Nonlinear stepsize control, Trust-Region and Regularization Algorithms for Unconstrained Optimization

Philippe Toint
(input by Stefania Bellavia, Coralia Cartis, Nick Gould and Benedetta Morini)

Department of Mathematics,University of Namur, Belgium

(philippe.toint@fundp.ac.be)

Veszprem, December 2008

Outline

1. Regularization techniques
   - Cubic
   - Quadratic
2. Nonlinear stepsize control
3. Conclusions

Outline

1. Regularization techniques
   - Cubic
   - Quadratic
2. Nonlinear stepsize control
3. Conclusions

Outline

1. Regularization techniques
   - Cubic
   - Quadratic
2. Nonlinear stepsize control
3. Conclusions

The problem

We consider the unconstrained nonlinear programming problem:

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

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

Unconstrained optimization — a “mature” area?

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

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 $\|s_k\| \leq \Delta$
  - 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

A useful theoretical observation

Consider trust-region method where

$$\boxed{\text{model} = \text{true objective function}}$$

Then

- model and objective always agree
- trust-region radius goes to infinity

$$\Rightarrow \boxed{\text{a linesearch method}}$$

Nice consequence:

$$\boxed{\text{A unique convergence theory!}}$$

(Shultz/Schnabel/Byrd, 1985, T., 1988, Conn/Gould/T., 2000)

The keys to convergence theory for trust regions

The Cauchy condition:

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{\mathrm{TR}} \|g_k\| \min \left[\frac{\|g_k\|}{1 + \|H_k\|}, \Delta_k \right]$$

The bound on the stepsize:

$$\|s\| \leq \Delta$$

And we derive:

Global convergence to first/second-order critical points

Is there anything more to say?

Regularization Techniques

Is there anything more to say?

Observe the following: if

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

Taylor, Cauchy-Schwarz and Lipschitz imply

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

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

The cubic regularization

Change from

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

to

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

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
  - 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 $\boxed{m_k(s_k) \le m_k(s_k^C)}$
  - 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 $\gamma_2 \ge \gamma_1 > 1$, set

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

c.f. trust-region methods

Local convergence theory for cubic regularization (1)

The Cauchy condition:

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

The bound on the stepsize:

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

(Cartis/Gould/T)

Local convergence theory for cubic regularization (2)

And therefore. . .

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

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 $s(\lambda)$:

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

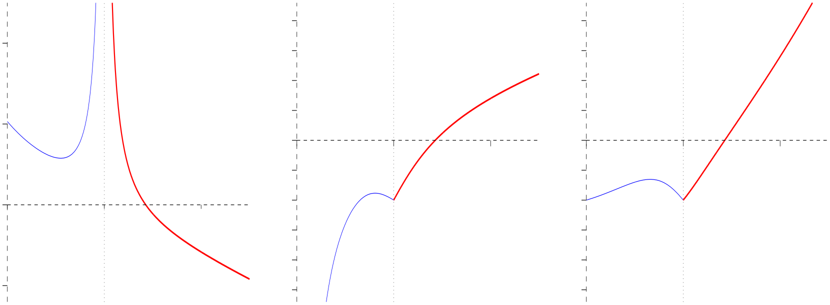
and find scalar λ as the root of secular equations

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

- 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

The main features of adaptive cubic regularization

And the result is. . .

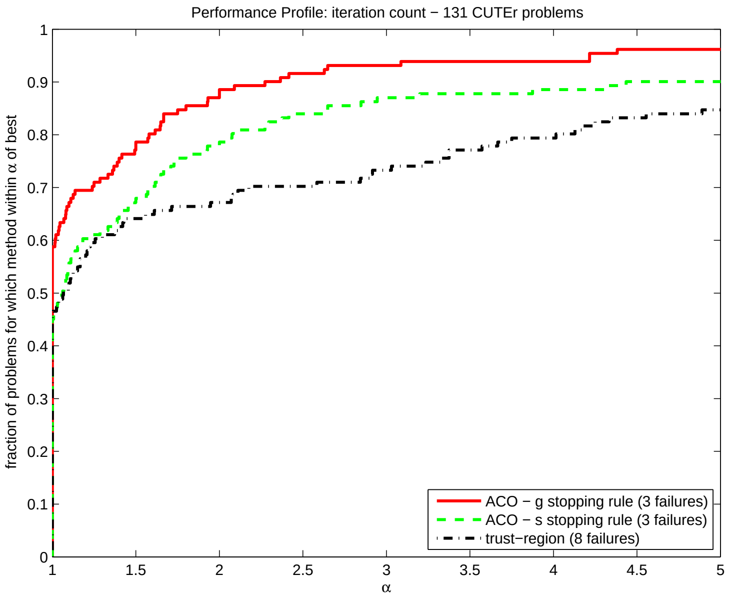
longer steps on ill-conditioned problems

similar (very satisfactory) convergence analysis

best known worst-case complexity for nonconvex problems

excellent performance and reliability

Numerical experience — small problems using Matlab



The quadratic regularization for NLS

Consider the Gauss-Newton method for nonlinear least-squares problems. Change from

$$\min_s \quad \tfrac{1}{2}\|c(x)\|^2 + \langle s, J(x)^T c(x)\rangle + \tfrac{1}{2}\langle s, J(x)^T J(x) s\rangle \;\; \text{s.t.} \;\; \|s\| \leq \Delta$$

to

$$\min_s \quad \|c(x) + J(x)s\| + \tfrac{1}{2}\sigma\|s\|^2$$

σ is the (adaptive) regularization parameter

(idea by Nesterov)

Quadratic regularization: reformulation

Note that

$$\min_{s} \quad \|c(x) + J(x)s\| + \tfrac{1}{2}\sigma\|s\|^2$$

$$\Leftrightarrow$$

$$\min_{\nu,s} \quad \nu + \tfrac{1}{2}\sigma\|s\|^2 \quad \text{such that} \quad \|c(x) + J(x)s\|^2 = \nu^2$$

exact penalty function for the problem of minimizing $\|s\|$ subject to $c(x) + J(x)s = 0$.

