Chapter 25 Trust Region Method

1. Trust Region Model
2. Trust Region Radius and Subproblem

Trust Region Method

revisit Newton method

$\because$ $f(\boldsymbol{x}) \approx q_k(\boldsymbol{x}) = f(\boldsymbol{x}^{(k)}) + (\boldsymbol{g}^{(k)})^\top(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{F}_k(\boldsymbol{x} - \boldsymbol{x}^{(k)})$, where $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$, $\boldsymbol{F}_k = \nabla^2 f(\boldsymbol{x}^{(k)})$.

$\therefore$ Newton method: given $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} q_k(\boldsymbol{x})$, i.e.,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{F}_k)^{-1}\boldsymbol{g}^{(k)}$$

Disadvantages:

- Newton method is locally convergent, i.e., $\boldsymbol{x}^{(0)}$ should be near to $\boldsymbol{x}^*$.
- Newton method may fail if $\boldsymbol{F} \not\succ 0$.

motivation: restrict line search in neighborhood of $\boldsymbol{x}^{(k)}$

trust region: $\Omega_k = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le r_k\}$, where $r_k > 0$ is trust region radius.
trust region method: given $\boldsymbol{x}^{(k)}$,
$\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}\in\Omega_k} q_k(\boldsymbol{x})$ and tune radius r_k dynamically at each step.

★ $\|\cdot\|$ can be any ℓ^p-norm. ℓ^2-norm is popular.

★ $\boldsymbol{F}_k$ in $q_k(\boldsymbol{x})$ can be replaced by any approximation, e.g., BFGS, DFP, $\cdots$

Trust Region Method

(w.l.o.g) we use
$f(\boldsymbol{x}) \approx q_k(\boldsymbol{x}) = f(\boldsymbol{x}^{(k)}) + (\boldsymbol{g}^{(k)})^\top(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{B}_k(\boldsymbol{x} - \boldsymbol{x}^{(k)})$,
where $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$, $\boldsymbol{B}_k \approx \nabla^2 f(\boldsymbol{x}^{(k)})$.

trust region subproblem with general approximation

$\min q_k(\boldsymbol{x}) = f(\boldsymbol{x}^{(k)}) + (\boldsymbol{g}^{(k)})^\top(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{B}_k(\boldsymbol{x} - \boldsymbol{x}^{(k)})$
s.t. $\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le r_k$

By defining $\boldsymbol{s} = \boldsymbol{x} - \boldsymbol{x}^{(k)}$, trust region subproblem reduces to:

$$\boxed{\min (\boldsymbol{g}^{(k)})^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{B}_k \boldsymbol{s}, \quad \text{s.t. } \|\boldsymbol{s}\| \le r_k.} \tag{1}$$

properties of trust region method

❶ rapid locally convergent (rate), and ideal globally convergent.
❷ Instead of first direction then stepsize in linear search, trust region method first limit stepsize and then determine direction.
❸ trust region method can be efficient for nonconvex problem.

Trust Region Subproblem

Quadratic problem: $\min\ f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}, \quad \text{s.t. } \|\boldsymbol{x}\|_2 \le 1.$

- If $\boldsymbol{b} = 0$, then the optimum is

$$\boldsymbol{x}^* = \begin{cases} 0, & \text{if } \lambda_{\min}(\boldsymbol{Q}) \ge 0 \\ \text{unit eigenvector of } \boldsymbol{Q} \text{ for } \lambda_{\min}(\boldsymbol{Q}), & \text{if } \lambda_{\min}(\boldsymbol{Q}) < 0 \end{cases}$$

- If $\boldsymbol{b} \neq 0$, the KKT condition is

$$\boldsymbol{Q}\boldsymbol{x}^* + \boldsymbol{b} + \lambda^* \boldsymbol{x}^* = 0,\ (\|\boldsymbol{x}^*\|^2 - 1)\lambda^* = 0,\ \lambda^* \ge 0,\ \boldsymbol{Q} + \lambda^* I \succeq 0. \tag{2}$$

proof. Optimum $\boldsymbol{x}^*$ always exists (why?). If $\|\boldsymbol{x}^*\| < 1$, then $\boldsymbol{x}^*$ is an unconstrained local minimizer of f. Moreover, it must be a global minimizer (why?), KKT condition (2) holds with $\lambda^* = 0$.

If $\|\boldsymbol{x}^*\| = 1$, let $h(\boldsymbol{x}) = \frac{1}{2}\max\{0, \|\boldsymbol{x}\|^2 - 1\}$, $\alpha > 0$ and $f_k(\boldsymbol{x}) = f(x) + k|h(\boldsymbol{x})|^2 + \frac{\alpha}{2}\|\boldsymbol{x} - bx^*\|^2$.

