Trust-Region Methods

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

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

(3) Check new iterate (func. eval.)

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

"good" agreement, accept step
~~redu~~

"bad" agreement, ~~R~~
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 quadr. model:

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p$$

where B_k symmetric

Based on Taylor:

$$f(x_k+p) = f_k + g_k^T p + \frac{1}{2} p^T \{\nabla^2 f(x_k+tp)\} p$$

$$f_k = f(x_k), \; g_k = \nabla f_k, \text{ and } t \in (0,1)$$

$m_k(p) - f(x_k+p) = O(\|p\|^2)$ assuming $\nabla^2 f(\,)$ bounded in suff. large region around x_k.

Taking $B_k = \nabla^2 f(x_k) \rightarrow$ error $O(\|p\|^3)$ (assuming $\nabla^3 f$ exists and bounded)

$\rightarrow$ Trust-Region Newton method

$\rightarrow$ very acc. if $\|p\|$ small

In general

$$\min_{\substack{p \in \mathbb{R} \\ \|p\| \leq \Delta_k}} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p$$

As for LS really need only good enough appr. to such p (actual min of $m_k(p)$)

Grow / Shrink TR based on

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

* grow or shrink Δ_k by fixed fraction based on ρ_k
* accept step p_k if ρ_k larger than some modest $\eta \in [0, \eta_{max})$ (eg. $\eta_{max} = \frac{1}{4}$)
  otherwise reject, shrink Δ and resolve *
* do not increase Δ if $\|p_k\| < \Delta$
  (choice, others do increase Δ)

See <u>alg</u> 4.1 in book
* parameters such that rejection of p_k always coincides with reducing Δ.

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 <u>Cauchy point</u>

Cauchy point: minimizer of the trust region model problem

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p \,, \quad \|p\| \le \Delta_k$$

in direction of negative gradient (steepest descent direction).

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

$$p_k^c = \tau_k p_k^s \quad \text{where}$$

$$\tau_k = \arg\min_{\tau \geq 0} m_k(\tau p_k^s) \quad \text{s.t.} \quad \|\tau p_k^s\| \le \Delta_k$$

The solution for τ_k depends on whether

$g_k^T B_k g_k \le 0$ or $g_k^T B_k g_k > 0$

$$\tau_k = \begin{cases} 1 & \text{if } g_k^T B_k g_k \le 0 \\ \min\left(\dfrac{\|g_k\|^3}{\Delta_k g_k^T B_k g_k}, 1 \right) & \text{otherwise} \end{cases}$$

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^c with improvement based on quadratic model.

Dog leg method

unconstrained minimizer of m_k: $p_B = -B^{-1}g$

if $\|p_B\| \le \Delta \rightarrow$ solution (for update)

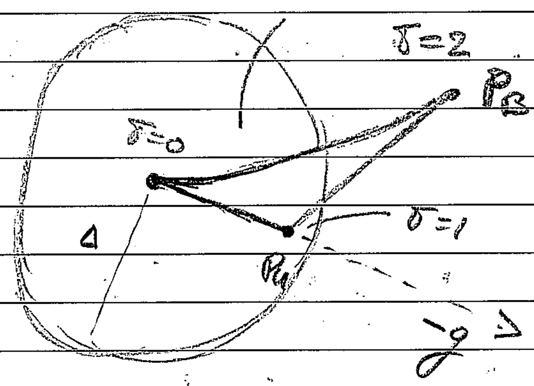
What if $\|p_B\| > \Delta$?

unconstrained minimizer along steepest descent direction:

$$p_u = -\frac{g^Tg}{g^TBg}\, g$$

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

optimal $p(\Delta)$



if $\|p_u\| \ge \Delta$ then

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

if $\|p_u\| < \Delta < \|p_B\|$

solve

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

quadratic problem in τ (or in $\tau - 1$)

The dogleg method finds p by minimization along path $\tilde{p}(\tau)$.

