

AI 中的数学

邓小铁 李翰禹

最后一次编译的时间：2025 年 2 月 14 日 14:56 (UTC+8)

前言

目录

第零章 引言	1
第一部分 AI 的逻辑	4
第一章 合情推理	5
§1.1 命题逻辑的演绎推理	6
§1.2 合情推理的数学模型	14
§1.2.1 合情推理的基本假设, 似然	15
§1.2.2 似然与概率	21
§1.2.3 先验与基率谬误	22
§1.3 合情推理的归纳强论证	25
§1.3.1 归纳强论证	25
§1.3.2 有效论证和归纳强论证的比较	29
§1.4 先验模型的存在性	35
§1.5 习题	38

第二章 Markov 链与模型	42
§2.1 Markov 链	43
§2.2 Markov 奖励过程 (MRP)	54
§2.3 Markov 决策过程 (MDP)	60
§2.4 隐 Markov 模型 (HMM)	70
§2.4.1 评估问题	72
§2.4.2 解释问题	75
§2.5 扩散模型	78
§2.5.1 采样逆向过程	83
§2.5.2 训练逆向过程	84
§2.6 习题	87
第二部分 信息与数据	92
第三章 熵与 Kullback-Leibler 散度	93
§3.1 熵	94
§3.1.1 概念的导出	94
§3.1.2 概念与性质	99
§3.2 Kullback-Leibler 散度	109
§3.2.1 定义	109
§3.2.2 两个关于信息的不等式	111
§3.3 编码理论	113
§3.3.1 熵与编码	113
§3.3.2 K-L 散度、交叉熵与编码	117
§3.4 在机器学习中的应用：语言生成模型	120

§3.5 附录：Shannon 定理的证明	122
§3.6 习题	125
第四章 高维几何, Johnson-Lindenstrauss 引理	130
§4.1 高维几何	132
§4.1.1 高维球体	132
§4.1.2 Stein 悖论	136
§4.1.3 为什么我们要正则化? 远有潜龙, 勿用	144
§4.2 集中不等式	145
§4.3 J-L 引理的陈述与证明	152
§4.4 J-L 引理的应用	157
§4.5 习题	160
第五章 差分隐私	163
§5.1 数据隐私问题	164
§5.2 差分隐私的定义与性质	168
§5.3 差分隐私的应用	177
§5.3.1 随机反应算法	177
§5.3.2 全局灵敏度与 Laplace 机制	179
§5.3.3 DP 版本 Lloyd 算法	182
§5.4 习题	185
第三部分 决策与优化	189
第六章 凸分析	190
§6.1 决策与优化的基本原理	191

§6.1.1 统计决策理论	191
§6.1.2 优化问题与算法	194
§6.1.3 例子：网格搜索算法	201
§6.2 凸函数	206
§6.3 凸集	214
§6.3.1 基本定义和性质	214
§6.3.2 分离超平面定理	219
§6.4 习题	222
第七章 对偶理论	226
§7.1 约束的几何意义	229
§7.2 条件极值与 Lagrange 乘子法	237
§7.3 Karush-Kuhn-Tucker 条件	240
§7.4 Lagrange 对偶	246
§7.4.1 原始规划与对偶规划	246
§7.4.2 对偶的几何意义	250
§7.4.3 弱对偶定理	254
§7.4.4 Slater 条件, 强对偶定理	256
§7.5 应用：支持向量机（SVM）	263
§7.6 习题	265
第八章 不动点理论	268
§8.1 Banach 不动点定理	269
§8.2 Brouwer 不动点定理	280
§8.3 习题	287

第四部分 博弈与逻辑	289
第九章 博弈与逻辑	290
§9.1 博弈的基本语言：以井字棋为例	292
§9.2 输赢博弈	294
§9.2.1 博弈的不同维度	294
§9.2.2 Zermelo 定理与 AlphaGo Zero	297
§9.3 正则形式博弈	304
§9.3.1 定义	304
§9.3.2 理性与均衡	306
§9.3.3 生成对抗网络	310
§9.3.4 混合策略	313
§9.4 随机博弈（Markov 博弈）	319
§9.5 习题	329
第五部分 认知与逻辑	332
第十章 共同知识，Bayes 博弈，Aumann 知识算子	333
§10.1 “泥泞的孩童”谜题	336
§10.2 不完全信息博弈（Bayes 博弈）	341
§10.3 电子邮件博弈	351
§10.4 Aumann 知识算子	356
§10.5 习题	365
第十一章 模态逻辑，知识的逻辑	367
§11.1 知识逻辑的形式语言	369

§11.2 Kripke 语义	374
§11.3 模态可定义性	382
§11.4 知识逻辑的基本模型与性质	384
§11.4.1 知识逻辑的 Kripke 模型与公理	385
§11.4.2 Kripke 模型与 Aumann 结构	392
§11.4.3 “泥泞的孩童”再回顾：形式化解法	394
§11.5 对不一致达成一致	397
§11.5.1 模型	397
§11.5.2 定理及其证明	401
§11.6 习题	403
第六部分 附录：预备知识	405
附录 A 线性代数基础	406
§A.1 线性空间	406
§A.2 线性映射	413
§A.3 矩阵	419
§A.4 双线性型与二次型	428
§A.5 带内积的线性空间	434
§A.6 行列式	443
§A.7 算子范数与谱理论	447
附录 B 微分学基础	455
§B.1 点集拓扑	455
§B.1.1 度量空间，范数	455
§B.1.2 开集与闭集	460

§B.1.3 紧致性, 收敛性, 完备性	464
§B.1.4 连续映射	468
§B.1.5 与实数序有关的性质	473
§B.2 一元函数的微分学	476
§B.2.1 导数与微分的定义	477
§B.2.2 微分学基本定理	482
§B.3 多元函数的微分学	484
§B.3.1 微分、偏导数与导数的定义	484
§B.3.2 微分学基本定理	494
§B.3.3 隐函数定理	497
附录 C 概率论基础	503
§C.1 从朴素概率论到公理化概率论	503
§C.1.1 Kolmogorov 概率论	503
§C.1.2 条件概率, 独立性	509
§C.2 随机变量, 分布函数	515
§C.2.1 基本定义	515
§C.2.2 离散型随机变量	521
§C.2.3 连续型随机变量	522
§C.2.4 随机向量, 条件分布, 独立性	527
§C.2.5 随机变量（向量）的函数	534
§C.3 随机变量的数字特征, 条件数学期望	538
§C.3.1 数学期望, Lebesgue 积分	539
§C.3.2 数学期望的性质	545
§C.3.3 随机变量的内积空间	549
§C.3.4 特征函数	552

