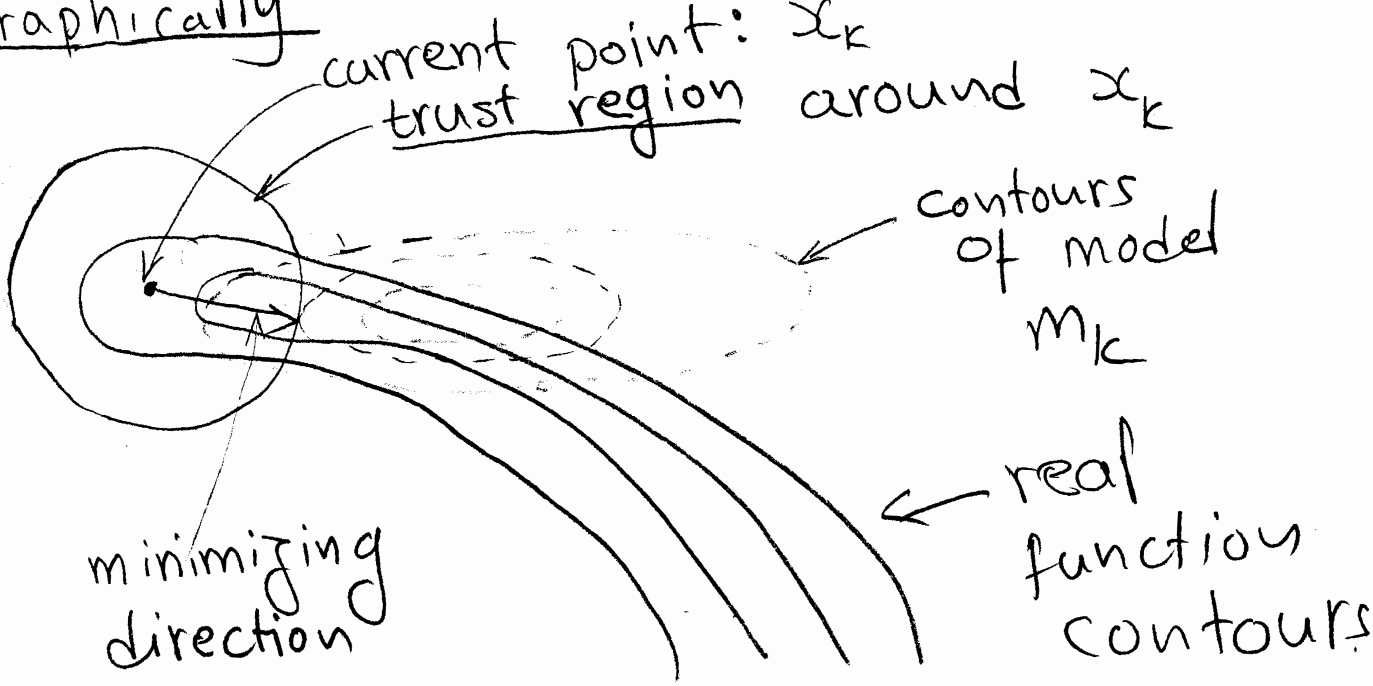
Chapter 4 Trust-Region Methods "First part"

Basic idea:

* Fit a model around the current location.
* Reduce size if inadequate
* Increase size if model fits well
* minimize function within region

Graphically



Fit a quadratic model:

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

* If $B_k = \nabla^2 f(x_k + tp)$, then the error is $O(\|p\|^3)$. Else, the error is $O(\|p\|^2)$.
* When B_k is the Hessian, we have the trust-region Newton method.

Basic minimization problem is:

$$\min_{p \in \mathbb{R}^n} m_k(p) \quad \text{such that} \quad \|p\| \leq \Delta_k.$$

$$\text{or: } \left(p^T p \leq \Delta_k^2\right)$$

Suppose:

* B_k is positive definite,
* $\|B_k^{-1} \nabla f_k\| \leq \Delta_k$

Then $p_k^B = -B_k^{-1} \nabla f_k$ (the unconstrained minimum)

Else approximate

Reduction Ratio

For the step p_k:

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

We have $m_k(0) - m_k(p_k) > 0$ by design.

