

上海交通大学 · 数学科学学院

数值分析

NUMERICAL ANALYSIS

——— **第零章 · 简介** ———

马征

按空格键继续 →

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

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

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

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	<i>Convex Optimization</i> , 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	<i>Scientific Computing with Python</i> (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！

上海交通大学 · 数学科学学院

数值分析

NUMERICAL ANALYSIS

第一章 · 多项式插值

马征

按空格键继续 →

第一节 多项式插值

Polynomial Interpolation

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

插值方法

◆ 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

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

- $P(x)$ 称为 $f(x)$ 的**插值函数**， $x_0, x_1, \dots, x_n$ 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

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

⚠ 思考

还可以采用哪些类型的函数（模型）？

多项式插值的存在唯一性

► 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

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

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。

Lagrange 插值

◆ 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:

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

Lagrange 插值基函数的性质

► 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 $i, j = 0, 1, \dots, n$ 满足

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

Lagrange 插值示例（一）

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 ($n = 2$):

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

Lagrange 插值示例（二）

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 $\ln(0.6)$ 。

Lagrange 插值余项

► 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 $\omega_{n+1}(x) = \prod_{i=0}^n (x - x_i)$ 。

Lagrange 插值余项上界

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

Lagrange 插值余项示例

例 1.3

1 次和 2 次 Lagrange 插值的余项分别为

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

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

Lagrange 插值逼近误差

► 定理 1.4

对复函数 $f(z)$, 如果存在 $r_0 > \frac{3}{2}(b-a)$, 使得 $f(z)$ 在 $B_{r_0}\left(\frac{a+b}{2}\right)$ 内解析, 则 $L_n(x)$ 在 $[a, b]$ 内一致收敛到 $f(x)$ 的 n 次 Lagrange 插值多项式满足逼近误差估计. 这里 $B_{r_0}\left(\frac{a+b}{2}\right)$ 表示以 $\frac{a+b}{2}$ 为中心, 半径为 r_0 的圆.

Lagrange 插值性质

例 1.4

证明

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

Lagrange 插值综合示例

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。

第二节 牛顿插值

Newton Interpolation

差商

◆ 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 $x_0, x_1, \dots, x_k$ 的 **k 阶差商**递归定义为

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

差商的对称性

► 定理 1.5

差商关于节点具有对称性：

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

差商的线性组合表示

► 定理 1.6

k 阶差商有如下等价表达式：

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

差商与导数的关系

► 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

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

差商的多项式性质

► 定理 1.8

若 $f[x, x_0, x_1, \dots, x_k]$ 是关于 x 的 m 次多项式, 则 $f[x, x_0, x_1, \dots, x_k, x_{k+1}]$ 是关于 x 的 $m - 1$ 次多项式。

↳ 推论

若 f 是 n 次多项式, 则 $f[x, x_0, x_1, \dots, x_n] \equiv 0$ 。

牛顿插值多项式

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

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

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。

牛顿插值与 Lagrange 插值的比较

▶ 等价性

牛顿插值与 Lagrange 插值是同一插值多项式的不同表达形式。

⚠ 效率优势

牛顿插值比 Lagrange 插值更高效——增加新节点时只需追加一项，无需重新计算，可用于在线（递增节点）插值。

⚠ 光滑性要求

牛顿插值通过差商定义，对目标函数光滑性要求较低。

牛顿插值示例

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 $f(1.566)$ 。

差分

◆ 前向差分 / 后向差分

$$\text{前向差分: } \Delta f_k = f_{k+1} - f_k, \quad \Delta^m f_k = \Delta^{m-1} f_{k+1} - \Delta^{m-1} f_k$$

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

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 4.9

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

等距节点牛顿前插公式

令 $x = x_0 + th$, $h > 0$, $0 \leq t \leq 1$, 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

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

等距节点牛顿后插公式

令 $x = x_n + th$, $-1 \leq t \leq 0$, $h > 0$, 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

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

等距插值示例

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

第三节 Hermite 插值

Hermite Interpolation

Hermite 插值

◆ HERMITE 插值多项式

设 $a \leq x_0 < x_1 < \cdots < x_n \leq b$, 寻找多项式 $H_{2n+1} \in \mathcal{P}_{2n+1}$ 满足

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

称 H_{2n+1} 为数据 $\{(x_j, f(x_j), f'(x_j))\}_{j=0}^n$ 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。

Hermite 插值基函数 (一)

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

► 定理

基函数基于 Lagrange 基, 满足:

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

Hermite 插值基函数（二）

基函数的具体形式为：

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

其中 $l_j(x)$ 为 Lagrange 插值基函数。

Hermite 插值余项

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

Hermite 插值计算示例

例 4.8

计算 3 次 Hermite 插值多项式 $H_3(x)$ 及其余项。

三次 Hermite 插值示例

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 $\ln(1.5)$ 。

Hermite 插值误差

► 定理 4.12

三次 Hermite 插值满足误差估计

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

第四节 分段多项式插值

Piece-wise Polynomial Interpolation

Runge 现象

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

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

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

Demo

Runge 现象

⚠ 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

分段多项式插值

◆ 分段多项式

设 $[a, b]$ 的剖分为 $a = x_0 < x_1 < \cdots < x_n = b$, 若函数 $\varphi_h(x)$ 在每个子区间 $[x_k, x_{k+1}]$ ($k = 0, \dots, n-1$) 上均为多项式, 则称 $\varphi_h(x)$ 为 $[a, b]$ 上的**分段多项式**。

分段线性插值

◆ 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：

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

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**:

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

连续模与收敛性

◆ 连续模

函数 $f(x)$ 在 $[a, b]$ 上的连续模 $\omega(h) = \omega(f, h)$ 满足：对任意 $x_1, x_2 \in [a, b]$ ，若 $|x_1 - x_2| \leq h$ ，则 $|f(x_1) - f(x_2)| \leq \omega(h)$ 。

► 命题 4.1

对 $f(x) \in C[a, b]$ ， $\lim_{h \rightarrow 0} \omega(h) = 0$ 。

分段线性插值误差

► 定理 4.13

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

► 定理 4.14

设 $f(x) \in C^2[a, b]$, 则

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

分段三次 Hermite 插值

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

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

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

分段三次 Hermite 插值误差

► 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

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

三次样条函数

◆ 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. $S(x_j) = f(x_j)$, $j = 0, 1, \dots, n$;
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。

三次样条的边界条件

◆ 第一类边界条件（匹配一阶导数）

$$S'(x_0) = f'(x_0), \quad S'(x_n) = f'(x_n).$$

◆ 第二类边界条件（匹配二阶导数）

$$S''(x_0) = f''(x_0), \quad S''(x_n) = f''(x_n).$$

◆ 第三类边界条件（周期条件）

$$S(x_0) = S(x_n), \quad S'(x_0) = S'(x_n), \quad S''(x_0) = S''(x_n).$$

三次样条的计算（一阶导数条件）

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

$$\begin{bmatrix} 2 & \mu_1 & & \\ \lambda_2 & 2 & \mu_2 & \\ & \ddots & \ddots & \ddots \\ & & \lambda_{n-1} & 2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} g_1 - \lambda_1 f'_0 \\ g_2 \\ \vdots \\ g_{n-1} - \mu_{n-1} f'_n \end{bmatrix}$$

三次样条的计算（二阶导数条件）

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

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

三次样条的收敛性与误差

► 定理 4.14 (收敛性)

$S(x)$ 在 $[a, b]$ 上一致收敛到 $f(x)$ 。

► 定理 4.15 (误差估计)

对一阶导数匹配边界条件及 $f \in C^4[a, b]$,

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

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）

上海交通大学 · 数学科学学院

数值分析

NUMERICAL ANALYSIS

第一章 · 第四节

马征

按空格键继续 →

第四节 函数逼近

Function Approximation

函数逼近的误差度量

◆ 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

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

常用情形：

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

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

最佳逼近

◆ 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

Weierstrass 定理

► 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。

⚠ 说明

常见证明思路使用 Bernstein 多项式。

Bernstein 多项式

◆ 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

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

最佳一致逼近 (minimax)

◆ 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。

⚠ 几何理解

交错定理 (Chebyshev Alternation)

▶ 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

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

使误差函数 $e(x) = f(x) - P_n(x)$ 在这些点上满足

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

即：最大误差至少在 $n+2$ 个点达到，且符号交替。

交错定理证明思路

⚠ 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

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

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

交错定理证明思路

⚠ 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 $\deg r \leq n$ ，除非 $r \equiv 0$ ，从而不可能得到更小的 $\|f - Q_n\|_\infty$ 。

唯一性与计算

► 存在唯一性

对任意 $f \in C[a, b]$ 与固定 n ，最佳一致逼近多项式在 Π_n 中存在且唯一。

⚠ 计算方法

实际求解常用 Remez 交换算法：迭代更新交错点并解等振荡方程组。

⚠ 与 CHEBYSHEV 多项式的关系

在标准区间 $[-1, 1]$ 上，Chebyshev 多项式的等振荡性质是构造近似最优节点与近似最优多项式的核心。

例题： x^2 的一次最佳一致逼近

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

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

并有

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

内积空间

◆ 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. $\langle u, u \rangle \geq 0$ ，且当且仅当 $u = 0$ 时取等号

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

内积空间性质 I

► 定理 4.11 (CAUCHY-SCHWARZ 不等式)

对任意 $u, v \in X$, 有

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

内积空间性质 II

► 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。

权函数与加权 L_2 空间

◆ 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 $\int_a^b x^k \rho(x) dx$ 存在且有限
2. 对任意非负连续函数 g , 若 $\int_a^b g(x) \rho(x) dx = 0$, 则 $g \equiv 0$

则称 ρ 为权函数。

由此可定义

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

线性无关与 Gram 行列式

► 定理 4.14

设 $\varphi_1, \dots, \varphi_n \in C[a, b]$ 。它们线性无关当且仅当

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

正交多项式

◆ 正交性

设 $g_0, g_1, \dots, g_m$ 为多项式, 若在权函数 ρ 下

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

则称它们关于 ρ 正交。

正交多项式三项递推

► 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

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

正交多项式零点性质

► 定理 4.15

对 $n \geq 1$, $\varphi_n(x)$ 在 (a, b) 上有 n 个互异零点。

Legendre 多项式定义

◆ LEGENDRE 多项式

在区间 $[-1, 1]$ 上定义

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

其中 $P_0(x) = 1$ 。

Legendre 多项式性质 I

► 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

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

Legendre 多项式性质 II

► 性质 4.2 (正交性)

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

► 性质 4.3 (奇偶性)

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

Legendre 多项式性质 III

► 性质 4.4 (递推)

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

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

Legendre 多项式结论与例题

► 定理 4.16

P_n 在 $(-1, 1)$ 上有 n 个互异零点。

⚠ 最优化性质

在首项系数固定为 1 的 n 次多项式中, 适当归一化的 Legendre 多项式对零函数具有最小平方误差。

🔑 例 4.8

写出次数不超过 6 的 Legendre 多项式。

Chebyshev 多项式定义

◆ 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

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

Chebyshev 多项式性质 I

► 性质 4.5 (正交性)

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

► 性质 4.6 (递推)

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

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

Chebyshev 多项式性质 II

► 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} 。

► 性质 4.8 (奇偶性)

T_{2k} 仅含偶次幂, T_{2k+1} 仅含奇次幂。

Chebyshev 多项式性质 III

► 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 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。

Chebyshev 零点插值误差

► 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

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

🔗 例 4.11

对 $f(x) = \frac{1}{1+x^2}$ 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。

最小二乘逼近

◆ 最小二乘逼近

设子空间

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

寻找 $S^* \in \Phi$ 使

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

⚠ 等价正交条件

误差函数 $e = f - S^*$ 与子空间 Φ 正交: $\langle e, \varphi_k \rangle = 0$ 。

正规方程

► 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

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

正交基下的最小二乘系数

► 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

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

例 4.12

例 4.12

计算 $f(x) = \sqrt{1+x^2}$ 在 $[0, 1]$ 上的最小二乘逼近多项式。

例 4.13

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。

第五节 三角逼近与傅里叶方法

Trigonometric Approximation and Fourier Methods

三角多项式与三角逼近

◆ 三角多项式空间

以周期区间 $[-\pi, \pi]$ 为例，定义

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

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。

Fourier 部分和与最佳平方逼近

► L_2 最优性

在内积

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

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

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

其中

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

三角插值与三角逼近的联系

◆ 等距节点三角插值

取 $N = 2n + 1$ 个节点

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

求 $I_n \in \mathcal{T}_n$ 使

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

⚠ 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。

从三角插值到 DFT

► 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

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

逆变换

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

⚠ 解释

DFT 本质是把离散信号投影到离散三角基（复指数基）上。

DFT 与三角插值系数

▶ 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

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

故三角插值系数可由 DFT 一次性计算。

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

第六节 快速傅立叶变换

Fast Fourier Transform

FFT 的核心思想

► 偶奇分解 (\$N=2^M\$)

将 DFT

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

按偶、奇下标拆分：

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

即把一个 N 点 DFT 化为两个 $N/2$ 点 DFT，再做蝶形合并。

复杂度与算法结构

► 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

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

⚠ 实现要点

1. 位逆序重排 (bit-reversal permutation)
2. 分层蝶形迭代 (stage-by-stage butterfly)

FFT 计算流程（迭代版）

✎ 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:

$$\begin{aligned} u &= f[p], & v &= \omega_L^q f[p + L/2], \\ f[p] &= u + v, & f[p + L/2] &= u - v. \end{aligned}$$

4. 输出即为 $\hat{f}$ 。

FFT 的数值计算应用

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 $\hat{a} = \text{FFT}(a)$, $\hat{b} = \text{FFT}(b)$
3. 频域逐点乘积 $\hat{c}_k = \hat{a}_k \hat{b}_k$
4. 逆 FFT 得卷积结果 c 。

总复杂度由 $O(nm)$ 降为 $O(N \log N)$ 。

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 $O(N \log N)$

上海交通大学 · 数学科学学院

数值分析

NUMERICAL ANALYSIS

第五章 · 数值积分

马征

按空格键继续 →

第一节 数值积分

Numerical Integration

数值积分

◆ 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 $w_j \in \mathbb{R}$ ($j = 0, 1, \dots, n$) 的有限求和来近似积分：

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

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

近似误差与代数精度

◆ 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 $m = 0, 1, \dots, k$ 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。

代数精度示例

例 5.1

求下列数值积分的代数精度：

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

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

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

其中权重

$$w_j = \int_a^b l_j(x) \, dx.$$

称此类方法为**插值型数值积分**。

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

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

其中 $\xi \in (a, b)$ 依赖于 x ，且

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

插值型数值积分的代数精度

► 定理 5.1

数值积分至少具有 n 阶代数精度，当且仅当它是插值型的。

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

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

Newton-Cotes 公式的特殊情形

- $n = 1$ (梯形公式)

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

- $n = 2$ (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)

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

Newton-Cotes 公式的误差 (n 为偶数)

► 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 $M_n = \int_a^b x \omega_{n+1}(x) dx < 0$ 。

此时 Newton-Cotes 公式至少具有 $(n+1)$ 阶代数精度。

Newton-Cotes 公式的误差 (n 为奇数)

► 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 $M'_n = \int_a^b \omega_{n+1}(x) dx < 0$ 。

