Trust Region with a Cubic Model

Trond Steihaug
Department of Informatics
University of Bergen, Norway
and
Humboldt Universität zu Berlin

Workshop on Automatic Differentiation, Nice April 15-15, 2005

Outline

Higher order is commonly used on convergence and on derivatives in optimization. First order methods are gradient based and have Q-order 1 or Q-super-linear (for Quasi-Newton methods) rate of convergence. Second order methods are using the Hessian and have Q-order 2 rate of convergence. *Rate of convergence* (Q-order) and the degree of the derivatives will not match for 'difficult' problems.

- Regularization $\Rightarrow$ Trust-region Subproblem (TRS)
- Trust region Methods in Unconstrained Optimization $\rightarrow$ TRS
- AD can give higher order
- Higher Order TRS

Linear Least Squares (LLS)

Given $m \times n$ matrix A and $b \in \mathbb{R}^m$ where $m \geq n$. Compute $x \in \mathbb{R}^n$ so that

$$\min \frac{1}{2}\|Ax - b\|_2$$

Let $A = V\Sigma U^T$ be the singular value decomposition and let

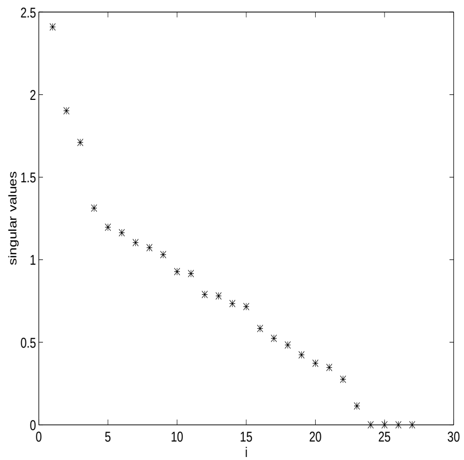
$$\Sigma^\dagger = \mathrm{diag}(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0), \quad r = \mathrm{rank}(A).$$

Define $A^\dagger = V\Sigma^\dagger U^T$. The solution x is

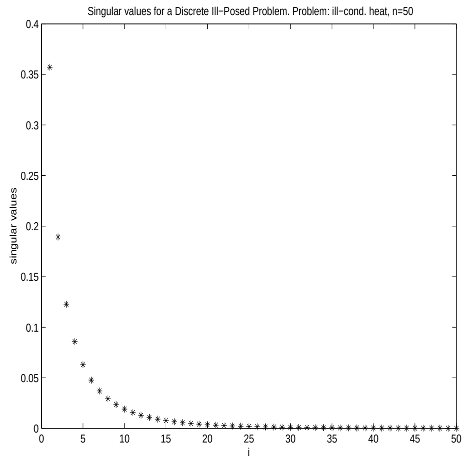
$$x = A^\dagger b = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

where $U = [u_1 \cdots u_n]$ and $V = [v_1 \cdots v_m]$.

Singular Values σ_i for Rank Deficient Problem



Singular Values σ_i for Discrete Ill-posed Problem



Discrete Picard Condition

A, b come from discretization from an ill-posed problem. All $\sigma_i > 0$ so formally

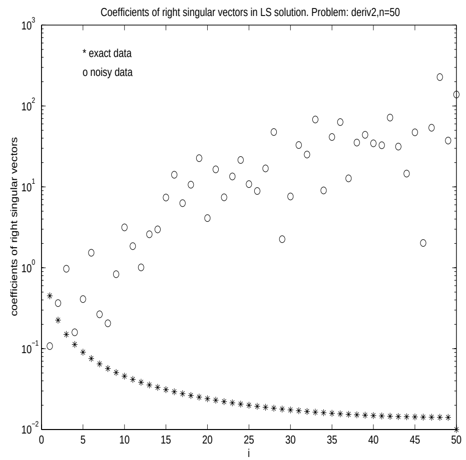
$$x = A^{\dagger} b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i$$

However

$$\frac{u_i^T b}{\sigma_i} \searrow 0 \text{ as } i \text{ increases } (\textbf{the discrete Picard condition}.)$$

Introduce noise in problem $b = \tilde{b} + \varepsilon$.

Coefficients $\frac{u_i^T b}{\sigma_i}$ for exact data and noisy data



One Solution to the Noisy Problem: Regularization

The following three problems are equivalent and make the 'noisy' problem smooth

Given $\mu \geq 0$ solve $\min \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_2^2$.

Given $\lambda \geq 0$ solve $(A^T A + \lambda I)x = A^T b$.

