

标题 title

作者 *author*

2023 年 8 月 23 日

前言

目录

前言	i
第一部分 科学的逻辑	1
第一章 合情推理	2
§1.1 回顾：命题逻辑的演绎推理	2
§1.2 合情推理的数学模型	4
1.2.1 似然，合情推理的原则	4
1.2.2 似然与概率	6
§1.3 合情推理的归纳强论证	8
1.3.1 先验与基率谬误	8
1.3.2 归纳强论证	9
1.3.3 有效论证和归纳强论证的比较	12
第二章 Markov 链与决策	15
§2.1 Markov 链	15
§2.2 Markov 奖励过程 (MRP)	19
§2.3 Markov 决策过程 (MDP)	22
§2.4 隐 Markov 模型 (HMM)	26
2.4.1 评估问题	27
2.4.2 解释问题	28
第二部分 信息与数据	30
第三章 信息论基础	31

§3.1 熵	31
3.1.1 概念的导出	31
3.1.2 概念与性质	34
3.1.3 熵与通信理论	39
§3.2 Kullback-Leibler 散度	42
3.2.1 定义	42
3.2.2 两个关于信息的不等式	44
3.2.3 在机器学习中的应用：语言生成模型	45
§3.3 附录：Shannon 定理的证明	46
§3.4 习题	47
§3.5 章末注记	49
第四章 Johnson-Lindenstrauss 引理	51
§4.1 机器学习中的数据	51
§4.2 矩法与集中不等式	52
§4.3 J-L 引理的陈述与证明	56
§4.4 J-L 引理的应用	60
§4.5 习题	61
§4.6 章末注记	61
第五章 差分隐私	62
§5.1 数据隐私问题	62
§5.2 差分隐私的定义与性质	64
§5.3 差分隐私的应用	68
5.3.1 随机反应算法	68
5.3.2 全局灵敏度与 Laplace 机制	69
5.3.3 DP 版本 Llyod 算法	71
§5.4 差分隐私与信息论	72
§5.5 习题	73
§5.6 章末注记	73
第三部分 决策与优化	74
第六章 凸分析	75

§6.1 决策与优化的基本原理	75
6.1.1 统计决策理论	75
6.1.2 优化问题	76
6.1.3 例子：网格搜索算法	79
§6.2 凸函数	81
§6.3 凸集	84
6.3.1 基本定义和性质	84
6.3.2 分离超平面定理	86
第七章 对偶理论	88
§7.1 条件极值与 Lagrange 乘子法	89
§7.2 Karush–Kuhn–Tucker 条件	92
§7.3 Lagrange 对偶	95
7.3.1 Lagrange 定理	95
7.3.2 弱对偶定理，强对偶定理	99
§7.4 应用：支持向量机 (SVM)	103
第八章 不动点理论	106
§8.1 Banach 不动点定理	106
§8.2 Brouwer 不动点定理	109
§8.3 不动点的一般视角	112
第四部分 逻辑与博弈	113
第九章 动态博弈	114
§9.1 输赢博弈	114
§9.2 随机博弈 (Markov 博弈)	119
第十章 静态博弈	125
§10.1 正则形式博弈	125
10.1.1 生成对抗网络	126
10.1.2 混合策略	128
§10.2 不完全信息博弈 (Bayes 博弈)	129

第五部分 认知逻辑	134
第十一章 模态逻辑基础	135
§11.1 模态逻辑的起源	135
11.1.1 三段论	135
11.1.2 非经典逻辑	136
§11.2 模态语言	137
§11.3 Kripke 语义与框架语义	140
§11.4 模态可定义性	145
第十二章 认知逻辑与共同知识	147
§12.1 “泥泞的孩童”谜题	147
§12.2 认知逻辑的基本模型与性质	149
12.2.1 “泥泞的孩童”再回顾	153
12.2.2 Aumann 结构	154
§12.3 对不一致达成一致	155
§12.4 Rubinstein 电子邮件博弈	158
附录 A 线性代数基础	162
§A.1 线性空间	162
§A.2 线性映射	166
§A.3 矩阵	171
§A.4 双线性型与二次型	176
§A.5 带内积的线性空间	180
§A.6 行列式	185
§A.7 算子范数与谱理论	187
附录 B 微分学基础	193
§B.1 点集拓扑	193
B.1.1 度量空间, 范数	193
B.1.2 开集与闭集	196
B.1.3 紧集, 收敛性, 完备性	199
B.1.4 连续映射	202
§B.2 一元函数的微分学	205
§B.3 多元函数的微分学	205

§B.4 线性赋范空间的微分学，矩阵微分学	205
附录 C 概率统计基础	194

第一部分

科学的逻辑

第二部分

信息与数据

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

附录 B 微分学基础

本书中的积分学使用非常少，并且集中在概率论部分，所以在本附录中我们只讨论微分学，积分学的内容在概率论中简单介绍。尽管我们的视角非常一般且抽象，我们主要讨论的是 Euclid 空间 \mathbb{R}^n 相关的微分学。

§B.1 点集拓扑

本部分讨论极限、连续、紧致等概念，这些概念是微分学的基础。

B.1.1 度量空间，范数

实数集 \mathbb{R} 上面的元素可以被看成一些点，这些点之间有距离的概念。这是 \mathbb{R} 最重要的几个性质之一。我们把这种性质抽象出来，得到度量空间的概念。

