梯度法

1 梯度法

利用一阶梯度信息进行算法的迭代，迭代模式为

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$$

其中 t_k 是由线搜索方法确定的步长，是一个常量
算法的优点: 每次计算代价小，不需要计算二阶导数
算法的优点: 收敛速度慢，且当目标函数不微时不可以使用梯度法

2 凸函数

定义 1. 如果 f 的定义域是凸的，并且满足 *Jensen* 不等式，即

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \quad \forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, \theta \in [0, 1]$$

则函数 f 是凸函数

定理 1. 对于连续可微函数，如果 f 为凸函数，则成立

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T + (\mathbf{y} - \mathbf{x}) \quad \forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$$

定理 2. 对于二次可微函数，如果 f 为凸函数，则成立

$$\nabla^2 f(\mathbf{x}) \succeq 0$$

2.1 严格凸函数

定义 2. f 是严格凸的，如果 dom f 是一个凸集，并且满足

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \quad \forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, \mathbf{x} \neq \mathbf{y}, \theta \in (0, 1)$$

注 1. 严格凸函数表示如果 f 存在最小值点，则最小值点是唯一的

定理 3. 对于连续可微函数，如果 f 为凸函数，则成立

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^T + (\mathbf{y} - \mathbf{x}) \quad \forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, \mathbf{x} \neq \mathbf{y}$$

注 2. $\nabla^2 f(\mathbf{x}) \succ 0$ 不是严格凸函数的必要条件，即若 f 为严格凸函数，不能得到 $\nabla^2 f(\mathbf{x}) \succ 0$，但如果满足 $\nabla^2 f(\mathbf{x}) \succ 0$，则函数为严格凸函数。比如函数 $f(x) = x^4$，其为严格凸函数，但在 $x = 0$ 处，$\nabla^2 f(x) = 0$

2.2 梯度的单调性

函数的凸性也可以利用函数梯度的单调性来定义

定理 4. *f 是凸函数，如果 $\mathrm{dom}\ f$ 是一个凸集，并且满足*

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$$

即梯度 $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ 是单调映射

证明. 必要性的证明：如果 f 是可微凸函数，则 $\forall\ \mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$ 成立

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \delta f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \quad (i)$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \delta f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \quad (ii)$$

$(i) + (ii)$ 得

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq 0$$

充分性的证明：如果 ∇f 是单调的，令

$$g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

则

$$g^{'}(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^T(\mathbf{y} - \mathbf{x})$$

可以得到

$$g^{'}(t) - g^{'}(0) = \left[\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\right](\mathbf{y} - \mathbf{x})$$

因为 ∇f 是单调的，则有

$$t\left[\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\right](\mathbf{y} - \mathbf{x}) \geq 0$$

又因为 $t > 0$，因此

$$g^{'}(t) - g^{'}(0) = \left[\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\right](\mathbf{y} - \mathbf{x}) \geq 0$$

即 $g^{'}(t) \geq g^{'}(0)$。因为

$$\begin{aligned} f(\mathbf{y}) = g(1) = g(0) + \int_0^1 g^{'}(t)\,dt &\geq g(0) + \int_0^1 g^{'}(0)\,dt \\ &= g(0) + g^{'}(0) \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \end{aligned}$$

因此 f 是凸函数 □

定理 5. f 是强凸函数，如果 dom f 是一个凸集，并且满足

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) > 0 \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f, \mathbf{x} \neq \mathbf{y}$$

即梯度 $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ 是严格单调映射

3 Lipschitz 连续梯度

定义 3. 如果函数 f 的梯度满足

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom} f$$

则称 f 的梯度是 *Lipschitz* 连续的，其中 L 被称为 *Lipschitz* 常数

注 3. 如果函数具有上面的性质，也称这个函数是 $L - smooth$ 的
定义不要求 f 是凸的，并且如果对 f 是成立的，那么对 $-f$ 也同样成立
定义中 $\|.\|_*$ 和 $\|.\|$ 互为共偶范数，共偶范数的定义为

$$\|\mathbf{u}\|_* = \sup_{\mathbf{v} \neq 0} \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{v}\|} = \sup_{\|\mathbf{v}\|=1} = \mathbf{u}^T \mathbf{v}$$

