Trust-Region Methods

1. Maintain region (typ. ball in some approx. norm) around current iterate in which we "trust" approximate (often quadratic) model.

2. Optimize (approx.) model within this region. As for line search fairly modest quality approx. are good enough (cheap/fast).

3. Check new iterate (func. eval.)

- "very good" agreement model/func.
  - Increase trust region (and poss. make bigger step)

- "good" agreement, accept step return

- "bad" agreement, p
  - Reduce trust region
  - Accept or discard step depending on function improvement
  
  (typically discard)

Typically, TR methods converge faster than LS methods but more work per iteration.
**Typical quadratic model:**

\[ m_k(p) = p_k + g_k^T p + \frac{1}{2} p^T B_k p \]

where \( B_k \) symmetric.

Based on Taylor:

\[ P(x_k + p) = P_k + g_k^T p + \frac{1}{2} p^T \nabla^2_P(x_k + tp) p \]

\( P_k = P(x_k), g_k = \nabla P_k, \) and \( t \in (0,1) \)

\[ m_k(p) - P(x_k + p) = O(||p||^2) \text{ assuming } \nabla^2_P(\cdot) \text{ bounded in sup. large region around } x_k. \]

Taking \( B_k = \nabla^2_P(x_k) \implies \text{error } O(||p||^3) \]

(assuming \( \nabla^2_P \) exists and bounded)

\( \rightarrow \) Trust-Region Newton method

\( \rightarrow \) very acc. if \( ||p|| \) small

In general

\[ \min_{p \in \mathbb{R}} m_k(p) = P_k + g_k^T p + \frac{1}{2} p^T B_k p \]

\( ||p|| \leq \Delta_k \)

As for LS really need only good enough approx. to such \( p \) (actual min of \( m_k(p) \))
Grow/Shrink TR based on

\[ p_h = \frac{p(x_h) - p(x_{h+1})}{m_k(0) - m_k(p_h)} \]

* grow or shrink \( \delta_h \) by fixed fraction based on \( p_h \)

* accept step \( p_h \) if \( p_h \) larger than some modest \( \eta \in [0, \eta_{\text{max}}] \)

** or \( \eta_{\text{max}} = \frac{1}{2} \)**

otherwise reject, shrink \( \delta \) and resolve

* do not increase \( \delta \) of \( ||\nabla f|| < \delta \)

(choice, others do increase \( \delta \))

See algo 4.1 in book

* parameters such that rejection of \( p_h \)

always coincides with reducing \( \delta \)

In order to allow approx. sol. \( p \) (of model problem) at low cost and to prove convergence under weak assumptions, need a "suff. decrease cond."
Consider Cauchy point.

Cauchy point: minimizer of the trust region model problem

\[ m_k(p) = f_k + g_k^T p + \frac{1}{2} g_k^T B_k g_k, \|p\| \leq \Delta_k \]

in direction of negative gradient (steepest descent direction).

\[ p_k^s = -\frac{g_k}{\|g_k\|} \Delta_k \]

\[ p_k^c = \tau_k p_k^s \text{ where} \]

\[ \tau_k = \arg \min_{\tau \geq 0} m_k(\tau p_k^s) \text{ s.t.} \]

\[ \|\tau p_k^s\| \leq \Delta_k \]

The solution for \( \tau_k \) depends on whether

\[ g_k^T B_k g_k < 0 \text{ or } g_k^T B_k g_k > 0 \]

\[ \tau_k = \begin{cases} 1 & g_k^T B_k g_k < 0 \\ \min \left( \frac{\|g_k\|^3}{(\Delta_k g_k^T B_k g_k)}, 1 \right) & \text{otherwise} \end{cases} \]

Since \( p_k^s \) is scaled (negative) gradient, the Cauchy point will in general not lead to fast convergence. (but enough for global conv.)
In general, we want to use $B_k$ to determine step $p_k$. Typical strategy combines $p_k$ with improvement based on quadratic model.

