Trust region

- Choice of radius of a trust region.
- Cauchy point.
- Dogleg, two-dimensional subspace.
- Exact minimization for quadratic function.
- Sufficient reduction. Convergence.

**Review**

The line-search methods do not require the optimal choice of step-length for convergence.

*Question* What is the intuition behind the definitions of convergence? Geometric series.

**Corollary.** Linear convergence requires a constraint on $M$:

$$||x_{k+1} - x^*|| \leq M||x_k - x^*||, M < 1$$

**Corollary.** For steepest descent

$$||x_{k+1} - x^*||_Q \leq M||x_k - x^*||_Q$$

where

$$M \approx \sqrt{1 - 1/\kappa^2(Q)}.$$  

**Line-search:** decide on an algorithm, focus on choice of length step.

**Trust region:** replace the objective function by an approximation in a ball.

Trust region is “good”, if

a) not too big, objective function is well approximated by the replacement,

b) not too small, objective function is well approximated by the replacement in a larger ball.

Define the ratio of actual reduction, the decrease of the objective function, to the predicted reduction, the decrease of the approximating function:

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$  

Trust region radius $\Delta_k$ is determined “inductively” successful previous step - larger size of the next ball, unsuccessful - reduce current.

**Algorithm**

if $\rho_k > 3/4$, then $\Delta_{k+1} = \min \Delta_0, 2\Delta_k$, elseif $\rho_k > 1/4\Delta_{k+1} = \Delta_k$ else $\Delta_k = ||p_k||/4$.

where $\Delta_0$ is a maximal size of a trust region.

Sufficient reduction of the objective function, a measure of performance of the choice of $\Delta_k$ is comparison to reduction by Cauchy point.

Cauchy point = steepest descent + step-length: minimum of the trust region size or step-length from proof of convergence of steepest descent.
Since Cauchy point \( \approx \) steepest descent, which has only linear rate of convergence, we want to improve on Cauchy point search:

\[
\min_p m(p), \quad m(p) = f + \langle g, p \rangle + \frac{1}{2} \langle p, Bp \rangle, \quad ||p|| \leq \Delta
\]

The unconstrained by the ball radius size minimizer direction is

\[
p^B = -B^{-1}g, \quad \text{Newton.}
\]

If \( \Delta \) is tiny

\[
p \approx -\frac{\Delta}{||g||} g, \quad \text{steepest descent} \quad p^U = -\frac{||g||^2}{||g||^2_B} g.
\]

**Dogleg method** is a combination of steepest descent and Newton:

\[
p(\tau) = \begin{cases} 
\tau p^U, & 0 \leq \tau \leq 1, \\
p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2 
\end{cases}
\]

**Lemma** Dogleg intersects the trust-region boundary at most once, because (if \( B \) is positive-definite)

(i) \( ||p(\tau)|| \) is increasing,

(ii) \( m(p(\tau)) \) is decreasing.

**Proof:** (i) follows if angle between \( p^U \) and \( p^B - p^U \) is not obtuse:

\[
\langle p^U, p^B - p^U \rangle \geq 0.
\]

which is true by direct computations:

\[
\langle p^U, p^B - p^U \rangle = ||g||^2 ||g||^2_B^{-1} \left[ 1 - \frac{||g||^4}{||g||^2_B^{-1} ||g||^2_B} \right].
\]

(ii) follows if we observe that any convex combination of the Newton and steepest descent

\[
\alpha p^U + (1 - \alpha) p^B, \quad 0 \leq \alpha \leq 1
\]

is a descent direction.

**Two-dimensional subspace minimization** is a generalization of dogleg to the space spanned by \( p^U \) and \( p^B \).

Use it when \( B \) is not positive-definite, by replacing the Newton direction for \( B \) by the Newton direction for \( B' = B + \alpha I \) where \( \alpha > 0 \) is chosen to make \( B' \) positive-definite.

