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)

IRCM Conference, Luxemburg, October 2008
Outline

1. Regularization techniques
   - Cubic
   - Quadratic

2. Nonlinear stepsize control

3. Conclusions
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:

\[
\text{minimize } f(x)
\]

for \( x \in \mathbb{R}^n \) and \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) smooth.

Important special case: the nonlinear least-squares problem

\[
\text{minimize } f(x) = \frac{1}{2} \|F(x)\|^2
\]

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

Work in progress...
Unconstrained optimization — a “mature” area?

\[
\text{minimize } f(x) \text{ where } f \in C^1 \text{ (maybe } C^2) \]

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 \( \Delta \)
A useful theoretical observation

Consider trust-region method where

\[ \text{model} = \text{true objective function} \]

Then

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

\[ \Rightarrow \text{a linesearch method} \]

Nice consequence:

A unique convergence theory!

The keys to convergence theory for trust regions

The Cauchy condition:

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

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 \textbf{globally Lipschitz continuous} Hessian $H$ with constant $2L$

Taylor, Cauchy-Schwarz and Lipschitz imply

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

$\Rightarrow$ reducing $m$ from $s = 0$ improves $f$ since $m(0) = f(x)$. 
The cubic regularization

Change from

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

to

\[
\min_s \ f(x) + \langle s, g(x) \rangle + \frac{1}{2} \langle s, H(x)s \rangle + \frac{1}{3} \sigma \|s\|^3
\]

\(\sigma\) is the (adaptive) regularization parameter

(ideas from Griewank, Weiser/Deuflhard/Erdmann, Nesterov/Polyak, Cartis/Gould/T)
The Cauchy condition:

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

The bound on the stepsize:

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

(Cartis/Gould/T)
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

Performance Profile: iteration count – 131 CUTEr problems

- ACO – g stopping rule (3 failures)
- ACO – s stopping rule (3 failures)
- trust–region (8 failures)
Consider the **Gauss-Newton** method for **nonlinear least-squares** problems. Change from

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

to

\[
\min_s \quad \|c(x) + J(x)s\| + \frac{1}{2} \sigma \|s\|^2
\]

\(\sigma\) is the (adaptive) **regularization parameter**

(idea by **Nesterov**)
Regularization techniques

Quadratic regularization: reformulation

Note that

\[
\min_s \| c(x) + J(x)s \| + \frac{1}{2} \sigma \| s \|^2
\]

\[\iff\]

\[
\min_{\nu, s} \nu + \frac{1}{2} \sigma \| s \|^2
\]

such that

\[
\| c(x) + J(x)s \|^2 = \nu^2
\]

exact penalty function for the problem of minimizing \(\|s\|\) subject to \(c(x) + J(x)s = 0\).
The keys to convergence theory for quadratic regularization

The Cauchy condition:

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

The bound on the stepsize:

\[ \|s\| \leq \frac{1}{2} \frac{\|J_k^T c_k\|}{\sigma_k \|c_k\|} \]
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.)
Computing regularization steps

Iterative techniques...

solve the problem in nested Krylov subspaces

- Lanczos → basis of the Krylov subspace
- → factorization of tridiagonal matrices
- different scalar secular equation (solution by Newton’s method)

Approach valid for

- trust-region (GLTR),
- cubic and quadratic regularizations

(details in CGT techreport)
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**
Given

- two continuous first-order criticality measures $\psi(x)$ and $\psi(x)\chi(x)$
- an adaptive stepsize parameter $\delta$

define a generalized radius $\Delta(\delta, \chi(x))$ such that

- $\Delta(\cdot, \chi)$ is $C^1$, strictly increasing and concave,
- $\Delta(0, \chi) = 0$ for all $\chi$,
- $\Delta(\delta, \cdot)$ is non-increasing

- $\psi(x)$ bounded above
- ...
the generalized Cauchy condition:

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

the generalized bound on the stepsize:

\[
\|s\| \leq \Delta(\delta_k, \chi_k)
\]
Algorithm 2.1: Nonlinear Stepsize Control Algorithm

Step 0: Initialization: \( x_0 \in \mathbb{R}^n, \delta_0 \) 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

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

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...

$$\liminf_{k \to +\infty} \psi_k = 0 \quad \text{or} \quad \lim_{k \to +\infty} \chi_k = 0$$

(both true limits if $\psi$ is non-increasing)

Unified first-order convergence theory!
Covers all previous cases

trust regions:

\[ \chi_k = \|g_k\|, \quad \psi_k = 1, \quad \Delta(\delta, \chi) = \delta \]

cubic regularization:

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

quadratic regularization:

\[ \chi_k = \frac{\|J_k^T F_k\|}{\|F_k\|}, \quad \psi_k = \|F_k\|, \quad \delta_k = \frac{1}{\sigma_k}, \quad \Delta(\delta, \chi) = \delta \chi \]
Conclusions

- Much left to do... but very interesting
- Could lead to very untypical methods
  Example:
  \[ \chi_k = \| g_k \|, \quad \Delta(\delta, \chi) = \sqrt{\delta \chi} \]
- 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)