Newton-Cotes 公式示例

例 5.2

分别用 $n = 1, 2, 3, 4$ 的 Newton-Cotes 公式计算积分

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

第二节 复合求积公式

Composite Quadrature Rule

复合求积公式

将区间 $[a, b]$ 均匀划分为若干子区间 $[a_k, b_k]$ ，在每个子区间上分别应用求积公式，再求和，即得**复合求积公式**。

⚠ 动机

高阶 Newton-Cotes 公式的 Cotes 系数可能出现负值，导致数值不稳定。复合低阶公式更稳定、精度更高。

复合梯形公式

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

◆ 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

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

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。

◆ 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

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

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。

◆ 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

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

复合求积公式示例

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

$$\int_0^1 \frac{\sin x}{x} dx.$$

第三节 Romberg 方法与 Richardson 外推

Romberg Rule and Richardson Extrapolation

Romberg 方法

Romberg 求积对梯形公式作递推加速，递归定义为

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

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

► 定理 5.4

Romberg 公式的误差满足

$$(T_{j+1} f)(h) - \int_a^b f(x) \, dx = \mathcal{O}\left(h^{2(j+1)}\right).$$

梯形公式的二分递推

梯形公式满足如下二分递推公式：

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

⚠ 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。

Richardson 外推

► 定理 5.5

梯形公式具有如下渐近展开：

$$T(h) = I + a_1 h^2 + a_2 h^4 + a_3 h^6 + \cdots + a_k h^{2k} + \cdots$$

由此得到误差阶递增的组合：

$$T_1(h) := \frac{4}{3}T\left(\frac{h}{2}\right) - \frac{1}{3}T(h) = I + \beta_1 h^4 + \beta_2 h^6 + \cdots,$$

$$T_2(h) := \frac{16}{15}T_1\left(\frac{h}{2}\right) - \frac{1}{15}T_1(h) = I + \gamma_1 h^6 + \gamma_2 h^8 + \cdots$$

Richardson 外推的一般公式

一般地, 对 $m = 1, 2, \dots$:

$$T_m(h) := \frac{4^m}{4^m - 1} T_{m-1}\left(\frac{h}{2}\right) - \frac{1}{4^m - 1} T_{m-1}(h) = I + \delta_1 h^{2(m+1)} + \delta_2 h^{2(m+2)} + \dots$$

⚠ 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。

Romberg 算法

设 $T_0^{(k)}$ 为步长二分 k 次后的梯形公式值。

1. 计算 $T_0^{(0)} = \frac{b-a}{2} [f(a) + f(b)]$ 。

2. 利用二分递推公式计算 $T_0^{(k)}$ 。

3. 递归计算外推值：

$$T_m^{(k)} = \frac{4^m}{4^m - 1} T_{m-1}^{(k+1)} - \frac{1}{4^m - 1} T_{m-1}^{(k)}, \quad k = 1, 2, \dots$$

4. 若 $|T_k^{(0)} - T_{k-1}^{(0)}| < \varepsilon$, 输出 $T_k^{(0)}$; 否则令 $k+1 \rightarrow k$, 返回步骤 2。

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 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：

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

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

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：

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

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： $B_{2k+1} = 0$ 。

Euler-Maclaurin 公式的证明 (2)

续前

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:

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(x) dx &= h \int_0^1 f(x_i + th) B_0(t) dt = h \int_0^1 f(x_i + th) dB_1(t) \\ &= h [f(x_{i+1}) B_1(1) - f(x_i) B_1(0)] - h^2 \int_0^1 f'(x_i + th) B_1(t) dt \\ &= \frac{h}{2} [f(x_{i+1}) + f(x_i)] - h^2 \int_0^1 f'(x_i + th) B_1(t) dt.\end{aligned}$$

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。

Euler-Maclaurin 公式的证明 (3)

续前

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

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

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。■

第四节 Gauss 求积公式

Gauss Quadrature Rule

Gauss 节点

考虑以 $n + 1$ 个节点的插值型数值积分

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

◆ 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**。

Gauss 节点的特征

► 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

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

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：

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

定理 5.5 的证明 (充分性)

续前

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:

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

因此节点为 Gauss 节点。■

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：

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

◆ 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 ($j = 0, \dots, n$) 为关于权函数 $\rho(x)$ 的 **Gauss 节点**。

带权 Gauss 节点的确定

► 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 ($j = 0, 1, \dots, n$) 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:

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

其中 a_n 为 $\varphi_n(x)$ 的首项系数。

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。

定理 5.6 的证明 (2)

续前

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。

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 $\int_{-1}^1 f(x) dx$ 的 Gauss 节点, 对应的求积公式称为 **Gauss-Legendre 公式**。
- 当 $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ 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 $\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx$ 的 Gauss 节点, 对应公式称为 **Gauss-Chebyshev 公式**。

Gauss-Legendre 公式示例

例 5.4

求两节点 Gauss-Legendre 公式的节点和权重：

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

带权 Gauss 公式示例

例 5.5

构造如下带权积分的两节点 Gauss 公式：

$$\int_0^1 \frac{f(x)}{\sqrt{x}} dx \approx w_0 f(x_0) + w_1 f(x_1).$$

Gauss 公式的稳定性

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

► 定理 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

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

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：

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

定理 5.7 的证明 (2)

续前

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

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

结论得证。■

Gauss 公式的收敛性

► 定理 5.8 (收敛性)

设 $f(x)$ 在 $[a, b]$ 上连续, 则 Gauss 求积公式收敛:

$$\lim_{n \rightarrow \infty} Q_n(f) = \int_a^b f(x) \rho(x) dx.$$

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 $2n + 1 \geq \deg(P)$ 。利用公式对 $P(x)$ 精确、权重 $w_j > 0$ 的性质可估算积分误差:

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

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

上海交通大学 · 数学科学学院

蒙特卡洛方法

MONTE CARLO METHODS

数值积分与随机模拟

马征

按空格键继续 →

第一节 维度灾难与引入

The Curse of Dimensionality

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 $E = O(N^{-r})$ 。

⚠ 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)**。

⚠ 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。

蒙特卡洛方法的引入

◆ 蒙特卡洛估计量

将积分视作期望：

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

独立抽取

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

构造 MC 估计量：

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

无偏性与方差

► 基本统计性质

设

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

则

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

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., $\mathbb{E}[Y_i] = I$, $\text{Var}(Y_i) = |\Omega|^2\sigma^2$, $\hat{I}_M = \bar{Y}_M$ 。

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):

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

收敛性：大数定律

► 强大数定律 (SLLN)

若 $f \in L^1(\Omega)$ (即 $\mathbb{E}[|f(X)|] < \infty$), 则

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

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 $X_1, \dots, X_M \stackrel{\text{i.i.d.}}{\sim} U(\Omega)$, 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 $f \in L^1(\Omega)$, $\mathbb{E}[|Y_1|] = |\Omega| \mathbb{E}[|f(X)|] = \int_{\Omega} |f| \, d\mathbf{x} < \infty$. ✓
3. **期望：** $\mathbb{E}[Y_1] = |\Omega| \mathbb{E}[f(X)] = \int_{\Omega} f \, d\mathbf{x} = I$. ✓

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

收敛性：中心极限定理

► 中心极限定理 (CLT)

若 $f \in L^2(\Omega)$, $\sigma^2 = \text{Var}(f(X)) < \infty$, 则

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

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 $Y_1, \dots, Y_M$ i.i.d., 且

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

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

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

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。□

半阶收敛率

↳ 半阶收敛率

MC 估计的均方根误差 (RMSE) 满足

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

与维度 d 完全无关，从根本上克服维度灾难。

半阶收敛率：与传统方法的对比

⚠ 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？

第二节 减少方差的方法

Variance Reduction Techniques

为什么需要减少方差？

⚠ 收敛速率的代价

由中心极限定理，MC 的误差（以 RMSE 度量）为

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

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。

常用减少方差方法

重要性抽样

Importance Sampling

选取与被积函数形状相似的抽样分布，集中采样于高贡献区域

控制变量法

Control Variates

引入已知期望的辅助函数，通过相关性抵消估计误差

分层抽样

Stratified Sampling

将积分域划分为子层分别采样，消除层间波动对方差的贡献

对偶变量法

Antithetic Variates

构造负相关配对样本，通过协方差抵消降低整体方差

重要性抽样

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

重要性抽样：最优密度

最优密度推导

目标。 最小化 $\text{Var}_g(f/g) = \int \frac{f^2(x)}{g(x)} dx - I^2$, 即最小化泛函 $J[g] = \int \frac{f^2}{g} dx$ 。

变分推导。 引入 Lagrange 乘子 λ 处理约束 $\int g dx = 1$, 对泛函取变分:

$$\frac{\delta}{\delta g(x)} \left[\frac{f^2(x)}{g(x)} - \lambda g(x) \right] = 0 \implies -\frac{f^2(x)}{g^2(x)} - \lambda = 0 \implies g(x) \propto |f(x)|.$$

结论。 归一化后得**最优重要性密度** $g^*(x) = \frac{|f(x)|}{\int |f(x)| dx}$ 。当 $f \geq 0$ 时, $f(x)/g^*(x) = \text{const}$, $\text{Var}_{g^*}(f/g^*) = 0$, 方差完全消除。□

实践中 g^* 依赖于未知的 I , 需选取形状与 $|f|$ 相似的可采样分布 g 近似代替。

控制变量法

◆ 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 $\text{Var}(Y_i)$ 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。

控制变量法：最优系数

⚠ 最优系数推导

展开方差：

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

对 λ 求导置零，得最优系数与对应方差：

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

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

分层抽样 (Stratified Sampling)

◆ 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

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

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

分层抽样：方差分析

► 分层抽样方差优越性（比例分配）

令 $\sigma_k^2 = \text{Var}(f(X)|X \in \Omega_k)$, $M_k = p_k M$ （比例分配），则

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

分层抽样：方差分析（证明）

证明

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

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

$$\begin{aligned} \sigma^2 = \text{Var}(f) &= \underbrace{\mathbb{E}[\text{Var}(f|K)]}_{= \sum_k p_k \sigma_k^2} + \underbrace{\text{Var}(\mathbb{E}[f|K])}_{= \sum_k p_k (\mu_k - \bar{\mu})^2} \geq 0 \end{aligned}$$

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

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 $\text{Var}(\hat{I}_{\text{strat}})$ 。

对偶变量法 (Antithetic Variates)

◆ 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):

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

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

对偶变量法：方差减少的证明

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

证明

设 $U, V \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 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 :

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

对期望展开 (注意 $1-V \sim U(0, 1)$ 与 U 独立):

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

即 $\text{Cov}(f(U), f(1-U)) \leq 0$, 故 $\text{Var}(\hat{I}_{\text{AV}}) \leq \frac{\sigma^2}{2M} < \frac{\sigma^2}{M}$. $\square$

第三节 随机数生成

Random Number Generation

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。

◆ 线性同余发生器 (LCG)

通过迭代公式生成 $\{0, 1, \dots, m-1\}$ 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。

⚠ 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。

逆变换定理

► 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 $\mathbb{P}(X \leq x) = F(x)$ 。

逆变换法（证明）

证明

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

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

第二步: F 连续且严格单调, 故

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

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

$$\mathbb{P}(U \leq F(x)) = F(x). \quad \square$$

⚠ 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)。

舍选法 (Rejection Sampling)

◆ 舍选法原理

欲从密度 $f(x)$ 采样。寻找包络密度 $g(x)$ 及常数 $c \geq 1$, 使

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

算法:

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

⚠ 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。

舍选法：正确性证明

被接受样本服从密度 $f(x)$

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

接受概率:

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

联合概率:

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

条件 CDF:

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

第四节 拓展与应用

Extensions and Applications

拟蒙特卡洛 (QMC)

◆ 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

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

► KOKSMA-HLAWKA 不等式

设 f 在 $[0, 1]^d$ 上的 Hardy-Krause 变差为 $V(f) < \infty$, 则

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

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 $O(M^{-1/2})$ 。

马尔可夫链蒙特卡洛 (MCMC)

◆ 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 $\alpha(x, y) = \min\left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right)$ 接受 y ，否则保持 x

细致平衡（正确性）

细致平衡

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 $x \neq y$ ：

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

第二步：推导平稳性。 对细致平衡条件关于 x 积分：

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

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。

蒙特卡洛的现实应用

金融工程：欧式 / 美式期权定价，风险价值 (VaR) 计算，高维随机利率模型

统计物理：大分子配分函数与热力学量，Ising 模型临界现象模拟

机器学习：强化学习策略评估，贝叶斯深度学习后验积分，变分推断

计算机图形学：路径追踪 (Path Tracing) 通过随机光线求解全局光照渲染方程

小结

MC 积分误差为 $O(M^{-1/2})$, 与维度无关, 从根本上克服维度灾难

减少方差: 重要性抽样、控制变量法、分层抽样 (Neyman 分配)、对偶变量法 (负相关配对)

随机数生成: LCG / Mersenne Twister 生成均匀伪随机数; 逆变换法与舍选法转换为目标分布

QMC 通过低差异序列提升光滑函数的收敛阶; MCMC 将方法推广到无法直接采样的高维分布

上海交通大学 · 数学科学学院

常微分方程数值解

NUMERICAL METHODS FOR ODES

第六章 · 常微分方程数值解

马征

按空格键继续 →

第一节 Euler 方法

Euler Methods

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系列节点 $x_0 < x_1 < x_2 < \cdots$ 处求近似值 $y_n \approx y(x_n)$ 。取等距步长 $h > 0$ ：

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

⚠ 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。

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**：

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

⚠ 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。

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：

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

此为关于 y_{n+1} 的隐式方程，一般需迭代求解。

⚠ 与显式格式的对比

	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

Euler 方法的局部截断误差

► 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。

梯形方法

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

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

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 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：

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

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

定理 6.2 的证明

证明

设 y_{n+1} 为真实隐式解，则

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

两式作差：

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

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

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

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

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：

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

⚠ 精度

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

两步 Euler 方法

利用二阶中心差商，得到**两步 Euler 方法**：

$$\bar{y}_{n+1} = y_{n-1} + 2h f(x_n, y_n),$$

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

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 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**。

第二节 Runge-Kutta 方法

Runge-Kutta Methods

p 阶精度的定义

◆ 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

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 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。

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

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

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:

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

△ 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。

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**：

$$y_{n+1} = y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

$$K_1 = f(x_n, y_n),$$

$$K_2 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_1\right),$$

$$K_3 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_2\right),$$

$$K_4 = f(x_n + h, y_n + hK_3).$$

► 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**。

Runge-Kutta 方法示例

例 6.2

用 $h = 0.5$ 的显式 Euler 和 四阶 RK 方法求解：

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

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}

第三节 收敛性与稳定性

Convergence and Stability

收敛性与单步方法

◆ 收敛性

若对固定的 $x_n = x_0 + nh$, 当 $h \rightarrow 0$ ($n \rightarrow \infty$, nh 有界) 时有

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

则称该数值方法**收敛**。

◆ 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 $\varphi = \frac{1}{4}(K_1 + 2K_2 + 2K_3 + K_4)$ 。

单步方法的收敛定理

► 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：

$$|\varphi(x, y, h) - \varphi(x, y', h)| \leq C_\varphi |y - y'|,$$

则该方法收敛，且整体误差（global error）满足：

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

证明思路

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

稳定性

◆ 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

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

(C 与 h, m 无关)，则称该方法**稳定**。

⚠ 稳定性 VS 收敛性

- **收敛性**描述 $h \rightarrow 0$ 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。

显式 Euler 稳定性分析

以模型方程 $y' = \lambda y$ ($\lambda \in \mathbb{C}$, $\operatorname{Re}(\lambda) < 0$) 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：