Lemma 4.2 : Let B be positive definite. Then
(i) $\|\tilde{p}(\tau)\|$ is increasing function of τ
(ii) $m(\tilde{p}(\tau))$ is decreasing function of τ

(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 minizer satisfying second order suff. conditions the Newton TR 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. in $\text{span}\{p_U, p_B\} = \text{span}\{-g, -B^{-1}g\}$

$$\min_p m(p) = f + g^T p + \tfrac{1}{2} p^T B p \quad \text{s.t.}$$
$$\|p\| \leq \Delta, \quad p \in \text{span}\{g, B^{-1}g\}$$

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

When $\|(B+\alpha I)^{-1}g\| \leq \Delta$ we simply take $p = -(B+\alpha I)^{-1}g + v$ where $v^T(B+\alpha I)^{-1}g \leq 0$

When B has zero eigenvalues but no neg. eigenvalues, we take $p = p^c$

Global Convergence

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

DL & $\nabla\Phi$ min satisfy — crucial assump. (but alt. exist)

(*1) $m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$

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 $c_1 = 1/2$.

Note that DL and $\nabla\Phi M$ have $m_k(p_k) \leq m_k(p_k^c)$ — by constr.
hence they have $c_1 = \frac{1}{2}$ (at least)

Further assumptions:

- B_k uniformly bounded in norm
- f bounded below on level set

$$S = \{x \mid f(x) \le f(x_0)\}$$

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

Slight generalization $\|p_k\| \le \gamma \Delta_k$ for some
$\gamma \ge 1$

(Algu.1) Case $\eta = 0$ $\rightarrow$ step accepted when $\rho_k > 0$

Theo 4.5: let $\eta = 0$ (alg 4.1). Suppose $\|B_k\| \le \beta$ for some β, f bnd below on S and Lip. cont. diff in $S(R_0)$ (some $R_0 > 0$), and that all approx. sol.s of (4.3) (quadr. model probl.) satisfy suff. decrease (4.20) and step size constr. (4.25) for some pos. c_1 and γ. Then

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

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

(i) given some Δ_k suff. progress is made if step accepted

(ii) there is lower bnd on Δ_k and for that Δ_k steps always accepted.

<u>Proof</u> $$|\rho_k - 1| = \left| \frac{f(x_k) - f(x_k+p_k)}{m_k(0) - m_k(p_k)} - \frac{m_k(0) - m_k(p_k)}{m_k(0) - m_k(p_k)} \right|$$

$$= \left| \frac{m_k(p_k) - f(x_k+p_k)}{m_k(0) - m_k(p_k)} \right|$$

Taylor: $$f(x_k+p_k) = f(x_k) + g(x_k)^T p_k + \int_0^1 [g(x_k+tp_k) - g(x_k)]^T p_k \, dt$$

$$m_k(p_k) - f(x_k+p_k) = \cancel{f_k} + \cancel{g_k^T p_k} + \tfrac{1}{2} p_k^T B_k p_k - \cancel{f_k} - \cancel{g_k^T p_k} - \int_0^1 [g(x_k+tp_k) - g(x_k)]^T p_k \, dt$$

$$|m_k(p_k) - f(x_k+p_k)| = \left| \tfrac{1}{2} p_k^T B_k p_k - \int_0^1 [g(x_k+tp_k) - g(x_k)]^T p_k \, dt \right|$$

$$\le \tfrac{1}{2} \beta \|p_k\|^2 + \beta_1 \|p_k\|^2$$

where $2\beta_1$ is Lip constant for $g(\,)$ on $S(R_0)$.

We also assume that $\|p_k\| \le R_0$, so that x_k and $x_k + tp_k$ in $S(R_0)$.

Now assume $\liminf_{k\to\infty} \|g_k\| \neq 0 \rightarrow$

There is $\varepsilon > 0$ and $K > 0$ s.t.

$\|g_k\| \geq \varepsilon$ for all $k \geq K$.

