02610
Optimization and Data Fitting
Week 3: Trust-Region Methods

Yiqiu Dong

DTU Compute
Technical University of Denmark

Numerical optimization algorithms

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n \to \mathbb{R},\ f\in\mathcal{C}^2(\mathbb{R}^n)$$

The iteration step in most algorithms that we will introduce in this course is essentially in the form of

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k.$$

- $\boldsymbol{x}_k$ is the iterate.
- $\boldsymbol{p}_k$ is the search direction.
- α_k is the step size or step length.
- **Goal:** Expect as $k \to +\infty$ we have $\boldsymbol{x}_k \to \boldsymbol{x}^*$.
- **Aim:** Ensure that the limit of $\{\boldsymbol{x}_k\}$ satisfies the necessary optimality conditions.

Line search vs. trust-region

- **Line search:**
  - considered as descent methods, i.e., $f_{k+1} < f_k$
  - search direction $\boldsymbol{p}_k$ first, step length α_k second
  - ensure that the search direction is a descent direction, i.e., $\boldsymbol{g}_k^T \boldsymbol{p}_k < 0$
  - the computation of α_k may itself require an iterative procedure
  - generic update is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$
- **Trust-region:**
  - only need $f_{k+1} \leq f_k$
  - step length first (trust-region radius Δ_k), search direction second ("solve" a subproblem for $\boldsymbol{p}_k$)
  - find $\boldsymbol{p}_k$ by minimizing an approximation of $f(\boldsymbol{x}_k + \boldsymbol{p})$
  - generic update is

$$\boldsymbol{x}_{k+1} = \begin{cases} \boldsymbol{x}_k + \boldsymbol{p}_k, & \text{if } f(\boldsymbol{x}_k + \boldsymbol{p}_k) \text{``} <\text{''} f_k \\ \boldsymbol{x}_k, & \text{otherwise} \end{cases}$$

Notations: $f_k = f(\boldsymbol{x}_k)$, $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$

Why need trust-region methods?

Review of the line search methods:

- Steepest descent methods: need only the gradient and has global convergence, but converges very slow.
- Newton and quasi-Newton: converge fast, but has only local convergence

Question: Can we improve Newton and quasi-Newton methods to obtain global convergence and still keep fast convergence rate?

One possible solution: Use trust-region strategy.

Newton and quasi-Newton

Newton and quasi-Newton at each iteration (approximately) solve the minimization problem

$$\min_{\boldsymbol{p}} m_k(\boldsymbol{p}) = f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p}.$$

- If B_k is positive definite, the minimizer is $\boldsymbol{p} = -B_k^{-1}\boldsymbol{g}_k$.
- If $B_k = \nabla^2 f(\boldsymbol{x}_k)$ and positive definite, we obtain the Newton iteration.
- If B_k satisfies quasi-Newton condition, we obtain the quasi-Newton iteration.
- **Problem:** The model $m_k(\boldsymbol{p})$ is an approximation of $f(\boldsymbol{x}_k + \boldsymbol{p})$, but this approximation is only valid in a small neighborhood of $\boldsymbol{x}_k$.
- **Solution:** We add a constraint on the minimization probelm:

$$\min_{\boldsymbol{p}} m_k(\boldsymbol{p}), \quad \text{subject to } \|\boldsymbol{p}\| \le \Delta_k,$$

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

Trust-region methods

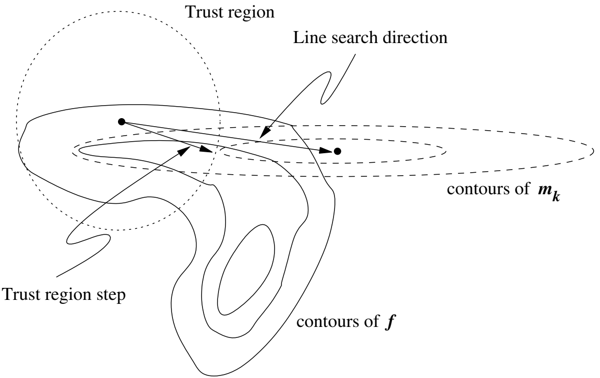
$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m_k(\boldsymbol{p}), \quad \text{s. t. } \|\boldsymbol{p}\| \le \Delta_k.$$