Thus:

* If $\rho_k < 0$, we are actually moving in a direction of increasing $f(\cdot)$, and the step is rejected, and reduce trust region
* if $\rho_k \simeq 1$ then we have very good agreement. If we are also taking the maximum possible step: $\|p_k\| = \Delta_k$, then increase the trust region

Algorithm 4.1

Given:

$\bar{\Delta} > 0$: maximum allowable radius.

$\Delta_0 \in (0, \bar{\Delta})$: initial value for trust-region

$\eta \in [0, 1/4)$: minimum acceptable reduction threshold.

for $k = 0, 1, 2, \ldots$

Obtain p_k by approximately solving:

$$\begin{cases} \min\limits_{p \in \mathbb{R}^n} m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \\ \text{such that: } \|p\| \le \Delta_k \end{cases}$$

Evaluate: $\rho_k = \dfrac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$ ← actual reduction / ← model reduction

if $\rho_k < 1/4$ ← (insufficient reduction or perhaps increase!)

$\Delta_{k+1} = \frac{1}{4}\Delta_k$ ← Reduce trust-region, where model is more applicable

Correct me

else

if $\rho_k > 3/4$ and $\|p_k\| = \Delta_k$ ← if sufficient reduction, increase region within bound

$\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$;

else $\Delta_{k+1} = \Delta_k$; ← else, keep the size.

if $\rho_k > \eta$ then $x_{k+1} = x_k + p_k$; ← and take the step...

else $x_{k+1} = x_k$

end

4.1 The Cauchy Point & Related Alg.

Ch 4–5 / 12

From <u>steepest descent</u>, solve

$$p_k^s = \arg\min_{p \in \mathbb{R}^n} f_k + \nabla f_k^T \underline{p}, \quad \|p\| \leq \Delta_k$$

using: $\boxed{p_k^s = \frac{-\Delta_k}{\|\nabla f_k\|} \nabla f_k}$

Then test this on the local min.:

$$\circledast - \tau_k = \arg\min_{\tau > 0} m_k(\tau p_k^s) \quad \|\tau p_k^s\| \leq \Delta_k$$

to compute the Cauchy direction:

$$\boxed{p_k^c = \tau_k p_k^s}$$

which is line seach subject to $\circledast$.

The proposed solution is:

$$\boxed{\tau_k = \begin{cases} 1 & \underline{\text{if}} \quad \nabla f_k^T B_k \nabla f_k \leq 0 \\ \min\left(\|\nabla f_k\|^3 / (\Delta_k \nabla f_k^T B_k \nabla f_k), 1\right), & \underline{\text{otherwise}}. \end{cases}}$$

Improving on the Cauchy Point

We want to improve convergence by minimizing:

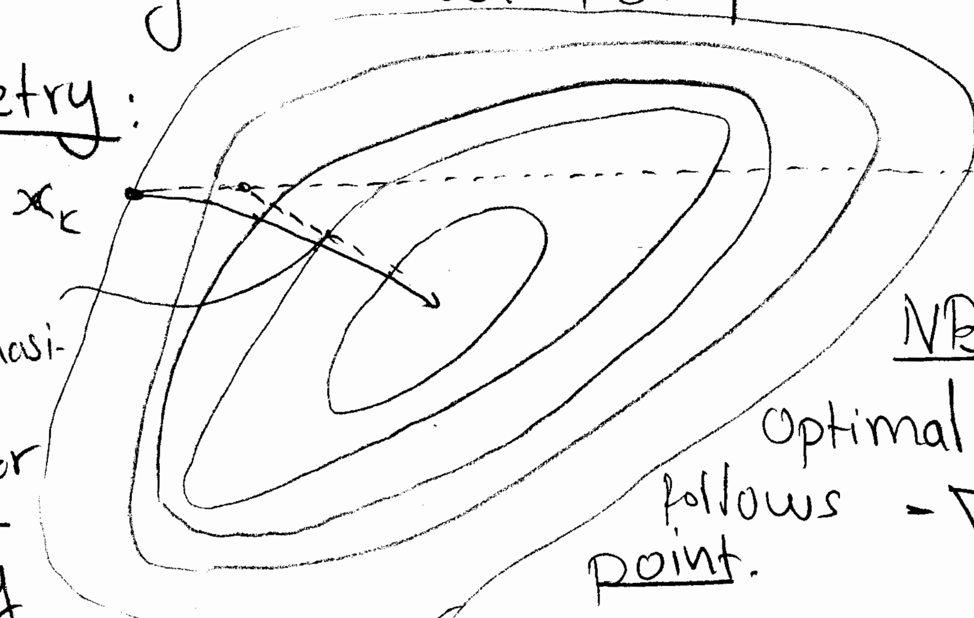
$$\min_{p \in \mathbb{R}^n} m(\underline{p}) = f + \nabla f^T \underline{p} + \tfrac{1}{2} \underline{p}^T B \underline{p}, \quad \|\underline{p}\| \le \Delta$$

at every step. Solution is $\underline{p}^*(\Delta)$. Dropping k

Dogleg Method

Basic idea: approximate optimal trajectory by first taking a steepest descent step along $-\nabla f$, and then taking a Newton-step at $-B^{-1}\nabla f$.