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

⚠ 实数情形

当 $\lambda < 0$ 为实数时，稳定条件化为 $-2 \leq \lambda h \leq 0$ ，即步长须满足

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

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 $\left|\frac{1}{1 - h\lambda}\right| \leq 1$ ，即 $|1 - h\lambda| \geq 1$ 。

当 $\text{Re}(\lambda) < 0$ 时，对任意 $h > 0$ 均有 $|1 - h\lambda| > 1$ ，故后退 Euler 方法无条件稳定。

⚠ 例 6.3 小结

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

稳定性示例

例 6.4

用显式和后退 Euler 方法（步长 $h = 0.05$ ）求解：

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

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 $\frac{1}{|1 - (-5)|} = \frac{1}{6} < 1$ ，格式**稳定**，数值解单调趋零。

⚠ 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。

第四节 线性多步方法

Linear Multi-step Methods

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**。

⚠ 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。

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**：

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

⚠ 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。

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**：

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

避免差分计算的等价形式：

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

⚠ 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**。

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：

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

直接系数形式：

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

⚠ 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。

Taylor 展开法与系数确定

利用 Taylor 展开匹配，对格式

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

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

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

⚠ 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。

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**:

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

► 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**。

Hamming 方法与预测-校正系统

Hamming 格式（隐式，4 阶）：

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

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

⚠ 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_{n+1}^P ；
2. E（估计）：计算 $f(x_{n+1}, y_{n+1}^P)$ ；
3. C（校正）：Hamming 隐式格式用 y_{n+1}^P 计算 y_{n+1}^C ；
4. E（估计）：计算 $f(x_{n+1}, y_{n+1}^C)$ 供下一步使用。

上海交通大学 · 数学科学学院

凸优化

CONVEX OPTIMIZATION

第七章 · 凸优化

马征

按空格键继续 →

第一节 凸集

Convex Sets

优化问题的基本形式

◆ 优化问题的标准形式

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

其中 $x \in \mathbb{R}^n$ 为**优化变量**， $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ 为**目标函数**， $f_i(x) \leq 0$ 为**不等式约束**， $h_i(x) = 0$ 为**等式约束**。

定义可行集 $\mathcal{F} = \{x \mid f_i(x) \leq 0, h_i(x) = 0\}$ 。**最优值** $p^* = \inf\{f_0(x) \mid x \in \mathcal{F}\}$ 。若存在 $x^* \in \mathcal{F}$ 使 $f_0(x^*) = p^*$ ，则称 x^* 为**最优解** (optimal point)。

仿射集合

◆ 仿射集合

若对任意 $x_1, x_2 \in C$ 及 $\theta \in \mathbb{R}$, 有 $\theta x_1 + (1 - \theta)x_2 \in C$, 则称 $C \subseteq \mathbb{R}^n$ 为**仿射集合** (affine set)。即仿射集合包含其中任意两点所确定的整条直线。

◆ 仿射包

集合 D 的**仿射包** $\text{aff } D$ 是包含 D 的最小仿射集合：

$$\text{aff } D = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in D, \sum_{i=1}^k \theta_i = 1, k \geq 1 \right\}.$$

⚠ 注

线性方程组的解集 $\{x \mid Ax = b\}$ 是仿射集合。反之，每个仿射集合都可表示为线性方程组的解集。

凸集

◆ 凸集

若对任意 $x_1, x_2 \in C$ 及 $\theta \in [0, 1]$, 有 $\theta x_1 + (1 - \theta)x_2 \in C$, 则称 C 为**凸集** (convex set)。几何上, 凸集中任意两点之间的线段全部包含在集合中。

凸集的例子:

- 空集、单点集、全空间 $\mathbb{R}^n$
- 直线、线段、超平面与半空间
- Euclid 球 $\{x \mid \|x - x_c\|_2 \leq r\}$ 、椭球
- 多面体 $\{x \mid Ax \preceq b, Cx = d\}$
- 半正定锥 $\mathbf{S}_+^n$
- 概率单纯形

非凸集的例子:

- 离散点集、整数格 $\mathbb{Z}^n$
- 圆环 $\{x \mid r \leq \|x\|_2 \leq R\}$
- 球面 $\{x \mid \|x\|_2 = 1\}$
- 非凸多边形的补集
- 可逆矩阵集合 (非凸)

锥与凸锥

◆ 锥

若对任意 $x \in C$ 及 $\theta \geq 0$ 有 $\theta x \in C$ ，则称 C 为**锥** (cone)。若锥 C 同时为凸集，则称为**凸锥** (convex cone)。

◆ 凸锥的等价刻画

C 为凸锥当且仅当对任意 $x_1, x_2 \in C$ 及 $\theta_1, \theta_2 \geq 0$ ，有 $\theta_1 x_1 + \theta_2 x_2 \in C$ 。

凸锥的例子：

- 非负象限 $\mathbb{R}_+^n = \{x \mid x_i \geq 0\}$
- 半正定锥 $\mathbf{S}_+^n = \{X \in \mathbf{S}^n \mid X \succeq 0\}$
- 二阶锥 $\mathcal{K}_{\text{soc}} = \{(x, t) \in \mathbb{R}^{n+1} \mid \|x\|_2 \leq t\}$

范数锥：

$$C = \{(x, t) \in \mathbb{R}^{n+1} \mid \|x\| \leq t\}$$

当范数为 ℓ_2 范数时称为**二阶锥** (second-order cone)，又称冰淇淋锥 (ice-cream cone) 或 Lorentz 锥。

常见的凸集 —— 多面体与单纯形

◆ 多面体

多面体 (polyhedron) 是有限多个线性不等式与线性等式的交集：

$$\mathcal{P} = \{x \mid a_j^\top x \leq b_j, j = 1, \dots, m, c_j^\top x = d_j, j = 1, \dots, p\} = \{x \mid Ax \preceq b, Cx = d\}.$$

有界的多面体称为**多胞形** (polytope)，可等价地表示为其顶点的凸包。

◆ 单纯形

$k + 1$ 个仿射独立点 $v_0, \dots, v_k \in \mathbb{R}^n$ ($k \leq n$) 的凸包构成 k 维**单纯形** (simplex)：

$$\Delta_k = \mathbf{conv}\{v_0, \dots, v_k\} = \left\{ \sum_{i=0}^k \theta_i v_i \mid \theta_i \geq 0, \sum_{i=0}^k \theta_i = 1 \right\}.$$

单位单纯形 $\Delta_n = \{x \succeq 0 \mid \mathbf{1}^\top x = 1\}$ 在概率与优化中最为常见。

常见的凸集 —— 范数球与椭球

◆ 范数球

以 x_c 为中心、 $r > 0$ 为半径的**范数球**：

$$B_{\|\cdot\|}(x_c, r) = \{x \in \mathbb{R}^n \mid \|x - x_c\| \leq r\}.$$

◆ 椭球

设 $P \in \mathbf{S}_{++}^n$ (对称正定)。以 x_c 为中心的**椭球** (ellipsoid) 为

$$\mathcal{E}(x_c, P) = \{x \in \mathbb{R}^n \mid (x - x_c)^\top P^{-1} (x - x_c) \leq 1\}.$$

当 $P = r^2 I$ 时退化为 Euclid 球。等价地， $\mathcal{E}(x_c, P) = \{x_c + Au \mid \|u\|_2 \leq 1\}$ ，其中 A 满足 $P = AA^\top$ 。

保持凸性的运算

► 任意交集

任意多个凸集的交集仍然是凸集。 $\bigcap_{\alpha \in \mathcal{A}} C_\alpha$ 为凸。

► 仿射变换下的像与原像

设 $f(x) = Ax + b$ 为仿射函数。若 C 为凸集，则像 $f(C) = \{f(x) \mid x \in C\}$ 为凸集；若 D 为凸集，则原像 $f^{-1}(D) = \{x \mid f(x) \in D\}$ 亦为凸集。

保持凸性的运算

► 凸包

集合 C 中任意有限个点的凸组合全体构成 C 的**凸包**:

$$\mathbf{conv} C = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in C, \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1, k \geq 1 \right\}.$$

凸包是包含 C 的最小凸集。

► MINKOWSKI 和与笛卡尔积

设 C_1, C_2 为凸集, 则 $C_1 + C_2 = \{x + y \mid x \in C_1, y \in C_2\}$ 及 $C_1 \times C_2$ 均为凸集。

透视函数与线性分式函数

◆ 透视函数

定义 $\mathbb{R}^{n+1}$ 上的透视函数 $P(z, t) = z/t$, $\text{dom } P = \{(z, t) \mid t > 0\}$ 。透视函数将凸集的像和原像均保持为凸集。

▶ 线性分式函数

设 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}$ 。线性分式函数

$$f(x) = \frac{Ax + b}{c^\top x + d}, \quad \text{dom } f = \{x \mid c^\top x + d > 0\}$$

将凸集映射为凸集。该性质源自线性分式函数可分解为仿射函数与透视函数的复合。

透视函数与线性分式函数

例 7.1 —— 条件概率

若 (u, v) 的联合分布为有限个点的凸组合，则条件期望 $\mathbb{E}[u \mid v \in C]$ 可表示为线性分式形式，且取值范围构成凸集。

超平面与支撑超平面

◆ 超平面

设 $a \in \mathbb{R}^n \setminus \{0\}$, $b \in \mathbb{R}$ 。集合 $\mathcal{H} = \{x \in \mathbb{R}^n \mid a^\top x = b\}$ 称为**超平面** (hyperplane)，它是 $\mathbb{R}^n$ 的一个 $n - 1$ 维仿射子空间。两个相伴的闭半空间为 $\{x \mid a^\top x \leq b\}$ 与 $\{x \mid a^\top x \geq b\}$ 。

◆ 支撑超平面

设 $C \subseteq \mathbb{R}^n$ 为凸集, $x_0 \in \partial C$ 。若存在 $a \neq 0$ 使得

$$a^\top x \leq a^\top x_0, \quad \forall x \in C,$$

则称超平面 $\{x \mid a^\top x = a^\top x_0\}$ 为 C 在 x_0 处的**支撑超平面** (supporting hyperplane)。它将凸集完全置于其一侧，且与凸集在边界点处相切。

分离超平面定理

► 分离超平面定理 (SEPARATING HYPERPLANE THEOREM)

设 $C, D \subseteq \mathbb{R}^n$ 为两个不相交的非空凸集 ($C \cap D = \emptyset$)。则存在 $a \neq 0$ 及 $b \in \mathbb{R}$, 使得

$$a^\top x \leq b \quad (\forall x \in C), \quad a^\top x \geq b \quad (\forall x \in D).$$

► 严格分离

若 C 为闭凸集, D 为紧凸集, 且 $C \cap D = \emptyset$, 则存在 $a \neq 0$ 及 $b \in \mathbb{R}$ 使得

$$a^\top x < b \quad (\forall x \in C), \quad a^\top x > b \quad (\forall x \in D).$$

► 超平面严格分离点与闭凸集

设 $C \subseteq \mathbb{R}^n$ 为非空闭凸集, $x_0 \notin C$ 。则存在 $a \neq 0$ 及 $b \in \mathbb{R}$ 将 x_0 与 C 严格分离。

对偶锥与广义不等式

◆ 对偶锥

设 $K \subseteq \mathbb{R}^n$ 为锥。其**对偶锥** (dual cone) 定义为

$$K^* = \{y \in \mathbb{R}^n \mid y^\top x \geq 0, \forall x \in K\}.$$

常见对偶锥：

- $(\mathbb{R}_+^n)^* = \mathbb{R}_+^n$ (自对偶)
- $(\mathbf{S}_+^n)^* = \mathbf{S}_+^n$ (自对偶)
- 二阶锥 $\mathcal{K}_{\text{soc}}$ 自对偶
- 半正定锥的自对偶性是 SDP 对偶理论的基础

基本性质：

- K^* 总是闭凸锥 (即使 K 非闭非凸)
- 若 $K_1 \subseteq K_2$, 则 $K_2^* \subseteq K_1^*$
- 若 K 为闭凸锥, 则 $(K^*)^* = K$
- $K^* = (\text{cl } K)^* = (\text{conv } K)^*$

⚠ 广义不等式

由正常锥 K (闭凸、尖、有非空内部) 可定义广义不等式: $x \preceq_K y \iff y - x \in K$, $x \prec_K y \iff y - x \in \text{int } K$ 。分量不等式和矩阵不等式均为其特例。

第二节 凸函数

Convex Functions

凸函数的定义与基本性质

◆ 凸函数

设函数 $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$, 其有效定义域 $\mathbf{dom} f = \{x \mid f(x) < +\infty\}$ 为凸集。若对任意 $x, y \in \mathbf{dom} f$ 及 $\theta \in [0, 1]$, 有

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y),$$

则称 f 为凸函数。若 $x \neq y$ 且 $\theta \in (0, 1)$ 时不等式严格成立, 则称 f 为严格凸函数。若 $-f$ 为凸, 则称 f 为凹函数。

上图刻画 (epigraph):

f 为凸函数 $\iff$ 其上图 of 凸集:

$$\mathbf{epi} f = \{(x, t) \in \mathbb{R}^{n+1} \mid x \in \mathbf{dom} f, f(x) \leq t\}.$$

扩展值函数的好处:

将 f 在 $\mathbf{dom} f$ 之外定义为 $+\infty$, 可使 f 定义在全空间 $\mathbb{R}^n$ 上, 简化理论表述。

一阶条件与二阶条件

► 定理 7.1 —— 一阶条件

设 f 在开凸集 $\mathbf{dom} f$ 上一阶可微。则 f 为凸函数当且仅当

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in \mathbf{dom} f.$$

即一阶 Taylor 近似构成 f 的**全局下估计** (global underestimator)。

► 定理 7.2 —— 二阶条件

设 f 在开凸集 $\mathbf{dom} f$ 上二阶可微。则 f 为凸函数当且仅当

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \mathbf{dom} f.$$

若 $\nabla^2 f(x) \succ 0$ 对所有 x 成立, 则 f **严格凸**; 但逆向不成立 (反例: $f(x) = x^4$ 在 $x = 0$ 处 Hessian 为零矩阵, 但函数严格凸)。

一阶条件与二阶条件

↳ 推论

仿射函数 $f(x) = a^\top x + b$ 既凸又凹；二次函数 $f(x) = \frac{1}{2}x^\top Px + q^\top x + r$ 为凸 $\iff P \succeq 0$ 。

一阶条件的证明

充分性 (一阶条件 $\implies$ 凸性)

设一阶条件成立。对任意 $x, y \in \text{dom } f$ 及 $\theta \in [0, 1]$, 令 $z = \theta x + (1 - \theta)y \in \text{dom } f$ 。由一阶条件,

$$\begin{aligned}f(x) &\geq f(z) + \nabla f(z)^\top (x - z), \\f(y) &\geq f(z) + \nabla f(z)^\top (y - z).\end{aligned}$$

第一式乘以 θ , 第二式乘以 $1 - \theta$, 相加得

$$\theta f(x) + (1 - \theta)f(y) \geq f(z) + \nabla f(z)^\top (\theta x + (1 - \theta)y - z) = f(z),$$

此即 $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$ 。证毕。

必要性 (凸性 + 可微 $\implies$ 一阶条件)

设 f 凸且可微。对 $t \in (0, 1]$, 定义方向导数。由凸性定义, 对任意 x, y 及 $t \in (0, 1]$,

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

令 $t \rightarrow 0^+$, 左边趋于方向导数 $\nabla f(x)^\top (y - x)$, 故得 $\nabla f(x)^\top (y - x) \leq f(y) - f(x)$ 。证毕。

