MATH 3011 运筹学 **秋季 2023**

Lecture 18: 次梯度和次梯度法

Lecturer: 陈士祥 *Scribes:* 陈士祥

1 次梯度

许多优化问题，目标函数都是不可微的，例如前面我们见到的基追踪问题和矩阵补全问题，目标函数分别是最小化 ℓ_1 范数和矩阵变量的核范数。为了研究不可微时问题的最优条件，我们可以定义一般非光滑凸函数的次梯度。

非光滑优化在多个其他领域都有应用，例如：

- 机器学习：L1 正则化就是一个典型的非光滑优化问题。
- 信号处理：稀疏信号恢复中常用的 L1 范数最小化。
- 经济学：在某些经济模型中，效用函数或成本函数可能是非光滑的。
- 深度学习：深度学习中，损失函数其实是非光滑的。由于其复杂性太高，故实际中常常忽略了这点。

1.1 次梯度的定义

回顾可微凸函数 f 的一阶等价条件:

$$f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x).$$

这表明，f 在点 x 处的一阶近似是 f 的一个全局下界。我们这里的想法是，将上述不等式拓展到一般不可微的情形。我们先考虑简单的函数 $f(x) = |x|, x \in \mathbb{R}$. $f(x)$ 在 $x = 0$ 处不可导，因为其左右导数分别为

$$\lim_{t \to 0^-} \frac{|t|}{t} = -1, \quad \lim_{t \to 0^+} \frac{|t|}{t} = 1.$$

可以验证, 对于任意 $g \in [-1, 1]$, 下面的不等式成立

$$|y| \geq 0 + g \cdot y,$$

此即

$$f(y) \geq f(0) + g \cdot (y - 0).$$

上面的不等式的几何意义为，任意斜率为 $g \in [-1,1]$ 过原点的直线，均为函数 f 的一个下界。

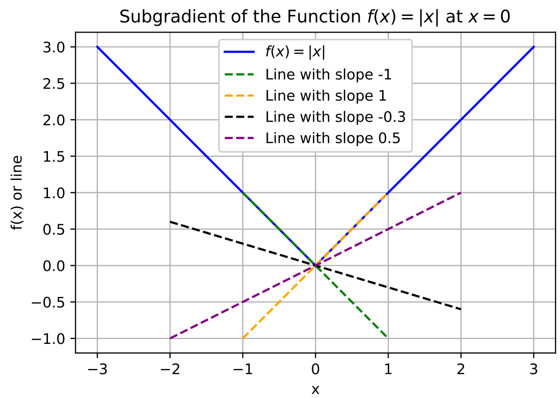


图 18.1: 函数 $f(x) = |x|$ 次梯度示意图。任意斜率为 $g \in [-1,1]$ 过原点的直线，均为函数 f 的一个下界。

我们引出如下定义。

Definition 18.1 (次梯度和次微分) 设 f 为适当凸函数，x 为定义域 $\mathbf{dom}\ f$ 中的一点. 若向量 $g \in \mathbb{R}^n$ 满足

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

则称 g 为函数 f 在点 x 处的一个**次梯度** *(subgradient)*. 进一步地，称集合

$$\partial f(x) = \{g \,|\, g \in \mathbb{R}^n, f(y) \geq f(x) + g^{\mathrm{T}}(y-x), \forall y \in \mathbf{dom}\ f\}$$

为 f 在点 x 处的**次微分** *(subdifferential)*.

注 18.1

- 定义中的凸函数，值域可以为广义实数 $\mathbb{R} \cup \{+\infty\}$ 空间。适当函数是指，存在 x 使得 $f(x) < +\infty$.
- 由定义可知，次微分是一个集合，次梯度是某个次微分的元素。

