THE UNIVERSITY OF CHICAGO

Section 4: Trust Region Methods

Mihai Anitescu STAT 310

Reference:

Chapter 4 in Nocedal and Wright.

More and Sorensen paper.

4.1 Trust Region Fundamentals

- Notations

$$f^k = f(x^k), \quad \nabla f^k = \nabla f(x^k)$$

- Quadratic Model

$$m_k(p) = f^k + p^T g^k + \frac{1}{2} p^T B^k p$$

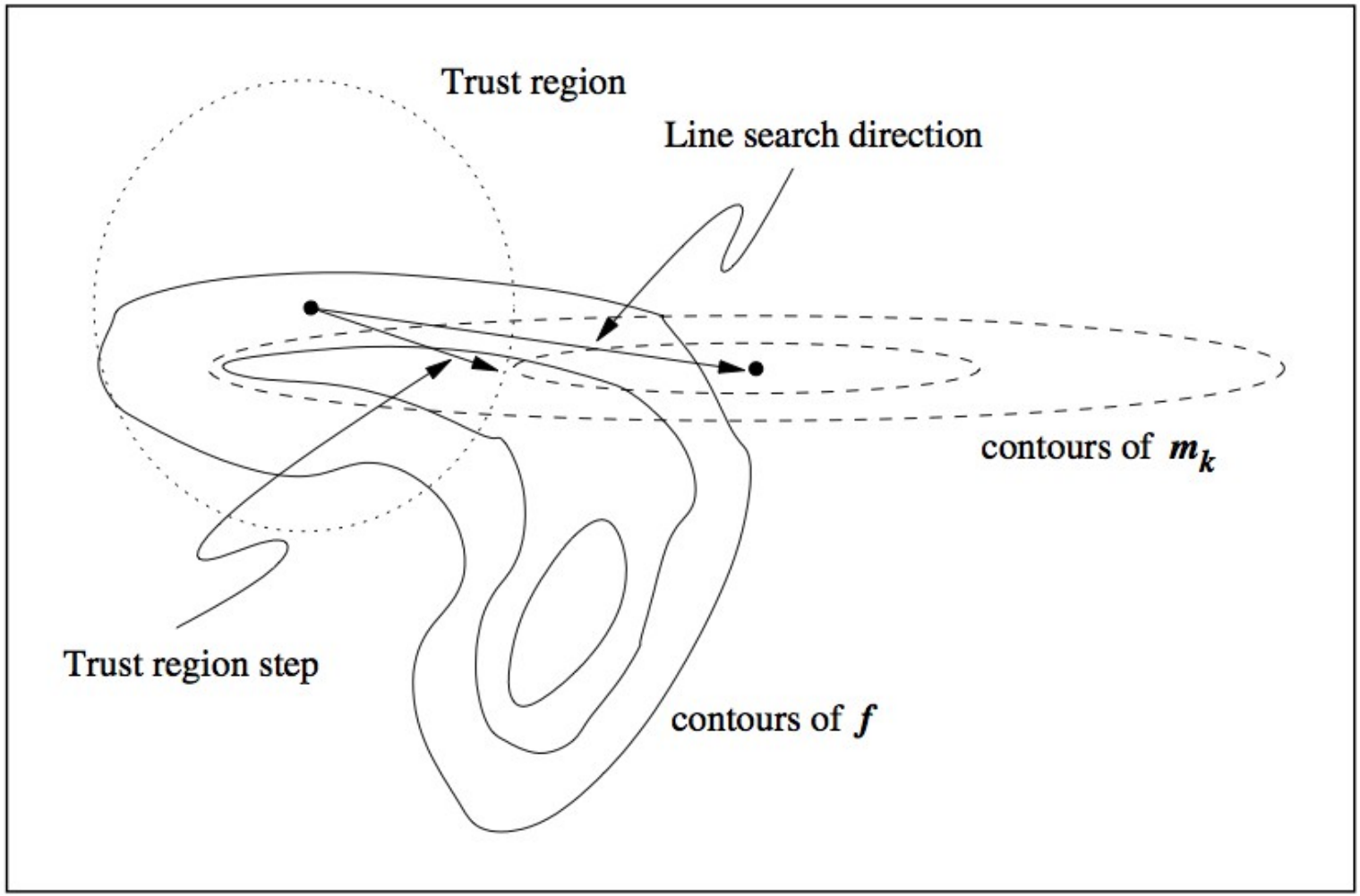
- Interesting choices: B^k is the Hessian, or 0.
- The TR problem.

$$\min_{p \in R^n} \quad m_k(p)$$

$$\text{subject to} \quad \|p\| \le \Delta^k$$

- Solve inexactly (w/o care of eigenvalues) -- Cauchy Points (dogleg or truncated CG).
- Solve exactly -- need to ”take care” of eigenvalues.