- m_k can be any "reasonable" approximation of $f(\boldsymbol{x}_k + \boldsymbol{p})$.
- For simplicity, we focus on the 2nd-order quadratic model

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

  and 2-norm on the trust region.
- Global convergence results do not depend on which norm to be used.
- If B_k is positive definite and $\|B_k^{-1}\boldsymbol{g}_k\|_2 \le \Delta_k$, we obtain $\boldsymbol{p}_k^B = -B_k^{-1}\boldsymbol{g}_k$, which is called the *full step*. In other cases, we need solve a constrained minimization problem.

Example



- A line search method with the minimizer of m_k as the search direction will yield at most a small reduction in f.
- The trust-region method yields a more significant reduction in f and better progress toward the solution.

Outline of the trust-region

Algorithm

Set $\Delta_0 > 0$ and $\boldsymbol{x}_0$.

loop

Obtain $\boldsymbol{p}_k$ by (approximately) solving

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m_k(\boldsymbol{p}) = f_k + \boldsymbol{g}_k^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\| \le \Delta_k \tag{1}$$

$\boldsymbol{x}_{new} = \boldsymbol{x}_k + \boldsymbol{p}_k$

Update Δ_k to get Δ_{k+1}

if $\boldsymbol{x}_{new}$ is acceptable **then**

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{new}$

else

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$

end if

end loop

How to update Δ_k

We update the trust-region radius Δ_k at each iteration based on the agreement between the model function m_k and the objective function f at previous iterations.

$$\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)}{m_k(\boldsymbol{0}) - m_k(\boldsymbol{p}_k)}$$

- The numerator, $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)$, is called the actual reduction.
- The denominator, $m_k(\boldsymbol{0}) - m_k(\boldsymbol{p}_k)$ is the predicted reduction.
- $m_k(\boldsymbol{0}) \geq m_k(\boldsymbol{p}_k)$
- If ρ_k is close to 1, there is good agreement between m_k and f, so it is safe to expand the trust region for the next iteration.
- If ρ_k is close to zero or negative, we shrink the trust region at the next iteration.

When accept the step?

$$\boldsymbol{x}_{new} = \boldsymbol{x}_k + \boldsymbol{p}_k$$

The point $\boldsymbol{x}_{new}$ can be accepted or rejected. The acceptance criterium can be

- Armijo condition:
$$f(\boldsymbol{x}_{new}) \le f_k + c_1 \boldsymbol{g}_k^T \boldsymbol{p}_k$$
- by using ρ_k: We accept $\boldsymbol{x}_{new}$ if $\rho_k > \eta$, in order to ensure sufficient decreasing in f.

Outline of the trust-region

Algorithm

Set $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, $\boldsymbol{x}_0$ and $\eta \in [0, \frac{1}{4})$.
loop
  Obtain $\boldsymbol{p}_k$ by (approximately) solving (1);
  $\boldsymbol{x}_{new} = \boldsymbol{x}_k + \boldsymbol{p}_k$;
  $\rho_k = (f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k))/(m_k(\boldsymbol{0}) - m_k(\boldsymbol{p}_k))$;

  if $\rho_k < \frac{1}{4}$ **then**
    $\Delta_{k+1} = \frac{1}{4}\Delta_k$;
  else
    if $\rho_k > \frac{3}{4}$ and $\|\boldsymbol{p}_k\|_2 = \Delta_k$ **then**
      $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta})$;
    else
      $\Delta_{k+1} = \Delta_k$;
    end if
  end if

  if $\rho_k > \eta$ **then**
    $\boldsymbol{x}_{k+1} = \boldsymbol{x}_{new}$;
  else
    $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$;
  end if
end loop

Cauchy point

Similar as line search methods, even $\boldsymbol{p}_k$ is not the the minimizer of (1), if $\boldsymbol{p}_k$ lies within the trust region and gives a *sufficient reduction* in the model, the global convergence still can be achieved. The sufficient reduction can be quantified in terms of the Cauchy point.