**Steihaug’s approach** is related to Conjugate Gradient which we discuss in the next lecture.

**Question** What guarantees that these methods have good comparison with the Cauchy point iterations?

**Another point of view** Use of Newton and steepest descent is basically look for \( p \) that satisfies

\[
(B + \lambda I)p = -g, \quad \text{for some} \quad \lambda \geq 0(\ast)
\]

**Lemma** Even for an indefinite matrix \( B \), if \( p \) is a global solution for a trust-region problem if and only if (\( \ast \)) holds, \( B + \lambda I \) is positive semi-definite, and \( \lambda(\Delta - ||p||) = 0 \).
Proof Denote the eigenvalue $B$ has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ with corresponding eigenvectors $e_1, e_2, \ldots, e_n$.

Assume for simplicity that $\langle g, e_1 \rangle \neq 0$. The condition $\lambda(\Delta - ||p||) = 0$ is simply a short-write for $p$ being at the boundary of the trust region or being at the minimum of a quadratic form.

The only issue is why there is $\lambda \geq 0$ such that $(B + \lambda I)p = -g$ if $p$ lies on the boundary of the trust region (Why?).

Observe that when $\lambda \to \infty$, the solution of

$$\min_p \hat{m}(p, \lambda) = f + \langle g, p \rangle + \frac{1}{2} \langle p, (B + \lambda I)p \rangle = m(p) + \frac{\lambda}{2} ||p||^2, \quad ||p|| \leq \Delta$$

must be strictly inside the trust region. By decreasing $\lambda$ we can achieve $p$ to be exactly on the boundary of the trust region. Note that when $p$ is strictly inside the trust region, $B + \lambda I$, must be positive-definite, because

$$||p(\lambda)||^2 = \sum_j \frac{|\langle e_j, g \rangle|^2}{(\lambda_j + \lambda)^2} \to \infty$$

as $\lambda \to -\lambda_1$, in all the cases except when $g$ is perpendicular to $e_1$. The claim is that then the minimizers of $m$ and $\hat{m}$ are the same (see p.85 in book). Indeed suppose $p$ is a minimizer of $\hat{m}(p, \lambda)$ where $\lambda$ is chosen so that $||p|| = \Delta$. Then

$$\hat{m}(x, \lambda) > \hat{m}(p, \lambda)$$

for any $x$, in particular for all $x$ in the trust region.

Since $\hat{m}(x, \lambda) = m(x) + \lambda/2||x||^2$

$$m(x) - m(p) = \left(\hat{m}(x, \lambda) - \hat{m}(p, \lambda)\right) + \lambda/2(||p||^2 - ||x||^2) > 0.$$ 

This idea is not only a trick but also an example of an important reduction of problems of constraint optimization to problems of unconstrained optimization with Lagrange multipliers.

Observation If $x$ and $p$ are both on the boundary of the trust region

$$m(x) - m(p) = \left(\hat{m}(x, \lambda) - \hat{m}(p, \lambda)\right)$$

Example of the case when $\langle g, e_1 \rangle$. Consider

$$m(x) = -x_1 - \frac{1}{2}x_1^2 - x_2^2.$$ 

Reduction A sufficient reduction estimates typically are

$$m_k(p_k) \leq m_k(0) - c_1 ||\nabla f(x_k)|| \min \left(\Delta_k, \frac{||\nabla f(x_k)||}{||B_k||}\right)(**).$$

Lemma Cauchy point $p_k^C$ satisfies (**) with $c_1 = 1/2$. See proof book pp.81-82.

Other sufficient reduction estimates typically are compared to Cauchy point reduction:

$$m_k(0) - m_k(p_k) \geq c_2 (m_k(0) - m_k(p_k^C)).$$
Question what is $c_2$ for dogleg and two-dimensional subspace algorithms?

Answer $c_2 = 1$.

Theorem Our trust-region algorithm gives $\nabla f(x_k) \to 0$ for dogleg and 2d reduction under a simplifying assumption that for a given objective function we can always find $\Delta_1$ so that shrinkage of the trust-region never occurs beyond $\Delta_1$.

Proof By sufficient reduction condition combined with Cauchy point estimates gives (How? the simplifying assumption was used here twice)

$$f(x_k) - f(x_{k+1}) \geq C \min(||\nabla f(x_k)||, ||\nabla f(x_k)||^2),$$

Therefore $||\nabla f(x_k)|| \to 0$.  