Given $\Delta \geq 0$ solve $\min_{\|x\| \leq \Delta} \frac{1}{2}\|Ax - b\|_2$. **TRS**

Equivalence from the Karush-Kuhn-Tucker conditions. (There exits open intervals for the three parameters μ, λ, Δ so that x is the solution to all three problems)

Where is AD?

Gauss - Newton and Nonlinear Least Squares

Given a nonlinear function $F : \mathbb{R}^n \to \mathbb{R}^m$.

Inexact Gauss-Newton Method:

Given x^0
while not converged do
  Compute $F'(x^i)$
  Find approximate solution s_i of $\min_{s\in\mathbb{R}^n} \frac{1}{2}\|F'(x^i)s + F(x^i)\|_2^2$
  Update $x^{i+1} = x^i + s^i$
end-while

$F'(x)$ is the $m \times n$ Jacobian matrix at x

Noise is inherit in the LLS problem!

unless high accuracy of F and F'

Higher Order Model Function

Gauss-Newton is based on 1.order approximation of F at x, i.e. $F(x+s) \approx M_1(s) = F(x) + F'(x)s$ and solve for the step s

$$\min_{s\in\mathbb{R}^n} \|M(s)\|_2^2.$$

Finding approximate solution s^i by constraining $\|s\| \le \Delta$ leads to Levenberg - Marquard methods. These are *trust-region methods* that use a linear model $M(s) = F'(x^i)s + F(x^i)$ at x^i of $F(x^i + s)$ with approximate solution

$$\min_{\|s\|\le\Delta} \quad \|M(s)\|_2^2.$$

Use more accurate model

$$M_2(s) = F(x^i) + F'(x^i)s + \frac{1}{2}(\mathcal{T}s)s, \;\; \mathcal{T} = F''(x^i)$$

Higher Order Model Function (2)

Let $m(s) \approx f(x+s) = F(x+s)^T F(x+s)$ and solve

$$\min_{\|s\|\le\Delta} m(s)$$

where

$$m_2(s) = f(x) + \nabla f(x)^T s + \frac{1}{2}s^T \nabla^2 f(x) s$$

$$m_3(s) = f(x) + \nabla f(x)^T s + \frac{1}{2}s^T \nabla^2 f(x) s + \frac{1}{6}s^T(\mathcal{T}s)s, \;\; \mathcal{T} = \nabla^3 f(x)$$

The Basic Trust Region Method

Given x^0 and Δ_0 $(0 \le \gamma_2 < \gamma_1 < 1,\ 0 \le \gamma_4 \le \gamma_5 < 1 \le \gamma_3)$
while not converged do
  Compute model $m^i(s)$.
  Compute approximate solution s^i of TRS:
    $\min_{\|s\|\le\Delta} \quad m^i(s)$.
  Compute $f(x^i + s^i)$, $m^i(s^i)$ and $\rho_i = \frac{f(x^i) - f(x^i + s^i)}{f(x^i) - m^i(s^i)} = \frac{\text{actual}}{\text{predicted}}$
  Update $x^{i+1} = \begin{cases} x^i + s^i \text{ if } \rho \ge \gamma_2 \\ x^i \text{ otherwise} \end{cases}$
  Update Δ_{i+1}: $\|s^i\| \le \Delta_{i+1} \le \gamma_3\|s^i\|$ if $\rho_i \ge \gamma_1$
    $\gamma_4\|s^i\| \le \Delta_{i+1} \le \gamma_5\|s^i\|$ if $\rho_i < \gamma_1$
end-while

Properties

$m_1(s) = f(x) + \nabla f(x)^T s$ - linear model
$m_2(s) = m_1(s) + \frac{1}{2} s^T \nabla^2 f(x) s$- quadratic model
$m_3(s) = m_2(s) + \frac{1}{6} s^T (\mathcal{T} s) s$ - cubic model.

Under 'reasonable' conditions the basic trust region algorithm be globally convergent, i.e. for given $\varepsilon > 0$ and any x^0 there exists an index i so that $\|\nabla f(x^i)\| \leq \varepsilon$. for the models.