Example 18.1 $f(x) = \|x\|_2$ 为凸函数。若 $x \neq 0$, $f(x)$ 可微，故

$$\partial f(x) = \frac{1}{\|x\|_2} x.$$

若 $x = 0$, 我们下面证明 $\partial f(x) = \{g | \|g\|_2 \leq 1\}$. 由定义可知，

$$\|y\|_2 \geq g^T y, \quad \forall y.$$

首先，若 $\|g\|_2 \leq 1$, 由 *Cauchy* 不等式得 $g^T y \leq \|g\|_2 \|y\|_2 \leq \|y\|_2$, 故

$$\{g | \|g\|_2 \leq 1\} \subset \partial f(0).$$

反之，若 $g \in \partial f(0)$, 故

$$\max_{\|y\|_2=1} g^T y = \|g\|_2 \le \|y\|_2 = 1.$$

故

$$\partial f(0) \subset \{g | \|g\|_2 \le 1\}.$$

1.2 次梯度存在性

为了说明定义 18.1中的次梯度对于一般凸函数存在，我们引入如下定义。

Definition 18.2 设 $f(x)$ 为 $\mathbb{R}^n$ 上的实值函数，函数的上方图 **epi** f 定义为

$$\textbf{epi } f := \{ \begin{bmatrix} x \\ z \end{bmatrix} \in \mathbb{R}^{n+1} \mid z \ge f(x) \}.$$

Lemma 18.1 函数 $f(x)$ 是凸函数，当且仅当其上方图是凸集。

当 f 可微时，我们有

$$f(x) + \nabla f(x)^T (y - x) \le f(y) \le z.$$

即

$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ z \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0 \quad \forall\ (y, z) \in \textbf{epi } f$$

这表明，$\nabla f(x)$ 可以诱导出上方图 **epi** f 在点 $(x, f(x))$ 处的支撑超平面, 如下图所示。

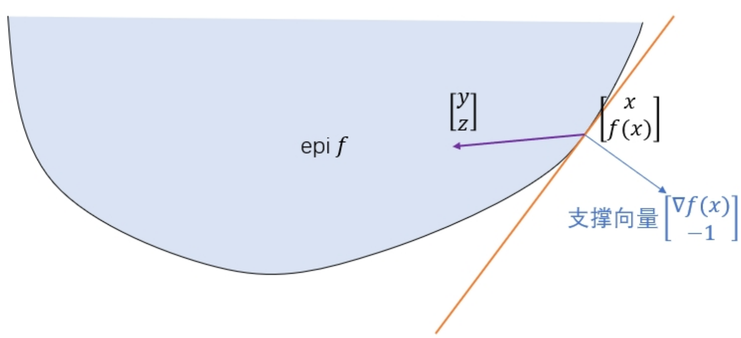


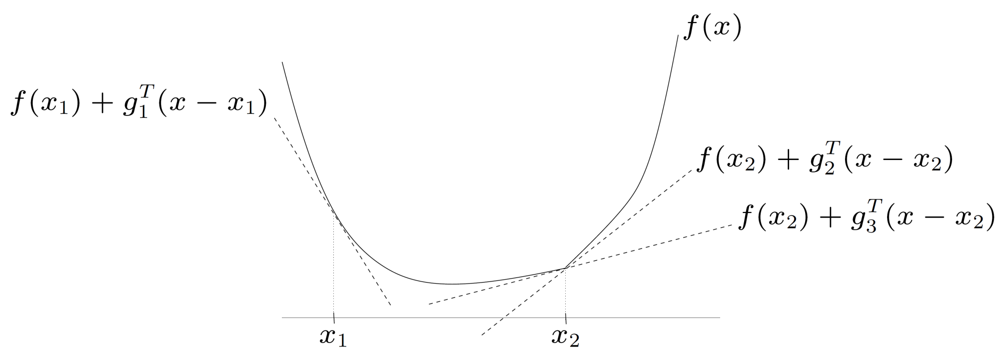
图 18.2: 对于凸函数 $f(x)$, 其上方图 **epi** f 是一个凸集。$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ 是 **epi** f 的支撑向量。

由次梯度的定义 18.1可知，

- $f(x)+g^{\mathrm{T}}(y-x)$ 是 $f(y)$ 的一个全局下界
- g 可以诱导出上方图 **epi** f 在点 $(x,f(x))$ 处的一个支撑超平面

$$\begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ z \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0 \quad \forall\ (y,z) \in \textbf{epi } f$$

- 如果 f 是可微凸函数, 那么 $\nabla f(x)$ 是 f 在点 x 处的一个次梯度
- 例：g_2, g_3 是点 x_2 处的次梯度；g_1 是点 x_1 处的次梯度



图片来源：《最优化计算方法》文再文等讲义。

次梯度的存在性主要依赖于凸集的下述性质：

Lemma 18.2 任意凸集的边界点处都存在支撑超平面。

Theorem 18.1 设 f 为凸函数, **dom** $f=\{x: f(x)<\infty\}$ 为其定义域. 如果 $x \in$ **int dom** f, 则 $\partial f(x)$ 是非空的, 其中 **int dom** f 的含义是集合 **dom** f 的所有内点.

