次梯度算法

宋晓良

大连理工大学数学科学学院

教材《最优化：建模、算法与理论》配套电子教案

http://bicmr.pku.edu.cn/~wenzw/optbook.html

回顾：梯度下降算法

设 $f(x)$ 是可微凸函数且 $\mathbf{dom}\, f = \mathbb{R}^n$，考虑如下问题：

$$\min_x \; f(x).$$

- 梯度下降法：选择初始点 $x^0 \in \mathbb{R}^n$，然后重复：

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, 2, \ldots$$

  其中 $\alpha_k > 0$为步长，可取为固定常数或者通过线搜索确定.
- 若 $\nabla f(x)$ 利普西茨连续，则梯度下降法的收敛速度是 $\mathcal{O}\left(\frac{1}{k}\right)$.

如果 $f(x)$不可微呢?

提纲

非光滑优化的例子

- 极小极大问题：

$$\min_{x\in X}\ \max_{1\leq i\leq m}\ f_i(x)$$

- 求解非线性方程组：

$$f_i(x)=0,\quad i=1,\cdots,m$$

可以把它化为一个极小化问题：

$$\min_{x\in X}\|(f_1(x),\cdots,f_m(x))\|$$

特别地，$\|\cdot\|=\|\cdot\|_1$ 对应 L_1 极小化问题，$\|\cdot\|=\|\cdot\|_\infty$对应切比雪夫近似问题.

- LASSO问题：

$$\min_x\ \|Ax-b\|^2+\mu\|x\|_1$$

梯度下降法失败的例子

考虑函数 $f:\mathbb{R}^2 \to \mathbb{R}^1, x=(u,v)^{\mathrm{T}}$，

$$f(x)=\max\left[\frac{1}{2}u^2+(v-1)^2,\frac{1}{2}u^2+(v+1)^2\right].$$

- 假设迭代点 x^k 的形式为

$$x^k=\begin{pmatrix} 2\left(1+|\epsilon_k|\right) \\ \epsilon_k \end{pmatrix},\quad \text{其中}\,\epsilon_k\neq 0.$$

- 可以计算迭代点 x^k 处的梯度：

$$\nabla f\left(x^k\right)=\begin{pmatrix} 2\left(1+|\epsilon_k|\right) \\ 2\left(1+|\epsilon_k|\right)t_k \end{pmatrix}=2\left(1+|\epsilon_k|\right)\begin{pmatrix} 1 \\ t_k \end{pmatrix},$$

其中 $t_k=\operatorname{sign}\left(\epsilon_k\right)$.

梯度下降法失败的例子

下面我们考虑直接用梯度下降法进行迭代.

- 在负梯度方向$-\nabla f\left(x^k\right)$ 上做精确线搜索，可得

$$x^{k+1}=x^k+\alpha_k\left(-\nabla f\left(x^k\right)\right)=\begin{bmatrix} 2\left(1+\left|\epsilon_k\right|/3\right) \\ -\epsilon_k/3 \end{bmatrix}=\begin{bmatrix} 2\left(1+\left|\epsilon_{k+1}\right|\right) \\ \epsilon_{k+1} \end{bmatrix}$$

  其中 $\epsilon_{k+1}=-\epsilon_k/3\neq 0$. 所以显然有$\epsilon_k\rightarrow 0$.

- 给定一个初始点 $x^0=(2+2|\delta|,\delta)^{\mathrm{T}}$，我们有 $x^k\rightarrow(2,0)^{\mathrm{T}}$.
- 然而 $(2,0)^{\mathrm{T}}$ 并不是稳定点.
- 这表明对非光滑问题直接使用梯度法可能会收敛到一个非稳定点.

提纲

问题设定

假设$f(x)$ 为凸函数，但不一定可微，考虑如下问题：

$$\min_{x} f(x)$$

- 一阶充要条件：

$$x^* \text{是一个全局极小点} \quad \Leftrightarrow \quad 0 \in \partial f(x^*)$$

- 因此可以通过计算凸函数的次梯度集合中包含0 的点来求解其对应的全局极小点.

为了极小化一个不可微的凸函数f，可类似梯度法构造如下次梯度算法的迭代格式：

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k),$$

其中$\alpha_k > 0$为步长．它通常有如下四种选择：

1. 固定步长$\alpha_k = \alpha$；
2. 固定$\|x^{k+1} - x^k\|$，即$\alpha_k \|g^k\|$ 为常数；
3. 消失步长$\alpha_k \to 0$且$\sum_{k=0}^{\infty} \alpha_k = +\infty$；
4. 选取α_k使其满足某种线搜索准则．

下面我们讨论在不同步长取法下次梯度算法的收敛性质．

提纲

假设条件

(1) f为凸函数；

(2) f至少存在一个有限的极小值点x^*，且$f(x^*) > -\infty$；

(3) f为利普希茨连续的，即

$$|f(x) - f(y)| \le G\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

其中$G > 0$为利普希茨常数.