Cauchy point

Consider the univariate minimization problem subject to the trust-region bound:

$$\min_{\tau \geq 0} m_k(\tau) = f_k + \tau \boldsymbol{g}_k^T \boldsymbol{p}_k^s + \frac{\tau^2}{2} \boldsymbol{p}_k^{sT} B_k \boldsymbol{p}_k^s \quad \text{s. t. } \|\tau \boldsymbol{p}_k^s\|_2 \leq \Delta_k,$$

where

$$\boldsymbol{p}_k^s = \arg\min_{\boldsymbol{p}} f_k + \boldsymbol{g}_k^T \boldsymbol{p} \quad \text{s. t. } \|\boldsymbol{p}\|_2 \leq \Delta_k,$$

and τ_k denotes the minmizer. The point $\boldsymbol{p}_k^C = \tau_k \boldsymbol{p}_k^s$ is called Cauchy point (step).

Cauchy point

1. The minimizer

$$\boldsymbol{p}_k^s = \arg\min_{\boldsymbol{p}} f_k + \boldsymbol{g}_k^T \boldsymbol{p} \quad \text{s. t. } \|\boldsymbol{p}\|_2 \leq \Delta_k.$$

has a closed form, i.e.,

$$\boldsymbol{p}_k^s = -\frac{\Delta_k}{\|\boldsymbol{g}_k\|_2} \boldsymbol{g}_k$$

2. Consider

$$\min_{\tau \geq 0} m_k(\tau) = f_k + \tau \boldsymbol{g}_k^T \boldsymbol{p}_k^s + \frac{\tau^2}{2} \boldsymbol{p}_k^{sT} B_k \boldsymbol{p}_k^s \quad \text{s. t. } \|\tau \boldsymbol{p}_k^s\|_2 \leq \Delta_k.$$

According to the constraint $\|\tau \boldsymbol{p}_k^s\|_2 \leq \Delta_k$ and $\tau \geq 0$, we have $\tau \in [0, 1]$.

Cauchy point

1. The minimizer

$$\boldsymbol{p}_k^s = \arg\min_{\boldsymbol{p}} f_k + \boldsymbol{g}_k^T \boldsymbol{p} \quad \text{s. t. } \|\boldsymbol{p}\|_2 \leq \Delta_k.$$

has a closed form, i.e.,

$$\boldsymbol{p}_k^s = -\frac{\Delta_k}{\|\boldsymbol{g}_k\|_2} \boldsymbol{g}_k$$

2. Consider

$$\min_{\tau \in [0,1]} m_k(\tau) = f_k + \tau \boldsymbol{g}_k^T \boldsymbol{p}_k^s + \frac{\tau^2}{2} \boldsymbol{p}_k^{sT} B_k \boldsymbol{p}_k^s.$$

Its minimizer is

$$\tau_k = \begin{cases} 1, & \text{if } \boldsymbol{g}_k^T B_k \boldsymbol{g}_k \leq 0; \\ \min(\|\boldsymbol{g}_k\|_2^3 / (\Delta_k \boldsymbol{g}_k^T B_k \boldsymbol{g}_k), 1), & \text{otherwise.} \end{cases}$$

3. Cauchy point is $\boldsymbol{p}_k^C = \tau_k \boldsymbol{p}_k^s$.

Reduction obtained by the Cauchy point

Lemma

Consider

$$m_k(\boldsymbol{p}) = f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p},$$

then the Cauchy point $\boldsymbol{p}_k^C$ satisfies

$$m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k^C) \geq \frac{1}{2}\|\boldsymbol{g}_k\|_2 \min\left(\Delta_k, \frac{\|\boldsymbol{g}_k\|_2}{\|B_k\|_2}\right).$$