Iterative techniques. . . as for the cubic case (Cartis, Gould,T.):

$$\boxed{\text{solve the problem in nested Krylov subspaces}}$$

- Lanczos $\rightarrow$ factorization of tridiagonal matrices
- different scalar secular equation (solution by Newton's method)

The keys to convergence theory for quadratic regularization

The Cauchy condition:

$$m(x_k) - m(x_k + s_k) \geq \kappa_{\rm QR} \frac{\|J_k^T c_k\|}{\|c_k\|} \min \left[\frac{\|J_k^T c_k\|}{1 + \|J_k^T J_k\|}, \frac{\|J_k^T c_k\|}{\sigma_k \|c_k\|} \right]$$

The bound on the stepsize:

$$\|s_k\| \leq \frac{1}{2} \frac{\|J_k^T c_k\|}{\sigma_k \|c_k\|}$$

Convergence theory for the quadratic regularization

Convergence results:

Global convergence to first-order critical points

Quadratic convergence to roots

Valid for

- general values of m and n,
- exact/approximate subproblem solution

(Bellavia/Cartis/Gould/Morini/T.)

A unifying concept: Nonlinear stepsize control

Towards a unified global convergence theory

Objectives:

- recover a unified global convergence theory
- possibly open the door for new algorithms

Idea:

- cast all three methods into a generalized TR framework
- allow this TR to be updated nonlinearly

Towards a unified global convergence theory (2)

Given

- 3 continuous first-order criticality measures $\psi(x)$, $\phi(x)$, $\chi(x)$
- an adaptive stepsize parameter δ

define a $\boxed{\text{generalized radius } \Delta(\delta, \chi(x))}$ such that

- $\Delta(\cdot, \chi)$ is C^1, strictly increasing and concave,
- $\Delta(0, \chi) = 0$ for all χ,
- $\Delta(\delta, \cdot)$ is non-increasing
- $$\boxed{\delta \frac{\partial \Delta}{\partial \delta}(\delta, \chi) \leq \kappa_\Delta \Delta(\delta, \chi)}$$
- . . .

Towards a unified global convergence theory (3)

- the generalized Cauchy condition:

$$m(x_k) - m(x_k + s_k) \geq \kappa_{\rm N} \phi_k \min \left[\frac{\psi_k}{1 + \|H_k\|}, \Delta(\delta_k, \chi_k) \right]$$

- the generalized bound on the stepsize:

$$\|s_k\| \leq \Delta(\delta_k, \chi_k)$$

The nonlinear stepsize control algorithm

Algorithm 2.1: Nonlinear Stepsize Control Algorithm

Step 0: Initialization: $x_0 \in \mathbb{R}^n$, δ_0 given. Set $k = 0$.

Step 1: Step computation: Choose a model $m_k(x_k + s)$ and find a step s_k satisfying generalized Cauchy and $\|s_k\| \leq \Delta(\delta_k, \chi_k)$.

Step 2: Step acceptance: Compute $f(x_k + s_k)$ and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$$

Set $x_{k+1} = x_k + s_k$ if $\rho_k \geq \eta_1$; $x_{k+1} = x_k$ otherwise.

Step 3: Stepsize parameter update: Choose

$$\delta_{k+1} \in \begin{cases} [\gamma_1 \delta_k, \gamma_2 \delta_k] & \text{if} \quad \rho_k < \eta_1, \\ [\gamma_2 \delta_k, \delta_k] & \text{if} \quad \rho_k \in [\eta_1, \eta_2), \\ [\delta_k, +\infty] & \text{if} \quad \rho_k \geq \eta_2. \end{cases}$$

Set $k \leftarrow k + 1$ and go to Step 1.

Resulting convergence theory

Similar to trust-region convergence theory, but

more work to prove that $\Delta(\delta_k, \chi_k)$ remains bounded away from zero

(assumptions of $\Delta(\delta, \chi)$ crucial here)
and the result is . . .

$$\lim_{k \to +\infty} \min[\phi_k, \psi_k, \chi_k] = 0$$

Unified first-order convergence theory!

Covers all previous cases

trust regions:

$$\phi_k = \psi_k = \chi_k = \|g_k\|, \qquad \Delta(\delta, \chi) = \delta$$

cubic regularization:

$$\phi_k = \psi_k = \chi_k = \|g_k\|, \qquad \delta_k = \frac{1}{\sigma_k}, \quad \Delta(\delta, \chi) = \sqrt{\delta \chi}$$

quadratic regularization:

$$\phi_k = \chi_k = \frac{\|J_k^T F_k\|}{\|F_k\|}, \; \psi_k = \|J_k^T F_k\|, \quad \delta_k = \frac{1}{\sigma_k}, \quad \Delta(\delta, \chi) = \delta \chi$$

a method by Fan and Yuan:

$$\phi_k = \chi_k = \psi_k = \|g_k\|, \qquad \Delta(\delta, \chi) = \delta \chi$$

Conclusions

- Much left to do. . . but very interesting
- Could lead to very untypical methods
  Example:

$$\psi_k = \phi_k = \chi_k = \|g_k\|, \qquad \Delta(\delta, \chi) = \sqrt{\delta\chi}$$

- Meaningful numerical evaluation still needed
- Many issues regarding regularizations still unresolved

Thank you for your attention !

(see `http://perso.fundp.ac.be/~phtoint/publications.html` for references)