Artificial Intelligence

5. 从线性模型到统计学习框架

罗晓鹏

xpluo@nju.edu.cn

工管 · 南京大学 · 2022 秋

目录

统计学习导言

统计学习导言

- 统计学习与自然语言

- 统计学习与自然语言
- 统计学习与模式识别

统计学习导言

- 统计学习与自然语言
- 统计学习与模式识别
- 统计学习与元学习

统计学习导言

- 统计学习与自然语言
- 统计学习与模式识别
- 统计学习与元学习
- 统计学习中的法律问题

从数据拟合到监督学习

单变量最小二乘拟合问题

针对数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，考虑

(1) q 阶多项式模型：

$$h_w(x) = \sum_{i=0}^{q} w_i x^i$$

(2) 最小二乘策略：

$$f(w) = \frac{1}{M} \sum_{j=1}^{M} \Big(h_w(X_j) - Y_j \Big)^2$$

(3) 训练算法：

$$w_* = \arg \min_{w \in \mathbb{R}^{q+1}} f(w)$$

单变量最小二乘拟合问题

针对数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，考虑

(1) q 阶多项式模型：

$$h_w(x) = \sum_{i=0}^{q} w_i x^i$$

(2) 最小二乘策略：

$$f(w) = \frac{1}{M}\sum_{j=1}^{M}\Big(h_w(X_j) - Y_j\Big)^2$$

(3) 训练算法：

$$w_* = \arg\min_{w\in\mathbb{R}^{q+1}} f(w)$$

单变量最小二乘拟合问题

针对数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，考虑

(1) q 阶多项式模型：

$$h_w(x) = \sum_{i=0}^{q} w_i x^i$$

(2) 最小二乘策略：

$$f(w) = \frac{1}{M} \sum_{j=1}^{M} \Big(h_w(X_j) - Y_j \Big)^2$$

(3) 训练算法：

$$w_* = \arg \min_{w \in \mathbb{R}^{q+1}} f(w)$$

单变量最小二乘拟合问题

针对数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，考虑

(1) q 阶多项式模型：

$$h_w(x) = \sum_{i=0}^{q} w_i x^i$$

(2) 最小二乘策略：

$$f(w) = \frac{1}{M}\sum_{j=1}^{M}\Big(h_w(X_j) - Y_j\Big)^2$$

(3) 训练算法：

$$w_* = \arg\min_{w\in\mathbb{R}^{q+1}} f(w)$$

监督学习

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

监督学习

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

监督学习

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

监督学习

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

拓展问题：

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^M$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

拓展问题：

- 问题本身：可学习性、数据需求下界、关联与隶属……

监督学习

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

拓展问题：

- 问题本身：可学习性、数据需求下界、关联与隶属……
- 数据相关：获取、质量、迁移……

监督学习问题

针对一组给定的数据 $\{(X_j, Y_j)\}_{j=1}^{M}$，通过步骤：

(1) 模型：选择一类“合适”模型，

(2) 策略：确定一个准则以定义“最优”模型，

(3) 算法：求解相应策略定义的“最优”模型；

建立能够“充分”表达数据信息的模型.

拓展问题：

- 问题本身：可学习性、数据需求下界、关联与隶属……
- 数据相关：获取、质量、迁移……
- 智能模式：给定模型、策略和算法框架所蕴含的实质

监督学习的模型

监督学习的模型

- 网络模型：全连接、卷积、循环、对抗生成、……

监督学习的模型

- 网络模型：全连接、卷积、循环、对抗生成、……
- 传统模型：核模型 (如SVM)，树模型 (如决策树)，……

- 网络模型：全连接、卷积、循环、对抗生成、……
- 传统模型：核模型 (如SVM)，树模型 (如决策树)，……

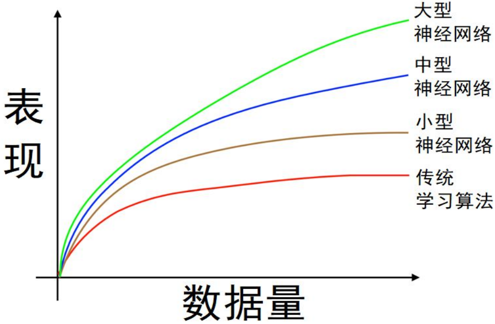


Fig 1: 引自吴恩达《Machine Learning Yearning》

监督学习的策略

监督学习的误差度量

定义 1 (损失函数)

对于模型 $h_w(x)$，一个损失函数 $L := L(h_w(X_j), Y_j)$ 是 $h_w(x)$ 在数据对 (X_j, Y_j) 处的偏差大小的一种度量.

定义 1 (损失函数)

对于模型 $h_w(x)$，一个损失函数 $L := L(h_w(X_j), Y_j)$ 是 $h_w(x)$ 在数据对 (X_j, Y_j) 处的偏差大小的一种度量.

定义 2 (经验风险)

给定一个目标 $y(x)$，一个数据集 $\{(X_j, Y_j)\}_{j=1}^{M}$，以及一个模型 $h(x)$，则 $h(x)$ 对于给定损失函数 $L(h(X_j), Y_j)$ 的经验风险被定义为

$$\hat{R}_M(h) = \frac{1}{M}\sum_{j=1}^{M} L(h(X_j), Y_j).$$

定义 3 (泛化误差)

给定一个目标 $y(x)$ 且 (x, y) 服从密度为 $p(x, y)$ 的分布，$h(x)$ 是目标的一个模型，则 $h(x)$ 对于给定损失函数 $L(h(X_j), Y_j)$ 的泛化误差被定义为

$$R(h) = \int_{XY} L(h(x), y)p(x, y)\mathrm{d}x\mathrm{d}y.$$

监督学习的误差度量

定义 3 (泛化误差)

给定一个目标 $y(x)$ 且 (x,y) 服从密度为 $p(x,y)$ 的分布，$h(x)$ 是目标的一个模型，则 $h(x)$ 对于给定损失函数 $L(h(X_j),Y_j)$ 的泛化误差被定义为

$$R(h) = \int_{XY} L(h(x),y)p(x,y)\mathrm{d}x\mathrm{d}y.$$

定理 1 (回顾·统计收敛性)

设 $\{(X_j,Y_j)\}_{j=1}^{M}$ 来自于方差有限的分布 $P(x,y)$，则

$$\hat{R}_M = R + \mathcal{O}\Big(M^{-\frac{1}{2}}\Big).$$

定理 2 (回归函数的性质)

若选择平方损失函数 $L(h(x),y)=(h(x)-y)^2$，则回归函数

$$\mathbb{E}[y|x]=\frac{\int_Y y(x)p(x,y)\mathrm{d}y}{\int_Y p(x,y)\mathrm{d}y}$$

最小化泛化误差

$$R(h)=\int_{XY}(h(x)-y)^2 p(x,y)\mathrm{d}x\mathrm{d}y.$$

定理 2 (回归函数的性质)

若选择平方损失函数 $L(h(x), y) = (h(x) - y)^2$，则回归函数

$$\mathbb{E}[y|x] = \frac{\int_Y y(x)p(x,y)\mathrm{d}y}{\int_Y p(x,y)\mathrm{d}y}$$

最小化泛化误差

$$R(h) = \int_{XY} (h(x) - y)^2 p(x,y)\mathrm{d}x\mathrm{d}y.$$

Proof.

