CS Space

Unconstrained Nonlinear Optimization
Lecture 5: Trust-region methods 2

Feodor Pisnitchenko

CMCC/UFABC

Momoe Sakamori

IMECC/UNICAMP

Summary

- Iterative Solution of the Subproblem
- Local Convergence of Trust-Region Newton Methods
- Scaling

Trust-Region Theorem

Theorem 1

The vector p^* is a global solution of the trust-region problem

$$\min_{p\in\mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \text{ s.t. } \|p\| \le \Delta, \tag{1}$$

if and only if p^* is feasible and there exists a scalar $\lambda \ge 0$ such that:

(i) $(B + \lambda I)p^* = -g$,

(ii) $\lambda(\Delta - \|p^*\|) = 0$,

(iii) $B + \lambda I$ is positive semidefinite.

Iterative solution of the subproblem

- For small-scale problems (when n is not too large), more accurate solutions can be exploited.
- A closer approximation may justify a slightly higher computational cost.
- Cost: few factorizations of the matrix B (typically three factorization).
- This approach is based on the characterization of the exact solution from Trust-Region Theorem together with a clever application of Newton's method in one variable.
- Essentially, the algorithm tries to identify the value of λ for which $(B + \lambda I)p^* = -g$ is satisfied by the solution of the subproblem (1).

Algorithm for Finding p

- If $\lambda = 0$ satisfies (i) and (iii) with $\|p\| \leq \Delta$, use it.
- Otherwise, define $p(\lambda) = -(B + \lambda I)^{-1} g$ for λ where $B + \lambda I$ is positive definite.
- Solve $\|p(\lambda)\| = \Delta$, a one-dimensional root-finding problem in λ.

Eigendecomposition of B (Part 1)

- B (symmetric) has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and eigenvectors $q_1, q_2, \ldots, q_n$.
- Spectral decomposition:

$$B = \sum_{j=1}^{n} \lambda_j q_j q_j^T = Q \Lambda Q^T,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $Q = [q_1 | \ldots | q_n]$.

Eigendecomposition of B (Part 2)

- Then:

$$B + \lambda I = Q(\Lambda + \lambda I)Q^T.$$

- For $\lambda \neq -\lambda_j$:

$$p(\lambda) = -Q(\Lambda + \lambda I)^{-1}Q^T g = -\sum_{j=1}^{n} \frac{q_j^T g}{\lambda_j + \lambda} q_j.$$

- Norm:

$$\begin{aligned}
\|p(\lambda)\|^2 &= [-Q(\Lambda + \lambda I)^{-1}Q^T g]^T[-Q(\Lambda + \lambda I)^{-1}Q^T g] \\
&= g^T Q(\Lambda + \lambda I)^{-1}Q^T Q(\Lambda + \lambda I)^{-1}Q^T g \\
&= g^T Q[(\Lambda + \lambda I)^{-1}]^2 Q^T g \\
&= \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \lambda)^2}
\end{aligned}$$

Properties of $\|p(\lambda)\|$

- Consider the expression

$$\|p(\lambda)\| = \sqrt{\sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \lambda)^2}}.$$

- If $\lambda \geq -\lambda_1$, then $\lambda_j + \lambda > 0$ for all j, and so $\|p(\lambda)\|$ is continous, nonincreasing function of λ on the interval $(-\lambda_1, \infty)$. Thus, we have

$$\lim_{\lambda \to \infty} \|p(\lambda)\| = 0.$$

  Moreover, if $q_j^T g \neq 0$, then

$$\lim_{\lambda \to \lambda_j} \|p(\lambda)\| = \infty.$$

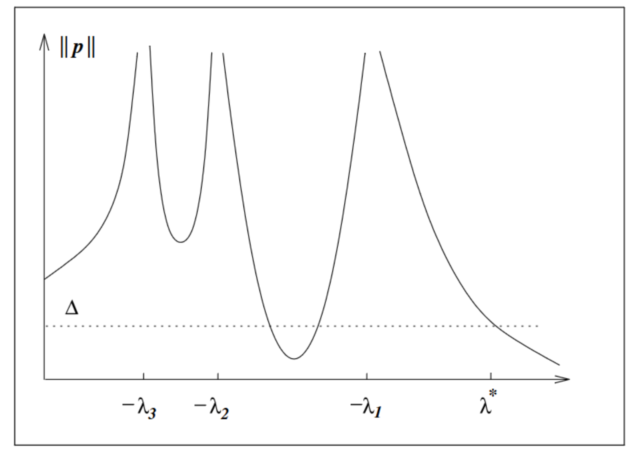

Figure: $\|p(\lambda)\|$ as a function of λ in a case in which $q_j^T g \neq 0$, for $j = 1, 2, 3$.