- Any reasonable step in a trust region method should be not worse than the Cauchy point with a factor $c_2 > 0$. So we require

$$\begin{aligned} m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k) &\geq c_2\left(m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k^C)\right) \\ &\geq \frac{c_2}{2}\|\boldsymbol{g}_k\|_2 \min\left(\Delta_k, \frac{\|\boldsymbol{g}_k\|_2}{\|B_k\|_2}\right). \end{aligned}$$

- For an accepted step, we have $\rho_k > \eta$, then the actual reduction satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k) = \rho_k(m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k)) > \frac{\eta c_2}{2}\|\boldsymbol{g}_k\|_2 \min\left(\Delta_k, \frac{\|\boldsymbol{g}_k\|_2}{\|B_k\|_2}\right).$$

Global convergence

Theorem (The case $\eta = 0$)

Let $\eta = 0$. Suppose that $\|B_k\|_2 \leq \beta$ for some constant β, that f is bounded below on the level set S and Lipschitz continuously differentiable in the neighborhood $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions of (1) satisfy the inequalities

$$m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k) \geq c_1 \|\boldsymbol{g}_k\|_2 \min\left(\Delta_k, \frac{\|\boldsymbol{g}_k\|_2}{\|B_k\|_2}\right)$$

$$\|\boldsymbol{p}_k\|_2 \leq \gamma \Delta_k$$

for some positive constants c_1 and $\gamma \geq 1$. We then have

$$\liminf_{k \to \infty} \|\boldsymbol{g}_k\|_2 = 0.$$

- The level set: $S = \{\boldsymbol{x} | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$.
- An open neighborhood of S: $S(R_0) = \{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{y}\|_2 < R_0 \text{ for some } \boldsymbol{y} \in S\}$.
- It shows the convergence of a subsequence of the gradients.

Global convergence

Theorem (The case $\eta > 0$)

Let $\eta \in (0, 1/4)$. Suppose that $\|B_k\|_2 \leq \beta$ for some constant β, that f is bounded below on the level set S and Lipschitz continuously differentiable in $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions of (1) satisfy the inequalities

$$m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k) \geq c_1 \|\boldsymbol{g}_k\|_2 \min\left(\Delta_k, \frac{\|\boldsymbol{g}_k\|_2}{\|B_k\|_2}\right)$$

$$\|\boldsymbol{p}_k\|_2 \leq \gamma \Delta_k$$

for some positive constants c_1 and $\gamma \geq 1$. We then have

$$\lim_{k\to\infty} \boldsymbol{g}_k = 0.$$

- It shows the convergence of the gradients.
- It only shows that the method converges to a stationary point.

Solving the subproblem

The remaining question: How to solve the subproblem (1), i.e.,

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m_k(\boldsymbol{p}) = f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\|_2 \le \Delta_k$$

Similar as line search,

- we can solve accurately by using like iterative methods; or
- we can just approximate the solution.

We need note that

- the global convergence results show that an accurate solution is not really necessary;
- the Cauchy point is inexpensive to calculate, but it's the same as implementing the steepest descent method with a particular choice of step length.

Characterization of the exact solution

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m(\boldsymbol{p}) = f + \boldsymbol{g}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B\boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\|_2 \leq \Delta \tag{2}$$

Theorem

The vector $\boldsymbol{p}^*$ is a global solution of the trust-region problem (2) if and only if $\boldsymbol{p}^*$ is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied:

$$\begin{aligned} (B+\lambda I)\boldsymbol{p}^* &= -\boldsymbol{g}, \\ \lambda(\Delta - \|\boldsymbol{p}^*\|_2) &= 0, \\ (B+\lambda I) \quad &\text{is positive semidefinite.} \end{aligned}$$