**Dogleg method**

**unconstrained minimizer of $m_k$:**

$B_k g$

If $\|p_k\| \leq \Delta \rightarrow$ solution (for update)

What if $\|p_k\| > \Delta$?

**unconstrained minimizer along steepest descent direction:**

$p_u = -\frac{g^T g}{g^T B g} g$

Define $\tilde{p}(\tau) = \begin{cases} T p_u & 0 \leq \tau \leq 1 \\ p_u + \tau (p_k - p_u) & 1 \leq \tau \leq 2 \end{cases}$

Optimal $\tilde{p}(\alpha)$

If $\|p_u\| > \Delta$, then

$\tau = \frac{\Delta}{\|p_u\|} \quad (\leq 1)$

If $\|p_u\| < \Delta < \|p_k\|$

solve

$\|p_u + (\tau - 1)(p_k - p_u)\|^2 = \Delta^2$

Quadratic problem in $\Delta$ (or in $\tau - 1$)
The dogleg method finds ρ by minimization along path \( \tilde{ρ}(t) \).

Lemma 4.2: Let \( B \) be positive definite. Then

(i) \( \|\tilde{ρ}(t)\| \) is increasing function of \( t \)
(ii) \( m(\tilde{ρ}(t)) \) is decreasing function of \( t \)

(proof in book - mainly matter of diff.)

Assume \( \nabla^2 f(x_k) \) available. If Hessian SPD,

take \( B_k = \nabla^2 f(x_k) \). Otherwise take \( B_k \) to be one of the positive definite modified Hessian discussed before.

Near a minimizer satisfying second order suff. conditions the Newton-TF algorithm becomes Newton's method.

Dogleg most appropriate when objective function convex (hence \( \nabla^2 f \) always SPSD).
Alternative: Two-dim. Subspace Minimization

Extend Dogleg approach by two-dim. min.

\[ \min \{ \mathbf{p} \in \text{span} \{ \mathbf{p}_u, \mathbf{p}_B \} = \text{span} \{ -\mathbf{g}, -\mathbf{B}^{-1}\mathbf{g} \} \] 

\[ \min m(\mathbf{p}) = \mathbf{p}^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{B} \mathbf{p} \quad \text{s.t.} \]

\[ \| \mathbf{p} \| \leq \alpha, \quad \mathbf{p} \in \text{span} \{ \mathbf{g}, \mathbf{B}^{-1}\mathbf{g} \} \]

If \( \mathbf{B} \) indefinite (negative eigenvalues)
replace \( \mathbf{B} \mathbf{g} \) by \( (\mathbf{B} + \alpha \mathbf{I})^{-1} \mathbf{g} \) (as direction)
for some \( \alpha \in (-d_1, -2d_1] \) where \( d_1 \) largest neg. eigenvalue. So \( \mathbf{B} + \alpha \mathbf{I} \) SPD.

When \( \| (\mathbf{B} + \alpha \mathbf{I})^{-1} \mathbf{g} \| \leq \alpha \) we simply take
\[ \mathbf{p} = - (\mathbf{B} + \alpha \mathbf{I})^{-1} \mathbf{g} + \nu \quad \text{where} \quad \nu^T (\mathbf{B} + \alpha \mathbf{I})^{-1} \mathbf{g} \leq 0 \]

When \( \mathbf{B} \) has zero eigenvalues but no neg. eigenvalues, we take \( \mathbf{p} = \mathbf{p}_c \).
Global Convergence

Glob. conv. requires a fixed fraction of the decrease of the model $m(.)$ for Cauchy point

$$DL \& TD \min \text{ satisfy } \frac{\text{crucial assump}}{\text{but all exist}}(\text{but all exist})$$

$$(x_1) \quad m_k(0) - m_k(P_k) \geq c_1 \|g_k\| \min \left(\Delta_k, \frac{\|g_k\|}{\|\Delta_k\|}\right)$$

**Lemma 4.3**: For the Cauchy point, we have $c_1 = \frac{1}{2}$. 

We only need some fraction $c_2$ of reduction obtained by Cauchy point.