§C.3.5 条件数学期望	554
§C.4 多元正态分布 (Gauss 向量)	561
§C.5 大数定律	563
参考文献	566

第零章 引言

何谓智能？从历史的进程来看，生命的演化似乎就是智能的演化。生命最初的形态很可能是单细胞生物，它能够维持自身的结构，并通过吸收外界的物质来维持自己这样的结构。此时，生物已经可以选择性吸收外界的物质，而将不需要的物质排除在外，这已经可以被看成一种极其简单的智能。

随着的演化进行，生命的形态变得越来越复杂。在人的这条演化线上，我们经历了多细胞的、类似线虫的动物，到了爬行动物，然后变成哺乳动物，再到了灵长类动物，最后是人类。在这个演化的过程中，生命对于外界的反应变得越来越复杂，智能也变得越来越强大。

那么，应该怎么定义智能呢？困难在于，智能并不是一个非黑即白的东西，我们不能画一条线然后说，这边的是智能，那边的是非智能。正如同生命演化的历程，智能也是一个多层次渐变的过程。

从生命的观点来看，智能意味着一种对不断变化的外界的适应能力，越高级的智能能够考虑的外界因素就越复杂，适应变化的能力也越强。例如，秀丽隐杆线虫是一种简单的有神经系统的生物，它能够机械地对外界刺激做出反应，一切行为似乎是写好的，没有任何自主性。当我们把目光转移到蜥蜴，一种爬行动物，我们发现他们的行为似乎更

加灵活，蜥蜴似乎可以为了某个目标（比如进食）而做出一些行为。而到了人类，我们的行为准则中会把其他人纳入考虑，我们会考虑自己的行为对他人的影响，会去设身处地地考虑他人的感受，形成社会规范。[\[Tom22\]](#)

从演化的这个路径来看，智能有着两个维度：一个是从认知到决策，另一个是从个体到多体。

- 从认知到决策：智能体为了适应外界，必须要先知道外界的情况，然后根据自己的目标做出相应的选择。例如，如果我们感到口渴，我们会先扫视桌面，找到水杯，然后才会拿起水杯喝水。因此，智能需要解决的第一个问题是认知，即如何获取外界的信息，然后是决策，即如何根据自己的目标利用这些信息做出选择。
- 从个体到多体：在自然界中，生命面对的最大变数就是其他生命。因此，演化过程中，最大的选择压力其实来自其他生命。对于秀丽隐杆线虫，他们似乎只会对外界刺激做出反应，而不会考虑其他生命的行动。而到了人类（甚至猩猩），我们会根据自己的想象，去揣测其他人的意图，从而做出自己的选择。因此，智能不仅需要有能够面对变化环境的适应能力，更需要有能够面对其他生命的适应能力。

因此，如果我们要实现“人工智能”（AI, Artificial Intelligence），我们需要面对的问题就是，如何实现这两个维度的智能。本书的组织结构就是按照这两个维度来展开的。书一共被分为了五个部分：

- 第一部分，AI 的逻辑：这部分探究 AI 的认知是什么样的，如何被数学建模。我们将会探究如何用 Bayes 概率论建立 AI 的推理模型（第一章），以及如何用 Markov 链建立 AI 包含时间的认知与决策

模型（第二章）.

- 第二部分，信息与数据：这部分探究 AI 如何从环境中获得认知，而现实中，这就是关于信息与数据的问题. 我们将会展示信息论的基本事实（第三章），AI 面对数据的特性（第四章），以及如何在保护个人隐私的前提下让 AI 利用数据（第五章）.
- 第三部分，决策与优化：这部分探究 AI 如何面对环境做出决策以及优化. 我们将会给出优化的基本概念以及特别重要的一类优化问题——凸优化（第六章），以及如何处理带约束的优化问题（第七章），最后我们将会给出“多体优化”的基本工具——不动点理论（第八章）.
- 第四部分，博弈与逻辑：这部分探究 AI 如何面对其他个体做出决策以及优化，即博弈论，它也为多体智能的研究提供了一个标准的语言（第九章）.
- 第五部分，认知与逻辑：这部分探究 AI 如何形成对其他个体的认知，特别是如何在数学上建模这种认知. 我们将给出两种风格的数学模型：基于 Bayes 概率论的 Bayes 博弈（第十章）和基于形式逻辑的模态逻辑（第十一章）.

第一部分

AI 的逻辑

第一章 合情推理

1983 年，心理学家 Amos Tversky 和 Daniel Kahneman 进行了一个著名的实验 [TK83]，他们向一组被试提出了如下问题：Linda 是一位单身、外向且年龄为 31 岁的女性。在大学期间，她主修哲学，十分关注种族歧视和社会公正问题，而且曾参加过反核游行。请给下面几个选项发生的可能性排序，1 代表最可能，8 代表最不可能：

1. Linda 是一个小学老师。
2. Linda 在书店工作，并且参加瑜伽课程。
3. Linda 非常积极参加女权运动。
4. Linda 是一个精神病的护理员。
5. Linda 是美国女性选民联盟的成员。
6. Linda 是一名银行出纳员。
7. Linda 是一个保险销售员。
8. Linda 是一名银行出纳员，同时她还非常积极参加女权运动。

他们对以下三组人进行了同样的实验: (1) 88 名来自斯坦福大学或不列颠哥伦比亚大学 (UBC) 的本科生, 没有学过概率统计, (2) 53 名斯坦福大学的一年级研究生, 他们学过一门或更多的概率统计课程, (3) 32 名斯坦福大学商学院博士生, 他们学过很多概率统计和决策理论的进阶课程.

我们重点关注第三、六、八个选项. 抽象地来说, 我们可以把这三个陈述分别记为 $A, B, A \wedge B$. 从逻辑的角度来看 $A \wedge B$ 发生的时候, A 和 B 都一定发生, 因此 A (或 B) 更可能发生. 然而, 让人大跌眼镜的是, 实验结果显示, 所有三组人的打分中, 超过 85% 的人认为 $A \wedge B$ 更可能发生!

以上现象被称为合取谬误, 它表明人类的推理并不等同于我们所理解的逻辑推理 (被称为演绎推理). 合取谬误的背后隐藏了一种新的推理方式, 被称为合情推理. 演绎推理很早就有了非常严格的逻辑学框架, 得到了充分的研究. 然而, 合情推理同样甚至更加重要, 它不仅是人类与生俱来的能力, 更是科学的研究和 AI 的逻辑基础.

本章试图建立合情推理的一个数学模型, 这一模型与 Bayes 概率论紧密地联系在一起. 我们会用它来研究一些有趣的推理谬误, 包括合取谬误, 并与演绎推理进行对比. 最后, 我们会讨论 Bayes 概率论中先验为什么是存在的.