Trust Region Geometry



4.2 Solving the TR Subproblem (exactly)

Theorem 4.1.

The vector p^* is a global solution of the trust-region problem

$$\min_{p \in \mathbb{R}^n} m(p) = f + g^T p + \tfrac{1}{2} p^T B p, \quad \text{s.t. } \|p\| \leq \Delta, \tag{4.7}$$

if and only if p^* is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied:

$$(B + \lambda I) p^* = -g, \tag{4.8a}$$

$$\lambda(\Delta - ||p^*||) = 0, \tag{4.8b}$$

$$(B + \lambda I) \quad \text{is positive semidefinite.} \tag{4.8c}$$

Lemma 4.7.

Let m be the quadratic function defined by

$$m(p) = g^T p + \tfrac{1}{2} p^T B p, \tag{4.46}$$

where B is any symmetric matrix. Then the following statements are true.

(i) *m attains a minimum if and only if B is positive semidefinite and g is in the range of B. If B is positive semidefinite, then every p satisfying $Bp = -g$ is a global minimizer of m.*

(ii) *m has a unique minimizer if and only if B is positive definite.*

$$\min f(x) \quad \overset{\text{maybe}}{\iff} \quad \min f(x) + \lambda g(x)$$
$$\text{s.t } g(x) \le 0 \qquad\qquad \text{for some } \lambda$$

$$\nabla f(x) + \lambda \cdot \nabla g(x) = 0$$

$$\left.\begin{array}{l} g(x^*) < 0 \implies \lambda = 0 \\ \lambda > 0 \implies g(x^*) = 0 \end{array}\right\} \implies \lambda g(x^*) = 0$$

compl. slackness

Subproblem Solve

- Equation: $p(\lambda) = -(B + \lambda I)^{-1} g \qquad \|p(\lambda)\| = \Delta.$
- Now: $B = Q \Lambda Q^T \qquad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$

$$p(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^T g = -\sum_{j=1}^{n} \frac{q_j^T g}{\lambda_j + \lambda} q_j,$$

$$\|p(\lambda)\|^2 = \sum_{j=1}^{n} \frac{\left(q_j^T g\right)^2}{(\lambda_j + \lambda)^2}.$$

- Note that: $\lim_{\lambda \to \infty} \|p(\lambda)\| = 0.$
- Good case:

$$q_j^T g \neq 0 \Rightarrow \lim_{\lambda \to -\lambda_j} \|p(\lambda)\| = \infty.$$

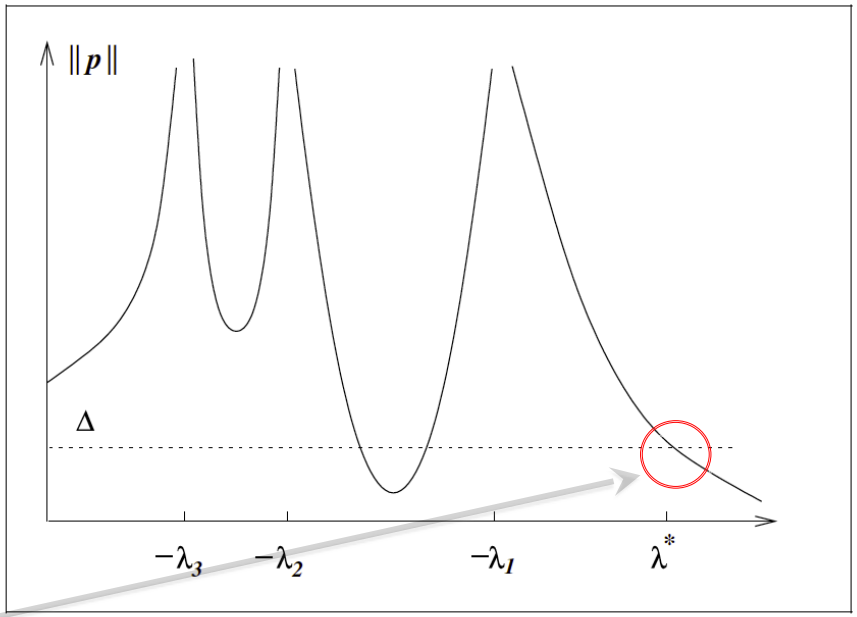


Figure 4.5 $\|p(\lambda)\|$ as a function of λ.