Geometry:



NB: Optimal trajectory follows $-\nabla f$ at every point.

Dogleg direction is given in terms of two line segments:

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

where the steepest-descent direction is simply:

* $$p^U = -\frac{g^T g}{\underbrace{g^T B g}} g, \quad g = \nabla f_k.$$

Note that the denominator can be negative, if B is <u>not</u> positive definite.
Book recommends that we don't apply the Dogleg method in this case! (set $|g^T B g|$?).

* $$p^B = -B^{-1} g, \quad g = \nabla f_k$$

* to compute τ:
  → <u>if</u> $\| p^U \| \leq \Delta$, <u>take</u> $\tilde{p} = \frac{\Delta}{\| p^U \|} \times p^U$
  → <u>Else</u>, solve:

$$\| p^U + (\tau^* - 1)(p^B - p^U) \|^2 = \Delta^2 \quad (A)$$

for τ^* and take $\tilde{p}(\tau^*)$.

Notes:

* The optimal solution is
$$p^*(\Delta) = p^B \quad \text{when} \quad \Delta \geq \|p^B\|$$

* The dogleg method allows to define a solution for all $\|p\|$, thus allowing us to change the trust-region size as we wish.

* Lemma 4.1 shows that everything works for B positive definite.

* two-dimensional subspace minimization replaces p by $p = \alpha_1 \nabla f + \alpha_2 B^{-1} \nabla f$ and computes the minimum of:
$$f + \nabla f^T p + \tfrac{1}{2} p^T B p \quad \text{and} \quad \Delta$$
by computing α_1, α_2 that satisfy both equations. "Book errata" claim that this leads to a 4th-order polynomial, for which we can use Maple/Mathematica to solve algebraically.

* if B is not positive definite, the method may somehow be corrected by using: $(B + \alpha I)$ instead of B, so that $(B + \alpha I)$ has positive eigenvalues, and it is positive-def. Yet, the improvements may not be as great as we would hope (see book).

Steihaug's Approach (excellent properties, if based on Newton).

* Correct algorithm 4.3 by changing the last line (see book errata):

$$d_{j+1} = -r_{j+1} + \beta_{j+1} d_j$$

CG-Steihaug

Given $\varepsilon > 0$;

Set $p_0 = 0$, $r_0 = -\nabla f_0$, $d_0 = -r_0$;

if $(\|r_0\| < \varepsilon)$ ← Careful on how to implement this! See Dennis & Schnabel.

return $p = p_0$; ← and. terminate the search!

for $j = 0, 1, 2, \ldots, n$ ← up to n, the dimension of B!

if $(d_j^T B d_j \leq 0)$ ← negative definite "behavior" for B, terminate.

Find τ for $m(p)$, $\|p\| = \Delta$, $p = p_j + \tau d_j$

return$(p = p_j + \tau d_j)$; ← see note 1.

$\alpha_j = r_j^T r_j / d_j^T B d_j$; ← CG-Step

$p_{j+1} = p_j + \alpha_j d_j$ ← direction CG

if $(\|p_{j+1}\| \geq \Delta)$ ← outside trust region?

