CS Space

Unconstrained Nonlinear Optimization
Lecture 5: Trust-region methods 1

Feodor Pisnitchenko

CMCC/UFABC

Momoe Sakamori

IMECC/UNICAMP

Summary

- Outline of the Trust-Region Strategy
- Algorithms Based on the Cauchy Point
- The Dogleg Method
- Global Convergence

Outline of Line Search Methods

- Choose an initial guess x_0 for the solution.
- Compute a descent direction p_k:
  - Steepest descent: $p_k = -\nabla f(x_k)$
  - Newton's method: $p_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$
  - Quasi-Newton methods (e.g., BFGS)
- Obtain a step size α_k along p_k satisfying:
  - Sufficient decrease (Armijo condition)
  - Curvature condition (Wolfe or strong Wolfe conditions)
- Update the iterate:

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

Outline of Trust-Region Methods: Key Concepts

- Define a trust region where the model is trusted to be a good approximation.
- Compute the step by minimizing the model within this region.
- Direction and step length are determined simultaneously.
- If a step is not acceptable, reduce the trust-region size and recompute the step.
- The step direction typically changes when the trust-region size is adjusted.

Trust-Region Methods: Radius Adjustment

- The **trust-region size** is **critical** for effectiveness:
  - **Too small**: Limits progress by missing significant steps.
  - **Too large**: Steps may stray far from the minimizer, making the model unreliable.
- The **trust-region radius** adjusts dynamically:
  - **Increase** if the step is successful and well-predicted by the model.
  - **Decrease** if the step fails (poor prediction or insufficient reduction).
- A failed step indicates the model inaccurately represents the objective function in the current region.

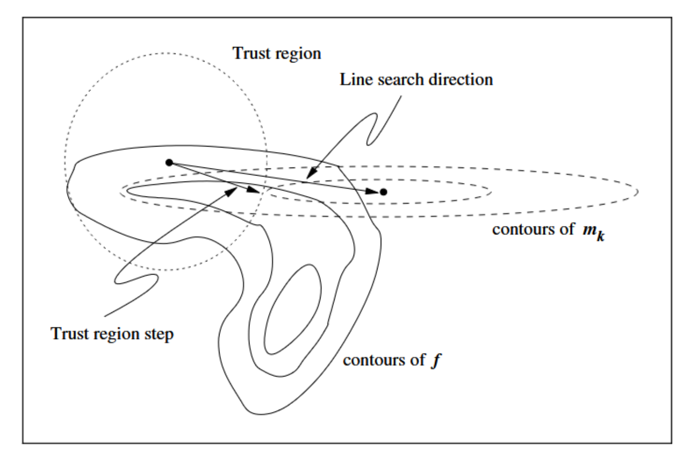

Figure: Trust-region and line search steps.

Trust-Region Approach: Taylor Expansion

- Using a second-order Taylor expansion of f around x_k:

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

  where $f_k = f(x_k)$, $g_k = \nabla f(x_k)$, and $t \in (0, 1)$.

- Replace the true Hessian with a symmetric matrix B_k to form the **quadratic model**:

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

- The error $m_k(p) - f(x_k + p) = \mathcal{O}(\|p\|^2)$, making the model a good local approximation near x_k.

Trust-Region Approach: Model Accuracy

- If $B_k = \nabla^2 f(x_k)$, the error in $m_k(p)$ is $\mathcal{O}(\|p\|^3)$, enhancing accuracy for small steps.
- This leads to the **Trust-Region Newton Method**.
- Generally, B_k is assumed symmetric and uniformly bounded.

Trust-Region Subproblem

- Solve the subproblem:

$$\begin{aligned} \min_{p \in \mathbb{R}^n} \quad & m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \\ \text{s.t.} \quad & \|p\| \leq \Delta_k, \end{aligned} \tag{1}$$

  where $\Delta_k > 0$ is the trust-region radius.

- Using the Euclidean norm, the feasible region is a ball of radius Δ_k.
- The constraint $p^T p \leq \Delta_k^2$ and the quadratic objective make this a quadratic program.
- *Note*: Exact solutions can be computationally expensive for large n, motivating approximate methods.

Special Case: Full Step

- If B_k is positive definite and $\|B_k^{-1} g_k\| \leq \Delta_k$, the solution is the unconstrained minimizer:

$$p_k^B = -B_k^{-1} g_k,$$

called the **full step**.

Trust-Region Radius Update

- The trust-region radius Δ_k is chosen iteratively based on model performance.
- The choice depends on how well $m_k(p)$ approximates $f(x)$.
- A ratio compares the actual reduction to the predicted reduction.

The Ratio ρ

- Define the ratio:

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}, \tag{2}$$

where:

  - Numerator: **Actual reduction** in f.
  - Denominator: **Predicted reduction** by m_k.