- There must be a solution!

The Hard Case

$$q_1^T g = 0 \quad \text{and} \quad \|p(\lambda)\| < \Delta$$

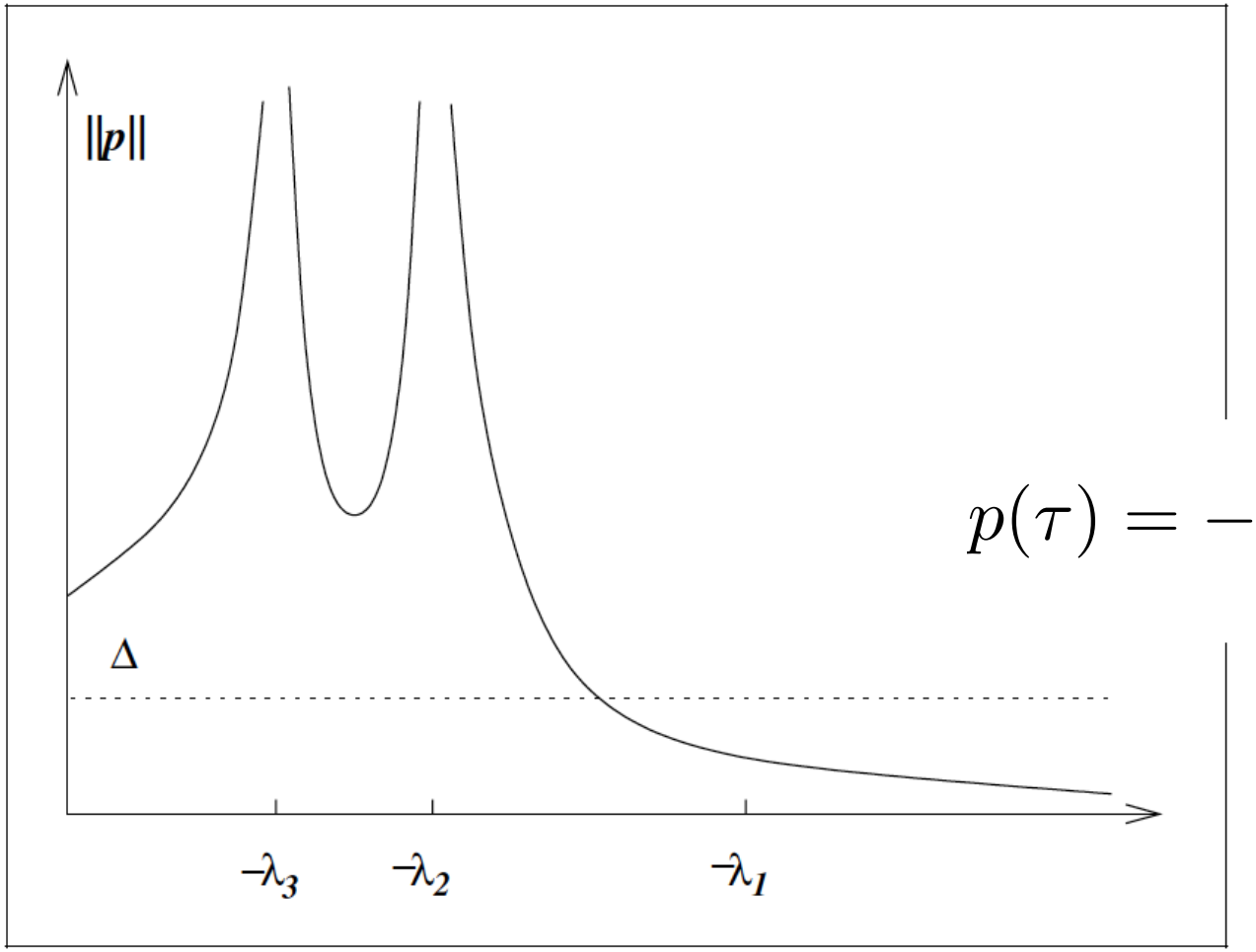


Figure 4.7 The hard case: $\|p(\lambda)\| < \Delta$ for all $\lambda \in (-\lambda_1, \infty)$.