**Theo 4.4**: Let $p_k$ be any vector s.t. $\|p_k\| \leq \Delta_k$

and $m_k(0) - m_k(P_k) \geq c_2 (m_k(0) - m_k(p_k^c))$.

Then $p_k$ gives $c_1 = c_2/2$. In particular, if $p_k$ is solution $p_k^*$ of (constr.) model problem (4.3)

then $c_1 = \frac{1}{2}$.

Note that DL and JOM. have $m_k(P_k) \leq m_k(p_k^c)$

hence they have $c_1 = \frac{1}{2}$ (at least) \( \uparrow \) by const.
Further assumptions:

- $B_k$ uniformly bounded in norm
- $P$ bounded below on level set $S = \{ x \mid f(x) \leq f(x_0) \}$

Define $S(R_0) = \{ x \mid \| x - y \| < R_0 \text{ for some } y \in S \}$

($R_0$ positive constant)

Slight generalization $\| P_k \| \leq y h_k$ for some $y \geq 1$

(Alg.1) Case $\eta = 0 \rightarrow$ step accepted when $p_k > 0$

Theo 4.5: Let $\eta = 0$ (Alg 4.1). Suppose $\| B_k \| \leq \beta$

for some $\beta$, $P$ bounded below on $S$ and Lep. cond.

diffe in $S(R_0)$ (some $R_0 > 0$), and that all approx.

sol.s of (4.3) (quadr. model probl.) satisfy

supp. decrease (4.20) and step size constr. (4.25)

for some pos. $\xi$ and $y$. Then

$\lim_{k \to \infty} \| g_k \| = 0$

$\exists \inf$ many steps $k$ s.t. $\| g_k \|$ arb. small
proof two important components:

(i) given some \( \Delta k \), suff. progress is made if step accepted

(ii) there is lower bound on \( \Delta k \) and for that \( \Delta k \) steps always accepted.

\[
\text{Proof: } |p_{k+1} - p_k| = \frac{|f(x_k) - f(x_k + \Delta k) - m_h(\Delta k)\Delta k - m_h(\Delta k)|}{m_h(0) - m_h(\Delta k)}
\]

\[
= \frac{|m_h(\Delta k) - f(x_k + \Delta k)|}{m_h(0) - m_h(\Delta k)}
\]

Taylor: \( f(x_k + \Delta k) = f(x_k) + g(x_k)^T \Delta k + \int_0^1 g(x_k + t\Delta k)^T \Delta k \, dt \)

\[
m_h(\Delta k) - f(x_k + \Delta k) = \frac{\Delta k}{2} \frac{\partial^2 f(x_k)}{\partial^2 \Delta k} \Delta k + \frac{1}{2} \Delta k^T B_k \Delta k
\]

\[
\leq \frac{1}{2} (\beta \| \Delta k \|^2 + \frac{\mu}{\beta} \| p_k \|^2)
\]

where \( 2 \beta \) is Lip constant for \( g(\cdot) \) on \( S(R_0) \).

We also assume that \( \| p_k \| \leq R_0 \), so that \( x_k \) and \( x_k + tp_k \) in \( S(R_0) \).
Now assume \( \lim_{k \to \infty} \|g_k\| = 0 \to 78 \)

There is \( \varepsilon > 0 \) and \( K > 0 \) s.t.

\[ \|g_k\| > \varepsilon \quad \text{for all } k > K. \]

(proof by contradiction)

For \( k \geq K \),

\[ m_k(0) - m_k(r_k) \geq \varepsilon \|g_k\| \min(A_k, \frac{\|g_k\|}{c_1}) \]

\[ = c_1 \varepsilon \min(A_k, \frac{\varepsilon}{c_1}) \]

Hence \( |f_k - 1| \leq \frac{\gamma^2 \Delta_k^2 (3/2 + \beta)}{c_1 \varepsilon \min(A_k, \varepsilon/c_1)} \)