常见凸函数示例

单变量凸函数：

- e^{ax} , $\forall a \in \mathbb{R}$ ($\mathbb{R}$ 上)
- x^a ($x > 0$): $a \geq 1$ 或 $a \leq 0$ 时凸; $0 \leq a \leq 1$ 时凹
- $|x|^p$, $p \geq 1$ ($\mathbb{R}$ 上)
- $x \log x$ ($x > 0$, 严格凸)
- $-\log x$ ($x > 0$, 严格凸)
- $\log(1 + e^x)$ (logistic 损失)

多变量凸函数：

- 任意范数 $\|\cdot\|$ (三角不等式 + 绝对齐次性)
- 二次型 $x^T P x + q^T x + r$ ($P \succeq 0$)
- $\log \sum_{i=1}^n e^{x_i}$ (log-sum-exp, softmax 的对数)
- $-\log \det X$ ($X \in \mathbf{S}_{++}^n$, 严格凸)
- $\lambda_{\max}(X)$ ($X \in \mathbf{S}^n$, 凸但非光滑)

例 7.2

证明 $\|Ax - b\|_2^2$ 和 $\log \sum_{i=1}^n e^{x_i}$ 均为凸函数。前者 Hessian 为 $2A^T A \succeq 0$; 后者 Hessian 为 $\text{diag}(z) - zz^T$ ($z_i = e^{x_i} / \sum_j e^{x_j}$), 可验证其半正定性。

保持凸性的函数运算 I

► 非负加权和

若 $f_1, \dots, f_m$ 为凸函数, $w_1, \dots, w_m \geq 0$, 则 $f(x) = \sum_{i=1}^m w_i f_i(x)$ 为凸函数。凸函数集合构成一个凸锥。

► 逐点上确界

若对每个 $y \in \mathcal{A}$, $f(x, y)$ 关于 x 为凸, 则 $g(x) = \sup_{y \in \mathcal{A}} f(x, y)$ 关于 x 为凸。特例: 一族凸函数的逐点最大值 $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ 为凸函数。

► 与仿射函数的复合

若 $f: \mathbb{R}^m \rightarrow \mathbb{R}$ 为凸, 则 $g(x) = f(Ax + b)$ 为凸函数。此运算是凸分析中使用最频繁的工具之一。

► 标量复合规则

设 $g: \mathbb{R} \rightarrow \mathbb{R}$ 凸且不减, $h: \mathbb{R}^n \rightarrow \mathbb{R}$ 凸, 则 $f(x) = g(h(x))$ 为凸函数。类似地, 若 g 凸且不增, h 凹, 则 $f = g \circ h$ 亦凸。

保持凸性的函数运算 II

► 下确界（部分极小化）

若 $f(x, y)$ 关于 (x, y) 联合凸，且 C 为非空凸集，则

$$g(x) = \inf_{y \in C} f(x, y)$$

关于 x 为凸函数（假设 $g(x) > -\infty$ ）。

► 透视函数

若 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 为凸，则其透视函数

$$g(x, t) = t f\left(\frac{x}{t}\right), \quad \text{dom } g = \{(x, t) \in \mathbb{R}^{n+1} \mid x/t \in \text{dom } f, t > 0\}$$

关于 (x, t) 联合凸。透视运算是联系凸分析与锥规划的核心桥梁。

保持凸性的函数运算 II

例 7.3 —— KL 散度的凸性

由负熵 $f(x) = x \log x$ ($x > 0$, 严格凸), 取透视函数得

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

该函数关于 (x, t) 联合凸, 是相对熵 (Kullback-Leibler 散度) 的构建块。

共轭函数

◆ FENCHEL 共轭

设 $f: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ 。其**共轭函数** (conjugate / Fenchel dual) $f^*: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ 定义为

$$f^*(y) = \sup_{x \in \text{dom } f} (y^\top x - f(x)).$$

常见共轭对:

$f(x)$	$f^*(y)$
$\frac{1}{2}x^\top Qx \ (Q \succ 0)$	$\frac{1}{2}y^\top Q^{-1}y$
$-\log x \ (x > 0)$	$-\log(-y) - 1 \ (y < 0)$
e^x	$y \log y - y \ (y > 0)$
$ x $	$\mathbb{I}_{ y \leq 1}$ (指示函数)
$\frac{1}{2}\ x\ _2^2$	$\frac{1}{2}\ y\ _2^2$

核心性质:

- f^* 总是**凸函数** (一族仿射函数的上确界)
- 即使 f 非凸, f^* 仍为凸
- 若 f 为正常闭凸函数, 则 $f^{**} = f$
- **Young 不等式:** $x^\top y \leq f(x) + f^*(y)$
- Fenchel 不等式由共轭定义直接推出

拟凸函数

◆ 拟凸函数

若对任意 $x, y \in \text{dom } f$ 及 $\theta \in [0, 1]$, 有

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

则称 f 为**拟凸函数** (quasiconvex function)。等价地, f 拟凸当且仅当所有下水平集 $S_\alpha = \{x \in \text{dom } f \mid f(x) \leq \alpha\}$ ($\forall \alpha \in \mathbb{R}$) 均为凸集。

一阶条件 (可微情形):

f 拟凸 $\iff f(y) \leq f(x) \implies \nabla f(x)^\top (y - x) \leq 0$ 。

即梯度 $\nabla f(x)$ 定义了 x 处下水平集的**支撑超平面**。

与凸函数的关系:

- 凸函数 $\implies$ 拟凸 (反之不然)
- 拟凸函数无凸函数的光滑性保障
- 典型非凸拟凸函数: $\sqrt{|x|}$, $\lceil x \rceil$
- 拟凸优化可能仍有局部非全局最优解

对数凹函数与对数凸函数

◆ 对数凹 / 对数凸

设 $f(x) > 0$ 对所有 $x \in \text{dom } f$ 成立。若 $\log f$ 为凹函数，则称 f 为**对数凹函数** (log-concave)；若 $\log f$ 为凸函数，则称 f 为**对数凸函数** (log-convex)。

对数凹函数的例子：

- 高斯密度 $\phi(x) \propto e^{-x^2/2}$
- 指数分布、Laplace 分布密度
- 多元正态密度函数
- $\det X$ ($X \in \mathbf{S}_{++}^n$)
- Gamma 函数 $\Gamma(x)$ ($x > 0$)

重要性质：

- 对数凹函数的乘积仍为对数凹
- 对数凹密度的边缘分布仍为对数凹 (Prékopa 定理)
- 对数凹性在凸几何、随机优化和高维概率中有深刻应用

第三节 凸优化问题

Convex Optimization Problems

凸优化问题的标准形式

◆ 凸优化问题

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & a_i^\top x = b_i, \quad i = 1, \dots, p,\end{array}$$

其中 $f_0, f_1, \dots, f_m$ 为凸函数，等式约束为仿射的。

核心性质（凸优化与一般优化的根本区别）：

1. 可行集 $\mathcal{F} = \{x \mid f_i(x) \leq 0, Ax = b\}$ 是凸集
2. 任意局部最优解即为全局最优解
3. 最优解集合 $\{x^* \in \mathcal{F} \mid f_0(x^*) = p^*\}$ 为凸集
4. 最优值 p^* 唯一；最优解可以不唯一（如 f_0 非严格凸）

凸优化与非凸优化的对比

特性	凸优化	非凸优化
局部最优与全局最优	局部最优 $\iff$ 全局最优	局部最优 $\nRightarrow$ 全局最优
最优解集合	凸集	无特殊结构
对偶间隙	通常为零（强对偶 + 约束规格）	通常非零
KKT 条件	充要条件 （可微 + 约束规格）	仅必要条件
计算复杂性	多项式时间（内点法）	一般 NP-hard
典型算法	内点法、梯度法、Newton 法	启发式、分支定界、局部搜索

⚠ **重要结论 (BOYD, 2004)**

如果能将实际问题**识别并建模**为凸优化问题，则原则上可以被**可靠、高效地**求解至全局最优。因此，凸优化是计算数学中连接建模与算法的桥梁。

线性规划 (LP)

◆ 线性规划的标准形式

$$\begin{aligned} &\text{minimize} && c^\top x + d \\ &\text{subject to} && Gx \preceq h \\ &&& Ax = b. \end{aligned}$$

典型应用：

- 资源分配与运输问题
- 生产计划与调度
- 投资组合（线性目标 + ℓ_1 约束）
- ℓ_∞ 回归（Chebyshev 逼近）
- 网络流问题（max-flow min-cut）

求解方法：

- **单纯形法** (Dantzig, 1947)：沿多面体棱移动，实践中很高效
- **内点法** (Karmarkar, 1984)：对大规模稀疏问题显著优于单纯形法
- 理论上两者均可达到多项式复杂度

LP 的应用示例

例 7.4 —— ℓ_∞ 回归 (CHEBYSHEV 逼近)

问题 $\min_{x \in \mathbb{R}^n} \|Ax - b\|_\infty$ 等价于如下 LP:

$$\begin{aligned} & \text{minimize} && t \\ & \text{subject to} && -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1}. \end{aligned}$$

例 7.5 —— ℓ_1 回归 (最小绝对偏差)

问题 $\min_x \|Ax - b\|_1$ 等价于:

$$\begin{aligned} & \text{minimize} && \mathbf{1}^\top t \\ & \text{subject to} && -t \preceq Ax - b \preceq t. \end{aligned}$$

二次规划 (QP) 与二次约束二次规划 (QCQP)

◆ 二次规划 (QP)

$$\begin{aligned} & \text{minimize} && \frac{1}{2}x^\top Px + q^\top x + r \\ & \text{subject to} && Gx \preceq h, \quad Ax = b, \end{aligned}$$

其中 $P \in \mathbf{S}_+^n$ (半正定) 确保目标函数为凸。当 $P \succ 0$ 时, 目标严格凸, 最优解唯一。

◆ 二次约束二次规划 (QCQP)

$$\begin{aligned} & \text{minimize} && \frac{1}{2}x^\top P_0 x + q_0^\top x + r_0 \\ & \text{subject to} && \frac{1}{2}x^\top P_i x + q_i^\top x + r_i \leq 0, \quad i = 1, \dots, m \\ & && Ax = b, \end{aligned}$$

其中 $P_i \succeq 0$ ($i = 0, 1, \dots, m$)。

二阶锥规划 (SOCP)

◆ 二阶锥规划

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

约束 $\|A_i x + b_i\|_2 \leq c_i^\top x + d_i$ 称为**二阶锥约束**，等价于 $(A_i x + b_i, c_i^\top x + d_i) \in \mathcal{K}_{\text{soc}}$ 。

问题层次结构： $\text{LP} \subset \text{QP} \subset \text{QCQP} \subset \text{SOCP} \subset \text{SDP}$

SOCP 将 QP 的"平方和"约束推广为 Euclid 范数约束，进而可处理**机会约束** (chance constraint) 和**鲁棒优化**问题。

🔍 例 7.6 —— 机会约束的 SOCP 表述

设 $a \sim \mathcal{N}(\bar{a}, \Sigma)$ 。概率约束 $\text{Prob}(a^\top x \leq b) \geq \Phi(\beta)$ 等价于二阶锥约束 $\bar{a}^\top x + \beta \|\Sigma^{1/2} x\|_2 \leq b$ 。

半正定规划 (SDP)

◆ 半正定规划的标准形式

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

其中优化变量 $X \in \mathbf{S}^n$ ($n \times n$ 对称矩阵), 数据 $C, A_i \in \mathbf{S}^n$, $b_i \in \mathbb{R}$ 。约束 $X \succeq 0$ 要求 X 属于半正定锥 $\mathbf{S}_+^n$ 。

为什么 SDP 如此重要?

- 统一框架: $\text{LP} \subset \text{SOCP} \subset \text{SDP}$
- 可处理矩阵不等式 (LMI)
- 在控制、组合优化、机器学习核方法中广泛应用
- 半正定松弛是 NP-hard 问题的有效近似

Schur 补引理 (SDP 建模的核心工具):

对 $X \succ 0$,

$$S = G - FX^{-1}F^T \succeq 0 \iff \begin{bmatrix} X & F^T \\ F & G \end{bmatrix} \succeq 0.$$

SDP 应用 —— 矩阵范数与特征值优化

例 7.7 —— 矩阵谱范数最小化

设 $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$ 。 $\|A(x)\|_2$ (最大奇异值) 是 x 的凸函数。 $\|A(x)\|_2 \leq t$ 等价于 LMI:

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

因此 $\min_x \|A(x)\|_2$ 可写为 SDP。

例 7.8 —— 最大特征值最小化

对 $A(x) = A_0 + \sum_{i=1}^n x_i A_i \in \mathbf{S}^m$, $\lambda_{\max}(A(x)) \leq t \iff A(x) \preceq tI$ 。 因此:

$$\min_x \lambda_{\max}(A(x)) \iff \begin{cases} \min t \\ \text{s.t. } A(x) \preceq tI. \end{cases}$$

此为 SDP (不等式约束为 LMI)。

几何规划 (GP)

◆ 单项式与正项式

设 $x = (x_1, \dots, x_n) \succ 0$ (变量严格正)。**单项式** (monomial): $f(x) = c x_1^{a_1} \cdots x_n^{a_n}$ ($c > 0$, $a_i \in \mathbb{R}$)。**正项式** (posynomial): 正系数单项式的和。

◆ 几何规划

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

其中 f_i 为正项式, h_i 为单项式。

几何规划 (GP)

⚠ 凸化变换

取变量代换 $y_i = \log x_i$, 正项式约束 $f_i(x) \leq 1$ 变为 $\log \sum_k e^{a_k^\top y + b_k} \leq 0$ (log-sum-exp 形式), 而 log-sum-exp 是凸函数。因此每个几何规划都等价于一个凸优化问题。

第四节 对偶理论

Duality

Lagrange 函数与对偶函数

◆ LAGRANGE 函数

考虑标准形式优化问题（不要求凸）。定义 **Lagrange 函数** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ ：

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

其中 $\lambda = (\lambda_1, \dots, \lambda_m)$ 为**对偶变量**（Lagrange 乘子）， $\nu = (\nu_1, \dots, \nu_p)$ 为等式约束的乘子。

◆ LAGRANGE 对偶函数

Lagrange 对偶函数（dual function）定义为 L 关于 x 的下确界：

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right).$$

Lagrange 对偶函数的基本性质

► 定理 7.3 —— 下界性质

对任意 $\lambda \succeq 0$ 及任意 $\nu \in \mathbb{R}^p$, 有

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

其中 p^* 为原始问题的最优值。即对偶函数为最优值提供了一族**全局下界**。

► 定理 7.4 —— 对偶函数的凹性

无论原始问题是否凸, Lagrange 对偶函数 $g(\lambda, \nu)$ 总是关于 (λ, ν) 的**联合凹函数**。

Lagrange 对偶函数的基本性质

证明思路

对任意固定的 x , $L(x, \lambda, \nu)$ 是关于 (λ, ν) 的仿射函数。 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必为凹函数。

Lagrange 对偶问题

◆ LAGRANGE 对偶问题

$$\begin{array}{ll}\text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0.\end{array}$$

关键术语：

- **原始问题** (primal)：以 x 为变量，目标 minimize
- **对偶问题** (dual)：以 (λ, ν) 为变量，目标 maximize
- p^* ：原始最优值； d^* ：对偶最优值
- **弱对偶** $d^* \leq p^*$ 恒成立（由下界性质直接可得）

对偶问题的优势：

- 对偶问题**总是凸的**（最大化的目标为凹函数，约束为线性不等式）
- 对偶变量的维数 m 等于约束个数
- 对偶最优值 d^* 提供了 p^* 的下界，可用于分支定界等方法

