近端梯度法

1 近端映射

定义 1. **闭凸函数**h 的近端映射定义为

$$\text{prox}_{th}(\mathbf{x}) = \arg\min_{\mathbf{u}} \left(h(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u}-\mathbf{x}\| \right), \quad t > 0$$

且 $u = \text{prox}_{th}(\mathbf{x})$ 存在且唯一,由无约束的最优性条件可以得到 $u = \text{prox}_{ht}(\mathbf{x})$ 等价于

$$\mathbf{x}-\mathbf{u} \in t\partial h(\mathbf{u}) \quad \Longleftrightarrow \quad th(\mathbf{z}) \geq th(\mathbf{u}) + (\mathbf{x}-\mathbf{u})^T(\mathbf{z}-\mathbf{u}) \quad \forall \mathbf{z}$$

定理 1. 近端映射是强非扩张的 *(firm nonexpansive)*,即

$$(\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq \|\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y})\|_2^2$$

证明. 若 $u = \text{prox}_{th}(\mathbf{x}), v = \text{prox}_{th}(\mathbf{y})$,则由最优性条件可得

$$\mathbf{x}-\mathbf{u} \in t\partial h(\mathbf{u}), \quad \mathbf{y}-\mathbf{v} \in t\partial h(\mathbf{v})$$

由次梯度的单调性可得

$$(\frac{\mathbf{x}-\mathbf{u}}{t} - \frac{\mathbf{y}-\mathbf{v}}{t})^T(\mathbf{u}-\mathbf{v}) \geq 0$$

因为 $t > 0$,上式也等价为

$$(\mathbf{x}-\mathbf{u}-\mathbf{y}+\mathbf{v})^T(\mathbf{u}-\mathbf{v}) \geq 0$$

即

$$(\mathbf{x} - \text{prox}_{th}(\mathbf{x}) - \mathbf{y} + \text{prox}_{th}(\mathbf{y}))^T(\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y})) \geq 0$$

整理可得

$$(\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \geq \|\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y})\|_2^2$$

应用 Cauchy-Schwarz 不等式可以得到近端映射的弱非扩张性

$$\|\text{prox}_{th}(\mathbf{x})-\text{prox}_{th}(\mathbf{y})\|_2^2 \leq (\text{prox}_{th}(\mathbf{x})-\text{prox}_{th}(\mathbf{y}))^T(\mathbf{x}-\mathbf{y}) \leq \|\mathbf{x}-\mathbf{y}\|_2\|\text{prox}_{th}(\mathbf{x})-\text{prox}_{th}(\mathbf{y})\|$$

即

$$\|\text{prox}_{th}(\mathbf{x}) - \text{prox}_{th}(\mathbf{y})\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2$$

□

例 1. $h(\mathbf{x})=0$，求 $\text{prox}_h(\mathbf{x})$

由定义

$$\begin{aligned}\text{prox}_h(\mathbf{x}) &= \arg\min_{\mathbf{u}} h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \\ &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2\end{aligned}$$

解得 $\mathbf{u} = \text{prox}_h(\mathbf{x}) = \mathbf{x}$

例 2. $h(\mathbf{x}) = \delta_C$，其中 C 为凸集，求 $\text{prox}_h(\mathbf{x})$

由定义

$$\begin{aligned}\text{prox}_h(\mathbf{x}) &= \arg\min_{\mathbf{u}} h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \\ &= \arg\min_{\mathbf{u}} \delta_C\mathbf{u} + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\| \\ &= \arg\min_{\mathbf{u}\in C} \frac{1}{2}\|\mathbf{u}-\mathbf{x}\| \\ &= P_C(\mathbf{x})\end{aligned}$$

例 3. $h(\mathbf{x}) = \|\mathbf{x}\|_1$，求 $\text{prox}_h(\mathbf{x})$

由定义

$$\begin{aligned}\text{prox}_h(\mathbf{x}) &= \arg\min_{\mathbf{u}} \|\mathbf{u}\|_1 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \\ &= \arg\min_{\mathbf{u}}\{\sum_i |u_i| + \frac{1}{2}(u_i - x_i)^2\}\end{aligned}$$