由 *Hölder* 不等式可以得到

$$\|.\| = \|.\|_* = \|.\|_2$$

由对偶范数的定义可以得到 *Cauchy–Schwarz* 不等式

$$|\mathbf{u}^T \mathbf{v}| \leq \|\mathbf{u}\|_* \|\mathbf{v}\| \quad \forall\ \mathbf{u}, \mathbf{v}$$

证明. 因为 $\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\|=1} \mathbf{u}^T \mathbf{v}$，可以得到

$$\mathbf{u}^T \mathbf{v} \leq \|\mathbf{u}\|_* \|\mathbf{v}\|$$

因为 $\mathbf{u}^T \mathbf{v}$ 具有对称性，因此可以得到

$$|\mathbf{u}^T \mathbf{v}| \leq \|\mathbf{u}\|_* \|\mathbf{v}\| \quad \forall\ \mathbf{u}, \mathbf{v}$$

□

3.1 二次上界

假设 ∇f 是 Lipschitz 连续的，并且常数为 L，由 Cauchy-Schwarz 不等式可以得到

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\|^2 \quad \forall\ \mathbf{x}, \mathbf{y} \in \text{dom } f \tag{1}$$

证明. 由定义可以得到

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

两边同乘 $\|\mathbf{x} - \mathbf{y}\|$，即

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\|^2$$

由 Cauchy-Schwarz 不等式可以得到

$$|(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})^T)(\mathbf{x} - \mathbf{y})| \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \|\mathbf{x} - \mathbf{y}\|$$

因此

$$|(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})^T)(\mathbf{x} - \mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|^2$$

即

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\|^2$$

□

如果 dom f 是凸集，则 (1) 式等价于

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \quad \forall\ \mathbf{x}, \mathbf{y} \in \text{dom} f \tag{2}$$

证明. 即证明当 dom f 是凸集时，(1) $\Longleftrightarrow$ (2)

(1) $\Longrightarrow$ (1)：假设 dom$f$ 是凸集，定义如下函数

$$g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom } f$$

其中 $t \in [0, 1]$，因为 dom f 是凸集，则

$$g^{'}(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^T(\mathbf{y} - \mathbf{x})$$

因此

$$g^{'}(t) - g^{'}(0) = (\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^T(\mathbf{y} - \mathbf{x})$$

(1) 成立，则

$$t(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \geq L\|t(\mathbf{y}-\mathbf{x})\|^2$$

即

$$(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) \geq tL\|\mathbf{y}-\mathbf{x}\|^2$$

因此可以得到

$$g^{'}(t)-g^{'}(0) \geq tL\|\mathbf{y}-\mathbf{x}\|^2$$

由积分性质

$$\begin{aligned} f(\mathbf{y}) = g(1) = g(0) + \int_0^1 g^{'}(t)\,dt \leq g(0) \int_0^1 (tL\|\mathbf{y}-\mathbf{x}\|^2 + g^{'}(0))\,dt \\ = g(0) + g^{'}(0) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \\ = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \end{aligned}$$

下证 (2) $\Longrightarrow$ (1)
交换 (2) 中 $\mathbf{x}$ 和 $\mathbf{y}$ 的位置，即

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f^T(\mathbf{y})(\mathbf{x}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

与 (2) 式相加得到

$$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \leq L\|\mathbf{x}-\mathbf{y}\|^2$$

□

推论 1. 如果 dom $f = \mathbb{R}^n$ 并且 f 有一个最小值点 $\mathbf{x}^*$，则

$$\frac{1}{2L}\|\nabla f(\mathbf{z})\|_*^2 \leq f(\mathbf{z})-f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{z}-\mathbf{x}^*\|^2 \quad \forall\ \mathbf{z} \tag{3}$$

证明. (2) 式中，令 $\mathbf{x}=\mathbf{x}^*, \mathbf{y}=\mathbf{z}$，因为 $\mathbf{x}^*$ 是最小值点，因此 $\nabla f(\mathbf{x}^*)=0$，可以得到不等式右边

$$f(\mathbf{z})-f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{z}-\mathbf{x}^*\|^2$$