(proof by contradiction)

For $k \geq K$, $m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$

$$= c_1 \varepsilon \min\left(\Delta_k, \frac{\varepsilon}{\beta}\right)$$

Hence $|\rho_k - 1| \leq \dfrac{\gamma^2 \Delta_k^2 (\beta/2 + \beta_1)}{c_1 \varepsilon \min(\Delta_k, \varepsilon/\beta)}$

To derive bnd for rhs, we show Δ_k does not get arbitrarily small.

$$\bar{\Delta} = \min\left(\frac{1}{2} \frac{c_1 \varepsilon}{\gamma^2(\beta/2 + \beta_1)}, \frac{R_0}{\gamma}\right)$$

for suff. small steps

R_0/γ ensures $x_k + p_k$ inside $S(R_0)$, since

$\|p_k\| \leq \gamma \Delta_k \leq \gamma \bar{\Delta} \leq R_0$ for suff. small Δ_k

Consider $\Delta_k \leq \bar{\Delta}$. Since $c_1 \leq 1$ and $\gamma \geq 1$

we have $\bar{\Delta} \leq \varepsilon/\beta$. So, for $\Delta_k \in [0, \bar{\Delta}]$

we also have $\Delta_k \leq \varepsilon/\beta$ and $\min(\Delta_k, \varepsilon/\beta) = \Delta_k$

$$|\rho_k - 1| \leq \frac{\gamma^2 \Delta_k^2 (\beta/2 + \beta_1)}{c_1 \varepsilon \Delta_k} = \frac{\gamma^2 \Delta_k (\beta/2 + \beta_1)}{c_1 \varepsilon} \leq$$

$$\frac{\gamma^2 \bar{\Delta} (\beta/2 + \beta_1)}{c_1 \varepsilon} \leq \frac{1}{2} \Rightarrow \rho_k > \frac{1}{4}$$

Therefore $\Delta_{k+1} \geq \Delta_k$ and $x_k + p_k$ accepted, whenever $\Delta_k \leq \bar{\Delta}$. Hence $\Delta_k \geq \min(\Delta_K, \bar{\Delta}/4)$ for all $k \geq K$.

(Note lower bnd on $\|g_k\|$ and on Δ_k implies some fixed reduction in $f(x_{k+1})$ each time $\rho_k \geq \frac{1}{4}$ (step accepted). Inf. number of such steps implies f not bnd below. However, finitely many such its implies inf. many steps $\rho_k < \frac{1}{4}$, implying $\Delta_k \to 0$: contradiction)

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

$$\frac{1}{4}(m_k(0) - m_k(p_k)) \geq \frac{1}{4} c_1 \varepsilon \min(\Delta_k, \varepsilon/\beta)$$

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

Hence no such subseq. $\mathcal{K}$ can exist.

But then we must have $\rho_k < \frac{1}{4}$ for all k suff. large, again implying $\Delta_k \to 0$ (because TR decreased by factor $\frac{1}{4}$). This again leads to a contradiction.

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

<u>Theo 4.6</u>: If we take, in addition to cond.s theo 4.5, $\eta > 0$ ($\eta \in (0, 1/4)$), ~~we~~ 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 $p_k \to p_k^N$ for large enough k we get (eventually) quadratic convergence, as long as p_k^N fits inside trust region.

Intuitively, should be true. As we saw Δ_k is bnd away from zero under mild con.d.s while $p_k \to 0$ if we conv.

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

Theo 4.9 let f twice Lip. cont. diff. in nbh of x^* satisf. 2nd order suff. cond.s, $x_k \to x^*$ and (k large enough) TR with $B_k = \nabla^2 f_k$ (std quadr. model) takes steps p_k s.t.

$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$ and asympt similar to Newton steps whenever $\|p_k\| \leq \frac{1}{2}\Delta_k$: $\|p_k - p_k^N\| = o(\|p_k^N\|)$

↳ statement about alg.

Then TR inactive for k suff. large and

$x_k \to x^*$ superlin.