当 $u_i > 0$ 时，$\text{prox}_h(\mathbf{x})_i = \arg\min_{u_i} u_i + \frac{1}{2}(u_i-x_i)^2$，关于 u_i 求偏导得到

$$u_i = x_i - 1, \quad x_i > 1$$

同理当 $u_i < 0$ 时，$u_i = x_i + 1, \quad x_i < -1$
当 $u_i = 0$ 时，$-1 \le x_i \le 1$，综上

$$\text{prox}_h(\mathbf{x})_i = \begin{cases} x_i - 1 & x_i > 1 \\ 0 & |x_i| \le 1 \\ x_i + 1 & x_i < 1 \end{cases}$$

2 近端梯度法

近端梯度法针对的具有如下形式的无约束优化问题

$$\min f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$$

其中 g 为可微凸函数，定义域 $\mathrm{dom}g = \mathbb{R}^n$，$h$ 为凸函数且其近端映射计算成本低

2.1 算法迭代模式

算法的迭代模式为

$$\mathbf{x}_{k+1} = \mathrm{prox}_{t_k h}(\mathbf{x}_k - t_k \nabla g(\mathbf{x}_k))$$

其中 t_k 为步长，可以为常数，也可以通过线搜索法确定

注 1. *算法的初始点可以从不可行域开始迭代，当 $k \geq 1$ 时，$\mathbf{x}_k \in \mathrm{dom}f = \mathrm{dom}h$，即从不可行域开始迭代，迭代一次后就会回到可行域上*

2.2 解释

由近端映射的定义，$\mathbf{x}^+ = \mathrm{prox}_{th}(\mathbf{x} - t\nabla g(\mathbf{x}))$ 等价于求解如下问题

$$\begin{aligned}
\mathbf{x}^+ &= \arg\min_{\mathbf{u}} \left(h(\mathbf{u}) + \frac{1}{2t}\|\mathbf{u} - \mathbf{x} + t\nabla g(\mathbf{x})\|_2^2 \right) \\
&= \arg\min_{\mathbf{u}} \left(h(\mathbf{u}) + g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{u} - \mathbf{x}) + \frac{1}{2t}\|\mathbf{u} - \mathbf{x}\|_2^2 \right)
\end{aligned}$$

即 $\mathbf{x}^+$ 最小化 $h(\mathbf{u})$ 加 $g(\mathbf{u})$ 在 $\mathbf{x}$ 附近的二次模型

例 4. *$\min g(\mathbf{x}) + h(\mathbf{x})$，其中 g, h 满足前面的条件*

迭代点为 $\mathbf{x}^+ = \mathbf{x} - t\nabla g(\mathbf{x})$
1. 当 $h(\mathbf{x}) = 0$ 时，由前面的结论可得，$\mathbf{x}^+ = \mathbf{x} - t\nabla g(\mathbf{x})$，此时近端梯度法即为梯度法
2. 当 $h(\mathbf{x}) = \delta_C(\mathbf{x})$ 时，由前面的结论可得 $\mathbf{x}^+ = P_C(\mathbf{x} - t\nabla g(\mathbf{x}))$，此时近端梯度法即为投影梯度法
3. 当 $h(\mathbf{x}) = \|\mathbf{x}\|_1$ 时，由定义

$$\begin{aligned}
\mathrm{prox}_{th}(\mathbf{x}) &= \arg\min_{\mathbf{u}}\{\|\mathbf{u}\|_1 + \frac{1}{2t}\|\mathbf{u} - \mathbf{x}\|_2^2\} \\
&= \arg\min_{\mathbf{u}}\{\sum_i |u_i| + \frac{1}{2t}(u_i - x_i)^2\}
\end{aligned}$$

