数值分析讲义

上海交通大学 · 数学科学学院

马征

2026

目录

第一章　课程简介

1.0.1　课程简介

本课程是**二年级数学学生的基础课**，以讲授科学计算学科的基本概念和基础计算方法的分析为主，穿插一些数据科学（包括最优化、机器学习等）的入门介绍。

课程内容包含：

- 非线性方程求根、多项式插值、函数逼近论
- 数值微分与数值积分
- 常微分方程数值解、蒙特卡罗方法
- 凸优化方法、机器学习入门介绍、科学计算中的 Python

课程目标：

- 熟练掌握数值分析中的**基础及部分高级方法**，学会分析**稳定性和精度**
- 掌握**常微分方程（组）数值求解**方法
- 学会使用 **Python 编程**实现算法，初步了解**数据科学**

1.0.2　内容安排

共四个章节，每章约 4 周，具体以实际情况为准。

章节	主题	内容
第一章	数值分析中的基本方法	插值与逼近、数值微分与积分、非线性方程求根
第二章	数值分析中的高级方法	快速傅立叶变换、蒙特卡洛等
第三章	常微分方程数值解	方法与分析
第四章	数据科学入门	优化、机器学习简介

1.0.3　考核方式

由平时作业、课程项目、期末考试组成

项目	占比	说明
平时作业	**20%**	每周 4-5 道题
课程项目	**40%**	程序大作业
期末考试	**40%**	闭卷笔试

1.0.4 教材及参考资料

内容按章节来源：

章节	教材
第 1–3 章	《数值分析》张平文、李铁军（北京大学出版社）
第 4 章	*Convex Optimization*，Stephen Boyd & Lieven Vandenberghe
第 4 章	《机器学习》（西瓜书）周志华（清华大学出版社）
Python	*Scientific Computing with Python* (2nd Ed.)，Führer, Solem, Verdier

其他：本课程配有习题课（时间待定），主要讲解程序实现及作业问题。请扫描二维码加入课程微信群，谢谢！

第二章　多项式插值

2.0.1　第一节多项式插值

Polynomial Interpolation

2.0.2　为什么需要插值？

在科学计算与工程实践中，目标函数 $f(x)$ 常常**仅在有限个离散点处已知**，或表达式过于复杂难以直接使用。**插值**提供了一种系统方法，用简单函数 $P(x)$ 在节点处精确匹配数据，并在其余位置给出合理近似。

主要需求

- 从实验/采样数据重建连续函数
- 为数值积分、微分提供可计算的近似
- 增密稀疏数据、估算中间值

典型应用

- **数值积分/微分**：构造求积公式
- **工程数据处理**：重建连续曲线
- **计算机图形学**：样条曲线造型
- **科学计算**：有限元形函数构造

2.0.3　插值方法

定义 2.0.1 (插值)**.** 设目标函数 $f(x)$ 定义在区间 $[a,b]$ 上，**插值**是指在 $[a,b]$ 上寻找函数 $P(x)$，使其在 $n+1$ 个样本点 $\{x_i\}_{i=0}^n$ 处满足

$$P(x_i) = f(x_i) \quad \forall\, i = 0, 1, \ldots, n.$$

- $P(x)$ 称为 $f(x)$ 的**插值函数**，$x_0, x_1, \ldots, x_n$ 称为**插值节点**
- 闭区间 $[a,b]$ 称为**插值区间**
- 通常令 $y_i = f(x_i)$，数据 $\{(x_i, y_i)\}_{i=0}^n$ 即为插值数据

2.0.4　插值方法的分类

根据插值函数 $P(x)$ 的类型，插值方法分为：

插值方法	$P(x)$ 的类型
多项式插值	全局多项式
分段插值	分段多项式
三角插值	三角多项式

注 (思考). 还可以采用哪些类型的函数（模型)?

2.0.5 多项式插值的存在唯一性

定理 2.0.2 (定理)**.** 多项式插值唯一存在。

证明. 设多项式 $P(x) = a_0 + a_1x + \cdots + a_nx^n$ 的系数满足线性方程组

$$\begin{cases} a_0 + a_1x_0 + \cdots + a_nx_0^n = y_0 \\ a_0 + a_1x_1 + \cdots + a_nx_1^n = y_1 \\ \qquad\qquad\qquad\qquad\quad \vdots \\ a_0 + a_1x_n + \cdots + a_nx_n^n = y_n \end{cases}$$

系数矩阵的行列式为 **Vandermonde 行列式** $\prod_{i=1}^{n}\prod_{j=0}^{i-1}(x_i - x_j) \neq 0$，故方程组有唯一解。

□

2.0.6 Lagrange 插值

定义 2.0.3 (Lagrange 插值多项式)**.** 对 $n+1$ 个插值数据 $\{(x_i, f(x_i))\}_{i=0}^{n}$，**Lagrange 插值多项式**定义为

$$L_n(x) = \sum_{i=0}^{n} f(x_i)\, l_i(x),$$

其中 $l_i(x)$ 为 **n 次插值基函数 **:

$$l_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 0, 1, \ldots, n.$$

2.0.7 Lagrange 插值基函数的性质

定理 2.0.4 (定理)**.** *Lagrange* 插值多项式是次数为 n 的多项式；每个插值基函数是次数为 n 的多项式，且对 $i, j = 0, 1, \ldots, n$ 满足

$$l_j(x_i) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

2.0.8 Lagrange 插值示例（一）

例 2.0.5 (例 1.1)**.** 写出线性和二次 Lagrange 插值多项式。

** 线性插值（$n = 1$）：**

$$L_1(x) = f(x_0)\frac{x - x_1}{x_0 - x_1} + f(x_1)\frac{x - x_0}{x_1 - x_0}$$

** 二次插值（$n = 2$）：**

$$L_2(x) = f(x_0)\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + f(x_1)\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + f(x_2)\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

2.0.9 Lagrange 插值示例（二）

例 2.0.6 (例 1.2)**.** 对插值节点 x_j，$j = 0, 1, 2, 3$：

x_j	0.40	0.50	0.70	0.80
$\ln x_j$	-0.916291	-0.693147	-0.356675	-0.223144

用 3 次 Lagrange 插值计算 $\ln(0.6)$。

2.0.10 Lagrange 插值余项

定理 2.0.7 (定理 1.2)**.** *n 次 Lagrange 插值的余项为*

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x),$$

其中 $\xi \in (a, b)$ 依赖于 x，且 $\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$。

2.0.11 Lagrange 插值余项上界

定理 2.0.8 (定理 1.3)**.** *n 次 Lagrange 插值余项满足上界估计*

$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!}\,|\omega_{n+1}(x)|,$$

其中 $M_{n+1} = \max_{a \leq x \leq b} |f^{(n+1)}(x)|$。

2.0.12 Lagrange 插值余项示例

例 2.0.9 (例 1.3)**.** 1 次和 2 次 Lagrange 插值的余项分别为

$$R_1(x) = \frac{1}{2}f''(\xi)(x-x_0)(x-x_1), \quad \xi \in [x_0, x_1],$$

$$R_2(x) = \frac{1}{6}f'''(\xi)(x-x_0)(x-x_1)(x-x_2), \quad \xi \in [x_0, x_2].$$

2.0.13 Lagrange 插值性质

例 2.0.10 (例 1.4)**.** 证明

$$\sum_{i=0}^{n} l_i(x)\, x_i^k = x^k, \quad k = 0, 1, \ldots, n.$$

2.0.14 Lagrange 插值综合示例

例 2.0.11 (例 1.5)**.** 已知数据

x_j	0.32	0.34	0.36
$\sin x_j$	-0.314567	-0.333487	-0.352274

分别用线性插值和二次插值计算 sin(0.3367)，并估计截断误差。

2.0.15 第二节牛顿插值

Newton Interpolation

2.0.16 差商

定义 2.0.12 (差商)**.** 对连续函数 f，关于点 x_0, x_k 的一**阶差商**定义为

$$f[x_0, x_k] = \frac{f(x_k) - f(x_0)}{x_k - x_0}.$$

关于点 $x_0, x_1, \ldots, x_k$ 的 **k 阶差商 ** 递归定义为

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_0, x_1, \ldots, x_{k-2}, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_{k-1}}.$$

2.0.17 差商的对称性

定理 2.0.13 (定理 1.5)**.** 差商关于节点具有对称性：

$$f[x_0, x_1, x_2, \ldots, x_k] = f[x_1, x_0, x_2, \ldots, x_k] = \cdots = f[x_1, x_2, \ldots, x_k, x_0].$$

2.0.18 差商的线性组合表示

定理 2.0.14 (定理 1.6)**.** k 阶差商有如下等价表达式:

$$f[x_0,x_1,\dots,x_k]=\sum_{j=0}^{k}\frac{f(x_j)}{\prod\limits_{\substack{i=0\\ i\neq j}}^{k}(x_j-x_i)}=\frac{f[x_1,x_2,\dots,x_k]-f[x_0,x_1,\dots,x_{k-1}]}{x_k-x_0}.$$

2.0.19 差商与导数的关系

定理 2.0.15 (定理 1.7)**.** 若 f 在 $[a,b]$ 上存在 n 阶导数，且 $x_0,x_1,\dots,x_n\in[a,b]$，则

$$f[x_0,x_1,\dots,x_n]=\frac{f^{(n)}(\xi)}{n!},\quad \xi\in[a,b].$$

2.0.20 牛顿插值多项式

定义 2.0.16 (牛顿插值多项式)**.** 对 $[a,b]$ 上的连续函数 f，**n 次牛顿插值多项式 ** 为

$$N_n(x)=\sum_{k=0}^{n}a_k\prod_{i=0}^{k-1}(x-x_i),$$

其中 $a_k=f[x_0,x_1,\dots,x_k]$，$a_0=f(x_0)$。

即：$N_n(x)=f(x_0)+f[x_0,x_1](x-x_0)+\cdots+f[x_0,\dots,x_n](x-x_0)\cdots(x-x_{n-1})$。

2.0.21 牛顿插值余项

余项 $R_n(x):=f(x)-N_n(x)$ 满足

$$R_n(x)=f[x_0,x_1,\dots,x_n,x]\,\omega_{n+1}(x),$$

其中 $\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$。

注 (注). 牛顿插值与 Lagrange 插值余项相同，因为两者是同一插值多项式的不同表示形式。

2.0.22 牛顿插值示例

例 2.0.17 (例 4.6)**.** 设 $x_k=k$，$k=0,1,2,3$，$f(x)=\cos x$。

1. 画出差商表。
2. 计算 1、2、3 次牛顿插值多项式 N_1、N_2、N_3 及其误差界。
3. 计算 $f(1.566)$。

2.0.23 差分

定义 2.0.18 (前向差分 / 后向差分)**. 前向差分**：$\Delta f_k=f_{k+1}-f_k$，$\quad\Delta^m f_k=\Delta^{m-1}f_{k+1}-\Delta^{m-1}f_k$

后向差分：$\nabla f_k=f_k-f_{k-1}$，$\quad\nabla^m f_k=\nabla^{m-1}f_k-\nabla^{m-1}f_{k-1}$

2.0.24 等距节点的差分与差商

等距节点：设 $h>0$，$x_k=x_0+kh$，$k=0,1,\dots,n$。

定理 2.0.19 (定理 4.9)**.**

$$f[x_k,x_{k+1}]=\frac{\Delta f_k}{h},\quad f[x_k,\dots,x_{k+m}]=\frac{\Delta^m f_k}{m!\,h^m},\quad \Delta^n f_k=h^n f^{(n)}(\xi),\ \xi\in(x_k,x_{k+n}).$$

2.0.25 等距节点牛顿前插公式

令 $x=x_0+th$，$h>0$，$0\le t\le 1$，得**牛顿前插公式**：

$$N_n(x_0+th)=f_0+t\Delta f_0+\frac{t(t-1)}{2!}\Delta^2 f_0+\cdots+\frac{t(t-1)\cdots(t-n+1)}{n!}\Delta^n f_0,$$

余项为

$$R_n(x)=\frac{t(t-1)\cdots(t-n)}{(n+1)!}\,h^{n+1}f^{(n+1)}(\xi),\quad \xi\in(x_0,x_n).$$

2.0.26 等距节点牛顿后插公式

令 $x=x_n+th$，$-1\le t\le 0$，$h>0$，逆序节点 $x_n,\dots,x_0$，得**牛顿后插公式**：

$$N_n(x_n+th)=f_n+t\nabla f_n+\frac{t(t+1)}{2!}\nabla^2 f_n+\cdots+\frac{t(t+1)\cdots(t+n-1)}{n!}\nabla^n f_n,$$

余项为

$$R_n(x)=\frac{t(t+1)\cdots(t+n)}{(n+1)!}\,h^{n+1}f^{(n+1)}(\xi),\quad \xi\in(x_0,x_n).$$

2.0.27 等距插值示例

例 2.0.20 (例 4.7)**.** 设 $x_0=1.0$，$h=0.05$，给定 $f(x)=\sqrt{x}$ 在 $x_j=x_0+jh$ $(j=0,1,\dots,6)$ 处的函数值，用 3 次等距插值计算 $f(1.01)$ 和 $f(1.28)$。

前插公式：$N_3(1.01)\approx 1.00499$；后插公式：$N_3(1.28)\approx 1.13137$

真值：1.00498756 和 1.13137085

2.0.28 第三节 Hermite 插值

Hermite Interpolation

2.0.29 Hermite 插值

定义 2.0.21 (Hermite 插值多项式)**.** 设 $a \le x_0 < x_1 < \cdots < x_n \le b$, 寻找多项式 $H_{2n+1} \in \mathcal{P}_{2n+1}$ 满足

$$H_{2n+1}(x_j) = f(x_j), \quad H'_{2n+1}(x_j) = f'(x_j), \quad \forall\, j = 0, 1, \ldots, n.$$

称 H_{2n+1} 为数据 $\left\{(x_j, f(x_j), f'(x_j))\right\}_{j=0}^{n}$ 的 **n 次 Hermite 插值多项式 **。

共有 $2n+2$ 个插值条件，恰好确定 $\mathcal{P}_{2n+1}$ 的 $2n+2$ 个系数。

2.0.30 Hermite 插值基函数（一）

$$H_{2n+1}(x) = \sum_{j=0}^{n} \big[f(x_j)\, \alpha_j(x) + f'(x_j)\, \beta_j(x) \big].$$

定理 2.0.22 (定理)**.** 基函数基于 *Lagrange* 基，满足：

$$\alpha_j(x_i) = \delta_{ij}, \quad \alpha'_j(x_i) = 0, \quad \beta_j(x_i) = 0, \quad \beta'_j(x_i) = \delta_{ij}, \quad \forall\, i, j = 0, \ldots, n.$$

2.0.31 Hermite 插值基函数（二）

基函数的具体形式为：

$$\alpha_j(x) = \left[1 - 2(x - x_j) \sum_{\substack{i=0 \\ i \neq j}}^{n} \frac{1}{x_j - x_i} \right] l_j^2(x), \qquad \beta_j(x) = (x - x_j)\, l_j^2(x),$$

其中 $l_j(x)$ 为 Lagrange 插值基函数。

2.0.32 Hermite 插值余项

定理 2.0.23 (定理 4.10)**.** 设目标函数 $f(x)$ 在 (a, b) 上存在 $2n+2$ 阶导数，则 *Hermite* 插值的余项为

$$R_{2n+1}(x) = f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\, \omega_{n+1}^2(x),$$

其中 $\xi \in (a, b)$, $\omega_{n+1}(x) = \prod_{j=0}^{n} (x - x_j)$。

2.0.33 Hermite 插值计算示例

例 2.0.24 (例 4.8)**.** 计算 3 次 Hermite 插值多项式 $H_3(x)$ 及其余项。

2.0.34 三次 Hermite 插值示例

例 2.0.25 (例 4.9)**.** 设 $f(x) = \ln x$，已知数据：

$f(1.0)$	$f(2.0)$	$f'(1.0)$	$f'(2.0)$
0.0	0.693147	1.0	0.5

用三次 Hermite 插值多项式估计 ln(1.5)。

2.0.35 Hermite 插值误差

定理 2.0.26 (定理 4.12)**.** *三次 Hermite 插值满足误差估计*

$$\max_{x\in[x_k,x_{k+1}]} |f(x) - H_3(x)| \le \frac{(x_{k+1}-x_k)^4}{384} \max_{x\in[x_k,x_{k+1}]} |f^{(4)}(x)|.$$

2.0.36 第四节分段多项式插值

Piece-wise Polynomial Interpolation

2.0.37 Runge 现象

例 2.0.27 (考虑如下插值问题)**.** 设函数 $R(x)$ 为

$$R(x) = \frac{1}{1+x^2}, \quad x \in [-5, 5].$$

如果我们用等距插值节点

$$x_i = -5 + \frac{10i}{n}, \quad i = 0, 1, \ldots, n,$$

上的 Lagrange 插值多项式 $L_n(x)$ 来近似 $R(x)$，会发生什么现象?

2.0.38 Runge 现象

注 (Runge 现象). 高次多项式插值在区间端点处可能出现**剧烈振荡**，导致插值误差随节点增多而增大，而非减小, 即不一致收敛。

问题根源

- 全局高次多项式
- 函数本身性质
- 区间
- 等距节点

解决方案

- 选择**非等距节点**（如 Chebyshev 节点）
- 使用**分段低次多项式插值**
- 在局部保持光滑性即可

2.0.39 分段线性插值

定义 2.0.28 (分段线性插值)**.** 若 $\varphi_h(x)$ 在每个子区间 $[x_k, x_{k+1}]$ 上均为线性函数，则称其为**分段线性插值函数**。在每个子区间上：

$$\varphi_h(x) = \frac{x - x_{k+1}}{x_k - x_{k+1}} f_k + \frac{x - x_k}{x_{k+1} - x_k} f_{k+1}, \quad x_k \le x \le x_{k+1}.$$

2.0.40 分段线性插值的整体表示

$\varphi_h(x) = \sum_{j=0}^{n} f_j \, l_j(x)$，其中基函数 $l_j(x)$ 具有**局部非零性质**：

$$l_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j \quad (j \neq 0) \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1} \quad (j \neq n) \\ 0, & x \in [a,b] \setminus [x_{j-1}, x_{j+1}]. \end{cases}$$

2.0.41 分段线性插值误差

定理 2.0.29 (定理 4.13)**.** 设 $f(x) \in C[a,b]$，则 $\max_{a \le x \le b} |f(x) - \varphi_h(x)| \le \omega(h)$，从而 $\varphi_h(x)$ 在 $[a,b]$ 上一致收敛到 $f(x)$.

定理 2.0.30 (定理 4.14)**.** 设 $f(x) \in C^2[a,b]$，则

$$\|f - \varphi_h\|_\infty \le \frac{h^2}{8} \|f''\|_\infty.$$

2.0.42 分段三次 Hermite 插值

$$\varphi_h(x) = \sum_{j=0}^{n} \big[f(x_j) \, \alpha_j(x) + f'(x_j) \, \beta_j(x) \big],$$

$$\alpha_j(x) = \begin{cases} \left(\dfrac{x - x_{j-1}}{x_j - x_{j-1}} \right)^2 \left(1 + 2 \dfrac{x - x_j}{x_{j-1} - x_j} \right), & x_{j-1} \le x \le x_j \ (j \neq 0) \\ \left(\dfrac{x - x_{j+1}}{x_j - x_{j+1}} \right)^2 \left(1 + 2 \dfrac{x - x_j}{x_{j+1} - x_j} \right), & x_j \le x \le x_{j+1} \ (j \neq n) \\ 0, & \text{其他} \end{cases}$$

$$\beta_j(x)=\begin{cases}\left(\dfrac{x-x_{j-1}}{x_j-x_{j-1}}\right)^2(x-x_j), & x_{j-1}\le x\le x_j\ (j\neq 0)\\ \left(\dfrac{x-x_{j+1}}{x_j-x_{j+1}}\right)^2(x-x_j), & x_j\le x\le x_{j+1}\ (j\neq n)\\ 0, & \text{其他}\end{cases}$$

2.0.43 分段三次 Hermite 插值误差

定理 2.0.31 (定理 4.14)**.** 对目标函数 $f\in C^4[a,b]$，分段三次 *Hermite* 插值满足误差估计

$$\|f-\varphi_h\|_\infty\le\frac{h^4}{384}\|f^{(4)}\|_\infty.$$

2.0.44 三次样条函数

定义 2.0.32 (三次样条函数)**.** 对 $[a,b]$ 的剖分 $a=x_0<x_1<\cdots<x_n=b$，**三次样条函数** $S(x)$ 满足：

1. $S(x_j)=f(x_j)$，$j=0,1,\ldots,n$；
2. $S(x)\in C^2[a,b]$；
3. $S(x)$ 在每个子区间 $[x_k,x_{k+1}]$ 上是三次多项式。

2.0.45 三次样条的计算（一阶导数条件）

设 $S(x)=\sum_{j=0}^{n}\big[f(x_j)\alpha_j(x)+m_j\beta_j(x)\big]$，**一阶导数匹配**边界条件下 m_j 满足三对角方程组：

$$\begin{bmatrix}2&\mu_1&&\\ \lambda_2&2&\mu_2&\\ &\ddots&\ddots&\ddots\\ &&\lambda_{n-1}&2\end{bmatrix}\begin{bmatrix}m_1\\ m_2\\ \vdots\\ m_{n-1}\end{bmatrix}=\begin{bmatrix}g_1-\lambda_1f_0'\\ g_2\\ \vdots\\ g_{n-1}-\mu_{n-1}f_n'\end{bmatrix}$$

2.0.46 三次样条的计算（二阶导数条件）

设 $S(x)=\sum_{j=0}^{n}\big[f(x_j)\alpha_j(x)+m_j\beta_j(x)\big]$，**二阶导数匹配**边界条件下 m_j 满足：

$$\begin{bmatrix}2&1&&\\ \lambda_1&2&\mu_1&\\ &\ddots&\ddots&\ddots\\ &&1&2\end{bmatrix}\begin{bmatrix}m_0\\ m_1\\ \vdots\\ m_n\end{bmatrix}=\begin{bmatrix}g_0\\ g_1\\ \vdots\\ g_n\end{bmatrix}$$

2.0.47 三次样条的收敛性与误差

定理 2.0.33 (定理 4.14（收敛性）)**.** $S(x)$ 在 $[a,b]$ 上一致收敛到 $f(x)$。

定理 2.0.34 (定理 4.15（误差估计）)**.** 对一阶导数匹配边界条件及 $f \in C^4[a,b]$,

$$\|f - S\|_\infty \le \frac{5}{384}\, h^4 \, \|f^{(4)}\|_\infty.$$

2.0.48 总结

整体插值

- Lagrange 插值（基函数构造）
- 牛顿插值（差商递推，前/后插）
- Hermite 插值（同时匹配函数值和导数值）

分段插值

- 分段线性插值（$O(h^2)$ 误差）
- 分段三次 Hermite 插值（$O(h^4)$ 误差）
- 三次样条插值（C^2 光滑，$O(h^4)$ 误差）

参考文献：张平文、李铁军《数值分析》；李庆扬、王能超、易大义《数值分析》（第五版）

第三章　函数逼近

3.0.1　第四节函数逼近

Function Approximation

3.0.2　函数逼近的误差度量

定义 3.0.1 (函数逼近)**.** 设目标函数为 f，近似函数为 P。在区间 $[a,b]$ 上定义误差范数

$$\|f-P\|_p, \quad 1 \le p \le \infty.$$

常用情形：

$$\|f-P\|_\infty = \max_{x\in[a,b]} |f(x)-P(x)|,$$

$$\|f-P\|_2 = \left(\int_a^b |f(x)-P(x)|^2 \,\mathrm{d}x\right)^{1/2}.$$

3.0.3　最佳逼近

定义 3.0.2 (最佳逼近)**.** 设 $H \subseteq C[a,b]$。若 $P^* \in H$ 满足

$$\|f-P^*\|_p = \min_{P\in H} \|f-P\|_p,$$

则称 P^* 为 f 在 H 中关于 $\|\cdot\|_p$ 的最佳逼近。

- $p=\infty$：最佳一致逼近
- $p=2$：最佳平方逼近

3.0.4　Weierstrass 定理

定理 3.0.3 (定理 4.10)**.** 设 $f \in C[a,b]$。对任意 $\varepsilon>0$，存在代数多项式 P，使得

$$\|f-P\|_\infty < \varepsilon.$$

即：多项式在一致范数意义下可以逼近任意连续函数。

注 (说明). 常见证明思路使用 Bernstein 多项式。

3.0.5 Bernstein 多项式

定义 3.0.4 (Bernstein 逼近)**.** 在区间 $[0,1]$ 上定义

$$B_n(f,x)=\sum_{k=0}^{n} f\left(\frac{k}{n}\right)\ P_{k,n}(x),$$

其中

$$P_{k,n}(x)=\binom{n}{k}x^k(1-x)^{n-k}.$$

3.0.6 最佳一致逼近（minimax）

定义 3.0.5 (最佳一致逼近多项式)**.** 记

$$\Pi_n=\{\text{次数不超过}n\text{ 的多项式}\},\quad E_n(f)=\inf_{P\in\Pi_n}\|f-P\|_\infty.$$

若 $P_n^*\in\Pi_n$ 满足

$$\|f-P_n^*\|_\infty=E_n(f),$$

则称 P_n^* 为 f 在 Π_n 上的最佳一致逼近多项式。

注 (几何理解)*.* 它是在所有 n 次以下多项式中，使最大绝对误差最小的“最平坦”逼近。

3.0.7 交错定理（Chebyshev Alternation）

定理 3.0.6 (最佳一致逼近充要条件)**.** *设 $f\in C[a,b]$，$P_n\in\Pi_n$。则 P_n 是最佳一致逼近当且仅当存在*

$$a\le x_0<x_1<\cdots<x_{n+1}\le b,$$

使误差函数 $e(x)=f(x)-P_n(x)$ 在这些点上满足

$$e(x_k)=(-1)^k\sigma\,\|e\|_\infty,\quad \sigma\in\{\pm1\},\quad k=0,1,\dots,n+1.$$

即：最大误差至少在 $n+2$ 个点达到，且符号交替。

3.0.8 交错定理证明思路

注 (必要性思路)*.* 若 P_n^* 最优，但误差极值点不足 $n+2$ 个交错点，则可构造

$$q\in\Pi_n$$

使 $e^*=f-P_n^*$ 在主要极值点处与 q 同号。取

$$\widetilde{P} = P_n^* + \varepsilon q$$

并选足够小的 $\varepsilon > 0$，可使最大误差进一步下降，矛盾。

3.0.9 交错定理证明思路

注 (充分性思路). 若存在 $n+2$ 个交错点，任取另一多项式 $Q_n \in \Pi_n$，令

$$r = Q_n - P_n \in \Pi_n.$$

由交错符号可推出 r 在相邻区间内至少有 $n+1$ 次变号，迫使 r 至少有 $n+1$ 个零点。但 $\deg r \le n$，除非 $r \equiv 0$，从而不可能得到更小的 $\|f - Q_n\|_\infty$。

例题：x^2 的一次最佳一致逼近

例 3.0.7 (例). 在 Π_1 中，最佳一致逼近为

$$P_1^*(x) = \frac{1}{2},$$

对应误差函数

$$e(x) = x^2 - \frac{1}{2}, \quad \|e\|_\infty = \frac{1}{2}.$$

交错点可取

$$x_0 = -1,\ x_1 = 0,\ x_2 = 1,$$

并有

$$e(-1) = \frac{1}{2}, \quad e(0) = -\frac{1}{2}, \quad e(1) = \frac{1}{2}.$$

3.0.10 内积空间

定义 3.0.8 (内积空间). 设向量空间 X 定义在实域或复域 K 上，若映射 $\langle \cdot, \cdot \rangle$ 满足：

1. $\langle u, v \rangle = \overline{\langle v, u \rangle}$
2. $\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$
3. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$
4. $\langle u, u \rangle \ge 0$，且当且仅当 $u = 0$ 时取等号

则称 $(X, \langle \cdot, \cdot \rangle)$ 为内积空间。

3.0.11 内积空间性质 I

定理 3.0.9 (定理 4.11（Cauchy-Schwarz 不等式）). 对任意 $u, v \in X$，有

$$|\langle u, v\rangle|^2 \le \langle u, u\rangle \langle v, v\rangle.$$

3.0.12 内积空间性质 II

定理 3.0.10 (定理 4.12（Gram 矩阵）). 给定 $u_1, \dots, u_n \in X$，定义

$$\mathbf{G} = (g_{ij})_{n\times n}, \quad g_{ij} = \langle u_i, u_j\rangle.$$

则 $\mathbf{G}$ 非奇异，当且仅当 $u_1, \dots, u_n$ 线性无关。

3.0.13 权函数与加权 L_2 空间

定义 3.0.11 (权函数). 设区间 $[a, b]$（可有限或无限）上函数 $\rho(x)$ 满足：

1. 对 $k = 0, 1, 2, \dots$，积分 $\int_a^b x^k \rho(x)\,\mathrm{d}x$ 存在且有限
2. 对任意非负连续函数 g，若 $\int_a^b g(x)\rho(x)\,\mathrm{d}x = 0$，则 $g \equiv 0$

则称 ρ 为权函数。

由此可定义

$$\langle f, g\rangle = \int_a^b f(x)g(x)\rho(x)\,\mathrm{d}x, \quad \|f\|_2 = \sqrt{\langle f, f\rangle}.$$

3.0.14 线性无关与 Gram 行列式

定理 3.0.12 (定理 4.14). 设 $\varphi_1, \dots, \varphi_n \in C[a, b]$。它们线性无关当且仅当

$$\det \mathbf{G}(\varphi_1, \dots, \varphi_n) \neq 0.$$

3.0.15 正交多项式

定义 3.0.13 (正交性). 设 $g_0, g_1, \dots, g_m$ 为多项式，若在权函数 ρ 下

$$\langle g_i, g_j\rangle = \int_a^b g_i(x)g_j(x)\rho(x)\,\mathrm{d}x = A_i\,\delta_{ij},$$