From theory we MUST have

$$\lambda \geq -\lambda_1 \Rightarrow \lambda = -\lambda_1$$

By inspection, we have a family of solutions to

$$(B_k - \lambda_1 \mathbb{I}_n)\, p(\tau) = -g$$

$$p(\tau) = -\sum_{j:\lambda_j \neq \lambda_1} \frac{q_j^T g}{\lambda_j - \lambda_1} q_j + \tau q_1$$

One of the choices MUST satisfy The trust-region constraints.

$$\exists \tau \quad \|p(\tau)\| = \Delta^k$$

If double root, things continue to be complicated …

Practical (INCOMPLETE) algorithm

$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|p(\lambda)\|},$$

Newton's Method:

$$\lambda^{(\ell+1)} = \lambda^{(\ell)} - \frac{\phi_2\left(\lambda^{(\ell)}\right)}{\phi_2'\left(\lambda^{(\ell)}\right)}.$$

Algorithm 4.3 (Trust Region Subproblem).
Given $\lambda^{(0)}$, $\Delta > 0$:
for $\ell = 0, 1, 2, \ldots$
Factor $B + \lambda^{(\ell)} I = R^T R$;
Solve $R^T R p_\ell = -g$, $R^T q_\ell = p_\ell$;

Set

$$\lambda^{(\ell+1)} = \lambda^{(\ell)} + \left(\frac{\|p_\ell\|}{\|q_\ell\|}\right)^2 \left(\frac{\|p_\ell\| - \Delta}{\Delta}\right);$$

end (for).

- It uses that for lambda large enough lambda the function is convex.
- Need to start with large enough values of lambda (that is the non-elegant part) (see More Sorensen)
- It generally gives a machine precision solution in 2-3 iterations (Cholesky)

Summary: trust region method

Outer Loop

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } \|p\| \le \Delta_k$$

Algorithm 4.1 (Trust Region).

Given $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \in [0, \frac{1}{4})$:

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

  Obtain p_k by ( ) solving (4.3);

  Evaluate ρ_k from (4.4);

  if $\rho_k < \frac{1}{4}$

    $\Delta_{k+1} = \frac{1}{4}\Delta_k$

  else

    if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$

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

    else

      $\Delta_{k+1} = \Delta_k$;

  if $\rho_k > \eta$

    $x_{k+1} = x_k + p_k$

  else

    $x_{k+1} = x_k$;

end (for).

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

Discussion

- The efficiency of this algorithm revolves around the Cholesky factorization being much more efficient that eigenvalue decomposition.
- In dense linear algebra, this is true with a factor of about 6. However, if you need more than 6 Cholesky you loose.
- In sparse linear algebra, the eigenvalue computation is relatively speaking much more expensive than Cholesky. So eigenvalue calculation directly is hopeless on large problems. You can afford hundreds of Cholesky's compared to eigenvalue calculation.
- Example the cute problem, n=1000, and n=3000
- On the other hand, for data-driven problems eigenvalue

```
>> tic;chol(H);toc
Elapsed time is 0.000437 seconds.
>> tic;eig(H);toc
Elapsed time is 1.853798 seconds.
```

```
>> tic;chol(H);toc
Elapsed time is 0.001033 seconds.
>> tic;eig(H);toc
Elapsed time is 110.560001 seconds.
```

4.3 "InExact" Trust Region algorithms

- Based upon solving exactly a possibly nonconvex model of the function.
- Essential Element: The Cauchy Point as Yardstick (sufficient descent ..)

Algorithm 4.2 (Cauchy Point Calculation).

Find the vector p_k^s that solves a linear version of (4.3), that is,

$$p_k^s = \arg\min_{p \in \mathrm{R}^n} \; f_k + g_k^T p \qquad \text{s.t.} \;\; \|p\| \leq \Delta_k;$$

Calculate the scalar $\tau_k > 0$ that minimizes $m_k(\tau p_k^s)$ subject to satisfying the trust-region bound, that is,

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

Set $p_k^c = \tau_k p_k^s$.