Find $\tau \geq 0$ s.t. $p = p_j + \tau d_j$ satisfies $\|p\| = \Delta$ ← see note 2

return p;

$r_{j+1} = r_j + \alpha_j B d_j$;

if $\|r_{j+1}\| < \varepsilon \|r_0\|$ ← if sufficient residual reduction, terminate, (watch implementation!)

return $p = p_{j+1}$;

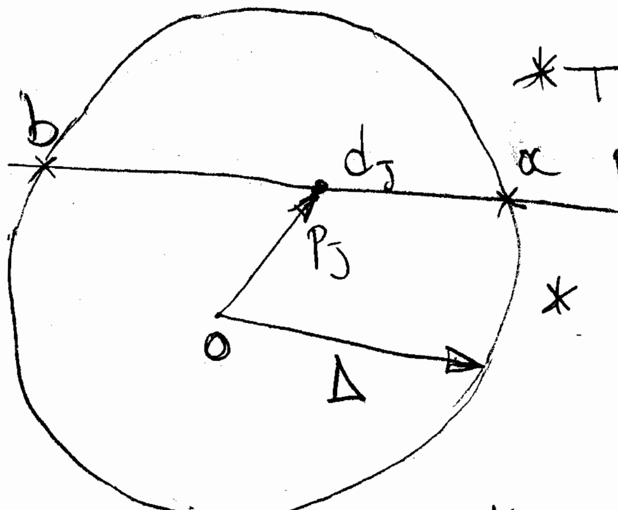
$\beta_{j+1} = r_{j+1}^T r_{j+1} / r_j^T r_j$; ← new step-length

$d_{j+1} = -r_{j+1} + \beta_{j+1} d_j$;

correct the text (see book errata).

end

Note 1:



* There are only two points to consider.

* Find p_1, p_2 from:
$\|p_j + \tau d_j\|^2 = \Delta^2$

* Bisection with $\tau = 0$, τ large > 0, or $\tau = 0$, τ large < 0, and then evaluate $f + \nabla f^T p + \frac{1}{2} p^T B p$ to find which one gives the minimum (or use Newton ...)

Note 2: We have the same situation, but now, we only look for the solution for $\tau > 0$.

4.2 Using Nearly Exact Solutions

Start with B! (memorize!)

* Note that B is always symmetric.
* If the estimated B is not symmetric, then set B to $B = \frac{1}{2}\hat{B} + \frac{1}{2}\hat{B}^T$ where $\hat{B}$ denotes an estimate.
* Recall that every symmetric matrix has an eigen decomposition of the form:

$$B = \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ e_1 & \cdots & e_n \\ \downarrow & & \downarrow \end{bmatrix}}_{\substack{Q \\ \text{eigenvector matrix}}} \underbrace{\begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \; \lambda_n \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} \leftarrow e_1^T \rightarrow \\ \vdots \\ \leftarrow e_n^T \rightarrow \end{bmatrix}}_{Q^T}$$

Note that: $QQ^T = I$

Alternatively, we write:

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n), \quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

$$Q = [q_1 \; q_2 \cdots q_n]$$

Now, add λI:

$$
\begin{aligned}
B + \lambda I &= Q \Lambda Q^T + \lambda I \\
&= Q \Lambda Q^T + \lambda Q Q^T \\
&= Q \left(\Lambda Q^T + \lambda I \cdot Q^T \right) \\
&= Q \underbrace{\left(\Lambda + \lambda I \right)}_{\text{new eigenvalue matrix. } \Lambda'} Q^T
\end{aligned}
$$

$$\Lambda' = \Lambda + \lambda I$$

$$\Rightarrow \Lambda' = \operatorname{diag}\left(\lambda_1 + \lambda, \lambda_2 + \lambda, \dots, \lambda_n + \lambda \right)$$

Clearly, if we look at the minimum:

$$\underbrace{\min_i \lambda_i + \lambda \geq 0}_{\substack{\text{require this for} \\ B+\lambda I \text{ to be positive} \\ \text{semi-definite (all } \lambda_i > 0)}} \Rightarrow \boxed{\lambda \geq -\min_i \lambda_i}$$

(*)

⇒ From ⊛, if B was positive definite,

$-\min_i \lambda_i < 0$ and $\lambda = 0$ is a solution

⇒ If B has negative eigenvalues, the $\min_i \lambda_i$ is the <u>smallest</u> eigenvalue: $\boxed{\min_i \lambda_i = \lambda_1}$

$\Rightarrow \boxed{\lambda \geq -\lambda_1}$

We have thus shown that $(B+\lambda I)$ is positive semidefinite for $\lambda \geq -\lambda_1$.

{ Recall that a <u>feasible</u> point refers to a point for which all required conditions are satisfied.

The key to finding a global solution is theorem 4.3.

It applies to all B!

Thm 4.3

For $\min_{p \in \mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p$,

with $\|p\| \leq \Delta$,

we have a solution p^* iff

(I) $(B + \lambda I) p^* = -g$, — (Δ)

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

(III) $(B + \lambda I)$ is positive semidefinite,

and (I)+(II)+(III) are satisfied for some $\lambda \geq 0$.