定义 B.1 (度量空间) 设 X 是一个集合， $d: X \times X \rightarrow \mathbb{R}$ 是一个函数，如果满足

1. 非负性：对任意 $x, y \in X$ ， $d(x, y) \geq 0$ ，且 $d(x, y) = 0$ 当且仅当 $x = y$ ；
2. 对称性：对任意 $x, y \in X$ ， $d(x, y) = d(y, x)$ ；
3. 三角不等式：对任意 $x, y, z \in X$ ， $d(x, z) \leq d(x, y) + d(y, z)$ 。

则称 (X, d) 是一个度量空间， d 称为度量。

下面给出一些度量的例子，但我们不给出验证。

例 B.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： $\forall x_1 \neq x_2 \ d(x_1, x_2) \equiv 1, d(x, x) = 0$ 。
- $d(x_1, x_2) = |x_1 - x_2|$ 。

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) : $d(x_1, x_2) = (\sum_{i=1}^n |x_1^i - x_2^i|^p)^{1/p}$ ($p \geq 1$).
- *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) : $d(x_1, x_2) = \max_i |x_1^i - x_2^i| = \lim_{p \rightarrow \infty} (\sum_{i=1}^n |x_1^i - x_2^i|^p)^{1/p}$.

再看一个抽象的例子. 假设 (X, d_X) 和 (Y, d_Y) 是两个度量空间，我们可以定义 $X \times Y$ 上的度量 d 为

$$d((x_1, y_1), (x_2, y_2)) = d_{\mathbb{R}^2}(0, (d_X(x_1, x_2), d_Y(y_1, y_2))).$$

其中 $d_{\mathbb{R}^2}$ 为 \mathbb{R}^2 上的某个度量. 容易验证这这也是一个度量.

上面关于 \mathbb{R}^n 的例子都有一个特点，他们都是用向量 $x_1 - x_2$ 某种意义上的长度定义的，这种长度的概念在数学中有一个统一的抽象，就是范数的概念.

定义 B.2 (范数, 赋范空间) 设 X 是一个向量空间， $\|\cdot\| : X \rightarrow \mathbb{R}$ 是一个函数，如果满足

1. 非负性与非退化：对任意 $x \in X$ ， $\|x\| \geq 0$ ，且 $\|x\| = 0$ 当且仅当 $x = 0$ ；
2. 齐次性：对任意 $x \in X$ ， $\lambda \in \mathbb{R}$ ， $\|\lambda x\| = |\lambda| \|x\|$ ；
3. 三角不等式：对任意 $x, y \in X$ ， $\|x + y\| \leq \|x\| + \|y\|$.

则称 $\|\cdot\|$ 是 X 上的一个范数， $(X, \|\cdot\|)$ 称为一个赋范空间.

我们不进行验证，但是指出，例 B.1 中的度量都自然地导出了一个范数，即 $\|x\| = d(x, 0)$. 我们可以自然地称呼这些范数，例如 L^p 范数就是 L^p 度量所有诱导的范数. 实际上，很多无穷维线性空间都是先有范数才有空间本身的. 例如， ℓ^p 实际上就是由 L^p 范数划定的：

$$\ell^p = \left\{ x \in \mathbb{C}^\infty : \|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p} < \infty \right\}.$$

此外，函数空间 $C[a, b]$ 也可以定义范数，例如

$$\|f\|_\infty = \sup_{x \in [a, b]} |f(x)|.$$

反之，任何一个范数都可以导出一个度量，即 $d(x, y) = \|x - y\|$. 这一结论可以总结为如下性质：

命题 B.1 设 X 是一个向量空间, $\|\cdot\|$ 是 X 上的一个范数, 则 $d(x, y) = \|x - y\|$ 是 X 上的一个度量, 称之为范数诱导的度量. 反之, 如果 d 是 X 上的一个度量, 则 $\|x\| = d(x, 0)$ 是 X 上的一个范数当且仅当对任意 $x, y, z \in X, \lambda \in \mathbb{R}$, 有

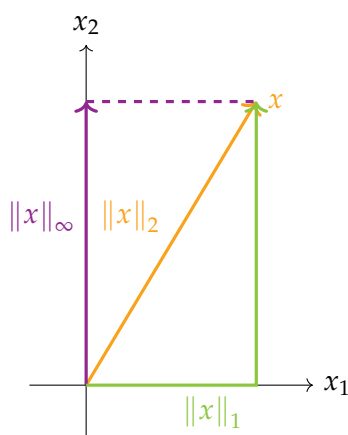
1. 平移不变性: $d(x + z, y + z) = d(x, y)$;
2. 相似性: $d(\lambda x, \lambda y) = |\lambda|d(x, y)$.

