

标题 title

作者 *author*

2024 年 8 月 30 日

前言

目录

前言	i
第一部分 AI 的逻辑	1
第一章 合情推理	2
§1.1 命题逻辑的演绎推理	3
§1.2 合情推理的数学模型	8
§1.2.1 合情推理的基本假设, 似然	9
§1.2.2 似然与概率	12
§1.2.3 先验与基率谬误	14
§1.3 合情推理的归纳强论证	15
§1.3.1 归纳强论证	15
§1.3.2 有效论证和归纳强论证的比较	18
§1.4 先验模型的存在性	21
§1.5 章末注记	23
§1.6 习题	23
第二章 Markov 链与决策	24
§2.1 Markov 链	24
§2.2 Markov 奖励过程 (MRP)	32
§2.3 Markov 决策过程 (MDP)	36
§2.4 隐 Markov 模型 (HMM)	43
§2.4.1 评估问题	45
§2.4.2 解释问题	46
§2.5 扩散模型	48

§2.5.1 采样逆向过程	51
§2.5.2 训练逆向过程	52
§2.6 章末注记	54
§2.7 习题	54
 第二部分 信息与数据	 55
第三章 熵与 Kullback-Leibler 散度	56
§3.1 熵	56
§3.1.1 概念的导出	56
§3.1.2 概念与性质	60
§3.2 Kullback-Leibler 散度	66
§3.2.1 定义	66
§3.2.2 两个关于信息的不等式	67
§3.3 编码理论	68
§3.3.1 熵与编码	68
§3.3.2 K-L 散度、交叉熵与编码	70
§3.4 在机器学习中的应用：语言生成模型	72
§3.5 附录：Shannon 定理的证明	73
§3.6 习题	75
§3.7 章末注记	77
 第四章 高维几何， Johnson-Lindenstrauss 引理	 78
§4.1 高维几何	79
§4.1.1 高维球体	79
§4.1.2 Stein 悖论	82
§4.1.3 为什么我们要正则化？远有潜龙，勿用	86
§4.2 集中不等式	87
§4.3 J-L 引理的陈述与证明	91
§4.4 J-L 引理的应用	95
§4.5 附录：Stein 悖论的证明	97
§4.6 习题	97
§4.7 章末注记	97

第五章 差分隐私	98
§5.1 数据隐私问题	99
§5.2 差分隐私的定义与性质	101
§5.3 差分隐私的应用	107
§5.3.1 随机反应算法	107
§5.3.2 全局灵敏度与 Laplace 机制	108
§5.3.3 DP 版本 Llyod 算法	111
§5.4 习题	113
§5.5 章末注记	113
 第三部分 决策与优化	 114
第六章 凸分析	115
§6.1 决策与优化的基本原理	116
§6.1.1 统计决策理论	116
§6.1.2 优化问题	118
§6.1.3 例子：网格搜索算法	122
§6.2 凸函数	124
§6.3 凸集	128
§6.3.1 基本定义和性质	129
§6.3.2 分离超平面定理	132
§6.4 习题	133
§6.5 章末注记	133
 第七章 对偶理论	 128
§7.1 条件极值与 Lagrange 乘子法	129
§7.2 Karush–Kuhn–Tucker 条件	132
§7.3 Lagrange 对偶	135
§7.3.1 Lagrange 定理	135
§7.3.2 弱对偶定理，强对偶定理	139
§7.4 应用：支持向量机 (SVM)	143
 第八章 不动点理论	 146
§8.1 Banach 不动点定理	146

§8.2 Brouwer 不动点定理	154
§8.3 习题	158
§8.4 章末注记	158
第四部分 逻辑与博弈	159
第九章 动态博弈	154
§9.1 输赢博弈	154
§9.2 随机博弈 (Markov 博弈)	159
第十章 静态博弈	165
§10.1 正则形式博弈	165
§10.1.1 生成对抗网络	166
§10.1.2 混合策略	168
§10.2 不完全信息博弈 (Bayes 博弈)	169
第五部分 认知逻辑	174
第十一章 模态逻辑基础	175
§11.1 模态逻辑的起源	175
§11.1.1 三段论	175
§11.1.2 非经典逻辑	176
§11.2 模态语言	177
§11.3 Kripke 语义与框架语义	181
§11.4 模态可定义性	185
第十二章 认知逻辑与共同知识	188
§12.1 “泥泞的孩童”谜题	188
§12.2 认知逻辑的基本模型与性质	190
§12.2.1 “泥泞的孩童”再回顾	194
§12.2.2 Aumann 结构	195
§12.3 对不一致达成一致	196
§12.4 Rubinstein 电子邮件博弈	199