* Note that $\lambda(\Delta - \|p^*\|) = 0$ requires that either $\lambda = 0$ or $\Delta = \|p^*\|$.

* for the solution:

$$-\nabla m(p^*) = \underbrace{-Bp^* - g = \lambda p^*}_{\text{from (Δ)}}$$

Now, I + III have been discussed, and if $(B+\lambda I)$ is positive definite, both can be satisfied for some $\lambda > \lambda_1$.

For II, we simply require:

$$\|p(\lambda)\| = \Delta \quad — \quad \mathrm{IV}$$

to be solved (unless B was already positive definite $\Rightarrow \lambda = 0$).

$\Rightarrow$ Solve IV using root-finding methods.

Now, from I, we have:

$$p(\lambda) = -(B+\lambda I)^{-1} g$$

$$\Rightarrow (B+\lambda I)^{-1} \cdot (B+\lambda I) = \underbrace{(B+\lambda I)^{-1}}_{\text{symmetric}} \underbrace{Q(\Lambda+\lambda I)Q^T}_{\uparrow \text{ invert}} = I$$

$$\Rightarrow (B+\lambda I)^{-1} = Q(\Lambda+\lambda I)^{-1} Q^T$$

$$\Rightarrow p(\lambda) = -\underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ q_1 & \cdots & q_n \\ \downarrow & & \downarrow \end{bmatrix}}_{Q} \begin{bmatrix} \frac{1}{\lambda_1+\lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_n+\lambda} \end{bmatrix} \begin{bmatrix} \leftarrow q_1^T \rightarrow \\ \vdots \\ \leftarrow q_n^T \rightarrow \end{bmatrix} g$$

$$\Rightarrow \underline{p}(\lambda) = -\sum_{j=1}^{n} \underbrace{\left(\frac{q_j^T g}{\lambda_j + \lambda}\right) q_j}_{} \quad \text{— } (\Delta)$$

from the orthogonality of the eigenvectors. (Note for eg:

$$B = \sum_{j=1}^{n} \lambda_j q_j q^T, \ldots)$$

The key in (Δ) is to note that when $\lambda \to -\lambda_1$ (as we want), we have:

$$\underline{p}(\lambda) \approx \frac{q_1^T g}{(\lambda_1 + \lambda)} q_1 \quad \begin{pmatrix}\text{small} \\ \text{terms} \\ \text{dropped}\end{pmatrix}$$

Following this discussion, we eventually come to algorithm 4.4 for computing λ using Newton's algorithm.

4.3 Global Convergence

For convergence, we note the following model reduction by the Cauchy point (Lemma 4.5):

$$m_k(0) - m_k(p_k^c) \geq \frac{1}{2}\|\nabla f_k\| \min\left(\Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|}\right) \quad \Delta$$

For convergence to stationary points, recall the following step from algorithm 4.1:

$$\begin{cases} \underline{\text{if}} \quad \rho_k > \eta & \leftarrow \text{threshold reduction} \\ \quad x_{k+1} = x_k + p_k; & \leftarrow \text{accept} \\ \underline{\text{else}} & \leftarrow \text{reject} \\ \quad x_{k+1} = x_k; & \end{cases}$$

Thm 4.7

For $\eta = 0$, $\|B_k\|$ bounded, the level set $\{x \mid f(x) \leq f(x_0)\}$ is bounded, the reduction in Δ (above) is attained, and $\|p_k\| \leq \gamma \Delta_k$, some $\gamma \geq 1$. Then:

$$\liminf_{k\to\infty} \|\nabla f_k\| = 0. \quad \blacksquare$$

Note that the algorithm is 4.1, where the following line: 4II 8/9

{Obtain p_k by minimizing $m_k(p)$

has been replaced by the Cauchy-point method, or better.

For $\underline{\eta > 0}$, theorem 4.8 says that $\lim_{k \to \infty} \nabla f_k = 0$, which is convergence to a stationary point.

*A similar result follows for nearly exact solutions.

* <u>Scaling issues</u>:

$\|Dp\| \leq \Delta$ for some diagonal matrix D can be used for ellipsoidal regions.

Algorithm 1.5 gives the Cauchy point for such ellipsoidal regions.

For <u>non-Euclidean</u> regions, we need to wait till Quadratic Programming