Trust Region Subproblem

3. Optimum $\boldsymbol{x}^*$ always exists (why?). If $\|\boldsymbol{x}^*\| < 1$, then $\boldsymbol{x}^*$ is an unconstrained local minimizer of f. Moreover, it must be a global minimizer (why?), KKT condition (2) holds with $\lambda^* = 0$.
4. If $\|\boldsymbol{x}^*\| = 1$, let $h(\boldsymbol{x}) = \frac{1}{2}\max\{0, \|\boldsymbol{x}\|^2 - 1\}$, $\alpha > 0$ and $f_k(\boldsymbol{x}) = f(x) + k|h(\boldsymbol{x})|^2 + \frac{\alpha}{2}\|\boldsymbol{x} - bx^*\|^2$.
5. Let $x^k = \arg\min\{f_k(x) \mid \|x - x^*\| \le 1\}$. We will show that x^k is an unconstrained local min of f_k for all large k.
6. Taking limit $k \to \infty$ of $f_k(x^k) = f(x^k) + k|h(x)|^2 + \frac{\alpha}{2}\|x^k - x^*\|^2 \le f_k(x^*) = f(x^*)$, along any convergent subsequence of $\{x^k\}$, we get $h(\bar{x}) = \lim_{k\to\infty} h(x^k) \to 0$.
7. Furthermore, taking limit of $f(x^k) + \frac{\alpha}{2}\|x^k - x^*\|^2 \le f(x^*)$ shows $f(\bar{x}) + \frac{\alpha}{2}\|\bar{x} - x^*\|^2 \le f(x^*)$
8. Since $h(\bar{x}) = 0$, it follows that $f(x^*) \le f(\bar{x})$. Thus, we have $\bar{x} = x^*$ and $f(x^*) = f(\bar{x})$.
9. Since $\bar{x}$ is any limit point, we have $x^k \to x^*$, so $\|x^k - x^*\| < 1$ for large k,$\Rightarrow$ x^k is an unconstrained local min of f_k, $\nabla f_k(x^k) = 0$, $\nabla^2 f_k(x^k) \succeq 0$.

Trust Region Subproblem

10. Taking limit of $0 = \nabla f(x^k) + 2kh(x^k)\nabla h(x^k) + \alpha(x^k - x^*) = \boldsymbol{Q}x^k + b + 2kh(x^k)x^k + \alpha(x^k - x^*)$. shows $2kh(x^k) = -\frac{(x^k)^\top(\boldsymbol{Q}x^k + b + \alpha(x^k - x^*))}{\|x^k\|^2} \to -(x^*)^\top(\boldsymbol{Q}x^* + b) \equiv \lambda^*$. Taking limit in (3) yields $\boldsymbol{Q}x^* + b + \lambda^* x^* = 0$, $\lambda^* \geq 0$.
11. The remaining condition $\boldsymbol{Q} + \lambda^* I \succeq 0$ follows from a contra-positive argument: if $\boldsymbol{Q} + \lambda^* I \not\succeq 0$, then there exists a $u \neq 0$ (e.g., the eigenvector of $\boldsymbol{Q} + \lambda^* I$ corresponding to a negative eigenvalue, perturbed if necessary) s.t.

$$u^\top(\boldsymbol{Q} + \lambda^* I)u < 0, \quad \langle u, x^* \rangle < 0.$$

Then, perturb x^* by u to derive a contradiction to the optimality of x^*.

12. Finally, check the sufficiency: suppose (2) holds for some x^*, λ^*. Consider the following quadratic function

$$f_{\lambda^*}(x) = f(x) + \frac{1}{2}\lambda^*(\|x\|^2 - 1)$$

Claim: x^* is a global min of f_{λ^*}. So x^* is a global min of f over the sphere $\|x\|^2 = 1$.

Solving Trust Region Subproblem

- Observation: if $\boldsymbol{Q} + \lambda I \succ 0$, then

$$\boldsymbol{Q}x + b + \lambda x = 0 \quad \Rightarrow \quad x(\lambda) = (\boldsymbol{Q} + \lambda I)^{-1} b.$$

  Moreover, $\|x(\lambda)\|$ decreases w.r.t λ.

