梯度下降法（最速下降法）

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

Thm：二次正定的函数 $f(x) = \frac{1}{2}x^TAx - b^Tx$

① ⇒ 精确搜索步长 $\arg\min_{\alpha>0} \phi(\alpha) = f(x^k - \alpha\nabla f(x^k))$

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T A \nabla f(x^k)}$$

② Q-线性收敛：$\|x^{k+1} - x^*\|_A \le \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)\|x^k - x^*\|_A$

其中 λ_1, λ_n 是 A 的最小、最大特征值，$\|x\|_A = \sqrt{x^TAx}$

Proof：

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \alpha\nabla f(x^k) \\ &= x^k - x^* - \alpha(Ax^k - b) \\ &= x^k - x^* - \alpha A(x^k - x^*) \end{aligned}$$

$\nabla f(x^k) = Ax^k - b$

$\nabla f(x^*) = 0$

$= Ax^* - b = 0$

令 $e^k = x^k - x^*$，

$$e^{k+1} = (I - \alpha A)e^k$$

$\|x\|_A^2 = x^TAx$

$$\begin{aligned} \|e^{k+1}\|_A^2 &= (e^{k+1})^T A e^{k+1} \\ &= e^{kT}(I - \alpha A)^T A (I - \alpha A)e^k \\ &= e^{kT}(A - 2\alpha A^2 + \alpha^2 A^3)e^k \\ &= e^{kT} Q(\Lambda - 2\alpha\Lambda^2 + \alpha^2\Lambda^3)Q^T e^k \\ &= \sum_{i=1}^n \lambda_i (1 - \alpha\lambda_i)^2 y_i^2 \end{aligned}$$

$A = Q\Lambda Q^T \quad Q^TQ = I$

$A^2 = Q\Lambda Q^T Q \Lambda Q^T$

$= Q\Lambda^2 Q^T$

$A^3 = Q\Lambda^3 Q^T$

$y = Q^T e^k$

$$\begin{aligned} &\sum_{i=1}^n \lambda_i y_i^2 \\ =& \sum_{i=1}^n \lambda_i (Q^T e^k)_i^2 \\ =& e^{kT} A e^k \\ =& \|e^k\|_A^2 \end{aligned}$$

由于

$$\begin{aligned} \|Ax^k - b\|_A^2 &= \mathrm{tr}\left((Ax^k-b)^T A (Ax^k - b)\right) \\ &= \mathrm{tr}\left((Ax^k-b)^T Q^T\Lambda Q (Ax^k - b)\right) \end{aligned}$$

I

$$\lambda_1\|Ax^k - b\|_2 \le \|Ax^k - b\|_A \le \lambda_n\|Ax^k - b\|_2$$

又因为 $\alpha = \frac{\|\nabla f(x^k)\|_2^2}{\nabla f(x^k)^T A \nabla f(x^k)} = \frac{\|Ax^k - b\|_2^2}{\|Ax^k - b\|_A^2} \ge \frac{2}{\lambda_1 + \lambda_n}$

$$1 - \alpha\lambda_i \le 1 - \frac{2}{\lambda_1+\lambda_n}\lambda_i \le 1 - \frac{2}{\lambda_1+\lambda_n}\lambda_n = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$$

$$\|e^{k+1}\|_A^2 \le \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 \|e^k\|_A^2$$

Defn (Condition Number 条件数) 对于可逆矩阵A，其条件数为

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$ → 任意范数

最常用：基于谱范数 (spectral norm)

$$\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$$

对于正定矩阵：$\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}}$

$\kappa(A)\uparrow$，收敛 $\downarrow$　　$\kappa(A)\downarrow$，收敛 $\uparrow$

最优性条件

梯度下降法

初始化：初始点 x^0，步长 α，最大迭代数 K，$\varepsilon \in (0,1)$

For $i = 1, 2, \cdots K$

　　$x^{k+1} = x^k - \alpha \nabla f(x^k)$

　　If $\|\nabla f(x^k)\| < \varepsilon$ then → 近似一阶最优性条件.