(2) 式中，令 $\mathbf{x}=\mathbf{z}$，两边关于 $\mathbf{y}$ 取最小

$$\inf_{\mathbf{y}} f(\mathbf{y}) \leq \inf_{\mathbf{y}} \left(f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{y}-\mathbf{z}) + \frac{L}{2}\|\mathbf{y}-\mathbf{z}\|^2 \right)$$

令 $\mathbf{y}-\mathbf{z}=t\mathbf{v}$，可以得到

$$\begin{aligned}
\inf_{\mathbf{y}}\left(f(\mathbf{z})+\nabla f(\mathbf{z})^T(\mathbf{y}-\mathbf{z})+\frac{L}{2}\|\mathbf{y}-\mathbf{z}\|^2\right) &= \inf_{\|\mathbf{v}\|=1}\inf_{t}\left(f(\mathbf{z})+t\nabla f(\mathbf{z})^T\mathbf{v}+\frac{Lt^2}{2}\right)\\
&= \inf_{\|v\|=1}\left(f(\mathbf{z})-\frac{1}{2L}(\nabla f(\mathbf{z})^T\mathbf{v})^2\right)\\
&= f(\mathbf{z})-\frac{1}{2L}\sup_{\|\mathbf{v}\|=1}(\nabla f(\mathbf{z})^T\mathbf{v})^2\\
&= f(\mathbf{z})-\frac{1}{2L}\|\nabla f(\mathbf{z})\|_*
\end{aligned}$$

因此

$$f(\mathbf{x}^*)=\inf_{\mathbf{y}} f(\mathbf{y})\leq f(\mathbf{z})-\frac{1}{2L}\|\nabla f(\mathbf{z})\|_*$$

整理得

$$\frac{1}{2L}\|\nabla f(\mathbf{z})\|_*\leq f(\mathbf{z})-f(\mathbf{x}^*)$$

□

3.2 梯度的 Co-coercivity

性质 1. 如果 f 是定义域为 $\mathrm{dom}=\mathbb{R}^n$ 的凸函数，并且 $\nabla f(\mathbf{x})$ 是 *L-Lipschitz* 的，则有

$$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y})\geq\frac{1}{L}\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_*^2 \quad \forall\ \mathbf{x},\mathbf{y} \tag{4}$$

证明. 定义如下两个函数

$$f_{\mathrm{x}}(\mathbf{z})=f(\mathbf{z})-\nabla f(\mathbf{x})^T\mathbf{z},\quad f_{\mathrm{y}}(\mathbf{z})=f(\mathbf{z})-\nabla f(\mathbf{y})^T\mathbf{z}$$

易证 $f_{\mathrm{x}}(\mathbf{z}), f_{\mathrm{y}}(\mathbf{z})$ 为凸函数，且因为 ∇f 具有 L-Lipschitz 连续，因此 $f_{\mathrm{x}}(\mathbf{z}), f_{\mathrm{y}}(\mathbf{z})$ 具有 L-Lipschitz 连续梯度。因为 $\mathbf{x}=\arg\min f_{\mathrm{x}}(\mathbf{z})$，由 (3) 可以得到

$$\begin{aligned}
f(\mathbf{y})-f(\mathbf{x})-\nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) &= f_{\mathrm{x}}(\mathbf{y})-f_{\mathrm{x}}(\mathbf{x})\\
&\geq \frac{1}{2L}\|\nabla f_{\mathrm{x}}(\mathbf{y})\|_*^2\\
&= \frac{1}{2L}\|\nabla f(\mathbf{y})-\nabla f(\mathbf{x})\|_*^2
\end{aligned}$$

同理 $\mathbf{z}=\mathbf{y}$ 最小化 $f_{\mathrm{y}}(\mathbf{z})$，因此

$$f(\mathbf{x})-f(\mathbf{y})-\nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y})\geq\frac{1}{2L}\|\nabla f(\mathbf{y})-\nabla f(\mathbf{x})\|_*^2$$

两式相加得到

$$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2 \quad \forall\ \mathbf{x},\mathbf{y}$$