(板书·选讲) □

定理 3 (泛化误差的分解)

若 $L(h(x), y) = (h(x) - y)^2$，则泛化误差满足以下分解：

$$R(h) = \int_X \Big(h(x) - \mathbb{E}[y|x]\Big)^2 p(x)\mathrm{d}x + \int_{XY} \Big(y - \mathbb{E}[y|x]\Big)^2 p(x, y)\mathrm{d}x\mathrm{d}y.$$

定理 3 (泛化误差的分解)

若 $L(h(x),y)=(h(x)-y)^2$，则泛化误差满足以下分解：

$$R(h)=\int_X\Big(h(x)-\mathbb{E}[y|x]\Big)^2p(x)\mathrm{d}x+\int_{XY}\Big(y-\mathbb{E}[y|x]\Big)^2p(x,y)\mathrm{d}x\mathrm{d}y.$$

Proof.

提示：$\int_{XY}(h(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y)p(x,y)\mathrm{d}x\mathrm{d}y=0.$ □

泛化误差的分解

定理 3 (泛化误差的分解)

若 $L(h(x),y)=(h(x)-y)^2$，则泛化误差满足以下分解：

$$R(h)=\int_X\Big(h(x)-\mathbb{E}[y|x]\Big)^2p(x)\mathrm{d}x+\int_{XY}\Big(y-\mathbb{E}[y|x]\Big)^2p(x,y)\mathrm{d}x\mathrm{d}y.$$

Proof.

提示：$\int_{XY}(h(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y)p(x,y)\mathrm{d}x\mathrm{d}y=0.$ □

推论 1

$$\min_h R(h)=\int_X\Big(y-\mathbb{E}[y|x]\Big)^2p(x)\mathrm{d}x.$$

定理 3 (泛化误差的分解)

若 $L(h(x),y)=(h(x)-y)^2$，则泛化误差满足以下分解：

$$R(h)=\int_X\Big(h(x)-\mathbb{E}[y|x]\Big)^2p(x)\mathrm{d}x+\int_{XY}\Big(y-\mathbb{E}[y|x]\Big)^2p(x,y)\mathrm{d}x\mathrm{d}y.$$

Proof.

提示：$\int_{XY}(h(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y)p(x,y)\mathrm{d}x\mathrm{d}y=0.$ □

推论 1

$$\min_h R(h)=\int_X\Big(y-\mathbb{E}[y|x]\Big)^2p(x)\mathrm{d}x.$$

注记 1

监督学习的典型问题之一：偏差与方差的权衡 - $\mathbb{E}_{\mathcal{D}}[h(x,D)]$.

(小规模)监督学习的梯度算法

最小化问题的设置

- 给定 $f:\Omega\subset\mathbb{R}^d\to\mathbb{R}$
- 存在 $x_*\in\Omega$ 使得 $f(x_*)\leqslant f(y),\ \forall y\in\Omega$，即

$$x_*=\arg\min_{y\in\Omega} f(y)$$

一般问题的迭代法

最小化问题的设置

- 给定 $f:\Omega\subset\mathbb{R}^d\to\mathbb{R}$
- 存在 $x_*\in\Omega$ 使得 $f(x_*)\leqslant f(y),\ \forall y\in\Omega$，即

$$x_*=\arg\min_{y\in\Omega} f(y)$$

最小化问题的设置

- 给定 $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$
- 存在 $x_* \in \Omega$ 使得 $f(x_*) \leqslant f(y),\ \forall y \in \Omega$，即

$$x_* = \arg\min_{y \in \Omega} f(y)$$

一般问题的迭代法

最小化问题的设置

- 给定 $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$
- 存在 $x_* \in \Omega$ 使得 $f(x_*) \leqslant f(y), \forall y \in \Omega$，即

$$x_* = \arg\min_{y \in \Omega} f(y)$$

一般目标

构造序列 $\{x_t\}$ 收敛到 x_*，其中，x_* 可以是任何期待的解.

一般问题的迭代法

最小化问题的设置

- 给定 $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$
- 存在 $x_* \in \Omega$ 使得 $f(x_*) \leqslant f(y)$, $\forall y \in \Omega$，即

$$x_* = \arg\min_{y \in \Omega} f(y)$$

一般目标

构造序列 $\{x_t\}$ 收敛到 x_*，其中，x_* 可以是任何期待的解.

思考

给定 $t \in \mathbb{N}$，什么条件能使得 $\|x_{t+1} - x_*\| < \|x_t - x_*\|$ 成立?

定理 4

对于 $\mathbb{R}^n$ 内任意的序列 $\{x_t\}$ 和一点 x_*, 当 $t \in \mathbb{N}$ 给定时,

$$\|x_{t+1} - x_t\|_2 < 2\|x_t - x_*\|_2 \cos\theta,$$

等价于

$$\|x_{t+1} - x_*\|_2 < \|x_t - x_*\|_2,$$

其中, $\theta \in (-\pi/2, \pi/2)$ 是向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角.

定理 4

对于 $\mathbb{R}^n$ 内任意的序列 $\{x_t\}$ 和一点 x_*, 当 $t \in \mathbb{N}$ 给定时,

$$\|x_{t+1} - x_t\|_2 < 2\|x_t - x_*\|_2 \cos\theta,$$

等价于

$$\|x_{t+1} - x_*\|_2 < \|x_t - x_*\|_2,$$

其中, $\theta \in (-\pi/2, \pi/2)$ 是向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角.

提示

考虑由 x_*, x_t, x_{t+1} 三点确定的三角形，思考<u>方向</u>与<u>步长</u>.

定理 4

对于 $\mathbb{R}^n$ 内任意的序列 $\{x_t\}$ 和一点 x_*, 当 $t \in \mathbb{N}$ 给定时,

$$\|x_{t+1} - x_t\|_2 < 2\|x_t - x_*\|_2 \cos\theta,$$

等价于

$$\|x_{t+1} - x_*\|_2 < \|x_t - x_*\|_2,$$

其中, $\theta \in (-\pi/2, \pi/2)$ 是向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角.

提示

考虑由 x_*, x_t, x_{t+1} 三点确定的三角形，思考<u>方向</u>与<u>步长</u>.

Proof.

(练习) $\square$

迭代函数生成的序列

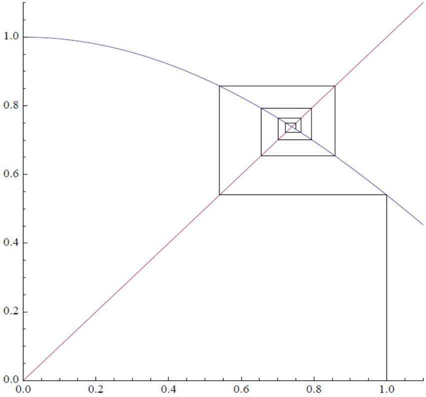
定义 4 (回顾·不动点迭代)

令 $x_0 \in \mathbb{R}^n$ 且 $T : \mathbb{R}^n \to \mathbb{R}^n$，映射 T 对应的迭代序列 $\{x_t\}$ 由迭代公式 $x_{t+1} = T(x_t)$ 递归地定义.

迭代函数生成的序列

定义 4 (回顾·不动点迭代)

令 $x_0 \in \mathbb{R}^n$ 且 $T : \mathbb{R}^n \to \mathbb{R}^n$，映射 T 对应的迭代序列 $\{x_t\}$ 由迭代公式 $x_{t+1} = T(x_t)$ 递归地定义.



下降方法的一般框架

定义 5 (下降方法的一般框架)

令 $x_0 \in \mathbb{R}^n$ 且 $f : \mathbb{R}^n \to \mathbb{R}^n$，迭代公式

$$x_{t+1} = x_t + \alpha_t d_t,$$

定义的迭代序列 $\{x_t\}$ 被称为一个下降序列，如果

$$f(x_{t+1}) < f(x_t), \ \forall t \in \mathbb{N},$$

其中，$\alpha_t > 0$ 称为步长，单位向量 d_t 称为方向.