　　　　STOP

　　End If

End For

输出 x^{k+1}

Thm（一阶必要条件）假设 f 在全空间 $\mathbb{R}^n$ 可微，如果 x^* 是一个局部极小点，那么

$$\nabla f(x^*) = 0$$

Proof：设 x^* 为局部极小点，

$\forall d \in \mathbb{R}^n$ $\quad f(x^*+td) = f(x^*) + \nabla f(x^*)^T(td) + o(t)$

$$\Rightarrow \frac{f(x^*+td) - f(x^*)}{t} = d^T \nabla f(x^*) + o(1)$$

$$\lim_{t\to 0^+} \frac{f(x^*+td) - f(x^*)}{t} = d^T\nabla f(x^*) \geq 0 \qquad \forall d \in \mathbb{R}^n$$

$$\lim_{t\to 0^-} \frac{f(x^*+td) - f(x^*)}{t} = d^T\nabla f(x^*) \leq 0$$

$d^T\nabla f(x^*) = 0 \quad \Rightarrow \nabla f(x^*) = 0$ → 也称为稳定点（stationary point）

仅为必要条件． $f(x) = x^2$， $f(x) = -x^2$ $\quad f(x) = x^3$

$f(x, y) = x^2 - y^2$

Thm（二阶最优性条件）假设 f 在点 x^* 的一个开邻域内是二阶连续可微的．则：

①（必要性）若 x^* 是 f 的一个局部极小点，则

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0$$

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

②（充分性）若在点 x^*，有

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0$$

则 x^* 是一个局部极小点． ← 全局会怎么样呢．

Proof：证①，设 x^* 为局部极小点，在 x^* 处二阶泰勒展开： $\forall d \in \mathbb{R}^n$

$$f(x^*+d) = f(x^*) + \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2)$$

由于 $\nabla f(x^*) = 0$．反证法设 $\nabla^2 f(x) \succeq$ 不成立．即存在特征值 $\lambda_- < 0$．令 d 为 λ_- 对应的特征向量．

$$\frac{f(x^*+d) - f(x^*)}{\|d\|^2} = \frac{1}{2}\frac{1}{\|d\|^2} d^T\nabla^2 f(x^*) d + o(1)$$

$$= \frac{1}{2}\lambda_- + o(1)$$

$\nabla^2 f(x^*) d = \lambda_- d$

$d^T\nabla^2 f(x^*) d = d^T \lambda_- d$

$= \lambda_- d^T d$

$= \lambda_- \|d\|^2$

当 $\|d\| \to 0$， $\dfrac{f(x^*+d) - f(x^*)}{\|d\|^2} = \frac{1}{2}\lambda_- < 0$

$\Rightarrow f(x^*+d) < f(x^*)$ $\quad$ x^* 最优点，矛盾

② (充分性) 若点 x^*，有

$$\nabla f(x^*)=0,\quad \nabla^2 f(x^*) \succ 0$$

则 x^* 是 f 局部极小点. ← 全局需要凸性.

设 $\nabla f(x^*)=0$, $d^T\nabla^2 f(x^*)d \geq \lambda_{min}\|d\|^2 > 0 \quad \forall d\neq 0$

$$\Rightarrow \frac{f(x^*+d)-f(x^*)}{\|d\|^2} = \frac{1}{2}\frac{1}{\|d\|^2}d^T\nabla^2 f(x^*)d \geq \frac{1}{2}\lambda_{min}+o(1)$$

当 $\|d\|\to 0$

$$\frac{f(x^*+d)-f(x^*)}{\|d\|^2} \geq \frac{1}{2}\lambda_{min} > 0$$

$$\Rightarrow f(x^*+d) > f(x^*) \quad \forall d\neq 0 \qquad \therefore f(x^*) \text{ 是局部最优}$$

Defn (梯度 Lipschitz 连续) 设 f 为可微函数，若存在 $L>0$，

$$\|\nabla f(x)-\nabla f(y)\| \leq L\|x-y\| \qquad \forall x,y\in \mathrm{dom} f$$

则称 f 是(全局的) L-梯度 Lipschitz 连续函数

E.g. $f(x)=x^2$ $\quad \|\nabla f(x)-\nabla f(y)\| = \|2x-2y\| = 2\|x-y\|$

$f(x)=x^3$ $\quad \|\nabla f(x)-\nabla f(y)\| = \|3x^2-3y^2\| \leq 3\|x+y\|\,\|x-y\|$

→ 局部梯度 Lipschitz 连续，非全局 ....

