Numerical Optimization
Lecture Notes #9 — Trust-Region Methods
Global Convergence and Enhancements

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Fall 2018
Recap & Introduction
- Recap: Iterative “Nearly Exact” Solution of the Subproblem
- Quick Lookahead

Global Convergence
- Tool #1 — A Lemma: The Cauchy Point
- Tool #2 — A Theorem
- Recall: The Trust Region Algorithm

Global Convergence...
- Convergence to Stationary Points

Enhancements
- Scaling
Last time we looked at **nearly exact solution** of the subproblem

\[
\min_{\bar{p} \in T_k} m_k(\bar{p}) = \min_{\bar{p} \in T_k} f(\bar{x}_k) + \bar{p}^T \nabla f(\bar{x}_k) + \frac{1}{2} \bar{p}^T B_k \bar{p}
\]

This approach is viable for problems with few degrees of freedom, e.g. \(T_k \subseteq \mathbb{R}^n\), \(n\) “small.” Where “small” means that the **unitary diagonalization** \(Q_k \Lambda_k Q_k^T = B_k\) is computable in a “reasonable” amount of time.

From a theoretical characterization of the exact problem, we derived an algorithm which finds a nearly exact solution at a cost per iteration approximately **three** times that of dogleg and 2D-subspace minimization.

The scheme was based on a 1-D Newton iteration (with some clever tricks), and some careful analysis of special (hard) cases.
On Today’s Menu

We wrap up the first pass of Trust Region methods —

– We briefly discuss global convergence properties for trust region methods.
– We look at some theorems, but leave the proofs as “exercises.”
– For second order \((B_k \neq \nabla^2 f(\bar{x}_k))\) models we can show convergence to a stationary point.
– For trust-region Newton methods \((B_k = \nabla^2 f(\bar{x}_k))\) models we can show convergence to a point where the second order necessary conditions hold.
– We look at modifications for poorly scaled problems, as well as the use of non-spherical trust regions.

Theorem (Second Order Necessary Conditions)

If \(\bar{x}^*\) is a local minimizer of \(f\) and \(\nabla^2 f\) is continuous in an open neighborhood of \(\bar{x}^*\), then \(\nabla f(\bar{x}^*) = 0\) and \(\nabla^2 f(\bar{x}^*)\) is positive semi-definite.
Recall: The trust-region subproblem is

$$\bar{p}_k = \arg \min_{\|\bar{p}\| \leq \Delta_k} m_k(\bar{p}) = \arg \min_{\|\bar{p}\| \leq \Delta_k} f(\bar{x}_k) + \bar{p}^T \nabla f(\bar{x}_k) + \frac{1}{2} \bar{p}^T B_k \bar{p}.$$  

The following lemma gives us a lower bound for the decrease in the model at the Cauchy point:

**Lemma (Cauchy point descent)**

The Cauchy point $\bar{p}^c_k$ satisfies

$$m_k(\bar{0}) - m_k(\bar{p}^c_k) \geq \frac{1}{2} \|\nabla f(\bar{x}_k)\| \min \left[ \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right].$$
Proof of Lemma

We recall the explicit expressions for the Cauchy point (from lecture 7)

\[ \bar{p}^c_k = -\tau_k \frac{\Delta_k}{\| \nabla f(\bar{x}_k) \|} \nabla f(\bar{x}_k) \]

where

\[ \tau_k = \begin{cases} 
1 & \text{if } \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k) \leq 0 \\
\min \left(1, \frac{\| \nabla f(\bar{x}_k) \|^3}{\Delta_k \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)} \right) & \text{otherwise}
\end{cases} \]

Figure: The three possible scenarios for selection of \( \tau \).
Case#1 \( (\nabla f(\bar{x}_k)B_k \nabla f(\bar{x}) \leq 0) \):

In this scenario \( m_k(\bar{p}_k^c) - m_k(\bar{0}) = \)

\[
m_k \left( -\Delta_k \frac{\nabla f(\bar{x}_k)}{\|\nabla f(\bar{x}_k)\|} \right) - m_k(\bar{0}) \\
= -\Delta_k \|\nabla f(\bar{x}_k)\| + \frac{1}{2} \frac{\Delta_k^2}{\|\nabla f(\bar{x}_k)\|^2} \left( \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k) \right) \leq 0 \\
\leq -\Delta_k \|\nabla f(\bar{x}_k)\| \\
\leq -\|\nabla f(\bar{x}_k)\| \min \left( \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right)
\]