方向选择简述

负梯度方向

- (Traced to) Augustin Louis Cauchy '1847
- Bernhard Riemann '1892
- Peter Debye '1909

负随机梯度方向

- H Robbins and S Monro '1951
- 2000$\sim$

角与内积$_1$

(回顾)向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角

(回顾)向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角

定义 6 (内积)

任取 $a, b \in \mathbb{R}^n$，两者的内积定义为

$$a \cdot b = \sum_{i=1}^{n} a_i b_i, \quad \text{特别地，} |a| = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} = \sqrt{a \cdot a}.$$

角与内积$_1$

(回顾)向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角

定义 6 (内积)

任取 $a, b \in \mathbb{R}^n$，两者的<u>内积</u>定义为

$$a \cdot b = \sum_{i=1}^{n} a_i b_i, \quad \text{特别地，} |a| = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} = \sqrt{a \cdot a}.$$

命题

任取 $a, b \in \mathbb{R}^n$, $|a \pm b|^2 = |a|^2 \pm 2a \cdot b + |b|^2$.

角与内积$_1$

(回顾)向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角

定义 6 (内积)

任取 $a, b \in \mathbb{R}^n$，两者的内积定义为

$$a \cdot b = \sum_{i=1}^{n} a_i b_i, \quad \text{特别地，} |a| = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} = \sqrt{a \cdot a}.$$

命题

任取 $a, b \in \mathbb{R}^n$, $|a \pm b|^2 = |a|^2 \pm 2a \cdot b + |b|^2$.

Proof.

(练习) □

(回顾)向量 $x_t - x_*$ 与 $x_{t+1} - x_t$ 之间的夹角

定义 6 (内积)

任取 $a, b \in \mathbb{R}^n$，两者的内积定义为

$$a \cdot b = \sum_{i=1}^{n} a_i b_i, \quad \text{特别地,} \quad |a| = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} = \sqrt{a \cdot a}.$$

命题

任取 $a, b \in \mathbb{R}^n$, $|a \pm b|^2 = |a|^2 \pm 2a \cdot b + |b|^2$.

Proof.

(练习) □

后面会用到：$-2a \cdot b = |a|^2 + |b|^2 - |a + b|^2$ (内积与模长的关系).

角与内积$_2$

命题 (内积与夹角的关系)

若 $a, b \in \mathbb{R}^n$，则

$$a \cdot b = a^{\mathrm{T}} b = |a||b| \cos\theta, \quad \theta \text{ 为 } a, b \text{ 间夹角}.$$

命题 (内积与夹角的关系)

若 $a, b \in \mathbb{R}^n$，则

$$a \cdot b = a^{\mathrm{T}} b = |a||b| \cos\theta, \quad \theta \text{ 为 } a, b \text{ 间夹角}.$$

Proof.

令 $\overrightarrow{AB} = (b_1 - a_1, \cdots, b_n - a_n)$, 由余弦定理

$$\begin{aligned} |a||b| \cos\theta = & \frac{1}{2}(|a|^2 + |b|^2 - |\overrightarrow{AB}|^2) \\ = & \frac{1}{2} \sum_{i=1}^{n} [a_i^2 + b_i^2 - (b_i - a_i)^2] = \sum_{i=1}^{n} a_i b_i. \end{aligned}$$

$\square$

角与内积$_2$

命题 (内积与夹角的关系)

若 $a, b \in \mathbb{R}^n$，则

$$a \cdot b = a^{\mathrm{T}} b = |a||b| \cos\theta, \quad \theta \text{ 为 } a, b \text{ 间夹角}.$$

Proof.

令 $\overrightarrow{AB} = (b_1 - a_1, \cdots, b_n - a_n)$, 由余弦定理

$$\begin{aligned} |a||b|\cos\theta =& \frac{1}{2}(|a|^2 + |b|^2 - |\overrightarrow{AB}|^2) \\ =& \frac{1}{2}\sum_{i=1}^{n}[a_i^2 + b_i^2 - (b_i - a_i)^2] = \sum_{i=1}^{n} a_i b_i. \end{aligned}$$

□

推论 (Cauchy 不等式)

令 $a, b \in \mathbb{R}^n$，有 $|a^{\mathrm{T}} b| \leqslant |a||b|$.

最速下降方向$_0$：单变量函数的局部性质

- 设定

- 设定
  - ▷ 给定一个函数 $f(x) : \mathbb{R} \to \mathbb{R}$

最速下降方向$_0$：单变量函数的局部性质

- 设定
  - ▷ 给定一个函数 $f(x):\mathbb{R}\to\mathbb{R}$
  - ▷ 给定某一个固定的点 $x_0\in\mathbb{R}$

最速下降方向$_0$：单变量函数的局部性质

- 设定
  - 给定一个函数 $f(x):\mathbb{R}\to\mathbb{R}$
  - 给定某一个固定的点 $x_0\in\mathbb{R}$
  - $f(x)$ 在 $x_0\in\mathbb{R}$ 处可导

最速下降方向$_0$：单变量函数的局部性质

- 设定
  - ▷ 给定一个函数 $f(x) : \mathbb{R} \to \mathbb{R}$
  - ▷ 给定某一个固定的点 $x_0 \in \mathbb{R}$
  - ▷ $f(x)$ 在 $x_0 \in \mathbb{R}$ 处可导
- 局部性质

- 设定
  - ▷ 给定一个函数 $f(x): \mathbb{R} \to \mathbb{R}$
  - ▷ 给定某一个固定的点 $x_0 \in \mathbb{R}$
  - ▷ $f(x)$ 在 $x_0 \in \mathbb{R}$ 处可导
- 局部性质
  - ▷ 存在 x_0 的一个邻域，对该邻域内的任意一点 x，有

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \boldsymbol{o}(x - x_0).$$

最速下降方向$_0$：单变量函数的局部性质

- 设定
  - ▷ 给定一个函数 $f(x):\mathbb{R}\to\mathbb{R}$
  - ▷ 给定某一个固定的点 $x_0\in\mathbb{R}$
  - ▷ $f(x)$ 在 $x_0\in\mathbb{R}$ 处可导
- 局部性质
  - ▷ 存在 x_0 的一个邻域，对该邻域内的任意一点 x，有

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\boldsymbol{o}(x-x_0).$$

  - ▷ 局部线性展开与 $f'(x_0)$ 的符号

最速下降方向$_0$：单变量函数的局部性质

- 设定
  - ▷ 给定一个函数 $f(x):\mathbb{R}\to\mathbb{R}$
  - ▷ 给定某一个固定的点 $x_0\in\mathbb{R}$
  - ▷ $f(x)$ 在 $x_0\in\mathbb{R}$ 处可导
- 局部性质
  - ▷ 存在 x_0 的一个邻域，对该邻域内的任意一点 x，有

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\boldsymbol{o}(x-x_0).$$

  - ▷ 局部线性展开与 $f'(x_0)$ 的符号
  - ▷ 多变量函数的梯度向量 $\nabla f(x_0)$: x_0 邻域内的变化趋势

- 设定
  - 给定一个函数 $f(x):\mathbb{R}\to\mathbb{R}$
  - 给定某一个固定的点 $x_0\in\mathbb{R}$
  - $f(x)$ 在 $x_0\in\mathbb{R}$ 处可导
- 局部性质
  - 存在 x_0 的一个邻域，对该邻域内的任意一点 x，有

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\boldsymbol{o}(x-x_0).$$

  - 局部线性展开与 $f'(x_0)$ 的符号
  - 多变量函数的梯度向量 $\nabla f(x_0)$: x_0 邻域内的变化趋势
  - Fermat 引理: x_0 是极值 $\Rightarrow \nabla f(x_0)=0$

最速下降方向$_1$

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$

最速下降方向$_1$

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$
- 考虑 $f(x+\Delta x)$ 的局部线性展开