第六部分 附录：预备知识	203
附录 A 线性代数基础	204
§A.1 线性空间	204
§A.2 线性映射	208
§A.3 矩阵	213
§A.4 双线性型与二次型	219
§A.5 带内积的线性空间	223
§A.6 行列式	229
§A.7 算子范数与谱理论	232
附录 B 微分学基础	238
§B.1 点集拓扑	238
§B.1.1 度量空间，范数	238
§B.1.2 开集与闭集	241
§B.1.3 紧致性，收敛性，完备性	244
§B.1.4 连续映射	247
§B.1.5 与实数序有关的性质	250
§B.2 一元函数的微分学	252
§B.2.1 导数与微分的定义	253
§B.2.2 微分学基本定理	256
§B.3 多元函数的微分学	258
§B.3.1 微分、偏导数与导数的定义	258
§B.3.2 微分学基本定理	264
§B.3.3 隐函数定理	266
附录 C 概率论基础	270
§C.1 从朴素概率论到公理化概率论	270
§C.1.1 Kolmogorov 概率论	270
§C.1.2 条件概率，独立性	274
§C.2 随机变量，分布函数	278
§C.2.1 基本定义	278
§C.2.2 离散型随机变量	282
§C.2.3 连续型随机变量	282

§C.2.4 随机向量, 条件分布, 独立性	286
§C.2.5 随机变量 (向量) 的函数	290
§C.3 随机变量的数字特征, 条件数学期望	293
§C.3.1 数学期望, Lebesgue 积分	293
§C.3.2 数学期望的性质	297
§C.3.3 随机变量的内积空间	300
§C.3.4 特征函数	302
§C.3.5 条件数学期望	303
§C.4 多元正态分布 (Gauss 向量)	307

第一部分

AI 的逻辑

第二部分

信息与数据

第三部分

决策与优化

第八章 不动点理论

如果有一个长满毛发的球体，你能够把它所有的毛发都梳理平顺吗？做个实验就会发现，这好像是做不到的，总会有一根毛发直立不倒，或某个地方没有毛发覆盖。实际上，早在 1912 年，Brouwer 就从数学上严格证明了上述现象，我们现在称之为毛球定理。

你是否在大型商场或者公园里经常看到“您在此处”的地图标识牌？为什么可以有这样的标识，它真的表明了你的位置吗？

你是否相信，地球上有两个地方，它们分别位于地球的对径点，并且温度和湿度完全相同？

这些问题看似毫无关联，但它们都有一个共同的数学背景：不动点理论。

不动点的定义是非常直接的，考虑一个集合 X 以及它到自身的映射 $f: X \rightarrow X$ ，元素 $a \in X$ 称为映射 $f: X \rightarrow X$ 的不动点，如果 $f(a) = a$ 。

除了生活中，不动点理论对于优化来说也是非常重要的。考虑优化算法 A ，它在函数 f 上的收敛性如何？算法运行所产生的点列记为 $\{x_n\}$ ，它满足

$$x_{n+1} = A(x_n).$$

如果关注序列 x_n 本身，要分析收敛性，我们需要通过寻找不同量之间的联系，比如 $f(x_n)$ 和 $f(x_{n+1})$ 之间的关系。在数学中，这样的思路被归类到了数学分析中。

一种更为抽象的做法是，我们直接看算法 A 本身的性质。此时，要想说明 A 收敛，我们要说明 A 有一个“吸收点”，即不管从何处出发，经过若干次迭代，都会收敛到这个点附近。这样的思路是更加现代的数学方法，它被归类到了算子法和泛函分析中。

我们将看到，从算子的角度来理解收敛性，最终问题就归结到了不动点理论。本章将介绍两种不动点存在性定理，并介绍他们的应用。

§8.1 Banach 不动点定理

首先，我们需要引入一些度量空间相关的概念，更系统的讨论请参阅附录 B。

定义 8.1 (度量与度量空间) 集合 X 上的度量 (或距离) d 是映射

$$d : X \times X \rightarrow \mathbb{R}$$

满足条件

- 非负性: $d(x_1, x_2) \geq 0$, 并且 $d(x_1, x_2) = 0 \iff x_1 = x_2$.
- 对称性: $d(x_1, x_2) = d(x_2, x_1)$.
- 三角不等式: $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$.

其中 x_1, x_2, x_3 是 X 的任意元素.

此时, (X, d) 或 X 被称为度量空间.

度量是一个非常直观的概念. 实际上, 它就是 Euclid 空间中“距离”概念的抽象化.

下面, 我们不加证明地给出一些度量的例子, 他们的证明见习题[\[lhy: 出一下\]](#).