$\square$

注 4. 由梯度的 *Co-coercivity*，利用 *Cauchy-Schwarz* 不等式可以得到

$$\frac{1}{L}\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_*^2 \leq |(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y})| \leq \|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|_*\|\mathbf{x}-\mathbf{y}\|$$

即

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x}-\mathbf{y}\|$$

因此对一个定义域为 $\mathbb{R}^n$ 的可微凸函数 f，以下三条性质是等价的

1. ∇f 的 *Lipschitz* 连续
2. 二次上界性质
3. ∇f 的 *Co-coercivity* 性质

3.3 欧几里得范数的 Lipschitz 连续性

因为 $\|.\|_2 = \|\|_*$，因此由函数 f 梯度的 Lipschitz 的连续性可得

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x}-\mathbf{y}\|_2 \quad \forall\ \mathbf{x},\mathbf{y}$$

不等式 (1) 等价于

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \leq L(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \quad \forall\ \mathbf{x},\mathbf{y}$$

可以得到

$$\left((L\mathbf{x} - \nabla f(\mathbf{x})) - (L\mathbf{y} - \nabla f(\mathbf{y}))\right)^T (\mathbf{x}-\mathbf{y}) \geq 0$$

由凸函数的性质可以得到

$$\frac{L}{2}\|\mathbf{x}\|_2^2 - f(\mathbf{x})$$

是凸函数，若 f 二次可微，则 Hessian 矩阵 $LI - \nabla^2 f(\mathbf{x}) \succeq 0$，即

$$\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L \quad \forall\ \mathbf{x}$$

该不等式也是梯度 L-Lipschitz 连续的刻画

4 强凸函数

定义 4. f 是系数为 $m>0$ 的强凸函数，如果 dom f 是凸集，并且成立

$$f(\theta\mathbf{x}+(1-\theta)\mathbf{y})\leq\theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})-\frac{m}{2}\theta(1-\theta)\|\mathbf{x}-\mathbf{y}\|^2$$

注 5. $\|.\|$ 可以选择 $\|\|_2$，强凸系数 m 与范数的选择有关

定理 6. 若 f 为系数为 m 的强凸函数，则

$$h(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\frac{m}{2}t^2\|\mathbf{x}-\mathbf{y}\|^2$$

是凸函数

证明. 若 h 为凸函数，则对任意 $\theta\in[0,1]$ 成立

$$h(\theta t_1+(1-\theta)t_2)\leq\theta h(t_1)+(1-\theta)h(t_2)$$

代入 h 的表达式得到

$$\begin{aligned}f(\mathbf{x}+(\theta t_1+(1-\theta)t_2)(\mathbf{y}-\mathbf{x}))-\frac{m}{2}(\theta t_1+(1-\theta)t_2)^2\|\mathbf{x}-\mathbf{y}\|^2\\ \leq\theta\left(f(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))-\frac{m}{2}t_1^2\|\mathbf{x}-\mathbf{y}\|^2\right)\\ +(1-\theta)\left(f(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))-\frac{m}{2}t_2^2\|\mathbf{x}-\mathbf{y}\|^2\right)\end{aligned}$$

整理得

$$f(\mathbf{x}+(\theta t_1+(1-\theta)t_2))\leq\theta f(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))+(1-\theta)f(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))+\frac{m}{2}\theta(1-\theta)\|\mathbf{x}-\mathbf{y}\|^2(t_1-t_2)^2$$

其中 $\mathbf{x}+(\theta t_1+(1-\theta)t_2)=\theta(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))+(1-\theta)(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))$，并且

$$\|\mathbf{x}-\mathbf{y}\|^2(t_1-t_2)^2=\|(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))-(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))\|^2$$

即

$$\begin{aligned}f(\theta(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))+(1-\theta)(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x})))\leq{}&\theta f(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))\\ &+(1-\theta)f(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))\\ &-\frac{m}{2}\theta(1-\theta)\|(\mathbf{x}+t_1(\mathbf{y}-\mathbf{x}))-(\mathbf{x}+t_2(\mathbf{y}-\mathbf{x}))\|^2\end{aligned}$$