Proof: $(x,f(x))$ 是 **epi** f 边界上的点. 因此凸集 **epi** f 在点 $(x,f(x))$ 处存在支撑超平面:

$$\exists\ (a,b) \neq 0, \quad \begin{bmatrix} a \\ b \end{bmatrix}^{\mathrm{T}} \left(\begin{bmatrix} y \\ z \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0 \quad \forall\ (y,z) \in \textbf{epi } f$$

令 $z \to +\infty$, 可知 $b \le 0$. 由于 $x \in$ **int dom** f, 取 $y = x+\epsilon a \in$ **dom** f, $\epsilon>0$, 可知 $b \neq 0$.

因此 $b<0$, 令 $g = a/|b|$, 则 g 是 f 在点 x 处的次梯度。

■

Example 18.2 (反例) 如下函数在点 $x=0$ 处不是次可微的*:*

- $f : \mathbf{R} \to \mathbf{R}, \mathbf{dom}\ f = \mathbf{R}_+ = \{x \in \mathbf{R} \mid x \geq 0\}$

$$x = 0 \text{ 时，} f(x) = 1, x > 0 \text{ 时，} f(x) = 0$$

- $f : \mathbf{R} \to \mathbf{R}, \mathbf{dom}\ f = \mathbf{R}_+$

$$f(x) = -\sqrt{x}$$

$\mathbf{epi}\ f$ 在点 $(0, f(0))$ 处的唯一支撑超平面是垂直的

1.3 次梯度的计算法则

- **可微凸函数：** 若凸函数 f 在点 x 处可微，则 $\partial f(x) = \{\nabla f(x)\}$.

- **凸函数的非负线性组合：** 设凸函数 f_1, f_2 满足 $\mathbf{int}\ \mathbf{dom}\ f_1 \cap \mathbf{dom}\ f_2 \neq \emptyset$，而 $x \in \mathbf{dom}\ f_1 \cap \mathbf{dom}\ f_2$. 若

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x), \quad \alpha_1, \alpha_2 \geq 0,$$

  则 $f(x)$ 的次微分

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x).$$

- **线性变量替换：** 设 h 为适当凸函数，f 满足 $f(x) = h(Ax + b)$. 若存在 $x^\sharp \in \mathbb{R}^m$，使得 $Ax^\sharp + b \in \mathbf{int}\ \mathbf{dom}\ h$，则

$$\partial f(x) = A^{\mathrm{T}} \partial h(Ax + b), \quad \forall\ x \in \mathbf{int}\ \mathbf{dom}\ f.$$

- **多个函数取上界：** 设 $f_1, f_2, \cdots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ 均为凸函数，令

$$f(x) = \max\{f_1(x), f_2(x), \cdots, f_m(x)\}, \quad \forall x \in \mathbb{R}^n.$$

  对 $x_0 \in \bigcap_{i=1}^{m} \mathbf{int}\ \mathbf{dom}\ f_i$，定义 $I(x_0) = \{i \mid f_i(x_0) = f(x_0)\}$，则

$$\partial f(x_0) = \mathbf{conv} \bigcup_{i \in I(x_0)} \partial f_i(x_0).$$

  - $I(x_0)$ 表示点 x_0 处"有效"函数的指标
  - $\mathbf{conv}S$ 表示集合 S 的凸包, 即集合 S 的所有点的凸组合构成的点集.
  - $\partial f(x_0)$ 是点 x_0 处"有效"函数的次微分并集的凸包
  - 如果 f_i 可微, $\partial f(x_0) = \mathbf{conv}\{\nabla f_i(x_0) \mid i \in I(x_0)\}$

- **固定分量的函数极小值**

$$f(x) = \inf_y h(x,y), \quad h \text{ 关于 } (x,y) \text{ 联合凸}$$

计算点 $\hat{x}$ 处的一个次梯度:

  - 设 $\hat{y} \in \mathbb{R}^m$ 满足 $h(\hat{x},\hat{y}) = f(\hat{x})$, 即 $\hat{y}$ 是固定 $x = \hat{x}$ 后 $h(\hat{x},y)$ 的极小解。
  - 存在 $g \in \mathbb{R}^n$ 使得 $(g,0) \in \partial h(\hat{x},\hat{y})$, 则 $g \in \partial f(\hat{x})$, 即若 $g \in \partial_x h(\hat{x},\hat{y})$, 则 $g \in \partial f(\hat{x})$.