$f(x)=e^x \quad x\in[0,100]$ 微分中值定理

$$\|\nabla f(x)-\nabla f(y)\| = \|\nabla^2 f(\xi)(x-y)\| \qquad \xi\in[0,100]$$

$$\leq \|\nabla^2 f(\xi)\|\,\|x-y\|$$

$$L = \max_{\xi\in[0,100]} \|\nabla^2 f(\xi)\| = e^{100}$$

梯度 Lipschitz 表明 $\nabla f(x)$ 的变化可被 x 的变化控制.

Lemma (二次上界) 设 f 为可微函数, $\mathrm{dom} f = \mathbb{R}^n$, 且 f 是 L-梯度 Lipschitz 连续函数, 则

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 \quad \forall x,y \in \mathrm{dom} f$$

↘ 二次上界.

① 意味 f 的增长不超过二次.

Thm (梯度法在凸函数上的收敛性) 设 f 为凸, 梯度 L-Lipschitz 连续, $f^* = \inf_x f(x)$ 可达. 如果步长 α 为常数且 $0<\alpha<\frac{1}{L}$, 那么

① 梯度下降法生成的序列 $\{x^k\}$ 收敛到最优值 → 全局最优

② 函数值意义下收敛速度为 $O(\frac{1}{k})$

Proof: 令 $\tilde{x} = x - \alpha\nabla f(x)$

$$f(\tilde{x}) \le f(x) + \nabla f(x)^T(-\alpha\nabla f(x)) + \frac{L}{2}\|-\alpha\nabla f(x)\|^2$$
$$= f(x) - \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2$$

令 $0<\alpha<\frac{1}{L} \Rightarrow \left(1-\frac{L\alpha}{2}\right) > \frac{1}{2}$

$f^* \ge f(x) + \nabla f(x)^T(x^*-x)$

$\Rightarrow f(x) \le f^* + \nabla f(x)^T(x-x^*)$

$$\therefore f(\tilde{x}) \le f(x) - \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2$$
$$\le f(x) - \frac{\alpha}{2}\|\nabla f(x)\|^2$$
$$\le f^* + \nabla f(x)^T(x-x^*) - \frac{\alpha}{2}\|\nabla f(x)\|^2$$
$$= f^* + \frac{1}{2\alpha}\left(\|x-x^*\|^2 - \|x-x^*\|^2 + 2\alpha\nabla f(x)^T(x-x^*) - \alpha^2\|\nabla f(x)\|^2\right)$$
$$= f^* + \frac{1}{2\alpha}\left(\|x-x^*\|^2 - \|x-x^*-\alpha\nabla f(x)\|^2\right)$$
$$= f^* + \frac{1}{2\alpha}\left(\|x-x^*\|^2 - \|\tilde{x}-x^*\|^2\right)$$

$$\Rightarrow f(\tilde{x}) - f^* \le \frac{1}{2\alpha}\left(\|x-x^*\|^2 - \|\tilde{x}-x^*\|^2\right)$$

$x = x^k$, $\tilde{x} = x^{k+1}$, $k=0,\dots K$, 求和:

$$\Rightarrow \sum_{k=0}^{K}\left(f(x^{k+1}) - f^*\right) \le \frac{1}{2\alpha}\sum_{k=0}^{K}\left(\|x^k-x^*\|^2 - \|x^{k+1}-x^*\|^2\right)$$
$$= \frac{1}{2\alpha}\left(\|x^0-x^*\|^2 - \|x^{K+1}-x^*\|^2\right)$$
$$\le \frac{1}{2\alpha}\|x^0-x^*\|^2$$

$$\Rightarrow f(x^{K+1}) - f^* \le \frac{1}{K}\sum_{k=1}^{K}\left(f(x^k)-f^*\right) \le \frac{1}{K}\frac{1}{2\alpha}\|x^0-x^*\|^2$$