最速下降方向$_1$

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$
- 考虑 $f(x+\Delta x)$ 的局部线性展开

$$f(x+\Delta x) = f(x) + \underbrace{\nabla f(x) \cdot \Delta x}_{\text{内积}} + \boldsymbol{o}(|\Delta x|).$$

最速下降方向$_1$

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$
- 考虑 $f(x+\Delta x)$ 的局部线性展开

$$f(x+\Delta x) = f(x) + \underbrace{\nabla f(x) \cdot \Delta x}_{\text{内积}} + \boldsymbol{o}(|\Delta x|).$$

- 期待 $\nabla f(x) \cdot \Delta x < 0$ 且尽可能的小

最速下降方向$_1$

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$
- 考虑 $f(x+\Delta x)$ 的局部线性展开
$$f(x+\Delta x) = f(x) + \underbrace{\nabla f(x) \cdot \Delta x}_{\text{内积}} + \boldsymbol{o}(|\Delta x|).$$
- 期待 $\nabla f(x) \cdot \Delta x < 0$ 且尽可能的小
- 如何选择 Δx？

- 目的：从 x 出发，寻求 Δx 使 $f(x+\Delta x) < f(x)$
- 考虑 $f(x+\Delta x)$ 的局部线性展开

$$f(x+\Delta x) = f(x) + \underbrace{\nabla f(x)\cdot\Delta x}_{\text{内积}} + \boldsymbol{o}(|\Delta x|).$$

- 期待 $\nabla f(x)\cdot\Delta x < 0$ 且尽可能的小
- 如何选择 Δx? 方向与步长的分解

最速下降方向$_2$

- 假设 $\Delta x = \alpha v$，其中 $\alpha \in \mathbb{R}_+, v \in \mathbb{R}^n$ 且 $|v| = 1$

最速下降方向$_2$

- 假设 $\Delta x = \alpha v$，其中 $\alpha \in \mathbb{R}_+, v \in \mathbb{R}^n$ 且 $|v| = 1$
- 单位向量 v 称为方向，α 称为步长

- 假设 $\Delta x = \alpha v$，其中 $\alpha \in \mathbb{R}_+, v \in \mathbb{R}^n$ 且 $|v| = 1$
- 单位向量 v 称为方向，α 称为步长
- **定义** 在局部意义下，称

$$v_* = \arg \min_{v \in \mathbb{R}^n, |v|=1} \underbrace{v \cdot \nabla f(x)}_{\text{方向导数}}$$

为可微函数 f 在 x 处的<u>最速下降方向</u>

最速下降方向$_2$

- 假设 $\Delta x = \alpha v$，其中 $\alpha \in \mathbb{R}_+, v \in \mathbb{R}^n$ 且 $|v| = 1$
- 单位向量 v 称为方向，α 称为步长
- **定义** 在局部意义下，称

$$v_* = \arg\min_{v\in\mathbb{R}^n, |v|=1} \underbrace{v \cdot \nabla f(x)}_{\text{方向导数}}$$

为可微函数 f 在 x 处的最速下降方向

- 最速下降方向

$$v_* = -\nabla f(x)/|\nabla f(x)|.$$

- 假设 $\Delta x = \alpha v$，其中 $\alpha \in \mathbb{R}_+, v \in \mathbb{R}^n$ 且 $|v| = 1$
- 单位向量 v 称为方向，α 称为步长
- **定义** 在局部意义下，称

$$v_* = \arg \min_{v \in \mathbb{R}^n, |v|=1} \underbrace{v \cdot \nabla f(x)}_{\text{方向导数}}$$

为可微函数 f 在 x 处的<u>最速下降方向</u>

- 最速下降方向

$$v_* = -\nabla f(x)/|\nabla f(x)|.$$

- (课堂练习) 连续可微函数的梯度垂直于其等值面的切平面.

最速下降法$_1$

- 取

$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$

- 取

$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$

- 最速下降迭代格式：

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

其中，α 称为“步长”或“学习率”

最速下降法$_1$

- 取

$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$

- 最速下降迭代格式：

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

其中，α 称为“步长”或“学习率”

- 自然策略的近似方法

- 取
$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$
- 最速下降迭代格式：
$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$
其中，α 称为“步长”或“学习率”
- 自然策略的近似方法
- 当 f 给定，确定一个迭代序列还需：

最速下降法$_1$

- 取
$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$
- 最速下降迭代格式：
$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$
其中，α 称为“步长”或“学习率”
- 自然策略的近似方法
- 当 f 给定，确定一个迭代序列还需：
    - 步长参数 α

最速下降法$_1$

- 取

$$\Delta x = \alpha' v_* = -\alpha' \frac{\nabla f(x)}{|\nabla f(x)|} = -\alpha \nabla f(x)$$

- 最速下降迭代格式：

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

其中，α 称为“步长”或“学习率”

- 自然策略的近似方法
- 当 f 给定，确定一个迭代序列还需：
  - ▷ 步长参数 α
  - ▷ 初始位置 x_1

最速下降法$_2$

Algorithm 3 定步长最速下降算法

给定初始位置 x_1 和学习率 α.
for $t = 1 : T$ **do**
    计算梯度 $\nabla f(x_t)$.
    更新位置 $x_{t+1} = x_t - \alpha \nabla f(x_t)$.
end for

Algorithm 3 定步长最速下降算法

1: 给定初始位置 x_1 和学习率 α.
2: **for** $t = 1 : T$ **do**
3: 计算梯度 $\nabla f(x_t)$.
4: 更新位置 $x_{t+1} = x_t - \alpha \nabla f(x_t)$.
5: **end for**

- 算法的实质中止: $\nabla f(x_s) = 0$

最速下降法$_2$

Algorithm 3 定步长最速下降算法

1: 给定初始位置 x_1 和学习率 α.
2: **for** $t = 1 : T$ **do**
3: 　计算梯度 $\nabla f(x_t)$.
4: 　更新位置 $x_{t+1} = x_t - \alpha \nabla f(x_t)$.
5: **end for**

- 算法的实质中止: $\nabla f(x_s) = 0$
- 步长 α 的取值：过大 (迭代序列 $\{x_t\}$ 发散的例子)；过小？

Algorithm 3 定步长最速下降算法

1: 给定初始位置 x_1 和学习率 α.
2: **for** $t = 1 : T$ **do**
3: 计算梯度 $\nabla f(x_t)$.
4: 更新位置 $x_{t+1} = x_t - \alpha \nabla f(x_t)$.
5: **end for**

- 算法的实质中止: $\nabla f(x_s) = 0$
- 步长 α 的取值：过大 (迭代序列 $\{x_t\}$ 发散的例子)；过小？
- 问题：步长策略？$\{x_t\}$ 与 x_* 关系？计算花费的度量？

Algorithm 3 定步长最速下降算法

1: 给定初始位置 x_1 和学习率 α.
2: **for** $t = 1 : T$ **do**
3: 计算梯度 $\nabla f(x_t)$.
4: 更新位置 $x_{t+1} = x_t - \alpha \nabla f(x_t)$.
5: **end for**

- 算法的实质中止: $\nabla f(x_s) = 0$
- 步长 α 的取值：过大 (迭代序列 $\{x_t\}$ 发散的例子)；过小？
- **问题**：步长策略？$\{x_t\}$ 与 x_* 关系？计算花费的度量？
- 在何种前提下，可以回答上述问题？暂时限于讨论凸条件

凸集与凸函数

- 有限线性张成集：点、线段、三角形、四面体……

- 有限线性张成集：点、线段、三角形、四面体……
- **定义** (从单纯形到凸集) 称 $K \subset \mathbb{R}^n$ 是凸集，倘若

$$\forall x, y \in K \ \Rightarrow \ (1-\lambda)x + \lambda y \in K, \ \forall \lambda \in [0,1]. \tag{1}$$