则称它们关于 ρ 正交。

3.0.16 正交多项式三项递推

定理 3.0.14 (递推关系). 一组正交多项式 $\{\varphi_n\}$ 满足

$$\varphi_{n+1}(x) = (x - \alpha_n)\varphi_n(x) - \beta_n\varphi_{n-1}(x), \quad n \geq 0,$$

$$\varphi_0(x) = 1, \quad \varphi_{-1}(x) = 0,$$

其中

$$\alpha_n = \frac{\langle x\varphi_n, \varphi_n\rangle}{\langle \varphi_n, \varphi_n\rangle}, \quad \beta_n = \frac{\langle \varphi_n, \varphi_n\rangle}{\langle \varphi_{n-1}, \varphi_{n-1}\rangle}.$$

3.0.17 Legendre 多项式定义

定义 3.0.15 (Legendre 多项式). 在区间 $[-1, 1]$ 上定义

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - 1)^n, \quad n = 0, 1, 2, \ldots$$

其中 $P_0(x) = 1$。

3.0.18 Legendre 多项式性质 I

定理 3.0.16 (性质 4.1). 其最高次项系数为

$$a_n = \frac{(2n)!}{2^n(n!)^2}.$$

若将最高次项归一化为 1，则可写为

$$\widehat{P}_n(x) = \frac{n!}{(2n)!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - 1)^n.$$

3.0.19 Legendre 多项式性质 II

定理 3.0.17 (性质 4.2（正交性）).

$$\int_{-1}^{1} P_n(x)P_m(x)\,\mathrm{d}x = \frac{2}{2n+1}\delta_{nm}.$$

定理 3.0.18 (性质 4.3（奇偶性）).

$$P_n(-x) = (-1)^n P_n(x).$$

3.0.20 Legendre 多项式性质 III

定理 3.0.19 (性质 4.4（递推）)**.**

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x), \quad n \geq 1,$$

$$P_0(x) = 1, \quad P_1(x) = x.$$

3.0.21 Chebyshev 多项式定义

定义 3.0.20 (Chebyshev 多项式（第一类）)**.** 取权函数

$$\rho(x) = \frac{1}{\sqrt{1-x^2}}, \quad x \in [-1, 1],$$

定义

$$T_n(x) = \cos(n \arccos x), \quad -1 \leq x \leq 1,$$

等价地

$$T_n(\cos\theta) = \cos(n\theta), \quad 0 \leq \theta \leq \pi.$$

3.0.22 Chebyshev 多项式性质 I

定理 3.0.21 (性质 4.5（正交性）)**.**

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \begin{cases} 0, & n \neq m, \\ \pi/2, & n = m \neq 0, \\ \pi, & n = m = 0. \end{cases}$$

定理 3.0.22 (性质 4.6（递推）)**.**

$$T_0(x) = 1, \quad T_1(x) = x,$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1.$$

3.0.23 Chebyshev 多项式性质 II

定理 3.0.23 (性质 4.7（极小偏差）)**.** 在首项系数固定为 1 的 n 次多项式集合中，

$$2^{1-n}T_n(x)$$

到零函数的一致范数距离最小，最小值为 2^{1-n}。

定理 3.0.24 (性质 4.8（奇偶性）)**.** T_{2k} 仅含偶次幂，T_{2k+1} 仅含奇次幂。

3.0.24 Chebyshev 多项式性质 III

定理 3.0.25 (性质 4.9（零点）). T_n 的 n 个零点为

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, \dots, n.$$

例 3.0.26 (例 4.9). 求 $f(x) = 2x^3 + x^2 + 2x - 1$ 在 $[-1, 1]$ 上的最佳平方逼近多项式。

3.0.25 Chebyshev 零点插值误差

定理 3.0.27 (定理 4.17). 在 *Chebyshev* 零点上构造的 n 次 *Lagrange* 插值 L_n, 对 $f \in C^{n+1}[-1, 1]$ 满足

$$\|f - L_n\|_\infty \le \frac{1}{2^n(n+1)!}\,\|f^{(n+1)}\|_\infty.$$

例 3.0.28 (例 4.11). 对 $f(x) = \dfrac{1}{1+x^2}$ 在 $[-5, 5]$ 上分别使用 Chebyshev 节点与等距节点做插值，比较误差与振荡现象。

3.0.26 最小二乘逼近

定义 3.0.29 (最小二乘逼近). 设子空间

$$\Phi = \operatorname{span}\{\varphi_1, \varphi_2, \dots, \varphi_n\}.$$

寻找 $S^* \in \Phi$ 使

$$\|f - S^*\|_2 = \min_{S \in \Phi} \|f - S\|_2.$$

注 (等价正交条件). 误差函数 $e = f - S^*$ 与子空间 Φ 正交：$\langle e, \varphi_k \rangle = 0$。

3.0.27 正规方程

定理 3.0.30 (最小二乘解的线性系统). 令

$$S^*(x) = \sum_{k=1}^{n} a_k \varphi_k(x),$$

则系数满足正规方程

$$\mathbf{G}\mathbf{a} = \mathbf{d},$$

其中

$$G_{ij} = \langle \varphi_i, \varphi_j \rangle, \quad d_i = \langle f, \varphi_i \rangle.$$

3.0.28 正交基下的最小二乘系数

定理 3.0.31 (定理 4.18)**.** 若 $\{\varphi_k\}$ 是正交基，则

$$a_k^* = \frac{\langle f, \varphi_k \rangle}{\langle \varphi_k, \varphi_k \rangle}.$$

离散加权情形（节点 x_i，权重 w_i）可写为

$$a_k^* = \frac{\sum_{i=0}^{m} w_i f(x_i) \varphi_k(x_i)}{\sum_{i=0}^{m} w_i \varphi_k^2(x_i)}.$$

3.0.29 例 4.13

例 3.0.32 (例 4.13)**.** 对数据

x_i	1	2	3	4	5
f_i	4	4.5	6	8	8.5
w_i	2	1	3	1	1

构造加权最小二乘拟合曲线。

3.0.30 第五节三角逼近与傅里叶方法

Trigonometric Approximation and Fourier Methods

3.0.31 三角多项式与三角逼近

定义 3.0.33 (三角多项式空间)**.** 以周期区间 $[-\pi, \pi]$ 为例，定义

$$\mathcal{T}_n = \left\{ \frac{a_0}{2} + \sum_{k=1}^{n} \left(a_k \cos kx + b_k \sin kx \right) \right\}.$$

对周期函数 f，称

$$\min_{T_n \in \mathcal{T}_n} \|f - T_n\|_p$$

为 f 在 $\mathcal{T}_n$ 上的三角逼近问题。

注 (优势)*.* 对周期问题，三角基天然满足周期边界，较多项式逼近更稳定。

3.0.32 Fourier 部分和与最佳平方逼近

定理 3.0.34 (L_2 最优性)**.** 在内积

$$\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)g(x)\,\mathrm{d}x$$

下，$\mathcal{T}_n$ 的最佳平方逼近为 *Fourier* 部分和

$$S_n(f,x) = \frac{a_0}{2} + \sum_{k=1}^{n}\big(a_k\cos kx + b_k\sin kx\big),$$

其中

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\,\mathrm{d}x, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx\,\mathrm{d}x.$$

3.0.33 三角插值与三角逼近的联系

定义 3.0.35 (等距节点三角插值)**.** 取 $N = 2n+1$ 个节点

$$x_j = -\pi + \frac{2\pi j}{N}, \quad j = 0, 1, \ldots, N-1,$$

求 $I_n \in \mathcal{T}_n$ 使

$$I_n(x_j) = f(x_j), \quad j = 0, 1, \ldots, N-1.$$

注 (联系). 连续情形用积分得到 Fourier 系数；离散节点情形用求和得到插值系数。当采样足够细且函数光滑时，离散系数逼近连续 Fourier 系数。

3.0.34 从三角插值到 DFT

定理 3.0.36 (DFT 基本公式)**.** 设采样值 $f_j = f(x_j)$，记 $\omega_N = e^{-2\pi i/N}$，则离散 *Fourier* 变换为

$$\widehat{f}_k = \sum_{j=0}^{N-1} f_j\,\omega_N^{jk}, \quad k = 0, 1, \ldots, N-1.$$

逆变换

$$f_j = \frac{1}{N}\sum_{k=0}^{N-1} \widehat{f}_k\,\omega_N^{-jk}.$$

注 (解释). DFT 本质是把离散信号投影到离散三角基（复指数基）上。

3.0.35 DFT 与三角插值系数

定理 3.0.37 (系数对应关系)**.** 复形式三角插值可写为

$$I_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

在等距节点上有

$$c_k = \frac{1}{N}\widehat{f}_k \quad (\text{按频率重排后}).$$

故三角插值系数可由 DFT 一次性计算。

3.0.36 DFT 典型应用

- 频谱分析：识别主频、谐波、噪声能量分布
- 周期问题数值求解：谱方法中的导数与积分计算
- 卷积快速计算：时域卷积对应频域乘法
- 图像与信号处理：滤波、压缩、去噪、特征提取

3.0.37 第六节快速傅立叶变换

Fast Fourier Transform

3.0.38 FFT 的核心思想

定理 3.0.38 (偶奇分解（$N = 2^m$）)**.** 将 DFT

$$\widehat{f}_k = \sum_{j=0}^{N-1} f_j \omega_N^{jk}$$

按偶、奇下标拆分：

$$\widehat{f}_k = \sum_{r=0}^{N/2-1} f_{2r}\omega_{N/2}^{rk} + \omega_N^k \sum_{r=0}^{N/2-1} f_{2r+1}\omega_{N/2}^{rk}.$$

即把一个 N 点 DFT 化为两个 $N/2$ 点 DFT，再做蝶形合并。

3.0.39 复杂度与算法结构

定理 3.0.39 (复杂度下降)**.** 直接 DFT 计算量为 $O(N^2)$；FFT 满足递推

$$T(N) = 2T\left(\frac{N}{2}\right) + O(N),$$

因此

$$T(N) = O(N \log_2 N).$$

注 (实现要点). 1. 位逆序重排（bit-reversal permutation）

2. 分层蝶形迭代（stage-by-stage butterfly）
3. 预计算旋转因子（twiddle factors）

3.0.40 FFT 计算流程（迭代版）

例 3.0.40 (算法框架)**.** 设输入为复向量 $f[0:N-1]$:

1. 按位逆序重排数组
2. 对层数 $s=1,2,\dots,m$: 块长 $L=2^s$
3. 在每个块内执行蝶形:

$$u=f[p],\quad v=\omega_L^q f[p+L/2],$$

$$f[p]=u+v,\quad f[p+L/2]=u-v.$$

1. 输出即为 $\widehat{f}$。

3.0.41 FFT 的数值计算应用

例 3.0.41 (卷积加速)**.** 对长度分别为 n,m 的序列 a,b:

1. 零填充到长度 $N\geq n+m-1$（常取 2 的幂）
2. 计算 $\widehat{a}=\mathrm{FFT}(a),\ \widehat{b}=\mathrm{FFT}(b)$
3. 频域逐点乘积 $\widehat{c}_k=\widehat{a}_k\widehat{b}_k$
4. 逆 FFT 得卷积结果 c。

总复杂度由 $O(nm)$ 降为 $O(N\log N)$。

3.0.42 小结

- 连续函数可由多项式一致逼近（Weierstrass）
- 正交多项式为逼近与计算提供稳定基底
- Legendre 与 Chebyshev 是最常用的两类经典正交多项式
- 最小二乘逼近可由正规方程或正交展开直接求解
- 三角逼近与三角插值通过离散采样自然导向 DFT
- FFT 通过偶奇分解把 DFT 从 $O(N^2)$ 降到 $O(N\log N)$

第四章　数值积分

4.0.1　第一节数值积分

Numerical Integration

4.0.2　数值积分

定义 4.0.1 (数值积分)**.** 对区间 $[a,b]$ 上的连续函数 f，**数值积分**是指利用一组**节点** $x_j \in [a,b]$ 与**权重** $w_j \in \mathbb{R}$（$j=0,1,\ldots,n$）的有限求和来近似积分：

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{j=0}^{n} w_j f(x_j).$$

称 $\{(x_j,w_j)\}_{j=0}^{n}$ 为数值积分的**求积公式**。

4.0.3　近似误差与代数精度

定义 4.0.2 (近似误差与代数精度)**.** 数值积分的**近似误差**定义为

$$E_n(f) = \int_a^b f(x)\,\mathrm{d}x - \sum_{j=0}^{n} w_j f(x_j).$$

若 $E_n(x^m)=0$ 对 $m=0,1,\ldots,k$ 成立，而 $E_n(x^{k+1})\neq 0$，则称该求积公式具有 **k 阶代数精度 **。

4.0.4　代数精度示例

例 4.0.3 (例 5.1)**.** 求下列数值积分的代数精度：

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx 2f(0).$$

4.0.5　插值型数值积分

利用函数 $f(x)$ 的插值多项式可以近似其积分。设以 $\{(x_j,f(x_j))\}_{j=0}^{n}$ 为插值数据的 Lagrange 插值多项式为 $L_n(x)=\sum_{j=0}^{n} f(x_j)\,l_j(x)$，则

$$I_n := \int_a^b L_n(x)\,\mathrm{d}x = \sum_{j=0}^{n} w_j f(x_j) \approx \int_a^b f(x)\,\mathrm{d}x =: I,$$

其中权重

$$w_j = \int_a^b l_j(x)\,\mathrm{d}x.$$

称此类方法为**插值型数值积分**。

4.0.6 插值型数值积分的误差

由 Lagrange 插值余项，插值型数值积分的近似误差为

$$I_n - I = \int_a^b \big[L_n(x) - f(x)\big]\,\mathrm{d}x = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x)\,\mathrm{d}x,$$

其中 $\xi \in (a,b)$ 依赖于 x，且

$$\omega_{n+1}(x) = \prod_{j=0}^{n}(x - x_j).$$

4.0.7 Newton-Cotes 公式

取等距插值节点 $x_j = a + jh$，$h = (b-a)/n$，代入插值型数值积分得 **Newton-Cotes 公式**:

$$I_n = (b-a)\sum_{j=0}^{n} C_j^{(n)} f(x_j),$$

其中 **Cotes 系数**为

$$C_j^{(n)} = \frac{1}{b-a}\int_a^b l_j(x)\,\mathrm{d}x = \frac{(-1)^{n-j}}{j!\,(n-j)!}\cdot\frac{1}{n}\int_0^n \prod_{\substack{i=0\\ i\neq j}}^{n}(t-i)\,\mathrm{d}t.$$

4.0.8 Newton-Cotes 公式的特殊情形

- **$n = 1$（梯形公式）**

$$T = \frac{b-a}{2}\big[f(a) + f(b)\big].$$

- **$n = 2$（Simpson 公式）**

$$S = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

- **$n = 4$（Cotes 公式）**

$$C = \frac{b-a}{90}\big[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\big].$$

4.0.9 Newton-Cotes 公式的误差（n 为偶数）

定理 4.0.4 (定理 5.2)**.** 当 n **为偶数**时，设 $f \in C^{n+2}[a,b]$，*Newton-Cotes* 公式的误差为

$$E_n(f) = \frac{M_n}{(n+2)!} f^{(n+2)}(\xi), \quad \xi \in [a,b],$$

其中 $M_n = \int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x < 0$。

此时 *Newton-Cotes* 公式至少具有 $**(n+1)$ 阶代数精度 $**$。

4.0.10 Newton-Cotes 公式的误差（n 为奇数）

定理 4.0.5 (定理 5.3)**.** 当 n **为奇数**时，*Newton-Cotes* 公式的误差为

$$E_n(f) = \frac{M_n'}{(n+1)!} f^{(n+1)}(\xi), \quad \xi \in [a,b],$$

其中 $M_n' = \int_a^b \omega_{n+1}(x)\,\mathrm{d}x < 0$。

4.0.11 Newton-Cotes 公式示例

例 4.0.6 (例 5.2)**.** 分别用 $n = 1, 2, 3, 4$ 的 Newton-Cotes 公式计算积分

$$\int_{1.1}^{1.5} e^x\,\mathrm{d}x.$$

4.0.12 第二节复合求积公式

Composite Quadrature Rule

4.0.13 复合梯形公式

将 $[a,b]$ 等分为 n 个子区间 $[x_k, x_{k+1}]$，步长 $h = (b-a)/n$。

定义 4.0.7 (复合梯形公式)**.**

$$T_n = \frac{h}{2}\left[f(a) + 2\sum_{k=1}^{n-1} f(x_k) + f(b)\right].$$

近似误差为

$$E_n(f) = -\frac{b-a}{12} h^2 f''(\eta), \quad \eta \in [a,b].$$

4.0.14 复合 Simpson 公式

将 $[a,b]$ 等分为 n 个子区间，设 $x_{k+\frac{1}{2}}$ 为 $[x_k, x_{k+1}]$ 的中点。

定义 4.0.8 (复合 Simpson 公式)**.**

$$S_n = \frac{h}{6}\left[f(a) + 4\sum_{k=0}^{n-1} f\left(x_{k+\frac{1}{2}}\right) + 2\sum_{k=1}^{n-1} f(x_k) + f(b)\right].$$

近似误差为

$$E_n(f) = -\frac{b-a}{180}\left(\frac{h}{2}\right)^4 f^{(4)}(\eta), \quad \eta \in [a,b].$$

4.0.15 复合 Cotes 公式

将 $[a,b]$ 等分为 n 个子区间，设 $x_{k+\frac{1}{4}}$，$x_{k+\frac{1}{2}}$，$x_{k+\frac{3}{4}}$ 分别为各子区间的四等分点。

定义 4.0.9 (复合 Cotes 公式)**.**

$$C_n = \frac{h}{90}\left[7f(a) + 32\sum_{k=0}^{n-1} f\left(x_{k+\frac{1}{4}}\right) + 12\sum_{k=0}^{n-1} f\left(x_{k+\frac{1}{2}}\right) + 32\sum_{k=0}^{n-1} f\left(x_{k+\frac{3}{4}}\right) + 14\sum_{k=1}^{n-1} f(x_k) + f(b)\right].$$

近似误差为

$$E_n(f) = -\frac{2(b-a)}{945}\left(\frac{h}{4}\right)^6 f^{(6)}(\eta), \quad \eta \in [a,b].$$

4.0.16 复合求积公式示例

例 4.0.10 (例 5.3)**.** 分别用 $n=4$ 的复合梯形公式、复合 Simpson 公式和复合 Cotes 公式计算积分

$$\int_0^1 \frac{\sin x}{x}\,\mathrm{d}x.$$

4.0.17 第三节 Romberg 方法与 Richardson 外推

Romberg Rule and Richardson Extrapolation

4.0.18 Romberg 方法

Romberg 求积对梯形公式作递推加速，递归定义为

$$(T_1 f)(h) = \frac{h}{2}\sum_{k=1}^{n}\big[f(x_{k-1}) + f(x_k)\big],$$

$$(T_{j+1} f)(h) = \frac{4^j (T_j f)\left(\frac{h}{2}\right) - (T_j f)(h)}{4^j - 1}, \quad j = 1, 2, \ldots$$

定理 4.0.11 (定理 5.4)**.** *Romberg* 公式的误差满足

$$(T_{j+1}f)(h) - \int_a^b f(x)\,\mathrm{d}x = \mathcal{O}\Big(h^{2(j+1)}\Big).$$

4.0.19 梯形公式的二分递推

梯形公式满足如下二**分递推公式**:

$$T_{2n} = \frac{1}{2}T_n + \frac{h}{2}\sum_{k=0}^{n-1} f\Big(x_{k+\frac{1}{2}}\Big).$$

注 (意义). 每次步长减半时，只需计算新增中点处的函数值，无需重复计算已有节点，从而大幅节省计算量。

4.0.20 Richardson 外推

定理 4.0.12 (定理 5.5)**.** 梯形公式具有如下渐近展开:

$$T(h) = I + a_1h^2 + a_2h^4 + a_3h^6 + \cdots + a_kh^{2k} + \cdots$$

由此得到误差阶递增的组合:

$$T_1(h) := \frac{4}{3}T\big(\tfrac{h}{2}\big) - \frac{1}{3}T(h) = I + \beta_1h^4 + \beta_2h^6 + \cdots,$$

$$T_2(h) := \frac{16}{15}T_1\big(\tfrac{h}{2}\big) - \frac{1}{15}T_1(h) = I + \gamma_1h^6 + \gamma_2h^8 + \cdots$$

4.0.21 Richardson 外推的一般公式

一般地，对 $m = 1, 2, \ldots$:

$$T_m(h) := \frac{4^m}{4^m-1}T_{m-1}\big(\tfrac{h}{2}\big) - \frac{1}{4^m-1}T_{m-1}(h) = I + \delta_1h^{2(m+1)} + \delta_2h^{2(m+2)} + \cdots$$

注 (说明). 每提高一级外推，误差阶提高 2 阶，从而以较少的函数值求值实现高精度近似。

4.0.22 Romberg 算法

设 $T_0^{(k)}$ 为步长二分 k 次后的梯形公式值。

1. 计算 $T_0^{(0)} = \dfrac{b-a}{2}\big[f(a)+f(b)\big]$。
2. 利用二分递推公式计算 $T_0^{(k)}$。
3. 递归计算外推值:
$$T_m^{(k)} = \frac{4^m}{4^m-1}T_{m-1}^{(k+1)} - \frac{1}{4^m-1}T_{m-1}^{(k)}, \quad k = 1, 2, \ldots$$
1. 若 $\left|T_k^{(0)} - T_{k-1}^{(0)}\right| < \varepsilon$，输出 $T_k^{(0)}$；否则令 $k+1 \to k$，返回步骤 2。

4.0.23 Euler-Maclaurin 公式

Richardson 外推的理论基础是 **Euler-Maclaurin 公式**，它给出了梯形公式近似误差的精确渐近展开。

定理 4.0.13 (定理 5.6（Euler-Maclaurin 公式）)**.** *设 $f \in C^{2m+2}[a,b]$，复合梯形公式的计算值 $T(h)$ 满足如下展开式：*

$$T(h) = \int_a^b f(x)\,\mathrm{d}x + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} h^{2k} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right] + E_m(h),$$

其中 B_{2k} 为 Bernoulli 数，误差项 $E_m(h) = \mathcal{O}\big(h^{2m+2}\big)$。

4.0.24 Euler-Maclaurin 公式的证明（1）

证明. 为了利用分部积分提取出高阶导数的误差项，构造 **Bernoulli 多项式** $B_k(t)$，满足：

$$B_0(t) = 1, \quad B_k'(t) = kB_{k-1}(t), \quad \int_0^1 B_k(t)\,\mathrm{d}t = 0 \quad (k \geq 1).$$

重要性质：

- $B_1(t) = t - \frac{1}{2}$，边界值为 $B_1(0) = -\frac{1}{2}$，$B_1(1) = \frac{1}{2}$。
- 对 $k \geq 2$，边界值相等：$B_k(0) = B_k(1) = B_k$（即为 **Bernoulli 数**）。
- 对于所有的奇数 $k \geq 1$，Bernoulli 数为零：$B_{2k+1} = 0$。

□

4.0.25 Euler-Maclaurin 公式的证明（2）

证明. 在单个子区间 $[x_i, x_{i+1}]$ 上，引入变换 $x = x_i + th$（$t \in [0,1]$）。利用 $B_0(t) = B_1'(t)$ 进行第一步分部积分：

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x &= h \int_0^1 f(x_i + th) B_0(t)\,\mathrm{d}t = h \int_0^1 f(x_i + th)\,\mathrm{d}B_1(t) \\
&= h\left[f(x_{i+1})B_1(1) - f(x_i)B_1(0) \right] - h^2 \int_0^1 f'(x_i + th) B_1(t)\,\mathrm{d}t \\
&= \frac{h}{2}\left[f(x_{i+1}) + f(x_i) \right] - h^2 \int_0^1 f'(x_i + th) B_1(t)\,\mathrm{d}t.
\end{aligned}$$

第一项恰为该子区间上的**局部梯形公式**，第二项为误差余项。

□

4.0.26 Euler-Maclaurin 公式的证明（3）

证明. 对余项应用 $B_{k-1}(t) = \frac{1}{k} B_k'(t)$ 反复分部积分。由于 $B_k(1) = B_k(0) = B_k$：

$$\int_0^1 f^{(k)} B_k(t)\,\mathrm{d}t = \frac{B_{k+1}}{k+1} h \left[f^{(k)}(x_{i+1}) - f^{(k)}(x_i) \right] - \frac{h}{k+1} \int_0^1 f^{(k+1)} B_{k+1}(t)\,\mathrm{d}t.$$

对所有子区间 $i=0,\dots,n-1$ 求和，内部节点处的导数值互相**抵消（Telescoping）**，仅留下端点 a,b 处的导数值。

由于奇数阶 $B_{2k+1}=0$（$k\geq 1$），展开式中只含偶数阶幂次 h^{2k}，即得定理结论。■

□

4.0.27 第四节 Gauss 求积公式

Gauss Quadrature Rule

4.0.28 Gauss 节点

考虑以 $n+1$ 个节点的插值型数值积分

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{j=0}^{n} w_j f(x_j).$$

定义 4.0.14 (Gauss 节点)**.** 若上述数值积分具有 $2n+1$ 阶代数精度，则称节点 $x_j\,(j=0,1,\dots,n)$ 为 **Gauss 节点**，相应的求积公式称为 **Gauss 型求积公式**。

4.0.29 Gauss 节点的特征

定理 4.0.15 (定理 5.5)**.** *对于插值型数值积分，节点 x_j（$j=0,1,\dots,n$）为 Gauss 节点，当且仅当多项式*

$$\omega_{n+1}(x)=\prod_{j=0}^{n}(x-x_j)$$

与任意次数不超过 n 的多项式 $P(x)$ 正交，即

$$\int_a^b P(x)\,\omega_{n+1}(x)\,\mathrm{d}x=0,\quad \forall P\in\mathcal{P}_n.$$

4.0.30 定理 5.5 的证明（必要性）

证明. **必要性**：设 x_j 为 Gauss 节点，求积公式具有 $2n+1$ 阶代数精度。对于任意 $P\in\mathcal{P}_n$，多项式 $P(x)\omega_{n+1}(x)$ 的次数不超过 $2n+1$。由于公式对其积分精确成立，且 $\omega_{n+1}(x_j)=0$，故：

$$\int_a^b P(x)\omega_{n+1}(x)\,\mathrm{d}x=\sum_{j=0}^{n} w_j P(x_j)\omega_{n+1}(x_j)=0.$$

□

4.0.31 定理 5.5 的证明（充分性）

证明. **充分性**：设 $\omega_{n+1}(x)$ 与所有 $P \in \mathcal{P}_n$ 正交。对任意次数不超过 $2n+1$ 的多项式 $f(x)$，作带余除法得 $f(x) = Q(x)\omega_{n+1}(x) + R(x)$，其中 $Q, R \in \mathcal{P}_n$。利用正交性及插值型求积公式对 $R(x)$ 精确成立的事实（且 $f(x_j) = R(x_j)$），有：

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b R(x)\,\mathrm{d}x = \sum_{j=0}^{n} w_j R(x_j) = \sum_{j=0}^{n} w_j f(x_j).$$

因此节点为 Gauss 节点。■

□

4.0.32 带权函数的 Gauss 节点

设 $\rho(x) \geq 0$ 为 $[a,b]$ 上的权函数，考虑带权积分的近似：

$$\int_a^b \rho(x) f(x)\,\mathrm{d}x \approx \sum_{j=0}^{n} w_j f(x_j).$$

定义 4.0.16 (带权 Gauss 节点)**.** 若上式对所有次数不超过 $2n+1$ 的多项式 $f(x)$ 精确成立，则称 x_j（$j = 0, \ldots, n$）为关于权函数 $\rho(x)$ 的 **Gauss 节点**。

4.0.33 带权 Gauss 节点的确定

定理 4.0.17 (定理 5.6)**.** 设非负权函数 $\rho(x)$ 定义在 $[a,b]$ 上，$\{\varphi_k(x)\}$ 为关于 $\rho(x)$ 的正交多项式系。节点 x_j（$j = 0, 1, \ldots, n$）为 *Gauss* 节点，当且仅当它们是正交多项式 $\varphi_{n+1}(x)$ 的零点。权重由下式确定：

$$w_j = -\frac{a_{n+2}}{a_{n+1}} \frac{\displaystyle\int_a^b \rho(x)\big[\varphi_{n+1}(x)\big]^2\,\mathrm{d}x}{\varphi_{n+2}(x_j)\,\varphi'_{n+1}(x_j)}, \quad j = 0, 1, \ldots, n,$$

其中 a_n 为 $\varphi_n(x)$ 的首项系数。

4.0.34 定理 5.6 的证明（1）

证明. **1. 节点的确定**：根据正交多项式的性质，带权非负内积下的正交多项式 $\varphi_{n+1}(x)$ 之根必然都在 $[a,b]$ 内且互异，可作为合法的插值节点。仿照定理 5.5 的逻辑，这组节点构成的基多项式恰与 $\varphi_{n+1}(x)$ 仅相差非零常数倍。由于 $\varphi_{n+1}(x)$ 与所有 $\mathcal{P}_n$ 正交，因此这组节点满足带权 Gauss 节点的充要条件。

□

4.0.35 定理 5.6 的证明（2）

证明. **2. 权重的公式**：权重也可借由正交多项式的递推关系以及 Christoffel-Darboux 公式推导。此处 w_j 往往由 Lagrange 基函数的投影带权积分给出。结合求导公式求极限后，其值可被表示成定理给出的分式形式。此证明计算量大且核心依赖于正交多项式的三项递推律。