Dogleg Methods: Improve CP

- If Cauchy point is on the boundary I have a lot of decrease and I accept it (e.g if $g^{k,T} B_k g^k \leq 0;$)
- If Cauchy point is interior, $g^{k,T} B_k g^k > 0; \quad p^{k,c} = -\frac{\|g_k\|^2}{g^{k,T} B_k g^k} g^k$
- Take now "Newton" step $p_B = -B_k^{-1} g^k$ (note, B need not be pd, all I need is nonsingular).
- Define dogleg path: $\tilde{p}(\tau) = \begin{cases} \tau p^{k,c} & \tau \leq 1 \\ p^{k,c} + (\tau - 1)\left(p^B - p^{k,c}\right) & 1 \leq \tau \leq 2 \end{cases}$
- Find dogleg point solution: $\tilde{p}(\tau_D); \quad \tau_D = \arg\min_{\tau; \|\tilde{p}(\tau)\| \leq \Delta_k} m_k\left(\tilde{p}(\tau)\right)$
- Need to minimize one bounded quadratic (second segment)
- Has at least Cauchy point decrease, allows Newton point direction when close to the solution.

The key to convergence for Cauchy Point

- Algorithmic requirement : improvement in the MODEL is bounded below by some coercive function of the gradient.

$$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \; c_1 \in (0, 1] \quad (4.20)$$

- Includes dogleg (proof):

Lemma 4.3.

The Cauchy point p_k^c satisfies (4.20) with $c_1 = \frac{1}{2}$, that is,

$$m_k(0) - m_k(p_k^c) \geq \tfrac{1}{2} \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right).$$

- Sufficient requirement to satisfy above becomes (see also Theorem 4.4)

$$m_k(0) - m_k(p_k) \geq c_2 \left(m_k(0) - m_k(p_k^c)\right)$$

Summary: trust region method

Outer Loop

Algorithm 4.1 (Trust Region).
Given $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \in [0, \frac{1}{4})$:
for $k = 0, 1, 2, \ldots$
  Obtain p_k by (approximately) solving (4.3);
  Evaluate ρ_k from (4.4);
  if $\rho_k < \frac{1}{4}$
    $\Delta_{k+1} = \frac{1}{4}\Delta_k$
  else
    if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$
      $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta})$
    else
      $\Delta_{k+1} = \Delta_k$;
  if $\rho_k > \eta$
    $x_{k+1} = x_k + p_k$
  else
    $x_{k+1} = x_k$;
end (for).

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } \|p\| \le \Delta_k$$

When solving approximately, use Cauchy point as fallback to ensure Sufficient decrease in the model.

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

$$m_k(0) - m_k(p_k^c) \ge \tfrac{1}{2} \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right).$$

$$m_k(0) - m_k(p_k) \ge c_2 \left(m_k(0) - m_k(p_k^c)\right)$$

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

Global Convergence, Case 1: \eta>0

Theorem 4.6.

Let $\eta \in \left(0, \frac{1}{4}\right)$ in Algorithm 4.1. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is bounded below on the level set S (4.24) and Lipschitz continuously differentiable in $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions p_k of (4.3) satisfy the inequalities (4.20) and (4.25) for some positive constants c_1 and γ. We then have

$$\lim_{k \to \infty} g_k = 0. \tag{4.33}$$

$$S \stackrel{\text{def}}{=} \{x \mid f(x) \leq f(x_0)\}. \tag{4.24}$$

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } \|p\| \leq \Delta_k, \tag{4.3}$$

$$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \tag{4.20}$$

$$\|p_k\| \leq \gamma \Delta_k, \qquad \text{for some constant } \gamma \geq 1. \tag{4.25}$$

- Bounded level set
- Trust Region subproblem
- Fraction of Cauchy Point decrease
- Approximate solution to the TR problem

Global Convergence, Case 1: \eta=0

Theorem 4.5.

Let $\eta = 0$ in Algorithm 4.1. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is bounded below on the level set S defined by (4.24) and Lipschitz continuously differentiable in the neighborhood $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions of (4.3) satisfy the inequalities (4.20) and (4.25), for some positive constants c_1 and γ. We then have

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{4.26}$$

Proof

$$S \stackrel{\text{def}}{=} \{x \mid f(x) \leq f(x_0)\}. \tag{4.24}$$

$$\min_{p\in\mathbb{R}^n} m_k(p) = f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } \|p\| \leq \Delta_k, \tag{4.3}$$

$$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \tag{4.20}$$

$$\|p_k\| \leq \gamma \Delta_k, \qquad \text{for some constant } \gamma \geq 1. \tag{4.25}$$

- Bounded level set
- Trust Region subproblem
- Fraction of Cauchy Point decrease
- Approximate solution to the TR problem

Theorem 4.6 , map of the proof.