Hence,

\[
m_k(\bar{0}) - m_k(\bar{p}_k^c) \geq \|\nabla f(\bar{x}_k)\| \min \left( \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right) \geq \frac{1}{2} \|\nabla f(\bar{x}_k)\| \min \left( \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right)
\]
Case#2 \((\nabla f(\bar{x}_k)B_k \nabla f(\bar{x}) > 0, \text{ and } \frac{\|\nabla f(\bar{x}_k)\|^3}{\Delta_k \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)} \leq 1)\):

In this scenario the Cauchy point is in the interior of the trust region, and

\[
m_k(\bar{p}^C_k) - m_k(\bar{0}) =
\]

\[
= -\frac{\|\nabla f(\bar{x}_k)\|^4}{\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)} + \frac{1}{2}\left(\frac{\|\nabla f(\bar{x}_k)\|^4}{\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)}\right)^2 \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)
\]

\[
= -\frac{1}{2}\frac{\|\nabla f(\bar{x}_k)\|^4}{\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)} \leq -\frac{1}{2}\frac{\|\nabla f(\bar{x}_k)\|^2}{\|B_k\| \|\nabla f(\bar{x}_k)\|^2} = -\frac{1}{2}\frac{\|\nabla f(\bar{x}_k)\|^2}{\|B_k\|}
\]

\[
\leq -\frac{1}{2}\|\nabla f(\bar{x}_k)\| \min \left(\Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|}\right)
\]

Use the minus sign to flip the inequality, and we’re there!
Proof of Lemma

Case#3 \((\nabla f(\bar{x}_k)B_k \nabla f(\bar{x}) > 0),\) and \(\frac{||\nabla f(\bar{x}_k)||^3}{\Delta_k \nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)} > 1\):

We note that in this scenario \(\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k) < \frac{||\nabla f(\bar{x}_k)||^3}{\Delta_k}\), and

\[
m_k(\bar{p}_k^c) - m_k(\bar{0}) =
\]

\[
= -\frac{\Delta_k}{||\nabla f(\bar{x}_k)||} \frac{\Delta_k^2}{2 \ ||\nabla f(\bar{x}_k)||^2} \Delta_k^2 \frac{\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)}{\Delta_k} \\
\leq -\Delta_k \frac{\nabla f(\bar{x}_k)}{||\nabla f(\bar{x}_k)||} + \frac{1}{2} \frac{\Delta_k^2}{||\nabla f(\bar{x}_k)||^2} \frac{\nabla f(\bar{x}_k)^T B_k \nabla f(\bar{x}_k)}{\Delta_k} \\
= -\frac{1}{2} \Delta_k ||\nabla f(\bar{x}_k)|| \\
\leq -\frac{1}{2} ||\nabla f(\bar{x}_k)|| \min \left( \Delta_k, \frac{||\nabla f(\bar{x}_k)||}{||B_k||} \right)
\]

Use the minus sign to flip the inequality, and we’re there!
Global Convergence: Tool #2 — A Theorem

Theorem

Let $\bar{p}_k$ be any vector, $\|\bar{p}_k\| \leq \Delta_k$, such that

$$m_k(\bar{0}) - m_k(\bar{p}_k) \geq c_2(m_k(\bar{0}) - m_k(\bar{p}_k^c))$$

then

$$m_k(\bar{0}) - m_k(\bar{p}_k) \geq \frac{c_2}{2} \|\nabla f(\bar{x}_k)\| \min \left[ \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right].$$

Both the dogleg, and 2-D subspace minimization algorithms (as well as Steihaug’s algorithm) fall into this category, with $c_2 = 1$, since they all produce $\bar{p}_k$ which give at least as much descent as the Cauchy point, i.e. $m_k(\bar{p}_k) \leq m_k(\bar{p}_k^c)$.

We are going to use this result to show convergence for the trust region algorithm (see next slide).
The Trust Region Algorithm

Algorithm: Trust Region

[1] Set $k = 1$, $\Delta > 0$, $\Delta_0 \in (0, \Delta)$, and $\eta \in [0, \frac{1}{4}]$
[2] While optimality condition not satisfied
[3] Get $\bar{p}_k$ (approximate solution)
[4] Evaluate $\rho_k$
[5] if $\rho_k < \frac{1}{4}$
[6] $\Delta_{k+1} = \frac{1}{4} \Delta_k$
[7] else
[8] if $\rho_k > \frac{3}{4}$ and $\|\bar{p}_k\| = \Delta_k$
[9] $\Delta_{k+1} = \min(2\Delta_k, \Delta)$
[10] else
[11] $\Delta_{k+1} = \Delta_k$
[12] endif
[13] endif
[14] if $\rho_k > \eta$
[15] $\bar{x}_{k+1} = \bar{x}_k + \bar{p}_k$
[16] else
[17] $\bar{x}_{k+1} = \bar{x}_k$
[18] endif
[19] $k = k + 1$
[20] End-While
Convergence to Stationary Points

**Case \( \eta = 0 \)**

accept any step which produces descent in \( f \) — we can show that the sequence of gradients \( \{ \nabla f(\bar{x}_k) \} \) has a **limit point** at zero.