To derive bound for \( \|r_k\| \) rhs, we show \( A_k \) does not get arbitrarily small.

\[ A = \min\left( \frac{1}{2} \frac{c_1 \varepsilon}{\gamma^2 (3/2 + \beta)}, \frac{R_0}{\gamma} \right) \quad \text{for suff. small steps} \]

\( R_0/\gamma \) ensures \( x_k + p_k \) inside \( S(R_0) \), since

\[ \|P_k\| \leq \gamma \Delta_k \leq \gamma \overline{A} \leq R_0 \quad \text{for suff. small } A_k \]

Consider \( A_k \leq \overline{A} \). Since \( c_1 \varepsilon \leq 1 \) and \( \gamma > 1 \)

we have \( \overline{A} \leq \varepsilon/c_1 \). So, for \( A_k \leq [0, \overline{A}] \)

we also have \( A_k \leq \varepsilon/c_1 \) and \( \min(A_k, \varepsilon/c_1) = A_k \)

\[ |f_k - 1| \leq \frac{\gamma^2 \Delta_k^2 (3/2 + \beta)}{c_1 \varepsilon \Delta_k} = \frac{\gamma^2 \Delta_k (3/2 + \beta)}{c_1 \varepsilon} \]

\[ \frac{\gamma^2 \Delta \left(3/2 + \beta_1\right)}{c_1 \varepsilon} \leq \frac{1}{2} \Rightarrow p_k \geq \frac{1}{4} \]
Therefore $\Delta k + 1 \geq \Delta k$ and $x_k + p_k$ accepted, for all $k \geq k$.

(Nota lower bound on $11g_{b1}$ and on $a_k$ implies some fixed reduction in $f(x_{k+1})$ each time $p_k \geq \frac{1}{4}$ (step accepted). Inf. number of such steps implies $f$ not bounded below.

However, finitely many such $11g_{b1}$ implies every many steps $p_k < \frac{1}{4}$, implying $\Delta k \to 0$; contradiction.)

Suppose, there is inf. subseq. $K$ s.t. $p_k \geq \frac{1}{4}$ for $k \in K$. Then for $k \in K$, $k \geq k$ we have $f(x_k) - f(x_{k+1}) = \frac{f(x_k)}{4} f(x_k + p_k) \geq \frac{1}{4} (m_k(0) - m_k(p_k)) \geq \frac{1}{4} \epsilon \Delta k \geq \epsilon / \Delta k$, $K$ exists.

Since $f$ bnd below, this implies $\lim_{k \to \infty} \Delta k = 0$. This gives contradiction.

Hence no such subseq. $K$ can exist.
But then we must have $\frac{p_k}{k} < \frac{1}{2}$ for all $k$ sufficiently large, again implying $A_k \to 0$
(because TR decreased by factor $\frac{1}{2}$). This again leads to a contradiction.

Hence assumption $\|g_k\| > \varepsilon$ for all $k > k$ must be false.

**Theo 4.6:** If we take, in addition to conditions Theo 4.5, $\eta > 0 \ (\eta \in (0, 1))$, then

$$\lim_{k \to \infty} \|g_k\| = 0$$

Proof: extension of proof Theo 4.5

(Read section 4.5 on scaling and use of alternative norms.)
Local Convergence (Just state main ideas)

If \( p_k \rightarrow p_{-\infty} \) for large enough \( k \), we get (eventually) quadratic convergence, as long as \( p_{-\infty} \) fits inside trust region.

Intuitively, should be true. As we saw, \( s_k \) is bad away from zero under mild cond.s while \( p_k \rightarrow 0 \) if we conv. to model.

If TR solution (approx. sol.) is inside TR and conv. to Newton step, we say updates asympt. similar to Newton steps.