因此 f 是强凸函数，反过来若 f 为系数为 m 的强凸函数，可以得到 $h(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\frac{m}{2}t^2\|\mathbf{x}-\mathbf{y}\|^2$ 为凸函数 □

4.1 二次下界

如果 f 是可微的系数为 m 的强凸函数，则

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|^2 \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{dom} f \tag{5}$$

证明. TODO □

如果 f 是闭的，则有唯一的最小值 $\mathbf{x}^*$，则满足

$$\frac{m}{2}\|\mathbf{z}-\mathbf{x}^*\| \leq f(\mathbf{z}) - f(\mathbf{x}^*) \leq \frac{1}{2m}\|\nabla f(\mathbf{z})\|_*^2 \quad \forall\ \mathbf{z} \in \mathrm{dom} f$$

证明. 在 (5) 中，取 $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{z}$，最小值点满足 $\nabla f(\mathbf{x}^*) = 0$，可以得到

$$f(\mathbf{z}) \geq f(\mathbf{x}^*) + \|\mathbf{z}-\mathbf{x}^*\|^2$$

即

$$f(\mathbf{z}) - f(\mathbf{x}^*) \geq \|\mathbf{z}-\mathbf{x}^*\|^2$$

下证不等式右边，(5) 中取 $\mathbf{x} = \mathbf{z}$，两边同时关于 $\mathbf{y}$ 取最小，即

$$\inf_{\mathbf{y}}\{f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{y}-\mathbf{z}) + \frac{m}{2}\|\mathbf{y}-\mathbf{z}\|^2\} \leq_{\mathrm{y}} f(\mathbf{y})$$

同理可以得到

$$f(\mathbf{z}) - \frac{1}{2m}\|\nabla f(\mathbf{z})\|_* \leq f(\mathbf{x}^*)$$

□

4.2 强凸函数梯度的单调性

可微函数 f 是强凸的当且仅当 dom f 是凸集，并且成立

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq m\|\mathbf{x}-\mathbf{y}\|^2 \quad \forall \mathbf{x},\mathbf{y} \in \mathrm{dom}\ f$$

证明. 必要性的证明：已知 f 是强凸函数，则满足 (4) 式，交换 $\mathbf{x},\mathbf{y}$ 的位置

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|^2 \quad (i)$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{m}{2}\|\mathbf{x}-\mathbf{y}\|^2 \quad (ii)$$

$(i) + (ii)$ 得到

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq m\|\mathbf{x}-\mathbf{y}\|^2$$

充分性的证明：定义

$$g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \frac{m}{2}t^2\|\mathbf{x} - \mathbf{y}\|^2$$

可以得到 g' 单调不增 (TODO)，由凸函数的性质可以得到 g 是凸函数，由强凸函数的性质可以得到 g 为强凸函数 □

4.3 欧几里得范数的强凸性

假设 f 是关于欧几里得范数的系数为 m 的强凸函数，即

$$f(\theta\mathbf{x}+(1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})-\frac{m}{2}\theta(1-\theta)\|\mathbf{x}-\mathbf{y}\|_2^2 \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathrm{dom} f, \theta \in [0,1]$$

等价于

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

为凸函数

证明. 因为 f 是系数为 m 的强凸函数，因此

$$\begin{aligned} h(\theta\mathbf{x} + (1-\theta)\mathbf{y}) &= f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) - \frac{m}{2}\|\theta\mathbf{x} + (1-\theta)\mathbf{y}\|_2^2 \\ &\leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) - \frac{m}{2}\theta(1-\theta)\|\mathbf{x} - \mathbf{y}\|_2^2 - \frac{m}{2}\|\theta\mathbf{x} + (1-\theta)\mathbf{y}\|_2^2 \\ &= \theta f(\mathbf{x}) - \frac{m}{2}\theta\|\mathbf{x}\|_2^2 + (1-\theta)f(\mathbf{y}) - \frac{m}{2}(1-\theta)\|\mathbf{y}\|_2^2 \\ &= \theta h(\mathbf{x}) + (1-\theta)h(\mathbf{y}) \end{aligned}$$