弱对偶与强对偶

► 定理 7.5 —— 弱对偶定理

对任意原始可行解 x ($f_i(x) \leq 0$, $h_i(x) = 0$) 及对偶可行解 (λ, ν) ($\lambda \succeq 0$), 恒有

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

取上确界和下确界即得 $d^* \leq p^*$ 。差值 $p^* - d^* \geq 0$ 称为**对偶间隙** (duality gap)。

► 强对偶

若 $d^* = p^*$ (对偶间隙为零), 则称**强对偶** (strong duality) 成立。强对偶**不是无条件成立的**——对一般非凸问题, 对偶间隙可能任意大。对凸优化问题, 若满足适当的**约束规格** (constraint qualification), 则强对偶成立。

△ 强对偶的重要性

当强对偶成立时, 可通过求解对偶问题获得原始最优值 p^* , 进而通过 KKT 条件恢复原始最优解 x^* 。

Slater 约束规格

► 定理 7.6 —— SLATER 条件

考虑凸优化问题

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m, \quad Ax = b.\end{array}$$

若存在 $x \in \text{relint } \mathcal{D}$ (定义域的**相对内部**) 使得

$$f_i(x) < 0, \quad \forall i = 1, \dots, m, \quad Ax = b,$$

则**强对偶成立** ($p^* = d^*$)。

特殊情形：

- 若所有不等式约束 f_i 均为仿射函数，Slater 条件退化为：存在 $x \in \mathcal{D}$ 满足 $Ax = b$ (仅需可行性即可)

实践意义：

- Slater 条件是验证凸优化强对偶的最常用方法
- 对绝大多数实际凸优化问题，Slater 条件自然满足
- Boyd & Vandenberghe: "almost all convex problems have zero duality gap"

KKT 最优性条件

► 定理 7.7 — KKT 条件 (KARUSH-KUHN-TUCKER)

设 $f_0, f_1, \dots, f_m$ 可微, 强对偶成立。则 x^* 与 (λ^*, ν^*) 分别为原始最优解与对偶最优解, 当且仅当满足以下 KKT 条件:

(原始可行性) $f_i(x^*) \leq 0, i = 1, \dots, m; \quad h_i(x^*) = 0, i = 1, \dots, p$

(对偶可行性) $\lambda_i^* \geq 0, i = 1, \dots, m$

(互补松弛性) $\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m$

(稳定性) $\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0.$

凸优化中的 KKT: 对可微凸优化问题, 若 Slater 条件成立, 则 KKT 条件是 x^* 为全局最优解的**充要条件**。即找到满足 KKT 的点即可确认其为全局最优。

KKT 条件的解读

互补松弛性 $\lambda_i^* f_i(x^*) = 0$:

- 若 $\lambda_i^* > 0$, 则必有 $f_i(x^*) = 0$ (第 i 个约束在最优解处**活跃**)
- 若 $f_i(x^*) < 0$, 则必有 $\lambda_i^* = 0$ (约束**不活跃**, 不参与最优性条件)
- 只有活跃约束才对最优解施加影响

稳定性条件的几何解释:

重写为 $-\nabla f_0(x^*) = \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*)$ 。

目标函数的负梯度方向处于活跃约束梯度与等式约束梯度所生成的**锥**中。无约束时退化为 $\nabla f_0(x^*) = 0$ 。

⚠ 互补松弛性的经济学解释 (影子价格)

λ_i^* 度量了第 i 个约束的"影子价格"——约束放宽一个单位对最优目标的边际改善。互补松弛性表示: 非活跃约束的影子价格为零。

KKT 条件 —— 等式约束二次规划示例

例 7.9

$$\begin{aligned} & \text{minimize} && \frac{1}{2}x^\top Px + q^\top x + r \\ & \text{subject to} && Ax = b, \end{aligned}$$

其中 $P \succeq 0$, $A \in \mathbb{R}^{p \times n}$ (假设 $\text{rank } A = p$)。

由 KKT 条件, 存在 $\nu^* \in \mathbb{R}^p$ 满足:

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

此即 **KKT 系统** (saddle-point system)。其系数矩阵非奇异当且仅当 $P \succ 0$ 在 $\ker A$ 上 (即约化 Hessian $Z^\top P Z \succ 0$)。

⚠ 扩展到含不等式约束的情形

含不等式约束时, 互补松弛条件 $\lambda_i^* f_i(x^*) = 0$ 将问题从光滑方程转变为**混合互补问题**, 通常需 active-set 方法或内点法迭代求解。

第五节 无约束优化算法

Unconstrained Optimization Algorithms

梯度下降法

◆ 梯度下降迭代

对无约束问题 $\min_x f(x)$ (f 凸且可微), 梯度下降法的迭代格式为

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

其中 $t_k > 0$ 为步长 (step size 或 learning rate)。

精确线搜索 (exact line search):

$$t_k = \operatorname{argmin}_{t \geq 0} \phi(t), \quad \phi(t) = f(x^{(k)} - t \nabla f(x^{(k)})).$$

一维凸优化问题, 通常使用黄金分割搜索或二分法。

回溯线搜索 (backtracking line search):

给定 $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$:

1. $t := t_0$
2. while $f(x - t \nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$:
3. $t := \beta t$

Armijo 条件确保充分下降。

梯度下降的收敛性分析

► 定理 7.8 —— 梯度下降收敛率

设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 为凸函数, ∇f 以常数 $L > 0$ 为 Lipschitz 连续:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y.$$

则取固定步长 $t = 1/L$ (或 t_k 由回溯线搜索确定), 有如下收敛率:

一般凸情形 (非强凸):

$$f(x^{(k)}) - p^* \leq \frac{L\|x^{(0)} - x^*\|_2^2}{2k} = O\left(\frac{1}{k}\right).$$

子线性收敛, 达到精度 ε 需 $O(1/\varepsilon)$ 次迭代。

强凸情形 (μ -强凸, $\mu > 0$):

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{\mu}{L}\right)^k \frac{L}{2} \|x^{(0)} - x^*\|_2^2.$$

线性收敛, 每次迭代以常数因子 $(1 - \mu/L)$ 缩小误差。

梯度下降完整算法与条件数

算法 5.1 —— 梯度下降法（回溯线搜索）

给定初始点 $x^{(0)} \in \text{dom } f$, 参数 $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$, 容忍度 $\varepsilon > 0$, **repeat** $k = 0, 1, \dots$: (1) $g := \nabla f(x^{(k)})$; 若 $\|g\|_2 \leq \varepsilon$ 则停止。(2) 回溯: $t := 1$; **while** $f(x^{(k)} - tg) > f(x^{(k)}) - \alpha t \|g\|_2^2$, 令 $t := \beta t$ 。(3) $x^{(k+1)} := x^{(k)} - tg$ 。

条件数与收敛速度

$\kappa = L/\mu$ 越大, 收敛越慢: 强凸线性收敛率为 $(1 - 1/\kappa)^k$ 。对二次函数 $f(x) = \frac{1}{2}x^\top Qx$, $\kappa = \lambda_{\max}(Q)/\lambda_{\min}(Q)$ 。 $\kappa = 100$ 时约需 460 步, $\kappa = 10^4$ 时约需 4.6×10^4 步 (精度 10^{-4})。

Nesterov 加速梯度法（最优一阶方法, $O(1/k^2)$):

$$y^{(k+1)} = x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)}),$$

$$x^{(k+1)} = y^{(k+1)} - \frac{1}{L}\nabla f(y^{(k+1)}).$$

收敛率 $O(1/k^2)$, 达到凸函数一阶方法的信息论下界。

梯度下降 $O(1/k)$ 收敛性证明

定理 7.8 的证明 —— 一般凸情形

设 f 为凸且 L -光滑。取固定步长 $t = 1/L$ 。以下分三步证明 $f(x^{(k)}) - p^* \leq \frac{L\|x^{(0)} - x^*\|_2^2}{2k}$ 。

第一步： L -光滑性给出函数值下降。

对任意 x, y , L -光滑性等价于 $f(y) \leq f(x) + \nabla f(x)^\top(y - x) + \frac{L}{2}\|y - x\|_2^2$ 。取 $x = x^{(k)}$, $y = x^{(k+1)} = x^{(k)} - t\nabla f(x^{(k)})$:

$$f(x^{(k+1)}) \leq f(x^{(k)}) - t\|\nabla f(x^{(k)})\|_2^2 + \frac{Lt^2}{2}\|\nabla f(x^{(k)})\|_2^2 = f(x^{(k)}) - \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2.$$

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

第二步：利用凸性建立迭代点距离与函数值的关系。

$$\begin{aligned}\|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - t\nabla f(x^{(k)}) - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2t\nabla f(x^{(k)})^\top(x^{(k)} - x^*) + t^2\|\nabla f(x^{(k)})\|_2^2 \\ &\leq \|x^{(k)} - x^*\|_2^2 - 2t\Delta_k + t^2\|\nabla f(x^{(k)})\|_2^2 \quad (\text{凸性: } \nabla f(x^{(k)})^\top(x^{(k)} - x^*) \geq \Delta_k).\end{aligned}$$

第三步：整理得递推关系并求和。

由上两步消去 $\|\nabla f(x^{(k)})\|_2^2$, 代入 $t = 1/L$, 得 $\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - \frac{2}{L}\Delta_{k+1}$ 。重排即 $\Delta_{k+1} \leq \frac{L}{2}(\|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2)$ 。对 $t = 0, \dots, k-1$ 求和, 利用 Δ_t 不增:

$$k\Delta_k \leq \sum_{t=0}^{k-1} \Delta_{t+1} \leq \frac{L}{2}\|x^{(0)} - x^*\|_2^2 \implies \Delta_k \leq \frac{L\|x^{(0)} - x^*\|_2^2}{2k}. \quad \blacksquare$$

梯度下降线性收敛性证明（强凸情形）

定理 7.8 的证明 —— μ -强凸情形

设 f 为 μ -强凸且 L -光滑，取步长 $t = 1/L$ 。证明：
$$f(x^{(k)}) - p^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x^{(0)}) - p^*).$$

第一步： L -光滑性给出函数值单步下降。（与一般凸情形相同）

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

第二步： μ -强凸给出梯度范数的下界。

μ -强凸 $\Leftrightarrow$ 对任意 x, y : $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|_2^2$ 。

取 $y = x^*$ ，对不等式右端关于 $x^{(k)} - x^*$ 最小化（完全平方），得

$$p^* \geq f(x^{(k)}) - \frac{1}{2\mu} \|\nabla f(x^{(k)})\|_2^2 \quad \implies \quad \|\nabla f(x^{(k)})\|_2^2 \geq 2\mu (f(x^{(k)}) - p^*).$$

第三步：组合两步得线性递推并迭代。

代入第一步，记 $\Delta_k = f(x^{(k)}) - p^*$ ：

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2L} \|\nabla f(x^{(k)})\|_2^2 \leq \Delta_k - \frac{\mu}{L} \Delta_k = \left(1 - \frac{\mu}{L}\right) \Delta_k.$$

对 $k = 0, 1, \dots, k-1$ 迭代即得 $\Delta_k \leq \left(1 - \frac{\mu}{L}\right)^k \Delta_0 = \left(1 - \frac{\mu}{L}\right)^k (f(x^{(0)}) - p^*)$. ■

最速下降法

◆ 最速下降方向

最速下降法 (steepest descent) 通过选择范数自适应搜索方向。在一般范数 $\|\cdot\|$ 下, 最速下降方向为

$$\Delta x_{\text{sd}} = \|\nabla f(x)\|_* \cdot \operatorname{argmin}_{\|v\| \leq 1} \nabla f(x)^\top v,$$

其中 $\|\cdot\|_*$ 为对偶范数: $\|y\|_* = \sup_{\|x\| \leq 1} y^\top x$ 。

常见范数下的最速下降方向:

范数	最速下降方向
$\ \cdot\ _2$	$-\nabla f(x)$ (标准梯度)
$\ \cdot\ _P$	$-P^{-1}\nabla f(x)$
$\ \cdot\ _1$	按最大梯度分量坐标下降
$\ \cdot\ _\infty$	符号梯度方向

关键洞察:

Newton 方法可视为在由局部 Hessian 诱导的范数

$$\|v\|_{\nabla^2 f(x)} = \sqrt{v^\top \nabla^2 f(x) v}$$

下的**最速下降法**。

Newton 方法

◆ NEWTON 步

对二阶可微目标函数 f ，**Newton 步** Δx_{nt} 为线性系统

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

的解。迭代格式： $x^{(k+1)} = x^{(k)} + t_k \Delta x_{\text{nt}}^{(k)}$ 。

几何来源：

Δx_{nt} 是最小化 f 在 x 处的二阶 Taylor 逼近：

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

Newton 减量（停止准则）：

$$\begin{aligned} \lambda(x) &= \left(\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x) \right)^{1/2} \\ &= \sqrt{-\nabla f(x)^\top \Delta x_{\text{nt}}}. \end{aligned}$$

$$f(x) - p^* \leq \lambda(x)^2 / (2m) \quad (\text{当 } \nabla^2 f \succeq mI).$$

Newton 方法完整算法

算法 5.2 —— 带回溯线搜索的 NEWTON 方法

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

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: $\lambda^2 := -\nabla f(x^{(k)})^\top \Delta x_{\text{nt}} = \Delta x_{\text{nt}}^\top \nabla^2 f(x^{(k)}) \Delta x_{\text{nt}}$
3. **if** $\lambda^2/2 \leq \varepsilon$: **break** ($f(x^{(k)}) - p^*$ 的估计已充分小)
4. 回溯线搜索: $t := 1$; **while** $f(x^{(k)} + t\Delta x_{\text{nt}}) > f(x^{(k)}) - \alpha t \lambda^2$: 令 $t := \beta t$
5. $x^{(k+1)} := x^{(k)} + t\Delta x_{\text{nt}}$

每步主要计算量:

操作	代价
Hessian 计算 $\nabla^2 f(x)$	$O(n^2)$
Cholesky 分解 (正定)	$O(n^3/3)$
前/后向替换 Δx_{nt}	$O(n^2)$
总计	$O(n^3)/\text{步}$

停止准则的可靠性:

Newton 减量满足 $\lambda(x)^2/2 \leq f(x) - p^*$ (精确上界), 停止判断无需知道 p^* 。由二次收敛性, 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 ($\approx 10^{-15}$)。

Newton 方法的收敛性分析

► 定理 7.9 —— NEWTON 方法的收敛性

设 f 在最优解 x^* 附近二阶连续可微, $\nabla^2 f(x^*) \succ 0$, 且 $\nabla^2 f$ Lipschitz 连续 (常数 L)。则 Newton 方法收敛分两个阶段:

阻尼 Newton 阶段 (damped phase):

远离 x^* 时, 使用回溯线搜索。每次迭代函数值下降至少一个常数 $\gamma > 0$ 。此阶段最多 $\frac{f(x^{(0)}) - p^*}{\gamma}$ 次迭代。

纯 Newton 阶段 (quadratically convergent phase):

进入 $\|\nabla f(x)\|_2 \leq \eta = \frac{2m^2}{L}$ 区域后, 取 $t_k = 1$:

$$\frac{\|\nabla f(x^{(k+1)})\|_2}{\|\nabla f(x^{(k)})\|_2^2} \leq \frac{L}{2m^2}.$$

即**二次收敛**——每次迭代正确数字位数翻倍。

⚠ 仿射不变性

Newton 步与坐标选取无关 ($\tilde{x} = Tx$ 变换下 $\Delta \tilde{x}_{\text{nt}} = T \Delta x_{\text{nt}}$)。这是其区别于梯度类方法的核心理论优势。