可以得到

$$(\mathrm{prox}_{th}(\mathbf{x}))_i = \arg\min_{u_i}\{\sum_i |u_i| + \frac{1}{2t}(u_i - x_i)^2\}$$

同前面一样进行分类讨论，当 $u_i > 0$ 时，关于 u_i 求导得到

$$1 + \frac{1}{t}(u_i - x_i) = 0$$

解得 $u_i = x_i - t$，同理可以得到 $u_i = 0$ 和 $u_i < 0$ 的情况，综上

$$(\mathrm{prox}_{th}(\mathbf{x}))_i = \begin{cases} x_i - t & x_i > t \\ 0 & |x_i| \leq t \\ x_i + t & x_i < -t \end{cases}$$

3 收敛性分析

3.1 假设条件

考虑如下优化问题

$$\min f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$$

假设：
1.h 是闭凸函数
2.g 是可微的，定义域 $\mathrm{dom} g = \mathbb{R}^n$，并且满足 L-smooth，m-strong 凸，若选择欧几里得范数，L-smooth 的等价刻画为

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

为凸函数，同时可以得到一个二次下界

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \tag{1}$$

m-strong 凸的等价刻画为

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

为凸函数，同时可以得到一个二次上界

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \tag{2}$$

3. 在 $\mathbf{x}^*$ 处取到最优解 f^*，最优解可以是非唯一的

3.2 梯度映射

记

$$G_t(\mathbf{x}) = \frac{1}{t}\left(x - rmprox_{th}(\mathbf{x} - t\nabla g(\mathbf{x}))\right)$$

可以得到

$$\begin{aligned}\mathbf{x}^+ &= \mathrm{prox}_{th}(\mathbf{x} - t\nabla g(\mathbf{x})) \\ &= \mathbf{x} - tG_t(\mathbf{x})\end{aligned}$$

写成这样的形式就和前面的梯度法的迭代模式很类似了，但是值得注意的是 $G_t(\mathbf{x})$ 既不是 $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$ 的梯度也不是次梯度
$\mathbf{x} - tG_t(\mathbf{x}) = \mathrm{prox}_{th}(\mathbf{x} - t\nabla g(\mathbf{x}))$ 由最优性条件可得

$$\mathbf{x} - t\nabla g(\mathbf{x}) - \mathbf{x} + tG_t(\mathbf{x}) \in t\partial h(\mathbf{x} - tG_t(\mathbf{x}))$$

即

$$G_t(\mathbf{x}) \in \nabla g(\mathbf{x}) + \partial h(\mathbf{x} - tG_t(\mathbf{x}))$$

当 $G_t(\mathbf{x})$ 时，即

$$0 \in \nabla g(\mathbf{x}) + \partial h(\mathbf{x})$$

由最优性条件可知，此时 $\mathbf{x}$ 最小化 $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$
用 $\mathbf{y} = \mathbf{x} - tG_t(\mathbf{x})$ 替换不等式 (1),(2) 可以得到对 $\forall t$ 有下面不等式成立

$$\frac{mt^2}{2}\|G_t(\mathbf{x})\|_2^2 \le g(\mathbf{x} - tG_t(\mathbf{x})) - g(\mathbf{x}) + t\nabla g(\mathbf{x})^T G_t(\mathbf{x}) \le \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2$$

若 $0 < t \le 1/L$，则不等式右边变为

$$g(\mathbf{x} - tG_t(\mathbf{x})) \le g(\mathbf{x}) - t\nabla g(\mathbf{x})^T G_t(\mathbf{x}) + \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 \tag{3}$$

若不等式 (3) 成立且 $G_t(\mathbf{x}) \neq 0$，可以得到 $mt \le 1$。同时可以得到对 $\forall \mathbf{z}$ 有

$$f(\mathbf{x} - tG_t(\mathbf{x})) \le f(\mathbf{z}) + G_t(\mathbf{x})^T(\mathbf{x} - \mathbf{z}) - \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 - \frac{m}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 \tag{4}$$