§1.1 命题逻辑的演绎推理

作为回顾, 本节讨论演绎推理, 它最主要的应用场景是数学证明和一些逻辑论证. 演绎推理定义了数学中一些非常重要的概念, 例如 “公理” “定理” “证明”. 我们会看到, 与演绎推理密切相关的是形式推理

系统，它尝试将推理的过程抽象为一堆字符串的变换。作为介绍，我们这里给出命题逻辑的形式推理系统，它是最基本的形式推理系统。

命题是一种陈述，它可以被判定为真或者假。例如，“北京是中国的首都。”是一个命题，而“北京是中国的首都吗？”是一个疑问句，所以不是命题。然而，利用自然语言定义命题，很难判断一些奇怪的句子是不是命题：

例 1.1 考虑句子 A : “这句话是假的。”是一个命题吗？首先， A 是一个陈述句。那它的真假呢？

- 如果 A 是真的， A 说的是“这句话是假的”，这与 A 真矛盾，所以它不是真的。
- 如果 A 是假的， A 的含义是“这句话是真的”，这与 A 假矛盾，所以它不是假的。

因此，这句话既不是真的也不是假的！ □

以上的现象被称为自指，也就是一个东西在描述自己。

自指为讨论命题的概念带来了非常大的麻烦。为了避免这些麻烦，我们回避命题到底是什么这一哲学问题，而是把他们抽象为一些字符串，这就是命题公式：

定义 1.1 (命题公式) 命题公式（或简称公式）是由一些命题变元和连接词组成的字符串，它们满足以下递归定义：

- 命题变元 p, q, r, \dots 是命题公式。
- 如果 ϕ 和 ψ 是命题公式，那么 $(\neg\phi)$, $(\phi \vee \psi)$, $(\phi \wedge \psi)$, $(\phi \leftrightarrow \psi)$ 和 $(\phi \rightarrow \psi)$ 都是命题公式。

$\neg, \vee, \wedge, \leftrightarrow, \rightarrow$ 被称为连接词， \neg 被称为否定， \vee 被称为析取（或）， \wedge 被称为合取（与）， \leftrightarrow 被称为双向蕴含（等价）， \rightarrow 被称为蕴含（推出）。在不产生混淆的时候也会省略括号。 $A \wedge B$ 也常写作 AB ， $\neg A$ 也常写作 \overline{A} . \square

例 1.2 $(p \vee (q \rightarrow r))$ 是命题公式，但是 $p \vee \vee q$ 不是. \square

命题最重要的概念是“真假”，因而命题逻辑一个重要的问题是：什么样的公式是真的？给定一个公式集 Γ ，对一个公式 ϕ ，我们有两种真的概念：

- 语义： $\Gamma \vDash \phi$.
- 语形： $\Gamma \vdash \phi$.

首先看语义。所谓语义，就是“句子的含义”。在命题逻辑中，句子就是命题公式，而含义就是真假。既然命题公式是递归定义的，命题公式的真假我们也应该递归定义。

定义 1.2 (赋值) 给定命题变元集合 V ，它形成的命题公式的集合是 L ，一个赋值是一个从 L 到 $\{\top, \perp\}$ 的映射 v ，递归定义如下：

- 如果 $\phi = p \in V$ ，那么 $v(\phi) = v_p$.
- 如果 $\phi = \neg\psi$ ，那么 $v(\phi) = \begin{cases} \top, & v(\psi) = \perp, \\ \perp, & v(\psi) = \top. \end{cases}$
- 如果 $\phi = \psi \vee \chi$ ，那么 $v(\phi) = \begin{cases} \top, & v(\psi) = \top \text{ 或 } v(\chi) = \top, \\ \perp, & v(\psi) = \perp \text{ 且 } v(\chi) = \perp. \end{cases}$

- 如果 $\phi = \psi \wedge \chi$, 那么 $v(\phi) = \begin{cases} \top, & v(\psi) = \top \text{ 且 } v(\chi) = \top, \\ \perp, & v(\psi) = \perp \text{ 或 } v(\chi) = \perp. \end{cases}$
- 如果 $\phi = \psi \rightarrow \chi$, 那么 $v(\phi) = \begin{cases} \top, & v(\psi) = \perp \text{ 或 } v(\chi) = \top, \\ \perp, & v(\psi) = \top \text{ 且 } v(\chi) = \perp. \end{cases}$
- 如果 $\phi = \psi \leftrightarrow \chi$, 那么 $v(\phi) = \begin{cases} \top, & v(\psi) = v(\chi), \\ \perp, & v(\psi) \neq v(\chi). \end{cases}$ □

以上定义看起来比较冗长, 但是符合我们的直觉. 一种更为直观的表达方式是真值表. 注意, 连接词可以被理解为一个真值向量到真值的映射, 例如 \wedge 可以被理解为 $\{\top, \perp\}^2 \rightarrow \{\top, \perp\}$ 的映射, 这一映射可以用真值表表示:

p	q	$p \wedge q$
\top	\top	\top
\top	\perp	\perp
\perp	\top	\perp
\perp	\perp	\perp

于是, 定义中的递归可以用直接用 \wedge 来表示: 如果 $\phi = \psi \wedge \chi$, 那么 $v(\phi) = v(\psi) \wedge v(\chi)$.

我们看一个例子.

例 1.3 tt

□

接下来, 我们从赋值到推理. 推理是从前提出发, 得到结论的过程. 和赋值相关的推理概念有如下定义:

定义 1.3 (语义蕴含) 给定公式集 Γ 和公式 ϕ , 对任意赋值 v , 如果 $v(\psi) = \top$ 对任意 $\psi \in \Gamma$ 都成立, 那么 $v(\phi) = \top$, 则称 Γ 语义蕴含 ϕ , 记作 $\Gamma \vDash \phi$. \square

换言之, 给定一个真的前提 Γ , 我们所关注的公式 ϕ 也是真的. 这一概念只关心真值之间的关系, 但是至于 Γ 如何导致 ϕ 为真, 不是这一概念可以刻画的.

例 1.4 如果 $\Gamma = \{p \rightarrow q, q \rightarrow r\}$, 那么 $\Gamma \vDash p \rightarrow r$. 我们来证明这个, 对任意赋值 v , 如果 $v(p \rightarrow q) = \top$, 那么 $v(p) = \perp$ 或者 $v(q) = \top$,

- 如果 $v(p) = \perp$, 那么 $v(p \rightarrow r) = \top$.
- 如果 $v(q) = \top$, 那么 $v(q \rightarrow r) = v(r)$. 因为 $v(q \rightarrow r) = \top$, 所以根据定义 $v(r) = \top$, 于是 $v(p \rightarrow r) = \top$.