证明: 对任意 $x \in \mathbb{R}^n, y \in \mathbb{R}^m$

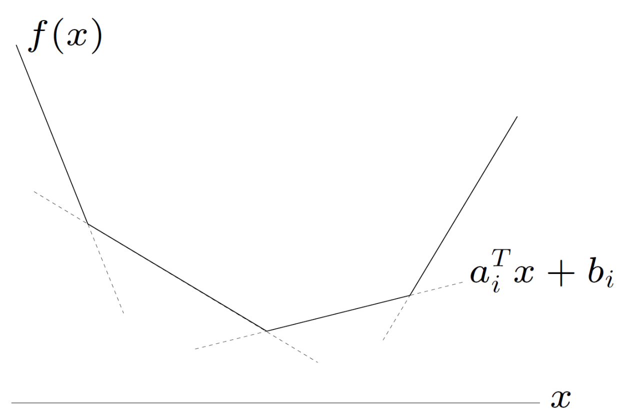
$$\begin{aligned} h(x,y) &\geq h(\hat{x},\hat{y}) + g^{\mathrm{T}}(x-\hat{x}) + 0^{\mathrm{T}}(y-\hat{y}) \\ &= f(\hat{x}) + g^{\mathrm{T}}(x-\hat{x}). \end{aligned}$$

于是

$$f(x) = \inf_y h(x,y) \geq f(\hat{x}) + g^{\mathrm{T}}(x-\hat{x})$$

Example 18.3 分段线性函数

$$f(x) = \max_{i=1,2,\cdots,m} \{a_i^{\mathrm{T}} x + b_i\}$$



- 点 x 处的次微分是一个多面体

$$\partial f(x) = \mathbf{conv}\{a_i \mid i \in I(x)\}$$

其中 $I(x) = \{i \mid a_i^{\mathrm{T}} x + b_i = f(x)\}$

Example 18.4 鲁棒线性回归: 求函数 $f(x) = \|Ax-b\|_1$ 的次微分, 这里 $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$。

解: 首先考虑函数 $h(y) = \|y\|_1 = \max_{s\in\{-1,1\}^m} s^T y, y \in \mathbb{R}^m$. 故,

$$\partial h(y) = J_1 \times \cdots \times J_m, \quad J_k = \begin{cases} [-1,1], & y_k = 0 \\ \{1\}, & y_k > 0 \\ \{-1\}, & y_k < 0 \end{cases}$$

对于 $f(x)$,

$$\partial f(x) = A^T \left(\partial h(y)|_{y=Ax-b}\right).$$

Example 18.5 设 C 是 $\mathbb{R}^n$ 中一闭凸集，令

$$f(x) = \min_{y\in C} \|x - y\|_2$$

计算点 $\hat{x}$ 处的一个次梯度:

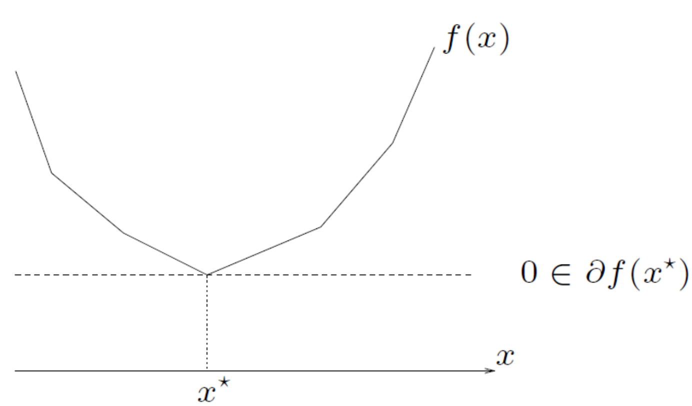
- 若 $f(\hat{x}) = 0$，则容易验证 $g = 0 \in \partial f(\hat{x})$;
- 若 $f(\hat{x}) > 0$，取 $\hat{y}$ 为 $\hat{x}$ 在 C 上的投影，即 $\hat{y} = \mathcal{P}_c(\hat{x})$，计算
$$g = \frac{1}{\|\hat{x} - \hat{y}\|_2}(\hat{x} - \hat{y}) = \frac{1}{\|\hat{x} - \mathcal{P}_c(\hat{x})\|_2}(\hat{x} - \mathcal{P}_c(\hat{x})).$$