□

4.0.36 Gauss-Legendre 公式与 Gauss-Chebyshev 公式

- 当 $\rho(x)\equiv 1$ 时，**Legendre 多项式** $P_{n+1}(x)$ 的零点即为 $\displaystyle\int_{-1}^{1} f(x)\,\mathrm{d}x$ 的 Gauss 节点，对应的求积公式称为 **Gauss-Legendre 公式**。
- 当 $\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$ 时，**Chebyshev 多项式** $T_{n+1}(x)$ 的零点

$$x_j=\cos\left(\frac{2j+1}{2n+2}\pi\right),\quad j=0,1,\dots,n$$

为 $\displaystyle\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x$ 的 Gauss 节点，对应公式称为 **Gauss-Chebyshev 公式**。

4.0.37 Gauss-Legendre 公式示例

例 4.0.18 (例 5.4)**.** 求两节点 Gauss-Legendre 公式的节点和权重：

$$\int_{-1}^{1} f(x)\,\mathrm{d}x\approx w_0 f(x_0)+w_1 f(x_1).$$

4.0.38 带权 Gauss 公式示例

例 4.0.19 (例 5.5)**.** 构造如下带权积分的两节点 Gauss 公式：

$$\int_{0}^{1}\frac{f(x)}{\sqrt{x}}\,\mathrm{d}x\approx w_0 f(x_0)+w_1 f(x_1).$$

4.0.39 Gauss 公式的稳定性

设 $Q_n(f)=\displaystyle\sum_{j=0}^{n} w_j f(x_j)$ 为关于权函数 $\rho(x)$ 的 Gauss 求积公式。

定理 4.0.20 (定理 5.7（稳定性）)**.** *设 $\tilde{f}(x_j)$ 是 $f(x_j)$ 在节点 $x_0,x_1,\dots,x_n$ 处的近似值，则*

$$\left|Q_n(\tilde{f})-Q_n(f)\right|\le\max_{0\le j\le n}\left|\tilde{f}(x_j)-f(x_j)\right|\int_a^b\rho(x)\,\mathrm{d}x.$$

4.0.40 定理 5.7 的证明（1）

证明. **权重的正性**：对于带权 Gauss 公式，所有权重 w_j 严格为正。由于 $l_j(x)^2\in\mathcal{P}_{2n}$，且 Gauss 公式至少对 $\mathcal{P}_{2n}$ 精确，由 $l_j(x_k)=\delta_{jk}$ 可得：

$$w_j = \sum_{k=0}^{n} w_k l_j(x_k)^2 = \int_a^b \rho(x) l_j(x)^2 \,\mathrm{d}x > 0.$$

□

4.0.41 定理 5.7 的证明（2）

证明. **稳定性不等式**：由于 $1 \in \mathcal{P}_0$，公式对其积分精确，故 $\sum_{j=0}^{n} w_j = \int_a^b \rho(x)\,\mathrm{d}x$。利用权重正性与三角不等式：

$$\left|Q_n(\tilde{f}) - Q_n(f)\right| \le \sum_{j=0}^{n} w_j \left|\tilde{f}(x_j) - f(x_j)\right| \le \max_j \left|\tilde{f}(x_j) - f(x_j)\right| \int_a^b \rho(x)\,\mathrm{d}x.$$

结论得证。■

□

4.0.42 Gauss 公式的收敛性

定理 4.0.21（定理 5.8（收敛性））. *设 $f(x)$ 在 $[a,b]$ 上连续，则 Gauss 求积公式收敛：*

$$\lim_{n\to\infty} Q_n(f) = \int_a^b f(x)\,\rho(x)\,\mathrm{d}x.$$

4.0.43 定理 5.8 的证明

证明. 根据 Weierstrass 逼近定理，对于任意给定的 $\varepsilon > 0$，存在一个多项式 $P(x)$ 使得在 $[a,b]$ 上：

$$\max_{x\in[a,b]} |f(x) - P(x)| < \varepsilon.$$

取够大的 n 使得公式对 $P(x)$ 的代数精度 $2n+1 \ge \deg(P)$。利用公式对 $P(x)$ 精确、权重 $w_j > 0$ 的性质可估算积分误差：

$$\begin{aligned}
\left|\int_a^b f\rho - Q_n(f)\right| &\le \left|\int_a^b (f-P)\rho\right| + \left|\int_a^b P\rho - Q_n(P)\right| + |Q_n(P) - Q_n(f)| \\
&\le \varepsilon \int_a^b \rho(x)\,\mathrm{d}x + 0 + \sum_{j=0}^{n} w_j |P(x_j) - f(x_j)| \\
&\le \varepsilon \int_a^b \rho(x)\,\mathrm{d}x + \varepsilon \sum_{j=0}^{n} w_j = 2\varepsilon \int_a^b \rho(x)\,\mathrm{d}x.
\end{aligned}$$

因为上式对任意 $\varepsilon > 0$ 均成立，故当 $n \to \infty$ 时积分的近似值一定收敛于精确值。

□

第五章　常微分方程数值解

5.0.1　第一节 Euler 方法

Euler Methods

5.0.2　一阶常微分方程初值问题

本章考虑如下**一阶常微分方程初值问题**：

$$\begin{cases} y' = f(x,\, y), \\ y(x_0) = y_0. \end{cases}$$

数值求解的目标是在一列节点 $x_0 < x_1 < x_2 < \cdots$ 处求近似值 $y_n \approx y(x_n)$。取**等距步长** $h > 0$：

$$x_n = x_0 + nh, \quad n = 0, 1, 2, \ldots$$

注 (存在唯一性). 若 f 关于 y 满足 Lipschitz 条件，则初值问题存在唯一解 $y(x)$，数值方法以此为逼近目标。

5.0.3　Euler 方法（显式）

对 $y(x_{n+1})$ 在 x_n 处作一阶 Taylor 展开：

$$y(x_{n+1}) = y(x_n) + h\, y'(x_n) + \frac{h^2}{2}\, y''(\xi_n), \quad \xi_n \in (x_n, x_{n+1}).$$

略去 $\mathcal{O}(h^2)$ 余项并用 $f(x_n, y_n)$ 代替 $y'(x_n)$，得**显式 Euler 格式**：

$$\boxed{y_{n+1} = y_n + h\, f(x_n,\, y_n).}$$

注 (几何意义). 沿 (x_n, y_n) 处切线方向走一步，即以**切线斜率**作为区间 $[x_n, x_{n+1}]$ 上的平均斜率。

5.0.4　后退 Euler 方法（隐式）

改用 x_{n+1} 处的导数进行 Taylor 展开（向后差分），得**后退 Euler 格式**：

$$\boxed{y_{n+1} = y_n + h\, f(x_{n+1},\, y_{n+1}).}$$

此为关于 y_{n+1} 的**隐式方程**，一般需迭代求解。

	显式 Euler	后退 Euler
类型	显式	隐式
计算量	低	较高（需迭代）
稳定性	条件稳定	无条件稳定

注 (与显式格式的对比).

5.0.5 Euler 方法的局部截断误差

定理 5.0.1 (定理 6.1)**.** 显式与后退（隐式）*Euler* 方法的**局部截断误差**均为

$$y(x_{n+1}) - y_{n+1} = \frac{h^2}{2}\, y''(x_n) + \mathcal{O}(h^3),$$

即两种方法均具有 ***1*** **阶精度**。

证明. 对显式 Euler：将 $y(x_{n+1})$ 在 x_n 处 Taylor 展开

$$y(x_{n+1}) = y(x_n) + h\, y'(x_n) + \frac{h^2}{2}\, y''(x_n) + \mathcal{O}(h^3).$$

而 $y_{n+1} = y_n + h\, f(x_n, y_n) = y(x_n) + h\, y'(x_n)$（假设 $y_n = y(x_n)$），相减即得。

□

5.0.6 梯形方法

梯形方法以梯形公式近似积分 $\int_{x_n}^{x_{n+1}} f\,\mathrm{d}x$，从而得迭代格式：

$$y_{n+1}^{(k+1)} = y_n + \frac{h}{2}\big[f(x_n,\, y_n) + f(x_{n+1},\, y_{n+1}^{(k)})\big], \quad k = 0, 1, 2, \ldots$$

以显式 Euler 结果 $y_{n+1}^{(0)} = y_n + h\, f(x_n, y_n)$ 作为初始迭代值。

定理 5.0.2 (定理 6.2)**.** 设 f 关于 y 满足 *Lipschitz* 条件（常数 L），则梯形迭代的误差满足：

$$\big|y_{n+1} - y_{n+1}^{(k+1)}\big| \le \frac{hL}{2}\, \big|y_{n+1} - y_{n+1}^{(k)}\big|.$$

当 $hL < 2$ 时，迭代**收缩**，方法收敛。梯形方法具有 ***2*** **阶精度**。

5.0.7 定理 6.2 的证明

证明. 设 y_{n+1} 为真实隐式解，则

$$y_{n+1} = y_n + \frac{h}{2}\big[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big].$$

两式作差：

$$y_{n+1} - y_{n+1}^{(k+1)} = \frac{h}{2}\big[f(x_{n+1}, y_{n+1}) - f(x_{n+1}, y_{n+1}^{(k)})\big].$$

由 Lipschitz 条件 $|f(\cdot, a) - f(\cdot, b)| \le L|a - b|$，取绝对值：

$$\big|y_{n+1} - y_{n+1}^{(k+1)}\big| \le \frac{hL}{2}\big|y_{n+1} - y_{n+1}^{(k)}\big|.$$

当 $\dfrac{hL}{2} < 1$，即 $h < \dfrac{2}{L}$ 时，迭代为压缩映射，收敛。■

□

5.0.8 改进 Euler 方法（预测-校正）

将显式 Euler（预测）与梯形法（校正）各用一次，得**改进 Euler 格式**：

$$\begin{aligned} y_p &= y_n + h\, f(x_n,\, y_n) \quad \text{（预测）}, \\ y_c &= y_n + h\, f(x_{n+1},\, y_p), \\ y_{n+1} &= \frac{1}{2}(y_p + y_c) \quad \text{（校正）}. \end{aligned}$$

等价地写成：

$$y_{n+1} = y_n + \frac{h}{2}\big[f(x_n, y_n) + f(x_{n+1},\, y_n + h\, f(x_n, y_n))\big].$$

注 (精度). 可验证改进 Euler 方法具有 **2 阶精度**：局部截断误差为 $\mathcal{O}(h^3)$。

5.0.9 两步 Euler 方法

利用二阶中心差商，得到**两步 Euler 方法**：

$$\begin{aligned} \bar{y}_{n+1} &= y_{n-1} + 2h\, f(x_n,\, y_n), \\ y_{n+1} &= y_n + \frac{h}{2}\big[f(x_n,\, y_n) + f(x_{n+1},\, \bar{y}_{n+1})\big]. \end{aligned}$$

第一步用**跨步格式**给出粗预测，第二步用梯形公式校正。

定理 5.0.3 (定理 6.3)**.** *两步 Euler 方法的局部截断误差为*

$$y(x_{n+1}) - y_{n+1} = -\frac{h^3}{12}\, y'''(x_n) + \mathcal{O}(h^4) = \mathcal{O}(h^3),$$

即具有 2 **阶精度**。

5.0.10 第二节 Runge-Kutta 方法

Runge-Kutta Methods

5.0.11 p 阶精度的定义

定义 5.0.4 (局部截断误差与 p 阶精度)**.** 假设前一步精确，即 $y_n = y(x_n)$。将 y_{n+1} 代入真实解的比较，称

$$T_{n+1} = y(x_{n+1}) - y_{n+1}$$

为**局部截断误差**。若 $T_{n+1} = \mathcal{O}(h^{p+1})$，则称该方法具有 **$p$ 阶精度 **。

各方法精度汇总：

方法	局部截断误差	精度阶
显式 / 后退 Euler	$\mathcal{O}(h^2)$	**1 阶**
改进 Euler / 两步 Euler	$\mathcal{O}(h^3)$	**2 阶**
四阶 Runge-Kutta	$\mathcal{O}(h^5)$	**4 阶**

5.0.12 Taylor 级数方法

反复利用 $y' = f(x, y)$ 计算高阶导数，得 p 阶 **Taylor 级数格式**：

$$y_{n+1} = y_n + h\, y_n' + \frac{h^2}{2!}\, y_n'' + \cdots + \frac{h^p}{p!}\, y_n^{(p)}.$$

定理 5.0.5 (定理 6.4)**.** *p 阶 Taylor 级数方法的局部截断误差为*

$$y(x_{n+1}) - y_{n+1} = \frac{h^{p+1}}{(p+1)!}\, y^{(p+1)}(\xi), \quad \xi \in (x_n, x_{n+1}),$$

*因此具有 **p 阶精度 **。*

注 (缺点). 每步需要计算 f 的高阶偏导数，当 f 复杂时计算代价大。Runge-Kutta 方法以多次函数值评估代替高阶导数计算。

5.0.13 二阶 Runge-Kutta 方法的推导

取两个斜率的加权组合 $y_{n+1} = y_n + h(\lambda_1 K_1 + \lambda_2 K_2)$，其中

$$K_1 = f(x_n, y_n), \qquad K_2 = f(x_n + ph,\ y_n + ph\, K_1).$$

对 K_2 作二元 Taylor 展开：

$$K_2 = f + ph\, f_x + ph\, f\, f_y + \mathcal{O}(h^2),$$

与二阶 Taylor 方法 $y_{n+1} = y_n + h(f + \frac{h}{2}(f_x + f\, f_y)) + \mathcal{O}(h^3)$ 比较系数，得约束：

$$\lambda_1 + \lambda_2 = 1, \qquad \lambda_2 p = \frac{1}{2}.$$

注 (自由度). 约束有无穷多解。取 $\lambda_1 = \lambda_2 = \frac{1}{2}$, $p = 1$ 得**改进 Euler 方法**; 取 $\lambda_1 = 0$, $\lambda_2 = 1$, $p = \frac{1}{2}$ 得**中点方法**。

5.0.14 四阶 Runge-Kutta 方法

引入四个斜率, 得到最常用的**经典四阶 RK 格式**:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4), \\ K_1 &= f(x_n,\, y_n), \\ K_2 &= f\left(x_n + \tfrac{h}{2},\, y_n + \tfrac{h}{2}K_1\right), \\ K_3 &= f\left(x_n + \tfrac{h}{2},\, y_n + \tfrac{h}{2}K_2\right), \\ K_4 &= f(x_n + h,\, y_n + hK_3). \end{aligned}$$

定理 5.0.6 (定理 6.5)**.** *经典四阶 Runge-Kutta 方法的局部截断误差为 $\mathcal{O}(h^5)$, 具有 **4 阶精度**。*

5.0.15 Runge-Kutta 方法示例

例 5.0.7 (例 6.2)**.** 用 $h = 0.5$ 的**显式 Euler** 和 **四阶 RK** 方法求解:

$$\begin{cases} y'(x) = 1 + \dfrac{y}{x}, & x \in [1, 2], \\ y(1) = 2. \end{cases}$$

精确解为 $y(x) = x(\ln x + 2)$。

x_n	Euler y_n	RK4 y_n	精确值 $y(x_n)$	Euler 误差	RK4 误差
1.5	3.000	3.608	3.608	6.1×10^{-1}	1.0×10^{-6}
2.0	4.750	5.386	5.386	6.4×10^{-1}	3.5×10^{-6}

5.0.16 第三节收敛性与稳定性

Convergence and Stability

5.0.17 收敛性与单步方法

定义 5.0.8 (收敛性)**.** 若对固定的 $x_n = x_0 + nh$，当 $h \to 0$（$n \to \infty$，nh 有界）时有

$$y_n \xrightarrow{h\to 0} y(x_n),$$

则称该数值方法**收敛**。

定义 5.0.9 (显式单步方法)**. 显式单步方法**可统一写成

$$y_{k+1} = y_k + h\,\varphi(x_k,\, y_k,\, h),$$

其中 φ 称为**增量函数**。Euler 方法的增量函数为 $\varphi = f$；RK4 的 $\varphi = \frac{1}{6}(K_1+2K_2+2K_3+K_4)$。

5.0.18 单步方法的收敛定理

定理 5.0.10 (定理 6.6)**.** 设显式单步方法具有 p 阶精度（局部截断误差为 $\mathcal{O}(h^{p+1})$），且增量函数 φ 关于 y 满足 *Lipschitz* 条件：

$$\bigl|\varphi(x,y,h) - \varphi(x,y',h)\bigr| \le C_\varphi\,|y - y'|,$$

则该方法**收敛**，且整体误差（*global error*）满足：

$$\max_{0\le n\le N}\bigl|y(x_n) - y_n\bigr| = \mathcal{O}(h^p).$$

证明. 设 $e_n = y(x_n) - y_n$。由精度条件知局部误差 $\le Ch^{p+1}$；由 Lipschitz 条件知误差逐步传播满足 $|e_{n+1}| \le (1 + hC_\varphi)|e_n| + Ch^{p+1}$。递推并利用 $1 + t \le e^t$，得 $|e_n| \le \dfrac{C}{C_\varphi}(e^{C_\varphi(x_n - x_0)} - 1)\,h^p = \mathcal{O}(h^p)$。

□

5.0.19 稳定性

定义 5.0.11 (稳定性)**.** 若在某步 y_n 处施加扰动 δ，后续各步 y_m（$m > n$）的变化量满足

$$|y_m - \tilde{y}_m| \le C\,|\delta|$$

（C 与 h, m 无关），则称该方法**稳定**。

注 (稳定性 vs 收敛性).

- **收敛性**描述 $h \to 0$ 时格式的渐近行为；
- **稳定性**描述固定 h 下误差不被放大的能力。

两者均是数值方法质量的重要指标。

5.0.20 显式 Euler 稳定性分析

以**模型方程** $y' = \lambda y$（$\lambda \in \mathbb{C}$，$\mathrm{Re}(\lambda) < 0$）为测试问题。
显式 Euler 格式给出：

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda)\, y_n = (1 + h\lambda)^{n+1}\, y_0.$$

要使数值解不随步数增长（即**数值稳定**），需：

$$|1 + h\lambda| \le 1.$$

注 (实数情形). 当 $\lambda < 0$ 为实数时，稳定条件化为 $-2 \le \lambda h \le 0$，即步长须满足

$$h \le \frac{2}{|\lambda|}.$$

5.0.21 后退 Euler 稳定性分析

后退 Euler 格式对 $y' = \lambda y$ 给出：

$$y_{n+1} = y_n + h\lambda y_{n+1} \implies y_{n+1} = \frac{1}{1 - h\lambda}\, y_n = \left(\frac{1}{1 - h\lambda}\right)^{n+1} y_0.$$

稳定条件为 $\left|\dfrac{1}{1 - h\lambda}\right| \le 1$，即 $|1 - h\lambda| \ge 1$。
当 $\mathrm{Re}(\lambda) < 0$ 时，对**任意** $h > 0$ 均有 $|1 - h\lambda| > 1$，故后退 Euler 方法**无条件稳定**。

方法	稳定区域	对 $\lambda = -100$ 的步长限制
显式 Euler	$h \le 2/\lvert\lambda\rvert$	$h \le 0.02$
后退 Euler	任意 $h > 0$	无限制

注 (例 6.3 小结).

5.0.22 稳定性示例

例 5.0.12 (例 6.4)**.** 用显式和后退 Euler 方法（步长 $h = 0.05$）求解：

$$\begin{cases} y' = -100\, y, \\ y(0) = 1. \end{cases}$$

精确解 $y(x) = e^{-100x}$，在 $x = 0.1$ 时 $y(0.1) \approx 4.54 \times 10^{-5}$。

- **显式 Euler**：$h = 0.05$ 时 $|1 + h\lambda| = |1 - 5| = 4 > 1$，格式**不稳定**，数值解振荡发散。
- **后退 Euler**：$h = 0.05$ 时 $\dfrac{1}{|1 - (-5)|} = \dfrac{1}{6} < 1$，格式**稳定**，数值解单调趋零。

注 (刚性方程). $|\lambda|$ 很大时方程称为**刚性方程**，显式方法受步长严格限制，隐式方法更为实用。

5.0.23 第四节线性多步方法

Linear Multi-step Methods

5.0.24 多步方法的动机

单步方法每步仅利用 (x_n, y_n) 一个历史点。若充分利用已计算的多个历史值 $y_n, y_{n-1}, \dots$，则在相同函数值评估次数下可达到更高精度。

线性多步方法的一般形式为：

$$\sum_{j=0}^{r} \alpha_j \, y_{n-j} = h \sum_{j=0}^{r} \beta_j \, f_{n-j}, \quad \alpha_0 = 1,$$

其中 $f_{n-j} = f(x_{n-j}, y_{n-j})$。当 $\beta_{-1} \neq 0$（习惯记 j 从 -1 起）时为**隐式**，否则为**显式**。

注 (启动问题). r 步方法需要 r 个起始值 $y_0, y_1, \dots, y_{r-1}$，通常用单步方法高精度地计算。

5.0.25 积分形式与梯形方法

对 ODE 在 $[x_n, x_{n+1}]$ 积分：

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x,\, y(x))\, \mathrm{d}x.$$

用**梯形公式**近似右端积分，得**梯形格式**：

$$y_{n+1} = y_n + \frac{h}{2}\big[f(x_n,\, y_n) + f(x_{n+1},\, y_{n+1})\big].$$

注 (精度). 梯形格式是隐式二步方法，具有 **2 阶精度**，且绝对稳定区域包含整个左半复平面。

5.0.26 Adams 显式方法

用过 $(x_n, f_n), (x_{n-1}, f_{n-1}), \dots, (x_{n-r}, f_{n-r})$ 的 r 次插值多项式代替被积函数，积分后得 **Adams 显式（外推）格式**：

$$y_{n+1} = y_n + h \sum_{j=0}^{r} \alpha_j \, \Delta^j f_n, \qquad \alpha_j = (-1)^j \int_0^1 \binom{-t}{j} \mathrm{d}t.$$

避免差分计算的等价形式：

$$y_{n+1} = y_n + h \sum_{i=0}^{r} \beta_i \, f_{n-i}, \qquad \beta_i = (-1)^i \sum_{j=i}^{r} \binom{j}{i} \alpha_j.$$

注 (常用系数（$r = 3$，四阶）). $\beta_0 = \frac{55}{24}$, $\beta_1 = -\frac{59}{24}$, $\beta_2 = \frac{37}{24}$, $\beta_3 = -\frac{9}{24}$，精度阶为 **4 阶**。

5.0.27 Adams 隐式方法

类似地，用过 $(x_{n+1}, f_{n+1}), (x_n, f_n), \dots, (x_{n-r+1}, f_{n-r+1})$ 的插值多项式，得 **Adams 隐式（内插）格式**：

$$y_{n+1} = y_n + h \sum_{j=0}^{r} \alpha'_j \, \Delta^j f_{n-j+1}, \qquad \alpha'_j = (-1)^j \int_{-1}^{0} \binom{-t}{j} \mathrm{d}t.$$

直接系数形式：

$$y_{n+1} = y_n + h \sum_{i=0}^{r} \beta'_i \, f_{n-i+1}, \qquad \beta'_i = (-1)^i \sum_{j=i}^{r} \binom{j}{i} \alpha'_j.$$

注 (常用系数（$r = 3$，四阶）). $\beta'_0 = \frac{9}{24}$, $\beta'_1 = \frac{19}{24}$, $\beta'_2 = -\frac{5}{24}$, $\beta'_3 = \frac{1}{24}$，精度阶同样为 **4 阶**，但稳定性更好。

5.0.28 Taylor 展开法与系数确定

利用 Taylor 展开匹配，对格式

$$y_{n+1} = \alpha_0 y_n + \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + h\big(\beta_0 y'_n + \beta_1 y'_{n-1} + \beta_2 y'_{n-2} + \beta_3 y'_{n-3}\big)$$

要求对 $y = x^k$（$k = 0, 1, 2, 3, 4$）精确成立，得方程组：

$$\begin{aligned}
&\alpha_0 + \alpha_1 + \alpha_2 = 1,\\
&-\alpha_1 - 2\alpha_2 + \beta_0 + \beta_1 + \beta_2 + \beta_3 = 1,\\
&\alpha_1 + 4\alpha_2 - 2\beta_1 - 4\beta_2 - 6\beta_3 = 1,\\
&-\alpha_1 - 8\alpha_2 + 3\beta_1 + 12\beta_2 + 27\beta_3 = 1,\\
&\alpha_1 + 16\alpha_2 - 4\beta_1 - 32\beta_2 - 108\beta_3 = 1.
\end{aligned}$$

注 (特殊情形). 取 $\alpha_1 = \alpha_2 = 0$ 退化为**四阶 Adams 方法**；方程组有多组解对应不同的四阶多步格式。

5.0.29 Milne 方法

在 Taylor 展开法中取 $\alpha_0 = 0$，$\alpha_1 = 0$，$\alpha_2 = -1$，并令 $\beta_3 = 0$（以 y_{n-3} 替换 y'_{n-3}），得 **Milne 显式格式**：

$$y_{n+1} = y_{n-3} + \frac{4h}{3}\big(2y'_n - y'_{n-1} + 2y'_{n-2}\big).$$

定理 5.0.13 (局部截断误差)**.** *Milne 方法的局部截断误差为*

$$y(x_{n+1}) - y_{n+1} = \frac{14h^5}{45}\, y^{(5)}(\xi), \quad \xi \in (x_{n-3}, x_{n+1}),$$

具有 ***4 阶精度***。

5.0.30 Hamming 方法与预测-校正系统

Hamming 格式（隐式，4 阶）：

$$y_{n+1} = \frac{1}{8}(9y_n - y_{n-2}) + \frac{3h}{8}\big(2y'_{n+1} + 2y'_n - y'_{n-1}\big).$$

其局部截断误差为 $-\dfrac{h^5}{40}y^{(5)}(\xi)$。

注 (Milne–Hamming 预测-校正系统). 将两者组合，构成实用的**预测-校正（P-C）系统**：

1. **P（预测）**：Milne 显式格式给出 y^P_{n+1}；
2. **E（估计）**：计算 $f(x_{n+1}, y^P_{n+1})$；
3. **C（校正）**：Hamming 隐式格式用 y^P_{n+1} 计算 y^C_{n+1}；
4. **E（估计）**：计算 $f(x_{n+1}, y^C_{n+1})$ 供下一步使用。

每步仅需 **2 次**函数值计算，同时兼顾精度与稳定性。

第六章　蒙特卡洛方法

6.0.1　第一节维度灾难与引入

The Curse of Dimensionality

6.0.2　传统数值积分的困境

对于低维积分，常用求积公式（复化梯形、辛普森、Gauss 型等）有良好的收敛速度：节点数为 N 时，误差 $E = O(N^{-r})$。

注 (维度灾难). 考虑 d 维积分区域（如单位立方体 $[0,1]^d$）。每个维度布置 N 个节点，总节点数为

$$M = N^d.$$

计算代价随维度 d **呈指数增长**，即**维度灾难 (Curse of Dimensionality)**。

注 (例子). 若 $d = 30$（如美式期权定价），每维仅取 $N = 2$ 个节点，总计算次数 $2^{30} \approx 10^9$，高精度时完全无法承受。

6.0.3　蒙特卡洛方法的引入

定义 6.0.1 (蒙特卡洛估计量)**.** 将积分视作期望：

$$I = \int_\Omega f(\mathbf{x})\,\mathrm{d}\mathbf{x} = |\Omega|\,\mathbb{E}[f(X)] \quad X \sim U(\Omega)$$

独立抽取

$$X_1, \ldots, X_M \overset{\text{i.i.d.}}{\sim} U(\Omega)$$

构造 **MC 估计量**：

$$\hat{I}_M = \frac{|\Omega|}{M} \sum_{i=1}^{M} f(X_i).$$

6.0.4 无偏性与方差

定理 6.0.2 (基本统计性质). 设

$$X_1, \dots, X_M \overset{i.i.d.}{\sim} U(\Omega) \quad \sigma^2 = \mathrm{Var}(f(X)) < \infty,$$

则

$$\mathbb{E}[\hat{I}_M] = I, \qquad \mathrm{Var}(\hat{I}_M) = \frac{|\Omega|^2\sigma^2}{M}.$$

6.0.5 无偏性与方差

证明. 令 $Y_i = |\Omega| f(X_i)$，则 $Y_1, \dots, Y_M$ 是 i.i.d.，$\mathbb{E}[Y_i] = I$，$\mathrm{Var}(Y_i) = |\Omega|^2\sigma^2$，$\hat{I}_M = \bar{Y}_M$。

无偏性（期望的线性性）：

$$\mathbb{E}[\hat{I}_M] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[Y_i] = \frac{1}{M}\cdot M \cdot I = I.$$

方差（Y_i 两两独立，方差可加）：

$$\mathrm{Var}(\hat{I}_M) = \frac{1}{M^2}\sum_{i=1}^{M}\mathrm{Var}(Y_i) = \frac{M|\Omega|^2\sigma^2}{M^2} = \frac{|\Omega|^2\sigma^2}{M}. \qquad \square$$

$\square$

6.0.6 收敛性：大数定律

定理 6.0.3 (强大数定律（SLLN）). 若 $f \in L^1(\Omega)$（即 $\mathbb{E}[|f(X)|] < \infty$），则

$$\hat{I}_M = \frac{|\Omega|}{M}\sum_{i=1}^{M} f(X_i) \xrightarrow{a.s.} I.$$

6.0.7 收敛性：大数定律（证明）

证明. 第一步：变量替换。令 $Y_i = |\Omega|\, f(X_i)$，$i = 1, \dots, M$。由于 $X_1, \dots, X_M \overset{\text{i.i.d.}}{\sim} U(\Omega)$，$f$ 的可测性保证 $Y_1, \dots, Y_M$ 也是 i.i.d.，且

$$\hat{I}_M = \frac{1}{M}\sum_{i=1}^{M} Y_i = \bar{Y}_M.$$

第二步：验证 SLLN 的条件。

1. **独立同分布**：$Y_i = |\Omega| f(X_i)$，X_i 互相独立且同为 $U(\Omega)$，故 Y_i i.i.d. ✓
2. **有限一阶矩**：由 $f \in L^1(\Omega)$，$\mathbb{E}[|Y_1|] = |\Omega|\,\mathbb{E}[|f(X)|] = \displaystyle\int_\Omega |f|\,\mathrm{d}\mathbf{x} < \infty$. ✓
3. **期望**：$\mathbb{E}[Y_1] = |\Omega|\,\mathbb{E}[f(X)] = \displaystyle\int_\Omega f\,\mathrm{d}\mathbf{x} = I$. ✓

$\square$

6.0.8 收敛性：大数定律（证明）

证明. 第三步：应用 Kolmogorov 强大数定律。对满足上述两条件的 i.i.d. 序列，

$$\bar{Y}_M = \frac{1}{M}\sum_{i=1}^{M} Y_i \xrightarrow{\text{a.s.}} \mathbb{E}[Y_1] = I.$$

因此 $\hat{I}_M = \bar{Y}_M \to I$ 以概率 1。□

□

6.0.9 收敛性：中心极限定理

定理 6.0.4 (中心极限定理（CLT）). *若 $f \in L^2(\Omega)$，$\sigma^2 = \mathrm{Var}(f(X)) < \infty$，则*

$$\sqrt{M}(\hat{I}_M - I) \xrightarrow{d} \mathcal{N}(0,\, |\Omega|^2\sigma^2).$$

6.0.10 收敛性：中心极限定理（证明）

证明. 第一步：变量替换。令 $Y_i = |\Omega| f(X_i)$，则 $Y_1, \ldots, Y_M$ i.i.d.，且

$$\mathbb{E}[Y_i] = I, \quad \mathrm{Var}(Y_i) = |\Omega|^2\sigma^2 =: \tau^2 < \infty, \quad \hat{I}_M = \bar{Y}_M.$$

第二步：标准化。对样本均值 $\bar{Y}_M$ 作中心化与归一化：

$$Z_M = \frac{\bar{Y}_M - \mathbb{E}[Y_1]}{\tau/\sqrt{M}} = \frac{\sqrt{M}(\hat{I}_M - I)}{|\Omega|\sigma}.$$

第三步：应用 Lindeberg-Lévy CLT。Y_i i.i.d.、$\mathbb{E}[Y_i^2] < \infty$，故

$$Z_M \xrightarrow{d} \mathcal{N}(0,1) \quad (M \to \infty).$$

第四步：还原尺度。由连续映射定理，$\sqrt{M}(\hat{I}_M - I) = \tau Z_M$ 的极限分布为

$$\sqrt{M}(\hat{I}_M - I) \xrightarrow{d} \mathcal{N}(0, \tau^2) = \mathcal{N}(0, |\Omega|^2\sigma^2).$$

即得 CLT 结论。□

□

6.0.11 半阶收敛率

推论 6.0.5 (半阶收敛率). *MC 估计的均方根误差（RMSE）满足*

$$\mathrm{RMSE}(\hat{I}_M) = \sqrt{\mathrm{Var}(\hat{I}_M)} = \frac{|\Omega|\sigma}{\sqrt{M}} = O\Big(M^{-1/2}\Big),$$

*** 与维度 d 完全无关 **，从根本上克服维度灾难。*

6.0.12 半阶收敛率：与传统方法的对比

注 (与传统求积公式的对比). 传统 d 维求积公式（网格法）误差为 $O(N^{-r/d})$，其中 N 为总节点数，r 为方法阶次。

	MC 方法	网格法（r 阶）
误差	$O(M^{-1/2})$	$O(N^{-r/d})$
维度依赖	**无**	指数衰减

随 d 增大，MC 的优势指数级扩大。

问题：$O(M^{-1/2})$ 的收敛率已经与维度无关，但实际计算中还面临什么困难?