不论在何种情况, 都有 $v(p \rightarrow r) = \top$, 即 $\Gamma \vDash p \rightarrow r$. \square

接下来, 我们看语形. 语形就是“句子的形态”, 或者说一个句子是如何构造出来的. 它更像我们所理解的“推理”, 也就是从一些公式出发, 得出另一些公式. 推导规则描述了从一些公式出发如何得到另外公式. 语法推导的形式是

$$\frac{\phi_1 \quad \phi_2 \quad \dots \quad \phi_n}{\psi}$$

横线上方的称为前提, 横线下方的称为结论.

在一些命题逻辑的形式推理系统 (比如 Hilbert 推理系统) 中, 推

导规则只有一条，即肯定前件 (MP)：

$$\frac{\phi \rightarrow \psi \quad \phi}{\psi}$$

然而，引入更多推导规则可以有助于简化推理的过程，这是另外一些形式推理系统（比如自然演绎系统）的做法。比如，我们可以加入合取和析取的推导规则：

- 引入新的连接词，例如

$$\frac{\phi \quad \psi}{\phi \wedge \psi}$$

- 消除连接词，例如

$$\frac{\phi \wedge \psi}{\phi}$$

我们再看一个非常重要的例子。

例 1.5 将归谬法 (RAA) 加入推导法则：将 $\neg\phi$ 作为前提，得到了结论 \perp ，那么可以推出 ϕ 才是结论。写作

$$\begin{array}{c} [\neg\phi] \\ \vdots \\ \perp \\ \hline \phi \end{array}$$

方括号表示假设 $[\neg\phi]$ 是前提，省略号表示推导的中间步骤。注意，这里是假设了一个前提，而不是真的前提，这一点是非常重要的。利用

这一推导规则，我们就可以得到反证法. \square

演绎推理所对应的数学模型即为形式推理系统：一系列字符串的集合（命题公式），以及一系列的推导规则，以及公理的集合（一些命题公式）.

在形式推理系统中，我们可以形式地定义什么是“推理”和“结论”，而这些东西构成了演绎推理的过程与结果：

定义 1.4 (演绎推理) 设 L 是公式集， $A \subseteq L$ 是公理的集合， \mathcal{R} 是推导规则， $\Gamma \subseteq L$ ，考虑 L 中的序列 ϕ_1, \dots, ϕ_n ，如果对任意 i ， ϕ_i 满足下面的其中一个条件：

1. $\phi_i \in A$ ，
2. $\phi_i \in \Gamma$ ，或者
3. 存在一个推导规则，和 $\phi_{i_1}, \dots, \phi_{i_k}$ ， $i_1, \dots, i_k < i$ 使得 ϕ_i 可以由 $\phi_{i_1}, \dots, \phi_{i_k}$ 推导出来，

那么称 ϕ_1, \dots, ϕ_n 是从 Γ 出发的一个演绎推理，记作 $\Gamma \vdash \phi_n$ ， ϕ_n 称为结论， Γ 称为前提. \square

简而言之，记号 $\Gamma \vdash \phi$ 的意思是以 Γ 作为前提，依据推导法则，通过一系列（但有限）的推理步骤，可以得到 ϕ 作为结论. 因而，语形定义的真值是“可以被推理出来”的概念. 对比语义，语形更加关注“过程”，而不是“关系”.

介绍完了语义和语形，我们现在来看他们之间的关系，其中最重要的是完备性定理：

定理 1.1 (完备性定理) 对任意公式集 Γ 和任意公式 ϕ ,

$$\Gamma \vDash \phi \iff \Gamma \vdash \phi.$$

完备性定理有两层意思：能推理出来的都是真的，真的都能被推理出来。并非所有的形式推理系统都有完备性定理，但是对于命题逻辑来说，这个定理成立的形式推理系统是更加符合直观的。

完备性定理的一个直接推论是，我们可以用真值表来检查从前提出发是否能推出结论。例如，在例 1.4 中我们验证了 $p \rightarrow q, q \rightarrow r \vDash p \rightarrow r^1$ ，所以也有

$$p \rightarrow q, q \rightarrow r \vdash p \rightarrow r,$$

即便我们完全不知道这一推理是如何进行的，我们依然知道这一推理是可行的！

如果 $\emptyset \vDash \phi$ ，那么称 ϕ 为重言式。重言式是不需要加任何假设也一定成立的公式，是这一推理系统所包含的“正确的废话”。如果 $\psi \leftrightarrow \phi$ 是重言式，我们就说 ψ, ϕ 是等值的。等值的公式在演绎推理中可以互相替代使用。例如： $p \rightarrow q$ 与 $\neg q \rightarrow \neg p$ 是等值的，所以证明一个命题也可以去证明它的逆否命题。

注. 对本节内容感兴趣的读者，有很多很好的教材可以参考。[\[VD13\]](#) 提供了全面但是简洁明了的讲述，[\[Ham88\]](#) 从数学家的角度讲述了数理逻辑的基本观点，[\[NS97\]](#) 则从应用、特别是计算机的角度来讲述数理逻辑。

¹在书写时，我们常常省略 Γ 中的花括号，因此这里写作 $p \rightarrow q, q \rightarrow r \vDash p \rightarrow r$ 而不是 $\{p \rightarrow q, q \rightarrow r\} \vDash p \rightarrow r$ 。对于符号 \vdash 我们遵循同样的规则。

§1.2 合情推理的数学模型

现在，我们将话题转到合情推理。出乎意料的是，人类先天所具备的推理能力，更接近合情推理而不是演绎推理。在婴儿时期，我们需要建立各式各样的经验规则，以便在这个世界生存下来。例如，如果婴儿碰到了带电的插座，那么他/她会感受到疼痛，这样就知道了插座是危险的。下次遇到插座的时候，他/她就会避开插座，因为他/她学到了“所有插座都是危险的”这一经验。上面的例子都是从一些特殊的事件（摸插座触电了）出发，得到了一些一般的结论（所有插座都是危险的），这就是合情推理的基本形式。

为了从更抽象的角度理解合情推理的形式，我们用经典的三段论来对比演绎推理和合情推理。

演绎推理包含两个强三段论：

$$\begin{array}{c} \phi \rightarrow \psi \\ \phi \text{ 是真的} \\ \hline \psi \text{ 是真的} \end{array} \qquad \begin{array}{c} \phi \rightarrow \psi \\ \psi \text{ 是假的} \\ \hline \phi \text{ 是假的} \end{array}$$

例如，从科学的角度，触电会疼痛 ($\phi \rightarrow \psi$)，摸插座会触电 (ϕ)，所以摸插座会疼痛 (ψ)。这是一个演绎推理的例子。

然而，合情推理是完全反过来的，即弱三段论，形式上写作

$$\begin{array}{c} \phi \rightarrow \psi \\ \psi \text{ 是真的} \\ \hline \phi \text{ 变得更合理} \end{array}$$

例如，触电会疼痛 ($\phi \rightarrow \psi$)，摸插座会疼痛 (ψ)，所以摸插座很

可能会触电 (ϕ) . 这样的推理更符合上面婴儿推理的模式. 虽然我们不能得知 ϕ 的真假，但是根据已有的事实 ψ ，我们可以得到 ϕ 的合理性推断. 这样没有严格真假的推理模式就是合情推理.