1.4 最优性条件

Theorem 18.2 对于凸函数 $f(x)$, x^* 是 $f(x)$ 的全局极小点当且仅当

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

Proof:



图片来源：《最优化计算方法》，文再文等讲义。

根据次梯度的定义以及最优性，我们有

$$f(y) \geq f(x^*),\ \forall y \quad \Longleftrightarrow f(y) \geq f(x^*) + 0^{\mathrm{T}}(y - x^*),\ \forall y \quad \Longleftrightarrow \quad 0 \in \partial f(x^*).$$

■

Example 18.6 (分段线性函数最优条件)

$$f(x) = \max_{i=1,\cdots,m}(a_i^{\mathrm{T}} x + b_i)$$

- **最优性条件**

$$0 \in \mathbf{conv}\{a_i \mid i \in I(x^*)\}, \quad 其中\ I(x) = \{i \mid a_i^{\mathrm{T}} x + b_i = f(x)\}$$

- 也就是说，x^* 是最优解当且仅当存在 λ 使得

$$\lambda \geq 0, \quad \mathbf{1}^{\mathrm{T}}\lambda = 1, \quad \sum_{i=1}^{m} \lambda_i a_i = 0, \quad \lambda_i = 0 \text{ for } i \notin I(x^*)$$

- 这是等价（思考：为何等价？）线性规划问题的最优性条件：$A = [a_1^{\top}; \ldots; a_m^{\top}]$

$$\begin{aligned} \min \quad & t \\ \text{s.t.} \quad & Ax + b \leq t\mathbf{1}. \end{aligned}$$

  其对偶为

$$\begin{aligned} \max \quad & b^{\mathrm{T}}\lambda \\ \text{s.t.} \quad & A^{\mathrm{T}}\lambda = 0, \\ & \lambda \geq 0, \quad \mathbf{1}^{\mathrm{T}}\lambda = 1, \end{aligned}$$

  强对偶定理为：原始可行：

$$Ax + b \leq t\mathbf{1} \Longleftrightarrow f(x) = \max_{i=1,\cdots,m}(a_i^{\mathrm{T}} x + b_i)$$

  互补松弛

$$\lambda^T(Ax + b - t\mathbf{1}) = 0 \quad \Longleftrightarrow \lambda_i = 0 \text{ for } i \notin I(x^*),$$

  对偶可行

$$A^{\mathrm{T}}\lambda = 0, \quad \lambda \geq 0, \quad \mathbf{1}^{\mathrm{T}}\lambda = 1.$$

作业 18.1 计算下面两个问题的一个次梯度

- $f(x) = \|Ax - b\|_2$.
- $f(x) = \min_y \|Ay - x\|_\infty$，这里 $\|x\|_\infty = \max_i |x_i|$ 表示无穷范数。假设存在 $\hat{y}$ 使得 $f(\hat{x}) = \min_y \|Ay - x\|_\infty$. 计算一个 $f(\hat{x})$ 的次梯度。

1.5 约束问题的次梯度和最优条件

给定约束 C 为 $\mathbb{R}^n$ 中的闭凸集，考虑问题

$$\begin{aligned} &\min f(x) \\ &\text{s.t.} \quad x \in C. \end{aligned} \tag{18.1}$$

可定义指示函数

$$\mathcal{I}_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

则问题(18.1)等价于

$$\min h(x) := f(x) + \mathcal{I}_C(x). \tag{18.2}$$

对于问题(18.2), 最优条件为

$$0 \in \partial f(x) + \partial \mathcal{I}_C(x), x \in C.$$

这里，若 $g \in \partial \mathcal{I}_C(x)$, 则 $\mathcal{I}_C(y) \geq \mathcal{I}_C(x) + g^T(y-x), \forall y \in C$, 即

$$0 \geq g^T(y-x), \quad \forall y \in C.$$

这里说明次梯度在 x 处的法锥中，法锥的定义为 $N_C(x) = \{g \mid 0 \geq g^T(y-x), \forall y \in C\}$. 事实上，法锥与切锥不相交，故这表明次梯度不在切锥中，与前面的课程一致。