尽管都是 \mathbb{R}^n , 但是不同的 p 对应的 L^p 范数大小是不一样的. 他们之间有如下的关系:

命题 B.2 设 $1 \leq p \leq q \leq \infty$, 则对任意 $x \in \mathbb{R}^n$, 有

$$\|x\|_p \geq \|x\|_q.$$

这一命题的证明依赖于 Hölder 不等式, 这里不给出细节了. 要想对这一不等式有更好的直观, 我们可以考虑 $n = 2$ 时 $p = 1, 2, \infty$ 的极端情形, 如下图所示, 想象我们要从原点到点 x . 绿色的是 $\|x\|_1 = |x_1| + |x_2|$, 相当于沿着坐标轴走; 而橙色的是 $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$, 相当于沿着对角线走, 肯定比沿着坐标轴走要快; 紫色的是 $\|x\|_\infty = \max\{|x_1|, |x_2|\}$, 相当于挑了较长的那条边走, 仿佛虫洞一样, 走完了就到了, 所以甚至比对角线还快.



然而, 从拓扑学的角度来说, 这些度量并没有本质的区别, 这是因为:

命题 B.3 设 $1 \leq p \leq q \leq \infty$, 则存在正常数 $c_{p,q}$ 和 $C_{p,q}$, 对任意 $x, y \in \mathbb{R}^n$,

$$c_{p,q} \|x\|_q \leq \|x\|_p \leq C_{p,q} \|x\|_q.$$

这一证明也依赖于 Hölder 不等式，所以也略去。这一命题说明，虽然不同的范数对应的度量不同，但是他们之间的关系是最多差个常数倍。我们后面会看到，这一性质表明 L^p 范数定义的所有拓扑性质都是完全相同的。这一性质也可以一般化：

定义 B.3 (等价范数) 设 X 是一个向量空间， $\|\cdot\|_1$ 和 $\|\cdot\|_2$ 是 X 上的两个范数，如果存在正常数 c, C ，使得对任意 $x \in X$ ，有

$$c \|x\|_1 \leq \|x\|_2 \leq C \|x\|_1,$$

则称 $\|\cdot\|_1$ 和 $\|\cdot\|_2$ 是等价的。

B.1.2 开集与闭集

接下来我们进一步进行讨论 \mathbb{R}^n 空间的拓扑性质。拓扑学是关于开集的学问，给定所有的开集，我们就可以研究一个空间的拓扑性质。

在 \mathbb{R} 中，很早就已经有了开区间的概念，它指的是集合 $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ 。实际上， \mathbb{R} 中的开集定义很简单，就是若干开区间的并集。在更一般的拓扑空间中，开集的定义也是如此。我们将视角聚焦在度量空间中。我们可以把开区间 (a, b) 看成一个圆心在 $(a + b)/2$ ，半径为 $(b - a)/2$ 的一维开球。从这个视角看，开集的定义是从开球给出的。这样的定义是有一般性的：

定义 B.4 (开球，开集，拓扑空间) 设 (X, d) 是一个度量空间， $x \in X$ ， $r > 0$ ，定义

$$B(x, r) = \{y \in X : d(x, y) < r\}.$$

则称 $B(x, r)$ 是以 x 为球心， r 为半径的开球。

集合 $U \subseteq X$ 被称为开集，如果它是若干开球的并集。

X 连同它的所有开集，被称为拓扑空间¹。

在通常的微积分教科书上，我们会看到另一种开集的定义，即开集是任意一点都可以找到一个开球包含在这个集合中。这两种定义是等价的：

命题 B.4 设 (X, d) 是一个度量空间， $U \subseteq X$ ，则 U 是开集当且仅当对任意 $x \in U$ ，存在 $r > 0$ ，使得 $B(x, r) \subseteq U$ 。

¹一般拓扑空间的定义是给出所有开集的集合，并要求他们满足某种封闭性，然而我们这里只关心度量空间，所以不具体给出这些封闭性条件了。

证明 \Rightarrow : 设 U 是开集, $x \in U$, $U = \bigcup_{i \in I} B(x_i, r_i)$, 则存在 $i \in I$, 使得 $x \in B(x_i, r_i)$, 取 $r = r_i - d(x, x_i)$, 显然 $r > 0$, 并且 $B(x, r) \subseteq B(x_i, r_i) \subseteq U$.

\Leftarrow : 设对任意 $x \in U$, 存在 $r_x > 0$, 使得 $B(x, r_x) \subseteq U$, 则 $U = \bigcup_{x \in U} B(x, r_x)$, 是开集. \square

我们给的定义是一个更拓扑、更整体的定义：开集就是由基本的开集（开球）经过任意次的并得到的集合，这一定义关心的集合而不是具体的元素。而等价的定义，我们称之为点定义，是更局部的定义，这一定义关心的是点而不是集合。今后的定义，我们都尝试用两种方式给出，特别地，拓扑的定义只使用开集而不使用度量。

我们给几个开集的例子：

例 B.2 (范等价拓扑空间) 设 X 是一个线性空间，它上面有两个等价的范数 $\|\cdot\|_1$ 和 $\|\cdot\|_2$ ²，则两个赋范空间 $(X, \|\cdot\|_1)$ 和 $(X, \|\cdot\|_2)$ 定义了相同的拓扑空间。因此，在拓扑意义下， \mathbb{R}^n 空间到底装备了哪个 L^p 范数是不重要的，因此对于同一个数学对象（集合、序列、函数）来说，收敛性、完备性以及连续性在 L^p 范数下都是完全一样的。