例 8.1 考虑实数集 \mathbb{R} , 要成为度量空间, 可以装备以下度量:

- 平凡的离散度量: $\forall x_1 \neq x_2 \ d(x_1, x_2) \equiv 1, d(x, x) = 0$.
- $d(x_1, x_2) = |x_1 - x_2|$.

这一例子告诉我们, 尽管我们熟悉的绝对值度量是最常见的度量, 但实数也可以具备其他度量.

考虑向量空间 \mathbb{R}^n , 要成为度量空间, 可以装备以下度量:

- Minkowski 度量 (L^p 度量):

$$d(x_1, x_2) = \left(\sum_{i=1}^n |x_1^i - x_2^i|^p \right)^{1/p} \quad (p \geq 1).$$

- Manhattan 度量 (L^1 度量):

$$d(x_1, x_2) = \sum_{i=1}^n |x_1^i - x_2^i|.$$

- Euclid 度量 (L^2 度量):

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^n |x_1^i - x_2^i|^2}.$$

• Chebyshev 度量 (L^∞ 度量):

$$d(x_1, x_2) = \max_i |x_1^i - x_2^i| = \lim_{p \rightarrow \infty} \left(\sum_{i=1}^n |x_1^i - x_2^i|^p \right)^{1/p}. \quad \square$$

我们的目标是找到一类和实数集非常像的度量空间. 实数集一个非常重要的性质是实数列收敛当且仅当它是 Cauchy 列. 我们把这一性质抽象出来, 就得到了如下定义:

定义 8.2 (Cauchy 列, 完备度量空间) 考虑度量空间 (X, d) 的点列 $\{x_n\}_{n \in \mathbb{N}}$, 如果对于任何 $\epsilon > 0$, 都可以找到序号 $N \in \mathbb{N}$, 使得对于任何大于 N 的序号 $m, n \in \mathbb{N}$,

$$d(x_m, x_n) < \epsilon,$$

那么我们称 $\{x_n\}$ 是 **Cauchy 列**.

如果度量空间 (X, d) 的任意 Cauchy 列 $\{x_n\}_{n \in \mathbb{N}}$ 都收敛, 即存在点 $a \in X$, 使得

$$\lim_{n \rightarrow \infty} d(a, x_n) = 0,$$

那么, 我们称度量空间 (X, d) 是**完备的**,

为了理解 Cauchy 列的含义, 我们先要理解序列的收敛性 (也就是极限). 一列实数 a_n 有极限 a , 指的是对任何 $\epsilon > 0$, 都可以找到序号 $N \in \mathbb{N}$, 使得对于任何大于 N 的序号 $n \in \mathbb{N}$,

$$|a_n - a| < \epsilon.$$

更直观一些的说法是, 不论给多小的精度, 除了有限项, a_n 都可以以这一精度逼近 a .

而 Cauchy 列描述了另一种形式的收敛性, 此时, 我们虽然不知道 a_n 离哪个实数比较近, 但是我们知道除了有限项, a_n 相互之间的差异都会小于这个精度. 直观上, 这说明 a_n 在靠近某个东西, 也就是收敛.

完备性这一概念就是说, 这两个收敛性的定义是等价的, 因此 a_n 的确是在靠近某个东西. 我们将它写作定理的形式:

定理 8.1 设 (X, d) 是一个完备度量空间, 对任意序列 $\{x_n\}_{n \in \mathbb{N}}$, 以下两个条件等价:

- $\{x_n\}$ 是 Cauchy 列.
- $\{x_n\}$ 收敛.

证明. 我们只需要证明收敛序列是 Cauchy 列. 设 $\{x_n\}$ 收敛到 a , 即对任意 $\epsilon > 0$, 存在 $N \in \mathbb{N}$, 使得对于任意 $n > N$, 有

$$d(x_n, a) < \epsilon/2.$$

于是对于任意 $m, n > N$, 有

$$d(x_m, x_n) \leq d(x_m, a) + d(a, x_n) < \epsilon/2 + \epsilon/2 = \epsilon.$$

因此 $\{x_n\}$ 是 Cauchy 列. □

下面, 我们不加证明地给出一些完备度量空间的例子, 证明见习题[\[lhy: 出一下\]](#).

例 8.2 • L^p 度量下 \mathbb{R}^n 是完备的.

- 使用度量 $d(x_1, x_2) = |x_1 - x_2|$, 则 $X = \mathbb{R} \setminus \{0\}$ 不是完备度量空间. 考虑

$$\left\{ x_n = \frac{1}{n} \right\}_{n \in \mathbb{N}},$$

它是 Cauchy 列, 但该点列在 X 中没有极限 (极限是 0).