Interpreting ρ_k

- Since p_k minimizes m_k and $p = 0$ is feasible, the predicted reduction $m_k(0) - m_k(p_k) \geq 0$.
- Interpretations:
  - $\rho_k < 0$: $f(x_k + p_k)$ increased; reject the step and shrink the trust region.
  - $\rho_k \approx 1$: Excellent model agreement; accept the step and expand the trust region.
  - $0 < \rho_k \ll 1$: Moderate agreement; accept the step, keep the trust region unchanged.
  - $\rho_k \approx 0$: Poor agreement; reject the step and reduce the trust region.

Algorithm 6: Trust-Region (Part 1)

- Given $\Delta_{\max} > 0$, $\Delta_0 \in (0, \Delta_{\max})$, $\eta \in [0, \frac{1}{4})$:
- For $k = 0, 1, 2, \ldots$:
  - Obtain p_k by (approximately) solving (1).
  - Compute ρ_k using (2).
  - Update Δ_{k+1}:
    - If $\rho_k < \frac{1}{4}$, then $\Delta_{k+1} = \frac{1}{4}\Delta_k$.
    - Else, if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$, then $\Delta_{k+1} = \min(2\Delta_k, \Delta_{\max})$.
    - Else, $\Delta_{k+1} = \Delta_k$.
  - Update x_{k+1}:
    - If $\rho_k > \eta$, then $x_{k+1} = x_k + p_k$.
    - Else, $x_{k+1} = x_k$.

Algorithm 6: Trust-Region (Part 2)

- Δ_{max} bounds step lengths.
- Dynamic radius adjustment:
  - **Increase**: Good model agreement and step hits the boundary.
  - **Unchanged**: Good agreement, step within the region (current Δ_k is adequate).
  - **Decrease**: Poor model agreement.

Trust-Region Theorem: Problem Restatement

- Drop subscript k and restate (1):

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

- The solution p^* satisfies:

$$(B + \lambda I) p^* = -g,$$

for some $\lambda \geq 0$ (Moré and Sorensen).

Trust-Region Theorem

Theorem

p^* is a global solution to:

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

if and only if p^* is feasible and there exists $\lambda \geq 0$ such that:

$$\begin{array}{rc} \text{(i)} & (B + \lambda I) p^* = -g, \\ \text{(ii)} & \lambda(\Delta - \|p^*\|) = 0, \\ \text{(iii)} & (B + \lambda I) \text{ is positive semidefinite.} \end{array}$$

Trust-Region Theorem: Explanation

- (ii) is a complementarity condition: Either $\lambda = 0$ or $\|p^*\| = \Delta$.
- If $\|p^*\| < \Delta$ (Δ_1 in Figure), then $\lambda = 0$, so $Bp^* = -g$, and B is positive semidefinite.
- If $\|p^*\| = \Delta$ (Δ_2, Δ_3 in Figure), $\lambda \geq 0$, and:

$$\lambda p^* = -Bp^* - g = -\nabla m(p^*),$$

  aligning p^* with $-\nabla m(p^*)$, normal to the contours.

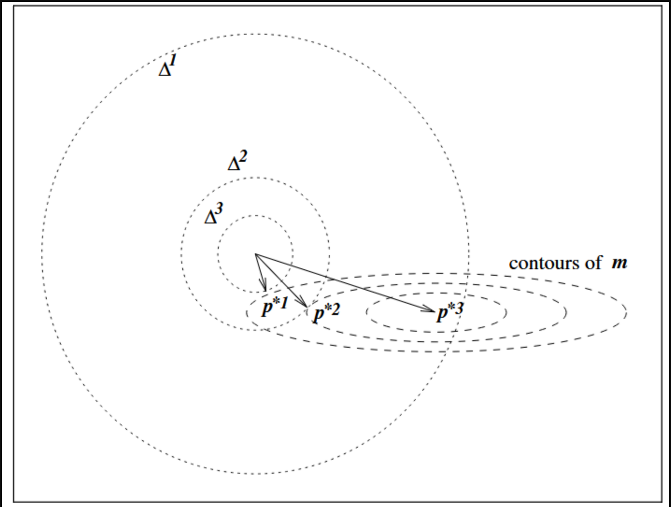

Figure: Solution for the trust-region subproblem for different radii $\Delta^1, \Delta^2, \Delta^3$.

Strategies to Solve the Subproblem

- Approximate solutions to (1) should reduce m_k at least as much as the Cauchy point.
- **Cauchy Point**: Minimizes m_k along $-g_k$ within the trust region.
- Strategies:
  - **Dogleg Method**: Effective for positive definite B_k.
  - Two-Dimensional Subspace Minimization: Handles indefinite B_k (needs the most negative eigenvalue).
  - Conjugate Gradient: Efficient for large, sparse Hessians.
  - Iterative Methods.