我们下面证明这等价于$f(x)$的次梯度是有界的，即

$$\|g\| \le G, \quad \forall\, g \in \partial f(x), x \in \mathbb{R}^n.$$

证明：

- 充分性：假设$\|g\|_2 \le G, \forall\, g \in \partial f(x)$；取$g_y \in \partial f(y), g_x \in \partial f(x)$:

$$g_x^{\mathrm{T}}(x-y) \ge f(x) - f(y) \ge g_y^{\mathrm{T}}(x-y)$$

  再由柯西不等式

$$G\|x-y\|_2 \ge f(x) - f(y) \ge -G\|x-y\|_2$$

- 必要性：反设存在x和$g \in \partial f(x)$，使得$\|g\|_2 > G$；取$y = x + \frac{g}{\|g\|_2}$

$$\begin{aligned} f(y) &\ge f(x) + g^{\mathrm{T}}(y-x) \\ &= f(x) + \|g\|_2 \\ &> f(x) + G \end{aligned}$$

  这与$f(x)$是G-利普希茨连续的矛盾.

收敛性分析

- 次梯度方法不是一个下降方法，即无法保证$f(x^{k+1}) < f(x^k)$；
- 收敛性分析的关键是分析$f(x)$历史迭代的最优点所满足的性质.
- 设x^*是$f(x)$的一个全局极小值点，$f^* = f(x^*)$，根据迭代格式，

$$
\begin{aligned}
\left\|x^{i+1} - x^*\right\|^2 &= \left\|x^i - \alpha_i g^i - x^*\right\|^2 \\
&= \left\|x^i - x^*\right\|^2 - 2\alpha_i \left\langle g^i, x^i - x^* \right\rangle + \alpha_i^2 \left\|g^i\right\|^2 \\
&\leqslant \left\|x^i - x^*\right\|^2 - 2\alpha_i \left(f\left(x^i\right) - f^*\right) + \alpha_i^2 G^2
\end{aligned}
$$

- 结合$i = 0, \cdots, k$ 时相应的不等式，并定义$\hat{f}^k = \min_{0 \leqslant i \leqslant k} f\left(x^i\right)$：

$$
\begin{aligned}
2\left(\sum_{i=0}^{k} \alpha_i\right)\left(\hat{f}^k - f^*\right) &\leqslant \left\|x^0 - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2 \\
&\leqslant \left\|x^0 - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2
\end{aligned}
$$

不同步长下的收敛性

(1) 取$\alpha_i = t$为固定步长，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时, $\hat{f}^k$ 近似为$G^2t/2$-次优的

(2) 取α_i使得$\|x^{i+1} - x^i\|$固定，即$\alpha_i\|g^i\| = s$为常数，则

$$\hat{f}^k - f^* \leq \frac{G\|x^0 - x^*\|^2}{2ks} + \frac{Gs}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时, $\hat{f}^k$ 近似为$Gs/2$- 次优的

不同步长下的收敛性

(3) 取α_i为消失步长，即$\alpha_i \to 0$且$\sum_{i=0}^{\infty} \alpha_i = +\infty$，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + G^2 \sum\limits_{i=0}^{k} \alpha_i^2}{2 \sum\limits_{i=0}^{k} \alpha_i};$$

进一步可得$\hat{f}^k$收敛到f^*.

- 和梯度法不同，只有当α_k取消失步长时$\hat{f}^k$才具有收敛性.
- 一个常用的步长取法是$\alpha_k = \frac{1}{k}$.

提纲

例：LASSO 问题求解

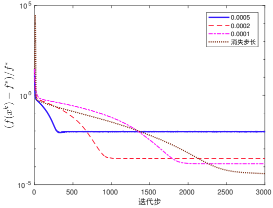
考虑LASSO 问题

$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1,$$

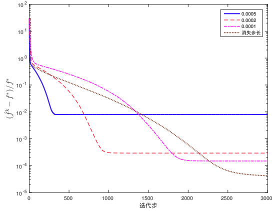
$f(x)$的一个次梯度为$g = A^{\mathrm{T}}(Ax - b) + \mu\mathrm{sign}(x)$, 其中$\mathrm{sign}(x)$是关于$x$逐分量的符号函数．因此的次梯度算法为

$$x^{k+1} = x^k - \alpha_k(A^{\mathrm{T}}(Ax^k - b) + \mu\mathrm{sign}(x^k)),$$

步长α_k可选为固定步长或消失步长．



(a) $f(x^k) - f^*$ 的相对变化



(b) $\hat{f}^k - f^*$的相对变化

提纲

投影次梯度法

考虑在凸集 C 上极小化一个不可微凸函数 f：

$$\min_x f(x) \quad \text{s.t. } x \in C$$

- 投影次梯度法和次梯度法的区别在于每步迭代都需要将迭代点投影到集合C 上：

$$x^{k+1} = P_C\left(x^k - \alpha_k g^k\right), \quad k = 0, 1, 2, \ldots$$

- 假设投影算子 P_C可以计算，那么投影次梯度法可以得到和次梯度法相同的收敛性保证.

投影次梯度法

容易计算投影的集合举例：

- 仿射变换的像：$\{Ax + b : x \in \mathbb{R}^n\}$
- 线性系统的解集：$\{x : Ax = b\}$
- 非负象限：$\mathbb{R}^n_+ = \{x : x \geq 0\}$
- 一些范数球：$\{x : \|x\|_p \leq 1\}, p = 1, 2, \infty$
- 一些简单的多面体和简单的锥

注意：存在很多简单的集合C，它们的投影算子 P_C 很难计算

例：任意一个多面体$C = \{x : Ax \leq b\}$

总结：次梯度法

- 能够处理一般的不可微凸函数
- 常能推导出非常简单的算法
- 收敛速度可能非常缓慢
- 理论复杂度：迭代$O(1/\epsilon^2)$ 步，得到ϵ-次优的点
- 一种“最优”的一阶方法：$O(1/\epsilon^2)$ 无法再改进