- 有限线性张成集：点、线段、三角形、四面体……
- **定义** (从单纯形到凸集) 称 $K \subset \mathbb{R}^n$ 是凸集，倘若

$$\forall x, y \in K \ \Rightarrow \ (1-\lambda)x + \lambda y \in K, \ \forall \lambda \in [0,1]. \tag{1}$$

- **定义** 假设 K 为凸集. 称 $f : K \subset \mathbb{R}^n \to \mathbb{R}$ 是凸集 K 上的凸函数，倘若 $\forall x, y \in K$, $\forall \lambda \in [0,1]$,

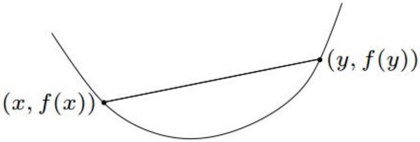
$$f((1-\lambda)x + \lambda y) \leqslant (1-\lambda)f(x) + \lambda f(y). \tag{2}$$

- 有限线性张成集：点、线段、三角形、四面体……
- **定义** (从单纯形到凸集) 称 $K \subset \mathbb{R}^n$ 是凸集，倘若

$$\forall x, y \in K \ \Rightarrow \ (1-\lambda)x + \lambda y \in K, \ \forall \lambda \in [0,1]. \qquad (1)$$

- **定义** 假设 K 为凸集. 称 $f : K \subset \mathbb{R}^n \to \mathbb{R}$ 是凸集 K 上的凸函数，倘若 $\forall x, y \in K, \ \forall \lambda \in [0,1]$,

$$f((1-\lambda)x + \lambda y) \leqslant (1-\lambda)f(x) + \lambda f(y). \qquad (2)$$



命题 (一阶凸条件)

设 K是凸集. 若 f 在 K 上有一阶导数，则 f 是凸函数等价于

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x),\ \forall x, y \in K. \tag{3}$$

命题 (一阶凸条件)

设 K是凸集. 若 f 在 K 上有一阶导数，则 f 是凸函数等价于

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x),\ \forall x, y \in K. \tag{3}$$

Proof.

(证明要点见补充内容) □

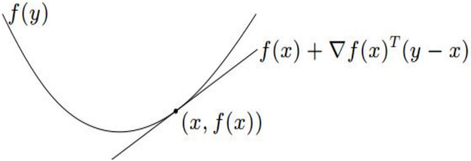
命题 (一阶凸条件)

设 K是凸集. 若 f 在 K 上有一阶导数，则 f 是凸函数等价于

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x),\ \forall x, y \in K. \tag{3}$$

Proof.

(证明要点见补充内容) □



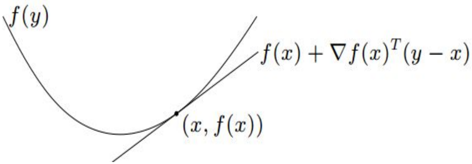
命题 (一阶凸条件)

设 K是凸集. 若 f 在 K 上有一阶导数，则 f 是凸函数等价于

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x),\ \forall x, y \in K. \tag{3}$$

Proof.

(证明要点见补充内容) □



有时还考虑二阶凸条件 (由 Taylor 公式 与 特征分解 得出)

凸函数性质$_2$

Lagrange 型余项 Taylor 公式

设 Ω 是凸区域. 若 $f:\Omega\to\mathbb{R}$ 在 Ω 上具有二阶连续偏导数，则对于任意的 $x,y\in\Omega$, 存在一点 $\xi\in\overline{xy}$ 使得

$$f(y)=f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{1}{2}(y-x)^{\mathrm{T}}\nabla^2 f(\xi)(y-x),$$

其中，Hessian 矩阵 (实对称方阵) $\nabla^2 f(x)$ 定义为

$$\begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n\partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

凸函数性质$_3$

引理 (特征分解)

任意的实对称矩阵 A 可被分解成 $A = Q\Lambda Q^{\mathrm{T}}$, 其中, $QQ^{\mathrm{T}} = I$ 且 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$.

引理 (特征分解)

任意的实对称矩阵 A 可被分解成 $A = Q\Lambda Q^{\mathrm{T}}$, 其中, $QQ^{\mathrm{T}} = I$ 且 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$.

Proof.

(略) □

引理 (特征分解)

任意的实对称矩阵 A 可被分解成 $A=Q\Lambda Q^{\mathrm{T}}$, 其中, $QQ^{\mathrm{T}}=I$ 且 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$.

Proof.

(略) □

定义

实对称矩阵 A 是半正定的，若 $x^{\mathrm{T}}Ax\geqslant 0,\ \forall x\in\mathbb{R}^n$.

引理 (特征分解)

任意的实对称矩阵 A 可被分解成 $A = Q\Lambda Q^{\mathrm{T}}$, 其中, $QQ^{\mathrm{T}} = I$ 且 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$.

Proof.

(略) $\square$

定义

实对称矩阵 A 是半正定的，若 $x^{\mathrm{T}}Ax \geqslant 0, \ \forall x \in \mathbb{R}^n$.

命题

实对称矩阵 A 为半正定等价于 A 的特征值均为非负.

引理 (特征分解)

任意的实对称矩阵 A 可被分解成 $A = Q\Lambda Q^{\mathrm{T}}$, 其中, $QQ^{\mathrm{T}} = I$ 且 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$.

Proof.

(略) $\square$

定义

实对称矩阵 A 是半正定的，若 $x^{\mathrm{T}}Ax \geqslant 0,\ \forall x \in \mathbb{R}^n$.

命题

实对称矩阵 A 为半正定等价于 A 的特征值均为非负.

Proof.

(板书) $\square$

定理 (二阶凸条件)

设 K是凸集. 如果 f 在 K 上具有二阶导数，那么 f 是凸函数等价于对于任意的 $x \in K$，Hessian 矩阵 $\nabla^2 f(x)$ 所有特征值均为非负.

定理 (二阶凸条件)

设 K是凸集. 如果 f 在 K 上具有二阶导数，那么 f 是凸函数等价于对于任意的 $x \in K$，Hessian 矩阵 $\nabla^2 f(x)$ 所有特征值均为非负.

Proof.

定理 (二阶凸条件)

设 K是凸集. 如果 f 在 K 上具有二阶导数，那么 f 是凸函数等价于对于任意的 $x \in K$，Hessian 矩阵 $\nabla^2 f(x)$ 所有特征值均为非负.

Proof.

由 Taylor 公式，可知

$$f(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{1}{2}(y - x)^{\mathrm{T}}\nabla^2 f(\xi)(y - x).$$

定理 (二阶凸条件)

设 K是凸集. 如果 f 在 K 上具有二阶导数，那么 f 是凸函数等价于对于任意的 $x \in K$，Hessian 矩阵 $\nabla^2 f(x)$ 所有特征值均为非负.

Proof.

由 Taylor 公式，可知

$$f(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \frac{1}{2}(y-x)^{\mathrm{T}}\nabla^2 f(\xi)(y-x).$$

所以，一阶凸条件

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x),\ \forall x, y \in K$$

等价于 $(y-x)^{\mathrm{T}}\nabla^2 f(\xi)(y-x) \geqslant 0,\ \forall x, y \in K.$ $\square$

定义 (l-强凸与 L-光滑)

f 是 l-强凸与 L-光滑的，若存在 $0 < l \leqslant L < \infty$ 使得 f 满足

$$\frac{l}{2}\|x-y\|^2 \leqslant f(x) - f(y) - \nabla f(y)^{\mathrm{T}}(x-y) \leqslant \frac{L}{2}\|x-y\|^2.$$

凸函数性质$_5$

定义 (l-强凸与 L-光滑)

f 是 l-强凸与 L-光滑的，若存在 $0 < l \leqslant L < \infty$ 使得 f 满足

$$\frac{l}{2}\|x-y\|^2 \leqslant f(x) - f(y) - \nabla f(y)^{\mathrm{T}}(x-y) \leqslant \frac{L}{2}\|x-y\|^2.$$

该定义隐含 f 具有唯一的极小值 x_*.