- If $\|\boldsymbol{p}^*\|_2 < \Delta$, then $\lambda^* = 0$ and $B\boldsymbol{p}^* = -\boldsymbol{g}$ with B positive semidefinite.
- If $\lambda^* > 0$, then $\|\boldsymbol{p}^*\|_2 = \Delta$, i.e, the global solution $\boldsymbol{p}^*$ reaches the trust-region boundary. Further, $\lambda^*\boldsymbol{p}^* = -B\boldsymbol{p}^* - \boldsymbol{g} = -\nabla m(\boldsymbol{p}^*)$, i.e., $\boldsymbol{p}^*$ is collinear with the negative gradient of m and normal to its contours.
- λ^* is actually a Lagrange multiplier of the constraint problem.

Exact solver for the subproblem

According to the previous theorem, we consider two cases:

- **First case:** $\lambda^* = 0$ and $\boldsymbol{p}^*$ satisfies $\|\boldsymbol{p}^*\|_2 < \Delta$ and $B\boldsymbol{p}^* = -\boldsymbol{g}$ with B positive semidefinite.
- **Second case:** λ sufficiently large that $B + \lambda I$ is positive definite. Then, we seek a value $\lambda > 0$ such that

$$\boldsymbol{p}(\lambda) = -(B + \lambda I)^{-1}\boldsymbol{g} \quad \text{and} \quad \|\boldsymbol{p}(\lambda)\|_2 = \Delta.$$

Consider the spectral decomposition of B, i.e., $B = Q\Lambda Q^T$ with an orthogonal matrix $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1 \le \cdots \le \lambda_n$ are the eigenvalues of B. Then, for $\lambda \neq -\lambda_j$ we have

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

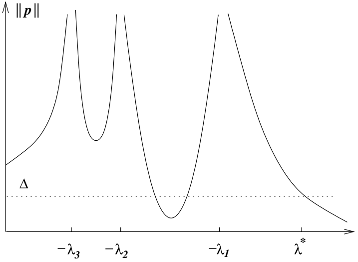
$$\|\boldsymbol{p}(\lambda)\|_2^2 = \sum_{j=1}^{n} \frac{\left(\boldsymbol{q}_j^T\boldsymbol{g}\right)^2}{(\lambda_j + \lambda)^2}$$

Exact solver for the subproblem

- **Second case:** We seek $\lambda > 0$ such that

$$\|\boldsymbol{p}(\lambda)\|_2^2 = \sum_{j=1}^{n} \frac{\left(\boldsymbol{q}_j^T \boldsymbol{g}\right)^2}{(\lambda_j + \lambda)^2} = \Delta^2$$

  - If $\lambda > -\lambda_1$, then $\lambda_j + \lambda > 0$ for all $j = 1, \ldots, n$. So $\|\boldsymbol{p}(\lambda)\|_2$ is a continuous, nonincreasing function of λ on the interval $(-\lambda_1, \infty)$.
  - **Easy case:** When $\boldsymbol{q}_1^T \boldsymbol{g} \neq 0$, then $\lim_{\lambda \to -\lambda_1} \|\boldsymbol{p}(\lambda)\|_2 = \infty$.

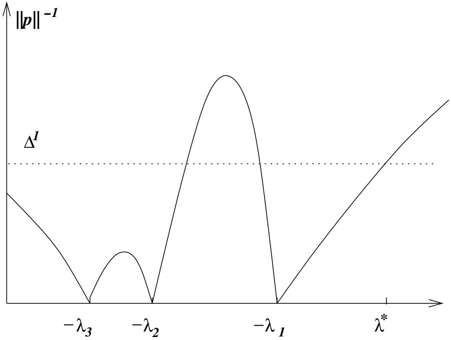


Exact solver for the subproblem

- **Second case:**
  - **Easy case:** When $\boldsymbol{q}_1^T \boldsymbol{g} \neq 0$, we only need find the value of $\lambda > -\lambda_1$ that solves
  $$\phi_1(\lambda) = \|\boldsymbol{p}(\lambda)\|_2 - \Delta = 0.$$
  But the value of $\|\boldsymbol{p}(\lambda)\|_2$ changes very fast, if λ is close to λ_1.

