Conclusions (2)

- "obvious" extensions to simple bounds, augmented Lagrangians etc.
- other regularizations ($p > 3$ or $p > 2$) possible (any reason?)
- use of semi-norm in the presence of linear equality constraints
- not known if (e.g.) trust-region methods have as good worst-case complexity (work in progress)

Thanks for your attention