6.0.13 第二节减少方差的方法

Variance Reduction Techniques

6.0.14 为什么需要减少方差？

注 (收敛速率的代价). 由中心极限定理，MC 的误差（以 RMSE 度量）为

$$\mathrm{RMSE}(\hat{I}_M) = \frac{|\Omega|\sigma}{\sqrt{M}} = O\Big(M^{-1/2}\Big).$$

精度提高 **10 倍**，需将采样数 M 增大 **100 倍**——计算代价以平方速率增长。

核心思路：在固定计算预算下，** 降低方差 σ^{2}** 比单纯增大 M 更高效。

6.0.15 重要性抽样

定义 6.0.6 (重要性抽样)**.** 若存在密度 $g(x)$ 满足 $f(x) \neq 0 \Rightarrow g(x) > 0$，则积分可改写为关于 g 的期望：

$$I = \int f(x)\,\mathrm{d}x = \int \frac{f(x)}{g(x)} \cdot g(x)\,\mathrm{d}x = \mathbb{E}_g\left[\frac{f(X)}{g(X)}\right], \quad X \sim g.$$

估计量：$\hat{I}_M = \dfrac{1}{M}\displaystyle\sum_{i=1}^{M} \frac{f(X_i)}{g(X_i)}$，其中 $X_i \overset{\text{i.i.d.}}{\sim} g$。选取与 $|f|$ 形状相似的 g 可显著减小方差。

6.0.16 控制变量法

定义 6.0.7 (控制变量法)**.** 设 $I = \mathbb{E}[f(X)]$，已知辅助函数 h 满足 $\mathbb{E}[h(X)] = c$（解析可得）。对任意 $\lambda \in \mathbb{R}$，令

$$Y_i = f(X_i) - \lambda\,(h(X_i) - c),$$

则 $\mathbb{E}[Y_i] = I$（无偏），但 $\mathrm{Var}(Y_i)$ 随 λ 的选取而变化。

直觉：若 h 与 f 高度相关，则 $h(X_i) - c$ 的波动可”对冲” $f(X_i)$ 的波动，从而降低方差。

6.0.17 控制变量法：最优系数

注 (最优系数推导). 展开方差：

$$\mathrm{Var}(Y) = \mathrm{Var}(f) + \lambda^2\,\mathrm{Var}(h) - 2\lambda\,\mathrm{Cov}(f,h).$$

对 λ 求导置零，得最优系数与对应方差：

$$\lambda^* = \frac{\mathrm{Cov}(f,h)}{\mathrm{Var}(h)}, \qquad \mathrm{Var}(Y^*) = (1-\rho^2)\,\mathrm{Var}(f),$$

其中 $\rho = \mathrm{Corr}(f(X), h(X))$。$|\rho|$ 越接近 1，方差缩减越显著。

6.0.18 分层抽样 (Stratified Sampling)

定义 6.0.8 (分层抽样估计量)**.** 将 Ω 划分为 K 个不相交子域 $\{\Omega_k\}$，令权重 $p_k = |\Omega_k|/|\Omega|$ $(\sum_k p_k = 1)$。在 Ω_k 中独立抽取 M_k 个样本 $X_{k,i} \sim U(\Omega_k)$，**分层估计量**为

$$\hat{I}_{\mathrm{strat}} = |\Omega| \sum_{k=1}^{K} p_k \cdot \frac{1}{M_k} \sum_{i=1}^{M_k} f(X_{k,i}).$$

无偏性：令 $\mu_k = \mathbb{E}[f(X) \mid X \in \Omega_k]$，则 $\mathbb{E}[\hat{I}_{\mathrm{strat}}] = |\Omega| \sum_k p_k \mu_k = I$。

6.0.19 分层抽样：方差分析

定理 6.0.9 (分层抽样方差优越性（比例分配）)**.** 令 $\sigma_k^2 = \mathrm{Var}(f(X)|X \in \Omega_k)$，$M_k = p_k M$（比例分配），则

$$\mathrm{Var}(\hat{I}_{strat}) = \frac{|\Omega|^2}{M} \sum_{k=1}^{K} p_k \sigma_k^2 \le \mathrm{Var}(\hat{I}_M) = \frac{|\Omega|^2 \sigma^2}{M}.$$

6.0.20 分层抽样：方差分析（证明）

证明. **分层方差**：各层独立，$\mathrm{Var}(\hat{I}_{\mathrm{strat}}) = |\Omega|^2 \sum_k p_k^2 \dfrac{\sigma_k^2}{M_k} = \dfrac{|\Omega|^2}{M} \sum_k p_k \sigma_k^2$。

全方差分解（条件方差公式）：令 K 为层指标，$\bar{\mu} = \sum_k p_k \mu_k = I/|\Omega|$，

$$\sigma^2 = \mathrm{Var}(f) = \underbrace{\mathbb{E}[\mathrm{Var}(f|K)]}_{=\sum_k p_k \sigma_k^2} + \underbrace{\mathrm{Var}(\mathbb{E}[f|K])}_{=\sum_k p_k(\mu_k - \bar{\mu})^2} \ge 0$$

故 $\sigma^2 \geq \sum_k p_k \sigma_k^2$，即 $\mathrm{Var}(\hat{I}_{\mathrm{strat}}) \leq \mathrm{Var}(\hat{I}_M)$。□

□

Neyman 最优分配：在总量 M 固定下令 $M_k \propto p_k \sigma_k$ 可进一步最小化 $\mathrm{Var}(\hat{I}_{\mathrm{strat}})$。

6.0.21 对偶变量法 (Antithetic Variates)

定义 6.0.10 (对偶估计量)**.** 利用 $U \sim U(0,1) \Rightarrow 1-U \sim U(0,1)$ 的对称性，构造配对估计量（以 $\Omega = [0,1]$ 为例）：

$$\hat{I}_{\mathrm{AV}} = \frac{1}{2M} \sum_{i=1}^{M} \big[f(U_i) + f(1-U_i) \big], \quad U_i \overset{\text{i.i.d.}}{\sim} U(0,1).$$

高维推广：用 $(\mathbf{U}_i, \mathbf{1} - \mathbf{U}_i)$ 配对，或通过 $F^{-1}(U)$ 与 $F^{-1}(1-U)$ 生成对偶样本。

方差公式：$\mathrm{Var}(\hat{I}_{\mathrm{AV}}) = \dfrac{\sigma^2 + \mathrm{Cov}(f(U), f(1-U))}{2M}$。当协方差为负时方差小于普通 MC 的 σ^2/M，详见下页证明。

6.0.22 对偶变量法：方差减少的证明

当 f 单调时 $\mathrm{Cov}(f(U), f(1-U)) \leq 0$

证明. 设 $U, V \overset{\text{i.i.d.}}{\sim} U(0,1)$，$f$ 单调不减。注意：

- 若 $U > V$，则 $f(U) \geq f(V)$（f 不减），而 $1-U < 1-V$ 故 $f(1-U) \leq f(1-V)$；
- 若 $U < V$，类似地两个因子异号。

因此对所有 U, V：

$$\big(f(U) - f(V)\big)\big(f(1-U) - f(1-V)\big) \leq 0.$$

对期望展开（注意 $1-V \sim U(0,1)$ 与 U 独立）：

$$2\mathbb{E}[f(U)f(1-U)] - 2\big(\mathbb{E}[f(U)]\big)^2 \leq 0,$$

即 $\mathrm{Cov}(f(U), f(1-U)) \leq 0$，故 $\mathrm{Var}(\hat{I}_{\mathrm{AV}}) \leq \dfrac{\sigma^2}{2M} < \dfrac{\sigma^2}{M}$。□

□

6.0.23 第三节随机数生成

Random Number Generation

6.0.24 伪随机数发生器

计算机生成**伪随机数**，使其统计特性与真实均匀独立样本无法区分。

定义 6.0.11 (线性同余发生器 (LCG))**.** 通过迭代公式生成 $\{0,1,\ldots,m-1\}$ 上的序列：

$$X_{n+1} = (aX_n + c) \bmod m.$$

其中 a（乘子）、c（增量）、m（模数）是精心选取的大整数。令 $U_n = X_n/m$ 即得 $[0,1)$ 的均匀伪随机数。

注 (现代生成器). **Mersenne Twister (MT19937)** 是主流标准，周期 $2^{19937}-1$，通过了所有标准统计测试。

6.0.25 逆变换定理

定理 6.0.12 (逆变换定理)**.** 设 $F(x)$ 为连续严格单调的 *CDF*，$U \sim \text{Uniform}(0,1)$，则

$$X = F^{-1}(U)$$

服从分布 F，即 $\mathbb{P}(X \le x) = F(x)$。

6.0.26 逆变换法（证明）

证明. 设 $X = F^{-1}(U)$，对任意 $x \in \mathbb{R}$ 验证 $\mathbb{P}(X \le x) = F(x)$：

第一步：由定义，$\mathbb{P}(X \le x) = \mathbb{P}(F^{-1}(U) \le x)$。

第二步：F 连续且严格单调，故

$$F^{-1}(U) \le x \iff U \le F(x), \quad \text{因此} \mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)).$$

第三步：$U \sim U(0,1)$ 的 CDF 为 $\mathbb{P}(U \le t) = t$，且 $F(x) \in [0,1]$，故

$$\mathbb{P}(U \le F(x)) = F(x). \qquad \square$$

$\square$

注 (局限性). 当 F^{-1} 无解析表达式时（如正态分布），需用其他方法（如 Box-Muller 变换或舍选法）。

6.0.27 舍选法 (Rejection Sampling)

定义 6.0.13 (舍选法原理)**.** 欲从密度 $f(x)$ 采样。寻找包络密度 $g(x)$ 及常数 $c \ge 1$，使

$$f(x) \le c \cdot g(x), \quad \forall x.$$

算法：

1. 产生 $Y \sim g$，独立地产生 $U \sim U(0,1)$
2. 若 $U \le \dfrac{f(Y)}{c\,g(Y)}$，令 $X = Y$（接受）；否则拒绝，回到第 1 步

注 (几何直观与效率). 在 $c \cdot g(x)$ 下方均匀投点，只保留落在 $f(x)$ 下方的点。接受概率为 $1/c$，故 c 越接近 1 越高效。

6.0.28 舍选法：正确性证明

证明. 设 $Y \sim g$，$U \sim U(0,1)$ 独立，接受事件 $A = \{U \le f(Y)/(cg(Y))\}$。

接受概率：

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \frac{f(y)}{cg(y)} \cdot g(y)\,\mathrm{d}y = \frac{1}{c}\int f(y)\,\mathrm{d}y = \frac{1}{c}.$$

联合概率：

$$\mathbb{P}(Y \le x,\, A) = \int_{-\infty}^{x} \frac{f(y)}{cg(y)} \cdot g(y)\,\mathrm{d}y = \frac{F(x)}{c}.$$

条件 CDF：

$$\mathbb{P}(Y \le x \mid A) = \frac{\mathbb{P}(Y \le x,\, A)}{\mathbb{P}(A)} = \frac{F(x)/c}{1/c} = F(x). \qquad \square$$

□

6.0.29 第四节拓展与应用

Extensions and Applications

6.0.30 拟蒙特卡洛 (QMC)

定义 6.0.14 (低差异序列与积分误差)**.** 使用确定性的 _ 低差异序列（Halton、Sobol、Niederreiter 等）_ 代替伪随机数，更均匀地填满多维空间, 星差

$$D_M^* = O((\log M)^d / M)$$

定理 6.0.15 (Koksma-Hlawka 不等式)**.** *设 f 在 $[0,1]^d$ 上的 Hardy-Krause 变差为 $V(f) < \infty$，则*

$$\left| \frac{1}{M}\sum_{i=1}^{M} f(\mathbf{x}_i) - \int_{[0,1]^d} f\,\mathrm{d}\mathbf{x} \right| \le V(f) \cdot D_M^*.$$

取 Sobol 序列时误差为 $O(M^{-1}(\log M)^d)$，在中等维度显著优于 MC 的 $O(M^{-1/2})$。

6.0.31 马尔可夫链蒙特卡洛 (MCMC)

定义 6.0.16 (Metropolis-Hastings 算法)**.** 目标：从密度 $\pi(x)$ 采样。给定当前状态 x：

1. 按建议分布 $q(\cdot|x)$ 生成候选 y
2. 以概率 $\alpha(x,y)=\min\left(1,\dfrac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right)$ 接受 y，否则保持 x

6.0.32 细致平衡（正确性）

证明. 第一步：验证细致平衡。设 $p(x,y)=q(y|x)\alpha(x,y)$ 为转移密度，对 $x\neq y$：

$$\pi(x)p(x,y)=\pi(x)q(y|x)\min\left(1,\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right)=\min\big(\pi(x)q(y|x),\,\pi(y)q(x|y)\big)=\pi(y)p(y,x).$$

第二步：推导平稳性。对细致平衡条件关于 x 积分：

$$\int \pi(x)\,p(x,y)\,\mathrm{d}x=\int \pi(y)\,p(y,x)\,\mathrm{d}x=\pi(y)\underbrace{\int p(y,x)\,\mathrm{d}x}_{=1}=\pi(y).\quad\square$$

因此 π 是 M-H 链的平稳分布——无需知道归一化常数即可采样。

□

第七章　凸优化

7.1　第七章凸优化—讲稿

本讲稿配合幻灯片 `optimization.md` 使用，按七个小节组织。每小节包含：教学目标、核心概念的详细解释、完整证明、例题精讲、易错点提醒和教学建议。

主要参考文献

：Stephen Boyd & Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.（以下简称"Boyd"）

7.2　第一节凸集

7.2.1　教学目标

学完本节后，学生应能：

1. 判断给定集合是否为凸集、仿射集或锥
2. 运用保凸运算构造新的凸集
3. 陈述分离超平面定理并理解其几何含义
4. 计算简单集合的对偶锥

7.2.2　1.1 优化问题的基本形式

讲解思路

这一页是整章的引子。不需要花太多时间，重点是让学生看到"优化问题的结构"和"为什么凸性重要"。

逐点讲解：

1. **标准形式**：

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

这是所有优化问题的统一写法。要点：- $x \in \mathbb{R}^n$ 是我们要找的决策变量 - f_0 是目标函数

——我们希望它尽可能小 - $f_i(x) \le 0$ 是不等式约束——可行解必须满足 - $h_i(x) = 0$ 是等式约束——同样是必须满足的硬性条件

2. **可行集**：$\mathcal{F} = \{x \mid f_i(x) \le 0,\ h_i(x) = 0\}$。可以问学生：” 可行集是什么形状？”——这就是本章要研究的核心问题。
3. **最优值与最优解**：- 最优值 $p^* = \inf\{f_0(x) \mid x \in \mathcal{F}\}$ 用的是下确界 inf，不是最小值 min。为什么？因为最小值可能达不到。举个例子：$\min_{x>0} 1/x$——下确界是 0，但永远达不到 0。- 如果存在 $x^* \in \mathcal{F}$ 使得 $f_0(x^*) = p^*$，x^* 就是最优解。

教学提示：可以花 2-3 分钟讲这个框架，让学生回忆一下线性规划中学过的标准形式。强调” 标准形式不是限制——任何优化问题都可以写成这种形式”。

7.2.3 1.2 仿射集合

定义回顾

仿射集合：集合 C 称为仿射的，如果对任意 $x_1, x_2 \in C$ 和任意 $\theta \in \mathbb{R}$，都有 $\theta x_1 + (1-\theta)x_2 \in C$。

详细解释

- 几何意义：仿射集合包含其中**任意两点所确定的整条直线**（不只是线段）。
- 凸集只需要包含线段（$\theta \in [0,1]$），而仿射集需要包含整条直线（$\theta \in \mathbb{R}$）。
- 因此：**仿射集一定是凸集，但凸集不一定是仿射集**。

例子：

- $\mathbb{R}^n$ 本身是仿射集
- 线性方程组的解集 $\{x \mid Ax = b\}$ 是仿射集。证：若 $Ax_1 = b, Ax_2 = b$，则 $A(\theta x_1 + (1-\theta)x_2) = \theta b + (1-\theta)b = b$。
- 反之，任意仿射集都可以写成线性方程组的解集。

仿射包：包含集合 D 的最小仿射集。

$$\mathbf{aff}\, D = \left\{ \sum_{i=1}^{k} \theta_i x_i \;\middle|\; x_i \in D,\ \sum_{i=1}^{k} \theta_i = 1 \right\}$$

注意这里的系数和必须为 1，但系数可以是任意实数（可正可负）。

教学提示：仿射集不是重点，但它是理解凸集的铺垫。关键是让学生区分” 仿射集（直线）” 和” 凸集（线段）”。

7.2.4 1.3 凸集

定义

集合 C 为凸集，如果对任意 $x_1, x_2 \in C$ 和 $\theta \in [0,1]$：

$$\theta x_1 + (1-\theta)x_2 \in C$$

核心讲解

这是整章最重要的定义之一。需要反复强调。

直观理解：

- 拿粉笔在黑板上画一个圆：圆内任意两点的连线全部在圆内 → 圆盘是凸集。
- 画一个月牙形：存在两点，其连线的某段在月牙外 → 月牙不是凸集。
- ”凸”在中文里的意思是”向外鼓出”，但数学中的 convex set 实际上是”内部无凹陷”的意思。

凸集和非凸集的对比讲解：

凸集的例子（逐一说明为什么是凸集）：

1. **空集**：空集平凡地满足凸性定义（没有两个点可以违反条件）。
2. **超平面** $\{x \mid a^\mathsf{T}x = b\}$：它是仿射集，因此也是凸集。
3. **半空间** $\{x \mid a^\mathsf{T}x \le b\}$：证——若 $a^\mathsf{T}x_1 \le b, a^\mathsf{T}x_2 \le b$，则 $a^\mathsf{T}(\theta x_1+(1-\theta)x_2) \le \theta b+(1-\theta)b = b$。
4. **Euclid 球** $\{x \mid \|x - x_c\|_2 \le r\}$：由三角不等式可证。
5. **多面体** $\{x \mid Ax \preceq b,\ Cx = d\}$：是半空间和超平面的交集，而凸集的交集仍是凸集。
6. **半正定锥** $\mathbf{S}^n_+$：对所有 $X, Y \succeq 0$ 和 $\theta \in [0,1]$，$\theta X + (1-\theta)Y \succeq 0$。

非凸集的例子（逐一说明为什么不凸）：

1. ** 整数格 $\mathbb{Z}^n$**：取 $(0,0)$ 和 $(1,1)$，中点 $(0.5, 0.5) \notin \mathbb{Z}^n$。
2. **圆环** $\{x \mid r \le \|x\|_2 \le R\}$：内圈和外圈之间的某些弦会穿过内圈的”洞”。
3. **球面** $\{x \mid \|x\|_2 = 1\}$：球面上两点的弦在球内部，不在球面上。

教学建议：这是一个互动的好时机。可以提问”下列集合哪些是凸的？”让学生举手回答，活跃课堂气氛。

7.2.5 1.4 锥与凸锥

定义

集合 C 为锥，如果对任意 $x \in C$ 和 $\theta \ge 0$，有 $\theta x \in C$。

集合 C 为凸锥，如果它既是锥又是凸集。等价地：对任意 $x_1, x_2 \in C$ 和 $\theta_1, \theta_2 \ge 0$，有 $\theta_1 x_1 + \theta_2 x_2 \in C$。

详细讲解

锥的几何含义：从原点出发的”射线族”。如果你在锥中取一个点，那么从原点经过该点的整条射线都在锥中。

凸锥的等价刻画值得强调：锥 + 凸 = 对非负线性组合封闭。这与线性子空间对比：

- 线性子空间：对任意线性组合封闭（系数可正可负）
- 凸锥：对非负线性组合封闭（系数只能非负）

三个最重要的凸锥（会在后面反复出现）：

1. ** 非负象限 $\mathbb{R}^n_+$**：$\{x \mid x_i \geq 0,\ \forall i\}$。在线性规划中，$x \geq 0$ 就是这个锥约束。
2. ** 半正定锥 $\mathbf{S}^n_+$**：全体 $n \times n$ 对称半正定矩阵。在 SDP 中，$X \succeq 0$ 就是这个锥约束。可以告诉学生：半正定锥是" 矩阵版本的非负象限"。
3. **二阶锥（Lorentz 锥/冰淇淋锥）**：$\mathcal{K}_{\text{soc}} = \{(x,t) \mid \|x\|_2 \leq t\}$。在 $\mathbb{R}^3$ 中，它的形状确实像一个冰淇淋蛋筒——顶点在原点，向上开口。在 SOCP 中会用到。

教学提示：画出二阶锥在 $\mathbb{R}^3$ 中的草图（或描述给学生听）。锥的概念是理解后续 SOCP 和 SDP 的关键。

7.2.6 1.5 多面体与单纯形

多面体

定义（再次强调）：多面体是有穷个线性等式和不等式的交集。

$$\mathcal{P} = \{x \mid Ax \preceq b,\ Cx = d\}$$

关键点：

- 多面体是凸集（半空间和超平面的交集）
- 有界多面体称为多胞形（polytope），可以用顶点的凸包表示
- 线性规划的可行域就是多面体

教学提示：可以回顾线性规划中单纯形法的几何本质——在多面体的顶点间移动。这是连接线性规划和凸优化的桥梁。

单纯形

定义：$k+1$ 个仿射独立点的凸包构成 k 维单纯形。

例子：

- 1 维单纯形：线段（2 个点）
- 2 维单纯形：三角形（3 个不共线的点）
- 3 维单纯形：四面体（4 个不共面的点）

单位单纯形：$\Delta_n = \{x \succeq 0 \mid \mathbf{1}^\mathsf{T} x = 1\}$。这是概率向量的集合——每个分量非负且和为 1。在机器学习的分类问题中，softmax 输出就是单位单纯形中的点。

7.2.7 1.6 范数球与椭球

范数球：$\{x \mid \|x - x_c\| \leq r\}$。注意这里的范数不一定是 Euclid 范数——$\ell_1$ 范数下的" 球" 是菱形，ℓ_∞ 范数下的" 球" 是正方形。

椭球：$\mathcal{E}(x_c, P) = \{x \mid (x-x_c)^\mathsf{T} P^{-1}(x-x_c) \leq 1\}$，其中 $P \succ 0$。

- P 决定了椭球的形状和方向：P 的特征向量给出椭球轴的方向，特征值的平方根给出各轴的长度。
- $P = r^2 I$ 时退化为半径为 r 的 Euclid 球。
- 等价表示 $\mathcal{E} = \{x_c + Au \mid \|u\|_2 \le 1\}$（$P = AA^\mathsf{T}$）揭示了椭球是 Euclid 球在仿射变换下的像。

教学建议：椭球的这两种表示都非常有用——前者方便验证点是否在椭球内，后者方便在椭球内生成随机点（先采样 u，再变换）。

7.2.8 1.7 保持凸性的运算

为什么重要？

这是凸分析中最实用的部分。如果你能识别一个复杂集合是由简单凸集通过保凸运算构造出来的，你就不需要从定义出发逐点验证凸性。

八大保凸运算

1. 任意交集：$\bigcap_\alpha C_\alpha$ 为凸。

证明很简单：若 x, y 在每个 C_α 中，则线段也在每个 C_α 中，从而在交集中。这个性质极其有用——多面体就是半空间的交集，因此是凸集。

2. 仿射变换：$f(x) = Ax + b$。凸集的像和原像都是凸集。这是" 线性变换保持凸性" 的形式化。

3. 凸包：包含一个集合的最小凸集。凸包运算把任意集合" 填满" 为凸集。

4. Minkowski 和：$C_1 + C_2 = \{x + y \mid x \in C_1,\ y \in C_2\}$。两个凸集的所有" 向量和" 构成凸集。

5. 笛卡尔积：$C_1 \times C_2$ 为凸。

6. 透视函数：$P(z, t) = z/t$ $(t > 0)$。透视函数模拟了" 针孔相机" 的投影过程。虽然它本身不是仿射的，但它保持凸性。

7. 线性分式函数：$f(x) = (Ax + b)/(c^\mathsf{T} x + d)$。这是透视函数与仿射函数的复合，因此保持凸性。条件概率和条件期望常具有这种形式。

教学建议：这些运算不需要全部详细证明。重点让学生理解**思想**——" 用已知的凸集通过保凸运算构造新凸集"。证明思路可以留给感兴趣的学生课后看 Boyd 的书。

7.2.9 1.8 超平面与支撑超平面

超平面

$\mathcal{H} = \{x \mid a^\mathsf{T} x = b\}$ $(a \neq 0)$。它是 $\mathbb{R}^n$ 中 $n-1$ 维的仿射子空间。

- 在 $\mathbb{R}^2$ 中是直线，在 $\mathbb{R}^3$ 中是平面。
- 它将全空间分为两个闭半空间：$\{x \mid a^\mathsf{T} x \le b\}$ 和 $\{x \mid a^\mathsf{T} x \ge b\}$。

支撑超平面

定义：C 在边界点 x_0 处的支撑超平面 $\{x \mid a^\mathsf{T}x = a^\mathsf{T}x_0\}$ 满足 $a^\mathsf{T}x \le a^\mathsf{T}x_0$ $(\forall x \in C)$。

几何理解：

- 想象一个凸的土豆。在土豆表面任意一点，你都可以放一把刀，刀面恰好与土豆相切，且土豆完全在刀的一侧。
- 这就是支撑超平面：a 是刀面的法向量，指向土豆的外部。

为什么重要？支撑超平面定理保证了凸集边界上的每一点都存在支撑超平面。这个性质是分离定理和对偶理论的基础。

7.2.10 1.9 分离超平面定理

定理陈述

若 C 和 D 是两个不相交的非空凸集，则存在分离超平面 $\{x \mid a^\mathsf{T}x = b\}$：

$$a^\mathsf{T}x \le b \ (\forall x \in C), \qquad a^\mathsf{T}x \ge b \ (\forall x \in D)$$