- Note that

$$\lim_{\lambda \to \infty} \|p(\lambda)\| = 0$$
$$\lim_{\lambda \to \lambda_j} \|p(\lambda)\| = \infty$$

  and $\|p(\lambda)\|$ is a nonincreasing function on λ on $(-\lambda_1, \infty)$.
- In particular, when $q_1^T g \neq 0$, there is a unique value $\lambda^* \in (-\lambda_1, \infty)$ such that $\|p(\lambda^*)\| = \Delta$.
- There may be other, smaller values of λ for which $\|p(\lambda)\| = \Delta$, but these do not satisfy the condition (iii) $(B + \lambda I)$ positive definite.

Procedure for identifying the λ^*

- Suppose $q_1^T g \neq 0$.
- When B is positive definite and $||B^{-1}g|| \leq \Delta$, the value $\lambda = 0$ satisfies the theorem's conditions (i-iii). Then the procedure terminates with $\lambda^* = 0$.
- Otherwise, use the root-finding Newton's method to find $\lambda > -\lambda_1$ that solves

$$\phi_1(\lambda) = ||p(\lambda)|| - \Delta = 0.$$

- The disadvantage os this approach is when λ is greater than, but close to, $-\lambda_1$. For this case, we can approximate ϕ_1 by a rational function, as follows:

$$\phi_1(\lambda) \approx \frac{C_1}{\lambda + \lambda_1} + C_2$$

  where $C_1 > 0$ and C_2 are constants.

- This approximation is highly nonlinear, so the root-finding Newton's method may not be effective.

- We can reformulate the problem

$$\phi_1(\lambda) = \|p(\lambda)\| - \Delta = 0.$$

as

$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|p(\lambda)\|}.$$

- It can be shown that for λ slightly greater than $-\lambda_1$, we have

$$\phi_2(\lambda) \approx \frac{1}{\Delta} - \frac{\lambda + \lambda_1}{C_3},$$

for some $C_3 > 0$.

- Since ϕ_2 is nearly linear near $-\lambda_1$, the root-finding Newton's method will perform well.
- This approach generates a sequence of iterates $\lambda^{(l)}$ by setting

$$\lambda^{(l+1)} = \lambda^{(l)} - \frac{\phi_2(\lambda^{(l)})}{\phi_2'(\lambda^{(l)})}$$

- After some manipulation, this updating formula can be implemented in the following practical way.

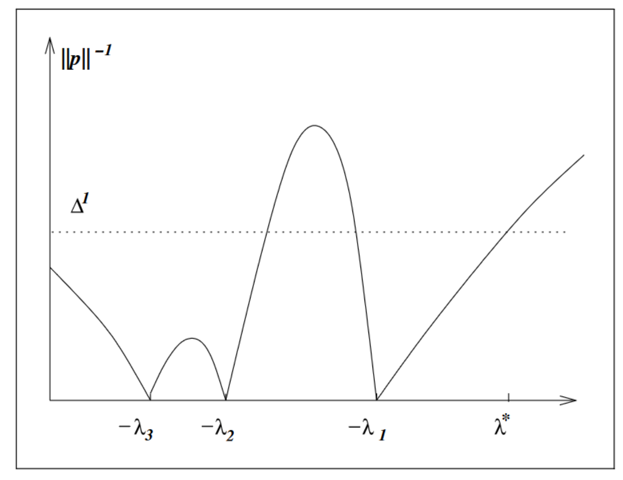

Figure: $1/\|p(\lambda)\|$ as a function of λ.

ALGORITHM 7: Trust-Region Subproblem (Part 1)

- Given $\lambda^{(0)}, \Delta > 0$:
- For $l = 0, 1, 2, \ldots$:
  - Factor $B + \lambda^{(l)} I = R^T R$.
  - Solve $R^T R p_l = -g$ and $R^T q_l = p_l$.
  - Update:

$$\lambda^{(l+1)} = \lambda^{(l)} - \left(\frac{\|p_l\|}{\|q_l\|}\right)^2 \left(\frac{\|p_l\| - \Delta}{\Delta}\right).$$

ALGORITHM 7: Trust-Region Subproblem (Part 2)

- Safeguards needed (e.g., Cholesky fails if $\lambda^{(l)} < -\lambda_1$).
- Main cost: Cholesky factorization per iteration.
- Practical use: Approximate solution with 2–3 iterations.

