Lecture 24: Trust-Region Methods

Yudong Chen

So far, we have been looking at methods of the form

$$x_{k+1} = x_k - \alpha_k \underbrace{B_k^{-1} \nabla f(x_k)}_{-p_k},$$

where $B_k \succ 0$. Examples:

- $B_k = I$: steepest descent;
- $B_k = \nabla^2 f(x_k)$: (damped) Newton's method
- B_k approximates $\nabla^2 f(x_k)$: quasi-Newton method.

In all these methods, we first determine the search direction p_k, then choose the stepsize α_k.

In Trust region (TR) methods, we first determine the size of the step, then the direction.

1 Trust region method

We want to compute the step p_k that gives the next iterate $x_{k+1} = x_k + p_k$.

Let $B_k \in \mathbb{R}^{d \times d}$ be given; typically, B_k equals $\nabla^2 f(x_k)$ or an approximation thereof obtained by Quasi-Newton (say SR1). Consider the following a quadratic approximate model of f around x_k:

$$m_k(p) := f(x_k) + \langle \nabla f(x_k), p \rangle + \frac{1}{2} p^\top B_k p.$$

Basic idea of TR: to compute p_k, we minimize $m_k(p)$ over a region (a ball centered at x_k) within which we trust that m_k is a good approximation of f.

Remark 1. We do *not* require $B_k \succ 0$. In particular, we can use an indefinite $\nabla^2 f(x_k)$ without modification.

Formally, the (exact) TR direction is given by

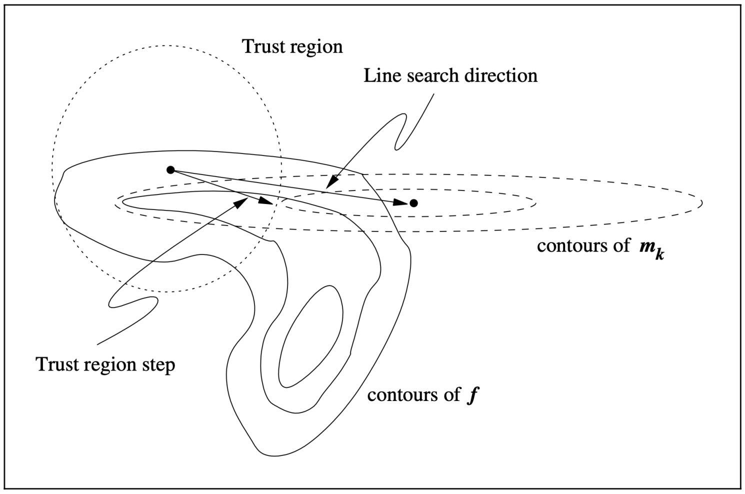
$$p_k := \operatorname*{argmin}_{p \in \mathbb{R}^d : \|p\| \le \Delta_k} m_k(p),$$

where Δ_k is the radius of the trust region.

Example 1. Suppose $f(x) = x_1^2 - x_2^2$, which is a nonconvex quadratic. The quadratic model is the function itself: $m_k(p) = f(x_k + p)$. If $x_k = 0$, then $\nabla f(x_k) = 0$, so gradient descent (GD) and Newton's method will stay at 0 (a stationary point). TR method will take the step

$$\begin{aligned} p_k &= \operatorname*{argmin}_{p : \|p\| \le \Delta_k} m_k(p) \\ &= \operatorname*{argmin}_{p : p_1^2 + p_2^2 \le \Delta_k^2} \left\{ (0 + p_1)^2 - (0 + p_2)^2 \right\} = (0, \Delta_k) \text{ or } (0, -\Delta_k). \end{aligned}$$

For more general functions, see the illustration below from Nocedal-Wright:



To completely specify the TR method, we need to decide:

1. how to choose the radius Δ_k,
2. how and to what accuracy to solve the minimization problem $\min_{p \in \mathbb{R}^d : \|p\| \leq \Delta_k} m_k(p)$.

2 Choosing the radius Δ_k

Define

$$\rho_k := \frac{\overbrace{f(x_k) - f(x_k + p_k)}^{\text{actual reduction}}}{\underbrace{m_k(0) - m_k(p_k)}_{\text{predicted reduction}, \geq 0}}.$$

The ratio ρ_k tells us whether we are making progress, and if so, how much.

General idea:

1. If $\rho_k \approx 1$, then f and m_k agree well for within the trust region $\|p\| \leq \Delta_k$. We can try increasing Δ_k in next iteration.
2. If $\rho_k < 0$, then f has increased. We should reject the step.
3. If ρ_k is small or negative, we should consider decreasing Δ_k (shrink the trust region).

The following algorithm describes the process.