Newton 方法二次收敛性证明（概要）

定理 7.9 证明概要（纯 Newton 阶段, $t_k = 1$ ）

设 f 满足 $\nabla^2 f \succeq mI$ （在 x^* 附近）， $\nabla^2 f$ Lipschitz 连续（常数 L ）， $e_k = \|x^{(k)} - x^*\|_2$ 。

第一步：Newton 迭代误差展开。

由 $x^{(k+1)} - x^* = x^{(k)} - x^* - [\nabla^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)})$ 及 $\nabla f(x^*) = 0$ ，在 x^* 处展开：

$$\nabla f(x^{(k)}) = \nabla^2 f(x^*)(x^{(k)} - x^*) + R_k, \quad \|R_k\|_2 = O(e_k^2).$$

代入得 $x^{(k+1)} - x^* = [\nabla^2 f(x^{(k)})]^{-1} (\nabla^2 f(x^{(k)}) - \nabla^2 f(x^*))(x^{(k)} - x^*) + [\nabla^2 f(x^{(k)})]^{-1} R_k$ 。

第二步：取范数得二次界。

记 $H_k = \nabla^2 f(x^{(k)})$ 。由 Lipschitz 条件 $\|H_k - \nabla^2 f(x^*)\|_2 \leq L e_k$ ，以及 $e_k \leq m/(2L)$ 时 $\|H_k^{-1}\|_2 \leq 2/m$ ，

$$e_{k+1} \leq \frac{2}{m} \cdot L e_k^2 + O(e_k^2).$$

精细分析（Boyd & Vandenberghe §9.5）给出精确常数：
$$e_{k+1} \leq \frac{L}{2m^2} e_k^2. \quad \blacksquare$$

△ 收敛意义

令 $c = L/(2m^2)$ ，若 $c e_0 < 1$ ，则 $e_k \leq c^{-1} (c e_0)^{2^k}$ ：正确有效数字位数每步翻倍。

拟牛顿方法 (BFGS)

◆ 拟牛顿方法

拟牛顿方法 (quasi-Newton) 使用近似的 Hessian 或其逆, 迭代格式为

$$\begin{aligned}\Delta x^{(k)} &= -B_k^{-1} \nabla f(x^{(k)}), \\ x^{(k+1)} &= x^{(k)} + t_k \Delta x^{(k)},\end{aligned}$$

其中 $B_k \succ 0$ 为 $\nabla^2 f(x^{(k)})$ 的近似, 每步以**低秩更新**修正。

BFGS 更新公式 (Broyden–Fletcher–Goldfarb–Shanno, 1970):

$$B_{k+1} = B_k + \frac{y_k y_k^\top}{y_k^\top s_k} - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k},$$

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

核心性质: 若 $B_k \succ 0$ 且 $y_k^\top s_k > 0$ (Wolfe 曲率条件保证), 则 $B_{k+1} \succ 0$ 。BFGS 满足**割线条件** $B_{k+1} s_k = y_k$ (离散 Newton 方程)。L-BFGS 仅存储最近 m 对 (s_k, y_k) , 实现 $O(mn)$ 存储。

L-BFGS 两环递推算法

L-BFGS 不显式存储 B_k^{-1} ，而是利用最近 m 对 (s_j, y_j) ，通过**两环递推**在 $O(mn)$ 时间计算搜索方向 $d_k = -H_k \nabla f(x^{(k)})$ (其中 $\rho_j = 1/(y_j^\top s_j)$):

前向循环 (修正梯度 q):

令 $q := \nabla f(x^{(k)})$, 对 $j = k-1, k-2, \dots, k-m$:

$$\alpha_j := \rho_j s_j^\top q; \quad q := q - \alpha_j y_j.$$

后向循环 (恢复方向 r):

令 $r := H^{(0)} q$ (初始近似), 对 $j = k-m, \dots, k-1$:

$$\beta := \rho_j y_j^\top r; \quad r := r + s_j(\alpha_j - \beta).$$

令 $d_k := -r$ 。

收敛性与实践要点: BFGS/L-BFGS 在强凸问题 (配合 Wolfe 线搜索) 上**超线性收敛** (Dennis-Moré 条件)。初始近似通常取 $H^{(0)} = \gamma_k I$, $\gamma_k = (s_{k-1}^\top y_{k-1}) / \|y_{k-1}\|_2^2$ (自适应尺度, 实践效果优于固定 I)。存储 m 对 (s_j, y_j) 共 $2mn$ 个浮点数, 实践取 $m = 5 \sim 20$, 适用于 $n > 10^4$ 的大规模问题。

无约束优化方法对比总结

方法	每步代价	收敛率	优势	局限
梯度下降	$O(n)$	子线性 $O(1/k)$	实现极简，每步代价最低	对条件数高度敏感
最速下降	$O(n)$	线性	可灵活选择范数	与梯度下降本质相似
Newton 法	$O(n^3)$	二次收敛	局部极快，仿射不变	Hessian 计算与分解代价高
BFGS	$O(n^2)$	超线性	无需显式 Hessian	需存储 $n \times n$ 矩阵
L-BFGS	$O(nm)$	线性/超线性	$O(nm)$ 存储 ($m \ll n$)	非光滑问题可能不稳定

⚠ 实践中如何选择？

- $n \leq 100$: Newton 方法（二次收敛的对数步代价足以抵消每步 $O(n^3)$ 成本）
- $100 < n \leq 10^4$: BFGS 或 L-BFGS
- $n > 10^4$: L-BFGS ($m \approx 10$) 或梯度下降及其变体
- 非光滑目标：次梯度方法（subgradient）或邻近梯度法（proximal gradient）

第六节 等式约束优化

Equality Constrained Optimization

等式约束凸优化问题

◆ 等式约束凸优化

$$\begin{array}{ll}\text{minimize} & f(x) \\ \text{subject to} & Ax = b,\end{array}$$

其中 f 为二阶可微凸函数, $A \in \mathbb{R}^{p \times n}$ 满足 $\text{rank } A = p < n$ (行满秩, 约束独立)。

KKT 条件 (充要):

存在 $\nu^* \in \mathbb{R}^p$ 使得

$$\nabla f(x^*) + A^\top \nu^* = 0, \quad Ax^* = b.$$

$n + p$ 个方程确定 $n + p$ 个未知量。

可行性结构:

解集 $\{x \mid Ax = b\}$ 构成 $n - p$ 维仿射子空间。任意可行点可写为 $x = x_0 + z$, 其中 $z \in \ker A$ 。最优解在 $\ker A$ 方向上的约化梯度为零。

等式约束 Newton 方法

◆ 等式约束 NEWTON 步

在可行点 x (满足 $Ax = b$) 处, 等式约束 Newton 步 Δx_{nt} 由以下 KKT 系统确定:

$$\begin{bmatrix} \nabla^2 f(x) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix},$$

其中 w 为对偶变量的 Newton 步, 收敛后 $\nu^* = w$ 。

关键性质:

- $A\Delta x_{\text{nt}} = 0$: Newton 步始终在 $\ker A$ 中, 迭代自动保持可行性
- Newton 减量 $\lambda(x) = \sqrt{\Delta x_{\text{nt}}^\top \nabla^2 f(x) \Delta x_{\text{nt}}}$ 作为停止准则
- 当 $\nabla^2 f(x) \succeq mI$ (在 $\ker A$ 上) 时, $f(x) - p^* \leq \lambda(x)^2 / (2m)$

算法框架:

1. 计算 Δx_{nt} 与 $\lambda(x)$
2. 若 $\lambda^2 / 2 \leq \varepsilon$, 停止
3. 回溯线搜索: 从 $t = 1$ 起, 满足 $f(x + t\Delta x_{\text{nt}}) \leq f(x) - \alpha t \lambda^2$ 时接受
4. $x := x + t\Delta x_{\text{nt}}$, 返回步骤 1

消元法 (Elimination Method)

◆ 变量消元

令 $F \in \mathbb{R}^{n \times (n-p)}$ 的列构成 $\ker A$ 的一组基, x_0 为 $Ax = b$ 的一个特解。则任意可行点可写为

$$x = Fz + x_0, \quad z \in \mathbb{R}^{n-p}.$$

代入原问题得**无约束等价问题**:

$$\text{minimize}_{z \in \mathbb{R}^{n-p}} \quad \tilde{f}(z) = f(Fz + x_0),$$

其梯度 $\nabla \tilde{f}(z) = F^\top \nabla f(Fz + x_0)$, Hessian $\nabla^2 \tilde{f}(z) = F^\top \nabla^2 f(Fz + x_0) F$ 。

优点:

- 把约束优化转化为无约束问题
- 可用任意无约束算法 (梯度下降、Newton、BFGS 等)
- 维数从 n 降至 $n - p$

缺点:

- F 的计算可能破坏矩阵稀疏结构
- Newton 方法的仿射不变性在变换后不再自然显现
- 当约束数量 p 较小时, KKT 系统直接求解通常更高效

第七节 内点法

Interior Point Methods

障碍函数思想

考虑含不等式约束的凸优化问题：

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b.\end{array}$$

障碍函数思想

◆ 对数障碍函数 (LOGARITHMIC BARRIER)

将不等式约束以**对数障碍函数**吸收至目标函数中：

$$\phi(x) = - \sum_{i=1}^m \log(-f_i(x)), \quad \text{dom } \phi = \{x \mid f_i(x) < 0, i = 1, \dots, m\}.$$

$\phi(x)$ 在可行域的**内部**光滑，当 x 趋近边界 ($f_i(x) \rightarrow 0^-$) 时 $\phi(x) \rightarrow +\infty$ ，自动将迭代点**隔离**在严格可行域内。

⚠ 障碍方法的核心哲学

用光滑障碍替代非光滑的"壁"（即指示函数 $\mathbb{I}_{(-\infty, 0]}(f_i(x))$ ），将带约束问题转化为一系列无约束（或等式约束）光滑问题的序列。

中心路径

◆ 障碍问题

对参数 $t > 0$ ，定义**障碍问题** (barrier problem)：

$$\begin{aligned} & \text{minimize} && t f_0(x) + \phi(x) \\ & \text{subject to} && Ax = b. \end{aligned}$$

其最优解 $x^*(t)$ 称为**中心点** (central point)。集合 $\{x^*(t) \mid t > 0\}$ 构成一条光滑曲线，称为**中心路径** (central path)。

中心路径的关键性质：

1. $x^*(t)$ 严格可行： $f_i(x^*(t)) < 0$ (对所有 i) 且 $Ax^*(t) = b$
2. 对每个 $t > 0$ ，存在 $\nu^*(t)$ 满足修改的 KKT 条件： $\nabla f_0(x^*(t)) + \sum_i \frac{1}{-tf_i(x^*(t))} \nabla f_i(x^*(t)) + A^T \nu^*(t) = 0$
3. 当 $t \rightarrow \infty$ 时， $x^*(t) \rightarrow x^*$ (原始最优解)
4. 对偶间隙上界： $f_0(x^*(t)) - p^* \leq m/t$

障碍方法 (Barrier Method)

算法 7.1 —— 障碍方法

给定严格可行初始点 $x^{(0)}$ ($f_i(x^{(0)}) < 0$, $Ax^{(0)} = b$), 初始参数 $t^{(0)} > 0$, 倍增因子 $\mu > 1$, 容忍度 $\varepsilon > 0$:

repeat

1. 以 $x^{(k)}$ 为初始点, 用**等式约束 Newton 方法**精确求解障碍问题 (参数 $t = t^{(k)}$), 得中心点 $x^*(t^{(k)})$
2. $x^{(k+1)} := x^*(t^{(k)})$
3. **if** $m/t^{(k)} \leq \varepsilon$: **break**
4. $t^{(k+1)} := \mu t^{(k)}$

收敛性 (自协调函数性质):

若 f_0 和 $-\log(-f_i)$ 均为自协调 (self-concordant) 函数, 则每次中心化仅需常数量 Newton 步。总 Newton 步数:

$$\# \text{Newton} \approx \left\lceil \frac{\log(m/(\varepsilon t^{(0)}))}{\log \mu} \right\rceil \times \text{const.}$$

参数选择建议:

- μ : 实践中取 $10 \sim 20$ 最稳健
- $t^{(0)}$: 使 $m/t^{(0)}$ 与初始对偶间隙同阶
- 阶段 I (找可行初始点): 可构造辅助问题, 将原问题嵌入

原始-对偶内点法

◆ 修改的 KKT 条件

障碍方法只更新原始变量 x 。原始-对偶内点法直接修改 KKT 条件，引入参数 $t > 0$ ，同时更新原始变量 x 和对偶变量 (λ, ν) ：

$$\begin{aligned}\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + A^T \nu &= 0, \\ -\lambda_i f_i(x) &= \frac{1}{t}, \quad i = 1, \dots, m, \\ Ax &= b, \\ f_i(x) &< 0, \quad \lambda_i > 0, \quad i = 1, \dots, m.\end{aligned}$$

与障碍方法的区别：

原始-对偶方法每步只做一次 **Newton 步**（不精确求解障碍问题），然后直接增大 t 。相比障碍方法的"外层循环 + 内层 Newton"，原始-对偶更高效。

实际地位：

原始-对偶内点法是现代求解器（Gurobi、Mosek、CVXOPT 等）的**核心算法**。对 LP、QP、SOCP 和 SDP，它兼具理论最优复杂度和实际效率。

原始-对偶搜索方向：Newton 线性系统

定义三类残差： $r_{\text{dual}} = \nabla f_0(x) + Df(x)^\top \lambda + A^\top \nu$ (对偶)， $r_{\text{cent},i} = -\lambda_i f_i(x) - 1/t$ (中心性)， $r_{\text{pri}} = Ax - b$ (原始)。每步求解如下 Newton 线性系统确定搜索方向 $(\Delta x, \Delta \lambda, \Delta \nu)$ ：

$$\begin{bmatrix} \nabla^2 f_0(x) + \sum_i \lambda_i \nabla^2 f_i(x) & Df(x)^\top & A^\top \\ -\text{diag}(\lambda) Df(x) & -\text{diag}(f(x)) & 0 \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \\ \Delta \nu \end{bmatrix} = - \begin{bmatrix} r_{\text{dual}} \\ r_{\text{cent}} \\ r_{\text{pri}} \end{bmatrix},$$

其中 $Df(x) \in \mathbb{R}^{m \times n}$ 为 f_i 的 Jacobi 矩阵 (第 i 行为 $\nabla f_i(x)^\top$)。系统规模 $(n + m + p) \times (n + m + p)$ ，非对称，通常通过消元化简。

原始-对偶搜索方向：约化对称系统

消去 $\Delta\lambda$ ：由中心性方程得 $\Delta\lambda = -[\text{diag}(f(x))]^{-1}(\text{diag}(\lambda)Df(x)\Delta x + r_{\text{cent}})$ ，代入后得关于 $(\Delta x, \Delta\nu)$ 的对称 KKT 型系统：

$$\begin{bmatrix} H_{\text{pd}} & A^{\top} \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\nu \end{bmatrix} = - \begin{bmatrix} \tilde{r}_{\text{dual}} \\ r_{\text{pri}} \end{bmatrix},$$

其中 $H_{\text{pd}} = \nabla^2 f_0(x) + \sum_i \lambda_i \nabla^2 f_i(x) + \sum_i \frac{\lambda_i}{-f_i(x)} \nabla f_i(x) \nabla f_i(x)^{\top}$ 。