证明.

$$
\begin{aligned}
f(\mathbf{x}-tG_t(\mathbf{x})) &\le g(\mathbf{x})-t\nabla g(\mathbf{x})^TG_t(\mathbf{x})+\frac{t}{2}\|G_t(\mathbf{x})\|_2^2+h(\mathbf{x}-tG_t(\mathbf{x}))\\
&\le g(\mathbf{z})-\nabla g(\mathbf{x})^T(\mathbf{z}-\mathbf{x})-\frac{m}{2}\|\mathbf{z}-\mathbf{x}\|_2^2-t\nabla g(\mathbf{x})^TG_t(\mathbf{x})+\frac{t}{2}\|G_t(\mathbf{x})\|_2^2\\
&+h(\mathbf{x}-tG_t(\mathbf{x}))\\
&\le g(\mathbf{z})-\nabla g(\mathbf{x})^T(\mathbf{z}-\mathbf{x})-\frac{m}{2}\|\mathbf{z}-\mathbf{x}\|_2^2-t\nabla g(\mathbf{x})^TG_t(\mathbf{x})+\frac{t}{2}\|G_t(\mathbf{x})\|_2^2\\
&+h(\mathbf{z})-(G_t(\mathbf{x})-\nabla g(\mathbf{x}))^T(\mathbf{z}-\mathbf{x}+tG_t(\mathbf{x}))\\
&=g(\mathbf{z})+h(\mathbf{z})+G_t(\mathbf{x})^T(\mathbf{x}-\mathbf{z})-\frac{t}{2}\|G_t(\mathbf{x})\|_2^2-\frac{m}{2}\|\mathbf{x}-\mathbf{z}\|_2^2
\end{aligned}
$$

(3) 两边同时加 $h(\mathbf{x}-tG_t(\mathbf{x}))$ 可以得到第一个不等式关系，通过 (1) 可以得到第二个不等式，第三个不等式则通过 $G_t(\mathbf{x})-\nabla g(\mathbf{x})\in\partial h(\mathbf{x}-tG_t(\mathbf{x}))$ 得到 □

令 $\mathbf{x}^+=\mathbf{x}-tG_t(\mathbf{x})$，(4) 式中取 $\mathbf{z}=\mathbf{x}$ 可以得到

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

和梯度下降法有相同的形式

(4) 式中取 $\mathbf{z}=\mathbf{x}^*$ 可以得到

$$
\begin{aligned}
f(\mathbf{x}^+)-f^* &\le G_t(\mathbf{x})^T(\mathbf{x}-\mathbf{x}^*)-\frac{t}{2}\|G_t(\mathbf{x})\|_2^2-\frac{m}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2\\
&=\frac{1}{2t}\left(\|\mathbf{x}-\mathbf{x}^*\|_2^2-\|\mathbf{x}-\mathbf{x}^*-tG_t(\mathbf{x})\|_2^2\right)\\
&=\frac{1}{2t}\left((1-mt)\|\mathbf{x}-\mathbf{x}^*\|-\|\mathbf{x}^+-\mathbf{x}^*\|\right) &(5)\\
&\le\frac{1}{2t}\left(\|\mathbf{x}-\mathbf{x}^*\|-\|\mathbf{x}^+-\mathbf{x}^*\|\right) &(6)
\end{aligned}
$$

3.3 固定步长

不等式 (6) 中取 $x=x_i, x^+=x_{i+1}$，固定步长 $t=t_i=1/L$，并从 $i=0$ 加到 $i=k-1$ 得到

$$
\begin{aligned}
\sum_{i=1}^{k}(f(\mathbf{x}_i)-f^*) &\le \frac{1}{2t}\sum_{i=0}^{k-1}\left(\|\mathbf{x}_i-\mathbf{x}^*\|_2^2-\|\mathbf{x}_{i+1}-\mathbf{x}^*\|_2^2\right)\\
&=\frac{1}{2t}\left(\|\mathbf{x}_0-\mathbf{x}^*\|_2^2-\|\mathbf{x}_k-\mathbf{x}^*\|_2^2\right)\\
&\le\frac{1}{2t}\|\mathbf{x}_0-\mathbf{x}^*\|_2^2
\end{aligned}
$$