Need to understand the Trust Region Subproblem (TRS)

$$\min_{\|s\| \leq \Delta} \quad m(s).$$

Exact Solution of TRS $m_i(s), i = 1, 2, 3$

The trust region subproblem with m_1

$$\min_{\|s\| \leq \Delta} \; f + g^T s$$

gives the Step Constrained Cauchy point $\tilde{s}$

$$\tilde{s} = -\frac{\Delta}{\|g\|} g$$

Exact Solution of TRS $m_2(s)$

$$\min_{\|s\| \leq \Delta} \; f + g^T s + \frac{1}{2} s^T H s$$

s is a solution with Lagrange multiplier δ *if and only if*

(i) $(H + \delta I)s + g = 0$

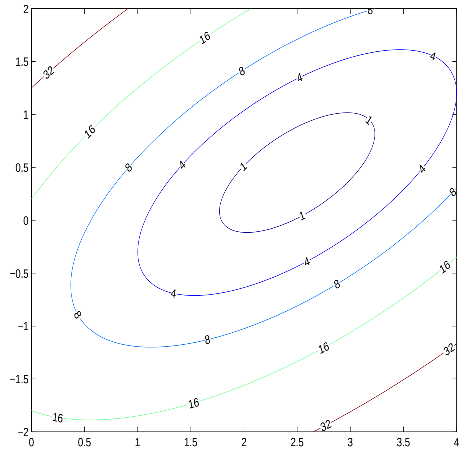
(ii) $H + \delta I$ is positive semi definite

(iii) $\delta \geq 0$ and $\delta(\|s\| - \Delta) = 0$.

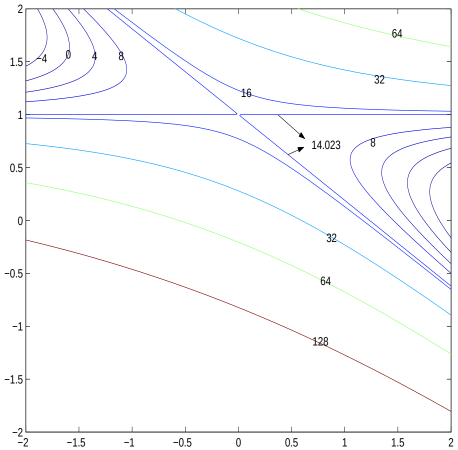
(Gay (1981) and Sorensen (1982))

The solution is on the form $s(\delta) = -(H + \delta I)^{-1} g$ provided $H + \delta I$ pos.def. and $s(\delta) = \Delta$ (i.e. small Δ gives large δ). For $H + \delta I$ positive semi-definite we have two cases: g is orthogonal to the null-space of $H + \delta I$ and we have the so called 'hard-case' and g not orthogonal in which case we have a smooth solution.

H positive definite



H semi definite



Exact Solution of TRS for LLS

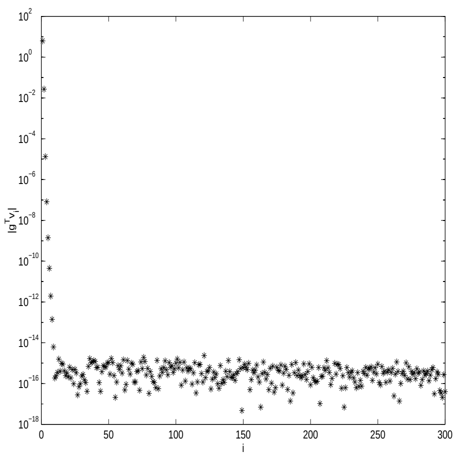
$$\min_{\|s\|\le\Delta} \; f + g^T s + \frac{1}{2} s^T H s$$

Let $\mathcal{S}_1 = \mathcal{N}(H + \lambda I)$ ($\mathcal{N}$ is the nullspace). We have the hard case when $g \perp \mathcal{S}_1$. For LLS recall that $H = A^T A$ and $g = -A^T b$ (so $\lambda = 0$ for the hard case)

$$g^T v_k = -\sigma_k b^T u_j, \; 1 \le j \le m_k$$

where u_j, v_j is associated with singular value σ_k with multiplicity m_k. (Rojas-Sorensen (2002))

The Hard Case is the Normal

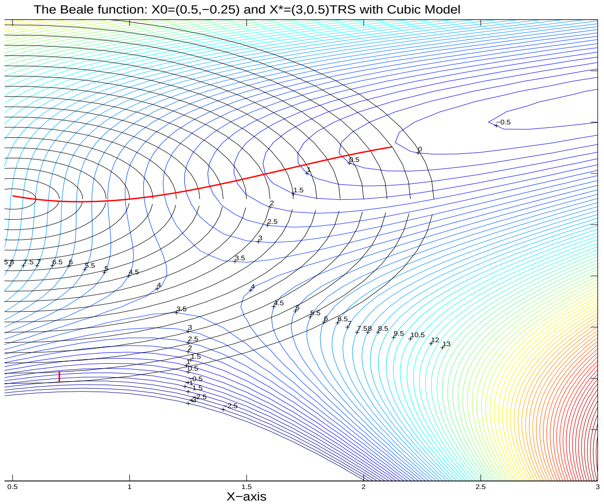


Note that $g^T v_j = 0$ is the (exact) hard case and $g^T v_j = -\sigma_j u_j^T b$.