系统规模与结构： $(n + p) \times (n + p)$ （较原系统维数大幅降低）， $H_{\text{pd}} \succ 0$ （正定），可用带稀疏结构的 Cholesky 分解高效求解。

步长选取： 取原始步长 $t_{\text{pri}}^{\max} = \sup\{t \mid f(x + t\Delta x) < 0\}$ 和对偶步长 $t_{\text{dual}}^{\max} = \sup\{t \mid \lambda + t\Delta\lambda > 0\}$ ，再做回溯线搜索确保残差范数充分下降。

内点法软件生态

求解器 (Solvers):

- **Mosek** —— LP、QP、SOCP、SDP 的顶级商业求解器，学术免费
- **Gurobi** —— LP、QP 领先求解器
- **IPOPT** —— 大规模非线性规划（原始-对偶内点法）
- **ECOS** —— 嵌入式 SOCP 求解器
- **SCS** —— 基于 ADMM 的锥规划求解器
- **CVXOPT** —— Python 原生 SDP 求解器

建模框架:

- **CVXPY** (Python) —— 声明式建模，自动检测凸性并选择合适求解器
- **CVX** (MATLAB) —— 经典凸优化建模工具
- **JuMP** (Julia) —— 高性能代数建模
- **YALMIP** (MATLAB) —— 通用优化建模

🔗 CVXPY 示例

```
import cvxpy as cp
x = cp.Variable(n)
objective = cp.Minimize(cp.sum_squares(A @ x - b))
constraints = [cp.norm(x, 1) <= 1, x >= 0]
prob = cp.Problem(objective, constraints)
prob.solve() # 自动选择求解器
```

凸优化的应用全景

机器学习与统计

- Lasso / Ridge / Elastic Net
- 支持向量机 (SVM)
- 逻辑回归 (交叉熵极小化)
- 稀疏主成分分析
- 矩阵补全 (核范数松弛)

信号与图像处理

- 压缩感知 (ℓ_1 极小化)
- 全变差 (TV) 去噪
- 稀疏编码与字典学习
- 波束成形
- 滤波器设计 (SDP)

金融工程

- Markowitz 投资组合优化 (QP)
- 风险平价
- 条件风险价值 (CVaR) 极小化
- 期权定价 (SDP 松弛)
- 交易策略优化

通信与网络

- 多用户功率分配 (GP)
- MIMO 预编码 (SOCP/SDP)
- 网络效用最大化
- 拥塞控制算法分析

控制与工程

- 模型预测控制 (QP)
- H_∞ 鲁棒控制 (LMI/SDP)
- 结构拓扑优化
- 电路晶体管尺寸设计 (GP)

运筹与能源

- 供应链网络优化
- 机组组合与电力调度
- 交通流分配
- 最优运输

总计超过 20 个应用领域，凸优化是连接数学理论与工程实践的通用语言。

总结

理论基础

- 凸集：仿射集、锥、多面体、保凸运算
- 凸函数：一阶/二阶条件、上图、保凸运算
- 共轭函数、拟凸函数、对数凹函数
- 分离超平面与支撑超平面定理

问题类型与理论

- $LP \subset QP \subset QCQP \subset SOCP \subset SDP$
- 几何规划及其凸化变换
- Lagrange 对偶、弱/强对偶
- Slater 条件、KKT 条件（充要）

算法体系

- 梯度下降、最速下降（一阶方法）
- Newton 方法（二阶，二次收敛）
- BFGS / L-BFGS（拟牛顿）
- 等式约束 Newton / 消元法
- 障碍法与原始-对偶内点法

参考文献：Stephen Boyd & Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

上海交通大学 · 数学科学学院

机器学习

MACHINE LEARNING

第八章 · 机器学习入门

马征

按空格键继续 →

第一节 机器学习概述

Introduction to Machine Learning

什么是机器学习

◆ 机器学习 (MACHINE LEARNING)

机器学习致力于研究如何通过**计算的手段**，利用**经验**来改善系统自身的性能。在计算机系统中，「经验」通常以「数据」的形式存在，而机器学习算法的主要任务是从数据中产生「模型」(model)。

机器学习的三要素：

- **数据**：经验的载体，算法的输入
- **算法**：从数据中学习规律的方法
- **模型**：学习的结果，用于预测或决策

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。

—— Tom Mitchell, 1997

基本术语

数据集 (Dataset)：一组记录的集合。每条记录称为一个**样本** (sample) 或**示例** (instance)。

特征/属性 (Feature/Attribute)：反映样本在某方面表现或性质的量。样本在特征空间中的坐标即为**特征向量** (feature vector)。

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id}) \in \mathcal{X}$$

其中 $\mathcal{X}$ 为**输入空间** (d 维样本空间)。

标记 (Label)：关于样本结果的信息。标记空间记为 $\mathcal{Y}$ 。

样例 (Example)：拥有标记信息的样本， $(\mathbf{x}_i, y_i)$ ，其中 $y_i \in \mathcal{Y}$ 是 $\mathbf{x}_i$ 的标记。

训练集 (Training Set)：用于训练模型的数据集合

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}.$$

测试样本 (Testing Sample)：被预测的样本， $\mathbf{x}$ ，学习后的模型 f 对其预测 $y = f(\mathbf{x})$ 。

△ 术语关系总结

数据集 → 样本 (特征向量) → 属性/特征 → 属性值。**训练** (training) → 从训练集学得模型；**测试** (testing) → 用学得的模型对测试样本进行预测。

分类与回归

◆ 分类 (CLASSIFICATION)

当预测的标记 y 为**离散值**时，学习任务称为**分类**。特别地，只有两个类别时称为**二分类** (binary classification)，多个类别时称为**多分类** (multi-class classification)。

$$y \in \{\text{类别1}, \text{类别2}, \dots, \text{类别C}\}$$

◆ 回归 (REGRESSION)

当预测的标记 y 为**连续值**时，学习任务称为**回归**。

$$y \in \mathbb{R}$$

⚠ 注

分类与回归是最常见的两类监督学习任务。此外还有**聚类** (clustering) 等无监督学习任务，其训练样本的标记信息未知。

监督学习与无监督学习

◆ 监督学习 (SUPERVISED LEARNING)

训练数据**拥有标记信息**。目标是学得一个模型，能够对未见样本的标记做出准确预测。

代表任务： 分类、回归

◆ 无监督学习 (UNSUPERVISED LEARNING)

训练数据**没有标记信息**。目标是揭示数据内在的结构与规律。

代表任务： 聚类、降维、密度估计

◆ 独立同分布 (I.I.D.) 假设

通常假设所有样本服从一个未知的相同分布 $\mathcal{D}$ ，且样本之间互相独立。即样本是独立地从分布 $\mathcal{D}$ 中采样得到的：

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}.$$

这是机器学习理论分析的**基本前提**。

假设空间与归纳偏好

◆ 假设空间 (HYPOTHESIS SPACE)

学习过程实际是在一个**假设空间** $\mathcal{H}$ 中搜索，寻找与训练集「匹配」的假设 $h \in \mathcal{H}$ 。假设空间的大小和结构决定了学习器的能力。

◆ 归纳偏好 (INDUCTIVE BIAS)

机器学习算法在学习过程中对某种类型假设的**偏好**，称为归纳偏好。任何一个有效的学习算法必有其归纳偏好，否则它将被假设空间中「等效」的假设所迷惑。

◆ 奥卡姆剃刀 (OCCAM'S RAZOR)

若有多个假设与观察一致，则选择**最简单**的那个。

「如无必要，勿增实体」

奥卡姆剃刀是一种常用的归纳偏好原则，但「简单」的定义依赖于具体的问题背景。

没有免费的午餐定理

▶ 没有免费的午餐定理 (NO FREE LUNCH THEOREM)

对于二分类问题，考虑所有可能的目标函数 $f: \mathcal{X} \rightarrow \{0, 1\}$ 上的均匀分布，则对任何学习算法 $\mathcal{L}_a$ ，都有：

$$\sum_f P(h \neq f \mid \mathcal{L}_a, f) = \text{与算法无关的常数.}$$

即：所有算法在所有问题上的平均性能相同。

NFL 定理的启示：

- 脱离具体问题，空泛谈论「什么学习算法更好」毫无意义
- 算法的优劣取决于其归纳偏好是否与具体问题匹配
- 若要证明算法 A 在某些问题上优于算法 B，则必然存在另一些问题，B 优于 A

实践指导： 针对具体任务选择最合适的算法，而非寻找「万能算法」。

机器学习发展简史

1. 推理期 (1950s–1960s)

- 逻辑理论家程序 (Logic Theorist)
- 符号主义：认为逻辑推理能力代表智能
- A. Newell & H. Simon 的「通用问题求解器」

2. 知识工程期 (1970s–1980s)

- 专家系统 (Expert Systems)
- 知识库 + 推理引擎
- 瓶颈：知识获取困难 (知识工程瓶颈)

3. 连接主义复兴 (1980s–1990s)

- BP 算法的提出 (Rumelhart, 1986)
- 神经网络从低谷中复兴
- 统计学习理论 (Vapnik, 1995) 奠定 SVM 基础

4. 统计学习时代 (1990s–2010s)

- 支持向量机 (SVM) 与核方法
- 集成学习 (Bagging, Boosting, Random Forest)
- 概率图模型

5. 深度学习时代 (2012–今)

- AlexNet (2012, ImageNet) 引爆深度学习浪潮
- Transformer (2017)、大语言模型 (GPT 系列)
- 多模态、强化学习、科学智能

第二节 模型评估与选择

Model Evaluation and Selection

误差与过拟合

◆ 错误率与精度

- **错误率** (error rate): 分类错误的样本数占总样本数的比例, $E = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$
- **精度** (accuracy): 分类正确的样本数占总样本数的比例, $\text{acc} = 1 - E$

◆ 过拟合 (OVERFITTING)

模型在训练集上表现优异, 但在测试集上表现欠佳。即模型把训练样本自身的特点当作所有潜在样本的普遍规律, 导致泛化能力下降。

◆ 欠拟合 (UNDERFITTING)

模型尚未充分学习训练数据的规律, 导致在训练集和测试集上表现均不理想。通常由模型表达能力不足引起。

⚠ 注

过拟合是机器学习中的核心挑战。**模型选择**的目标是在欠拟合与过拟合之间寻找最优平衡, 使模型具有最佳泛化能力 (generalization ability)。

评估方法

① 留出法 (Hold-out)

将数据集 D 划分为两个互斥的集合：

- 训练集 S (~70%-80%)
- 测试集 T (~20%-30%)

$$D = S \cup T, \quad S \cap T = \emptyset$$

- 需多次随机划分取平均
- 保持类别比例一致 (分层采样)

② 交叉验证法 (Cross Validation)

将 D 划分为 k 个大小相似的互斥子集：

$$D = D_1 \cup D_2 \cup \dots \cup D_k$$

每次用 $k - 1$ 个子集的并集训练，余下的子集测试。 k 折交叉验证取 k 次结果的均值。

- 常用： $k = 5$ 或 $k = 10$
- 留一法 (LOO)： $k = m$ (样本数)

③ 自助法 (Bootstrapping)

从包含 m 个样本的数据集 D 中，有放回地随机采样 m 次，得到训练集 D' ：

$$P(\text{样本不在 } D' \text{ 中}) = \left(1 - \frac{1}{m}\right)^m \rightarrow \frac{1}{e} \approx 0.368$$

- D' 中约 63.2% 为独立样本 (其余重复)
- 约 36.8% 的样本未出现，用作测试集
- 适用于数据集较小的情况

第三节 线性模型

Linear Models

线性模型基本形式

◆ 线性模型 (LINEAR MODEL)

线性模型试图学得一个通过**属性的线性组合**进行预测的函数：

$$f(\mathbf{x}) = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b = \mathbf{w}^\top \mathbf{x} + b,$$

其中 $\mathbf{w} = (w_1, w_2, \dots, w_d)^\top$ 为**权重向量**， b 为**偏置** (bias)。

优点：

- 形式简单，易于建模
- **可解释性强**：权重 w_i 直接反映属性 x_i 对预测的贡献
- 是许多复杂模型的基础（如神经网络、SVM）

向量形式：

令 $\tilde{\mathbf{w}} = (\mathbf{w}, b)$ ， $\tilde{\mathbf{x}} = (\mathbf{x}, 1)$ ，则

$$f(\mathbf{x}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}.$$

即扩维后化为齐次线性形式。

线性回归

◆ 线性回归 (LINEAR REGRESSION)

给定数据集 $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ 。线性回归试图学得 $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$, 使得 $f(\mathbf{x}_i) \approx y_i$ 。

最小二乘估计 (Least Squares):

$$(\mathbf{w}^*, b^*) = \arg \min_{(\mathbf{w}, b)} \sum_{i=1}^m (f(\mathbf{x}_i) - y_i)^2.$$

闭式解: 令 $\mathbf{X} \in \mathbb{R}^{m \times (d+1)}$ 为增广数据矩阵, $\mathbf{y} = (y_1, \dots, y_m)^\top$ 。当 $\mathbf{X}^\top \mathbf{X}$ 满秩时:

$$\tilde{\mathbf{w}}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

当 $\mathbf{X}^\top \mathbf{X}$ 不可逆时, 可用正则化 (如 Ridge 回归) 或伪逆求解。

对数几率回归（逻辑回归）

◆ 对数几率回归 (LOGISTIC REGRESSION)

对于二分类问题 $y \in \{0, 1\}$ ，使用**对数几率函数**（logistic/sigmoid function）将线性模型的输出映射到 $(0, 1)$ ：

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}.$$

对数几率（log-odds）：

$$\ln \frac{P(y = 1 \mid \mathbf{x})}{P(y = 0 \mid \mathbf{x})} = \mathbf{w}^\top \mathbf{x} + b.$$

参数估计 —— 极大似然法：

$$\ell(\mathbf{w}, b) = \sum_{i=1}^m \left[y_i \ln P(y_i = 1 \mid \mathbf{x}_i) + (1 - y_i) \ln P(y_i = 0 \mid \mathbf{x}_i) \right].$$

最大化对数似然等价于最小化**交叉熵损失**（cross-entropy loss）。该优化问题可用梯度下降或 Newton 方法求解（详见凸优化章节）。

第四节 决策树

Decision Trees

决策树的基本流程

◆ 决策树 (DECISION TREE)

决策树是一种基于**树结构**进行决策的模型。从根结点开始，在每个内部结点上根据某个**属性**的取值进行判断，沿分支进入子结点，直至到达**叶结点**，叶结点给出分类或回归结果。

决策树的结构：

- **根结点** (root node)：包含全体样本
- **内部结点** (internal node)：对应于一个属性测试
- **叶结点** (leaf node)：对应于决策结果
- **分支** (branch)：对应于属性测试的一个可能取值

核心思想 —— 分治法 (Divide and Conquer)：

自根至叶的递归过程，在每个结点上选择最优划分属性，将样本分配到子结点中，直到满足停止条件：

1. 当前结点样本全属同一类别
2. 当前属性集为空（或样本在所有属性上取值相同）
3. 当前结点样本集为空

信息增益与划分选择

◆ 信息熵 (INFORMATION ENTROPY)

设样本集 D 中第 k 类样本的比例为 p_k ($k = 1, 2, \dots, |\mathcal{Y}|$), 则 D 的**信息熵**为:

$$\text{Ent}(D) = - \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k.$$

$\text{Ent}(D)$ 越小, D 的**纯度**越高。当 $p_k \in \{0, 1\}$ 时, $\text{Ent}(D) = 0$ (纯度最高)。