For k large enough $\|p_k - p_k^N\| < \|p_k^N\|$

$$\left.\begin{aligned} \|p_k^N\| \le \tfrac{1}{2}\Delta_k &\Rightarrow \|p_k\| \le 2\|p_k^N\| \\ \|p_k^N\| > \tfrac{1}{2}\Delta_k &\Rightarrow \|p_k\| \le \Delta_k \le 2\|p_k^N\| \end{aligned}\right\} \Rightarrow$$

$$\|p_k\| \le 2\|\nabla^2 f(x_k)^{-1}\|\,\|g_k\| \Rightarrow$$

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

$$m_k(0) - m_k(p_k) \geq c_1\|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|\nabla^2 f_k\|}\right)$$

$$\geq c_1 \frac{1}{2}\frac{\|p_k\|}{\|\nabla^2 f_k^{-1}\|}\cdot \min\left(\|p_k\|, \frac{\|p_k\|}{\underbrace{2\|\nabla^2 f_k^{-1}\|\,\|\nabla^2 f_k\|}_{\geq 1}}\right)$$

$$\geq c_1 \frac{\|p_k\|}{2\|\nabla^2 f_k\|}\cdot\frac{\|p_k\|}{2\|\nabla^2 f_k^{-1}\|\,\|\nabla^2 f_k\|}$$

$$(\text{by conv.}) \geq c_1 \frac{\|p_k\|^2}{8\|\nabla^2 f_*^{-1}\|\,\|\nabla^2 f_*\|} \qquad \equiv c_3$$

$$m_k(0) - m_k(p_k) \geq c_3\|p_k\|^2$$

$$|m_k(0) - m_k(p_k) + f_k(x_k) - f(x_k+p_k)| =$$

$$\left|\frac{1}{2}p_k^T\nabla^2 f_k\, p_k - \frac{1}{2}\int_0^1 p_k^T\nabla^2 f(x_k+t p_k)\, p_k\, dt\right| \le$$

$$\frac{L}{4}\|p_k\|^3$$

$$|\rho_k - 1| \le \frac{\|p_k\|^3 (L/4)}{c_3 \|p_k\|^2} \le \frac{L}{4c_3} \Delta_k$$

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

Hence (again) Δ_k bnd away from zero.

Since $x_k \to x^*$, $p_k^N \to 0 \Rightarrow p_k \to 0$ (by assump. $\|p_k - p_k^N\| = o(\|p_k^N\|)$)

Hence TR inactive for k suff. large and hence $\|p_k^N\| \le \frac{1}{2}\Delta_k$ for k suff. large.

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

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

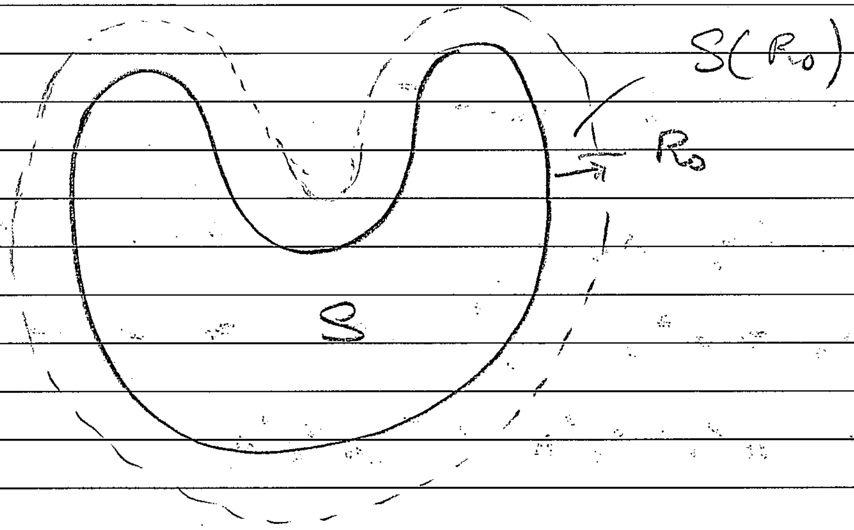
$$\|x_k + p_k - x^*\| \le \|x_k + p_k^N - x^*\| + \|p_k - p_k^N\| =$$

$$o(\|x_k - x^*\|^2) + o(\|p_k^N\|) = o(\|x_k - x^*\|)$$

Note that first part of proof shows that methods like DL give $p_k = p_k^N$ for k suff. large and hence yield quadr. conv.

Some Notes

$S(R_0)$ in general is not convex



However, for $\Delta_k \le \bar{\Delta}$ $\cancel{p_k}$ $x_k+p_k \in S(R_0)$