Exact solver for the subproblem

- **Second case:**
  - **Easy case:** When $\boldsymbol{q}_1^T \boldsymbol{g} \neq 0$, we only need find the value of $\lambda > -\lambda_1$ that solves
$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|\boldsymbol{p}(\lambda)\|_2} = 0.$$



It can be solved by Newton's method, i.e., $\lambda^{l+1} = \lambda^l - \frac{\phi_2(\lambda^l)}{\phi_2'(\lambda^l)}$.

Exact solver for the subproblem

- **Second case:**
  - **Easy case:** When $\boldsymbol{q}_1^T \boldsymbol{g} \neq 0$, we apply Newton's method to seek a root in $(-\lambda_1, \infty)$ of the function

$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|\boldsymbol{p}(\lambda)\|_2}.$$

Newton's method

Set $\Delta > 0$ and λ^0.
loop
Apply Cholesky factorization: $B + \lambda^l I = R^T R$;
Solve $R^T R \boldsymbol{p}_l = -\boldsymbol{g}$, $R^T \boldsymbol{q}_l = \boldsymbol{p}_l$;
Set

$$\lambda^{l+1} = \lambda^l + \left(\frac{\|\boldsymbol{p}_l\|_2}{\|\boldsymbol{q}_l\|_2}\right)^2 \left(\frac{\|\boldsymbol{p}_l\|_2 - \Delta}{\Delta}\right);$$

end loop

Exact solver for the subproblem

- **Second case:**
  - **Easy case:** When $\boldsymbol{q}_1^T \boldsymbol{g} \neq 0$, we apply Newton's method to seek a root in $(-\lambda_1, \infty)$ of the function

$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|\boldsymbol{p}(\lambda)\|_2}.$$

Newton's method

Set $\Delta > 0$ and λ^0.
loop
Apply Cholesky factorization: $B + \lambda^l I = R^T R$;
Solve $R^T R \boldsymbol{p}_l = -\boldsymbol{g}$, $R^T \boldsymbol{q}_l = \boldsymbol{p}_l$;
Set

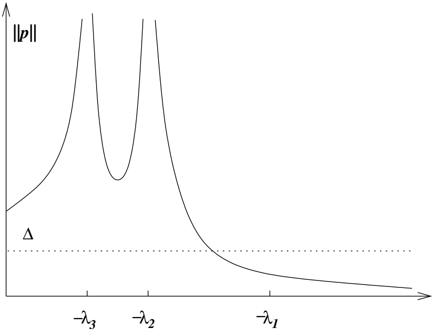
$$\lambda^{l+1} = \lambda^l + \left(\frac{\|\boldsymbol{p}_l\|_2}{\|\boldsymbol{q}_l\|_2}\right)^2 \left(\frac{\|\boldsymbol{p}_l\|_2 - \Delta}{\Delta}\right);$$

end loop

- ★ Each iteration requires 1 factorization and solving 3 triangular-system.
- ★ Safeguards must be added to ensure $\lambda > -\lambda_1$.
- ★ In practical, only 2 or 3 iterations are used to obtain an approximation.

Exact solver for the subproblem

- **Second case:**
  - **Hard case:** When $\boldsymbol{q}_1^T \boldsymbol{g} = 0$, then $\lim_{\lambda \to -\lambda_1} \|\boldsymbol{p}(\lambda)\|_2 = \infty$ does not hold, so there may not be a value $\lambda \in (-\lambda_1, \infty)$ such that $\|\boldsymbol{p}(\lambda)\|_2 = \Delta$.



Exact solver for the subproblem

- **Second case:**
  - **Hard case:** When $\boldsymbol{q}_1^T \boldsymbol{g} = 0$, then $\lim_{\lambda \to -\lambda_1} \|\boldsymbol{p}(\lambda)\|_2 = \infty$ does not hold, so there may not be a value $\lambda \in (-\lambda_1, \infty)$ such that $\|\boldsymbol{p}(\lambda)\|_2 = \Delta$.

