优化理论与算法

第八章 次梯度下降法

郭加熠|助理教授

目录

讲　　　员

郭加熠，江波，刘慧康

线性回归模型

- 考虑岭回归模型：

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|_2^2+\lambda\|x\|_2^2.$$

最优解为

$$x^\star=(A^TA+\lambda I)^{-1}A^Tb.$$

- 问题：LASSO 问题的最优性条件是什么？

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|_2^2+\lambda\|x\|_1.$$

次梯度

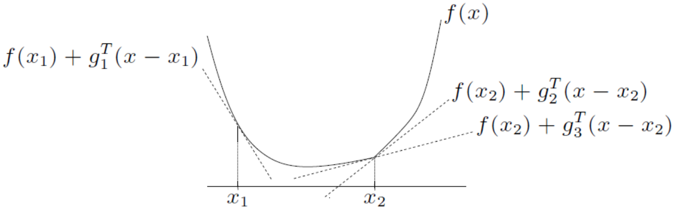
回忆凸可微函数的一阶条件：

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y.$$

定义. 对于凸函数 f，若满足：

$$f(y) \geq f(x) + g^T (y - x), \ \forall y.$$

则称 g 是 f 在 x 处的**次梯度** (subgradient)。



g_2, g_3 are subgradients at x_2; g_1 is a subgradient at x_1

次微分

f 在 x 处的**次微分** $\partial f(x)$ 是所有次梯度的集合：

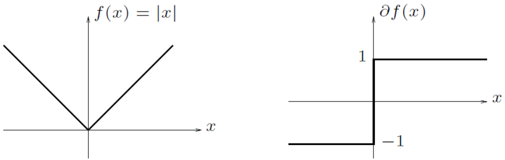
$$\partial f(x) = \{g \mid f(y) \geq f(x) + g^T(y - x),\ \forall y \in \mathrm{dom}(f)\}$$

- $\partial f(x)$ 是闭凸集（可能为空）
- 若 f 在 x 处可微，则 $\partial f(x) = \{\nabla f(x)\}$（或写作 $\partial f(x) = \nabla f(x)$）
- 若 $f(x)$ 是凸函数且定义域为开集，则 $\partial f(x)$ 在 $\mathrm{dom}(f)$ 上非空且有界

示例

绝对值函数 $f(x) = |x|$:

$$\partial f(x) = \begin{cases} 1, & 若 x > 0 \\ [-1, 1], & 若 x = 0 \\ -1, & 若 x < 0 \end{cases}$$



欧几里得范数 $f(x) = \|x\|_2$:

$$\partial f(x) = \begin{cases} \frac{x}{\|x\|_2}, & 若 x \neq 0 \\ \{g \mid \|g\|_2 \leq 1\}, & 若 x = 0. \end{cases}$$

基本规则：缩放与加法

- **缩放**: $\partial(tf) = t\partial f$ 当 $t > 0$，例如：

$$\partial(2|x|) = 2\partial(|x|), \quad \partial(2\|x\|_1) = 2\partial(\|x\|_1)$$

- **加法**: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$，例如：

$$\partial\left(\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1\right) = A^T(Ax - b) + \lambda\partial\|x\|_1$$

此外可以证明：

$$g \in \partial(\|x\|_1) \text{ 当且仅当} g_i \in \partial(|x_i|)$$

基本规则：逐点最大值

逐点最大值: 若 $f = \max_{i=1,\dots,m} f_i(x)$，则

$$\partial f(x) = \mathrm{Conv}\left(\bigcup \{\partial f_i(x) | f_i(x) = f(x)\}\right),$$

即在 x 处激活函数的次微分集合的凸包，例如:

- $|x| = \max\{x, -x\}$，则

$$\partial |x| \,|_{x=0} = \mathrm{Conv}\{-1, 1\} = [-1, 1]$$

- $\|x\|_2 = \max_{\|g\|_2=1}\{g^T x\}$，则

$$\partial \|x\|_2 \,|_{x=0} = \mathrm{Conv}\left(\{g \mid \|g\|_2 = 1\}\right) = \{g \mid \|g\|_2 \le 1\}$$

正式证明可参考"Ruszczyński, Andrzej. *Nonlinear Optimization*. Princeton University Press, Princeton, New Jersey, 2006"

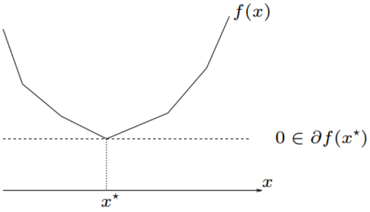
无约束问题的最优性条件

定理: 设 $f(x)$ 是**凸**且**不可微**函数，则 $x^\star$ 是最优解当且仅当

$$0 \in \partial f(x^\star).$$

证明.

