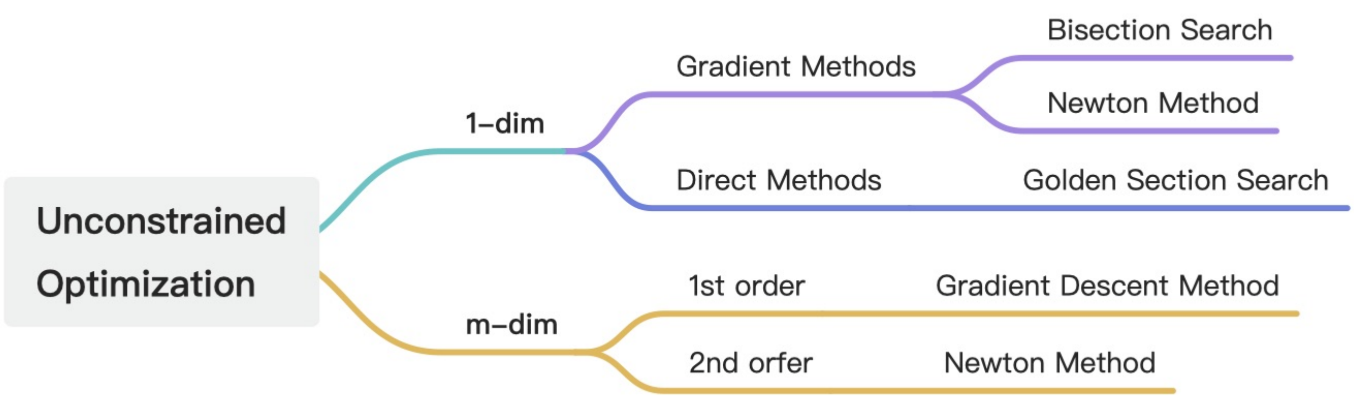
Unconstrained Minimisation



1–dim variable

Bisection Search （每次都取中点）

Bisection search algorithm

Choose an accuracy tolerance $\epsilon > 0$.

[Step 1] Choose an interval $[a_1, b_1]$ such that $f'(a_1)$ and $f'(b_1)$ have opposite signs.

[Step k] For $k = 1, 2, \cdots$,

(a) set $x_k = \frac{1}{2}(a_k + b_k)$.

(b) If $b_k - a_k \leq 2\epsilon$,

stop; use $x_k \in [a_k, b_k]$ as an approximate solution of x^*.

Else,

(i) If $f'(x_k)$ and $f'(b_k)$ have opposite sign, set $[a_{k+1}, b_{k+1}] = [x_k, b_k]$.

(ii) If $f'(x_k)$ and $f'(a_k)$ have opposite sign, set $[a_{k+1}, b_{k+1}] = [a_k, x_k]$.

Remark 5.1. *(a)* $|b_k - a_k| = |b_1 - a_1|/2^{k-1}$.

Thus, to get $|b_k - a_k| \leq 2\epsilon$, *we need to have*

$$k = \left\lceil \frac{\log\left((b_1 - a_1)/\epsilon\right)}{\log 2} \right\rceil$$

Here, $\lceil y \rceil$ *denotes the smallest integer greater than or equal to* y.

(b) At termination, $|x_k - x^*| \leq |b_k - a_k|/2 \leq \epsilon$.

Newton's Method（根据Taylor二阶展开式得到）

Newton's Method

[Step 0] Select initial point x_0, and an error of tolerance $\epsilon > 0$.

[Step k] For $k = 0, 1, 2, \cdots$,

(a) if $|f'(x_k)| < \epsilon$, stops; an appropriate critical point x_k is found.

(b) Else, compute $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$.

sensitive to the initial point

Golden Section Method

unimodal fun 只有一个全局最小值

$$\alpha^2 = 1 - \alpha, \alpha = \frac{\sqrt{5} - 1}{2}$$

Golden Section Method

[Step 0] Set $[a_0, b_0] = [a, b]$, and choose $\epsilon > 0$. Compute

$$\lambda_0 = b - \alpha(b-a), \quad \mu_0 = a + \alpha(b-a),$$

and evaluate $f(\lambda_0), f(\mu_0)$.

[Step k] For $k = 0, 1, 2, \cdots,$

(a) If $b_k - a_k < \epsilon$, stop; and $x^* \in [a_k, b_k]$.