- $[0, 1]$ 到自身的连续函数空间 $C([0, 1])$ 在 L^∞ 度量下是完备的. 此时

$$d(f, g) = \sup_{x \in [0, 1]} |f(x) - g(x)|.$$

□

特别注意最后一个例子, 我们这里给出了一类抽象的度量空间: 它的元素是函数. 这类空间是泛函分析中最主要的研究对象, 它关注的不再是函数局部的性质, 而是整体上研究函数之间的关系.

有了度量的概念, 我们就可以研究两个度量空间之间映射的性质: 连续性.

定义 8.3 (连续映射) 设 X 和 Y 是度量空间 $(X, d_X), (Y, d_Y)$, 考虑映射 $f: X \rightarrow Y$ 个点 $a \in X$, 如果对于任意 $\epsilon > 0$, 存在 $\delta > 0$, 使得对于任意 $x \in X$, 有

$$d_X(a, x) < \delta \Rightarrow d_Y(f(a), f(x)) < \epsilon,$$

那么我们称 f 在点 a 是连续的.

如果 f 在每个点 $x \in X$ 连续, 则称 f 为连续映射.

连续映射的定义也是非常直观的, 它的意思是, 如果 x 和 y 很接近, 那么 $f(x)$ 和 $f(y)$ 也应该很接近, 说明 f 变化得非常小.

下面我们给出与 Banach 不动点定理相关的概念:

定义 8.4 (压缩映射) 考虑度量空间 (X, d) 到自身的映射 $f: X \rightarrow X$. 如果存在 $q \in (0, 1)$, 使得 X 中的任何两个点 x_1, x_2 都成立不等式

$$d(f(x_1), f(x_2)) \leq q \cdot d(x_1, x_2),$$

那么我们称 f 是一个压缩映射.

压缩映射也是一个非常直观的概念, 它的意思是, 映射 f 的每次作用都会按照某个比例 q 缩小任意两点之间的距离. 比如, 考虑点 x_0 和 $f(x_0)$, 当压缩次数足够多之后, 两点之间的距离就会趋于零, 也就是

$$\underbrace{f(f(f(\cdots f(x_0) \cdots)))}_{n\text{次}} \approx \underbrace{f(f(\cdots f(x_0) \cdots))}_{n\text{次}}.$$

这就是压缩映射具有不动点的原因. 下面我们来严格证明这一点.

首先, 我们说明, 证明压缩映射一定是连续映射:

引理 8.1 压缩映射 $f: X \rightarrow X$ 是连续映射.

证明. 对于任意 $\epsilon > 0$, 取 $\delta = \epsilon/q$, 则对于任意 $x_1, x_2 \in X$, 有

$$d(x_1, x_2) < \delta \implies d(f(x_1), f(x_2)) \leq qd(x_1, x_2) < \epsilon.$$

因此 f 是连续的. □

接下来, 我们说明, 度量本身也是一个连续映射:

引理 8.2 度量 $d: X \times X \rightarrow \mathbb{R}$ 是连续映射.

证明. 对于任意 $x_1, x_2, y_1, y_2 \in X$, 有

$$|d(x_1, y_1) - d(x_2, y_2)| \leq d(x_1, y_1) + d(x_2, y_2) \leq 2 \max\{d(x_1, x_2), d(y_1, y_2)\}.$$

因此, 对于任意 $\epsilon > 0$, 取 $\delta = \epsilon/2$, 则对于任意 $x_1, x_2, y_1, y_2 \in X$, 有

$$d(x_1, x_2) < \delta, d(y_1, y_2) < \delta \implies |d(x_1, y_1) - d(x_2, y_2)| < \epsilon.$$

因此 d 是连续映射. □

接下来, 我们证明压缩映射一定有不动点, 这就是 Banach 不动点定理:

定理 8.2 (Banach 不动点定理, 压缩映像原理) 完备度量空间 (X, d) 到自身的压缩映射 $f: X \rightarrow X$ 具有唯一的不动点 a .