- Note: $x(-\lambda_{\min}(\boldsymbol{Q}))$ is not unique, and $\|x(-\lambda_{\min}(\boldsymbol{Q}))\|$ can be made arbitrarily large (why?).
- Strategy: binary search of λ over $[\lambda_b, \infty)$ with $\lambda_b = \max\{0, -\lambda_{\min}(\boldsymbol{Q})\}$,
  - ⋆ Check if $\|x(\lambda_b)\| \leq 1$. If yes, then $x(\lambda_b)$ is an optimmu (why?). Stop.
  - ⋆ Else, we must have $\|x(\lambda_b)\| > 1$. Let $\lambda_a = \max\{1, \lambda_b\}$, and check if $\|x(\lambda_a)\| \leq 1$. If not, update $\lambda_a := 2\lambda_a$ until $\|x(\lambda_a)\| \leq 1$.
  - ⋆ Let $\lambda = (\lambda_a + \lambda_b)/2$, If $\|x(\lambda)\| \leq 1$, then update $\lambda_a = \lambda$. Else, we update $\lambda_b = \lambda$. Stop when $|\lambda_b - \lambda_a| \leq \epsilon$.

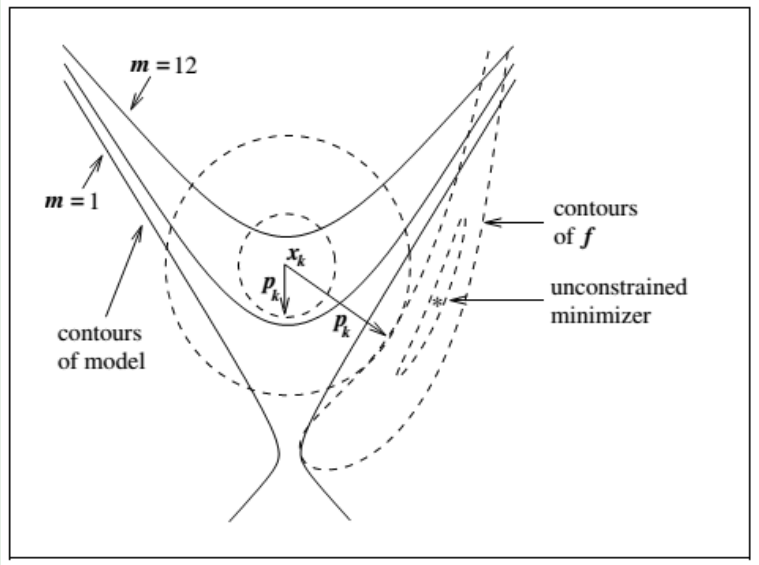
Trust Region Method

Example ($\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = 10(x_2 - x_1^2)^2 + (1 - x_1)^2$)

At the current point $\boldsymbol{x}^{(k)} = (0, 1)$, the gradient and Hessian of f at $\boldsymbol{x}^{(k)}$ are

$$\nabla f(\boldsymbol{x}^{(k)}) = \begin{pmatrix} -2 \\ 20 \end{pmatrix}, \quad \nabla^2 f(\boldsymbol{x}^{(k)}) = \begin{pmatrix} -38 & 0 \\ 0 & 20 \end{pmatrix}.$$

Two possible trust regions (circles) and their corresponding steps s_k.
The solid lines are contours of the model function in (1).



how to tune radius r_k?

when there is good agreement between the model $q_k(\boldsymbol{x}^{(k)})$ and the objective function $f(\boldsymbol{x}^{(k)})$, one should set r_k as large as possible.

Trust Region Radius

Definition (ratio for measuring agreement)

- actual reduction: $\text{Ared}_k = f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})$,
- predicted reduction: $\text{Pred}_k = q_k(\boldsymbol{x}^{(k)}) - q_k(\boldsymbol{x}^{(k+1)})$,
- ratio for measuring agreement: $\rho_k = \dfrac{\text{Ared}_k}{\text{Pred}_k}$.

procedure for tuning radius r_k

- if $\rho_k \approx 1$, it is good agreement between f and q_k. Then, dilate trust region;
- if $\rho_k \approx 0$ or $\rho_k \leq 0$, then reduce the trust region;
- otherwise, don't alter the trust region.

For example: let $\bar{r} > 0$ and $x(r_k)$ be the optimum of (1), denote

$$r_{k+1} := \begin{cases} 0.5r_k, & \text{if } \rho < 0.25 \\ r_k, & \text{if } \rho \in [0.25, 0.75] \\ \min\{2r_k, \bar{r}\}, & \text{if } \rho > 0.75 \end{cases}$$

Solve Trust Region Subproblem

By ignoring index, trust region subproblem (3) amounts to

$$\boxed{\min\ q(\boldsymbol{s}) = \boldsymbol{g}^\top \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{B}\boldsymbol{s}, \quad \text{s.t.}\ \ \|\boldsymbol{s}\| \le r.} \tag{3}$$