(b) Else,

(i) If $f(\lambda_k) > f(\mu_k)$, then set

$$\begin{aligned} a_{k+1} &= \lambda_k, & b_{k+1} &= b_k, \\ \lambda_{k+1} &= \mu_k, & \mu_{k+1} &= \lambda_k + \alpha(b_k - \lambda_k) \end{aligned}$$

Evaluate $f(\mu_{k+1})$

(ii) If $f(\lambda_k) \leq f(\mu_k)$, then set

$$\begin{aligned} a_{k+1} &= a_k, & b_{k+1} &= \mu_k, \\ \lambda_{k+1} &= \mu_k - \alpha(\mu_k - a_k), & \mu_{k+1} &= \lambda_k \end{aligned}$$

Evaluate $f(\lambda_{k+1})$

m-dim variable

1st order -- Gradient Descent Method

一般问题： $\min\limits_{X \in \mathbb{R}^n} \quad f(X)$

descent direction: $\left\langle \triangledown f(x^{(k)}), p^{(k)} \right\rangle < 0$

单位方向： $d = -\dfrac{\triangledown f(x^*)}{\|\triangledown f(x^*)\|}$

Steepest descent method with exact line search

[Step 0] Select an initial point $\mathbf{x}^{(0)}$, and $\epsilon > 0$.

[Step k] For $k = 0, 1, 2, 3 \cdots$,

(a) evaluate $\mathbf{d}^{(k)} := -\nabla f(\mathbf{x}^{(k)})$.

(b) if $\|\mathbf{d}^{(k)}\| < \epsilon$, stop the algorithm; $\mathbf{x}^{(k)}$ is an approximate solution.

(c) else,

(i) find the value t_k that minimizes the one-dimensional function

$$g(t) := f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}) \quad \text{over} \quad t \geq 0.$$

(ii) Set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)}$.

zig–zag path

Theorem 5.1. *The steepest descent method moves in perpendicular steps. More precisely, if $\mathbf{x}^{(k)}$ is a steepest descent sequence for a function $f(\mathbf{x})$, then, for each k, the vector joining $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k+1)}$ (i.e. $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$) is orthogonal (perpendicular) to the vector joining $\mathbf{x}^{(k+1)}$ to $\mathbf{x}^{(k+2)}$ (i.e. $\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}$).*

收敛性

Theorem 5.2 (Convergence of gradient descent method). *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on the set $S = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$, and that S is a closed and bounded set. Then every cluster point $\bar{x}$ of the sequence $\{x_k\}$ satisfies $\nabla f(\bar{x}) = 0$.*

Convex Quadratic Form $\min\limits_{X \in \mathbb{R}^n} \quad q(X) = \frac{1}{2} X^T Q X$

$Q \succ 0$, symmetric

收敛性

Proposition 5.1. *For a symmetric positive definite* $\mathbf{Q}$*, suppose that* $\{\mathbf{x}^{(k)}\}$ *is the sequence obtained from the steepest descent method with exact line search applied to the function* $q(\mathbf{x})$*. Then*
(a) Let $\mathbf{d}^k = \nabla q(\mathbf{x}^k) = \mathbf{Q}\mathbf{x}^k$.

$$\frac{q(\mathbf{x}^{k+1})}{q(\mathbf{x}^k)} = 1 - \frac{\langle \mathbf{d}^k, \mathbf{d}^k\rangle^2}{\langle \mathbf{d}^k, \mathbf{Q}\mathbf{d}^k\rangle\langle \mathbf{d}^k, \mathbf{Q}^{-1}\mathbf{d}^k\rangle}$$

(b)

$$\frac{q(\mathbf{x}^{(k+1)})}{q(\mathbf{x}^{(k)})} \leq \left[\frac{\kappa(\mathbf{Q})-1}{\kappa(\mathbf{Q})+1}\right]^2 =: \rho(\mathbf{Q}),$$

where $\kappa(\mathbf{Q}) = \lambda_n/\lambda_1$ *and* λ_n, λ_1 *are the largest and smallest eigenvalues of* $\mathbf{Q}$*, respectively. The number* $\kappa(\mathbf{Q})$ *is called the condition number of* $\mathbf{Q}$*. When* $\kappa(\mathbf{Q}) \geq 1$ *is small, say less than* 10^3*,* $\mathbf{Q}$ *is said to be well-conditioned, and ill-conditioned otherwise.*

Remark. (a) From Proposition 5.1, we see that the convergence rate $\rho(\mathbf{Q})$ of the steepest descent method depends on $\kappa(\mathbf{Q})$. When $\kappa(\mathbf{Q})$ is large, the convergence rate

$$\rho(\mathbf{Q}) \approx 1 - \frac{4}{\kappa(\mathbf{Q})}.$$

(b) In $\mathbb{R}^2$, the contours of $q(\mathbf{x})$ are ellipses. And $\sqrt{\kappa(\mathbf{Q})}$ is the ratio between the length of the principal axes of the ellipses. Thus the larger the value of $\kappa(\mathbf{Q})$, the more elongated the ellipses are.

(c) The number of iterations needed to reduce the relative error $q(\mathbf{x}_k)/q(\mathbf{x}_0)$ to smaller than ϵ is given by

$$k = \left[\frac{\log \epsilon}{\log \rho(\mathbf{Q})}\right] + 1$$

where $[a]$ denotes the largest integer less than or equal to a.