即 $h(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta h(\mathbf{x}) + (1-\theta)h(\mathbf{y})$，因此 $h(\mathbf{x})$ 是凸函数 □

若 f 二次可微，则 h 为凸函数当且仅当

$$\nabla^2 f(\mathbf{x}) - mI \succeq 0$$

即

$$\lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq m \quad \forall\, \mathbf{x} \in \mathrm{dom}\, f$$

该不等式也是 f 为系数为 m 的强凸函数的刻画

4.4 强凸函数的 co-coercivity 性

若 f 是关于 $\|.\|_2$ 的系数为 m 的强凸函数，且梯度满足 Lipschitz 连续，Lipschitz 系数为 L，则

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

是 (L-m)-smooth，即 ∇h 是 (L-m)-Lipschitz 连续的

证明. 因为 h 是凸函数 (前面已证)，由凸函数梯度的单调性可以得到

$$\begin{aligned} 0 &\le (\nabla h(\mathbf{x}) - \nabla h(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \\ &= (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) - m\|\mathbf{x} - \mathbf{y}\|_2^2 \end{aligned}$$

由凸函数的 Lipschitz 连续性可以得到

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \le L\|\mathbf{x} - \mathbf{y}\|_2^2$$

因此

$$(\nabla h(\mathbf{x}) - \nabla h(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \le (L - m)\|\mathbf{x} - \mathbf{y}\|_2^2$$

即 ∇h 是 (L-m)-Lipschitz 连续的 □

∇h 的 co-coercivity 可表示为

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \ge \frac{mL}{m+L}\|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

证明. 由 (4) 可得，若 h 是 (L-m)-Lipschitz 连续的，则可以得到

$$(\nabla h(\mathbf{x} - \nabla h(\mathbf{y})))^T(\mathbf{x} - \mathbf{y}) \ge \frac{1}{L-m}\|\nabla h(\mathbf{x} - \nabla h(\mathbf{y})\|_2^2$$

代入 ∇h 的表达式可以得到

$$\begin{aligned} (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) - m\|\mathbf{x} - \mathbf{y}\|_2^2 \ge{} & \frac{1}{L-m}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \\ & + \frac{m^2}{L-m}\|\mathbf{x} - \mathbf{y}\|_2^2 \\ & - \frac{2m}{L-m}(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \end{aligned}$$

整理得

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \ge \frac{mL}{m+L}\|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

□

5 梯度法的收敛性分析

梯度法的迭代模式为

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k), k = 0, 1, ...$$

假设:

1. f 是凸函数的，并且在 dom f 上都是可微的
2. $\nabla f(\mathbf{x})$ 关于欧几里得范数是 L-Lipschitz 连续的，且 Lipschitz 系数 $L > 0$
3. 最优值 $f^* = \inf_{\mathbf{x}} f(\mathbf{x})$ 是有限的，并且在 $\mathbf{x}^*$ 处取到最小值

令 $\mathbf{y} = \mathbf{x} - t\nabla f(\mathbf{x})$，由 (2) 式可以得到

$$f(\mathbf{x} - t\nabla f(x)) \leq f(\mathbf{x}) - t(1 - \frac{Lt}{2})\|\nabla f(\mathbf{x})\|_2^2$$

令 $\mathbf{x}^+ = \mathbf{x} - t\nabla f(\mathbf{x})$，当 $0 < t < \frac{1}{L}$ 时

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \frac{t}{2}\|\nabla f(\mathbf{x})\|_2^2$$

因为 f 是凸函数，因此有

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)$$

因此可以得到

$$\begin{aligned} f(\mathbf{x}^+) - f^* &\leq f(\mathbf{x}) - \frac{t}{2}\|\nabla f(\mathbf{x})\|_2^2 - f^* \\ &\leq \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) - \frac{t}{2}\|\nabla f(\mathbf{x})\|_2^2 \\ &= \frac{1}{2t}\left(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x} - \mathbf{x}^* - t\nabla f(\mathbf{x})\|_2^2\right) \\ &= \frac{1}{2t}\left(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_2^2\right) \end{aligned} \tag{6}$$