推理的一个重要概念是“真假”，或者说这个推理的正确性、合理性. 在演绎推理中，推理只有两种可能：正确的推理和错误的推理；而在合情推理中，推理的合理性有多种可能，例如“很可能正确”“可能正确”“很可能错误”等. 我们来看一个例子.

例 1.6 假如婴儿摸了插座，并且感到了疼痛，那么，你觉得他/她会怎么想？

- 选项 1：摸插座是危险的.
- 选项 2：摸的插座上面有微小不可见的细针，细针扎了他/她的手指.

□

虽然选项 1 和选项 2 都是可能的，但是他们的合理程度并不同：选项 1 明显比选项 2 要合理得多. 这里产生了一个自然的问题：如何刻画不同推理的合理程度？

§1.2.1 合情推理的基本假设，似然

接下来，我们给合情推理建立一个数学模型. 合情推理的基本方式是给定已知的命题 B ，推出命题 A 的合理程度，比如“非常可能是真的”“可能是真的”“可能是假的”“非常可能是假的”等，我们用符号 $A|B$ 表示这一合理程度.

在写出符号的时候，我们已经假设了合理程度所对应的数学对象： $A|B$ 应该是 $L \times L \rightarrow \mathbb{R}$ 的函数，这里 L 是命题公式的集合. 但是，这

一表述是有瑕疵的，因为不是所有的前提到结论的推理都可以用合理程度来描述。比如， B 是“太阳从西边升起”，前提根本就是一个假命题，从假命题出发，何谈“合情”？所以，符号 $A|B$ 应该只对一些特定的 A, B 有意义，我们称之为 $L \times L \rightarrow \mathbb{R}$ 的偏函数，并将这一函数写为 $f(A|B)$ ，称作似然。

我们还没有确定这一映射的具体形式。但是，这样的映射要满足两个原则：1. 符合我们对于合情推理的基本直观，2. 易于计算。

在这里，我们类比行星运动定律来理解这第二个原则的具体含义。最开始的时候，人们认为天体都是绕着地球转的。基于这样的思想，Ptolemaeus 提出了行星运动的模型。在这个模型中，Ptolemaeus 引入了“均轮”和“本轮”的概念。行星在一个较小的圆（本轮）上做匀速圆周运动，而这个本轮的中心又在一个更大的圆（均轮）上绕地球做匀速圆周运动。这一模型可以较为准确预测天体的运动，然而需要非常复杂的概念。

后来，Kepler 提出了完全不同的行星运动模型：行星绕着恒星做椭圆运动，与恒星的连线扫过的面积与运动时间成正比，行星轨道的半长轴的三次方与其公转周期的二次方的比值都相等。这就是著名的“Kepler 三定律”。在 Kepler 的模型中，我们放弃了“均轮”和“本轮”的概念，而是用简单的椭圆来描述行星的运动。尽管 Ptolemaeus 模型和 Kepler 模型都可以描述行星的运动，但是 Kepler 模型更加简单，更加容易计算，也正是在此基础上，Newton 才发现了万有引力定律。

同样的道理也适用于合情推理。或许会有很多的 f 可以符合直观地描述合理程度，但是我们希望找到一个尽可能容易使用和计算的 f ，它能有更广泛的应用，或许也蕴含了更深层次的规律。现在，我们暂时将所有符合合情推理直观的映射的集合记作 \mathcal{L} 。我们给出 \mathcal{L} 中的函数要具

备的性质（即假设），然后再确定具体的函数。

首先，在命题逻辑中，我们只关心命题的真假，并不关心命题具体是什么。在合情推理中，我们依然希望有这样的性质：推理的合理程度不依赖具体的命题形式。在数学上，这等价于如下的原则：

规则 1.1 (等值原则) 对任意 $f \in \mathcal{L}$ ，任意命题公式 $\phi_1, \phi_2, \psi \in L$ ，如果 ϕ_1 和 ϕ_2 等值（即 $\phi_1 \leftrightarrow \phi_2$ 是重言式），那么

$$f(\phi_1|\psi) = f(\phi_2|\psi).$$

例如， $\phi \rightarrow \psi$ 和 $\neg\phi \vee \psi$ 的具体含义是不同的：一个是蕴含式（即 ϕ 推出 ψ ），一个是两个命题 ($\neg\phi$ 和 ψ) 的或；但是他们等值，所以 $f(\phi \rightarrow \psi|\xi) = f(\neg\phi \vee \psi|\xi)$ 。更多的例子如下：

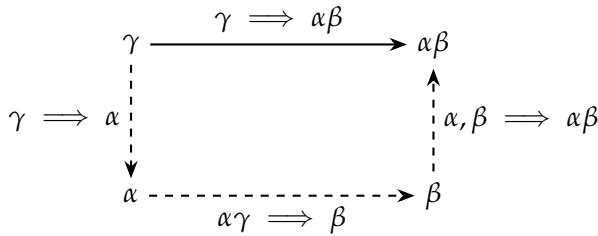
例 1.7 对任意 $f \in \mathcal{L}$ 以及公式 α, β, γ ，成立

- $f(\alpha \vee \beta|\gamma) = f((\beta \vee \alpha)|\gamma).$
- $f((\alpha\beta)\gamma|\delta) = f(\alpha(\beta\gamma)|\delta).$
- $f(\neg\neg\alpha|\beta) = f(\alpha|\beta).$ □

现在我们来看第二条假设，它与推理不同的路径有关。考虑从 γ 出发推出 $\alpha\beta$ 为真。我们有两种推理的链条：

1. 直接推理：直接从 γ 推出 $\alpha\beta$ 。
2. 分步推理：先从 γ 推出 α ，然后从 $\alpha\gamma$ 推出 β ，因此自然也推出了 $\alpha\beta$ 。

也可以用下面的图来直观表示，其中实线箭头表示直接推理，虚线箭头表示分步推理：



我们前面要求，“推理的合理程度不依赖具体的命题形式”，现在，我们将这一要求进一步细化为“推理的合理程度不依赖推理的具体形式”. 因为上面的直接推理和分步推理都是从 γ 推出 $\alpha\beta$, 所以我们要求他们的合理程度相同, 即

假设 1.1 对任意 $f \in \mathcal{L}$, 存在一个 $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 的函数 F , 对任意命题公式 α, β, γ ,

$$f(\alpha\beta|\gamma) = F(f(\alpha|\gamma), f(\beta|\alpha\gamma)).$$

换言之, 我们不关心 γ 具体是怎么推出来 α 的, 我们只要求 $\alpha\beta$ 的合理程度与 α 和 β 的合理程度有关.