By KKT conditions, $\boldsymbol{s}^*$ is an optimum of (3) $\iff$ $\exists$ scalar $\lambda^* \ge 0$ such that

$$(\boldsymbol{B} + \lambda^* I)\boldsymbol{s}^* = -\boldsymbol{g}, \ \ \|\boldsymbol{s}^*\|_2 \le r, \ \ \lambda^*(r - \|\boldsymbol{s}^*\|_2) = 0, \ \ \boldsymbol{B} + \lambda^* I \succ 0.$$

Cauchy point

Let $\boldsymbol{s}^G = -\frac{r\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$ (why?) be the optimum of linear version of (3), i.e.,

$$\boxed{\boldsymbol{s}^G = \arg\min_{\|\boldsymbol{s}\| \le r}\ q(\boldsymbol{s}) = \boldsymbol{g}^\top \boldsymbol{s}.}$$

The Cauchy point is defined by $\boldsymbol{s}^c = \tau \boldsymbol{s}^G = -\tau \frac{r\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$, where

$$\tau = \begin{cases} 1, & \boldsymbol{g}^\top \boldsymbol{B}\boldsymbol{g} \le 0; \\ \min\{\|\boldsymbol{g}\|_2^3/(r\boldsymbol{g}^\top \boldsymbol{B}\boldsymbol{g}), 1\} & \text{otherwise.} \end{cases}$$

Inexact Solver

Cauchy point (version 2)

- Cauchy point $\boldsymbol{x}^c$ is defined by
$$\begin{cases} \boldsymbol{z} := \arg\min \left\{ \boldsymbol{g}^\top (\boldsymbol{x} - \boldsymbol{x}^{(k)}) \mid \|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le r \right\}, \\ \boldsymbol{x}^c := \arg \min\limits_{\tau \in (0,1)} \left\{ \tau \boldsymbol{g}^\top (\boldsymbol{z} - \boldsymbol{x}^{(k)}) + \frac{\tau^2}{2} (\boldsymbol{z} - \boldsymbol{x}^{(k)})^\top \boldsymbol{B} (\boldsymbol{z} - \boldsymbol{x}^{(k)}) \right\}. \end{cases}$$
- Optima of $\boldsymbol{z}$ and $\boldsymbol{x}^c$ are
$$\begin{cases} \boldsymbol{z} = \boldsymbol{x}^{(k)} - \frac{r\boldsymbol{g}}{\|\boldsymbol{g}\|}, \\ \boldsymbol{x}^c = \boldsymbol{x}^{(k)} - \frac{\tau r \boldsymbol{g}}{\|\boldsymbol{g}\|}, \end{cases} \quad \text{where } \tau = \begin{cases} 1, & \text{if } \boldsymbol{g}^\top \boldsymbol{B} \boldsymbol{g} \le 0 \\ \min \left\{ \frac{\|\boldsymbol{g}\|^3}{r \boldsymbol{g}^\top \boldsymbol{B} \boldsymbol{g}}, 1 \right\}, & \text{else} \end{cases}$$

- Sufficient descent property: $\exists\, c_1 \in (0, 1]$ s.t. $q(\boldsymbol{x}^c) \le -c_1 \|\boldsymbol{g}\| \min\left(r, \frac{\|\boldsymbol{g}\|}{\|F\|}\right)$.
- Hessian information $\boldsymbol{F}$ can be used to accelerate convergence, e.g., find the Cauchy point $\boldsymbol{x}^c$ by
$$\min_{\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le r} \left\{ \boldsymbol{g}^\top (\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{F} (\boldsymbol{x} - \boldsymbol{x}^{(k)}) \mid \boldsymbol{x} - \boldsymbol{x}^{(k)} \in V \right\},$$
where $V := \text{span}\{\boldsymbol{g}, \boldsymbol{F}^{-1}\boldsymbol{g}\}$. It is simple to solve (reducible to root-finding of a 4th order polynomial which admits closed form solution).

Convergence of Trust Region Method

Theorem

Suppose the level set $\mathcal{S} := \{\boldsymbol{x} \mid f(\boldsymbol{x}) \leq f(\boldsymbol{x}^{(0)})\}$ *is bounded, Hessian* $\boldsymbol{F}$ *is bounded on* $\mathcal{S}$*. Assume* $\boldsymbol{x}^{(k)}$ *is at least as good as Cauchy point, i.e.,* $q(\boldsymbol{x}^{(k)}) \leq -c_1 \|\boldsymbol{g}\| \min\left(r, \frac{\|\boldsymbol{g}\|}{\|\boldsymbol{F}\|}\right)$*. Then,* $\boldsymbol{g} \to 0$ *as* $k \to \infty$*. In practice,* $\boldsymbol{F}$ *can ba replaced by uniformly bounded approximation* $\boldsymbol{B}$*.*