此外, 对于任何点 $x_0 \in X$, 迭代序列 $x_0, x_1 = f(x_0), \dots, x_{n+1} = f(x_n), \dots$ 收敛到 a . 收敛速度由以下估计给出:

$$d(a, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0).$$

证明. 首先证明存在性. $d(x_{n+1}, x_n) \leq qd(x_n, x_{n-1}) \leq \dots \leq q^n d(x_1, x_0)$. 从而

$$\begin{aligned} d(x_{n+k}, x_n) &\leq d(x_n, x_{n+1}) + \dots + d(x_{n+k-1}, x_{n+k}) \\ &\leq (q^n + \dots + q^{n+k-1})d(x_1, x_0) \leq \frac{q^n}{1-q} d(x_1, x_0). \end{aligned}$$

这一不等式对任意 k 都成立, 而因此 $\{x_n\}$ 是 Cauchy 列, 根据完备性的定义存在极限

$$\lim_{n \rightarrow \infty} x_n = a \in X.$$

结合压缩映射的连续性, 有

$$a = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} f(x_n) = f\left(\lim_{n \rightarrow \infty} x_n\right) = f(a).$$

然后证明唯一性. 若 f 还有其他不动点 a_1, a_2 , 则

$$0 \leq d(a_1, a_2) = d(f(a_1), f(a_2)) \leq qd(a_1, a_2).$$

而这当且仅当 $d(a_1, a_2) = 0$, 即 $a_1 = a_2$ 时才可能成立.

最后证明收敛速度. 对

$$d(x_{n+k}, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0),$$

取 $k \rightarrow \infty$, 根据 d 的连续性, 有

$$d(a, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0). \quad \square$$

在进入应用之前, 我们指出压缩映射在算子法中的表述, 这一部分的系统讨论需要线性代数的知识, 请参阅附录 A. 我们这里只做一个简单介绍.

首先, 如果我们把压缩映射 f 看成一个算子 \mathcal{A} , 即把 X 中的元素变换到 X 中的元素, 那么我们可以定义这一算子的范数:

定义 8.5 (算子范数) 设 $X = \mathbb{R}^n$, 对于算子 $\mathcal{A}: X \rightarrow X$, 它的范数定义为

$$\|\mathcal{A}\| = \sup_{x \neq 0} \frac{\|\mathcal{A}x\|}{\|x\|}.$$

其中 $\|\cdot\|$ 是 X 上的 L^2 范数, 即

$$\|x\| = \sqrt{\sum_{i=1}^n |x_i|^2}.$$

在这一概念下, 我们可以改写压缩映射的定义. 对任意 $x, y \in \mathbb{R}^n$, 有

$$\|\mathcal{A}x - \mathcal{A}y\| \leq q \|x - y\| \implies \frac{\|\mathcal{A}(x - y)\|}{\|x - y\|} \leq q.$$

根据 x, y 的任意性, 这其实就是说

$$\|\mathcal{A}\| \leq q < 1.$$

所以, 压缩映射其实就是算子范数小于 1 的算子.

反之, 如果一个算子 \mathcal{A} 的范数 q 小于 1, 那么对任意 $x, y \in \mathbb{R}^n$, 有

$$\|\mathcal{A}x - \mathcal{A}y\| \leq \|\mathcal{A}\| \|x - y\| \leq q \|x - y\|.$$

因此, \mathcal{A} 是一个压缩映射. 我们将这一讨论总结如下:

定理 8.3 设 $X = \mathbb{R}^n$, 对于算子 $\mathcal{A}: X \rightarrow X$, 以下两个条件等价:

- \mathcal{A} 是压缩映射.
- $\|\mathcal{A}\| < 1$.

对于很多算子, 直接验证压缩映射的定义比较困难, 而验证算子范数小于 1 则相对容易. 因此这是一个特别实用的表述方式.

例 8.3 (落在地面上的地图) 将一座公园的地图铺开在公园地面上, 则地面上恰有唯一一点与地图上对应的点重合.

设公园可以用有界的面闭区域 Ω 表示. 设地图的压缩比是 $\lambda \in (0, 1)$. 现在固定一个平面直角坐标系, 把地图铺在区域 Ω 内, 则从 Ω 内的点 x (公园中的地点) 到地图上对应点 x' 的变换由下面的公式给出:

$$x' = f(x) := \lambda R x + b.$$

其中 R 和 b 分别为旋转和平移变换.

根据旋转的定义, 容易看出 $\|Rx\| = \|x\|$, 因此

$$\|\lambda R\| = \sup_{\|x\|=1} \|\lambda Rx\| = \lambda < 1,$$

所以对任意 $x, y \in \Omega$, 有

$$\|f(x) - f(y)\| = \|\lambda Rx - \lambda Ry\| = \lambda \|Rx - Ry\| = \lambda \|x - y\|.$$

因此 f 是一个压缩映射.

由 Banach 不动点定理可知, 压缩映射 $f(x)$ 有唯一不动点 $a = f(a)$. \square

例 8.4 (梯度下降的收敛性) 这个例子研究如何利用算子法证明梯度下降的收敛性. 它需要较多微积分和线性代数的知识, 请参阅附录 B 和附录 A. 不过, 理解整个思路并不需要这些知识.