Theo. 9. Let \( f \) twice lip. cont. diff. in nbh of \( x^* \) sats. 2nd order suff. cond.s, \( x_k \rightarrow x^* \) and (k large enough) TR with \( B_k = \nabla^2 f_k \) (std quadr. model) takes steps \( p_k \) s.t.

\[ m_k(0) - m_k(p_k) \geq c \, \| g_k \| \, \min(\lambda_k, \frac{\| g_k \|}{\| s_k \|}) \]

and asympt. similar to Newton steps whenever

\[ \| p_k \| \leq \frac{1}{2} \Delta_k : \| p_{k+1} - p_k \| = o(\| p_k \|) \]

\( \rightarrow \) Statement about alg.
Then TR inactive for $k$ suff. large and

$x_k \to x^*$ superlin.

For $k$ large enough $\|P_k - p^n_k\| < \|p^n_k\|$

$\|P_k\| < \frac{1}{2} \Delta k \Rightarrow \|P_k\| < 2 \|P_k^n\|$

$\|P_k^n\| > \frac{1}{2} \Delta k \Rightarrow \|P_k\| \leq \Delta k < 2 \|P_k^n\|$

$\|P_k\| \leq 2 \|\nabla^2 f(x_k)\| \|g_k\| \Rightarrow$

$\|g_k\| \geq \frac{1}{2} \frac{\|P_k\|}{\|\nabla^2 f_k\|}$

$m_k(0) - m_k(P_k) \geq c_1 \|g_k\| \min(\Delta k, \frac{\|g_k\|}{\|\nabla^2 f_k\|})$

$\geq c_1 \frac{1}{2} \frac{\|P_k\|}{\|\nabla^2 f_k\|} \cdot \min(\|P_k\|, \frac{\|P_k\|}{2 \|\nabla^2 f_k\|})$

$\geq \frac{\|P_k\|}{2 \|\nabla^2 f_k\|} \cdot \frac{\|P_k\|}{2 \|\nabla^2 f_k\|} \|\nabla^2 f_k\| \\ \geq 1$

(by concl.) $\geq c_1 \frac{\|P_k\|^2}{8 \|\nabla^2 f_k\| \|\nabla^2 f_k\|} = c_3$

$\frac{m_k(0) - m_k(P_k)}{1} \geq c_3 \|P_k\|^2$

$\frac{m_k(0) - m_k(P_k) + P_k(x_k) - F(x_k + P_k)}{1} = \frac{\frac{1}{2} P_k^T \nabla^2 f_k P_k - \frac{1}{2} \int P_k^T \nabla^2 f(x_k + \tau P_k) P_k d\tau}{1} \leq \frac{1}{2} \frac{\|P_k\|^2}{4}$
\[ |p_k - 1| \leq \frac{\|p_k\|^2 (L / c)}{c_3 \|p_k\|^2} \leq \frac{L}{4c_3} \Delta_k \]

TR reduced if \( p_k < \frac{1}{4} \) (or some other choice < 1)

Hence (again) \( \Delta_k \) and away from zero.

Since \( x_k \to x^* \), \( p_k^N \to 0 \Rightarrow p_k \to 0 \) (by assumption, \( \|p_k - p_k^N\| = O(\|p_k^N\|) \))

Hence TR inactive for \( k \) sufficiently large and hence \( \|p_k^N\| < \frac{1}{2} \Delta_k \) for \( k \) sufficiently large.

As shown before, \( \|x_k - x^* + p_k^N\| = O(\|x_k - x^*\|^2) \)

So \( \|p_k^N\| = O(\|x_k - x^*\|) \) and

\[ \|x_k + p_k^N - x^*\| \leq \|x_k - x^*\| + \|p_k - p_k^N\| = \begin{cases} O(\|x_k - x^*\|^2) + O(\|p_k^N\|) = O(\|x_k - x^*\|) \\ \end{cases} \]

Note that first part of proof shows that methods like DL give \( p_k = p_k^N \) for \( k \) sufficiently large and hence yield quad. conv.
Some Notes

$S(R_0)$ in general is not convex

However, for $A_k < 0$, $P_k \times k + P_k \in S(R_0)$