- Step1 Around any iterate, find a neighborhood small enough, that the gradient is bounded below on it.
- Step 2 Prove that the algorithm must leave this neighborhood eventually (it NEEDS to, since if gradient keeps being large you will keep reducing the function value)
- Step 3 Prove that by the time it leaves this neighborhood it would have produce a decrease in the objective function that is proportional to either the norm of the gradient or the square of it.
- Step 4 Since the decreases add up and the total decrease is bounded below, this forces the gradient to converge to 0, which gives the proof.

4.4 Second order convergence

Global convergence away from saddle point

Theorem 4.8.

Suppose that the assumptions of Theorem 4.6 are satisfied and in addition that f is twice continuously differentiable in the level set S. Suppose that $B_k = \nabla^2 f(x_k)$ for all k, and that the approximate solution p_k of (4.3) at each iteration satisfies (4.52) for some fixed $\gamma > 0$. Then $\lim_{k\to\infty} \|g_k\| = 0$.

4.52: An ALMOST EXACT version of app. solution

If, in addition, the level set S of (4.24) is compact, then either the algorithm terminates at a point x_k at which the second-order necessary conditions (Theorem 2.3) for a local solution hold, or else $\{x_k\}$ has a limit point x^* in S at which the second-order necessary conditions hold.

Fast Local Convergence

Theorem 4.9.

Let f be twice Lipschitz continuously differentiable in a neighborhhod of a point x^* at which second-order sufficient conditions (Theorem 2.4) are satisfied. Suppose the sequence $\{x_k\}$ converges to x^* and that for all k sufficiently large, the trust-region algorithm based on (4.3) with $B_k = \nabla^2 f(x_k)$ chooses steps p_k that satisfy the Cauchy-point-based model reduction criterion (4.20) and are asymptotically similar to Newton steps p_k^N whenever $\|p_k^N\| \leq \frac{1}{2}\Delta_k$, that is,

$$\|p_k - p_k^N\| = o(\|p_k^N\|). \tag{4.53}$$

Then the trust-region bound Δ_k becomes inactive for all k sufficiently large and the sequence $\{x_k\}$ converges superlinearly to x^*.

Theorem 4.9 Map of the Proof.

- Step 0: Discussion of the "almost Newton" compatibility condition.
- Step 1: The gradient is an upper bound of a fixed factor of the iteration step. $\|g_k\| \geq \frac{1}{2}\|p_k\| / \left\|\nabla^2 f(x_k)^{-1}\right\|$
- Step 2: Lower bound on model decrease. $m_k(0) - m_k(p_k) \geq c_3\|p_k\|^2$
- Step 3: Upper bound on model discrepancy
- Step 4: Bound on the Actual Reduction/Predicted Reduction ratio (rho) $|\rho_k - 1| \leq \frac{L\Delta_k}{c_3}$

Conclude that the trust region is bounded below.

- Step 5: Eventually "almost Newton" steps are accepted, and prove superlinear convergence. Discussion

4.5 Variations and enhancements

- Badly scaled problems -- elliptical trust region:

$$\min_{p \in \mathrm{R}^n} m_k(p) \stackrel{\text{def}}{=} f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } \|Dp\| \le \Delta_k.$$

- Use other norms

$$\|p\|_1 \le \Delta_k \qquad \text{or} \qquad \|p\|_\infty \le \Delta_k,$$

$$\|Dp\|_1 \le \Delta_k \qquad \text{or} \qquad \|Dp\|_\infty \le \Delta_k,$$

- Particularly in bound constrained optimization:

$$\min_{x \in \mathrm{R}^n} f(x), \qquad \text{subject to } x \ge 0,$$

$$\min_{p \in \mathrm{R}^n} m_k(p) = f_k + g_k^T p + \tfrac{1}{2} p^T B_k p \qquad \text{s.t. } x_k + p \ge 0, \|p\| \le \Delta_k.$$

- If is not PD very hard problem (NP-hard).
- For =0, however, it is easier than elliptical!

Summary and Comparisons

- Line search problems have easier subproblems (if we modify Cholesky).
- But cannot guaranteed convergence to second order points.
- Trust-region problems can
- Empirically, they need about 2-3 Cholesky factorizations per step
- Both exhibit superlinear convergence
- Dogleg methods live “between” these two situations.