**Case \( \eta > 0 \)**

accept a step only if the decrease in \( f \) is at least some fixed fraction of the predicted decrease — we can show the stronger result \( \{ \nabla f(\bar{x}_k) \} \to \bar{0} \).

In order for the proof(s) to work, we must assume that the model Hessians \( B_k \) are uniformly bounded, *i.e.* \( \| B_k \| \leq \beta \), and that \( f \) is bounded below on the levelset \( \{ \bar{x} \in \mathbb{R}^n : f(\bar{x}) \leq f(\bar{x}_0) \} \).

The trust-region bound can be relaxed so that the results hold as long as the solution to the subproblems satisfy

\[
\| \bar{p}_k \| \leq \gamma \Delta_k, \quad \text{for some constant } \gamma \geq 1.
\]
Convergence to Stationary Points: $\eta = 0$

**Theorem**

Let $\eta = 0$ in the trust region algorithm. Suppose that $\|B_k\| \leq \beta$ for some constant $\beta$, that $f$ is continuously differentiable and bounded below on the bounded set $\{\bar{x} \in \mathbb{R}^n : f(\bar{x}) \leq f(\bar{x}_0)\}$, and that all approximate solutions to the trust-region subproblem satisfy the inequalities

$$m_k(\bar{0}) - m_k(\bar{p}_k) \geq c_1 \|\nabla f(\bar{x}_k)\| \min\left[\Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|}\right],$$

and

$$\|\bar{p}_k\| \leq \gamma \Delta_k,$$

for some positive constants $c_1$ and $\gamma$. Then we have

$$\liminf_{k \to \infty} \|\nabla f(\bar{x}_k)\| = 0.$$
Theorem

Let $\eta \in (0, \frac{1}{4})$ in the trust region algorithm. Suppose that $\|B_k\| \leq \beta$ for some constant $\beta$, that $f$ is Lipschitz continuously differentiable and bounded below on the bounded set $\{\bar{x} \in \mathbb{R}^n : f(\bar{x}) \leq f(\bar{x}_0)\}$, and that all approximate solutions to the trust-region subproblem satisfy the inequalities

$$m_k(\bar{0}) - m_k(\bar{p}_k) \geq c_1 \|\nabla f(\bar{x}_k)\| \min \left[ \Delta_k, \frac{\|\nabla f(\bar{x}_k)\|}{\|B_k\|} \right].$$

and

$$\|\bar{p}_k\| \leq \gamma \Delta_k$$

for some positive constants $c_1$ and $\gamma$. Then we have

$$\lim_{k \to \infty} \nabla f(\bar{x}_k) = \bar{0}.$$
The complete proofs are in NW$^{1\text{st}}$ pp.90–91, and pp.92–93; or NW$^{2\text{nd}}$ pp.80–82, and pp.82–83.

The proofs are based on manipulation of $\rho$ — the ratio of actual (objective) reduction and predicted (model) reduction; Taylor’s theorem; then deriving a contradiction from the supposition $\|\nabla f(\bar{x}_k)\| \geq \epsilon$ using careful selection of scalings and bounds for $\Delta_k$.

**Definition (lim sup and lim inf)**

Let $\{s_n\}$ be a sequence of real numbers. Let $E$ be the set of values $x$ so that $s_{n_k} \to x$ for some subsequence $\{s_{n_k}\}$. This set $E$ contains all sub-sequential limits, plus possibly $\pm \infty$; let

$$s^* = \sup E, \quad s_* = \inf E$$

The values $s^*$ and $s_*$ are the upper and lower limits of $\{s_n\}$, and we use the notation

$$\limsup_{n \to \infty} s_n = s^*, \quad \liminf_{n \to \infty} s_n = s_*$$
Theorem \((\text{NW}^2\text{nd} \ p.92, \text{proof in Moré & Sorensen (1983)})\)

Let \(\eta \in (0, \frac{1}{4})\) in the algorithm on slide 11, let \(B_k = \nabla^2 f(\bar{x}_k)\), and suppose that \(\bar{p}_k\) at each iteration satisfy

\[
m_k(\bar{0}) - m_k(\bar{p}_k) \geq c_1 (m_k(\bar{0}) - m_k(\bar{p}_k^*)),
\]

and \(\|\bar{p}_k\| \leq \gamma \Delta_k\), for some positive constant \(\gamma\), and \(c_1 \in (0, 1]\). Then

\[
\lim_{k \to \infty} \|\nabla f(\bar{x}_k)\| = 0.
\]