◆ 信息增益 (INFORMATION GAIN)

属性 a 对样本集 D 的**信息增益**为划分前后熵的减少:

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v),$$

其中 V 为属性 a 的可能取值数, D^v 为 D 中在属性 a 上取值为 a^v 的样本子集。**ID3 算法**选择使信息增益最大的属性进行划分。

其他划分准则

◆ 增益率 (GAIN RATIO, C4.5)

信息增益对可取值数目较多的属性有**偏好**。增益率在增益基础上除以属性的固有价值 (intrinsic value) 以减少此偏好：

$$\text{Gain_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)},$$

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}.$$

C4.5 算法先从候选属性中选出信息增益高于平均水平的属性，再从中选增益率最高者。

◆ 基尼指数 (GINI INDEX, CART)

基尼值度量数据集 D 的纯度：

$$\text{Gini}(D) = \sum_{k=1}^{|Y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|Y|} p_k^2.$$

$\text{Gini}(D)$ 越小，纯度越高。属性 a 的**基尼指数**：

$$\text{Gini_index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v).$$

CART 算法选择使划分后基尼指数最小的属性。

算法	划分准则	树结构	任务类型
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 (分类) / 平方误差 (回归)	二叉树	分类/回归

剪枝

◆ 剪枝 (PRUNING)

剪枝是缓解决策树**过拟合**的主要手段。通过删除一些分支来降低模型复杂度，提高泛化能力。

预剪枝 (Pre-pruning)

在决策树生成过程中，对每个结点在划分前进行估计：

- 若当前结点的划分**不能**带来泛化性能的提升，则停止划分并将当前结点标记为叶结点
- 优点：降低训练时间开销
- 缺点：存在**欠拟合**风险（基于贪心本质，可能阻止后续有价值的划分）

后剪枝 (Post-pruning)

先从训练集生成一棵完整的决策树，然后**自底向上**对非叶结点进行考察：

- 若将该结点对应的子树替换为叶结点能提升泛化性能，则进行剪枝
- 优点：欠拟合风险小，泛化性能通常优于预剪枝
- 缺点：训练时间开销较大（需先完全生成再修剪）

第五节 神经网络

Neural Networks

M-P 神经元模型

◆ M-P 神经元模型 (MCCULLOCH-PITTS NEURON)

神经元接收来自 n 个其他神经元的输入信号 $x_1, x_2, \dots, x_n$ ，通过**带权重的连接**传递，将总输入与**阈值**（threshold）比较，通过**激活函数**（activation function）产生输出：

$$y = \varphi \left(\sum_{i=1}^n w_i x_i - \theta \right),$$

其中 w_i 为连接权重， θ 为阈值， $\varphi(\cdot)$ 为激活函数。

常见激活函数：

函数	表达式	输出范围
阶跃函数	$\varphi(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$	$\{0, 1\}$
Sigmoid	$\sigma(z) = \frac{1}{1 + e^{-z}}$	$(0, 1)$
tanh	$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$(-1, 1)$
ReLU	$\varphi(z) = \max(0, z)$	$[0, +\infty)$

引入偏置的统一形式：

令 $b = -\theta$ （偏置），则：

$$y = \varphi(\mathbf{w}^T \mathbf{x} + b).$$

这等价于将阈值 θ 视为一个固定输入 -1 所对应的连接权重，从而可与权重的学习统一处理。

感知机

◆ 感知机 (PERCEPTRON)

感知机由两层神经元组成：**输入层**接收外界信号后传递给**输出层**。输出层为一个 M-P 神经元，使用阶跃函数作为激活函数。

$$y = \text{sgn}(\mathbf{w}^T \mathbf{x} - \theta).$$

几何意义：

感知机在特征空间中寻找一个**超平面** $\mathbf{w}^T \mathbf{x} + b = 0$ ，将正负样本分开。因此感知机只能解决**线性可分**问题。

局限性 (Minsky & Papert, 1969)：

单层感知机甚至无法解决简单的****异或 (XOR) ****问题，因为它不是线性可分的。这一发现导致了神经网络研究的第一次「寒冬」。

◆ 感知机学习规则

$$w_i \leftarrow w_i + \eta (y - \hat{y}) x_i,$$

其中 $\eta > 0$ 为**学习率** (learning rate)， y 为真实标记， $\hat{y}$ 为预测值。每次仅根据当前样本的误差调整权重。

多层前馈网络

◆ 多层前馈神经网络 (MULTI-LAYER FEEDFORWARD NEURAL NETWORK)

由**输入层**、**若干隐藏层**、**输出层**组成，每层神经元与下一层神经元**全互连**，神经元之间不存在同层连接和跨层连接（前馈结构）。

网络的表示：

设网络共有 L 层（输入层为第 0 层）。第 l 层有 s_l 个神经元。第 $l-1$ 层到第 l 层的权重矩阵为 $\mathbf{W}^{(l)} \in \mathbb{R}^{s_l \times s_{l-1}}$ ，偏置向量为 $\mathbf{b}^{(l)} \in \mathbb{R}^{s_l}$ 。

$$\mathbf{a}^{(l)} = \varphi^{(l)} \left(\mathbf{W}^{(l)} \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \right), \quad l = 1, 2, \dots, L.$$

其中 $\mathbf{a}^{(0)} = \mathbf{x}$ （输入）， $\mathbf{a}^{(L)} = \hat{\mathbf{y}}$ （输出）。

⚠ 通用逼近定理 (UNIVERSAL APPROXIMATION THEOREM)

只需一个包含足够多神经元的隐藏层，多层前馈网络可以以**任意精度**逼近任意复杂的连续函数。但这并不保证学习算法能找到这样的参数。

误差反向传播算法 (BP)

◆ BP 算法 (BACKPROPAGATION)

BP 算法是训练多层前馈网络的经典算法，包含两个阶段：

1. **前向传播**：输入信号逐层传递，产生输出
2. **反向传播**：误差从输出层反向传回，更新连接权重与阈值

均方误差目标函数：

$$E = \frac{1}{2} \sum_{j=1}^{s_L} (\hat{y}_j - y_j)^2.$$

梯度下降更新规则：

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial E}{\partial w_{ij}^{(l)}}.$$

链式法则计算梯度： $\frac{\partial E}{\partial w_{ij}^{(l)}} = \frac{\partial E}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}}$ ，其中 $z_j^{(l)} = \sum_k w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)}$ 。

BP 算法关键步骤：

1. 前向计算各层输出 $\mathbf{a}^{(l)}$
2. 计算输出层误差： $\delta_j^{(L)} = \varphi'(z_j^{(L)}) \cdot (\hat{y}_j - y_j)$
3. 反向传播误差： $\delta_j^{(l)} = \varphi'(z_j^{(l)}) \cdot \sum_{k=1}^{s_{l+1}} w_{kj}^{(l+1)} \delta_k^{(l+1)}$
4. 计算梯度并更新参数

总参数数量： 对于 d 输入、 q 隐藏、 l 输出的单隐藏层网络，共有 $(d+1)q + (q+1)l$ 个参数。

神经网络的训练技巧

学习率与早停

- 学习率 η : 太大则震荡甚至发散; 太小则收敛过慢
- 早停 (Early Stopping): 在验证集误差开始上升时停止训练, 防止过拟合
- 常用策略: 自适应学习率 (Adam, RMSprop 等)

逃离局部极小

由于目标函数高度非凸, BP 容易陷入局部极小:

- 多组不同初始参数, 选取误差最小者
- 模拟退火 (simulated annealing)
- 随机梯度下降 (SGD) 引入的随机性有助跳出
- 动量 (momentum) 方法帮助越过鞍点

⚠ 深度学习中的改进

BP 算法的基本思想仍是现代深度学习的核心。现代实践中常用的改进包括:

- Mini-batch SGD: 每次使用一小批样本更新
- 自适应优化器: Adam, RMSprop, AdaGrad 等
- 批归一化 (Batch Normalization): 加速训练、稳定收敛
- Dropout: 随机丢弃神经元抑制过拟合

其他常见神经网络

RBF 网络 (Radial Basis Function Network)

- 单隐藏层前馈网络
- 隐藏层使用径向基函数作为激活函数: $\varphi_i(\mathbf{x}) = e^{-\beta_i \|\mathbf{x} - \mathbf{c}_i\|^2}$
- 训练分两步: 先确定 RBF 中心 $\mathbf{c}_i$ (可用聚类), 再训练输出层权重 w_i

ART 网络 (Adaptive Resonance Theory)

- 无监督学习网络
- 由比较层和识别层组成
- 「胜者通吃」(winner-take-all) 竞争机制
- 通过识别阈值和重置模块实现稳定性-可塑性平衡

SOM 网络 (Self-Organizing Map)

- 无监督竞争学习网络
- 将高维输入映射到低维 (通常 2D) 输出层
- 保持输入数据的拓扑结构
- 邻近的神经元响应相似的输入

RNN 与 Boltzmann 机

- RNN (循环神经网络): 神经元存在环型连接, 状态随时间演化, 适合序列数据
- Boltzmann 机: 基于能量的随机神经网络, 包含可见层和隐藏层, 使用随机激活函数 $P(s_i = 1) = \sigma(\sum_j w_{ij}s_j - \theta_i)$

第六节 深度学习

Deep Learning

从神经网络到深度学习

◆ 深度学习 (DEEP LEARNING)

深度学习是**具有多个隐藏层**的神经网络。通过增加**深度**（层数）而非仅增加**宽度**（每层神经元数），使模型具备更强的表示学习能力，能够逐层提取从底层到高层的特征。

模型容量与训练难度：

- 深度模型容量（capacity）大，可以拟合复杂的函数
- 训练难度随深度急剧增加（梯度消失/爆炸）
- 需要**大数据**和**高性能计算**（GPU/TPU）
- 参数越多，过拟合风险越高，但大数据 + 正则化可以有效缓解

****权重共享****
Weight Sharing
(CNN 等)

****跳跃连接****
Skip Connections
(ResNet 等)

****注意力机制****
Attention
(Transformer 等)

深度学习的核心挑战与解决方案

梯度消失/爆炸 (Vanishing/Exploding Gradient)

深层网络中，梯度在反向传播中呈指数级衰减或增长：

$$\frac{\partial E}{\partial w^{(l)}} \propto \prod_{k=l+1}^L \mathbf{w}^{(k)} \cdot \varphi'(z^{(k)}).$$

解决方案：

- ReLU 激活函数 (缓解梯度消失)
- 批归一化 (Batch Normalization)
- 残差连接 (ResNet, 2015)
- 合适的权重初始化 (Xavier, He 初始化)

过拟合与泛化

深度网络参数量极大，容易过拟合训练数据。

正则化策略：

- **Dropout**: 训练时随机丢弃神经元
- **数据增强** (Data Augmentation): 扩充训练集
- **权重衰减** (Weight Decay): ℓ_2 正则化
- **早停** (Early Stopping)
- **迁移学习** (Transfer Learning): 在大规模数据上预训练

△ 表示学习 (REPRESENTATION LEARNING)

深度学习的核心优势是**自动学习特征表示**。浅层学习通常依赖人工设计特征 (feature engineering)，而深度学习通过逐层抽象，从数据中自动学习层次化表示——底层学到边缘/纹理等简单特征，高层学到语义概念等抽象特征。这彻底改变了计算机视觉、自然语言处理等领域的研究范式。

第七节 支持向量机

Support Vector Machines

间隔与支持向量

◆ 支持向量机 (SVM) 的基本思想

对于二分类问题, SVM 的目标是找到一个**超平面** $\mathbf{w}^T \mathbf{x} + b = 0$, 使得:

1. 正确划分两类样本
2. 两类样本到超平面的**最小距离 (间隔) 最大化**

函数间隔与几何间隔:

函数间隔 (functional margin):

$$\hat{\gamma}_i = y_i(\mathbf{w}^T \mathbf{x}_i + b).$$

几何间隔 (geometric margin):

$$\gamma_i = \frac{y_i(\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|}.$$

几何间隔消除了 $\mathbf{w}$ 缩放的影响, 是样本点到超平面的真实有向距离。

支持向量 (Support Vectors):

距离超平面最近的训练样本称为**支持向量**。它们满足:

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1.$$

△ 稀疏性

SVM 的解仅由**支持向量**决定, 与其他样本无关。这一性质使 SVM 具有天然的稀疏性, 在内存和计算上具有优势。

SVM 的优化形式

◆ 硬间隔 SVM 的原问题

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, m. \end{aligned}$$

这是一个凸二次规划 (convex QP) 问题。

◆ 软间隔 SVM

当数据**线性不可分**时，引入松弛变量 $\xi_i \geq 0$ ，允许一些样本不满足约束：

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

其中 $C > 0$ 为**惩罚参数**，平衡间隔最大化与误分类惩罚。

对偶问题与核技巧

◆ SVM 的对偶问题

通过 Lagrange 对偶性，SVM 原问题的对偶形式为：

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m. \end{aligned}$$

求解后得 $\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$ ，决策函数 $f(\mathbf{x}) = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$ 。

◆ 核函数 (KERNEL FUNCTION)

当原始空间中数据**非线性可分**时，通过映射 $\phi: \mathcal{X} \rightarrow \mathcal{F}$ 将数据映射到高维特征空间，在 $\mathcal{F}$ 中使用线性 SVM。**核技巧**允许在原始空间中直接计算内积：

常见核函数与 SVM 总结

常见核函数：

核名称	表达式
线性核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
多项式核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + c)^d$
高斯/RBF 核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\ \mathbf{x}_i - \mathbf{x}_j\ ^2}{2\sigma^2}\right)$
Sigmoid 核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i^T \mathbf{x}_j + \theta)$

核函数有效的条件：Mercer 定理——核矩阵 $\mathbf{K} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{m \times m}$ 必须是半正定的。

SVM 的核心优势：

- 凸优化**：目标函数是凸二次函数，全局最优解有保证
- 核技巧**：隐式映射到高维空间，避免显式计算 ϕ
- 稀疏性**：仅支持向量参与决策
- 泛化理论**：最大化间隔等价于最小化 VC 维上界
- 对偶形式**：可以使用 SMO (Sequential Minimal Optimization) 等高效算法

△ SVM 与凸优化的联系

SVM 的优化问题是一个典型的**凸二次规划**（见第七章凸优化），KKT 条件是求解对偶问题的理论基础。在现代深度学习中，SVM 的**最大间隔**思想也被推广到神经网络（如 Hinge Loss、Large Margin Softmax 等）。

总结

基础概念

- 数据集、样本、特征
- 标记、样例
- 分类与回归
- 监督与无监督学习
- 独立同分布 (i.i.d.)
- 训练 / 测试

理论基础

- 假设空间
- 归纳偏好
- 奥卡姆剃刀
- 没有免费的午餐 (NFL)
- 过拟合与欠拟合
- 偏差-方差权衡

经典模型

- 线性回归 (最小二乘)
- 逻辑回归 (交叉熵)
- 决策树 (ID3, C4.5, CART)
- 信息增益 / 增益率 / 基尼指数
- 剪枝 (预剪枝 & 后剪枝)

神经网络与前沿

- M-P 神经元模型
- 感知机 → 多层前馈网络
- BP 算法 / 梯度下降
- RBF, ART, SOM, RNN
- 深度学习
- 支持向量机 (SVM)

与凸优化的联系： 逻辑回归的交叉熵极小化、SVM 的对偶 QP 问题、BP 算法中的梯度下降，均属于/来源于凸优化理论体系。

模型评估： 留出法 → 交叉验证 → 自助法，为模型选择提供可靠的泛化性能估计。

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.