注意, 以上假设不能写成 $f(\alpha\beta|\gamma) = F(f(\beta|\gamma), f(\alpha|\gamma))$. 例如, 假设 α 是“小明的左眼是蓝色的”, β 是“小明的右眼是棕色的”. $f(\alpha|\gamma)$ 和 $f(\beta|\gamma)$ 都可以很合理, 但 $f(\alpha\beta|\gamma)$ 则非常罕见. 同样的道理, 我们可以排除其他形式的假设, 比如 $f(\alpha\beta|\gamma) = F(f(\gamma|\alpha), f(\beta|\gamma))$.

我们定义的 F 具有如下性质:

定理 1.2 如果 F 是二阶连续可微函数, 那么 F 充分必要地满足

$$cf(F(p, q)) = f(p)f(q),$$

这里, c 是某个常数, f 是某个函数. 因此, 对任意命题公式 α, β, γ , 成

立

$$cf(\alpha\beta|\gamma) = f(\alpha|\gamma)f(\beta|\alpha\gamma).$$

这一定理的证明见习题 4. 回忆，我们的目标是找到一个尽可能简单的 f ，它能够描述合情推理的合理程度。因此，我们可以接受 F 二阶连续可微的假设，然后让 $c = 1$ 。于是，我们得到了合情推理的一个基本原则：

规则 1.2 (与规则) 对命题公式 α, β, γ ，成立

$$f(\alpha\beta|\gamma) = f(\alpha|\gamma)f(\beta|\alpha\gamma).$$

第二个考虑的是似然 $f(\alpha|\beta)$ 和 $f(\neg\alpha|\beta)$ 的关系。两个似然应该存在某种函数关系，于是我们有如下假设：

假设 1.2 对任意 $f \in \mathcal{L}$ ，存在函数 $S : \mathbb{R} \rightarrow \mathbb{R}$ ，对任意命题公式 α, β ，成立

$$f(\neg\alpha|\beta) = S(f(\alpha|\beta)).$$

在与规则限制之下，我们可以证明 S 具有如下性质：

定理 1.3 如果 S 二阶连续可微，那么 S 充分必要地满足

$$S(x) = (1 - x^m)^{1/m},$$

这里的 m 是一个正常数。

这一定理的证明见习题 5. 由上面的定理，我们可以得到以下性质

推论 1.1 $f(\alpha|\beta)^m + f(\neg\alpha|\beta)^m = 1$.

证明. $f(\neg\alpha|\beta) = S(f(\alpha|\beta)) = (1 - f(\alpha|\beta)^m)^{1/m}$. □

同样, 从简洁性的角度出发, 我们可以接受 S 二阶连续可微的假设, 然后让 $m = 1$. 于是, 我们得到了合情推理的另一个基本原则:

规则 1.3 (否定规则) 对命题公式 α, β , 成立

$$f(\alpha|\beta) + f(\neg\alpha|\beta) = 1.$$

由上, 我们得到了合情推理的三个基本原则: 等值原则、与规则和否定规则. 这三个原则构成了合情推理的基本框架.

接下来, 我们说明, 这三个原则就已经足够用来计算任意的合情推理. 首先, 连接词 $\{\wedge, \neg\}$ 组成了演绎逻辑的完备集: 所有真值函数都可以用这个集合的词来表示, 因此, 根据等值规则, 只要可以计算包含 \wedge 和 \neg 的合情推理, 就可以计算任意合情推理. 其次, 我们可以用与规则来计算 \wedge , 否定规则计算 \neg , 因此, 任何命题都可以通过这三个原则来计算.

例 1.8 例如, 计算 $f(A \vee B|C)$,

$$\begin{aligned} f(A \vee B|C) &= 1 - f(\overline{A} \wedge \overline{B}|C) \\ &= 1 - f(\overline{A}|C)f(\overline{B}|\overline{A}C) \\ &= 1 - f(\overline{A}|C)(1 - f(B|\overline{A}C)) \\ &= f(A|C) + f(\overline{A}B|C) \\ &= f(A|C) + f(B|C)f(\overline{A}|BC) \\ &= f(A|C) + f(B|C)(1 - f(A|BC)) \\ &= f(A|C) + f(B|C) - f(AB|C). \end{aligned}$$
□

§1.2.2 似然与概率

我们已经知道，似然是一个合理性度量，在本节中，我们将似然与概率联系起来：他们实际上是等价的。

首先介绍 *Kolmogorov* 的概率论，详细的讨论见附录 C. 这一概率论研究事件空间上的概率测度。其研究对象的总体被称为样本空间，记为 Ω 。我们只能观测到某种可观测的特性 P ，而不能直接观测样本点，即我们只能观察事件，或者说集合

$$\{\omega \in \Omega : P(\omega)\}.$$

我们可以观测的所有事件的集合称为事件域，记为 \mathcal{F} 。

事件域 \mathcal{F} 中的事件之间互相有关联。我们自然可以观测到 Ω ，因此 $\Omega \in \mathcal{F}$. 如果我们可以观测到事件 A ，那么我们也可以通过没有观测到 A 来判断观测到了 $\Omega \setminus A$. 因此， $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$. 如果我们观测到了 A 或者 B ，我们其实也观测到了 $A \cup B$ ，即 $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$.

事实上，我们可以形式上定义概率：

定义 1.5 (概率) 概率 \Pr 是一个事件域 \mathcal{F} 到实数的映射，并且满足：

- 规范性： $\Pr(\Omega) = 1$.
- 非负性：对任意 $A \in \mathcal{F}$ ， $\Pr(A) \in [0, 1]$.
- 可列可加性：对 A_1, A_2, \dots 满足 $A_i \cap A_j = \emptyset, i \neq j$ ，有

$$\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i).$$

□

事件是命题的集合论描述. 具体来说, 有如下对应

事件	命题
Ω	\top
\emptyset	\perp
\overline{A}	$\neg A$
$A \cap B$	$A \wedge B$
$A \cup B$	$A \vee B$
$A \subseteq B$	$A \rightarrow B$
$A = B$	$A \leftrightarrow B$

回忆条件概率的定义, 我们可以得到链式法则:

$$\Pr(AB|C) = \frac{\Pr(ABC)}{\Pr(C)} = \frac{\Pr(B|AC)\Pr(AC)}{\Pr(C)} = \Pr(B|AC)\Pr(A|C).$$

回忆补事件公式: $\Pr(A|C) + \Pr(\overline{A}|C) = \Pr(\Omega|C) = 1$. 这两个公式恰好对应了合情推理的与规则以及否定规则!