一般来说，非光滑约束问题也有 KKT 条件。由于课程的设置，我们不再学习它们。

2 次梯度算法

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, g^k 为 x_k 处函数 f 任意的一个次梯度. 它通常有如下四种选择：

1. 固定步长 $\alpha_k = \alpha$；

2. 消失步长 $\alpha_k \to 0$ 且 $\sum_{k=0}^{\infty} \alpha_k = +\infty$；

下面我们讨论在不同步长取法下次梯度算法的收敛性质.

2.1 次梯度法的收敛结论

Theorem 18.3 假设 f 满足如下性质：

(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

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

其中 $G > 0$ 为利普希茨常数.

这等价于 $f(x)$ 的次梯度是有界的 (我们略过证明)，即

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

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

$$\hat{f}^k = \min_{0 \leq i \leq k} f(x_i),$$

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^*.

Proof: 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 $\hat{f}^k = \min_{0 \leqslant i \leqslant k} f\left(x^i\right)$：

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

■

注 18.2 • 若用常数步长 $\alpha_k = t$, 则

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

  – $\hat{f}^k$ 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

• 次梯度方法不是一个下降方法, 即无法保证 $f(x^{k+1}) < f(x^k)$

3 应用：LASSO 问题

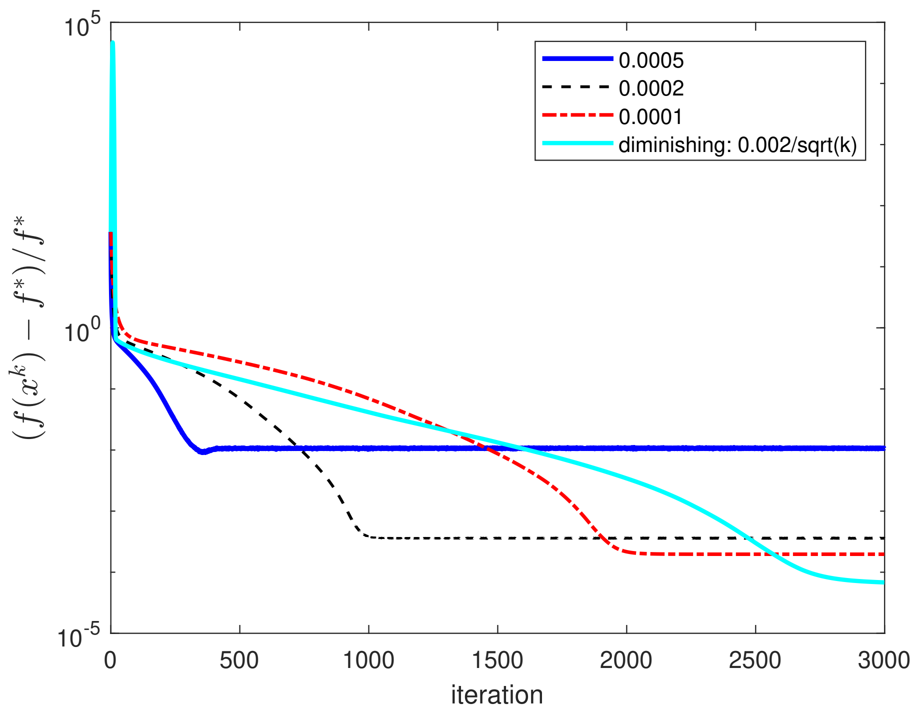
考虑 LASSO 问题

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

$f(x)$ 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. 图中, 选取 $\alpha_k = 0.0001, 0.0002, 0.0005$ 等固定步长, 到一定精度



无法继续下降。选取 $\alpha_k = \frac{0.002}{\sqrt{k}}$ 的 diminishing 步长求解精度更高。

参考资料：刘浩洋, 户将, 李勇锋，文再文，最优化：建模、算法与理论, 高教出版社，ISBN：9787040550351
H. Liu, J. Hu, Y. Li, Z. Wen, Optimization: Modeling, Algorithm and Theory (in Chinese)

Boyd, Stephen P., and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.