详细解释

几何意义：两个不重叠的凸集之间总是可以插入一张” 平面”，将两者完全分开。

严格分离：当 C 是闭凸集、D 是紧凸集时，分离可以是严格的（不等号严格）。特别地：

- 一个闭凸集和一个不在其中的点可以被严格分离。
- 这是” 点和闭凸集可以分离” 的几何事实。

证明草图（在课堂上可以简要提及）

严格分离定理的证明基于：点 x_0 到闭凸集 C 存在唯一的最近点（投影），由投影构造分离超平面。

重要性

分离超平面定理是凸优化的**几何基石**。后续的对偶理论、KKT 条件的几何解释，都根植于此。对偶变量本质上就是分离超平面的法向量！

教学建议：强调这个定理的” 存在性” 而非构造性——它告诉你分离超平面存在，但不告诉你如何找到它。在凸优化中，我们通过求解对偶问题来找到这个分离超平面。

7.2.11 1.10 对偶锥与广义不等式

对偶锥

$$K^* = \{y \mid y^\mathsf{T}x \ge 0, \ \forall x \in K\}$$

几何解释：y 属于 K^* 当且仅当 y 与 K 中每个向量的夹角不超过 90 度。K^* 是" 与 K 成锐角" 的所有向量的集合。

关键性质

1. $^{**}K^*$ 总是闭凸锥 **: 即使 K 不是闭的或不是凸的。因为 K^* 是一族闭半空间（$\{y \mid y^{\mathsf{T}}x \geq 0\}$）的交集。
2. **自对偶锥**：非负象限、半正定锥和二阶锥都是自对偶的。这意味着它们" 恰好等于与自身成锐角的向量集合"。自对偶性是这些锥在优化中如此重要的深层原因。
3. $^{**}(K^*)^* = K^{**}$: 当 K 是闭凸锥时，对偶的对偶回到自身。

广义不等式

由正常锥 K 可以定义广义不等式：

$$x \preceq_K y \iff y - x \in K, \qquad x \prec_K y \iff y - x \in \mathbf{int}\, K$$

特例：

- $K = \mathbb{R}^n_+$: 分量不等式 $x_i \leq y_i$
- $K = \mathbf{S}^n_+$: 矩阵不等式 $Y - X \succeq 0$

这统一了我们见过的所有" 不等式"——它们都只是某个锥的广义不等式。

教学提示：对偶锥可能比较抽象。可以多举几个例子：在 $\mathbb{R}^3$ 中，非负象限的对偶锥还是非负象限；在 $\mathbb{R}^2$ 中，第一象限（90 度角区域）的对偶锥就是它自己。

7.3 第二节凸函数

7.3.1 教学目标

学完本节后，学生应能：

1. 用定义、一阶条件、二阶条件判断函数的凸性
2. 利用保凸运算构造新的凸函数
3. 计算简单函数的共轭函数
4. 区分凸函数、拟凸函数、对数凹函数

7.3.2 2.1 凸函数的定义

定义

$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$，$\mathbf{dom}\, f$ 为凸集。f 为凸函数如果：

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \quad \forall x, y \in \mathbf{dom}\, f,\ \theta \in [0,1]$$

详细讲解

扩展实值函数：这里我们把函数定义在 $\mathbb{R}^n$ 上，但在 $\mathbf{dom}\, f$ 之外取 $+\infty$。这个技巧看起来奇怪，但实际上非常方便——它允许我们把" 约束 $x \in C$" 等价地写为目标函数加上一个指示函数：

$$\mathbb{I}_C(x) = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}$$

这样，带约束的优化问题 $\min_{x \in C} f(x)$ 就等价于无约束问题 $\min_x [f(x) + \mathbb{I}_C(x)]$。

凸函数的几何意义：函数图形上任意两点间的弦总在函数图形的上方。这个" 弦在上" 的直观是判断凸性最直接的方法。

上图刻画：f 凸 $\iff \mathbf{epi}\, f = \{(x,t) \mid f(x) \le t\}$ 为凸集。这是连接" 凸函数" 和" 凸集" 的桥梁——函数的凸性等价于其上图的集合凸性。

严格凸函数：不等式严格成立。直观上，严格凸函数的图形" 严格向下弯曲"——没有直线段。严格凸函数的最优解（如果存在）是唯一的。

教学建议：花足够时间讲上图刻画。这是从" 函数凸性" 到" 集合凸性" 的关键一步，后续很多证明都会用到。

7.3.3 2.2 一阶条件与二阶条件

一阶条件（定理 7.1）

设 f 可微。f 为凸函数 **当且仅当**：

$$f(y) \ge f(x) + \nabla f(x)^{\mathsf{T}}(y - x), \quad \forall x, y \in \mathbf{dom}\, f$$

通俗理解：一阶 Taylor 近似 $f(x) + \nabla f(x)^{\mathsf{T}}(y-x)$ 是 $f(y)$ 的一个**全局下界**。无论 y 离 x 多远，这个线性下界始终有效。这与一般函数的一阶近似只在局部有效形成鲜明对比。

几何理解：在函数图形上的任意一点 $(x, f(x))$ 处做切线（或切超平面），整条函数曲线都在切线的上方。对于凹函数，切线在曲线下方。

二阶条件（定理 7.2）

设 f 二阶可微。f 为凸函数 **当且仅当** $\nabla^2 f(x) \succeq 0$ 对所有 x 成立。

重要反例：$f(x) = x^4$ 是严格凸的，但在 $x = 0$ 处 Hessian 为 $f''(0) = 0$（不是正定的）。所以 Hessian 的正定性是严格凸性的充分条件而非必要条件。

教学建议：二阶条件是学生最常用的判断工具（直接算 Hessian 然后检查特征值）。但必须提醒他们反例的存在。

7.3.4 2.3 一阶条件的证明

充分性证明（一阶条件 $\Longrightarrow$ 凸性）

这个证明是典型的" 凸分析技巧"，值得在课堂上完整走一遍。

设定：假设一阶条件对任意 x, y 都成立。取任意 x, y 和 $\theta \in [0,1]$。

关键一步：令 $z = \theta x + (1-\theta)y$。$z$ 在 x 和 y 的连线上，因此也在 $\mathbf{dom}\, f$ 中（因为定义域是凸集）。

应用一阶条件：

- 在点 z 处，使用 x 作为" 另一个点"：$f(x) \geq f(z) + \nabla f(z)^{\mathsf{T}}(x - z)$
- 在点 z 处，使用 y 作为" 另一个点"：$f(y) \geq f(z) + \nabla f(z)^{\mathsf{T}}(y - z)$

加权求和：

$$\begin{aligned}\theta f(x) + (1-\theta) f(y) &\geq \theta[f(z) + \nabla f(z)^{\mathsf{T}}(x - z)] + (1-\theta)[f(z) + \nabla f(z)^{\mathsf{T}}(y - z)] \\ &= f(z) + \nabla f(z)^{\mathsf{T}}[\theta x + (1-\theta)y - z] \\ &= f(z) = f(\theta x + (1-\theta)y)\end{aligned}$$

最后一步是因为 z 的定义恰好使括号内为零。证毕。

必要性证明（凸性 + 可微 $\Longrightarrow$ 一阶条件）

构造方向导数：固定 x, y，考虑 $t \in (0,1]$，由凸性定义：

$$f(x + t(y-x)) = f((1-t)x + ty) \leq (1-t) f(x) + t f(y)$$

重排得：

$$\frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x)$$

令 $t \to 0^+$，左边趋于方向导数 $\nabla f(x)^{\mathsf{T}}(y - x)$，得：

$$\nabla f(x)^{\mathsf{T}}(y - x) \leq f(y) - f(x)$$

即一阶条件。证毕。

教学建议：这个证明的优美之处在于，充分性和必要性都用非常初等的方法完成——充分性用加权求和，必要性用方向导数。可以鼓励学生在课后自己尝试把这个证明复现一遍。

7.3.5 2.4 常见凸函数

单变量凸函数

对每个函数，简要说明为什么凸：

- e^{ax}：$f''(x) = a^2 e^{ax} \geq 0$（对所有 a 都凸！a 正负无所谓，因为平方后为正）

- x^a $(x>0)$：$f''(x)=a(a-1)x^{a-2}$。当 $a\geq 1$ 或 $a\leq 0$ 时 $f''\geq 0$；当 $0\leq a\leq 1$ 时 $f''\leq 0$（凹函数）
- $|x|^p$ $(p\geq 1)$：这是 $\mathbb{R}$ 上的范数，由三角不等式可直接推出凸性
- $x\log x$ $(x>0)$：$f''(x)=1/x>0$，严格凸。这是在信息论和机器学习中最重要的凸函数之一
- $-\log x$ $(x>0)$：$f''(x)=1/x^2>0$，严格凸。对数障碍函数的基础
- $\log(1+e^x)$：$f''(x)=e^x/(1+e^x)^2>0$。logistic 回归的损失函数

多变量凸函数

- **任意范数**：由三角不等式 $\|x+y\|\leq\|x\|+\|y\|$ 和绝对齐次性 $\|\alpha x\|=|\alpha|\|x\|$ 可推出凸性
- **二次型**：Hessian 为 P，凸 $\iff$ $P\succeq 0$
- **log-sum-exp**：$\log\sum e^{x_i}$。这是" 光滑的最大值"——当某个 x_i 远大于其他时，log-sum-exp 趋近于 $\max_i x_i$。在机器学习中被称为 softmax 的对数
- **$-\log\det X$**：矩阵负对数行列式。在参数估计和实验设计中非常重要
- **$\lambda_{\max}(X)$**：最大特征值。凸但非光滑——在某些点处不可微

教学提示：log-sum-exp 值得多花点时间讲。它是连接凸优化和机器学习（尤其是 softmax 分类器）的桥梁。

7.3.6 2.5 保持凸性的函数运算

六大运算

1. 非负加权和：$f=\sum w_i f_i$ $(w_i\geq 0)$。凸函数的非负线性组合仍是凸函数。这是最基本也最常用的运算。凸集 { } 本身构成一个凸锥。

2. 逐点上确界：$g(x)=\sup_y f(x,y)$。特别地，一族凸函数的逐点最大值是凸函数。

证明思路：g 的上图 $=\bigcap_y \mathbf{epi}\, f(\cdot,y)$，而凸集的任意交集是凸集。

这个性质解释了为什么很多重要的凸函数可以写成" 一族更简单凸函数的上确界"——例如共轭函数就是这种形式。

3. 与仿射函数的复合：$g(x)=f(Ax+b)$。如果 f 凸，则 g 凸。证：g 的上图是 f 的上图在仿射映射下的原像。

4. 标量复合规则：当 g 凸且不减，h 凸，则 $g\circ h$ 凸。需要额外的不减/不增条件，这是因为链式法则中的正负号需要匹配。

5. 下确界（部分极小化）：$g(x)=\inf_y f(x,y)$。如果 $f(x,y)$ 关于 (x,y) 联合凸，则 $g(x)$ 关于 x 凸。这个性质在" 消去变量" 和" 对偶理论" 中反复使用。

6. 透视函数：$g(x,t)=t\cdot f(x/t)$ $(t>0)$。很多重要函数都是通过透视运算从更简单的凸函数构造出来的——例如 KL 散度从负熵 $x\log x$ 得来。

KL 散度的凸性（例 7.3）

这是透视运算的一个经典例子：

负熵 $f(x) = x\log x$（$x > 0$）是严格凸的。取透视：

$$g(x,t) = t \cdot \frac{x}{t}\log\frac{x}{t} = x\log\frac{x}{t}, \quad x > 0,\ t > 0$$

g 关于 (x,t) 联合凸。而 KL 散度 $D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log(p_i/q_i)$ 正是这种形式的和，因此是凸函数。

教学建议：这些保凸运算中，" 非负加权和" 和" 与仿射函数的复合" 最常用，建议重点讲。" 标量复合规则" 条件和结论比较绕，可以简单提一下让学生知道有这个规则即可。

7.3.7 2.6 共轭函数

定义

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^{\mathsf{T}}x - f(x))$$

深度讲解

命名来源：Fenchel 共轭，也称为 Legendre 变换（在物理和力学中）。之所以叫" 共轭"，是因为它与对偶空间的自然对 $(x,y) \mapsto y^{\mathsf{T}}x$ 密切相关。

几何解释：$f^*(y)$ 是线性函数 $y^{\mathsf{T}}x$ 与 $f(x)$ 之间的**最大差距**。等价地说，$f^*(y)$ 是使得仿射函数 $y^{\mathsf{T}}x - \alpha$ 成为 f 的下界的 ** 最小 α**。

** 为什么 f^* 总是凸？** 因为对每个固定的 x，$y^{\mathsf{T}}x - f(x)$ 是关于 y 的仿射函数，而仿射函数族的上确界是凸函数。所以 f^* 总是凸的——即使 f 本身不是凸的！

常见共轭对的推导：

1. $f(x) = \frac{1}{2}x^{\mathsf{T}}Qx$（$Q \succ 0$）：对 $y^{\mathsf{T}}x - \frac{1}{2}x^{\mathsf{T}}Qx$ 关于 x 求导，得 $y - Qx = 0 \implies x = Q^{-1}y$。代入得 $f^*(y) = \frac{1}{2}y^{\mathsf{T}}Q^{-1}y$。注意共轭操作将 Hessian 从 Q 变为 Q^{-1}。
2. $f(x) = -\log x\ (x > 0)$：$f^*(y) = \sup_{x>0}(yx + \log x)$。对 x 求导：$y + 1/x = 0 \implies x = -1/y$（需 $y < 0$）。代入：$f^*(y) = y(-1/y) + \log(-1/y) = -1 - \log(-y) = -\log(-y) - 1$。
3. $f(x) = e^x$：$f^*(y) = \sup_x(yx - e^x)$。导数为 $y - e^x = 0 \implies x = \log y$（需 $y > 0$）。代入：$f^*(y) = y\log y - y$。这就是负熵的共轭。

Young 不等式：由共轭的定义直接得 $x^{\mathsf{T}}y \le f(x) + f^*(y)$。这是 Hölder 不等式的推广。当 $f(x) = \frac{1}{p}|x|^p$ 时，$f^*(y) = \frac{1}{q}|y|^q$（$1/p + 1/q = 1$），此时 Young 不等式就是 $xy \le \frac{x^p}{p} + \frac{y^q}{q}$。

教学建议：共轭函数是凸分析中比较高级的概念。重点讲清楚定义、几何解释和两三个最常见的共轭对。Young 不等式是连接共轭理论和对偶理论的纽带，务必讲清楚。

7.3.8 2.7 拟凸函数

定义

$$f(\theta x + (1-\theta)y) \le \max\{f(x), f(y)\}$$

与凸函数的区别

凸函数要求”函数值在弦的下方”，而拟凸函数只要求”函数值不超过两端点的最大值”。这是一个弱得多的条件。

等价刻画：f 拟凸当且仅当所有下水平集 $\{x \mid f(x) \le \alpha\}$ 为凸集。

例子：$\sqrt{|x|}$ 是拟凸的但不是凸的。画出它的图形——在 $x=0$ 处有一个尖点，图形位于连接任意两点的弦的上方，但下水平集 $[-\alpha^2, \alpha^2]$ 始终是区间（凸集）。

一阶条件：如果 $f(y) \le f(x)$，那么 $\nabla f(x)^{\mathsf{T}}(y-x) \le 0$。几何上：如果 y 的函数值不高于 x 的，那么从 x 出发向 y 移动时，函数值的一阶近似不增加。

实际意义：拟凸优化问题可能存在不是全局最优的局部最优解。这是与凸优化的关键区别。

7.3.9 2.8 对数凹函数

定义

$f(x) > 0$，$\log f$ 为凹函数。

重要性

对数凹性在统计学和概率论中无处不在。大多数常见概率分布的密度函数都是对数凹的：

- 正态分布：$\log \phi(x) = -\frac{x^2}{2} + \text{const}$，这是凹函数（二次项系数为负）
- 指数分布：$\log(\lambda e^{-\lambda x}) = \log\lambda - \lambda x$，这是凹函数（线性）
- 均匀分布：在支撑集上取常数，也是对数凹的

Prékopa 定理：对数凹密度的边缘分布仍为对数凹。这是概率论中一个深刻的定理，在凸几何中有广泛应用。

教学建议：对数凹函数可以相对简略地讲。重点是让学生知道”大多数常见概率密度都是对数凹的”。

7.4 第三节凸优化问题

7.4.1 教学目标

学完本节后，学生应能：

1. 识别一个优化问题是否为凸优化问题
2. 区分 LP、QP、QCQP、SOCP、SDP 和 GP
3. 将简单的实际问题建模为上述问题类型

7.4.2 3.1 凸优化问题的标准形式

定义

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0,\ i = 1, \ldots, m \\ & a_i^{\mathsf{T}} x = b_i,\ i = 1, \ldots, p \end{aligned}$$

其中 $f_0, \ldots, f_m$ 为凸函数，等式约束为仿射。

核心性质的详解

1. 可行集是凸集：$\{x \mid f_i(x) \leq 0\}$ 是凸集（凸函数的下水平集），$\{x \mid Ax = b\}$ 是仿射集。凸集和仿射集的交集是凸集。

2. 局部最优即全局最优（需要严格证明给学生看）：

假设 x^* 是局部最优解：存在邻域 N 使得 $f_0(x^*) \leq f_0(x)$ 对所有 $x \in N \cap \mathcal{F}$ 成立。

如果存在全局更优点 $y \in \mathcal{F}$ 使得 $f_0(y) < f_0(x^*)$，则考虑线段上的点 $z = \theta y + (1-\theta)x^*$。

由凸性：$f_0(z) \leq \theta f_0(y) + (1-\theta) f_0(x^*) < \theta f_0(x^*) + (1-\theta) f_0(x^*) = f_0(x^*)$。

当 θ 足够小时，z 属于 x^* 的邻域 N。这与 x^* 是局部最优矛盾。因此不存在 y 使得 $f_0(y) < f_0(x^*)$，x^* 是全局最优。

3. 最优解集为凸集：若 x_1^*, x_2^* 都是最优解，则 $f_0(x_1^*) = f_0(x_2^*) = p^*$。对 $\theta \in [0,1]$，$f_0(\theta x_1^* + (1-\theta)x_2^*) \leq \theta p^* + (1-\theta)p^* = p^*$。但又因为 p^* 是最优值，所以必须等于 p^*。因此所有凸组合也是最优解。

7.4.3 3.2 凸优化 vs 非凸优化

对比表讲解

这张表是本章最核心的信息之一。逐行讲解：

1. **局部最优**：凸优化中" 爬山" 不可能卡在局部最优（因为没有局部最优），而非凸优化中局部最优可能到处都是。
2. **最优解集合**：凸优化中如果最优解不唯一，所有最优解构成一个凸集（意味着有无限多个最优解，且它们可以连起来）。非凸优化中，多个最优解之间没有这种结构。
3. **对偶间隙**：这是区分凸和非凸最关键的标志之一。凸优化在温和条件下强对偶成立（间隙为零），而非凸优化的间隙可能任意大。
4. **KKT 条件**：凸优化中满足 KKT 条件等价于最优（充要），非凸优化中只是必要。
5. **计算复杂度**：凸优化是" 容易" 的（多项式时间可解），非凸优化一般是 NP-hard。

Boyd 的名言：*"If you formulate a practical problem as a convex optimization problem, you have solved it."* 这句话可以写在黑板上，让学生体会凸优化在工程实践中的核心地位。

7.4.4 3.3 线性规划（LP）

标准形式

$$\begin{aligned} \text{minimize} \quad & c^{\mathsf{T}}x + d \\ \text{subject to} \quad & Gx \preceq h, \quad Ax = b \end{aligned}$$

讲解要点

- LP 的目标和约束都是仿射的——仿射函数既凸又凹，因此 LP 是凸优化问题。
- 单纯形法的几何本质：沿多面体的棱从一个顶点走到另一个顶点，每次移动到使目标函数下降的邻接顶点。
- 内点法：在多面体内部走”捷径”，通过中心路径直接逼近最优解。

ℓ_∞ 回归（例 7.4）

$$\min_x \|Ax - b\|_\infty = \min_x \max_i |a_i^{\mathsf{T}}x - b_i|$$

等价 LP：

$$\begin{aligned} \text{minimize} \quad & t \\ \text{subject to} \quad & -t \le a_i^{\mathsf{T}}x - b_i \le t, \quad \forall i \end{aligned}$$

推导思路：$\max_i |a_i^{\mathsf{T}}x - b_i| \le t$ 等价于 $|a_i^{\mathsf{T}}x - b_i| \le t$ 对所有 i 成立，这又等价于 $-t \le a_i^{\mathsf{T}}x - b_i \le t$。这是将 ℓ_∞ 范数最小化写成 LP 的经典技巧。

ℓ_1 回归（例 7.5）

同理，$\|Ax - b\|_1$ 的最小化等价于：

$$\begin{aligned} \text{minimize} \quad & \sum_i t_i \\ \text{subject to} \quad & -t_i \le a_i^{\mathsf{T}}x - b_i \le t_i \end{aligned}$$

教学建议：这两个例子很重要。它们展示了如何将”看似非线性的范数最小化问题”等价转化为线性规划。这种建模技巧在实际问题中非常有用。

7.4.5 3.4 二次规划（QP 与 QCQP）

QP

目标为二次函数，约束为仿射。关键是 $P \succeq 0$ 确保目标凸。

QP 是最常见的凸优化问题类型。原因：很多问题的自然模型就是”最小化一个二次误差 + 线性约束”。

例子：支持向量机（SVM）的对偶问题、投资组合优化的 Markowitz 模型、线性最小二乘（$\|Ax-b\|_2^2$）、模型预测控制（MPC）。

QCQP

约束也是二次的，同样要求 $P_i \succeq 0$ 确保凸性。

注意：如果 P_i 中有非半正定的，那问题就不是凸 QCQP，而是更难的非凸二次规划。

7.4.6 3.5 二阶锥规划（SOCP）

标准形式

$$\begin{aligned} \text{minimize} \quad & f^\mathsf{T} x \\ \text{subject to} \quad & \|A_i x + b_i\|_2 \le c_i^\mathsf{T} x + d_i, \quad i = 1, \dots, m \\ & Fx = g \end{aligned}$$

讲解要点

SOCP 是 QP 和 QCQP 的推广。每个 QCQP 都可以写成 SOCP（但反过来不行）。

二阶锥约束 $\|Ax+b\|_2 \le c^\mathsf{T}x + d$ 要求点 $(Ax+b,\ c^\mathsf{T}x+d)$ 落在二阶锥 $\mathcal{K}_{\text{soc}}$ 中。

概率约束的 SOCP 形式：这是一个非常实用的例子。如果 $a \sim \mathcal{N}(\bar{a}, \Sigma)$，那么线性约束的概率界：

$$\text{Prob}(a^\mathsf{T} x \le b) \ge \eta$$

等价于 $\bar{a}^\mathsf{T}x + \Phi^{-1}(\eta)\|\Sigma^{1/2}x\|_2 \le b$。

推导：$a^\mathsf{T}x \sim \mathcal{N}(\bar{a}^\mathsf{T}x,\ x^\mathsf{T}\Sigma x)$。标准化后：

$$\text{Prob}\left(\frac{a^\mathsf{T}x - \bar{a}^\mathsf{T}x}{\|\Sigma^{1/2}x\|_2} \le \frac{b - \bar{a}^\mathsf{T}x}{\|\Sigma^{1/2}x\|_2}\right) = \Phi\left(\frac{b-\bar{a}^\mathsf{T}x}{\|\Sigma^{1/2}x\|_2}\right) \ge \eta$$

即 $\frac{b-\bar{a}^\mathsf{T}x}{\|\Sigma^{1/2}x\|_2} \ge \Phi^{-1}(\eta)$，重排便得二阶锥约束。

问题层次：LP $\subset$ QP $\subset$ QCQP $\subset$ SOCP $\subset$ SDP。层次越高，表达能力越强，但求解成本也越高。

7.4.7 3.6 半正定规划（SDP）

标准形式

$$\begin{aligned} \text{minimize} \quad & \text{tr}(C^\mathsf{T} X) \\ \text{subject to} \quad & \text{tr}(A_i^\mathsf{T} X) = b_i, \quad i = 1, \dots, m \\ & X \succeq 0 \end{aligned}$$

其中变量 X 是 $n \times n$ 对称矩阵。

为什么 SDP 重要？

1. **统一框架**：LP $\subset$ SOCP $\subset$ SDP。SDP 是表达力最强的" 易解" 优化问题。
2. **LMI（线性矩阵不等式）**：形如 $A(x) \succeq 0$ 的约束，在控制论中无处不在。Lyapunov 稳定性条件、H_∞ 控制器设计都可以写成 LMI。
3. **半正定松弛**：很多 NP-hard 的非凸二次优化问题，可以通过" 松弛 xx^T 为 $X \succeq xx^\mathsf{T}$" 转化为 SDP，得到最优值的下界（有时这个下界恰好是紧的!）。

Schur 补引理

$$\begin{bmatrix} X & F^\mathsf{T} \\ F & G \end{bmatrix} \succeq 0 \iff X \succeq 0,\ G - FX^{-1}F^\mathsf{T} \succeq 0$$

（假设 $X \succ 0$）

这个引理是 SDP 建模中最常用的工具。它允许我们将" 非线性的矩阵不等式" 转化为" 线性的增广矩阵半正定性"。几乎所有的 SDP 建模问题都会用到它。

矩阵谱范数最小化（例 7.7）

$\|A(x)\|_2 \le t$ 等价于 $A(x)^\mathsf{T}A(x) \preceq t^2 I$（最大奇异值不超过 t），再由 Schur 补等价于：

$$\begin{bmatrix} tI & A(x) \\ A(x)^\mathsf{T} & tI \end{bmatrix} \succeq 0$$

这是 LMI 的标准形式。

7.4.8 3.7 几何规划（GP）

定义

单项式：$cx_1^{a_1}\cdots x_n^{a_n}$（$c > 0$）。正项式：单项式的和（正系数）。

凸化变换

GP 本身不是凸的，但通过变量代换 $y_i = \log x_i$ 可以转化为凸优化问题。

正项式约束 $f(x) \le 1$ 变为：

$$\log \sum_k e^{a_k^\mathsf{T} y + b_k} \le 0$$

这是 log-sum-exp 函数，是凸的。

应用：GP 在电路晶体管尺寸设计中广泛使用——晶体管的电流、跨导等都可以用单项式/正项式近似。

教学建议：GP 可以快速带过，重点说清楚" 非凸问题通过变量变换可以转化为凸的" 这一思想。

7.5 第四节对偶理论

7.5.1 教学目标

学完本节后，学生应能：

1. 写出给定优化问题的 Lagrange 对偶函数和对偶问题
2. 理解弱对偶和强对偶的含义及区别
3. 使用 Slater 条件验证凸优化问题的强对偶性
4. 写出并解释 KKT 条件的四条内容

7.5.2 4.1 Lagrange 对偶函数

Lagrange 函数

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

关键：Lagrange 函数将约束以**加权**的方式加到目标函数上。

- 对不等式约束 $f_i(x) \leq 0$：加上 λ_i 倍的 $f_i(x)$，其中 $\lambda_i \geq 0$——如果 $f_i(x) > 0$（违反约束），则惩罚为正
- 对等式约束 $h_i(x) = 0$：加上 ν_i 倍的 $h_i(x)$，其中 $\nu_i \in \mathbb{R}$ 可正可负

重要：这里对原始问题不做凸性假设。Lagrange 函数的定义和基本性质对任意优化问题都成立。

Lagrange 对偶函数

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

** 为什么取 inf? ** 因为我们希望对于给定的乘子 (λ, ν)，找到 L 的最小可能值。这个下确界对于原始最优值 p^* 的估计至关重要。

计算对偶函数的技巧：通常通过令 $\nabla_x L(x, \lambda, \nu) = 0$ 解出 x 关于 (λ, ν) 的表达式，然后代回 L。

7.5.3 4.2 对偶函数的基本性质

下界性质（定理 7.3）

对任意 $\lambda \succeq 0$ 和 ν：

$$g(\lambda, \nu) \leq p^*$$

证明：设 x^* 为原始最优解。则：

$$\begin{aligned} g(\lambda,\nu) &= \inf_x \big(f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_i h_i(x)\big) \\ &\le f_0(x^*) + \sum \lambda_i f_i(x^*) + \sum \nu_i h_i(x^*) \\ &\le f_0(x^*) = p^* \end{aligned}$$

第一个不等号因为下确界不超过任意一点的值；第二个不等号因为 $\lambda_i \ge 0$，$f_i(x^*) \le 0$（所以 $\sum \lambda_i f_i(x^*) \le 0$），且 $h_i(x^*) = 0$。

凹性（定理 7.4）

$g(\lambda,\nu)$ 关于 (λ,ν) 是联合凹的。

证明的关键：对每个固定 x，$L(x,\lambda,\nu)$ 关于 (λ,ν) 是仿射函数（线性的）。g 是一族仿射函数的逐点下确界。而**仿射函数的逐点下确界一定是凹的**（类比：凸函数的逐点上确界是凸的）。

这对偶问题的意义：最大化凹函数 + 非负约束 $\lambda \succeq 0$ = **凸优化问题**。因此对偶问题总是凸的，无论原始问题是什么！

7.5.4 4.3 Lagrange 对偶问题

对偶问题

$$\begin{aligned} \text{maximize} \quad & g(\lambda,\nu) \\ \text{subject to} \quad & \lambda \succeq 0 \end{aligned}$$

原始与对偶的对称结构

	原始问题	对偶问题
变量	$x \in \mathbb{R}^n$	$(\lambda,\nu) \in \mathbb{R}^m \times \mathbb{R}^p$
目标	$\min f_0(x)$	$\max g(\lambda,\nu)$
约束	$f_i(x) \le 0$，$h_i(x) = 0$	$\lambda \succeq 0$
凸性	可能非凸	总是凸

对偶变量的维数：m 个不等式约束对应 m 个对偶变量 λ_i，p 个等式约束对应 p 个对偶变量 ν_i。总维数 $m+p$，通常远小于 n。

弱对偶：$d^* \le p^*$。这是对的，无论问题是否凸，强对偶是否成立。

7.5.5 4.4 弱对偶与强对偶

弱对偶定理

$$g(\lambda, \nu) \leq f_0(x)$$

对任意原始可行解 x 和对偶可行解 (λ, ν) 成立。

这给出了一个" 对偶间隙": $p^* - d^* \geq 0$。

强对偶

$p^* = d^*$，即对偶间隙为零。

什么时候强对偶成立？

- 对**一般非凸问题**：通常**不成立**。对偶间隙可以非常大。
- 对**凸优化问题**：如果满足约束规格（如 Slater 条件），则**成立**。

为什么强对偶重要？因为：

1. 对偶问题总是凸的，即使原始问题不是——我们可以通过求解对偶问题来获取原始最优值
2. 对偶变量 (λ^*, ν^*) 可以直接用于验证原始最优性（KKT 条件）