假设 $\nabla f(\mathbf{x}) \neq 0$ 并且 $0 < t \leq \frac{1}{L}$，则由 (5) 式可以得到

$$f(\mathbf{x}^+) \leq f(\mathbf{x})$$

因为

$$0 \leq f(\mathbf{x}^+) - f^* \leq \frac{1}{2t}\left(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|_2^2\right)$$

因此由不等式 (6) 可得

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_2 \leq \|\mathbf{x} - \mathbf{x}^*\|_2$$

即在梯度法中，函数值和到最优值的距离都在减小

在 (6) 中，取 $\mathbf{x}=\mathbf{x}_{i-1},\mathbf{x}^+=\mathbf{x}_i$，并求和得到

$$
\begin{aligned}
\sum_{i=1}^{k}(f(\mathbf{x}_i)-f^*) &\le \frac{1}{2t}\sum_{i=1}^{k}\left(\|\mathbf{x}_{i-1}-\mathbf{x}^*\|_2^2-\|\mathbf{x}_i-\mathbf{x}^*\|_2^2\right)\\
&= \frac{1}{2t}\left(\|\mathbf{x}_0-\mathbf{x}^*\|_2^2-\|\mathbf{x}_k-\mathbf{x}^*\|_2^2\right)\\
&= \frac{1}{2t}\left(\|(\mathbf{x}_0-\mathbf{x}_k)+(\mathbf{x}_k-\mathbf{x}^*)\|_2^2-\|\mathbf{x}_k-\mathbf{x}^*\|_2^2\right)\\
&\le \frac{1}{2t}\|\mathbf{x}_0-\mathbf{x}_k\|_2^2
\end{aligned}
$$

因为 $f(\mathbf{x}_k)$ 关于 k 单调不增，因此有

$$
f(\mathbf{x}_k)-f^* \le \frac{1}{k}\sum_{i=1}^{k}(f(\mathbf{x}_i)-f^*) \le \frac{1}{2kt}\|\mathbf{x}_0-\mathbf{x}_k\|_2^2
$$

结论：$f(\mathbf{x}_k)-f^* \le \epsilon$ 的迭代次数为 $O(1/\epsilon)$

5.1 强凸函数的梯度法

如果 $\mathbf{x}^+=\mathbf{x}-t\nabla f(\mathbf{x})$，且 $0<t\le 2/(m+L)$，则有

$$
\begin{aligned}
\|\mathbf{x}^+-\mathbf{x}\|_2^2 &= \|\mathbf{x}-t\nabla f(\mathbf{x})-\mathbf{x}\|_2^2\\
&= \|\mathbf{x}-\mathbf{x}^*\|_2^2+t^2\|\nabla f(\mathbf{x})\|_2^2-2t\nabla f(\mathbf{x})^T
\end{aligned}
$$

若 f 是系数为 m 的强凸函数，且是 $L-smooth$ 的，则有

$$
(\nabla f(\mathbf{x})-\nabla f(\mathbf{x}^*))^T(\mathbf{x}-\mathbf{x}^*) \ge \frac{mL}{m+L}\|\mathbf{x}-\mathbf{x}^*\|+\frac{1}{m+L}\|\nabla f(\mathbf{x})-\nabla f(\mathbf{x}^*)\|
$$

因此上述不等式变为

$$
\begin{aligned}
\|\mathbf{x}^+-\mathbf{x}\|_2^2 &\le \|\mathbf{x}-\mathbf{x}^*\|_2^2-\frac{2tmL}{m+L}\|\mathbf{x}-\mathbf{x}^*\|_2^2-\frac{2t}{m+L}\|\nabla f(\mathbf{x})\|_2^2+t^2\|\nabla f(\mathbf{x})\|_2^2\\
&= \left(1-\frac{2tmL}{m+L}\right)\|\mathbf{x}-\mathbf{x}^*\|_2^2+t\left(t-\frac{2}{m+L}\right)\|\nabla f(\mathbf{x})\|_2^2
\end{aligned}
$$