- $(\Rightarrow)$: $f(x) \geq f(x^\star)$ 推导出 $0 \in \partial f(x^\star)$
- $(\Leftarrow)$: $f(x) \geq f(x^\star) + g^T(x - x^\star)$ 取 $g = 0$



应用实例

考虑 LASSO 问题：

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|b-Ax\|_2^2+\lambda\|x\|_1.$$

其最优性条件为

$$0\in A^T(Ax^\star-b)+\lambda\partial\|x^\star\|_1$$

当 $A=I$ 时，最优性条件简化为

$$0\in x^\star-b+\lambda\partial\|x^\star\|_1$$

此时最优解为

$$x_i^\star=\begin{cases} b_i-\lambda & \text{若} b_i>\lambda \\ 0 & \text{若} |b_i|\le\lambda \\ b_i+\lambda & \text{若} b_i<-\lambda. \end{cases}$$

目录

讲　　　员

郭加熠，江波，刘慧康

次梯度方法

考虑不可微凸函数 f 的极小化问题：

$$\min_{x\in\mathbb{R}^n} f(x)$$

次梯度方法:

$$x^{k+1} = x^k - t_k g^k \quad \text{其中} \quad g^k \in \partial f(x^k)$$

- **固定步长**: t_k 较小
- **递减步长**: $t_k \to 0$, $\sum_{k=1}^{\infty} t_k = \infty$

示例

使用次梯度方法求解以下一维极小化问题:

$$\min_{x\in\mathbb{R}} |x|.$$

假设 $x^0 = 1$，若选择:

- **固定步长**: 例如 $t_k = 0.7$，则迭代序列为

  $$x^1 = x^3 = \cdots = 0.3, \quad x^2 = x^4 = \cdots = -0.4.$$

  序列可能不收敛到最优解

- **递减步长**: 例如 $t_k = \frac{1}{k+1} + \frac{1}{k+2}$，则可证明

  $$x^k = \frac{(-1)^k}{k+1}.$$

  序列将收敛到最优解

示例：分段线性函数极小化

极小化以下**分段线性函数**：

$$\min f(x) := \max_{i=1,\ldots,m} a_i^T x + b_i$$

在 x^k 处的次梯度：$g^k = a_i$（其中 $i \in I$），这里 I 是激活集

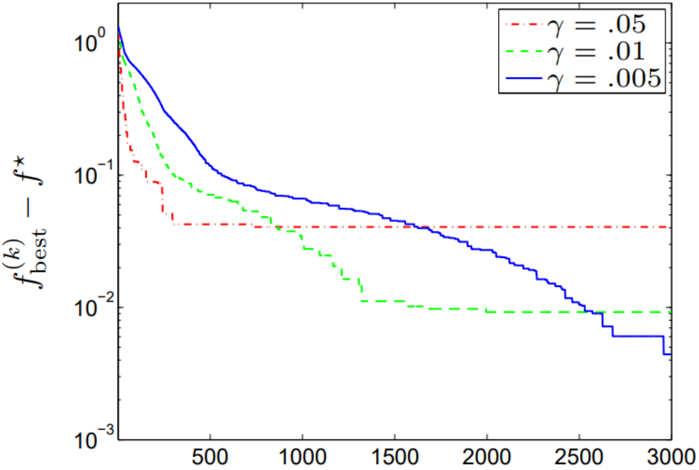
$$I = \{j : a_j^T x + b_j = \max_{i=1,\ldots,m} a_i^T x + b_i\}$$

次梯度方法：

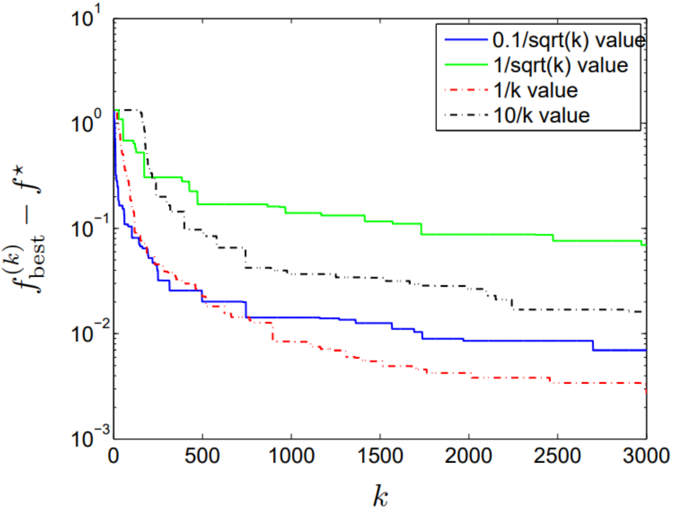
$$x^{k+1} = x^k - t_k g^k$$

问题实例：$n = 20$ 个变量，$m = 100$，$f^* \approx 1.1$

图示展示了 $f_{bs}^k - f^*$ 随迭代次数变化，固定步长 $\gamma = 0.05, 0.01, 0.005$



递减步长规则



0.1/sqrt(k) value
1/sqrt(k) value
1/k value
10/k value
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
500
1000
1500
2000
2500
3000
$f_{\mathrm{best}}^{(k)} - f^{\star}$
k