7.5.6 4.5 Slater 条件

条件陈述

如果存在严格可行点 x（$f_i(x) < 0$ 对所有 i，$Ax = b$），且 x 在定义域的**相对内部**中，则强对偶成立。

详细解释

" 严格可行" 的含义：不等式约束必须严格成立（$<$ 而非 $\leq$）。等式约束只需可行（$=$）。

" 相对内部" 的含义：考虑定义域 $\mathcal{D} = \bigcap_i \mathbf{dom}\, f_i$。$x$ 必须在 $\mathcal{D}$ 的相对内部（相对于 $\mathcal{D}$ 的仿射包）。对于大多数实际问题，$\mathcal{D}$ 是开的或者有内部点，这时 $\mathbf{relint}\, \mathcal{D} = \mathbf{int}\, \mathcal{D}$。

特殊情形——仿射约束：如果所有不等式约束 $f_i(x) \leq 0$ 都是仿射的（如 LP），那么 Slater 条件退化为：存在可行点即可。因为对于仿射约束，严格不等式可以放松为不等式。

实践意义：对大多数实际凸优化问题，Slater 条件自然满足。这就是为什么 Boyd 说"almost all convex problems have zero duality gap"。

7.5.7 4.6 KKT 条件

四条 KKT 条件

(1) 原始可行性：$f_i(x^*) \leq 0$，$h_i(x^*) = 0$。这一条只是说 x^* 必须在可行域内。

(2) 对偶可行性：$\lambda_i^* \geq 0$。对偶变量的非负性——因为我们惩罚的是 $f_i(x) \leq 0$ 的违反，乘子必须是惩罚而非" 奖励"。

(3) 互补松弛性：$\lambda_i^* f_i(x^*) = 0$。这是最重要的一条。它说：要么约束是活跃的（$f_i(x^*) = 0$），要么乘子为零（$\lambda_i^* = 0$），或者两者都为零。**不活跃的约束不参与最优性条件**。

(4) 稳定性：$\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(x^*) + \sum \nu_i^* \nabla h_i(x^*) = 0$。这是 Lagrange 函数关于 x 的梯度为零——在一阶条件下，x^* 使 L 达到最小。

KKT 条件的证法概述（凸情形）

充分性（KKT $\implies$ 最优）：由于 f_0 凸，$\lambda_i^* \geq 0$ 使得 $\sum \lambda_i^* f_i$ 凸，整个 $L(x, \lambda^*, \nu^*)$ 关于 x 是凸的。由稳定性条件知 x^* 是 L 的全局最小点。因此：

$$f_0(x^*) = L(x^*, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*) \leq f_0(x)$$

（最后一个不等号是因为对可行 x，$\lambda_i^* f_i(x) \leq 0$，$h_i(x) = 0$）。

必要性（最优 + 强对偶 $\implies$ KKT）：需要用到对偶最优解的性质，可参考 Boyd 书 pp. 243-244。

互补松弛性的经济学解释

λ_i^* 是第 i 个约束的**影子价格**。拉格朗日乘子法告诉我们：将约束放宽 ε（即 $f_i(x) \leq \varepsilon$ 代替 $f_i(x) \leq 0$），最优值的变化近似为 $-\lambda_i^* \varepsilon$。

如果约束在最优解处不活跃（$f_i(x^*) < 0$），那么稍微再放宽它不会改变最优值，因此其影子价格为零（$\lambda_i^* = 0$）。这就是互补松弛性的直觉！

7.5.8 4.7 KKT 条件应用——等式约束 QP

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^\mathsf{T} P x + q^\mathsf{T} x + r \\ \text{subject to} \quad & Ax = b \end{aligned}$$

其中 $P \succeq 0$。

由 KKT 条件推导：（没有不等式约束，所以只有原始可行性、稳定性）

稳定性：$\nabla_x L = Px^* + q + A^\mathsf{T} \nu^* = 0$。原始可行性：$Ax^* = b$。

组合成**鞍点系统**（saddle-point system）：

$$\begin{bmatrix} P & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

求解方法：

- 直接求解：如果 $P \succ 0$ 且 A 满秩，可以消去 x 得到关于 ν 的较小系统
- 如果 $P \succeq 0$ 但非正定，只能在 $\ker A$ 上正定，则系统仍是适定的
- 该矩阵非奇异当且仅当 $P \succ 0$ 在 $\ker A$ 上

7.6 第五节无约束优化算法

7.6.1 教学目标

学完本节后，学生应能：

1. 实现梯度下降并选择步长
2. 理解梯度下降的 $O(1/k)$ 收敛性证明
3. 说明 Newton 方法的二次收敛特性
4. 根据问题规模选择合适的优化算法

7.6.2 5.1 梯度下降法

迭代格式

$$x^{(k+1)} = x^{(k)} - t_k \nabla f(x^{(k)})$$

步长选择

精确线搜索：$t_k = \operatorname{argmin}_{t \geq 0} f(x^{(k)} - t\nabla f(x^{(k)}))$。

- 理论上最优但在实践中很少使用（一维搜索本身就有成本）
- 用于理论分析

回溯线搜索（Backtracking line search，实践中推荐）：

- 参数 $\alpha \in (0, 0.5)$（通常取 0.1 到 0.3），$\beta \in (0, 1)$（通常取 0.5 到 0.8）
- Armijo 条件：$f(x - t\nabla f(x)) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$
- 这个条件保证了" 充分下降"——函数值的实际下降量不低于一阶预测下降量的 α 倍
- 初始 t 可以取 1（对 Newton 方法）或取上一迭代的步长

固定步长：如果知道 Lipschitz 常数 L，可取 $t = 1/L$。

算法 5.1：带回溯线搜索的梯度下降

给定初始点 $x^{(0)}$，容忍度 $\varepsilon > 0$，参数 $\alpha \in (0, 0.5)$，$\beta \in (0, 1)$，**repeat** $k = 0, 1, \ldots$：

1. **if** $\|\nabla f(x^{(k)})\|_2 \leq \varepsilon$：**break**（满足停止准则）
2. 回溯线搜索：令 $t := 1$；while $f(x^{(k)} - t\nabla f(x^{(k)})) > f(x^{(k)}) - \alpha t \|\nabla f(x^{(k)})\|_2^2$：令 $t := \beta t$
3. 更新：$x^{(k+1)} := x^{(k)} - t\nabla f(x^{(k)})$

7.6.3 5.2 梯度下降的收敛性分析

假设条件

- f 凸且 L-光滑：$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$
- L-光滑等价于 $\nabla^2 f(x) \preceq LI$（当 f 二阶可微时）

收敛率

一般凸情形：$f(x^{(k)}) - p^* \le \dfrac{L\|x^{(0)} - x^*\|_2^2}{2k} = O(1/k)$

**μ-强凸情形 **：$f(x^{(k)}) - p^* \le \left(1 - \dfrac{\mu}{L}\right)^k \cdot \dfrac{L}{2}\|x^{(0)} - x^*\|_2^2$

条件数 $\kappa = L/\mu$ 决定收敛速度。κ 大（函数" 病态"）时收敛慢。这是梯度下降的主要局限。

Nesterov 加速：通过动量（momentum）技巧，一般凸情形可提升至 $O(1/k^2)$（Nesterov, 1983）：

$$x^{(k+1)} = y^{(k)} - t_k \nabla f(y^{(k)}), \quad y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3}(x^{(k+1)} - x^{(k)}).$$

这已是无梯度信息下一般凸函数最优化的**最优一阶方法**（Nesterov 最优性定理）。

7.6.4 5.3 梯度下降 $O(1/k)$ 收敛性证明

这是本章最重要的证明之一，建议在课堂上完整走一遍。

第一步：L-光滑性给出递降

L-光滑性等价于二次上界：

$$f(y) \le f(x) + \nabla f(x)^{\mathsf{T}}(y - x) + \frac{L}{2}\|y - x\|_2^2$$

（这是 L-光滑性的标准等价刻画，可以通过积分或中值定理证明。）

取 $x = x^{(k)}$，$y = x^{(k+1)} = x^{(k)} - t\nabla f(x^{(k)})$：

$$\begin{aligned} f(x^{(k+1)}) &\le f(x^{(k)}) - t\|\nabla f(x^{(k)})\|_2^2 + \frac{Lt^2}{2}\|\nabla f(x^{(k)})\|_2^2 \\ &= f(x^{(k)}) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x^{(k)})\|_2^2 \end{aligned}$$

取 $t = 1/L$，则 $1 - Lt/2 = 1/2$，得：

$$f(x^{(k+1)}) \le f(x^{(k)}) - \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$$

记 $\Delta_k = f(x^{(k)}) - p^*$，则 $\Delta_{k+1} \le \Delta_k - \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$。

第二步：建立距离递推

$$\begin{aligned} \|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - t\nabla f(x^{(k)}) - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2t\nabla f(x^{(k)})^{\mathsf{T}}(x^{(k)} - x^*) + t^2\|\nabla f(x^{(k)})\|_2^2 \end{aligned}$$

由凸性的一阶条件（以 x^* 作为 y，$x^{(k)}$ 作为 x）：

$$\nabla f(x^{(k)})^{\mathsf{T}}(x^{(k)} - x^*) \ge f(x^{(k)}) - f(x^*) = \Delta_k$$

所以：

$$\|x^{(k+1)} - x^*\|_2^2 \le \|x^{(k)} - x^*\|_2^2 - 2t\Delta_k + t^2\|\nabla f(x^{(k)})\|_2^2$$

第三步：消元与求和

由第一步得 $\|\nabla f(x^{(k)})\|_2^2 \le 2L(\Delta_k - \Delta_{k+1})$。代入第二步并令 $t = 1/L$：

$$\begin{aligned}\|x^{(k+1)} - x^*\|_2^2 &\le \|x^{(k)} - x^*\|_2^2 - \frac{2}{L}\Delta_k + \frac{2}{L}(\Delta_k - \Delta_{k+1}) \\ &= \|x^{(k)} - x^*\|_2^2 - \frac{2}{L}\Delta_{k+1}\end{aligned}$$

重排：

$$\Delta_{k+1} \le \frac{L}{2}\big(\|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2\big)$$

对 $t = 0, 1, \dots, k-1$ 求和（右端 telescoping）：

$$\sum_{t=0}^{k-1} \Delta_{t+1} \le \frac{L}{2}\|x^{(0)} - x^*\|_2^2$$

由于每一步函数值不增（$\Delta_{t+1} \le \Delta_t$），有 $\Delta_k \le \frac{1}{k}\sum_{t=0}^{k-1}\Delta_{t+1}$。因此：

$$\Delta_k \le \frac{L\|x^{(0)} - x^*\|_2^2}{2k}$$

证毕。■

教学建议：这个证明包含三个核心技巧——L-光滑性的二次上界、凸函数的一阶下界、telescoping 求和。每个技巧都值得单独强调。建议先在黑板上写出证明的整体结构（三步），然后逐步填充细节。

7.6.5 5.4 最速下降法

核心思想

梯度下降的方向依赖于范数的选择。改变内积（从而改变范数），就改变了”最速下降”的含义。

一般范数下的最速下降方向：

$$\Delta x_{\text{sd}} = \|\nabla f(x)\|_* \cdot \operatorname{argmin}_{\|v\|\le 1} \nabla f(x)^{\mathsf{T}} v$$

特殊情形：

- Euclid 范数：最速下降方向 = 负梯度（标准梯度下降）
- 二次范数 $\|x\|_P$：方向 $= -P^{-1}\nabla f(x)$（预处理梯度下降）
- ℓ_1 范数：方向 = 坐标方向（坐标下降法）

- ℓ_∞ 范数：方向 = 符号梯度（sign gradient）

Newton 方法作为最速下降：Newton 方法本质上是在由局部 Hessian 诱导的范数 $\|v\|_{\nabla^2 f(x)} = \sqrt{v^\mathsf{T}\nabla^2 f(x) v}$ 下的最速下降。从这个角度看，Newton 方法" 自动" 选择了最优的局部范数。

7.6.6 5.5 Newton 方法

Newton 步

$$\nabla^2 f(x)\Delta x_{\text{nt}} = -\nabla f(x)$$

几何来源

Newton 步最小化 f 在 x 处的二阶 Taylor 展开：

$$\hat{f}(x+v) = f(x) + \nabla f(x)^\mathsf{T} v + \frac{1}{2} v^\mathsf{T}\nabla^2 f(x) v$$

对 v 求导：$\nabla f(x) + \nabla^2 f(x) v = 0 \implies v = -\nabla^2 f(x)^{-1}\nabla f(x)$。

Newton 减量

$$\lambda(x) = \left(\nabla f(x)^\mathsf{T}\nabla^2 f(x)^{-1}\nabla f(x)\right)^{1/2} = \sqrt{-\nabla f(x)^\mathsf{T}\Delta x_{\text{nt}}}$$

$\lambda(x)$ 是 Newton 方法中的**天然停止准则**。当 $\nabla^2 f \succeq mI$ 时，$f(x) - p^* \le \lambda(x)^2/(2m)$，因此 $\lambda(x)^2/2$ 是当前函数值与最优值差距的精确上界（无需知道 p^*）。

算法 5.2：带回溯线搜索的 Newton 方法

给定初始点 $x^{(0)} \in \mathbf{dom}\, f$，容忍度 $\varepsilon > 0$，参数 $\alpha \in (0, 0.5)$，$\beta \in (0,1)$，**repeat** $k = 0, 1, \ldots$：

1. 求解线性系统 $\nabla^2 f(x^{(k)})\,\Delta x_{\text{nt}} = -\nabla f(x^{(k)})$，得 Newton 步 Δx_{nt}
2. 计算 Newton 减量：$\lambda^2 := -\nabla f(x^{(k)})^\mathsf{T}\Delta x_{\text{nt}} = \Delta x_{\text{nt}}^\mathsf{T}\nabla^2 f(x^{(k)})\Delta x_{\text{nt}}$
3. **if** $\lambda^2/2 \le \varepsilon$：**break**（$f(x^{(k)}) - p^*$ 的估计已充分小）
4. 回溯线搜索：令 $t := 1$；while $f(x^{(k)} + t\Delta x_{\text{nt}}) > f(x^{(k)}) - \alpha t\lambda^2$：令 $t := \beta t$
5. 更新：$x^{(k+1)} := x^{(k)} + t\Delta x_{\text{nt}}$

每步计算量：Hessian 计算 $O(n^2)$，Cholesky 分解 $O(n^3/3)$，前/后向替换 $O(n^2)$，总计 $O(n^3)$/步。实践中从合理初始点出发通常仅需 **5–10 步**即可达机器精度。

Newton 方法的收敛性

设 f 满足 $\nabla^2 f \succeq mI$，$\nabla^2 f$ Lipschitz 连续（常数 L）。收敛分两阶段：

- **阻尼阶段**：远离 x^* 时使用回溯线搜索，每步函数值至少下降常数 $\gamma > 0$
- **二次阶段**：进入 $\|\nabla f\|_2 \le \eta = 2m^2/L$ 区域后取 $t_k = 1$，误差 $e_k = \|x^{(k)} - x^*\|_2$ 满足

$$e_{k+1} \le \frac{L}{2m^2}\, e_k^2$$

令 $c = L/(2m^2)$，若 $ce_0 < 1$ 则 $e_k \le c^{-1}(ce_0)^{2^k}$：**正确有效数字位数每步翻倍**（如第 k 步有 3 位有效数字，第 $k+1$ 步约有 6 位，第 $k+2$ 步约 12 位）。

仿射不变性

Newton 步在变量的可逆仿射变换 $\tilde{x} = Ax$ 下不变：$\Delta\tilde{x}_{\mathrm{nt}} = A\Delta x_{\mathrm{nt}}$。这是梯度下降不具备的性质，也是 Newton 方法收敛速度与条件数无关的深层原因。

7.6.7 5.6 拟牛顿方法——BFGS 与 L-BFGS

动机

Newton 方法需要计算和分解 Hessian（$O(n^3)$ 成本），在高维问题中不现实。拟牛顿方法用**累积的梯度和迭代点信息**逐步构建 Hessian 的近似。

BFGS 更新

$$B_{k+1} = B_k + \frac{y_k y_k^{\mathsf{T}}}{y_k^{\mathsf{T}} s_k} - \frac{B_k s_k s_k^{\mathsf{T}} B_k}{s_k^{\mathsf{T}} B_k s_k}$$

其中 $s_k = x^{(k+1)} - x^{(k)}$，$y_k = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$。

关键性质：

1. **秩 2 更新**：每步只需 $O(n^2)$ 操作更新 B_k 或其逆
2. **保持正定性**：若 $B_k \succ 0$ 且 $y_k^{\mathsf{T}} s_k > 0$（Wolfe 曲率条件保证），则 $B_{k+1} \succ 0$
3. **割线条件**：$B_{k+1} s_k = y_k$（离散 Newton 方程 $\nabla^2 f(x^{(k)}) s_k \approx y_k$）
4. **超线性收敛**：强凸情形 + Wolfe 线搜索下 BFGS 超线性收敛（Dennis–Moré 定理）

L-BFGS 两环递推

只存储最近 m 对 (s_j, y_j)（m 通常取 5–20），每步代价 $O(mn)$，适用于 $n > 10^4$ 的大规模问题。记 $\rho_j = 1/(y_j^{\mathsf{T}} s_j)$，计算搜索方向 $d_k = -H_k \nabla f(x^{(k)})$：

前向循环（修正梯度）：令 $q := \nabla f(x^{(k)})$，对 $j = k-1, k-2, \ldots, k-m$：

$$\alpha_j := \rho_j\, s_j^{\mathsf{T}} q; \quad q := q - \alpha_j y_j.$$

初始缩放：令 $r := H^{(0)} q$，其中 $H^{(0)} = \gamma_k I$，$\gamma_k = \dfrac{s_{k-1}^{\mathsf{T}} y_{k-1}}{\|y_{k-1}\|_2^2}$（Barzilai–Borwein 初始缩放）。

后向循环（恢复方向）：对 $j = k-m, \ldots, k-1$：

$$\beta := \rho_j\, y_j^{\mathsf{T}} r; \quad r := r + s_j(\alpha_j - \beta).$$

令搜索方向 $d_k := -r$。

存储 m 对共 $2mn$ 个浮点数，两环递推共 $4mn$ 次浮点运算。

7.6.8 5.7 方法选择指南

问题规模	推荐方法	原因
$n \le 100$	Newton	二次收敛，总步数 ≤ 10；$O(n^3)$/步可接受
$100 < n \le 10^4$	BFGS / L-BFGS	超线性收敛 + $O(n^2)$ / $O(mn)$ 每步
$n > 10^4$	L-BFGS 或梯度下降	$O(mn)$ 每步；Nesterov 加速可用
非光滑	次梯度 / 邻近梯度	梯度不存在或 $f = g + h$（g 光滑，h 简单）

7.7 第六节等式约束优化

7.7.1 教学目标

学完本节后，学生应能：

1. 写出等式约束凸优化问题的 KKT 系统
2. 实现等式约束 Newton 方法
3. 理解消元法及其优缺点

7.7.2 6.1 等式约束凸优化问题

问题形式

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

其中 f 凸且二阶可微，A 行满秩（$\operatorname{rank} A = p < n$）。

KKT 条件

由于没有不等式约束，KKT 条件简化为（f 凸，强对偶自动成立因为只有仿射等式约束）：

$$\nabla f(x^*) + A^{\mathsf{T}}\nu^* = 0, \qquad Ax^* = b$$

$n + p$ 个方程，$n + p$ 个未知量。这个系统的结构是**对称但不定的**（saddle-point structure）。

7.7.3 6.2 等式约束 Newton 方法

Newton 步的计算

在可行点 x（$Ax = b$）处，求解 KKT 型系统：

$$\begin{bmatrix} \nabla^2 f(x) & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\mathrm{nt}} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

为什么第二块右边是 0？因为 $Ax = b$ 是线性约束，其 Taylor 展开没有高阶项。Newton 步必须满足 $A(x + \Delta x_{\mathrm{nt}}) = b \implies A\Delta x_{\mathrm{nt}} = 0$。

关键性质

1. **可行性保持**：$A\Delta x_{\mathrm{nt}} = 0$ 自动保证，因此每次迭代后 $Ax^{(k+1)} = b$ 仍然成立。
2. **Newton 减量**：$\lambda(x) = \sqrt{\Delta x_{\mathrm{nt}}^\mathsf{T} \nabla^2 f(x) \Delta x_{\mathrm{nt}}}$。
3. **对偶变量**：收敛后，w 的值就是最优 Lagrange 乘子 ν^*。

回溯线搜索

条件为 $f(x + t\Delta x_{\mathrm{nt}}) \le f(x) - \alpha t \lambda^2$。

注意这里用的是 λ^2 而不是梯度范数——因为约束优化中的停止准则是 $\lambda^2/2 \le \varepsilon$。

7.7.4 6.3 消元法

方法

将可行集参数化为 $x = Fz + x_0$，其中 F 的列是 $\ker A$ 的基，x_0 是特解。代回得无约束问题：

$$\min_z \tilde{f}(z) = f(Fz + x_0)$$

优缺点

优点：转化为无约束问题，维数降低，可用标准无约束算法。

缺点：

- F 的计算可能破坏稀疏性——如果 A 是稀疏的，F 通常不是
- 消元法的 Newton 步与直接求解 KKT 系统的 Newton 步是等价的——选择哪种取决于实现便利性

实践建议：当约束数量 p 较小时，直接求解 KKT 系统更好；当自由度数 $n - p$ 较小时，消元法可能更方便。

7.8 第七节内点法

7.8.1 教学目标

学完本节后，学生应能：

1. 理解对数障碍函数的思想

2. 描述中心路径及其性质
3. 对比障碍方法和原始-对偶内点法
4. 了解现代内点法求解器的生态

7.8.2 7.1 障碍函数思想

动机

如何用无约束/等式约束方法来求解带不等式约束的凸优化问题?

核心思想：用光滑的障碍函数替代不等式约束的" 硬壁"。

对数障碍函数

$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$$

性质分析：

- 当 $f_i(x) \to 0^-$（趋近约束边界），$-\log(-f_i(x)) \to +\infty$——障碍函数趋于无穷，阻止 x 越界
- 当 $f_i(x)$ 为很大的负数（在可行域内部深处），$-\log(-f_i(x))$ 较小——障碍函数允许探索
- $\phi(x)$ 是凸函数（$-\log(-u)$ 关于 u 凸且不减，f_i 凸，由标量复合规则知 $-\log(-f_i(x))$ 凸）

教学建议：可以用" 在房间里走路" 的比喻——墙壁是约束，对数障碍就像在离墙近时感受到的排斥力，越近越强。

7.8.3 7.2 中心路径

障碍问题

$$\begin{aligned} \text{minimize} \quad & t f_0(x) + \phi(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

参数 $t > 0$ 在目标和障碍之间做权衡：

- t 小：障碍占主导 $\to$ 解在可行域中心
- t 大：目标占主导 $\to$ 解逼近原始最优解 x^*

中心路径的性质

1. 对所有 $t > 0$，$x^*(t)$ 存在且唯一（在适当正则性条件下）
2. $x^*(t)$ 严格可行
3. $\lim_{t\to\infty} x^*(t) = x^*$（逼近原始最优解）
4. **对偶间隙**：$f_0(x^*(t)) - p^* \le m/t$。这是中心路径最重要的量化性质。

**m/t 公式的含义 **：每当我们把 t 增大到 μt（$\mu > 1$），对偶间隙至少缩小到原来的 $1/\mu$。这意味着需要倍增 t 大约 $\log(m/(\varepsilon t^{(0)}))/\log\mu$ 次即可达到精度 ε。

7.8.4 7.3 障碍方法

算法

外层循环（参数更新）：$t \leftarrow \mu t$ **内层循环**（中心化）：用等式约束 Newton 方法精确求解当前的障碍问题

收敛性

如果 f_0 和 $-\log(-f_i)$ 都是**自协调（self-concordant）**函数，则每次内层中心化仅需**常数量** Newton 步（与 t 和精度 ε 无关）。这是内点法多项式复杂度的理论基础。

总 Newton 步数 $\approx \left\lceil \frac{\log(m/(\varepsilon t^{(0)}))}{\log\mu} \right\rceil \times \text{const}$。

大多数实际中出现的凸函数（线性、二次、$-\log$、log-sum-exp 等）都是自协调的。

** 参数 μ 的选择 **：μ 小 $\rightarrow$ 中心化容易但需要更多外层迭代；μ 大 $\rightarrow$ 外层迭代少但中心化困难。实践中 $\mu \approx 10 \sim 20$ 取得最佳平衡。

7.8.5 7.4 原始-对偶内点法

修改的 KKT 条件

障碍方法的思想等价于” 修改” 互补松弛条件：

原 KKT：$\lambda_i f_i(x) = 0$ 修改后：$-\lambda_i f_i(x) = 1/t$（$t > 0$）

当 $t \to \infty$ 时，$1/t \to 0$，修改的 KKT 条件趋于原 KKT 条件。

与障碍方法的区别

障碍方法	原始-对偶方法
外层循环更新 t，内层循环精确求解障碍问题	每一步同时更新 (x, λ, ν) 和 t
内层需要多次 Newton 步	每步只做一次 Newton 步
结构清晰，理论分析方便	实践中效率更高

搜索方向

每一步求解关于 $(\Delta x, \Delta\lambda, \Delta\nu)$ 的 Newton 系统。通过分块消元可以简化为对称 KKT 型系统：

$$\begin{bmatrix} H_{\text{pd}} & A^{\mathsf{T}} \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\nu \end{bmatrix} = - \begin{bmatrix} \tilde{r}_{\text{dual}} \\ r_{\text{pri}} \end{bmatrix}$$

其中 H_{pd} 包含了目标函数的 Hessian、约束函数的 Hessian 和一个**障碍项** $\sum_i \frac{\lambda_i}{-f_i(x)} \nabla f_i \nabla f_i^{\mathsf{T}}$——这一项来自对数障碍的 Hessian，确保搜索方向指向可行域内部。

教学建议：搜索方向的计算是最技术性的部分。课堂上重点讲清楚修改 KKT 条件的思想，具体的矩阵形式可以留给学生课后看。

7.8.6 7.5 内点法软件生态

求解器

求解器	擅长的问题类型
Mosek	LP、QP、SOCP、SDP（学术免费）
Gurobi	LP、QP（商业领先）
IPOPT	大规模非线性规划
ECOS	嵌入式 SOCP
SCS	锥规划（基于 ADMM）
CVXOPT	Python 原生 SDP

建模框架

- **CVXPY**（Python）：声明式建模，你写数学问题，它自动选择求解器
- **CVX**（MATLAB）：经典工具，语法优雅
- **JuMP**（Julia）：高性能

教学建议：给学生看一个 CVXPY 的示例，让他们看到”写 5 行代码解决一个优化问题”有多容易。这能激发学习凸优化的兴趣。

7.8.7 7.6 凸优化的应用全景

六个主要应用方向

1. **机器学习与统计**：Lasso（ℓ_1 正则化的 QP）、SVM（QP 对偶）、逻辑回归（光滑凸无约束）
2. **信号与图像处理**：压缩感知（ℓ_1 极小化）、TV 去噪（SOCP）
3. **金融工程**：投资组合（QP）、风险极小化（LP/SOCP）
4. **通信**：功率分配（GP）、MIMO 预编码（SOCP/SDP）
5. **控制**：MPC（QP）、H_∞ 控制（LMI/SDP）
6. **运筹**：供应链、交通流、电力调度

教学建议：这里不需要展开讲每个应用，但应该让学生感受到凸优化的”无处不在”。可以挑两个最熟悉的领域稍微深入一些。

7.8.8 总结

本章涵盖了凸优化的核心知识体系：

1. **凸集与凸函数**：基础语言，判断凸性是第一步
2. **凸优化问题**：LP、QP、QCQP、SOCP、SDP、GP 的层次结构
3. **对偶理论**：Lagrange 对偶、Slater 条件、KKT 条件——理论与实践连接的桥梁
4. **无约束算法**：梯度下降（$O(1/k)$）、Newton（二次收敛）、BFGS/L-BFGS
5. **等式约束算法**：等式约束 Newton、消元法
6. **内点法**：障碍方法、原始-对偶方法——现代凸优化求解的核心引擎

核心信息：凸优化之所以重要，不仅因为它有优美的理论，更因为它有**可靠**、**高效的算法**来解决实际问题。能识别和建模凸优化问题，是连接数学与工程的桥梁。

第八章　机器学习入门

8.1　第八章机器学习入门—讲稿

本讲稿配合幻灯片 `machine-learning.md` 使用，按七个小节组织。每小节包含：教学目标、核心概念的详细解释、完整证明、例题精讲、易错点提醒和教学建议。本章内容取材于周志华《机器学习》第 1–6 章，并结合 Boyd《Convex Optimization》建立起优化与机器学习的联系。

主要参考文献

周志华，《机器学习》（西瓜书），清华大学出版社，2016.　Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997.　Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.　Stephen Boyd & Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.（以下简称"Boyd"）

8.2　第一节机器学习概述

8.2.1　教学目标

学完本节后，学生应能：

1. 陈述机器学习的定义和核心三要素（数据、算法、模型）
2. 区分分类、回归、聚类等基本学习任务
3. 解释监督学习与无监督学习的区别
4. 理解假设空间、归纳偏好和奥卡姆剃刀原则
5. 陈述「没有免费的午餐」定理及其启示