这并不是巧合, 实际上, 合情推理与事件域的公理化概率具有一一对应的关系, 见表 1.1.

从这个意义上说, 概率是似然唯一的数学模型! 从此, 我们将似然 $f(A|B)$ 定义为概率 $\Pr(A|B)$.

§1.2.3 先验与基率谬误

在前面, 我们有意模糊了条件概率和概率在合情推理中的区别. 然而, 这样的区别是非常重要的. 在合情推理中, 非条件的概率被称为先

		公理化概率论	合情推理
	事件		命题
	(条件) 概率		似然
	链式法则		与规则
	补事件公式		否定规则

表 1.1: 概率论与合情推理的对应

验概率或者基率，它表示了对这个命题合理程度的一种无条件的信念。对应地，条件概率就是后验概率或似然，它表示了对合情推理合理程度的一种信念。先验概率和后验概率有若干相互转化的公式。

条件概率有全概率公式：

定理 1.4 (全概率公式) 设 A_i 是彼此互斥的事件， $\cup_i A_i = \Omega$ ，那么

$$\Pr(B) = \sum_i \Pr(B|A_i) \Pr(A_i).$$

全概率公式表明了如何使用似然建立起不同先验概率之间的联系。

条件概率还有 Bayes 定理：

定理 1.5 (Bayes 定理)

$$\Pr(A|B) = \Pr(B|A) \frac{\Pr(A)}{\Pr(B)}.$$

Bayes 定理表明了两个不同的后验概率如何基于先验概率相互转化。在合情推理中，这表明了前提推结果的强三段论和结果推前提合理性的弱三段论之间的关系。

下面我们看一个合情推理的例子。一辆出租车在夜间发生了一起肇

事逃逸事故. 这座城市有两家出租车公司, 紫色和青色. 这个城市历史上肇事逃逸车辆 85% 是紫色的, 15% 是青色的. 一名目击者指认出租车是青色的, 这里的指认并不一定正确. 考虑一种理想的假设, 法庭知道这位证人 80% 的概率能正确识别颜色, 20% 的概率会把颜色识别错. 问: 事故车辆是那种颜色的可能性更大? 忽略先验概率会产生答案是青色车的结论.

下面我们考虑先验概率再次做计算. 记 C 为肇事逃逸的出租车为青色, P 为肇事逃逸的出租车为紫色, R 为目击者指认出租车是青色. 先验概率为 $\Pr(C) = 0.15$, $\Pr(P) = 0.85$. 似然为 $\Pr(R|C) = 0.8$, $\Pr(R|P) = 0.2$. 利用全概率公式, $\Pr(R) = \Pr(C)\Pr(R|C) + \Pr(P)\Pr(R|P) = 0.29$.

利用 Bayes 定理, $\Pr(C|R) = \Pr(R|C)\Pr(C)/\Pr(R) \approx 0.41$. 然而, $\Pr(P|R) = \Pr(R|P)\Pr(P)/\Pr(R) \approx 0.59$. 因此我们更应该倾向于认为肇事逃逸的出租车是紫色的!

上面的例子表明人类决策有一种忽略先验概率的倾向, 这种倾向被称为基率谬误, 已被大量心理学研究所证实 [KT73, TK74, Bar80]. 为了研究这种谬误, 我们将先验概率也纳入合情推理的模型之中. 这样, Kolmogorov 的概率论和合情推理就完全一致了. 反之, 如果我们将概率论理解为如上合情推理的数学模型, 我们就得到了 Bayes 概率论.

注. 本节主要讨论了概率论的“信念”和“主观”解释, 这是 Bayes 学派的基本观点. 这一观点有很多经典的论述, 例如 de Finetti 所著 [de 17], 此外, Cox 也有文章 [Cox46] 建立了似然的推理法则, 这正是本节的主要内容. 将这些观点集大成者则是 Jaynes 所著的教材 [Jay02], 他详细讨论了以似然为基础的概率论, 以及它是如何作为科学的逻辑基础的. 概率论还有其他很多解释 [Há23], 例如频率学派, 它认为概率是事件发生的频率, 这一观点在统计学中有很多应用.

§1.3 合情推理的归纳强论证

§1.3.1 归纳强论证

在基率谬误的例子中，我们看到合情推理必须要完整地考虑先验的影响。另一方面，我们看到最终做出决策的方式是最大似然，即似然更高的那个命题更有可能是对的。这说明合情推理中有很不同于演绎推理的论证方式。

我们首先指出合情推理包含了命题逻辑中的两个强三段论：

$$\begin{array}{c} A \rightarrow B \\ A \text{ 是真的} \\ \hline B \text{ 是真的} \end{array} \qquad \begin{array}{c} A \rightarrow B \\ B \text{ 是假的} \\ \hline A \text{ 是假的} \end{array}$$

我们以第一个三段论为例，记 $C \equiv A \rightarrow B$ 。由链式法则， $\Pr(B|AC) = \Pr(AB|C) / \Pr(A|C)$ 。 $A \rightarrow B$ 意味着作为事件 $A \subseteq B$ ，即 $AB = A$ ， $\Pr(AB|C) = \Pr(A|C)$ 。代入上式得到 $\Pr(B|AC) = 1$ ，这就是说，当 A 为真时， B 也为真。

除了演绎逻辑中的强三段论，合情推理还包含了弱三段论的定量形式：

$$\begin{array}{c} A \rightarrow B \\ B \text{ 是真的} \\ \hline A \text{ 变得更合理} \end{array}$$

$\Pr(A|C)$ 是 A 的似然，而 $\Pr(A|BC)$ 是假设 B 为真时， A 的似然。由链式法则， $\Pr(A|BC) = \Pr(A|C) \Pr(B|AC) / \Pr(B|C)$ 。因为 $\Pr(B|AC) = 1$ 且 $\Pr(B|C) \leq 1$ ，所以 $\Pr(A|BC) \geq \Pr(A|C)$ 。也就是当 B 为真时， A 的

合理程度会变大.

仿照三段论，现在我们将演绎推理中的若干概念推广到合情推理中.

回忆记号 $X \vdash Y$ 表示从前提为 X 出发可以根据推导规则得到结论 Y . 此时，我们说从 X 到 Y 的过程是一个有效论证，或者说 Y 是 X 的逻辑结论. 它与以下三个表述等价：

- $X \vdash Y$.
- $X \rightarrow Y$ 是重言式.
- $X \wedge \neg Y$ 是矛盾式.