事实上，设 U 是 $(X, \|\cdot\|_1)$ 中的开集， $x \in U$ ，则存在 $r > 0$ ，使得 $B_1(x, r) \subseteq U$ ，由范数等价，存在 $c, C > 0$ ，使得 $c\|x\|_2 \leq \|x\|_1 \leq C\|x\|_2$ ，则 $B_2(x, r/c) \subseteq B_1(x, r) \subseteq U$ ，所以 U 是 $(X, \|\cdot\|_2)$ 中的开集。反之亦然。

例 B.3 (乘积拓扑空间) 设 (X_1, d_1) 和 (X_2, d_2) 是两个度量空间，则 $X_1 \times X_2$ 上的开集有两种方式给出：

1. 规定 $X_1 \times X_2$ 上的度量 d ，然后利用这个度量给出开集；
2. 对任意开集 $U_1 \subseteq X_1$ 和 $U_2 \subseteq X_2$ ，定义 $U_1 \times U_2$ ，则 $U_1 \times U_2$ 是 $X_1 \times X_2$ 上的开集，然后利用这些基本的开集给出所有开集。

如果我们把度量 d 定义为

$$d((x_1, y_1), (x_2, y_2)) = \|(d_1(x_1, x_2), d_2(y_1, y_2))\|.$$

其中 $\|\cdot\|$ 是 \mathbb{R}^2 的某个 L^p 范数，可以证明，这两种方式给出的 $X_1 \times X_2$ 上的拓扑完全相同。因此，以后出现带有“拓扑空间 $X \times Y$ ”这暗示的地方，所指的拓扑空间都是由这两种等价方式给出的。这一结论可以推广到任意有限个度量空间的乘积。

开集的重要性质是：

²注意，对一般空间来说，这样的记号不意味着 L^1 或者 L^2 范数。

命题 B.5 设 (X, d) 是一个非空度量空间, 则

1. X 和 \emptyset 是开集;
2. 任意个开集的并集是开集;
3. 有限个开集的交集是开集.

证明 1. 取 $x \in X$, 则 $X = \bigcup_{r>0} B(x, r)$, 是开集. \emptyset 是零个 (也是若干个) 开集的并集, 是开集.

2. 设 $\{U_i\}_{i \in I}$ 是一族开集, $U_i = \bigcup_{j \in J_i} B(x_j, r_j)$, 显然 $U = \bigcup_{i \in I} U_i = \bigcup_{i \in I, j \in J_i} B(x_j, r_j)$, 是开集.

3. 设 U_1, \dots, U_n 是开集, $U = \bigcap_{i=1}^n U_i$, 对任意 $x \in U$, 对任意 $i = 1, \dots, n$, $x \in U_i$, 由开集的点定义, 存在 $r_i > 0$, 使得 $B(x, r_i) \subseteq U_i$, 取 $r = \min_{i=1}^n r_i$, 则 $B(x, r) \subseteq U_i$, 所以 U 是开集. \square

注意, 开集只对有限交封闭. 可以看一个简单的例子: $\bigcap_{n=1}^{\infty} (-1/n, 1/n) = \{0\}$, 但是 $\{0\}$ 不是开集, 因为这个集合不可能包含任何开球.

与开集相对应的是闭集的概念. 闭集的定义是:

定义 B.5 (闭集) 设 (X, d) 是一个度量空间, $F \subseteq X$, 如果 $X \setminus F$ 是开集, 则称 F 是闭集.

闭集的定义是开集的对偶, 所以有如下性质:

命题 B.6 设 (X, d) 是一个非空度量空间, 则

1. X 和 \emptyset 是闭集;
2. 任意个闭集的交集是闭集;
3. 有限个闭集的并集是闭集.

需要注意的是, 开集似乎可以简单理解为开区间的推广, 然而闭集完全不是这样的, 闭集是把若干开区间挖出来得到的集合, 并不是闭区间的简单推广, 所以比起把开区间拼起来会复杂得多, 例如 Cantor 集就是一个性质非常奇怪的闭集.

B.1.3 紧集，收敛性，完备性

接下来我们讨论一个更微妙的概念，紧集. 紧性与极限、收敛、连续等概念有着密切的联系，然而如何恰当的定义紧性是一个很难的问题. 我们这里不讨论历史，只给出历史的答案. 简单来说，紧这个词的概念是压缩，将无穷多的东西变成有限个. 我们的逻辑推理通常只能处理有限的东西，所以紧性是沟通无穷和有限的桥梁. 下面给出紧集的定义：

定义 B.6 (开覆盖，紧集) 设 (X, d) 是一个度量空间， $F \subseteq X$ ，如果存在一族开集 $\{U_i\}_{i \in I}$ ，使得 $F \subseteq \bigcup_{i \in I} U_i$ ，则称 $\{U_i\}_{i \in I}$ 是 F 的一个开覆盖.

如果对任意 F 的开覆盖 $\{U_i\}_{i \in I}$ ，都存在有限子覆盖 $\{U_{i_j}\}_{j=1}^n$ ，使得 $F \subseteq \bigcup_{j=1}^n U_{i_j}$ ，则称 F 是紧集.

这当然是一个非常抽象的定义. 然而，我们没有办法将它还原为更直观的定义了. 例如，即便在最基本的集合 \mathbb{R} 上，紧集的存在性也只能被作为与实数公理³等价的命题存在：

命题 B.7 (Heine-Borel 有限覆盖原理) 设 F 是 \mathbb{R} 的一个闭区间，对任意 F 开覆盖 $\{U_i\}_{i \in I}$ ，存在有限子覆盖 $\{U_{i_j}\}_{j=1}^n$.