8.2.2　1.1 什么是机器学习

讲解思路

这是整章的开篇。需要让学生建立对机器学习的基本认知：它是通过计算手段从数据中学习规律的方法论。

逐点讲解：

1. **Mitchell 的形式化定义**：称一个计算机程序从经验 E 中学习任务 T，并以性能度量 P 来衡量，若其在任务 T 上的性能随经验 E 的增加而提高。这一定义将机器学习的三个核心

要素绑定在一起。

2. **三要素的对应关系**：
   - 数据 ↔ 经验 E：训练样本的集合
   - 算法 ↔ 学习过程：如何从数据中提取规律
   - 模型 ↔ 学习结果：用于对新样本做预测的函数
3. **与传统编程的区别**：传统编程是「规则 + 数据 → 答案」，机器学习是「数据 + 答案 → 规则」。
4. **与优化问题的联系**：机器学习中的「学习」本质上是一个**优化问题**。例如，线性回归的最小二乘估计正是第七章中讨论的无约束凸优化问题。理解这一联系是本章的重要目标之一。

教学提示：可以用一个简单的例子引入——如何让计算机区分猫和狗的图片？传统方法需要人工编写规则（耳朵形状？毛色？），而机器学习让计算机从大量标注好的图片中自动学习判别规则。

8.2.3 1.2 基本术语

核心概念

机器学习有自己的术语体系，需要在一开始就建立清晰的概念。

数据集与样本：

数据集 D 是 m 个样本（示例）的集合。每个样本由 d 个属性（特征）描述，构成 d 维样本空间 $\mathcal{X}$ 中的一个特征向量：

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathcal{X}.$$

标记与样例：

标记 $y_i \in \mathcal{Y}$ 是关于样本 $\mathbf{x}_i$ 的结果信息。标记空间 $\mathcal{Y}$ 在分类问题中是离散的类别集合，在回归问题中是 $\mathbb{R}$。

样例（example）是拥有标记的样本：$(\mathbf{x}_i, y_i)$。

训练与测试：

训练集 $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ 用于训练（学习）模型。测试样本的标记未知，由模型预测。

重要约定：本章记数据矩阵 $\mathbf{X} \in \mathbb{R}^{m \times d}$（每行一个样本），标记向量 $\mathbf{y} \in \mathbb{R}^m$（回归）或 $\mathbf{y} \in \{0, 1\}^m$（二分类）。这与第七章凸优化中的记号一致。

西瓜数据集示例

例 8.2.1 (西瓜数据集)**.** 考虑周志华《机器学习》中的经典西瓜数据集。样本 $\mathbf{x}$ 包含属性：

- 色泽：{ 青绿, 乌黑, 浅白 }（离散）
- 根蒂：{ 蜷缩, 稍蜷, 硬挺 }（离散）
- 敲声：{ 浊响, 沉闷, 清脆 }（离散）
- 含糖率：连续值

- 密度：连续值

标记 y：好瓜 $= 1$，坏瓜 $= 0$。

教学提示：用西瓜数据集贯穿全章，使学生在具体例子上理解抽象概念。西瓜数据集的好处是：属性既有离散也有连续，二分类问题直观，数据量适中（17 个样本）。

8.2.4 1.3 分类与回归

核心区别

- **分类（Classification）**：预测离散的标记值。输出空间 $\mathcal{Y}$ 是有限集合。
  - 二分类：$\mathcal{Y} = \{+1, -1\}$ 或 $\{0, 1\}$
  - 多分类：$\mathcal{Y} = \{C_1, C_2, \ldots, C_k\}$
- **回归（Regression）**：预测连续的数值。输出空间 $\mathcal{Y} = \mathbb{R}$。
- **聚类（Clustering）**：无监督任务，将样本划分为若干组（簇），使得组内相似度高、组间相似度低。

分类的数学视角

从概率角度看，分类问题中我们试图估计后验概率 $P(y \mid \mathbf{x})$。对于二分类：

- 判别模型：直接学习决策边界（如 SVM、感知机）
- 概率模型：先估计 $P(y \mid \mathbf{x})$（如逻辑回归），再以 0.5 为阈值决策

易错点：对数几率回归（logistic regression）虽然名字中有「回归」，实际上是一个分类方法——它预测的是样本属于某类的概率。这一混淆在学生中非常普遍。

8.2.5 1.4 监督学习与无监督学习

核心区别

两者的根本区别在于训练数据是否拥有标记信息：

- **监督学习**：训练数据有标记。目标是学习从输入到输出的映射 $f : \mathcal{X} \to \mathcal{Y}$。
- **无监督学习**：训练数据无标记。目标是揭示数据内在的分布结构。

此外还有**半监督学习**（部分数据有标记）和**强化学习**（通过与环境交互获得奖励信号来学习策略）。

i.i.d. 假设

机器学习理论分析的基本前提：样本独立地从同一个未知分布 $\mathcal{D}$ 中采样。这一假设使得我们可以用训练误差来估计泛化误差。

形式化地：$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$。

教学提示：可以问学生：「如果训练数据和测试数据来自完全不同的分布会怎样?」——这引出了迁移学习、领域自适应等高级主题。实际应用中，i.i.d. 假设常常被违反。

8.2.6 1.5 假设空间与归纳偏好

详细解释

假设空间 $\mathcal{H}$ 是学习算法可能输出的所有假设（函数）的集合。学习过程本质上是在 $\mathcal{H}$ 中搜索与训练数据最匹配的假设。

数学表述：给定训练集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$，学习算法 $\mathfrak{L}$ 输出一个假设 $h\in\mathcal{H}$ 使得 $h(\mathbf{x}_i)\approx y_i$。若 $\mathcal{H}$ 包含目标函数，则称问题**可学习**。

归纳偏好（Inductive Bias）：

为什么需要归纳偏好？因为在有限训练样本下，可能有多个假设在训练集上表现相同（即这些假设是「等效」的），但它们在未见样本上的表现可能大相径庭。归纳偏好是算法对某类假设的先天倾向性。

例 8.2.2 (曲线拟合中的归纳偏好)**.** 给定 10 个数据点，在多项式空间中存在无穷多个不同次数（$d=9,10,11,\dots$）的多项式可以完美拟合这些点。若没有归纳偏好（如偏好低次多项式），算法无法在它们之间做出选择。实践中我们倾向于选择低次多项式（光滑性偏好）。

奥卡姆剃刀（Occam's Razor）：

「如无必要，勿增实体」——若有多个假设与观察一致，选择最简单的那个。奥卡姆剃刀是最常用的归纳偏好原则，但「简单」的定义方式（如参数个数、VC 维、描述长度等）取决于具体问题。

与优化的联系：归纳偏好通常由正则化项实现。例如，在最小二乘中加入 ℓ_2 正则化：$\min_{\mathbf{w}}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2$，其中 $\lambda>0$ 体现了对「小权重」的偏好。

8.2.7 1.6 没有免费的午餐定理

定理陈述

No Free Lunch Theorem (Wolpert & Macready, 1997)：

对于二分类问题，考虑**所有可能的目标函数**上的均匀分布，则任意两个学习算法的期望性能相同。形式化地：

$$\sum_{f} P(h\neq f\mid \mathfrak{L}_a,f)=\text{常数（与算法 }\mathfrak{L}_a\text{ 无关）}.$$

详细证明

证明思路（针对二分类、有限样本空间）：

设样本空间 $\mathcal{X}$ 有限，$|\mathcal{X}|=N$。考虑所有可能的目标函数 $f:\mathcal{X}\to\{0,1\}$，共有 2^N 个。对任意算法 $\mathfrak{L}_a$ 和训练集大小 $m<N$：

1. 训练集 D 包含 m 个样本及其标记，剩余 $N-m$ 个样本未见过。
2. 对于训练集外的每个样本 $\mathbf{x}$，在所有 2^N 个目标函数中，恰好有一半的函数给出 $f(\mathbf{x})=0$，另一半给出 $f(\mathbf{x})=1$。

3. 因此，考虑所有可能的目标函数，算法在测试样本上的期望错误率恒为 1/2。
4. 这意味着对所有算法的**平均**性能来说，没有优劣之分。

启示与应用

- **脱离具体问题谈算法优劣无意义**：算法的好坏取决于其归纳偏好是否与问题的真实规律一致。
- **不存在万能算法**：针对不同任务需要选择不同的算法。这一结论与优化中「没有适用于所有问题的优化算法」异曲同工。
- **理解问题上下文的重要性**：特征工程、领域知识、合适的模型选择比追求「最先进」算法更重要。

易错点：NFL 定理的前提是「在所有可能的问题上取平均」。在**实际应用中**，我们面对的问题通常具有特定结构（如平滑性、稀疏性、低维流形等），因此某些算法确实系统性地优于其他算法。另外，NFL 定理针对的是训练集外样本的泛化性能，而非训练误差。

8.2.8 1.7 机器学习发展简史

机器学习的发展经历了多个阶段：

1. **推理期（1950s–1960s）**：以逻辑推理为核心，代表工作有 A. Newell & H. Simon 的「逻辑理论家」程序。认为逻辑推理能力代表了智能的本质。
2. **知识工程期（1970s–1980s）**：专家系统盛行。将人类专家的知识编码为规则库，通过推理引擎进行决策。瓶颈：知识获取极其困难（知识工程瓶颈）。
3. **连接主义复兴（1980s–1990s）**：Rumelhart 等人在 1986 年重新发明了 BP 算法，使多层神经网络的训练成为可能，神经网络研究从第一次寒冬中复兴。同时期，Vapnik 的统计学习理论（1995）为 SVM 奠定了坚实的理论基础。
4. **统计学习时代（1990s–2010s）**：以 Vapnik 的统计学习理论和 SVM 为代表，强调泛化能力的理论保证。集成学习方法（Bagging, Boosting, Random Forest）在实践中广泛应用。
5. **深度学习时代（2012–至今）**：AlexNet 在 ImageNet 上的突破性表现开启了深度学习浪潮。关键里程碑包括：
   - 2012：AlexNet（8 层 CNN）在 ImageNet 上将 top-5 错误率从 26% 降至 15.3%
   - 2014：VGGNet、GoogLeNet（Inception）
   - 2015：ResNet（152 层，残差连接），top-5 错误率 3.57%（低于人类水平）
   - 2017：Transformer（Attention is All You Need）
   - 2022：ChatGPT / GPT-4 引爆大语言模型浪潮

教学提示：可以简要展示 ImageNet 分类错误率从 2010 年的 28% 降至 2015 年低于人类水平（约 5%）的过程，让学生感受到技术进步的震撼。

8.3 第二节模型评估与选择

8.3.1 教学目标

学完本节后，学生应能：

1. 区分训练误差与泛化误差
2. 解释过拟合和欠拟合的成因与现象
3. 熟练运用留出法、交叉验证法和自助法进行模型评估
4. 理解偏差-方差分解的基本思想

8.3.2 2.1 误差与过拟合

核心概念

错误率（Error Rate）：分类错误的样本数占总样本数的比例：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\mathbf{x}_i)\neq y_i).$$

精度（accuracy）$=1-E(f;D)$。

训练误差（Training Error）：学习器在训练集上的误差。也称为经验误差（empirical error），记为 $\hat{E}(f)$。

泛化误差（Generalization Error）：学习器在未见样本（来自同一分布）上的期望误差：

$$E(f)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{I}(f(\mathbf{x})\neq y)].$$

这是我们真正关心的量，但无法直接计算（因为 $\mathcal{D}$ 未知）。

过拟合与欠拟合

- **过拟合（Overfitting）**：训练误差很小，但泛化误差很大。学习器把训练样本自身的特性当成了普遍规律。成因：模型过于复杂（相对于训练数据量）。
- **欠拟合（Underfitting）**：训练误差本身都很大。学习器未能充分捕捉训练数据的规律。成因：模型过于简单。

模型选择的核心问题：在欠拟合与过拟合之间寻找最优平衡。

偏差-方差分解

对于回归问题（平方损失），泛化误差可以分解为：

$$\mathbb{E}_{\mathcal{D}}[(f(\mathbf{x};D)-y)^2]=\underbrace{(\mathbb{E}_D[f(\mathbf{x};D)]-\bar{y})^2}_{\text{偏差}^2}+\underbrace{\mathbb{E}_D[(f(\mathbf{x};D)-\mathbb{E}_D[f(\mathbf{x};D)])^2]}_{\text{方差}}+\underbrace{\text{噪声}^2}_{\text{不可约}}.$$

- **偏差**：模型预测的期望与真实值之间的差异。高偏差 → 欠拟合
- **方差**：模型对训练集变化的敏感度。高方差 → 过拟合

教学提示：多项式曲线拟合是最经典的例子。低次多项式（如一次）偏差大 → 欠拟合；高次多项式（如 $d=m-1$）方差大 → 完美通过所有训练点但剧烈震荡 → 过拟合。

8.3.3 2.2 评估方法

留出法（Hold-out）

将数据集 D 划分为互斥的训练集 S 和测试集 T：

$$D = S \cup T, \quad S \cap T = \varnothing.$$

- 常见比例：$|S| : |T| \approx 7:3$ 或 $8:2$
- 需保持类别比例（**分层采样**，stratified sampling）
- 单次划分不稳定，建议多次随机划分取平均

交叉验证法（Cross Validation）

将 D 划分为 k 个大小相似的互斥子集 $D_1, D_2, \ldots, D_k$。每次用其中一个子集测试，其余 $k-1$ 个子集训练，共进行 k 次。最终结果是 k 次测试结果的均值。

- $k = 10$：10 折交叉验证（最常用）
- $k = m$（样本数）：留一法（Leave-One-Out, LOO）。优点：受随机因素影响小；缺点：计算开销大（$O(m)$ 次训练）
- 分层 k 折交叉验证：保持每折的类别分布与原始数据一致

自助法（Bootstrapping）

从 m 个样本中有放回地随机采样 m 次，得到训练集 D'。样本在 m 次采样中从未被抽中的概率为：

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m = \frac{1}{e} \approx 0.368.$$

因此约 36.8% 的样本未出现在 D' 中，可作为测试集（称为**包外估计**，out-of-bag estimate）。

- 优点：适用于数据集较小、难以划分训练/测试集的情况
- 缺点：改变了原始数据分布（引入了重复样本），估计有偏差
- 自助法在**集成学习**（如随机森林）中广泛用于评估基学习器性能

三种方法的对比

评估方法对比

方法	数据利用	计算开销	适用场景
留出法	划分后互斥	低	大数据集
交叉验证	每个样本均作测试	中（k 次训练）	中等数据集
自助法	有放回，36.8% 作测试	中	小数据集

教学提示：可以让学生在计算机上亲自实现这三种评估方法，比较它们在不同大小数据集上的表现差异。

8.4 第三节线性模型

8.4.1 教学目标

学完本节后，学生应能：

1. 写出线性模型的基本形式并解释其可解释性优势
2. 推导线性回归的最小二乘解
3. 理解对数几率回归（逻辑回归）的概率解释和极大似然估计
4. 将线性模型与凸优化建立起联系（见第七章）

8.4.2 3.1 线性模型基本形式

定义与直觉

线性模型试图学得属性的线性组合：

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b = \mathbf{w}^{\mathsf{T}}\mathbf{x} + b.$$

线性模型的**核心优势**是可解释性强：权重 w_i 的大小和符号直接反映了对应属性对预测结果的贡献。

与优化问题的对应

线性模型的学习过程本质上是一个优化问题。$\mathbf{w}, b$ 是决策变量，目标函数由损失函数定义，这完全对应第七章中「优化问题的标准形式」。表 8.4.2 给出了几种常见线性模型的优化形式。

线性模型与优化问题的对应

模型	损失函数	优化问题类型	第七章对应
线性回归	ℓ_2 平方损失	无约束凸 QP	§7.4 无约束优化
Ridge 回归	ℓ_2 平方 + ℓ_2 正则	无约束凸 QP	§7.4 + ℓ_2 罚
Lasso	ℓ_2 平方 + ℓ_1 正则	无约束凸非光滑	§7.6 次梯度方法
逻辑回归	交叉熵损失	无约束凸光滑	§7.4 Newton 法
SVM（原）	Hinge 损失	约束凸 QP	§7.5 对偶理论

8.4.3 3.2 线性回归

最小二乘估计

目标：使预测值与真实标记之间的均方误差最小化：

$$(\mathbf{w}^*, b^*) = \arg\min_{(\mathbf{w},b)} \sum_{i=1}^{m} \big(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b - y_i\big)^2.$$

闭式解的推导

令 $\hat{\mathbf{w}} = (\mathbf{w}, b)$，$\hat{\mathbf{x}}_i = (\mathbf{x}_i, 1)$。目标函数：

$$E_{\hat{\mathbf{w}}} = \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 = (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y})^{\mathsf{T}}(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}),$$

其中 $\mathbf{X} \in \mathbb{R}^{m\times(d+1)}$ 为增广数据矩阵，$\mathbf{y} \in \mathbb{R}^m$ 为标记向量。

梯度：

$$\nabla_{\hat{\mathbf{w}}} E_{\hat{\mathbf{w}}} = 2\mathbf{X}^{\mathsf{T}}(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}).$$

令梯度为零，得**正规方程**（normal equations）：

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

当 $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ 满秩时：

$$\hat{\mathbf{w}}^* = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

数值稳定性

当 $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ 接近奇异时，直接求逆在数值上不稳定。在数值分析（第一章）中，我们学过更好的方法：

- **QR 分解**：$\mathbf{X} = \mathbf{QR}$，则 $\hat{\mathbf{w}}^* = \mathbf{R}^{-1}\mathbf{Q}^{\mathsf{T}}\mathbf{y}$，避免了计算 $\mathbf{X}^{\mathsf{T}}\mathbf{X}$
- **SVD 分解**：$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$，则 $\hat{\mathbf{w}}^* = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{y}$，可处理秩亏情形
- **正则化**：引入 ℓ_2 正则化（Ridge）：$\hat{\mathbf{w}}^* = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$，当 $\lambda > 0$ 时矩阵必然正定

Python 实现

```python
import numpy as np
from sklearn.linear_model import LinearRegression, Ridge, Lasso

# 生成合成数据
np.random.seed(42)
X = np.random.randn(100, 3)
true_w = np.array([2.0, -1.5, 0.5])
y = X @ true_w + 0.3 * np.random.randn(100)

# 线性回归（最小二乘）
lr = LinearRegression(fit_intercept=False)
lr.fit(X, y)
print(f"OLS: w = {lr.coef_}")
print(f"||w - true_w|| = {np.linalg.norm(lr.coef_ - true_w):.4f}")

# Ridge 回归（L2 正则化）
ridge = Ridge(alpha=1.0, fit_intercept=False)
```

```
ridge.fit(X, y)
print(f"Ridge: w = {ridge.coef_}")

# Lasso (L1 正则化，可产生稀疏解)
lasso = Lasso(alpha=0.1, fit_intercept=False)
lasso.fit(X, y)
print(f"Lasso: w = {lasso.coef_}")
```

教学提示：这是学生第一次在机器学习背景下系统见到最小二乘问题的求解。可以回顾数值分析中 QR 分解和 SVD 分解求解最小二乘问题的方法，建立与前面章节的联系。

8.4.4 3.3 对数几率回归（逻辑回归）

从线性回归到分类

对于二分类问题 $y \in \{0, 1\}$，直接使用线性回归不合适——输出是实数而非概率。解决方案是用一个**联系函数**（link function）将线性模型的输出映射到 $(0, 1)$。

最常用的联系函数是对数几率函数（sigmoid/logistic）：

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

于是：

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^{\mathsf{T}}\mathbf{x} + b)}}.$$

几率与对数几率

事件发生与不发生的概率之比称为**几率**（odds）：

$$\text{odds} = \frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})} = e^{\mathbf{w}^{\mathsf{T}}\mathbf{x} + b}.$$

取对数得**对数几率**（log-odds，也称 logit）：

$$\ln \frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})} = \mathbf{w}^{\mathsf{T}}\mathbf{x} + b.$$

关键理解：逻辑回归**不直接预测类别**，而是预测 $P(y = 1 \mid \mathbf{x})$。决策时通常以 0.5 为阈值：若 $P(y = 1 \mid \mathbf{x}) \geq 0.5$（即 $\mathbf{w}^{\mathsf{T}}\mathbf{x} + b \geq 0$），预测为正类；否则为负类。

极大似然估计

似然函数：

$$L(\mathbf{w}, b) = \prod_{i=1}^{m} P(y_i \mid \mathbf{x}_i; \mathbf{w}, b).$$

对二分类问题，$y_i \in \{0, 1\}$，可以统一写为：

$$P(y_i \mid \mathbf{x}_i; \mathbf{w}, b) = [\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b)]^{y_i} \cdot [1 - \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b)]^{1 - y_i}.$$

取负对数得**交叉熵损失**（cross-entropy loss）：

$$\ell(\mathbf{w}, b) = -\sum_{i=1}^{m}\Big[y_i \ln \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b) + (1 - y_i)\ln(1 - \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b))\Big].$$

凸性证明

可以证明，交叉熵损失 $\ell(\mathbf{w}, b)$ 是 $(\mathbf{w}, b)$ 的**凸函数**（严格凸当 $\mathbf{X}$ 列满秩时）。证明要点：

1. 负对数似然可以写为 $\ell(\mathbf{w}, b) = \sum_{i=1}^{m} \log(1 + e^{-y_i(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b)})$（取 $y_i \in \{\pm 1\}$ 形式）
2. $\log(1 + e^{-t})$ 的二阶导数为 $\frac{e^t}{(1+e^t)^2} > 0$，是凸函数
3. 凸函数与仿射函数的复合仍是凸函数，凸函数的非负加权和仍是凸函数

因此，同第七章中的无约束凸优化问题一样，逻辑回归的交叉熵极小化可以使用梯度下降、Newton 方法或 L-BFGS 等方法求解（参见第七章 §7.4–§7.5）。

Python 实现

```python
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# 二分类数据
X = np.random.randn(200, 5)
true_w = np.array([1.0, -2.0, 0.5, -1.5, 0.0])
y = (1 / (1 + np.exp(-X @ true_w)) > 0.5).astype(int)

# 逻辑回归（使用 L-BFGS 优化器）
clf = LogisticRegression(penalty=None, solver='lbfgs')
clf.fit(X, y)
print(f"Accuracy: {accuracy_score(y, clf.predict(X)):.4f}")
print(f"Estimated w: {clf.coef_.round(3)}")
print(f"True w:      {true_w}")
```

易错点：虽然名字中有「回归」，但逻辑回归是分类方法。学生容易混淆。另外，「logistic regression」中的 logistic 指 logistic 函数（sigmoid），而非回归任务。

教学提示：本节与第七章凸优化的联系非常紧密。建议明确指出：逻辑回归的交叉熵损失是凸函数，其梯度下降（第七章 §7.4）和 Newton 法（第七章 §7.5）是标准求解方法。

8.5 第四节决策树

8.5.1 教学目标

学完本节后，学生应能：

1. 描述决策树的分治学习框架

2. 计算信息增益、增益率和基尼指数
3. 比较 ID3、C4.5 和 CART 的划分准则差异
4. 区分预剪枝与后剪枝并分析各自的优缺点

8.5.2　4.1 决策树的基本流程

结构与思想

决策树由根结点、若干内部结点和若干叶结点组成。学习过程遵循**分治法**：在每个结点上选择一个最优属性进行划分，将样本分配至子结点，递归进行直至满足停止条件。

1. 若当前结点包含的样本全属于同一类别 → 标记为该类的叶结点
2. 若当前属性集为空，或样本在所有属性上取值相同 → 标记为样本数最多的类的叶结点
3. 否则，选择最优划分属性，为每个属性值产生一个分支

教学提示：在课堂上可以画一棵简单决策树的示意图（如用西瓜数据集），让学生直观理解树结构和分治过程。

8.5.3　4.2 信息熵与信息增益

信息熵

设样本集 D 中第 k 类样本的比例为 p_k，则 D 的**信息熵**定义为：

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k.$$

信息熵度量了样本集合的**纯度**：$\mathrm{Ent}(D)$ 越小，纯度越高。

极值分析：

- 当所有样本属于同一类（$p_k = 1$，其余 $p_{k'} = 0$）：$\mathrm{Ent}(D) = 0$（最小值，纯度最高）
- 当各类样本均匀混合（$p_k = 1/|\mathcal{Y}|$）：$\mathrm{Ent}(D) = \log_2 |\mathcal{Y}|$（最大值）

信息增益

属性 a 对 D 的信息增益为划分前后熵的减少量：

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v).$$

信息增益越大，意味着使用属性 a 进行划分所获得的「纯度提升」越大。ID3 算法选择信息增益最大的属性。

例 8.5.1 (西瓜数据集的 ID3 划分计算)**.** 考虑西瓜数据集中 17 个样本（8 个好瓜，9 个坏瓜）。

初始熵：

$$\mathrm{Ent}(D) = -\frac{8}{17}\log_2\frac{8}{17} - \frac{9}{17}\log_2\frac{9}{17} \approx 0.9975.$$

按「色泽」划分：

- 青绿（6 样本）：3 好 3 坏，$\mathrm{Ent} = 1.0$
- 乌黑（6 样本）：4 好 2 坏，$\mathrm{Ent} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} \approx 0.9183$
- 浅白（5 样本）：1 好 4 坏，$\mathrm{Ent} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} \approx 0.7219$

$$\mathrm{Gain}(D, \text{色泽}) = 0.9975 - \frac{6}{17}\times 1.0 - \frac{6}{17}\times 0.9183 - \frac{5}{17}\times 0.7219 \approx 0.1081.$$

类似地可计算其他属性的信息增益，选择最大的进行划分。

信息增益的偏好问题

信息增益准则对可取值数目较多的属性有偏好。极端例子：若「编号」属性每个样本取值都不同，则按编号划分后每个子集只有一个样本，熵为零，信息增益达到最大——但这样的划分毫无泛化能力。

8.5.4 4.3 增益率与基尼指数

增益率（C4.5）

C4.5 使用增益率来减少对多值属性的偏好：

$$\mathrm{Gain\_ratio}(D, a) = \frac{\mathrm{Gain}(D, a)}{\mathrm{IV}(a)},$$

其中 $\mathrm{IV}(a) = -\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$ 称为属性 a 的**固有值**（intrinsic value）。属性取值越多，IV 越大，增益率相应降低。

实际使用中，C4.5 并不直接选增益率最大的属性，而是采用**启发式策略**：先从候选属性中选出信息增益高于平均水平的，再从中选增益率最高者。这避免了增益率对可取值数目过少的属性的偏好。

基尼指数（CART）

CART 决策树使用基尼指数选择划分属性。数据集 D 的基尼值：

$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2.$$

基尼值反映了从 D 中随机抽取两个样本，其类别标记不一致的概率。基尼值越小，纯度越高。

基尼指数与信息熵的关系：当 $|\mathcal{Y}| = 2$ 时，$\mathrm{Gini}(D) = 2p(1-p)$，$\mathrm{Ent}(D) = -p\log_2 p - (1-p)\log_2(1-p)$。二者在 $p = 0.5$ 处同时达到最大值，在 $p = 0, 1$ 处同时达到最小值，曲线形状高度相似。

属性 a 的基尼指数定义为划分后各子集基尼值的加权和。CART 选择基尼指数最小的属性。

三种算法的对比

算法	划分准则	树结构	任务
ID3	信息增益	多叉树	分类
C4.5	增益率	多叉树	分类
CART	基尼指数 / 平方误差	二叉树	分类/回归

8.5.5 4.4 剪枝

为什么需要剪枝

决策树容易过拟合：如果一直划分直到每个叶结点只包含单个类别的样本，树将完美拟合训练数据但泛化能力很差。剪枝通过削减树的分支来对抗过拟合，其思想与优化中的**正则化**类似——通过降低模型复杂度来提高泛化能力。

预剪枝（Pre-pruning）

在生成过程中，对每个结点在划分前进行估计：若划分不能显著提升验证集上的性能，则停止划分，将当前结点标记为叶结点。

- 优点：减少训练时间，边生成边判断
- 缺点：基于贪心策略，可能阻止后续有价值的划分（**欠拟合风险**）

后剪枝（Post-pruning）

先完整生成一棵决策树，然后自底向上考察每个非叶结点：若将其子树替换为叶结点能在验证集上提升性能，则剪枝。

- 优点：保留更多分支的可能性，泛化性能通常优于预剪枝
- 缺点：训练时间较长（需先生成完整的树再回溯剪枝）

教学提示：对比预剪枝与后剪枝时，可以指出这与优化中的「早停」(early stopping) vs「先充分训练再压缩」的策略类似。

8.6 第五节神经网络

8.6.1 教学目标

学完本节后，学生应能：

1. 描述 M-P 神经元模型的结构
2. 理解感知机的工作原理及其局限性（XOR 问题）
3. 解释多层前馈网络的表示能力和通用逼近定理
4. 推导 BP 算法的误差反向传播公式

5. 认识到局部极小、学习率选择等训练挑战

8.6.2 5.1 M-P 神经元模型

数学模型

神经元接收 n 个输入信号，通过带权重的连接传递，汇总后与阈值比较，经激活函数输出：

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i - \theta\right).$$

引入偏置 $b = -\theta$，得统一形式：$y = \varphi(\mathbf{w}^{\mathsf{T}}\mathbf{x} + b)$。这与线性模型 $f(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x} + b$ 的唯一区别是多了一个非线性激活函数 φ。

常见激活函数

- **阶跃函数**：$\varphi(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$。理想但不连续，不可导。仅用于感知机。
- **Sigmoid**：$\sigma(z) = \frac{1}{1+e^{-z}}$。光滑、可导，输出在 $(0,1)$。导数为 $\sigma'(z) = \sigma(z)(1-\sigma(z))$，计算方便。但存在梯度饱和问题（$|z|$ 大时梯度接近于零）。
- **tanh**：$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$。零中心化（输出均值为 0），导数 $\tanh'(z) = 1 - \tanh^2(z)$。但仍存在梯度饱和。
- **ReLU**：$\varphi(z) = \max(0, z)$。现代深度学习的默认选择，计算简单（$\varphi'(z) = \mathbb{I}(z > 0)$），缓解梯度消失。缺点是「神经元死亡」——若某个神经元输出恒为零，它不再学习。