A major Challenge: The Cubic Model

$$\min_{\|s\|\le\Delta} \quad m_3(s).$$

- We can characterize (*if and only if*) the (local) solution of TRS.
- We can compute the local minimizers. *In a way*
- What do we know about the (global) solution path? *In the general case it bifurcates, stops and is not continuous*
- The solution path we want consists of local **and** global solutions.

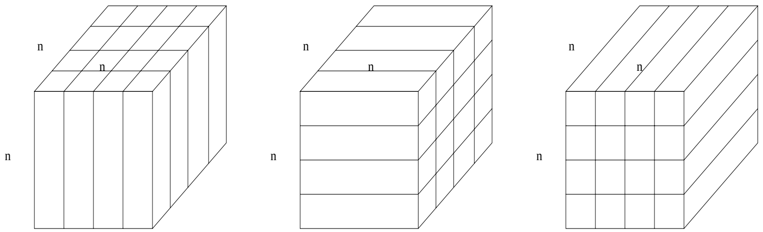


Global Convergence with the Cubic Model

$$\min_{\|s\|\leq\Delta} \quad m_3(s).$$

- Convergence results require $m_3(s^i) \leq \gamma_0 m_1(\tilde{s}^i)$. (Here $\tilde{s}^i$ is the step constrained Cauchy-point). *Not always the case for fixed* $\gamma_0 > 0$
- A problem arises in the proof of convergence when tensor is getting large. *Assume that the tensor is uniformly bounded*
- These results uses existence of s^i *No guaranteed working algorithm to compute* s^i.
- Can we say anything about the rate of convergence? *Except in the case when* f *is strictly convex at a (local) solution*

Vector Representation of Tensors



Columns - Rows - Tubes

Data structures for Super-symmetric tensors

$$\mathcal{T}_{ijk} = \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(x)$$

Clearly

$$\mathcal{T}_{ijk} = \mathcal{T}_{jik} = \mathcal{T}_{ikj} = \mathcal{T}_{jki} = \mathcal{T}_{kij} = \mathcal{T}_{kji}$$

To store the tensor we need to store $(n+2)(n+1)n/6$ (real) numbers

$$\mathcal{T}_{ijk} \quad 1 \leq k \leq j \leq i \leq n$$

Linear array:

$$\mathtt{T}((i-1)i(i+1)/6 + (j-1)*j/2 + k) \equiv \mathcal{T}_{ijk}, 1 \leq k \leq j \leq i \leq n$$

c# and java offer new possibilities to store the super-symmetric tensor and using standard notation $\mathtt{T}[i][j][k]$. Tube (i,j) is the array $\mathtt{T}[i][j]$

Sparse Tensors

Griewank-Toint (partial separability): The Hessian matrix is said to be sparse if

$$\nabla^2 f(x)_{ij} = 0 \text{ for all } x \in \mathbb{R}^n \quad (i,j) \in \mathcal{Z}.$$

Then *the sparsity structure of the tensor $\mathcal{T}$ is determined by the sparsity structure of the Hessian matrix.*

$$\mathcal{T}_{ijk} = 0 \text{ when } (i,j), (i,k) \text{ or } (j,k) \in \mathcal{Z}.$$

Symmetric skyline format is 'vector' based and can be extended to sparse super-symmetric tensors using array of arrays or a linear array with only n pointers as datastructure .

Tensor Methods are in Use

Around 50 papers in the database (in optimization and computational science). Around 50% of the papers 'Higher order methods have been considered by....'.

Brett W. Bader, PhD 2003, University of Colorado
Ali Bouaricha, PhD 1992, University of Colorado
Ta-Tung Chow, PhD 1989, University of Colorado
Paul D. Frank, PhD 1984, University of Colorado

Workshop on Tensor Decompositions and Applications August 2005 to discuss 'Large scale problems'.

Concluding Remarks

- AD has given us the opportunity to use higher derivatives. *Too messy for hand-coding*
- Very few classes of methods in optimization are capable to utilize 3rd derivative.
- Few efficient data structures for sparse super symmetric tensors.
- Are they right the researches that claim *tensor* methods can never compete with Newton's method in terms of speed of convergence when the Hessian matrix is nonsingular at the solution. $\Rightarrow$ Is there a big enough class of problems where 3rd derivative will be 'useful'.
- Ongoing work by *Geir Gundersen*, University of Bergen