我们优化目标是寻找二阶可微凸函数 $f(x), x \in \mathbb{R}^n$ 的最小值. 使用梯度下降方法:

$$x_{k+1} = x_k - \alpha_k f'(x_k),$$

其中 α_k 是第 k 步的步长, 在这个例子中, 我们假设 $\alpha_k = \alpha$ 是一个常数.

接下来, 我们给出对 f 的假设: 存在常数 $L > 0$, 对任意 $x \in \mathbb{R}^n$,

$$\lambda_{\min}(\nabla^2 f(x)) \geq L,$$

其中

- $\nabla^2 f(x)$ 是 f 的 Hessian 矩阵 (二次导数),
- $\lambda_{\min}(A)$ 表示矩阵 A 的最小特征值.

我们要证明: 对于足够小的 α , 梯度下降能收敛到最小值点, 且具有指数收敛速度.

先看一下证明的思路. 定义梯度下降算子:

$$\mathcal{T}^{(\alpha)} : x \mapsto x - \alpha \nabla f(x).$$

我们要设法证明梯度下降算法是完备度量空间中的一个压缩映射.

1. 首先, 根据??, 可微凸函数 f 的最小值点充分必要地满足

$$\nabla f(x) = 0.$$

2. 其次, 显然有

$$\nabla f(x^*) = 0 \iff \mathcal{T}^{(\alpha)} x^* = x^*.$$

因而最小值点是梯度下降算子的不动点.

3. 所以, 我们只需要说明 $\mathcal{T}^{(\alpha)}$ 是一个完备度量空间的压缩映射, 就可以用 Banach 不动点定理证明梯度下降的收敛性.

我们只需要证明 $\mathcal{T}^{(\alpha)}$ 是压缩映射, 并给出压缩系数. 由有限增量原理 (定理 B.14):

$$\|\mathcal{T}^{(\alpha)} x - \mathcal{T}^{(\alpha)} y\| \leq \sup_{z \in (x, y)} \|I - \alpha \nabla^2 f(z)\|_2 \cdot \|x - y\|_2.$$

注意到 $\|I - \alpha \nabla^2 f(z)\|_2$ 等于 $I - \alpha \nabla^2 f(z)$ 特征值的最大模, 根据条件可知特征值的最大模 $\leq 1 - L\alpha$. 因此, 只要 $\alpha < L^{-1}$, $\mathcal{T}^{(\alpha)}$ 就是一个压缩映射. \square

§8.2 Brouwer 不动点定理

下面我们考虑另一类不动点定理. 在 Banach 不动点定理中, 我们对映射的性质做出了限制. 在这一部分, 我们只要求映射是连续的, 但是对映射所在的集合做出了限制. 因此, 我们下面不加解释地给出几个技术性的概念, 更系统的讨论请参阅附录 B.

定义 8.6 (开集、闭集和紧集) 考虑度量空间 (X, d) , 定义 $a \in X$ 的邻域为

$$B(a, \delta) := \{x \in X \mid d(a, x) < \delta\}.$$

考虑一个集合 $K \subseteq X$,

- 如果对任意 $x \in G$, 都存在邻域 $B(x, \delta) \subseteq G$, 那么 G 是开集.
- 如果 $X \setminus G$ 是开集, 那么 G 是闭集.
- 如果对任何开集族 $\{G_\alpha\}$, 只要满足

$$K \subseteq \bigcup_{\alpha} G_{\alpha},$$

就存在 $G_{\alpha_1}, \dots, G_{\alpha_n}$ 使得

$$K \subseteq G_{\alpha_1} \cup \dots \cup G_{\alpha_n},$$

那么 K 是紧集. 换言之, 如果任何可以覆盖 K 的开集族都有一个有限子族可以覆盖 K , 那么 K 是紧集.

在 Euclid 空间中，我们有如下性质：

定理 8.4 考虑集合 $K \subseteq \mathbb{R}^n$ ，以下两个定义等价：

- K 是紧集.
- K 是有界闭集.

这里，有界的意思就是，存在一个半径 R ，使得 K 包含在半径为 R 的球内.

注意，定理 8.4 只在 \mathbb{R}^n 中成立，对于一般的度量空间，紧集和有界闭集不一定等价（见习题[thy: 出一下]）.

有了上面的准备，我们就可以叙述 Brouwer 不动点定理了：

定理 8.5 (Brouwer 不动点定理) 设 $M \subseteq \mathbb{R}^n$ 是一个非空紧凸集，而 $F: M \rightarrow M$ 是一个连续函数. 则存在 $x \in M$ 使得 $F(x) = x$ 成立.