The Hard Case

- So far, we assumed that $q_1^T g \neq 0$.
- When $q_1^T g = 0$. we have what is referred to by Moré and Sorensen as the hard case.
- In this case, the condition $\lim_{\lambda \to \lambda_j} \|p(\lambda)\| = \infty$ does not hold for $\lambda_j = \lambda_1$ and then there may not be a value $\lambda \in (-\lambda_1, \infty)$ such that $\|p(\lambda)\| = \Delta$.
- Theorem 1 guarantees that the right value of λ lies in the interval $[-\lambda_1, \infty)$, so the only possibility is $\lambda = -\lambda_1$.
- After some mathematical manipulations, it is possible to verify that the conditions (i),(ii),(iii) of the Theorem 1 are satisfied.

The hard case

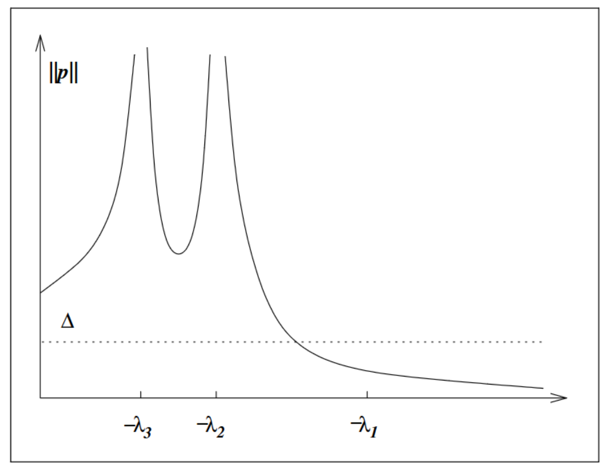

Figure: The hard case: $\|p(\lambda)\| < \Delta$ for all $\lambda \in (-\lambda_1, \infty)$.

Local Convergence of Trust-Region Methods

- Achieving the **fast convergence** rate typical of Newton's method relies on demonstrating that the **trust-region** constraint eventually **do not interfere** as we approach a solution.
- As we get **closer to the solution**, we want the **step** from the trust-region method to stay **inside the trust region** and get closer to the **Newton step**. When this happens, the steps are called **asymptotically similar** to Newton steps.

Theorem of Convergence Rate

Theorem

Let f be twice Lipschitz continuously differentiable in a neighborhood of a point x^* at which second-order sufficient conditions are satisfied. Suppose the sequence $\{x_k\}$ converges to x^* and that for all k sufficiently large, the trust-region algorithm based on $\min_{p\in\mathbb{R}^n} m_k(p) = f_k + g_k^T p + \frac{1}{2}p^T B_k p$, s.t. $\|p\| \le \Delta_k$, with $B_k = \nabla^2 f(x_k)$ chooses steps p_k that satisfy the Cauchy-point based model reduction criterion

$$m_k(0) - m_k(p_k) \ge c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \text{ for some } c_1 \in (0, 1]$$

and are asymptotically similar to Newton steps p_k^N whenever $\|p_k^N\| \le \frac{1}{2}\Delta_k$, that is,

$$\|p_k - p_k^N\| = o(\|p_k^N\|)$$

Then the trust-region bound Δ_k becomes inactive for all k sufficiently large and the sequence $\{x_k\}$ converges superlinearly to x^*.

Scaling

- Optimization problems often suffer from poor scaling, where the objective function is highly sensitive to changes in some variables and nearly insensitive to others. This creates elongated contour lines around the minimizer.
- Algorithms that do not account for this, like steepest descent, can perform poorly.
- A trust-region defines where the model m_k is a good approximation of the true function f.
- Spherical regions may not be effective if the function changes rapidly in some directions and slowly in others.
- Elliptical trust-regions, adapted to the scaling of f, provide more balanced confidence across all directions, leading to better performance.

Elliptical Trust Regions

- Ellliptical trust regions can be defined by

$$\|Dp\| \leq \Delta,$$

where D is a diagonal matrix with positive elements, yielding the scaled trust-region subproblem

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \text{ s.t. } \|Dp\| \leq \Delta_k.$$