定义 (l-强凸与 L-光滑)

f 是 l-强凸与 L-光滑的，若存在 $0 < l \leqslant L < \infty$ 使得 f 满足

$$\frac{l}{2}\|x-y\|^2 \leqslant f(x) - f(y) - \nabla f(y)^{\mathrm{T}}(x-y) \leqslant \frac{L}{2}\|x-y\|^2.$$

该定义隐含 f 具有唯一的极小值 x_*.

定理

若 $\forall x$ 有 $0 < l \leqslant \lambda_1(\nabla^2 f(x)) \leqslant \lambda_n(\nabla^2 f(x)) \leqslant L < \infty$，则 f 是 l-强凸与 L-光滑的.

定义 (l-强凸与 L-光滑)

f 是 l-强凸与 L-光滑的，若存在 $0 < l \leqslant L < \infty$ 使得 f 满足

$$\frac{l}{2}\|x-y\|^2 \leqslant f(x) - f(y) - \nabla f(y)^{\mathrm{T}}(x-y) \leqslant \frac{L}{2}\|x-y\|^2.$$

该定义隐含 f 具有唯一的极小值 x_*.

定理

若 $\forall x$ 有 $0 < l \leqslant \lambda_1(\nabla^2 f(x)) \leqslant \lambda_n(\nabla^2 f(x)) \leqslant L < \infty$，则 f 是 l-强凸与 L-光滑的.

Proof.

(练习) □

凸函数性质$_6$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i), \text{ 其中 } \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geqslant 0.$$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i), \text{ 其中 } \sum_{i=1}^{N} \lambda_i = 1, \ \lambda_i \geqslant 0.$$

Proof.

(数学归纳法) $N = 1, 2$ 时，上述不等式成立；

下面讨论 $N = k \to N = k+1$：

假设当 $N = k$ 时，有

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leqslant \sum_{i=1}^{k} \lambda_i f(x_i)$$

凸函数性质$_6$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i), \text{ 其中 } \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geqslant 0.$$

Proof.

(数学归纳法) $N = 1, 2$ 时，上述不等式成立；

下面讨论 $N = k \to N = k + 1$：

假设当 $N = k$ 时，有

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leqslant \sum_{i=1}^{k} \lambda_i f(x_i)$$

凸函数性质$_6$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i),\ \text{其中}\ \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geqslant 0.$$

Proof.

(数学归纳法) $N = 1, 2$ 时，上述不等式成立；

下面讨论 $N = k \to N = k + 1$:

假设当 $N = k$ 时，有

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leqslant \sum_{i=1}^{k} \lambda_i f(x_i)$$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i), \text{ 其中 } \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geqslant 0.$$

Proof.

(数学归纳法) $N = 1, 2$ 时，上述不等式成立；

下面讨论 $N = k \to N = k + 1$：

假设当 $N = k$ 时，有

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leqslant \sum_{i=1}^{k} \lambda_i f(x_i)$$

命题 1 (Jensen 不等式)

若 f 是凸集 K 上的凸函数且 $x_i \in K$，则

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leqslant \sum_{i=1}^{N} \lambda_i f(x_i), \text{ 其中 } \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geqslant 0.$$

Proof.

(数学归纳法) $N = 1, 2$ 时，上述不等式成立；

下面讨论 $N = k \to N = k + 1$：

假设当 $N = k$ 时，有

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leqslant \sum_{i=1}^{k} \lambda_i f(x_i)$$

凸函数性质$_7$

Proof (Contd.)

当 $N = k+1$ 时，

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) =& f\left(\lambda_{k+1}x_{k+1} + \sum_{i=1}^{k}\lambda_i x_i\right) \\
=& f\left(\lambda_{k+1}x_{k+1} + (1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right) \\
\leqslant& \lambda_{k+1}f(x_{k+1}) + (1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i' = \frac{\lambda_i}{1-\lambda_{k+1}}$，注意到 $\sum_{i=1}^{k}\lambda_i' = 1$，

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right) \leqslant \sum_{i=1}^{k}\lambda_i f(x_i).
$$

$\square$

Proof (Contd.)

当 $N = k+1$ 时,

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_{k+1}x_{k+1} + \sum_{i=1}^{k}\lambda_i x_i\right)\\
&= f\left(\lambda_{k+1}x_{k+1} + (1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
&\leqslant \lambda_{k+1}f(x_{k+1}) + (1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i' = \frac{\lambda_i}{1-\lambda_{k+1}}$, 注意到 $\sum_{i=1}^{k}\lambda_i' = 1$,

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right) \leqslant \sum_{i=1}^{k}\lambda_i f(x_i).
$$

$\square$

Proof (Contd.)

当 $N=k+1$ 时，

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_{k+1}x_{k+1}+\sum_{i=1}^{k}\lambda_i x_i\right)\\
&= f\left(\lambda_{k+1}x_{k+1}+(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
&\leqslant \lambda_{k+1}f(x_{k+1})+(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}$$

令 $\lambda_i'=\frac{\lambda_i}{1-\lambda_{k+1}}$，注意到 $\sum_{i=1}^{k}\lambda_i'=1$，

$$(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(x_i).$$

$\square$

Proof (Contd.)

当 $N=k+1$ 时，

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_{k+1}x_{k+1}+\sum_{i=1}^{k}\lambda_i x_i\right)\\
&= f\left(\lambda_{k+1}x_{k+1}+(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
&\leqslant \lambda_{k+1}f(x_{k+1})+(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i'=\frac{\lambda_i}{1-\lambda_{k+1}}$，注意到 $\sum_{i=1}^{k}\lambda_i'=1$，

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(x_i).
$$

□

凸函数性质$_7$

Proof (Contd.)

当 $N=k+1$ 时，

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) =& f\left(\lambda_{k+1}x_{k+1}+\sum_{i=1}^{k}\lambda_i x_i\right)\\
=& f\left(\lambda_{k+1}x_{k+1}+(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
\leqslant& \lambda_{k+1}f(x_{k+1})+(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i'=\frac{\lambda_i}{1-\lambda_{k+1}}$，注意到 $\sum_{i=1}^{k}\lambda_i'=1$，

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(x_i).
$$

$\square$

Proof (Contd.)

当 $N=k+1$ 时,

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) =& f\left(\lambda_{k+1}x_{k+1}+\sum_{i=1}^{k}\lambda_i x_i\right)\\
=& f\left(\lambda_{k+1}x_{k+1}+(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
\leqslant& \lambda_{k+1}f(x_{k+1})+(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i'=\frac{\lambda_i}{1-\lambda_{k+1}}$, 注意到 $\sum_{i=1}^{k}\lambda_i'=1$,

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(x_i).
$$

□

Proof (Contd.)