8.6.3 5.2 感知机

结构与学习规则

感知机是只有输入层和输出层（无隐藏层）的两层网络，使用阶跃激活函数。感知机学习规则（由 Rosenblatt 于 1957 年提出）：

$$\mathbf{w} \leftarrow \mathbf{w} + \eta(y - \hat{y})\mathbf{x}.$$

感知机收敛定理

定理（Novikoff, 1963）：若训练数据线性可分，则感知机学习算法在有限步内收敛于某个分离超平面。具体地，设 $R = \max_i \|\mathbf{x}_i\|$，$\gamma$ 为最大间隔，则误分类次数不超过 $(R/\gamma)^2$。

证明概要：跟踪 $\mathbf{w}^{(k)}$ 与最优超平面 $\mathbf{w}^*$（设 $\|\mathbf{w}^*\| = 1$）的内积变化：

- 每次误分类，$\mathbf{w}^{(k+1)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(k)} \cdot \mathbf{w}^* + \gamma$（向正确方向移动）
- 同时 $\|\mathbf{w}^{(k+1)}\|^2 \leq \|\mathbf{w}^{(k)}\|^2 + R^2$（长度增长有界）
- 由 Cauchy-Schwarz 不等式推出 $\mathbf{w}^{(k)} \cdot \mathbf{w}^* \leq \|\mathbf{w}^{(k)}\|$，得 $k\gamma \leq \|\mathbf{w}^{(0)}\| + kR^2$，推出 $k \leq (R/\gamma)^2$

几何意义与局限性

感知机在特征空间中寻找一个分离超平面 $\mathbf{w}^{\mathsf{T}}\mathbf{x}+b=0$。因此感知机只能解决**线性可分**问题。

XOR 问题的不可解性：异或(XOR)是最简单的非线性可分问题。四个点 $\{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$（前两维为输入）在平面上交叉分布，不存在一条直线分离正负样本。单层感知机无法解决 XOR——这一发现直接导致了神经网络研究的第一次寒冬（Minsky & Papert, 1969, *Perceptrons*）。

教学提示：让学生在黑板上画出 XOR 的四个点，亲自尝试画一条直线分开它们（会失败），然后展示两层网络如何解决 XOR（用两条线组合，形成折线分界）。

8.6.4 5.3 多层前馈网络

结构

多层前馈网络由输入层、若干隐藏层和输出层组成。层间全互连，同层和跨层无连接。信息单向流动（前馈）。

设网络有 L 层（输入层为第 0 层），第 l 层有 s_l 个神经元（$s_0=d$，$s_L=|\mathcal{Y}|$）。前向传播：

$$\mathbf{a}^{(l)}=\varphi^{(l)}\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)}+\mathbf{b}^{(l)}\right),\quad l=1,2,\dots,L,$$

其中 $\mathbf{a}^{(0)}=\mathbf{x}$（输入），$\mathbf{a}^{(L)}=\hat{\mathbf{y}}$（输出），$\mathbf{W}^{(l)}\in\mathbb{R}^{s_l\times s_{l-1}}$，$\mathbf{b}^{(l)}\in\mathbb{R}^{s_l}$。

参数总数：$\sum_{l=1}^{L}(s_{l-1}\times s_l+s_l)$。对于典型的单隐藏层网络（$d$ 输入，q 隐藏，l 输出），参数量为 $(d+1)q+(q+1)l$。

通用逼近定理

定理（Universal Approximation Theorem; Cybenko, 1989; Hornik, 1991）：

令 $\varphi(\cdot)$ 为非常数、有界、单调递增的连续函数。对任意连续函数 f 定义在紧集 $K\subset\mathbb{R}^d$ 上，和任意 $\varepsilon>0$，存在整数 q 和参数 $\{\mathbf{w}_j,b_j,\alpha_j\}_{j=1}^{q}$，使得单隐藏层网络

$$F(\mathbf{x})=\sum_{j=1}^{q}\alpha_j\varphi(\mathbf{w}_j^{\mathsf{T}}\mathbf{x}+b_j)$$

满足 $\sup_{\mathbf{x}\in K}|F(\mathbf{x})-f(\mathbf{x})|<\varepsilon$。

重要说明：

- 定理确认了网络的**表达能力**——理论上，单隐藏层就够了。
- 但它**不保证**学习算法能找到这样的参数（这就是 BP 算法的任务）。
- 实践中，深度（多层）比宽度（单层多神经元）更有效——深度网络可以用指数级更少的参数表示同样的函数。

8.6.5 5.4 误差反向传播算法（BP）

基本思想

BP 算法是训练多层前馈网络的经典方法，核心思想是利用链式法则高效计算梯度，以此指导参数更新。这本质上是第七章中**梯度下降法**在神经网络参数空间中的应用。

详细推导

以均方误差 $E=\frac{1}{2}\sum_{j=1}^{s_L}(\hat{y}_j-y_j)^2$ 为目标函数。定义第 l 层第 j 个神经元的带权输入：

$$z_j^{(l)}=\sum_{i=1}^{s_{l-1}} w_{ji}^{(l)}a_i^{(l-1)}+b_j^{(l)}.$$

输出层误差信号（$\delta^{(L)}$）：

$$\delta_j^{(L)}=\frac{\partial E}{\partial z_j^{(L)}}=\frac{\partial E}{\partial a_j^{(L)}}\cdot\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}=(a_j^{(L)}-y_j)\cdot\varphi'(z_j^{(L)}).$$

隐藏层误差信号的链式传播：

$$\delta_j^{(l)}=\frac{\partial E}{\partial z_j^{(l)}}=\sum_{k=1}^{s_{l+1}}\frac{\partial E}{\partial z_k^{(l+1)}}\cdot\frac{\partial z_k^{(l+1)}}{\partial a_j^{(l)}}\cdot\frac{\partial a_j^{(l)}}{\partial z_j^{(l)}}=\left(\sum_{k=1}^{s_{l+1}} w_{kj}^{(l+1)}\delta_k^{(l+1)}\right)\cdot\varphi'(z_j^{(l)}).$$

参数梯度：

$$\frac{\partial E}{\partial w_{ij}^{(l)}}=\frac{\partial E}{\partial z_j^{(l)}}\cdot\frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}}=\delta_j^{(l)}a_i^{(l-1)},\qquad \frac{\partial E}{\partial b_j^{(l)}}=\delta_j^{(l)}.$$

参数更新（小批量梯度下降）：

$$w_{ij}^{(l)}\leftarrow w_{ij}^{(l)}-\eta\frac{1}{|B|}\sum_{(\mathbf{x},y)\in B}\frac{\partial E_{(\mathbf{x},y)}}{\partial w_{ij}^{(l)}},$$

其中 B 为一个小批量（minibatch），η 为学习率。

一个简单的手算例子

例 8.6.1（单隐藏层 BP 数值示例）**.** 考虑一个极简网络：$d=2$（输入），$q=1$（隐藏神经元，sigmoid 激活），$l=1$（输出，恒等激活）。总参数：$(2\times1+1)+(1\times1+1)=5$ 个。

设初始参数：$w_{11}^{(1)}=0.5$, $w_{12}^{(1)}=-0.3$, $b_1^{(1)}=0.1$（隐藏层）；$w_{11}^{(2)}=0.8$, $b_1^{(2)}=-0.2$（输出层）。

训练样本：$\mathbf{x}=(1,2)$，$y=3$（回归）。

步骤 1 ——前向传播：

$$z_1^{(1)}=0.5\times1+(-0.3)\times2+0.1=0.5-0.6+0.1=0,$$

$$a_1^{(1)}=\sigma(0)=0.5,$$

$$\hat{y}=z_1^{(2)}=0.8\times0.5+(-0.2)=0.4-0.2=0.2.$$

步骤 2 ——输出层误差：

$$\delta_1^{(2)}=(\hat{y}-y)\cdot1=0.2-3=-2.8.$$

步骤 3 ——反向传播（$\sigma'(0) = 0.5 \cdot 0.5 = 0.25$）：

$$\delta_1^{(1)} = w_{11}^{(2)} \cdot \delta_1^{(2)} \cdot \sigma'(0) = 0.8 \times (-2.8) \times 0.25 = -0.56.$$

步骤 4 ——梯度：

$$\frac{\partial E}{\partial w_{11}^{(1)}} = \delta_1^{(1)} \cdot x_1 = -0.56, \quad \frac{\partial E}{\partial w_{12}^{(1)}} = \delta_1^{(1)} \cdot x_2 = -1.12, \quad \frac{\partial E}{\partial b_1^{(1)}} = \delta_1^{(1)} = -0.56,$$

$$\frac{\partial E}{\partial w_{11}^{(2)}} = \delta_1^{(2)} \cdot a_1^{(1)} = -1.4, \quad \frac{\partial E}{\partial b_1^{(2)}} = \delta_1^{(2)} = -2.8.$$

步骤 5 ——更新（设 $\eta = 0.1$）：

$$w_{11}^{(1)} \leftarrow 0.5 - 0.1 \times (-0.56) = 0.556, \quad w_{12}^{(1)} \leftarrow -0.3 - 0.1 \times (-1.12) = -0.188,$$

依此类推更新其余参数。

在这个简单例子中，$\hat{y} = 0.2$ 远小于真实值 $y = 3$，梯度为负，参数向增大 $\hat{y}$ 的方向调整。

算法流程

1. 前向传播：逐层计算输出 $\mathbf{a}^{(l)}$（$l = 1, \ldots, L$）
2. 计算输出层误差 $\boldsymbol{\delta}^{(L)} = \nabla_{\mathbf{a}^{(L)}} E \odot \varphi'(z^{(L)})$
3. 反向传播：逐层计算 $\boldsymbol{\delta}^{(l)} = [(\mathbf{W}^{(l+1)})^{\mathsf{T}} \boldsymbol{\delta}^{(l+1)}] \odot \varphi'(z^{(l)})$
4. 计算所有参数的梯度：$\nabla_{\mathbf{W}^{(l)}} E = \boldsymbol{\delta}^{(l)} (\mathbf{a}^{(l-1)})^{\mathsf{T}}$，$\nabla_{\mathbf{b}^{(l)}} E = \boldsymbol{\delta}^{(l)}$
5. 使用梯度下降（或 SGD、Adam 等变体）更新所有参数

与第七章的联系：BP 算法的核心是计算 $\nabla_{\mathbf{w}} E$，这正是第七章 §7.4 中梯度下降法所需的。BP 通过链式法则高效计算这一梯度（自动微分的原始形式）。

教学提示：建议在板书上完整推导单隐藏层网络的 BP 公式。学生亲手推导一次后会理解得更透彻。可以引导学生将 δ 的计算与第七章中 Lagrange 乘子法的对偶变量 y 做类比——它们都是「沿网络回传的信息」。

8.6.6 5.5 训练技巧与挑战

学习率选择

学习率 η 是关键超参数。太大导致震荡甚至发散；太小导致收敛过慢。实践中常用自适应学习率方法：

- **AdaGrad**：$\eta / \sqrt{\sum (\nabla E)^2}$，自动衰减，适合稀疏梯度
- **RMSprop**：使用指数移动平均代替历史梯度的累积和
- **Adam**：结合动量与 RMSprop，是目前最流行的默认选择

局部极小与鞍点问题

神经网络的目标函数 $E(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)})$ 高度非凸，存在大量局部极小和鞍点。幸运的是：

- 实践中大多数局部极小的质量（目标函数值）接近全局最优
- 鞍点（梯度为零但 Hessian 有正负特征值）在高维空间中远比局部极小普遍
- 随机梯度下降（SGD）的随机噪声有助于逃逸鞍点和差的局部极小
- 动量（momentum）方法：$\mathbf{v} \leftarrow \beta \mathbf{v} + (1-\beta)\nabla E$，$\mathbf{w} \leftarrow \mathbf{w} - \eta \mathbf{v}$，物理模拟帮助越过鞍点

早停（Early Stopping）

在验证集误差不再下降（甚至开始上升）时停止训练，是最简单有效的正则化手段之一。这等价于在参数空间的某个邻域内截断优化过程。

8.6.7 5.6 其他神经网络简介

此外还有多种重要的神经网络结构，本节仅作简要介绍：

- **RBF 网络（Radial Basis Function Network）**：使用径向基函数（如高斯核）作为隐藏层激活函数：$\varphi_j(\mathbf{x}) = \exp(-\beta_j \|\mathbf{x} - \mathbf{c}_j\|^2)$。训练通常分两步：先用无监督方法（如 k-means）确定中心和宽度，再用线性回归训练输出权重。可视为核方法（见第七节 SVM）与神经网络的桥梁。
- **ART 网络（Adaptive Resonance Theory）**：由 Grossberg 提出，旨在解决「稳定性-可塑性困境」（既能学习新知识又不遗忘旧知识）。采用竞争学习和「胜者通吃」机制，属于无监督学习。
- **SOM 网络（Self-Organizing Map，Kohonen 网络）**：将高维输入映射到低维（通常 2D）的神经元网格，同时保持拓扑结构，使得输入空间中相近的点在映射后仍相近。广泛用于高维数据的可视化。
- **循环神经网络（RNN）**：允许神经元之间存在环型连接，状态随时间演化。适合处理序列数据（时间序列、文本等）。训练使用沿时间展开的 BP（Backpropagation Through Time, BPTT），面临梯度消失/爆炸问题。LSTM（长短期记忆网络，Hochreiter & Schmidhuber, 1997）通过门控机制缓解了这一问题。
- **Boltzmann 机**：基于能量的随机神经网络，神经元状态按 Boltzmann 分布采样：$P(s_i = 1) = \sigma(\sum_j w_{ij} s_j - \theta_i)$。受限 Boltzmann 机（RBM）是深度信念网络（DBN）的基本构件，是深度学习复兴前的重要预训练方法。

8.7 第六节深度学习

8.7.1 教学目标

学完本节后，学生应能：

1. 解释深度学习的核心思想——层次化表示学习

2. 分析梯度消失/爆炸问题的成因和多种解决方案
3. 列举缓解过拟合的常用正则化策略并理解其原理

8.7.2 6.1 从神经网络到深度学习

深度学习是具有多个隐藏层的神经网络（通常 $L \geq 3$）。关键在于：增加**深度**（而非仅增加宽度）使网络能学到**层次化的特征表示**：

- 底层（第 1–2 层）：学到简单特征（边缘、纹理、颜色）
- 中层（第 3–5 层）：学到组合特征（形状、部件、图案）
- 高层（第 6+ 层）：学到语义特征（物体类别、场景概念）

为什么深度更好？以函数逼近为例，一个深度为 L、宽度为 w 的 ReLU 网络可以表示分段线性函数的个数随 L 指数增长，而浅层网络仅仅随宽度多项式增长。换言之，深度网络用指数级更少的参数达到了相同的表达能力。

8.7.3 6.2 梯度消失与爆炸

数学分析

考虑 L 层网络的梯度链式法则：

$$\frac{\partial E}{\partial w^{(l)}} = \delta^{(l)} \cdot a^{(l-1)}, \quad \delta^{(l)} = \left(\prod_{k=l+1}^{L} \mathbf{W}^{(k)} \cdot \mathrm{diag}(\varphi'(z^{(k-1)})) \right) \delta^{(L)}.$$

若每层 $\|\mathbf{W}^{(k)} \cdot \varphi'\| < 1$，$\|\delta^{(l)}\|$ 随 l 减小指数衰减（梯度消失）；若 $\|\mathbf{W}^{(k)} \cdot \varphi'\| > 1$，指数爆炸。

具体数值：Sigmoid 的导数最大为 0.25（在 $z = 0$ 处）。假设统一初始化 $w_{ij} \sim \mathcal{N}(0, 1)$，对于 10 层网络，前几层的梯度约为原始梯度的 $(0.25 \times 1)^{10} \approx 10^{-6}$ 倍！

多层次解决方案

1. **激活函数层面**：使用 ReLU 代替 Sigmoid/tanh。ReLU 对 $z > 0$ 的导数为 1，避免了梯度衰减。
2. **网络结构层面**：
   - **残差连接（ResNet, He et al., 2015）**：$\mathbf{a}^{(l+1)} = \varphi(\mathbf{a}^{(l)} + \mathcal{F}(\mathbf{a}^{(l)}; \mathbf{W}^{(l)}))$。梯度可通过恒等映射（$\frac{\partial \mathbf{a}^{(l+1)}}{\partial \mathbf{a}^{(l)}} \approx \mathbf{I}$）直接传递，避免了连乘导致的衰减。这是训练 100+ 层网络的关键技术。
   - 针对 ResNet 的训练可以看作是在上一层的输出上学习一个小的「修正」$\mathcal{F}$（残差），而非全新的表示。
3. **归一化层面**：
   - **批归一化（Batch Normalization, Ioffe & Szegedy, 2015）**：将每层的输入标准化到均值为 0、方差为 1，然后通过可学习的缩放和平移参数恢复网络的表示能力。这稳定了各层输入的分布，允许使用更大的学习率，并有轻微的正则化效果。

- **层归一化 (Layer Normalization)**：在特征维度上归一化，独立于 batch size，广泛用于 Transformer 架构。

4. **初始化层面**：
   - **Xavier/Glorot 初始化**：$w_{ij} \sim \mathcal{N}(0, \frac{2}{n_{\text{in}}+n_{\text{out}}})$，使前向和反向传播时方差保持一致
   - **He/Kaiming 初始化**（专为 ReLU 设计）：$w_{ij} \sim \mathcal{N}(0, \frac{2}{n_{\text{in}}})$

8.7.4 6.3 过拟合的正则化策略

深度网络参数量极大（百万至千亿级别），正则化至关重要。以下是实践中最重要的正则化方法：

- **Dropout (Srivastava et al., 2014)**：训练时以概率 p 随机丢弃（置零）神经元输出。测试时使用完整的网络（输出乘以 $1-p$ 以保持期望一致）。Dropout 可以理解为在训练过程中不断随机采样指数级数量的子网络，测试时对这些子网络进行模型平均。
- **权重衰减 (Weight Decay)**：等价于 ℓ_2 正则化（见第七章 §7.4 中的 Ridge 回归）。参数更新中引入衰减：$\mathbf{w} \leftarrow \mathbf{w} - \eta(\nabla E + \lambda \mathbf{w})$。在 SGD 中，权重衰减等价于 ℓ_2 正则化；在 Adam 等自适应优化器中，二者不同。
- **数据增强 (Data Augmentation)**：对训练数据做保持标签不变的随机变换（图像：随机裁剪、水平翻转、色彩抖动；文本：回译、同义词替换），有效扩充训练集。这是目前最有效的正则化手段之一——增加数据量的效果通常优于更复杂的算法。
- **早停 (Early Stopping)**：监控验证集误差，在误差不再下降时停止训练。
- **标签平滑(Label Smoothing)**:将硬标签(one-hot)替换为软标签:$y^{\text{smooth}} = (1-\varepsilon)y + \varepsilon/K$，其中 K 为类别数。抑制模型对训练标签的过度自信，已被广泛用于现代深度分类模型。
- **迁移学习 (Transfer Learning)**：在大规模通用数据集（如 ImageNet）上预训练模型，再在目标任务的小数据集上微调（fine-tune）最后的几层。这是在标注数据稀缺场景下最实用、最有效的方法。

8.8 第七节支持向量机

8.8.1 教学目标

学完本节后，学生应能：

1. 区分函数间隔与几何间隔
2. 写出硬间隔 SVM 的优化形式并解释最大间隔原理
3. 理解软间隔 SVM 中松弛变量和惩罚参数 C 的作用
4. 通过拉格朗日对偶推导 SVM 的对偶形式
5. 解释核技巧的基本思想并列举常见核函数
6. 将 SVM 与第七章的凸优化和对偶理论联系起来

8.8.2 7.1 间隔与支持向量

基本思想

SVM 的核心思想：寻找一个分离超平面，使得两类训练样本到该超平面的**最小距离最大化**。几何直觉：间隔越大，对新样本的容错能力越强。

函数间隔与几何间隔

函数间隔（依赖于 $\mathbf{w}$ 的尺度）：

$$\hat{\gamma}_i = y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b), \quad \hat{\gamma} = \min_i \hat{\gamma}_i.$$

若成比例缩放 $(\mathbf{w}, b) \to (\kappa\mathbf{w}, \kappa b)$，超平面不变但函数间隔变为 $\kappa\hat{\gamma}$。因此函数间隔不能直接用于优化。

几何间隔（尺度不变）：

$$\gamma_i = \frac{y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b)}{\|\mathbf{w}\|}, \quad \gamma = \min_i \gamma_i.$$

几何间隔是样本点到超平面的**真实欧氏距离**（有符号）。最大间隔原理就是使 γ 最大。

支持向量

使等号成立的样本点——即位于边界超平面 $y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) = 1$ 上的样本——称为**支持向量**。SVM 的解仅由支持向量决定（**稀疏性**），因为对偶变量 $\alpha_i = 0$ 的样本不参与决策函数。

8.8.3 7.2 硬间隔 SVM

问题形式

最大化间隔 $\frac{1}{\|\mathbf{w}\|}$ 等价于最小化 $\frac{1}{2}\|\mathbf{w}\|^2$。因此硬间隔 SVM 的原问题为：

$$\boxed{\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, m. \end{aligned}}$$

这是标准的**凸二次规划**（convex QP）问题（见第七章 §7.5）。目标函数 $\frac{1}{2}\|\mathbf{w}\|^2$ 是严格凸的二次函数，约束为仿射不等式。

注：最大化间隔等价于最小化 $\|\mathbf{w}\|$。为什么一定要除以 $\|\mathbf{w}\|$？因为超平面的方向只与 $\mathbf{w}/\|\mathbf{w}\|$ 有关，$\|\mathbf{w}\|$ 越小，几何间隔越大。

8.8.4 7.3 软间隔 SVM

动机

当数据**线性不可分**（或存在噪声）时，硬间隔 SVM 无可行解。软间隔 SVM 允许部分样本违反间隔约束，引入松弛变量 $\xi_i \geq 0$：

$$\boxed{\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}}$$

惩罚参数 $C > 0$ 控制间隔最大化与误分类容忍度之间的权衡：

- $C \to \infty$：趋于硬间隔 SVM（不允许违反约束）→ 低偏差、高方差
- $C \to 0$：允许较多违反 → 高偏差、低方差（模型更简单）

与正则化的类比：C 的作用类似于 ℓ_2 正则化中的 $1/\lambda$。大的 C 相当于弱正则化。

8.8.5 7.4 对偶问题

拉格朗日对偶推导

这是第七章 §7.5 对偶理论的直接应用。原问题（软间隔）的拉格朗日函数：

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{m}\mu_i\xi_i.$$

由 KKT 条件——鞍点条件：

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \boxed{\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i}, \tag{8.1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m}\alpha_i y_i = 0 \quad \Rightarrow \quad \boxed{\sum_{i=1}^{m}\alpha_i y_i = 0}, \tag{8.2}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \quad \Rightarrow \quad \mu_i = C - \alpha_i. \tag{8.3}$$

代入式 (8.1)–(8.3) 并对 $\mu_i \geq 0$ 化简，得**对偶问题**：

$$\boxed{\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\mathsf{T}\mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m. \end{aligned}}$$

这是一个关于 $\boldsymbol{\alpha}$ 的凸二次规划问题（变量数 m，约束简单）。

KKT 条件与解的结构

根据互补松弛条件：

$$\alpha_i[y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i+b)-1+\xi_i]=0,\quad \mu_i\xi_i=0.$$

由此可将 α_i 分为三类：

- $\alpha_i=0$：样本 $\mathbf{x}_i$ 不在边界上，被正确分类且远离超平面——非支持向量，不参与模型
- $0<\alpha_i<C$：样本 $\mathbf{x}_i$ 恰在间隔边界上（$\xi_i=0$，$y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i+b)=1$）——**支持向量**
- $\alpha_i=C$：样本 $\mathbf{x}_i$ 在间隔内或被误分类（$\xi_i\geq 0$）——也是支持向量，可能被惩罚

决策函数：$f(\mathbf{x})=\sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^\mathsf{T}\mathbf{x}+b$。仅支持向量（$\alpha_i>0$）参与求和。

对偶形式的意义

- 样本仅通过内积 $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ 出现——为核技巧铺路
- 决策函数中的 b 可由任意满足 $0<\alpha_i<C$ 的支持向量确定：$b=y_i-\sum_j \alpha_j y_j \mathbf{x}_j^\mathsf{T}\mathbf{x}_i$
- 对偶问题的优势：变量数从 $d+1+m$ 降为 m，且约束更简单（仅有框约束）

与第七章的联系：SVM 对偶问题的推导是第七章 §7.5 Lagrange 对偶理论的完美应用。Slater 条件自动满足（原问题为凸 QP，不等式约束为仿射），因此强对偶性成立。

8.8.6 7.5 核函数

核技巧

当原始空间中数据**非线性可分**时，通过映射 $\phi:\mathcal{X}\to\mathcal{F}$ 将数据映射到高维特征空间，在 $\mathcal{F}$ 中应用线性 SVM。核技巧的精妙之处在于：**不显式计算** $\phi(\mathbf{x})$，而直接在原始空间中计算高维内积：

$$\kappa(\mathbf{x}_i,\mathbf{x}_j)=\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle_{\mathcal{F}}.$$

常见核函数

- **线性核**：$\kappa(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$。退化为线性 SVM，适用于线性可分问题。
- **多项式核**：$\kappa(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j+c)^d$。参数 $c\geq 0$ 控制低阶项权重，d 为次数。特征空间维数为 $\binom{d+d_{\text{in}}}{d}$。
- **高斯/RBF 核**：$\kappa(\mathbf{x}_i,\mathbf{x}_j)=\exp\!\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right)$。这是最常用的核函数，对应**无限维**特征空间（Taylor 展开包含所有次数）。σ 控制平滑度：σ 越小，模型越灵活（过拟合风险增加）。
- **Sigmoid 核**：$\kappa(\mathbf{x}_i,\mathbf{x}_j)=\tanh(\beta\mathbf{x}_i^\mathsf{T}\mathbf{x}_j+\theta)$。源自神经网络中的激活函数，但仅在特定参数范围内满足 Mercer 条件。

Mercer 定理与核的有效性

Mercer 定理：设 $\mathcal{X}$ 为紧集，$\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 为连续对称函数。κ 可以作为某个 Hilbert 空间中内积的充要条件是，对任意平方可积函数 g：

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \kappa(\mathbf{x}, \mathbf{x}') g(\mathbf{x}) g(\mathbf{x}')\, d\mathbf{x}\, d\mathbf{x}' \geq 0.$$

等价地，对任意有限样本集，其 Gram 矩阵 $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ 为**半正定矩阵**。

SVM 的核心优势总结

1. **凸二次规划**：目标函数严格凸，全局最优解有保证（与第七章联系）
2. **核技巧**：隐式映射到高维（甚至无限维）特征空间，避免维数灾难
3. **稀疏性**：仅支持向量（通常远少于总样本数）参与决策函数
4. **泛化理论**：最大化间隔 $\leftrightarrow$ 最小化 VC 维上界，泛化保证有理论基础
5. **对偶可解**：SMO（Sequential Minimal Optimization）算法每次只优化两个变量，其他固定，解析求解，效率极高

教学提示：SVM 是对本课程所学内容的完美整合——涉及优化（凸二次规划，第七章 §7.5）、对偶理论（Lagrange 对偶、KKT 条件）、数值分析（SMO 算法、求解大规模 QP）。在本节中应有意识地回顾并联系第七章凸优化的内容。可以从第七章的 Lagrange 对偶出发，直接将 SVM 原问题的 KKT 条件写出，引导学生看到理论与应用的桥梁。

8.9 小结

本章介绍了机器学习的基本概念、经典模型和核心算法，从机器学习基础概念出发，依次覆盖模型评估、线性模型、决策树、神经网络、深度学习和支持向量机。

1. **基础概念**：数据集、样本、特征、标记、分类/回归、监督/无监督、假设空间、归纳偏好、奥卡姆剃刀、NFL 定理
2. **模型评估**：训练/测试误差、过拟合/欠拟合、偏差-方差分解、留出法、交叉验证法、自助法
3. **线性模型**：线性回归（最小二乘、闭式解、QR/SVD 联系）、逻辑回归（交叉熵极小化、凸性、极大似然估计）
4. **决策树**：ID3（信息增益）、C4.5（增益率）、CART（基尼指数）、剪枝策略（预剪枝与后剪枝）
5. **神经网络**：M-P 神经元模型、感知机（收敛定理、XOR 局限性）、多层前馈网络（通用逼近定理）、BP 算法（完整链式法则推导）
6. **深度学习**：层次化表示学习、梯度消失/爆炸（ReLU、残差连接、批归一化、权重初始化）、正则化策略（Dropout、数据增强、权重衰减、迁移学习）
7. **支持向量机**：最大间隔原理、硬/软间隔 SVM、Lagrange 对偶推导、KKT 条件与解结构、核技巧（线性/多项式/RBF/Sigmoid 核）、Mercer 定理

本章与第七章凸优化的联系构成了本课程「优化 + 学习」方法论的完整体系：

- 线性回归的最小二乘估计 → 无约束凸二次规划（第七章 §7.4）
- 逻辑回归的交叉熵极小化 → 无约束光滑凸优化（第七章 §7.4 Newton 法）
- 感知机和 SGD → 随机梯度下降法（第七章 §7.4）
- BP 算法 → 多元函数链式求导 + 梯度下降（第七章 §7.4）
- SVM 原问题 → 约束凸二次规划（第七章 §7.5）
- SVM 对偶问题 → Lagrange 对偶 + KKT 条件（第七章 §7.5）
- ℓ_1/ℓ_2 正则化 → Ridge/Lasso 的优化视角（第七章 §7.6）
- Dropout/早停 → 隐式正则化（第七章 §7.6）

建议在教学安排上将第七章（凸优化）与第八章（机器学习）作为连贯的教学单元，让数值分析、优化理论和机器学习三者相互印证、融会贯通。