收敛性分析：固定步长

定理: 设 $f(x)$ 是凸函数且对所有 $g \in \partial f$ 有 $\|g\| \leq G$，则固定步长 t 的次梯度下降法满足：

$$f_{bs}^k - f^* \leq \frac{R^2 + G^2 k t^2}{2kt} = \frac{R^2}{2kt} + \frac{G^2 t}{2}$$

其中 $f_{bs}^k = \min_{i=0,\ldots,k-1} f(x^i)$，$R = \left\|x^0 - x^*\right\|_2$。

- 右端收敛到 $G^2 t/2$
- 不保证收敛到最优解
- 若总迭代次数 K 已知，选择 $t = \frac{R}{G\sqrt{K}}$，复杂度为 $O(\frac{1}{\sqrt{K}})$

证明

$$
\begin{aligned}
\left\|x^{i+1}-x^*\right\|_2^2 &= \left\|x^i - t_i g^i - x^*\right\|_2^2 \\
&= \left\|x^i - x^*\right\|_2^2 - 2t_i (g^i)^T (x^i - x^*) + t_i^2 \left\|g^i\right\|_2^2 \\
&\le \left\|x^i - x^*\right\|_2^2 - 2t_i (f(x^i) - f^*) + t_i^2 \left\|g^i\right\|_2^2
\end{aligned}
$$

$$
\text{故有} \quad 2t_i(f(x^i)-f^*) \le \left\|x^i - x^*\right\|_2^2 - \left\|x^{i+1}-x^*\right\|_2^2 + t_i^2 \left\|g^i\right\|_2^2 .
$$

累加不等式 $i=0$ 到 $k-1$ 并错位相消，定义 $f_{bs}^k = \min_{i=0,\dots,k-1} f(x^i)$：

$$
\begin{aligned}
2\sum_{i=0}^{k-1} t_i (f_{bs} - f^*) &\le \left\|x^0 - x^*\right\|_2^2 - \left\|x^k - x^*\right\|_2^2 + \sum_{i=0}^{k-1} t_i^2 \left\|g^i\right\|_2^2 \\
&\le \left\|x^0 - x^*\right\|_2^2 + \sum_{i=0}^{k-1} t_i^2 \left\|g^i\right\|_2^2
\end{aligned}
$$

收敛性分析：递减步长

定理: 设 $f(x)$ 是凸函数且对所有 $g \in \partial f$ 有 $\|g\| \le G$，则次梯度下降法满足：

$$f_{bs}^k - f^* \le \frac{R^2 + G^2 \sum_{i=0}^{k-1} t_i^2}{2 \sum_{i=0}^{k-1} t_i}$$

其中 $f_{bs}^k = \min_{i=0,\dots,k-1} f(x^i)$，$R = \left\|x^0 - x^*\right\|_2$。

- **非可和递减步长**: $t_k \to 0$, $\sum_{i=0}^{\infty} t_i = \infty$ 保证收敛到最优解
- 若取 $t_k = \frac{R}{G\sqrt{k}}$，迭代复杂度为 $O(\frac{RG\log(k)}{\sqrt{k}})$

证明

根据 (P18) 中的推导可得如下不等式:

$$f_{bs}^k - f^* \quad \leq \quad \frac{R^2 + G^2 \sum_{i=0}^{k-1} t_i^2}{2 \sum_{i=0}^{k-1} t_i}$$

- 对任意 $\epsilon > 0$，存在 N_1 使得当 $i > N_1$ 时 $t_i < \epsilon / G^2$
- 存在 N_2 使得 $\sum_{i=0}^{N_2} t_i \geq \frac{1}{\epsilon} \left(R^2 + G^2 \sum_{i=0}^{N_1} t_i^2 \right)$
- 则对 $k \geq \max\{N_1, N_2\}$，有

$$\frac{R^2 + G^2 \sum_{i=0}^{k-1} t_i^2}{2 \sum_{i=0}^{k-1} t_i} = \frac{R^2 + G^2 \sum_{i=0}^{N_1} t_i^2}{2 \sum_{i=0}^{k-1} t_i} + \frac{G^2 \sum_{i=N_1+1}^{k-1} t_i^2}{2 \sum_{i=0}^{k-1} t_i} \leq \epsilon$$

当 f^* 已知时的最优步长（应用于可行性问题）

Polyak 步长:

$$t_k = \frac{f(x^k) - f^*}{\|g^k\|_2^2}$$

该步长的由来是最小化不等式的右端项:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2t_k(f(x^k) - f^*) + t_k^2 \|g^k\|_2^2$$

代入 t_k 得:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - \frac{(f(x^k) - f^*)^2}{\|g^k\|_2^2}$$

递归应用可得:

$$f_{bs}^k - f^* \leq GR/\sqrt{k}$$

Polyak 步长的性能