当 $N=k+1$ 时，

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_{k+1}x_{k+1}+\sum_{i=1}^{k}\lambda_i x_i\right)\\
&= f\left(\lambda_{k+1}x_{k+1}+(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\\
&\leqslant \lambda_{k+1}f(x_{k+1})+(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)
\end{aligned}
$$

令 $\lambda_i'=\frac{\lambda_i}{1-\lambda_{k+1}}$，注意到 $\sum_{i=1}^{k}\lambda_i'=1$，

$$
(1-\lambda_{k+1})f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leqslant\sum_{i=1}^{k}\lambda_i f(x_i).
$$

□

- 凸函数例子：
  - 指数函数：e^{ax}，$x \in \mathbb{R}, \forall a \in \mathbb{R}$
  - 幂函数：x^a，$x \in \mathbb{R}_{++}, \forall a \in \mathbb{R} - (0,1)$
  - 绝对值幂函数：$|x|^a$，$x \in \mathbb{R}, \forall a \geqslant 1$
  - 负对数函数：$-\log x$，$x \in \mathbb{R}_{++}$
  - 负熵：$x \log x$，$x \in \mathbb{R}_{++}$

- 凸函数例子：
  - ▷ 指数函数：e^{ax}, $x \in \mathbb{R}, \forall a \in \mathbb{R}$
  - ▷ 幂函数：x^a, $x \in \mathbb{R}_{++}, \forall a \in \mathbb{R} - (0,1)$
  - ▷ 绝对值幂函数：$|x|^a$, $x \in \mathbb{R}, \forall a \geqslant 1$
  - ▷ 负对数函数：$-\log x$, $x \in \mathbb{R}_{++}$
  - ▷ 负熵：$x \log x$, $x \in \mathbb{R}_{++}$
- 后面将针对凸条件，分析最速下降法：

凸函数性质$_8$

- 凸函数例子：
  - ▷ 指数函数：e^{ax}，$x \in \mathbb{R}$, $\forall a \in \mathbb{R}$
  - ▷ 幂函数：x^a，$x \in \mathbb{R}_{++}$, $\forall a \in \mathbb{R} - (0,1)$
  - ▷ 绝对值幂函数：$|x|^a$，$x \in \mathbb{R}$, $\forall a \geqslant 1$
  - ▷ 负对数函数：$-\log x$，$x \in \mathbb{R}_{++}$
  - ▷ 负熵：$x \log x$，$x \in \mathbb{R}_{++}$
- 后面将针对凸条件，分析最速下降法：
  - ▷ 步长策略

- 凸函数例子：
  - ▷ 指数函数：e^{ax}, $x \in \mathbb{R}$, $\forall a \in \mathbb{R}$
  - ▷ 幂函数：x^a, $x \in \mathbb{R}_{++}$, $\forall a \in \mathbb{R} - (0,1)$
  - ▷ 绝对值幂函数：$|x|^a$, $x \in \mathbb{R}$, $\forall a \geqslant 1$
  - ▷ 负对数函数：$-\log x$, $x \in \mathbb{R}_{++}$
  - ▷ 负熵：$x \log x$, $x \in \mathbb{R}_{++}$
- 后面将针对凸条件，分析最速下降法：
  - ▷ 步长策略
  - ▷ 以及相应的收敛速度

梯度法的收敛性

定理 (线性收敛)

若 $\forall x$ 有 $0 < l \leqslant \lambda_1(\nabla^2 f(x)) \leqslant \lambda_n(\nabla^2 f(x)) \leqslant L < \infty$，则当 $\alpha = \frac{2}{L+l}$ 时，迭代法 $x_{t+1} = x_t - \alpha \nabla f(x_t)$ 满足

$$\|x_{t+1} - x_*\| \leqslant \left(\frac{L-l}{L+l}\right)^t \|x_1 - x_*\|.$$

强凸条件下的线性收敛性

定理 (线性收敛)

若 $\forall x$ 有 $0 < l \leqslant \lambda_1(\nabla^2 f(x)) \leqslant \lambda_n(\nabla^2 f(x)) \leqslant L < \infty$，则当 $\alpha = \frac{2}{L+l}$ 时，迭代法 $x_{t+1} = x_t - \alpha \nabla f(x_t)$ 满足

$$\|x_{t+1} - x_*\| \leqslant \left(\frac{L-l}{L+l}\right)^t \|x_1 - x_*\|.$$

Proof.

(练习) $\square$

Polyak-Łojasiewicz 不等式与另一种证明

引理 1 (optimality gap · gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

引理 1 (optimality gap $\cdot$ gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

Proof.

令 $q(t) = f(x) + \nabla f(x)^{\mathrm{T}}(t-x) + \frac{l}{2}\|t-x\|_2^2$, 则 $f_* \geqslant q(x_*)$.

由 $\nabla q(t) = \nabla f(x) + l(t-x) = 0$ 可知 $t = x - \frac{1}{l}\nabla f(x)$ 时，$q(t)$ 取最小值 $q_* = f(x) - \frac{1}{2l}\|\nabla f(x)\|_2^2$. 故 $f_* \geqslant q(x_*) \geqslant q_*$. $\square$

Polyak-Łojasiewicz 不等式与另一种证明

引理 1 (optimality gap · gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

Proof.

令 $q(t) = f(x) + \nabla f(x)^{\mathrm{T}}(t - x) + \frac{l}{2}\|t - x\|_2^2$, 则 $f_* \geqslant q(x_*)$.

由 $\nabla q(t) = \nabla f(x) + l(t - x) = 0$ 可知 $t = x - \frac{1}{l}\nabla f(x)$ 时，$q(t)$ 取最小值 $q_* = f(x) - \frac{1}{2l}\|\nabla f(x)\|_2^2$. 故 $f_* \geqslant q(x_*) \geqslant q_*$. □

Polyak-Łojasiewicz 不等式与另一种证明

引理 1 (optimality gap · gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

Proof.

令 $q(t) = f(x) + \nabla f(x)^{\mathrm{T}}(t - x) + \frac{l}{2}\|t - x\|_2^2$, 则 $f_* \geqslant q(x_*)$.

由 $\nabla q(t) = \nabla f(x) + l(t - x) = 0$ 可知 $t = x - \frac{1}{l}\nabla f(x)$ 时，$q(t)$ 取最小值 $q_* = f(x) - \frac{1}{2l}\|\nabla f(x)\|_2^2$. 故 $f_* \geqslant q(x_*) \geqslant q_*$. □

Polyak-Łojasiewicz 不等式与另一种证明

引理 1 (optimality gap · gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

Proof.

令 $q(t) = f(x) + \nabla f(x)^{\mathrm{T}}(t-x) + \frac{l}{2}\|t-x\|_2^2$, 则 $f_* \geqslant q(x_*)$.

由 $\nabla q(t) = \nabla f(x) + l(t-x) = 0$ 可知 $t{=}x{-}\frac{1}{l}\nabla f(x)$ 时，$q(t)$ 取最小值 $q_* = f(x){-}\frac{1}{2l}\|\nabla f(x)\|_2^2$. 故 $f_* \geqslant q(x_*) \geqslant q_*$. □

Polyak-Łojasiewicz 不等式与另一种证明

引理 1 (optimality gap $\cdot$ gradient)

若 $0 < l \leqslant \lambda_i(\nabla^2 f)$，则 $f(x) - f_* \leqslant \frac{1}{2l}\|\nabla f(x)\|_2^2$.

Proof.

令 $q(t) = f(x) + \nabla f(x)^{\mathrm{T}}(t-x) + \frac{l}{2}\|t-x\|_2^2$, 则 $f_* \geqslant q(x_*)$.

由 $\nabla q(t) = \nabla f(x) + l(t-x) = 0$ 可知 $t{=}x{-}\frac{1}{l}\nabla f(x)$ 时，$q(t)$ 取最小值 $q_* = f(x){-}\frac{1}{2l}\|\nabla f(x)\|_2^2$. 故 $f_* \geqslant q(x_*) \geqslant q_*$. □

练习

根据P-L不等式证明 l-强凸与 L-光滑条件下梯度法满足

$$f(x_{t+1}) - f(x_*) \leqslant \left(\frac{L-l}{L+l}\right)^t \Big(f(x_1) - f(x_*)\Big).$$

凸条件下的次线性收敛性

定理 (平均次线性收敛)

设 f 是可微凸函数且 $|\nabla f(x)|\leqslant G, \forall x$. 如果步长 $\alpha = \frac{\alpha_0}{\sqrt{T}}$ 且 $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$，那么迭代法 $x_{t+1} = x_t - \alpha\nabla f(x_t)$ 满足

$$f(\bar{x}) - f(x_*) \leqslant \frac{1}{\sqrt{T}}\left(\frac{\|x_1 - x_*\|^2}{2\alpha_0} + \frac{\alpha_0 G^2}{2}\right).$$

注

若 $\alpha_0 = \frac{\|x_1 - x_*\|}{G}$, 则 $f(\bar{x}) - f(x_*) \leqslant \frac{G\|x_1 - x_*\|}{\sqrt{T}}$.