这一原理说明，闭区间是紧集，因而给出了紧集的存在性.

在度量空间上，紧集与收敛性密切相关. 为此，我们需要形式地定义度量空间中的收敛概念. 我们先使用 $\epsilon - N$ 语言定义：

定义 B.7 (收敛，极限) 设 (X, d) 是一个度量空间， $\{x_n\}_{n=1}^{\infty}$ 是 X 中的一个序列， $x \in X$ ，如果对任意 $\epsilon > 0$ ，存在 $N \in \mathbb{N}$ ，使得对任意 $n > N$ ， $d(x_n, x) < \epsilon$ ，则称 $\{x_n\}_{n=1}^{\infty}$ 收敛到 x ，记作 $\lim_{n \rightarrow \infty} x_n = x$ 或 $x_n \rightarrow x$ ， $n \rightarrow \infty$ ， x 称为 $\{x_n\}_{n=1}^{\infty}$ 的极限.

直观上说，这一定义在刻画一系列点越来越接近某个点 x . 如果我们将定义中的 N 取掉，这一直观会更清楚：对任意 $\epsilon > 0$ ，除掉有限个 n （也就是前 N 个），都有 $x_n \in B(x, \epsilon)$. 所谓越来越接近，指的就是画任意一个球 $B(x, \epsilon)$ ，除去有限个 x_n ，剩下的所有 x_n 都在这个球里面. 这一想法给出了只基于开集的等价定义：

命题 B.8 设 (X, d) 是一个度量空间， $\{x_n\}_{n=1}^{\infty}$ 是 X 中的一个序列， $x \in X$ ，则 $\{x_n\}_{n=1}^{\infty}$ 收敛到 x 当且仅当对任意包含 x 的开集 U ，存在 $N \in \mathbb{N}$ ，使得对任意 $n > N$ ， $x_n \in U$.

³当然，这样的说法把实数集作为一个数学对象，试图用公理定义出来，而不是从已有的数学对象构造出来（例如 Dedekind 分割）.

证明 \Rightarrow : 设 $\{x_n\}_{n=1}^{\infty}$ 收敛到 x , U 是包含 x 的开集, 由开集的点定义, 存在 $r > 0$, 使得 $B(x, r) \subseteq U$, 由收敛的定义, 存在 $N \in \mathbb{N}$, 使得对任意 $n > N$, $d(x_n, x) < r$, 所以 $x_n \in B(x, r) \subseteq U$.

\Leftarrow : 设对任意包含 x 的开集 U , 存在 $N \in \mathbb{N}$, 使得对任意 $n > N$, $x_n \in U$, 则对任意 $\epsilon > 0$, 取 $U = B(x, \epsilon)$, 则存在 $N \in \mathbb{N}$, 使得对任意 $n > N$, $x_n \in B(x, \epsilon)$, 即 $d(x_n, x) < \epsilon$, 所以 $\{x_n\}_{n=1}^{\infty}$ 收敛到 x . \square

在更一般的拓扑空间中, 甚至没有度量的概念, 然而, 开集定义收敛依然是可以的: 这正是这一命题的意义.

下面给一些收敛的经典例子:

例 B.4 • 在 \mathbb{R} 中, $\{1/n\}_{n=1}^{\infty}$ 收敛到 0, 然而, 序列 $\{n\}_{n=1}^{\infty}$ 则不收敛. 这个例子表明, 极限未必需要在序列中出现, 以及趋于无穷是一种特殊的不收敛.

- 在 \mathbb{R}^n 和 L^p 范数下, $\{x_n\}_{n=1}^{\infty}$ 收敛到 x , 当且仅当对任意 $i = 1, \dots, n$, $\{x_n^i\}_{n=1}^{\infty}$ 收敛到 x^i . 这个例子表明, 高维空间中的收敛性可以从每个分量看.
- 在 $C([0, 1])$ 和 L^∞ 范数下, $f_n \rightarrow f$ 实际上是所谓一致收敛的概念, 即对任意 $\epsilon > 0$, 存在不依赖 x 的 $N \in \mathbb{N}$, 使得对任意 $n > N$, 任意 $x \in [0, 1]$, $|f_n(x) - f(x)| < \epsilon$. 在这一概念下, $\{x^n\}_{n=1}^{\infty}$ 就不收敛 (尽管它逐点收敛).

度量空间中紧集可以完全由收敛性来刻画:

定理 B.1 设 (X, d) 是一个度量空间, $F \subseteq X$, 则 F 是紧集当且仅当 F 中的任意序列都有收敛子列.

这一定理的证明并不算困难, 但是需要陈述的事实较多, 且与本书关联不大, 所以这里都略去.

这一定理足以表明紧集这一概念的重要性了, 然而, 这一定理的成立只需要度量空间, 度量空间是一个非常弱的概念, 我们关心的 \mathbb{R}^n 空间实际上有更强的性质, 这一性质是完备性. 要定义完备性, 我们需要 Cauchy 列.