Brouwer 不动点定理可以通过该实际的例子来理解：将一张白纸平铺在桌面上，再将它揉成一团（不撕裂），放在原来白纸所在的地方，那么只要它不超出原来白纸平铺时的边界，那么白纸上一定有一点在水平方向上没有移动过. 这个断言依据 Brouwer 不动点定理在 \mathbb{R}^2 的情况，因为把纸揉皱是一个连续的变换过程.

另一个例子：大商场等地方可以看到的平面地图，上面标有“您在此处”的红点. 如果标注足够精确，那么这个点就是把实际地形映射到地图的连续函数的不动点.

下面我们看一个 Brouwer 不动点定理的应用例子，这一例子需要线性代数和 Markov 链的知识，请参阅附录 A 和第二章.

首先引入矩阵不可约的概念：

定义 8.7 (不可约矩阵) 考虑方阵 A ，定义操作 O_{ij} ：

- 将 A 的第 i 列和第 j 列交换，
- 同时将 A 的第 i 行和第 j 行交换.

如果经过有限次操作 O_{ij} （不同的 i, j ）后， A 变成分块上三角矩阵，那么 A 是可约的；否则， A 是不可约的.

下面我们来解释不可约矩阵在 Markov 链中的含义. 设 A 是某个 Markov 链的转移矩阵，假如 A 可约，通过行列交换的方法变成了分块上三角矩阵：

$$\begin{pmatrix} A_{11} & A_{12} \\ O & A_{22} \end{pmatrix},$$

设前半对应的状态集是 S_1 ，后半对应的状态集是 S_2 ，那么，这一转移矩阵的形式意味着，从 S_2 的任意状态出发，达到 S_1 的任意状态的概率都是 0。因此，这个 Markov 链的流动性是比较差的。

反之，如果 A 是不可约的，那么，不论从哪个状态出发，经过有限次转移，都可以到达任何一个状态。所以，这一 Markov 链的流动性是比较好的。

接下来，我们说明，如果 Markov 链不可约（也就是流动性很好），它会有一个平稳遍历分布（即所有状态都是正概率）。这个结论由以下定理给出：

定理 8.6 (Perron-Frobenius 定理) 设 $A = (a_{ij})$ 为 $n \times n$ 不可约实矩阵，所有元素均非负， $a_{ij} \geq 0$ ，则下列结论成立。

- 存在一个实特征值 r ，其他（左右）特征值 λ 的模均不超过 r ，即 $|\lambda| \leq r$ 。
- 存在一个与 r 对应的左特征向量和右特征向量，其所有元素恒正。
- $\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}$ 。

在开始证明之前，我们先说明它如何导出 Markov 链的性质。

推论 8.1 不可约有限状态 Markov 链必然存在平稳遍历分布。换言之，如果 P 是一个不可约有限状态 Markov 链的转移矩阵，那么存在一个分布 π ，使得 $\pi = \pi P$ 并且对任意 i 都有 $\pi_i > 0$ 。

证明。 根据定义， P 是非负实不可约方阵。由 Perron-Frobenius 定理， P 存在一个特征值 r 使得

$$1 = \min_i \sum_j P_{ij} \leq r \leq \max_i \sum_j P_{ij} = 1, \quad \square$$

即 $r = 1$ ，并且，它对应一个正的左特征向量

$$\pi_0 \in \left\{ x \in \mathbb{R}^n \mid x \geq 0, \sum_i x_i = 1 \right\}.$$

因此，

$$\pi_0 P = \pi_0.$$

即 π_0 是平稳遍历分布。 □

接下来，我们证明定理 8.6。

证明. (定理 8.6 的证明) 首先证明 A 存在一个正的特征值 $r > 0$. 考虑单纯形

$$S := \left\{ x \in \mathbb{R}^n \mid x \geq 0, \sum_i x_i = 1 \right\}.$$

任意 $x \in S$, 有 $Ax \geq 0$.

我们断言 $Ax > 0$. 若不然, A 存在某一列全 0 (由 $x \geq 0$ 和 A 非负可得). 此时可将该 0 列交换到第一列, 对应的行也交换, 得到的矩阵为分块上三角, 与不可约性矛盾.

可以在 S 上定义映射

$$T(x) = \frac{1}{\rho(x)} Ax,$$

其中 $\rho(x) > 0$ 使得 $T(x) \in S$. 具体来说,

$$\rho(x) = \sum_i (Ax)_i = \sum_{i,j} a_{ij} x_j.$$

显然 $T(x)$ 是 $S \rightarrow S$ 的连续映射. S 是一个有界凸闭集. 由 Brouwer 不动点定理, 存在 $x_0 \in S$ 使得

$$x_0 = T(x_0) = \frac{1}{\rho(x_0)} Ax_0.$$

令 $r = \rho(x_0)$, 则可得 r 为 A 的一个正的特征值.