    We set $\lambda = -\lambda_1$, and seek τ such that

    $$\|\boldsymbol{p}\|_2^2 = \sum_{j:\lambda_j \neq \lambda_1} \frac{\left(\boldsymbol{q}_j^T \boldsymbol{g}\right)^2}{(\lambda_j + \lambda)^2} + \tau^2 = \Delta^2,$$

    then

    $$\boldsymbol{p} = \sum_{j:\lambda_j \neq \lambda_1} \frac{\boldsymbol{q}_j^T \boldsymbol{g}}{\lambda_j + \lambda} \boldsymbol{q}_j + \tau \mathbf{z},$$

    where $\mathbf{z} = \boldsymbol{q}_1 / \|\boldsymbol{q}_1\|_2$.

Exact solver for the subproblem

Computing an exact solution for the subproblem, we need an iterative process, and in each iteration we may need

- one Cholesky factorization
- three triangular system solvings

Is it worth?

- Even an exact solution may be rejected.
- Factorization may not even be possible for large scale problems.
- Exact solution is not really necessary according to the global convergence results.

In practical, we only use exact solvers for small problems, and usually only use a few iterations to obtain an approximate solution. Inexact solvers are more often used.

The dogleg method

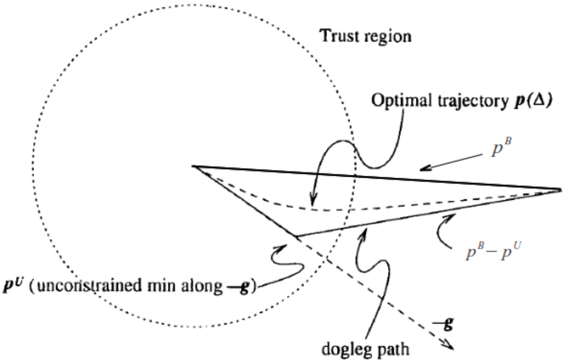
$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m(\boldsymbol{p}) = f + \boldsymbol{g}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B\boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\|_2 \leq \Delta \tag{2}$$

We consider the minimizer $\boldsymbol{p}^*$ as a function with respect to Δ, which shows the effect of the trust-region radius Δ on the solution $\boldsymbol{p}^*$ to (2).

- When B is positive definite, then the unique minimizer of the unconstrained problem is $\boldsymbol{p}^B = -B^{-1}\boldsymbol{g}$.
- When Δ is large such that $\Delta \geq \|\boldsymbol{p}^B\|_2$, i.e., $\boldsymbol{p}^B$ is feasible for the above minimization problem, then we have $\boldsymbol{p}^*(\Delta) = \boldsymbol{p}^B$.
- When Δ is small relative to $\boldsymbol{p}^B$, we can omit the quadratic term in (1), and obtain $\boldsymbol{p}^*(\Delta) \approx -\Delta\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$.
- For intermediate values of Δ, the solution $\boldsymbol{p}^*$ typically follows a curved trajectory.

The dogleg method

Idea: Approximate the curved trajectory for $\boldsymbol{p}^*(\Delta)$ by piecewise linear curve. Then, find an approximate solution on the piecewise linear curve.



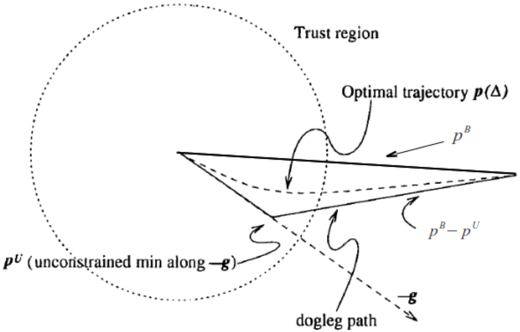
The dogleg method

The piecewise linear curve consists of two line segments:

1. For small Δ, the line segment runs from the origin to the minimizer of m along the steepest descent direction, i.e.,

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

2. For large Δ, the line segment runs from $\boldsymbol{p}^U$ to $\boldsymbol{p}^B = -B^{-1}\boldsymbol{g}$.



The dogleg method

The piecewise linear curve consists of two line segments:

1. For small Δ, the line segment runs from the origin to the minimizer of m along the steepest descent direction, i.e.,

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

2. For large Δ, the line segment runs from $\boldsymbol{p}^U$ to $\boldsymbol{p}^B = -B^{-1}\boldsymbol{g}$.

We denote this trajectory by $\tilde{\boldsymbol{p}}(\tau)$ for $\tau \in [0, 2]$, where

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

The dogleg method

Lemma

Let B be positive definite. Then,

1. $\|\tilde{\boldsymbol{p}}(\tau)\|_2$ is an increasing function of τ, and
2. $m(\tilde{\boldsymbol{p}}(\tau))$ is a decreasing function of τ.

Then, the approximate solution of the constrained minimization problem (2) can be obtained by this simple algorithm:

1. If $\Delta \leq \|\boldsymbol{p}^U\|_2$, we set $\boldsymbol{p} = \Delta \boldsymbol{p}^U / \|\boldsymbol{p}^U\|_2$;
2. If $\Delta \leq \|\boldsymbol{p}^B\|_2$, we set $\boldsymbol{p} = \boldsymbol{p}^U + (\tau - 1)(\boldsymbol{p}^B - \boldsymbol{p}^U)$ where τ is the root in the interval $[1, 2]$ of

$$\|\boldsymbol{p}^U + (\tau - 1)(\boldsymbol{p}^B - \boldsymbol{p}^U)\|_2^2 = \Delta^2.$$

3. If $\Delta \geq \|\boldsymbol{p}^B\|_2$, we set $\boldsymbol{p} = \boldsymbol{p}^B$.

Two-dimensional subspace minimization

Idea: Instead of looking for an approximate solution on the piecewise linear curve, we can search for on the entire 2D subspace spanned by $\boldsymbol{p}^U$ and $\boldsymbol{p}^B$.

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} m(\boldsymbol{p}) = f + \boldsymbol{g}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B\boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\|_2 \le \Delta,\ \boldsymbol{p} \in \mathrm{span}[\boldsymbol{g}, B^{-1}\boldsymbol{g}].$$

Equivalently,

$$\min_{[c_1,c_2]^T\in\mathbb{R}^2} m(c_1\boldsymbol{g} + c_2B^{-1}\boldsymbol{g}) \quad \text{s. t. } \|c_1\boldsymbol{g} + c_2B^{-1}\boldsymbol{g}\|_2 \le \Delta.$$

- The Cauchy point is feasible for this problem, so the optimal solution yields at least as much reduction in m as the Cauchy point, resulting in global convergence.
- For the case of indefinite B, we can replace the 2D subspace by $\mathrm{span}[\boldsymbol{g}, (B+\alpha I)^{-1}\boldsymbol{g}]$, where $\alpha \in (-\lambda_1, -2\lambda_1]$ with λ_1 as the most negative eigenvalue of B.

Summery of the trust-region

Algorithm

Set $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, $\boldsymbol{x}_0$.

loop

Obtain $\boldsymbol{p}_k$ by applying exact or inexact solvers to solve

$$\min_{\boldsymbol{p} \in \mathbb{R}^n} m_k(\boldsymbol{p}) = f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T B_k \boldsymbol{p}, \quad \text{s. t. } \|\boldsymbol{p}\| \leq \Delta_k$$

$\boldsymbol{x}_{new} = \boldsymbol{x}_k + \boldsymbol{p}_k$;
$\rho_k = (f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k))/(m_k(\boldsymbol{0}) - m_k(\boldsymbol{p}_k))$;

Update Δ_k to get Δ_{k+1};

Accept or reject $\boldsymbol{x}_{new}$ according to ρ_k;

end loop