Algorithms Based on the Cauchy Point

- Line search achieves global convergence with inexact steps if they meet criteria (e.g., Wolfe conditions).
- In trust-region methods, it suffices to find p_k within the trust region achieving sufficient reduction in m_k.
- The sufficient reduction can be quantified in terms of the Cauchy point p_k^C.

Cauchy Point Calculation: Overview

- Steps:
  - Find p_k^s solving the linear problem:

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

  - Compute $\tau_k > 0$ minimizing $m_k(\tau p_k^s)$ with $\|\tau p_k^s\| \leq \Delta_k$.
  - Set $p_k^C = \tau_k p_k^s$.

Cauchy Point: Finding p_k^s

- Solve:

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

- Solution:

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

Cauchy Point: Finding τ_k

- Recall:

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

- Minimize:

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

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

$$m_k(\tau p_k^s) = f_k - \Delta_k \|g_k\| \tau + \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} (g_k^T B_k g_k) \tau^2.$$

- Constraint: $\tau \leq 1$.

Cauchy Point: Computing τ_k

- Minimize:

$$m_k(\tau p_k^s) = f_k - \Delta_k \|g_k\| \tau + \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} (g_k^T B_k g_k) \tau^2,$$

  subject to $\tau \leq 1$.

- Cases:
  - If $g_k^T B_k g_k \leq 0$, $m_k(\tau p_k^s)$ decreases with τ, so $\tau_k = 1$.
  - If $g_k^T B_k g_k > 0$, $m_k(\tau p_k^s)$ is convex; $\tau_k = \min\left(\frac{\|g_k\|^3}{\Delta_k g_k^T B_k g_k}, 1\right)$.

Cauchy Point Summary

- The Cauchy point is:

$$p_k^C = -\tau_k \frac{\Delta_k}{\|g_k\|} g_k,$$

where:

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

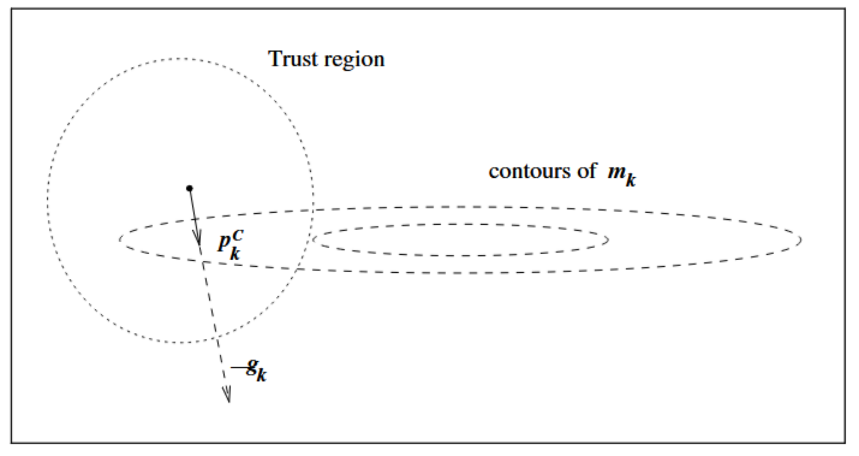

Figure: The Cauchy point for a subproblem with positive definite B_k. Here, p_k^C lies strictly inside the trust region.

Role of the Cauchy Point

- **Inexpensive**: Requires no matrix factorizations.
- **Crucial**: Evaluates the **acceptability of approximate solutions** to the trust-region subproblem.
- Ensures **global convergence** if p_k reduces m_k by a fixed fraction of the Cauchy step's reduction.

Limitations and Improvements

- Relying solely on the Cauchy point mimics steepest descent (poor performance).
- Direction is fixed at $-g_k$; B_k only affects step length.
- Faster convergence needs B_k to influence direction and curvature.
- Improvement: Use $p_k^B = -B_k^{-1} g_k$ when B_k is positive definite and $\|p_k^B\| \leq \Delta_k$.
- With true Hessian or quasi-Newton B_k, this yields superlinear convergence.

Dogleg Method: Construction

- Consider:

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

- When B is positive definite, the unconstrained minimizer of m is given by $p^B = -B^{-1}g$. When this point is feasible ($\|p^B\| \leq \Delta$), it is a obviously a solution, then we have

$$p^*(\Delta) = p^B = -B^{-1}g, \text{ when } \Delta \geq \|p^B\|.$$

  Note that we denote the solution by $p^*(\Delta)$ to emphasize the dependence on Δ

- When Δ is small relative to p^B, the restriction $\|p\| \leq \Delta$ ensures that the quadratic term in m has little effect on the solution of the subproblem.