定义 B.8 (Cauchy 列) 设 (X, d) 是一个度量空间, $\{x_n\}_{n=1}^{\infty}$ 是 X 中的一个序列, 如果对任意 $\epsilon > 0$, 存在 $N \in \mathbb{N}$, 使得对任意 $m, n > N$, $d(x_m, x_n) < \epsilon$, 则称 $\{x_n\}_{n=1}^{\infty}$ 是一个 **Cauchy 列**.

Cauchy 列描述了另一种收敛的概念, 它要求的是序列中的点越来越相互接近, 而不是越来越接近某个点. 注意, 这一定义没有办法像收敛性一样给一个纯拓扑的定义, 所以 Cauchy 列的概念是依赖于度量的. 自然, 收敛的点列是 Cauchy 列:

命题 B.9 设 (X, d) 是一个度量空间, $\{x_n\}_{n=1}^{\infty}$ 是 X 中的一个序列, 如果 $\{x_n\}_{n=1}^{\infty}$ 收敛, 则 $\{x_n\}_{n=1}^{\infty}$ 是 *Cauchy* 列.

证明 设 $\{x_n\}_{n=1}^{\infty}$ 收敛到 x , 则对任意 $\epsilon > 0$, 存在 $N \in \mathbb{N}$, 使得对任意 $n > N$, $d(x_n, x) < \epsilon/2$, 所以对任意 $m, n > N$, $d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \epsilon$, 所以 $\{x_n\}_{n=1}^{\infty}$ 是 *Cauchy* 列. \square

反过来, *Cauchy* 列是否一定收敛呢? 这一问题的答案是不一定. 实际上, 在 \mathbb{R} 上, 就如同有限覆盖原理, 这件事情的成立只能作为与实数公理等价的命题存在! 完备性指的就是 *Cauchy* 列一定收敛的性质:

定义 B.9 (完备度量空间) 设 (X, d) 是一个度量空间, 如果 X 中的任意 *Cauchy* 列都收敛, 则称 (X, d) 是一个完备度量空间.

我们不加证明地给出完备度量空间的例子:

例 B.5 • 有限维空间的例子: L^p 范数下 \mathbb{R}^n 是完备的.

- 反面的例子: 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 $\{x_n = \frac{1}{n} : n \in \mathbb{N}\}$, 它是 *Cauchy* 列, 但该点列在 X 中没有极限 (极限是 0).
- 无穷维空间的例子: $[0, 1]$ 到自身的连续函数空间 $C([0, 1])$ 在 L^∞ 范数下是完备的.
- 无穷维空间的另一个例子: ℓ^p 空间是完备的.

最后我们指出, 尽管完备度量空间已经足够发展微积分了, 但是它和 \mathbb{R}^n 依然有一个本质的区别, 这一区别在于紧集. 首先, 在有限维情况下, 紧集与有界闭集是等价的:

命题 B.10 设 \mathbb{R}^n 装备了 L^p 范数, 设 $F \subseteq \mathbb{R}^n$, 那么 F 是紧集当且仅当 F 是有界闭集, 有界指的是存在 $M > 0$, 使得对任意 $x \in F$, $\|x\|_p \leq M$.

这一命题的证明依赖于 Heine-Borel 有限覆盖原理, 这里就不给出细节了. 然而, 在无穷维空间中, 这一命题不一定成立:

命题 B.11 考虑 ℓ^2 的标准正交基 $\{e_i\}_{i=1}^{\infty}$, 考虑单位球面 $E = \{x \in \ell^2 : \|x\|_2 = 1\}$, 则 E 是有界闭集, 但不是紧集.

证明 因为对任意 $x \in E$, $\|x\| = 1$, 所以 $\|x\|_2 \leq 1$, 所以 E 是有界集. 取 $x \in \ell^2 \setminus E$. 如果 $\|x\| = r < 1$, 那么开球 $B(x, (1-r)/2) \subseteq B(0, 1) \subseteq \ell^2 \setminus E$; 对于 $r > 1$ 可以同理讨论. 这就证明了 E 是闭集. 最后证明 E 不是紧集. 考虑序列 $\{e_i\}_{i=1}^\infty$, 它是 E 中的序列, 因为对任意不同的 m, n , $\|e_m - e_n\| = 2$, 因此 $\{e_i\}$ 的任何子列都不是 Cauchy 列, 根据命题 B.9 的逆否命题, $\{e_i\}$ 没有任何收敛子列, 因而根据定理 B.1, E 不是紧集. \square

B.1.4 连续映射

接下来我们讨论两个拓扑空间之间的映射. 我们说过, 拓扑空间完全由开集给出, 所以某种程度保持拓扑性质的映射也会与开集有关系. 对于微积分来说, 连续性是其中最重要的一种. 遵循先前的惯例, 我们先给出更像微积分的 $\delta - \epsilon$ 语言的点定义, 然后再给出更像拓扑的定义.

定义 B.10 (连续映射) 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 是一个映射, 考虑点 $x \in X$, 对任意 $\epsilon > 0$, 存在 $\delta > 0$, 使得对任意 $y \in X$, 如果 $d_X(x, y) < \delta$, 则 $d_Y(f(x), f(y)) < \epsilon$, 则称 f 在点 x 处连续, 如果 f 在 X 的每一点都连续, 则称 f 是连续映射.