定理 (平均次线性收敛)

设 f 是可微凸函数且 $|\nabla f(x)|\leqslant G,\ \forall x$. 如果步长 $\alpha = \frac{\alpha_0}{\sqrt{T}}$ 且 $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$，那么迭代法 $x_{t+1} = x_t - \alpha\nabla f(x_t)$ 满足

$$f(\bar{x}) - f(x_*) \leqslant \frac{1}{\sqrt{T}}\left(\frac{\|x_1 - x_*\|^2}{2\alpha_0} + \frac{\alpha_0 G^2}{2}\right).$$

注

若 $\alpha_0 = \frac{\|x_1 - x_*\|}{G}$, 则 $f(\bar{x}) - f(x_*) \leqslant \frac{G\|x_1 - x_*\|}{\sqrt{T}}$.

Proof.

(板书) □

(1) 引入凸条件，由 x_t 处的信息建立 f 在 x_* 处的下界：

$$f(x_*) \geqslant f(x_t) + \nabla f(x_t)^{\mathrm{T}}(x_* - x_t)$$

(1) 引入凸条件，由 x_t 处的信息建立 f 在 x_* 处的下界：

$$f(x_*) \geqslant f(x_t) + \nabla f(x_t)^{\mathrm{T}}(x_* - x_t)$$

于是得到 $f(x_t)$ 与 $f(x_*)$ 的间隙上界：

$$f(x_t) - f(x_*) \leqslant \nabla f(x_t)^{\mathrm{T}}(x_t - x_*)$$

证明关键思路

(1) 引入凸条件，由 x_t 处的信息建立 f 在 x_* 处的下界：

$$f(x_*) \geqslant f(x_t) + \nabla f(x_t)^{\mathrm{T}}(x_* - x_t)$$

于是得到 $f(x_t)$ 与 $f(x_*)$ 的间隙上界：

$$f(x_t) - f(x_*) \leqslant \nabla f(x_t)^{\mathrm{T}}(x_t - x_*)$$

(2) 引入梯度的有界性，借助迭代格式建立上界：

$$\nabla f(x_t)^{\mathrm{T}}(x_t - x_*) \leqslant \frac{1}{2\alpha}(\|x_* - x_t\|^2 - \|x_* - x_{t+1}\|^2) + \frac{\alpha G^2}{2}$$

证明关键思路

(1) 引入凸条件，由 x_t 处的信息建立 f 在 x_* 处的下界：

$$f(x_*) \geqslant f(x_t) + \nabla f(x_t)^{\mathrm{T}}(x_* - x_t)$$

于是得到 $f(x_t)$ 与 $f(x_*)$ 的间隙上界：

$$f(x_t) - f(x_*) \leqslant \nabla f(x_t)^{\mathrm{T}}(x_t - x_*)$$

(2) 引入梯度的有界性，借助迭代格式建立上界：

$$\nabla f(x_t)^{\mathrm{T}}(x_t - x_*) \leqslant \frac{1}{2\alpha}(\|x_* - x_t\|^2 - \|x_* - x_{t+1}\|^2) + \frac{\alpha G^2}{2}$$

(3) Jensen 不等式与交错相消

非凸情况的收敛性

- 基于精确线搜索的最速下降：

- 基于精确线搜索的最速下降：
  - 在最速下降方向上，确定步长 α_t 使得

$$f\big(x_t - \alpha_t \nabla f(x_t)\big) = \min_{\alpha \geqslant 0} f\big(x_t - \alpha \nabla f(x_t)\big)$$

非凸情况的收敛性

- 基于精确线搜索的最速下降：
  - ▷ 在最速下降方向上，确定步长 α_t 使得

$$f\big(x_t - \alpha_t \nabla f(x_t)\big) = \min_{\alpha \geqslant 0} f\big(x_t - \alpha \nabla f(x_t)\big)$$

  - ▷ 基于精确线搜索的最速下降迭代格式

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

非凸情况的收敛性

- 基于精确线搜索的最速下降：
  - ▷ 在最速下降方向上，确定步长 α_t 使得

$$f\big(x_t - \alpha_t \nabla f(x_t)\big) = \min_{\alpha \geqslant 0} f\big(x_t - \alpha \nabla f(x_t)\big)$$

  - ▷ 基于精确线搜索的最速下降迭代格式

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

  - ▷ 收敛性 (课堂练习)：提示 - 考虑 $f(x_{t+1})$ 的上界

- 基于精确线搜索的最速下降：
  - ▷ 在最速下降方向上，确定步长 α_t 使得

$$f\big(x_t - \alpha_t \nabla f(x_t)\big) = \min_{\alpha \geqslant 0} f\big(x_t - \alpha \nabla f(x_t)\big)$$

  - ▷ 基于精确线搜索的最速下降迭代格式

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

  - ▷ 收敛性 (课堂练习)：提示 - 考虑 $f(x_{t+1})$ 的上界

定理 5 (非凸·精确线搜索·最速下降·收敛性)

如果 f 下方有界且 $\|\nabla^2 f\|_2 \leqslant L < \infty$，那么基于精确线搜索的最速下降迭代序列必然收敛到 f 的梯度零点.

- 注记：当 $\alpha = \frac{1}{L}$ 亦满足上述结论

监督学习梯度算法的花费

对于给定的(小规模)数据集，最小化相应的损失函数：

损失函数	步长/学习率	收敛速度	复杂度
凸 & 不光滑	$\alpha_t = \alpha_0 \ or \ \alpha_t = \frac{\alpha_0}{\sqrt{t}}$	$\boldsymbol{O}(\frac{1}{\sqrt{t}})$	$\boldsymbol{O}(\frac{1}{\epsilon^2})$
凸 & 光滑	$\alpha_t = \alpha_0 = \frac{1}{L}$	$\boldsymbol{O}(\frac{1}{t})$	$\boldsymbol{O}(\frac{1}{\epsilon})$
强凸 & 不光滑	$\alpha_t = \frac{1}{t}$	$\boldsymbol{O}(\frac{1}{t})$	$\boldsymbol{O}(\frac{1}{\epsilon})$
强凸 & 光滑	$\alpha_t = \alpha_0 = \frac{2}{L+l}$	$\boldsymbol{O}\left((\frac{L-l}{L+l})^t\right)$	$\boldsymbol{O}(\log \frac{1}{\epsilon})$

Table 1: 次梯度下降算法

补充

- 证明要点：仅讨论 $n=1$. 充分性：

$$\begin{aligned} f((1-\lambda)x+\lambda y) \leqslant& (1-\lambda)f(x)+\lambda f(y) \\ =& f(x)-\lambda f(x)+\lambda f(y) \end{aligned}$$

$$\Rightarrow f(y) \geqslant f(x)+[f(x+\lambda(y-x))-f(x)]/\lambda;$$

必要性：令 $z=(1-\lambda)x+\lambda y$,

$$f(x) \geqslant f(z)+\nabla f(z)(x-z), \quad f(y) \geqslant f(z)+\nabla f(z)(y-z)$$

$$\Rightarrow (1-\lambda)f(x)+\lambda f(y) \geqslant f(z).$$

END