If, in addition, the set \(\{\bar{x} \in \mathbb{R}^n : f(\bar{x}) \leq f(\bar{x}_0)\}\) is compact, then either the algorithm terminates at a point \(\bar{x}_k\) at which the second order necessary conditions for a local minimum hold, or \(\{\bar{x}_k\}\) has a limit point \(\bar{x}^* \in \{\bar{x} \in \mathbb{R}^n : f(\bar{x}) \leq f(\bar{x}_0)\}\) at which the conditions hold.
As we have seen before (in the context of steepest descent / line-search), **scaling** (ill-conditioning) can cause problems. — If the objective is more sensitive to changes in one variable than other, the contour lines stretch out to be narrow ellipses (in 2D).

Clearly, a circular trust-region may be quite limiting in this scenario. — The radius is limited by the sensitive variable.
Enhancement: Scaling — The Solution

The solution to the problem of poor scaling is to use **elliptical** trust regions. We define a diagonal scaling matrix

$$D = \text{diag}(d_1, d_2, \ldots, d_n), \quad d_i > 0.$$  

Then, the constraint $\|D\bar{p}\| \leq \Delta$ defines an elliptical trust region, and we get the following scaled trust-region subproblem:

$$\min_{\bar{p} \in \mathbb{R}^n : \|D\bar{p}\| \leq \Delta_k} f(\bar{x}_k) + \bar{p}^T \nabla f(\bar{x}_k) + \frac{1}{2} \bar{p}^T B_k \bar{p}.$$  

The scaling matrix can be built using information about the gradient $\nabla f(\bar{x}_k)$ and the Hessian $\nabla^2 f(\bar{x}_k)$ along the solution path. — We can allow $D = D_k$ to change from iteration to iteration.

All our analysis/algorithms still work with scaling added — but we get factors of $D^{-2}$, $D^{-1}$, $D$, and $D^2$ in our expressions.

Peter Blomgren, (blomgren.peter@gmail.com)
Feature: Non-Euclidean Trust Regions

Figure: Illustration of (unscaled) trust region boundaries for, from left-to-right: $\|\mathbf{p}\|_2 \leq \Delta_k$, $\|\mathbf{p}\|_1 \leq \Delta_k$, $\|\mathbf{p}\|_4 \leq \Delta_k$, and $\|\mathbf{p}\|_\infty \leq \Delta_k$.

Most of the time using trust regions based on norms with $q \neq 2$:

$$\|\mathbf{p}\|_q \leq \Delta_k \text{ (unscaled)}, \quad \|D\mathbf{p}\|_q \leq \Delta_k \text{ (scaled)}$$

cause us a giant head-ache. There are however some situations when such regions come in handy...
Feature: Non-Euclidean Trust Regions

Figure: Illustration of (unscaled) trust region boundaries for, from left-to-right: $\|\bar{p}\|_1 \leq \Delta_k$, $\|\bar{p}\|_{\frac{1}{2}} \leq \Delta_k$, $\|\bar{p}\|_{\frac{1}{4}} \leq \Delta_k$, and $\|\bar{p}\|_{\frac{1}{8}} \leq \Delta_k$.

Using $q < 1$ leads to non-convex trust regions, which may be a bit of a pain?!?

This may, however, be useful/necessary for non-convex optimization problems.
For **constrained** problems, e.g.

\[
\min_{\bar{x} \in \mathbb{R}^n} f(\bar{x}), \quad \text{subject to} \quad x_i \geq 0, \; i = 1, 2, \ldots, n
\]

the trust-region subproblem may be

\[
\min_{\bar{p} \in \mathbb{R}^n} m_k(\bar{p}), \quad \text{subject to} \quad \bar{x}_k + \bar{p} \geq 0, \text{ (component-wise), } \|\bar{p}\| \leq \Delta_k
\]

This trust region is the intersection of the disk centered at \(\bar{x}_k\) and the first quadrant. It could look like this:
Such a region is hard to describe, and hard to work with.

If, instead, we work with the $\| \cdot \|_\infty$-norm, the trust region is the intersection of the square with sides $\Delta_k$ centered at $\bar{x}_k$ and the first quadrant:

Much easier to work with...
Index

definition
  lim sup and lim inf, 15
lemma
  Cauchy point descent, 5
theorem
  Convergence (when $\eta = 0$), 13
  Convergence (when $\eta > 0$), 14
  Global trust-region Newton convergence ($\eta > 0$), 16
  Second order necessary conditions, 4

Reference(s):