直观上说, 连续映射是指, x 和 y 足够接近的时候 $f(x)$ 和 $f(y)$ 也足够接近. 然而, 数学定义其实是反过来的: 想让 $f(x)$ 和 $f(y)$ 足够接近, 我们只需要让 x 和 y 足够接近. 更精确一些来说, 如果我们画了一个 $f(x)$ 的任意小的范围, 我们只需要找到一个 x 的范围, 使得 x 的范围里的点都被映射到 $f(x)$ 的范围里. 这一定义可以用开集来表述, 为此, 我们需要先引入一些关于映射的概念.

定义 B.11 (像, 原像) 设 $f: X \rightarrow Y$ 是一个映射, $A \subseteq X$, 则 $f(A) = \{f(x) : x \in A\}$ 称为 A 的像, 如果 $B \subseteq Y$, 则 $f^{-1}(B) = \{x \in X : f(x) \in B\}$ 称为 B 的原像.

于是, 我们可以用开集表述连续性了:

命题 B.12 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 是一个映射, 则 f 在 $x \in X$ 处连续当且仅当对任意包含 $f(x)$ 的开集 $U \subseteq Y$, 存在包含 x 的开集 $V \subseteq X$, 使得 $f(V) \subseteq U$.

这一命题的证明非常类似命题 B.4, 我们这里就不给出了. 连续映射的定义也可以完全由拓扑给出:

命题 B.13 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 是一个映射, 则下列表述等价:

1. f 是连续映射;
2. 对任意 Y 中的开集 U , 原像 $f^{-1}(U)$ 是 X 中的开集;
3. 对任意 Y 中的闭集 F , 原像 $f^{-1}(F)$ 是 X 中的闭集.

利用命题 B.12、命题 B.6 以及开集的定义, 很容易证明这一命题, 我们就不再证明了.

例 B.6 在不给任何额外定义的时候, 我们有一个非常自然的连续函数的例子, 那就是度量. 设 (X, d) 是一个度量空间, 我们证明度量 $d: X \times X \rightarrow \mathbb{R}$ 是一个连续函数.

我们利用点连续的定义, 证明 d 在每一点都连续. 设 $(x_1, y_1) \in X \times X$. 我们利用命题 B.12 和原始定义的混合版本. 注意到要证明所有包含 $d_0 = d(x_1, y_1)$ 的开集 U 满足条件, 根据 U 的构造, 只需要证明, 对任意 $\epsilon > 0$, $B(d_0, \epsilon)$ 满足条件. 为此, 取一个包含 (x_1, y_1) 的开集 $V = B(x_1, \epsilon/2) \times B(y_1, \epsilon/2)$ (关于这个为什么是开集, 详细讨论见例 B.3), 则对任意 $(x_2, y_2) \in V$, 有 $d(x_1, x_2) < \epsilon/2$, $d(y_1, y_2) < \epsilon/2$, 所以根据三角不等式,

$$d(x_2, y_2) \leq d(x_1, y_1) + d(x_1, x_2) + d(y_1, y_2) < d_0 + \epsilon/2 + \epsilon/2 = d_0 + \epsilon.$$

另一方面,

$$\begin{aligned} d_0 = d(x_1, y_1) &\leq d(x_2, y_2) + d(x_1, x_2) + d(y_2, y_1) < d(x_2, y_2) + \epsilon \\ \implies d(x_2, y_2) &> d_0 - \epsilon. \end{aligned}$$

所以, $d(x_2, y_2) \in B(d_0, \epsilon)$, 即 $V \subseteq B(d_0, \epsilon)$, 所以 d 在 (x_1, y_1) 连续. 因为 (x_1, y_1) 是任意的, 所以 d 是连续的.

连续性的定义实际分为了两部分, 一个是局部的、点的连续性, 另一个是整体的、只依赖开集而不依赖具体点的定义. 他们也对应了连续不同的性质. 我们首先讨论局部连续的性质, 以下命题我们都不再给出证明.

首先, 连续也可以用收敛性刻画:

命题 B.14 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 是一个映射, 则下列表述等价:

1. f 是连续映射;
2. 对任意 $x \in X$, 对任意 $\{x_n\}_{n=1}^{\infty}$, 如果 $x_n \rightarrow x$, 则 $f(x_n) \rightarrow f(x)$.

其次, 连续对复合是封闭的:

命题 B.15 设 (X, d_X) 、 (Y, d_Y) 和 (Z, d_Z) 是三个度量空间, $f: X \rightarrow Y$ 在 $x \in X$ 连续, $g: Y \rightarrow Z$ 在 $f(x) \in Y$ 连续, 则 $g \circ f: X \rightarrow Z$ 在 x 连续.

利用以上两个性质, 在赋范空间中, 我们得到如下结论:

推论 B.1 设 $(X, \|\cdot\|_X)$ 是赋范空间, 则数乘是 $X \rightarrow X$ 的连续映射, 向量加法是 $X \times X \rightarrow X$ 的连续映射. 因此, 有限维线性空间到有限维线性空间的线性映射都是连续映射.

根据命题 B.14, 这一结论也有对应的序列版本, 我们就不再列出了. 特别要注意的是, 这一结论也适用于 \mathbb{R} , 从数乘来看, 对于 \mathbb{R} 来说, 乘法 $\times: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 和除法 $\div: \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \rightarrow \mathbb{R}$ 也都是连续映射⁴.