Strongly Convex function: $\min\limits_{X \in S} \quad f(X)$

S: convex set

f: strongly convex & M−Lipschitz continuous gradient on S. Hessian satisfies:

$$mI \preceq H_f(X) \preceq MI, \quad \forall X \in S$$

收敛性

Lemma 5.1. *Let* $\mathbf{x}^*$ *be a minimizer of* $\min\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. *Then*

$$f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|^2 \;\le\; f(\mathbf{x}^*) \;\le\; f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|^2 \quad \forall\, \mathbf{x} \in S.$$

Theorem 5.3. *Let* $f : \mathbf{R}^n \to \mathbf{R}$ *be strongly convex with parameter* m *and its gradient is* M*-Lipschitz. Let* $\mathbf{x}^*$ *be the unique minimizer of* f *over* $\mathbf{R}^n$. *Define* $E_k = f(\mathbf{x}^k) - f(\mathbf{x}^*)$. *Then*

$$E_{k+1} \;\le\; E_k - \frac{1}{2M}\|\nabla f(\mathbf{x}^k)\|^2 \quad \textit{(descent inequality)}$$

$$E_{k+1} \;\le\; E_k\Big(1 - \frac{m}{M}\Big).$$

Remark. (a) From Theorem 5.3, we see that

$$E(\mathbf{x}^{k+1})/E(\mathbf{x}^1) \;\le\; (1 - m/M)^k \;\le\; \varepsilon$$

implies that we need the number of iterations k to satisfy

$$k \;\ge\; \frac{\log \varepsilon^{-1}}{\log \rho^{-1}} \;\approx\; \frac{m}{M}\log \varepsilon^{-1} \quad (\text{if } m/M \ll 1)$$

where $\rho = 1 - m/M$.

Line Search Strategies

1. Minimization rule = exact line search: find

$$\alpha_k \;:=\; \operatorname{argmin}\Big\{ f(\mathbf{x}^k + \alpha \mathbf{d}^k) \mid \alpha \geq 0 \Big\}.$$

If the search interval is limited to $\alpha \in [0, \bar{\alpha}]$, it is called limited minimization rule.

2. Armijo rule: Let $\sigma \in (0, 0.5)$ and $\beta \in (0, 1)$. Start with $\bar{\alpha}$ and continue with $\alpha = \beta\bar{\alpha}, \beta^2\bar{\alpha}, \beta^3\bar{\alpha}, \ldots$ until the following inequality is satisfied:

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) \;\leq\; f(\mathbf{x}^k) + \alpha\sigma\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k\rangle.$$

Let r be the first integer satisfying the inequality. Set $\alpha_k = \beta^r\bar{\alpha}$.

3. Non-monotone line search:

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) \;\leq\; \max\{f(\mathbf{x}^{k-l}), \ldots, f(\mathbf{x}^k)\} + \alpha\sigma\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k\rangle.$$

2nd order –– Newton Method

用Taylor 二阶形式 $q(X)$ 去逼近原函数 $f(X)$ ，多维中依赖Hessian

Algorithm for Newton Method.

[Step 0] Select an initial point $\mathbf{x}^{(0)}$, and $\epsilon > 0$.

[Step k] For $k = 0, 1, 2, 3, \cdots$

(a) evaluate $\nabla f(\mathbf{x}^{(k)})$

(b) if $\|\nabla f(\mathbf{x}^{(k)})\| < \epsilon$, then stop; and $\mathbf{x}^{(k)}$ provides a good approximation to a critical point of $f(\mathbf{x})$.

(c) else, evaluate $H_f(\mathbf{x}^{(k)})$ and set

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - H_f(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Note that the direction of descent $\mathbf{p}^{(k)}$ is $-H_f(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)})$ and $\alpha_k = 1$.

收敛性

Assumption on f

(i) H_* is non-singular.

(ii) $H(\mathbf{x})$ is Lipschitz continuous in a neighborhood of $\mathbf{x}_*$, i.e., there exists a constant $L > 0$ and $\bar{\delta} > 0$ such that

$$\|H(\mathbf{x}) - H(\mathbf{y})\|_F \leq L\|x - y\| \quad \forall\, \mathbf{x}, \mathbf{y} \in B(\mathbf{x}_*, \bar{\delta}).$$

We pick $\bar{\delta}$ to be smaller enough so that $\bar{\delta} \leq 1/(2L\|H_*^{-1}\|_F)$.

Proposition 5.2. *Suppose Assumptions (i) and (ii) hold, and $\mathbf{x}_0$ is sufficiently close to $\mathbf{x}_*$. Then the sequence $\{\mathbf{x}_k\}$ generated by the Newton method converges to $\mathbf{x}_*$ quadratically, i.e., there exists a constant M such that*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \;\leq\; M\|\mathbf{x}_k - \mathbf{x}_*\|^2$$