我们接下来证明, 与 r 对应的右特征向量所有元素恒正. 由之前的证明, 与 r 对应的特征向量 $x_0 \in S$, 则 $x_0 \geq 0$. 我们证明 $x_0 > 0$.

我们将 A 的行列进行交换, 使得 Ax_0 非零的元素在上方. 具体来说, 设 $A = PBP^{-1}$, 其中 P 是置换矩阵, 则

$$PBP^{-1}x_0 = rx_0 \implies B(P^{-1}x_0) = r(P^{-1}x_0).$$

记 $\tilde{x}_0 = P^{-1}x_0$. 取 B 使得 $\tilde{x}_0 = (\xi, 0)^\top, \xi > 0$. 则

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \xi \\ 0 \end{pmatrix} = \begin{pmatrix} r\xi \\ 0 \end{pmatrix}.$$

此时 $B_{21}\xi = 0$, 由 $\xi > 0$ 可得 $B_{21} = 0$. 这与不可约矛盾, 因此 $x_0 > 0$.

以上过程可以对左特征值 r_1 和对应的左特征向量 x_1 重复, 得到 $r_1 > 0$ 且 $x_1 > 0$.

然后我们证明: 若 λ 是 A 的任意右特征值, 有 $|\lambda| \leq r$.

设 $0 \leq B \leq A$, 也就是 $0 \leq B_{ij} \leq A_{ij}$, 则 B 的特征值 β 和对应的特征向量 y 满足

$$|\beta| \leq r, \quad By = \beta y.$$

记 $y^* = |y| = (|y_i|)_i$. 于是有

$$|\beta|y^* = |\beta y| = |By| \leq By^* \leq Ay^*.$$

左乘 x_1^T , 有

$$|\beta|x_1^T y^* \leq x_1^T A y^* = r_1 x_1^T y^*.$$

由 $x_1^T y^* > 0$ 可得 $|\beta| \leq r_1$.

令 $B = A$ 可得 $|\lambda| \leq r_1$, 特别地 $r \leq r_1$.

如果 λ 是左特征值, 用同样的证明可以得到 $|\lambda| \leq r$, 特别地 $r_1 \leq r$.

综合以上两点, $r = r_1$, 于是我们说明了 x_0 和 x_1 是与 r 对应的左右特征向量, 并且其他左右特征值的模都不超过 r .

最后证明:

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}.$$

以这样的方式获得 \tilde{A} : 将 A 的每一行都扩增 (不减小某个元素), 使得每一行都达到 $\max_i \sum_j a_{ij}$. 此时 $\max_i \sum_j a_{ij}$ 成为 A 的一个正特征值, 且有右特征向量

$$\tilde{x}_0 = \frac{1}{n} \cdot \mathbf{1} \in S.$$

由之前的证明, 根据 $0 \leq A \leq \tilde{A}$, 可以得到 \tilde{A} 的正特征值 $\tilde{r} \geq r$. 因此

$$r \leq \max_i \sum_j a_{ij}.$$

同理缩小 A 可得

$$\min_i \sum_j a_{ij} \leq r.$$

□

§8.3 习题

[lhy: TODO]

§8.4 章末注记

[lhy: TODO]

第四部分

逻辑与博弈

第五部分

认知逻辑

第六部分

附录：预备知识

参考文献

- [Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.
- [CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.
- [Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.
- [Inf] Information | Etymology, origin and meaning of information by etymonline. <https://www.etymonline.com/word/information>.
- [Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.
- [KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [LLG⁺19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.
- [McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.
- [RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

- [Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.
- [Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.
- [Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.
- [Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.
- [Uff22] Jos Uffink. Boltzmann’s Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.
- [李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引

Cauchy 列, 148

Perron-Frobenius 定理, 156

不动点, 146

不动点定理

 Banach \sim , 151

 Brouwer \sim , 155

不动点理论, 146

不可约矩阵, 155

压缩映射, 150

压缩映射原理, 151

完备空间, 148

度量, 147

$L^1 \sim$, 147

$L^2 \sim$, 147

$L^\infty \sim$, 148

$L^p \sim$, 147

 Chebyshev \sim , 148

 Euclid \sim , 147

 Manhattan \sim , 147

 Minkowski \sim , 147

 离散 \sim , 147

 绝对值 \sim , 147

度量空间, 147

开集, 154

数学分析, 146

梯度下降, 153

步长, 153

毛球定理, 146

泛函分析, 146, 149

算子法, 146

紧集, 154

范数

 算子 \sim , 152

距离, 147

连续, 149

连续映射, 149

闭集, 154