Algorithm 1 Trust Region

Input: $\hat{\Delta} > 0$ (largest radius), $\Delta_0 \in (0, \hat{\Delta})$ (initial radius), $\eta \in [0, 1/4)$ (acceptance threshold)
for $k = 0, 1, 2, \ldots$
    $p_k = \operatorname{argmin}_{p:\|p\| \leq \Delta_k} m_k(p)$ (or approximate minimizer)
    $\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$
    if $\rho_k < \frac{1}{4}$: `\\ insufficient progress`

$$\Delta_{k+1} = \frac{1}{4}\Delta_k \quad \text{\\ reduce radius}$$

    else:
        if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$: `\\ sufficient progress, active trust region`

$$\Delta_{k+1} = \min\left\{2\Delta_k, \hat{\Delta}\right\} \quad \text{\\ increase radius}$$

        else: `\\ sufficient progress, inactive trust region`

$$\Delta_{k+1} = \Delta \quad \text{\\ keep radius}$$

    if $\rho_k > \eta$: `\\ sufficient progress`

$$x_{k+1} = x_k + p_k \quad \text{\\ accept step}$$

    else: `\\ insufficient progress`

$$x_{k+1} = x_k \quad \text{\\ reject step}$$

end for

3 Exact minimization of m_k

In each iteration of Algorithm 1, we need to solve the TR sub-problem

$$\min_{p:\|p\| \leq \Delta_k} m_k(p) := f_k + g_k^\top p + \frac{1}{2} p^\top B_k p, \tag{P_{m_k}}$$

where we introduce the shorthands $f_k := f(x_k)$ and $g_k := \nabla f(x_k)$.

The theorem below characterizes the exact minimizer $p_k^* = \operatorname{argmin}_{p:\|p\| \leq \Delta_k} m_k(p)$.

Theorem 1 (Characterizing the solution to (P_{m_k}))**.** *The vector $p^* \in \mathbb{R}^d$ is a global solution to the problem (P_{m_k}) if and only if p^* is feasible ($\|p^*\| \leq \Delta_k$) and there exists $\lambda \geq 0$ such that the following condition holds:*

1. $(B_k + \lambda I)p^* = -g_k$,
2. $\lambda(\Delta_k - \|p^*\|) = 0$ *(complementary slackness),*
3. $B_k + \lambda I \succcurlyeq 0$.

The proof of Theorem 1 makes use of the theory of constrained optimization and Lagrangian multipliers, which we will not delve into.

Exercise 1. Prove the necessity of part 1 above using the first-order optimality condition (Lecture 14, Theorem 1).

Some observations about Theorem 1:

- If $\|p^*\| < \Delta_k$, then the constraint is inactive/irrelevant. In this case, part 2 implies $\lambda = 0$, part 1 implies $B_k p^* = -g_k$, and part 3 implies $B_k \succcurlyeq 0$. See p^{*3} in the figure below.
- In the other case where $\|p^*\| = \Delta_k$, then $\lambda > 0$. From part 1:

$$\lambda p^* = -B_k p^* - g_k = -\nabla m_k(p^*),$$

hence p^* is parallel to $-\nabla m_k(p^*)$ and thus normal to contours of m_k; equivalently, $-\nabla m_k(p^*) \in N_{\mathcal{X}}(p^*)$, where $\mathcal{X} = \{p : \|p\| \leq \Delta_k\}$. See p^{*1} and p^{*2} in the figure below.

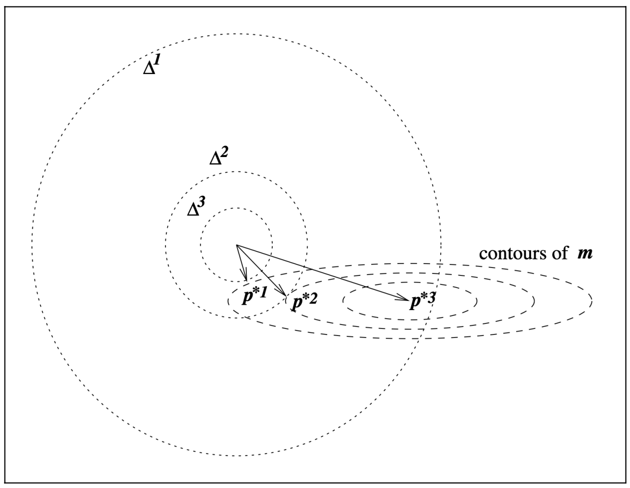


Figure 4.2 Solution of trust-region subproblem for different radii Δ^1, Δ^2, Δ^3.

To find the exact minimizer p_k^*, one may use an iterative method to search for the λ that satisfies the conditions in Theorem 1.

4 Approximate methods for minimizing m_k

Solving the TR subproblem (P_{m_k}) exactly is unnecessary. After all, m_k is only a local approximation of f.

4.1 Algorithms based on the Cauchy point

The *Cauchy point* p_k^C is defined by the following procedure.