因为 $f(\mathbf{x}_i)$ 单调不增，因此可以得到

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

即

$$f(\mathbf{x}_k) - f^* \leq \frac{1}{2kt}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

因为 $f(\mathbf{x}^+) \geq f^*$，由 (5) 可得

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

即

$$\begin{aligned}\|\mathbf{x}^+ - \mathbf{x}^*\|_2^2 &\leq ((1-mt)\|\mathbf{x} - \mathbf{x}^*\|_2^2 \\ &\leq \|\mathbf{x} - \mathbf{x}^*\|_2^2\end{aligned}$$

因此，当 $m = 0$ 时，即 f 为凸函数时，算法次线性收敛，当 $m > 0$ 时，即 f 为系数为 m 的强凸函数时，算法线性收敛

当 $t = t_i = 1/L$ 时有

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \leq c^k\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

其中 $c = 1 - \frac{m}{L}$

3.4 线搜索步长

当 $0 < t \leq 1/L$ 时，成立下式

$$g(\mathbf{x} - tG_t(\mathbf{x})) \leq g(\mathbf{x}) - t\nabla g(\mathbf{x})^T G_t(\mathbf{x}) + \frac{t}{2}\|G_t(\mathbf{x})\|_2^2$$

当 L 未知时，可以通过回溯线搜索法使上式成立。

当 (3) 式成立时，则成立 $f(\mathbf{x}_{i+1}) < f(\mathbf{x}_i)$，并且满足

$$t_i(f(\mathbf{x}_{i+1}) - f^*) \leq \frac{1}{2}\left(\|\mathbf{x}_i - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{i+1} - \mathbf{x}^*\|_2^2\right)$$

将上述不等式从 $i = 0$ 加到 $i = k-1$ 得到

$$(\sum_{i=0}^{k-1} t_i)(f(\mathbf{x}_k) - f^*) \leq \sum_{i=0}^{k-1} t_i(f(\mathbf{x}_{i+1}) - f^*) \leq \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

因为 f_i 单调不增可以得到第一个不等式。又因为 $t_i \geq t_{\min}$，因此可以得到

$$f(\mathbf{x}_k) - f^* \leq \frac{1}{2\sum_{i=0}^{k-1} t_i}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{1}{2kt_{\min}}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

且 (3) 式成立时由 (5) 式得

$$\begin{aligned}\|\mathbf{x}_{i+1}-\mathbf{x}^*\|_2^2 &\leq (1-mt_i)\|\mathbf{x}_i-\mathbf{x}^*\|_2^2 \\ &\leq (1-mt_{\min})\|\mathbf{x}_i-\mathbf{x}^*\|_2^2 \\ &= c\|\mathbf{x}_i-\mathbf{x}^*\|_2^2\end{aligned}$$

其中 $c=1-mt_{\min}=\max\{1-\frac{\beta m}{L},1-m\hat{t}\}$，因此当 $m=0$ 时，即 g 为凸函数但不是强凸函数时，算法满足次线性收敛，当 $m>0$ 时即 g 为强凸函数时，算法满足线性收敛。并且可以得到

$$\|\mathbf{x}_k-\mathbf{x}^*\|_2^2 \leq c^k\|\mathbf{x}_0-\mathbf{x}^*\|_2^2$$

4 小结

近端梯度法针对的是如下优化问题

$$\min f(\mathbf{x})=g(\mathbf{x})+h(\mathbf{x})$$

其中 g 为可微凸函数，h 为不可微凸函数。使用范围相对狭窄，当目标函数中存在不可微部分才会有用。收敛性和梯度下降法相似，但收敛速度比次梯度法快