最后, 连续意味着有界:

命题 B.16 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 在 $x \in X$ 连续, 则在 x 的某个邻域上 f 有界, 即存在 $r, M > 0$, 对任意 $y \in B(f(x), r)$, 有 $d_Y(f(x), y) \leq M$.

接下来我们讨论连续映射整体的性质, 这些性质都与紧集有关. 首先, 连续映射将紧集映射为紧集:

命题 B.17 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f: X \rightarrow Y$ 是一个连续映射, $F \subseteq X$ 是紧集, 则 $f(F)$ 是紧集.

然后要讨论的性质都限制映射的值是实数, 即 $f: X \rightarrow \mathbb{R}$. 这样的映射我们称之为函数. \mathbb{R}^n 区别于 \mathbb{R}^n 最大的不同是实数可以比大小而实数向量不行, 实数与大小相关的性质有很多, 我们列出其中两个与实数公理等价的. 这些性质需要用到界和确界的概念, 这一概念将会频繁出现在我们的讨论中, 所以这里单独给出:

定义 B.12 (上界, 上确界, 下界, 下确界) 设 $A \subseteq \mathbb{R}$, 如果存在 $M \in \mathbb{R}$, 使得对任意 $a \in A$, $a \leq M$, 则称 M 是 A 的一个上界, 如果 M 是 A 的上界, 且对任意 $M' < M$, 存在 $a \in A$, 使得 $a > M'$, 则称 M 是 A 的一个上确界, 记作 $\sup A$.

类似地, 如果存在 $M \in \mathbb{R}$, 使得对任意 $a \in A$, $a \geq M$, 则称 M 是 A 的一个下界, 如果 M 是 A 的下界, 且对任意 $M' > M$, 存在 $a \in A$, 使得 $a < M'$, 则称 M 是 A 的一个下确界, 记作 $\inf A$.

如果一个集合有上(下)界, 则称这个集合上(下)有界, 如果它既有上界又有下界, 则称这个集合有界.

⁴尽管从证明的逻辑顺序来说, 应该是先有了实数的四则运算连续性, 然后才有了赋范空间的连续性. 我们这样写是为了避免不必要的冗余, 直接从一个一般的视角出发来讨论.

上确界这个概念，直观上就是在说“最小可能的上界”，下确界也有类似的解读。

现在我们就可以阐述这两个实数的性质了。第一个是说单调有界的序列一定收敛。

命题 B.18 (单调有界原理) 设 $\{x_n\}$ 是一个单调有界的实数列，则 $\{x_n\}$ 收敛。

接下来一个是有上（下）界的实数集一定有上（下）确界，即最小可能的上（下）界是一个确实存在的实数，这也是一种完备性的体现。

命题 B.19 (确界原理) 设 $A \subseteq \mathbb{R}$ ，如果 A 有上界，则 $\sup A$ 存在；如果 A 有下界，则 $\inf A$ 存在。

确界原理给了一种求确界的方式：

命题 B.20 设 $A \subseteq \mathbb{R}$ ，如果 A 有上界，则存在一列 $\{a_n\}$ ，使得 $a_n \in A$ ，且 $\lim_{n \rightarrow \infty} a_n = \sup A$ 。

证明 设 $M = \sup A$ （由确界原理知 M 存在），对任意 $n \in \mathbb{N}$ ，由 $M - 1/n$ 不是 A 的上界，存在 $a_n \in A$ ，使得 $M - 1/n < a_n \leq M$ 。根据极限的定义易知 $\lim_{n \rightarrow \infty} a_n = M$ 。□

然后我们回到连续的整体性质上来。首先是 Weierstrass 最值定理：

定理 B.2 (Weierstrass 最值定理) 紧集上的连续函数 $f: F \rightarrow \mathbb{R}$ 在该紧集 F 的某个点取最大（最小）值。

然后是介值定理：

定理 B.3 (介值定理) 设 $f: [a, b] \rightarrow \mathbb{R}$ 是一个连续函数， $f(a) < f(b)$ ，则对任意 $y \in (f(a), f(b))$ ，存在 $x \in (a, b)$ ，使得 $f(x) = y$ 。

实际上，介值定理成立并不需要区间 $[a, b]$ ，任何一个连通的拓扑空间都可以，但是连通性的表述不是很直观，所以我们这里就不给出了。

§B.2 一元函数的微分学

§B.3 多元函数的微分学

§B.4 线性赋范空间的微分学，矩阵微分学

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, 200

上界, 204

上确界, 204

下界, 204

下确界, 204

像, 202

原像, 202

完备度量空间, 201

度量, 193

$L^1 \sim$, 194

$L^2 \sim$, 194

$L^\infty \sim$, 194

$L^p \sim$, 194

Chebyshev \sim , 194

Euclid \sim , 194

Manhattan \sim , 194

Minkowski \sim , 194

离散 \sim , 193

绝对值 \sim , 193

度量空间, 193

开球, 196

开覆盖, 199

开集, 196

拓扑空间, 196

收敛, 199

一致 \sim , 200

极限, 199

紧集, 199

范数, 194

等价 \sim , 196

赋范空间, 194

连续映射, 202

闭集, 198