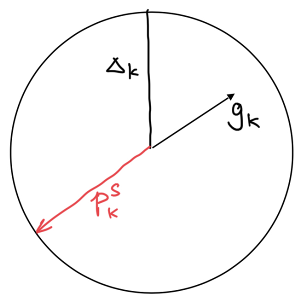
Algorithm 2 Cauchy Point Calculation

Compute

$$p_k^{\mathrm{S}} = \operatorname*{argmin}_{p:\|p\|\le\Delta_k} \left\{ f_k + g_k^\top p \right\},$$
$$\tau_k = \operatorname*{argmin}_{\tau\ge 0:\left\|\tau p_k^{\mathrm{S}}\right\|\le\Delta_k} m_k(\tau p_k^{\mathrm{S}}).$$

Return $p_k^{\mathrm{C}} = \tau_k p_k^{\mathrm{S}}$

Note that p_k^{S} is the minimizer of the *linear* model $f_k + g_k^\top p$ within the trust region; that is, p_k^{S} solves the linear version of the TR subproblem (P_{m_k}). The scalar τ_k is obtained by minimizing the *quadratic* model m_k along the direction of p_k^{S}.



Linear version, ignoring the quadratic part

The Cauchy point can be easily computed. First observe that

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

Hence

$$\begin{aligned} m_k(\tau p_k^{\mathrm{S}}) &= f_k + \tau \left\langle g_k, -\frac{\Delta_k}{\|g_k\|} g_k \right\rangle + \frac{\tau^2}{2} \left(\frac{\Delta_k}{\|g_k\|} g_k \right)^\top B_k \left(\frac{\Delta_k}{\|g_k\|} g_k \right) \\ &= f_k \underbrace{-\tau\Delta_k \|g_k\|}_{\le 0} + \frac{\tau^2}{2} \frac{\Delta_k^2}{\|g_k\|^2} g_k^\top B_k g_k. \end{aligned}$$

The RHS is a one-dimensional quadratic function of τ. Since $\left\|p_k^{\mathrm{S}}\right\| = \Delta_k$, the trust-region constraint $\left\|\tau p_k^{\mathrm{S}}\right\| \le \Delta_k$ is equivalent to $\tau \le 1$.

- Case 1: $g_k^\top B_k g_k \le 0$. Then $m_k(\tau p_k^{\mathrm{S}})$ is decreasing in τ, so the minimizer is on the boundary of the trust region, that is, $\tau_k = \frac{\Delta_k}{\|p_k^{\mathrm{S}}\|} = 1$.
- Case 2: $g_k^\top B_k g_k > 0$. Then $m_k(\tau p_k^{\mathrm{S}})$ is a convex quadratic in τ, hence τ_k is either the unconstrained minimizer of $m_k\left(\tau p_k^{\mathrm{S}}\right)$, or 1 (on the boundary), whichever is smaller.

Combining Case 1 + Case 2, we conclude that

$$\tau_k = \begin{cases} 1 & g_k^\top B_k g_k \le 0, \\ \min\left\{1, \frac{\|g_k\|^3}{\Delta_k g_k^\top B_k g_k}\right\}, & g_k^\top B_k g_k > 0. \end{cases}$$

The Cauchy point p_k^{C} can be used as a benchmark for an approximate solution p_k to the TR subproblem (P_{m_k}). As we will show later, for a TR method to converge globally, it is sufficient if p_k reduces m_k by at least some constant times the decrease from the Cauchy point, i.e.,

$$m_k(0) - m_k(p_k) \le c \cdot \left(m_k(0) - m_k(p_k^{\mathrm{C}})\right), \quad \text{where } c > 0 \text{ is a constant.}$$

Note that the RHS is roughly the progress made by gradient descent.

4.2 Improving the Cauchy point

If we simply using the Cauchy point, $p_k = p_k^{\mathrm{C}}$, then the TR method will move in the direction $-\nabla f(x_k)$ and hence converge no faster than gradient descent.

The Cauchy point only uses the matrix B_k to determine the length of the step but not the direction. To achieve faster convergence, we need to make more substantial use of B_k.

4.2.1 The dogleg method

The Dogleg method is used only when $B_k \succ 0$.

Intuition: consider two extremes.

- If Δ_k is small, then $\Delta_k^2 \ll \Delta_k$. Hence for $\|p\| \le \Delta_k$, the quadratic model is approximately linear: $m_k(p) \approx f_k + g_k^\top p$. In this case, it is approximately optimal to use the Cauchy point, i.e., $p_k^* \approx p_k^{\mathrm{C}}$.
- If Δ_k is large, then the constraint $\|p_k\| \le \Delta_k$ becomes irrelevant. In this case, p_k^* approximately equals the unconstrained minimizer of m_k, i.e., $p_k^* \approx -B_k^{-1} p_k =: p_k^{\mathrm{B}}$.

The dogleg method interpolates between these two extremes.