因为 $0<t\le 2/(m+L)$，因此 $t\left(t-\frac{2}{m+L}\right)\|\nabla f(\mathbf{x})\|_2^2 \ge 0$，可以得到

$$
\|\mathbf{x}^+-\mathbf{x}\|_2^2 \le \left(1-\frac{2tmL}{m+L}\right)\|\mathbf{x}-\mathbf{x}^*\|_2^2
$$

即

$$\|\mathbf{x}-\mathbf{x}^*\|_2^2 \leq c^k\|\mathbf{x}^0-\mathbf{x}\|$$

其中 $c = 1 - \frac{2tmL}{m+L}$，即当 f 是强凸函数时，梯度法具有线性收敛性
若 $t = 2/(m+L)$，可以得到 $c = (\frac{\gamma-1}{\gamma+1})^2$，其中 $\gamma = L/m$
由前面的二次上界不等式可以得到

$$f(\mathbf{x}_k) - f^* \leq \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \leq \frac{c^k L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

若当前迭代点的函数值和最优值相差 ϵ，即

$$f(\mathbf{x}_k) - f^* \geq \frac{c^k L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \epsilon$$

解得

$$k \geq -\log_c \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\epsilon}$$

结论：$f(\mathbf{x}_k)-f^* \leq \epsilon$ 的迭代次数为 $O(\log(1/\epsilon))$

5.2 一阶方法收敛率的局限性

需要满足前面的三个假设，算法才具有较好的收敛性

6 小结

梯度的 Lipschitz 连续/函数的 L-smooth：
定义：$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - y\|$，由 Cauchy-Schwarz 不等式上式也等价于

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T (\mathbf{x} - \mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\|^2$$

若 $\mathrm{dom}f$ 是凸集，则

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

若对函数 f 的要求更多一些，即要求 f 是凸的，又要求 L-smooth，且 $\mathrm{dom}f = \mathbb{R}^n$，则可以得到

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$$

若取欧几里得范数，即 $\|.\| = \|.\|_2$，则可以得到另外两个 L-smooth 的刻画
1. $\frac{L}{2}\|\mathbf{x}\|_2^2 - f(\mathbf{x})$ 为凸函数

2. 若 f 二次可微，则 $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L$

凸函数/严格凸函数/强凸函数：

表 1: 凸函数的判断条件

定义	$f(\theta\mathbf{x}+(1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})$
一阶条件	$f(\mathbf{y}) \geq f(\mathbf{x})+\nabla f^T(\mathbf{x})(\mathbf{y}-\mathbf{x})$
二阶条件	$\nabla^2 f(\mathbf{x}) \succeq 0$
单调性	$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq 0$
其他	$g(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$ 为凸函数

表 2: 严格凸函数的判断条件

定义	$f(\theta\mathbf{x}+(1-\theta)\mathbf{y}) < \theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})$
一阶条件	$f(\mathbf{y}) > f(\mathbf{x})+\nabla f^T(\mathbf{x})(\mathbf{y}-\mathbf{x})$
二阶条件	$\nabla^2 f(\mathbf{x}) \succ 0$ 为充分非必要条件
单调性	$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) > 0$
其他	$g(t)=f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$ 为严格凸函数

表 3: 强凸函数的判断条件

定义	$f(\theta\mathbf{x}+(1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x})+(1-\theta)f(\mathbf{y})-\frac{m}{2}\theta(1-\theta)\|\mathbf{x}-\mathbf{y}\|^2$
一阶条件	$f(y) \geq f(x)+\nabla f^T(x)(y-x)+\frac{m}{2}\|\mathbf{x}-\mathbf{y}\|^2$
单调性	$(\nabla f(\mathbf{x})-\nabla f(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq m\|\mathbf{x}-\mathbf{y}\|^2$
其他	$g(\mathbf{x})=f(\mathbf{x})-\frac{m}{2}\|\mathbf{x}\|^2$ 为凸函数

特别地,若取 $\|.\|_2$,则可以得到强凸函数的另外两个刻画,$h(\mathbf{x})=f(\mathbf{x})-\frac{m}{2}\|\mathbf{x}\|_2^2 2$ 为凸函数，若二阶可微可以得到 $\lambda_{\min}(\nabla f(\mathbf{x}) \geq m$