Dogleg Method: Construction

- For such Δ, we omitt the quadratic term from the subproblem ($m(p) \approx f + g^T p$) and obtain an approximate solution
$$p^*(\Delta) \approx -\Delta \frac{g}{\|g\|}, \text{ when } \Delta \text{ is small.}$$
- For intermediate values of Δ, the solution $p^*(\Delta)$ typically follows a curved trajectory like in the following Figure.

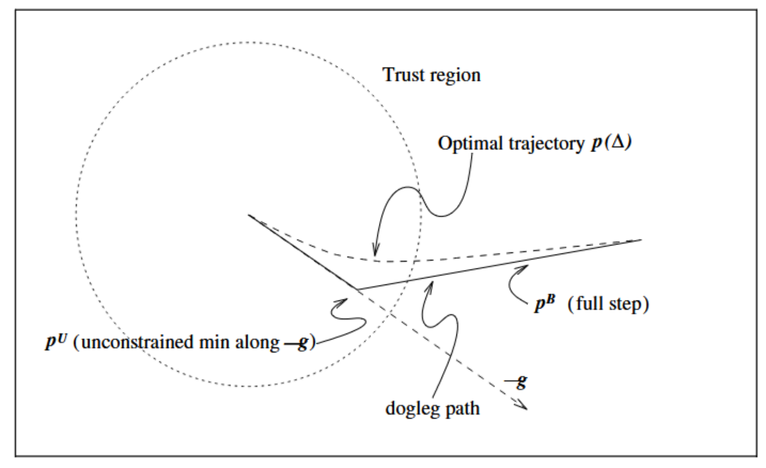

Figure: Exact trajectory and dogleg approximation.

Dogleg Path

- Approximates $p^*(\Delta)$ with two line segments.
- First segment: Origin to the steepest descent minimizer:

$$p^U = -\frac{g^T g}{g^T B g} g.$$

- Second segment: p^U to $p^B = -B^{-1} g$.

Dogleg Trajectory

- Define $\tilde{p}(\tau)$ for $\tau \in [0, 2]$:

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

- Chooses p minimizing m along this path within the trust region.
- *Note*: Balances steepest descent and Newton step efficiency.

Dogleg Lemma

Lemma

If B is positive definite:

(i) $\|\tilde{p}(\tau)\|$ increases with τ.

(ii) $m(\tilde{p}(\tau))$ decreases with τ.

- If $\|p^B\| < \Delta$, $p = p^B$.
- If $\|p^B\| \geq \Delta$, solve $\|\tilde{p}(\tau)\| = \Delta$.

Using the Exact Hessian

- If $\nabla^2 f(x_k)$ is positive definite, set $B = \nabla^2 f(x_k)$, so:

$$p^B = -(\nabla^2 f(x_k))^{-1} g_k.$$

- Otherwise, use a positive definite B.
- Near a solution, p^B becomes the Newton step, enabling rapid convergence.

Dogleg Decrease

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

The Cauchy point satisfies this with $c_1 = \frac{1}{2}$.

Cauchy Point Reduction

Lemma

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

Decrease Relative to Cauchy Point

Theorem

If $\|p_k\| \leq \Delta_k$ and $m_k(0) - m_k(p_k) \geq c(m_k(0) - m_k(p_k^C))$, then:

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

For the exact solution p^*, $c = 1$.

Global Convergence of the Trust-Region Methods

- We have two types of global convergence results depending on the choice of the parameter η in the Algorithm 6.
- When $\eta = 0$:
  - The algorithm accepts a step if it reduces the objective function value.
  - Guarantees that the gradient sequence g_k has a limit point at zero.
- When $\eta > 0$:
  - A step is accepted only if the actual decrease in f is at least a small fraction of the predicted decrease.
  - Provides a stronger result: the gradient $g_k \to 0$.

Global Convergence for $\eta = 0$

Theorem

Let $\eta = 0$ in Algorithm 6. Suppose that $\|B_k\| \le \beta$ for some constant β, f is bounded below on the level set $S = \{x | f(x) \le f(x_0)\}$ and Lipschitz continuously differentiable in the neighborhood $S(R_0) = \{x | \|x - y\| < R_0 \text{ for some } y \in S\}$ for some $R_0 > 0$, and that all approximate solutions of the subproblem (1) satisfy

$$m_k(0) - m_k(p_k) \ge c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \text{ for some } c_1 \in (0, 1],$$

and

$$\|p_k\| \le \gamma \Delta_k, \text{ for some constant } \gamma \ge 1.$$

Then we have

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

Global Convergence: $\eta > 0$

Theorem

If $\eta \in (0, \frac{1}{4})$, under the same conditions, then $\lim_{k\to\infty} g_k = 0$.