等价性证明由蕴含的推导法则和完备性定理得出. 那么，如何在合情推理中定义类似的概念呢？为此，我们引入随机真值表.

回忆，事件是命题的集合论描述. 合情推理中，每个事件被赋予一个概率（似然），对应的命题也会被赋予同样的概率. 于是对应于演绎推理中的语义真值表，合情推理中有随机真值表，例如

Pr	A	B	$A \vee B$
0.4	⊤	⊤	⊤
0.2	⊤	⊥	⊤
0.25	⊥	⊤	⊤
0.15	⊥	⊥	⊥

在合情推理中，我们也有和有效论证对应的归纳强论证. 考虑推理 $X \implies Y$ ，即从 X 出发推出 Y ，我们希望描绘这一推理是归纳强论证的条件. 利用随机真值表，我们可以尝试定义归纳强论证为 $X \wedge \neg Y$ 的

不太可能为真（或 $\neg X \vee Y$ 很可能为真）. 然而我们会看到，仅仅用随机真值表得到的概念是不符合合情推理的直觉的. 我们通过两类例子来引入归纳强论证的限制条件.

例 1.9 (奇怪的例子一) 记 X 为一个北京大学的同学今年 2000 岁, Y 为一个北京大学的同学今年 2000 岁, 并且有三条胳膊. 直观来讲, 如上 $X \implies Y$ 不是归纳强论证. 但是, 等效的 $\neg X \vee Y$, 成立的概率足够大.

考虑另外一个有效论证中的例子. 记

$$A_n = \{\text{圆周率 } \pi \text{ 的系数有 } n \text{ 个连续的 } 1\}.$$

那么, 推理 $A_{n+1} \implies A_n$ 看起来是显然的, 但等效的 $\neg A_{n+1} \vee A_n$ 却不太容易被一眼接受. \square

这两个例子都说明, 判断是否为归纳强论证不能只关注结论成立的概率. 于是, 我们得到限制条件: 证据支持, 它的定义如下.

定义 1.6 (证据支持) 假设 X 和 Y 是命题公式, 实数 $t \in [0.5, 1]$. 如果 $\Pr(Y|X) > t$, 我们称证据 X 支持 Y 的强度大于 t . \square

证据支持是比最大似然准则的更进一步的要求. 然而, 证据支持并不能解决所有问题. 考虑下面的例子.

例 1.10 (奇怪的例子二) 记 X 为小明是一位北京大学的学生, Y 为小明不会飞. 表面看来, $\Pr(Y|X) = 1$, 但直觉上, $X \implies Y$ 并不应该是归纳强论证. 问题出在了 $\Pr(Y)$ 本身就等于 1, 所以 $\Pr(Y|X) = 1$ 并没有什么实际意义. \square

从这个例子出发, 我们得到另一个限制条件: 正相关性, 它的定义如下.

定义 1.7 (相关性) 对随机变量 X 和 Y , 如果 $\Pr(Y|X) > \Pr(Y)$, 我们称 X 与 Y 正相关. 等价地, 示性函数 $I(X)$ 和 $I(Y)$ 相关系数大于 0. 类似地, 如果 $\Pr(Y|X) < \Pr(Y)$ (或 $I(X)$ 和 $I(Y)$ 的相关系数小于 0), 那么 X 和 Y 负相关. 如果 $\Pr(Y|X) = \Pr(Y)$ (或 $I(X)$ 和 $I(Y)$ 的相关系数等于 0), 那么 X 和 Y 不相关. 在归纳强论证中, 我们要求 X 和 Y 正相关. \square

加上以上两个限制条件, 我们可以得到归纳强论证的严格定义.

定义 1.8 (归纳强论证) 如果推理 $X \implies Y$ 满足以下三个条件, 我们称之为归纳强论证:

- X 证据支持 Y : $\Pr(Y|X) > 0.5$.
- X 与 Y 正相关: $\Pr(Y|X) > \Pr(Y)$.
- $X \rightarrow Y$ 不是有效论证. \square

不将有效论证定义为归纳强论证得原因之一是: 一个论证可以在前提 X 是矛盾式时有效. 例如: $P \wedge \neg P \vdash Q$ 是有效论证. 但是, 由于 $\Pr(P \wedge \neg P) = 0$, $\Pr(Q|P \wedge \neg P)$ 是无定义的, 所以 $P \wedge \neg P$ 并不证据支持 Q , 也不和 Q 正相关.

进一步, 我们还希望能够衡量前提 X 在多大程度上确认结论 Y 成立, 我们可以通过如下两个量来衡量:

定义 1.9 (认可度) 考虑命题 X 和 Y . 我们定义认可概率增量和认可度似然比如下:

- 认可概率增量 $d(X, Y)$: X 发生后给 Y 的发生增加了多大的概率.

$$d(X, Y) = \Pr(Y|X) - \Pr(Y).$$

- 认可度似然比 $\ell(X, Y)$: Y 发生时 X 的似然会比 Y 没发生时 X 的似然增加多少. 该差值越大表示观测到 X 的话越应该发生了 Y . 分母归一化使得 $\ell(X, Y) \in [-1, 1]$.

$$\ell(X, Y) = \frac{\Pr(X|Y) - \Pr(X|\neg Y)}{\Pr(X|Y) + \Pr(X|\neg Y)}. \quad \square$$

这两个认可度的定义其实相互可以转化 (见习题 13). 利用这一转化, 认可度和相关性的关系可以表达如下:

命题 1.1 设 $0 < \Pr(X), \Pr(Y) < 1$, 下列等价成立:

- X 和 Y 正相关 $\iff d(X, Y) > 0 \iff \ell(X, Y) > 0$.
- X 和 Y 不相关 $\iff d(X, Y) = 0 \iff \ell(X, Y) = 0$.
- X 和 Y 负相关 $\iff d(X, Y) < 0 \iff \ell(X, Y) < 0$.

认可度与有效论证的关系可以表达如下:

命题 1.2 设 $0 < \Pr(X), \Pr(Y) < 1$, 那么

$$d(X, Y) = \begin{cases} \Pr(\neg Y), & \text{如果 } X \vdash Y, \\ -\Pr(Y), & \text{如果 } X \vdash \neg Y. \end{cases}$$

$$\ell(X, Y) = \begin{cases} 1, & \text{如果 } X \vdash Y, \\ -1, & \text{如果 } X \vdash \neg Y. \end{cases}$$

§1.3.2 有效论证和归纳强论证的比较

考虑一个论证 $X \implies Y$, 我们已经有三种方式评估 X 如何支持 Y :

1. $X \implies Y$ 是一个演绎推理: $X \vdash Y$.
2. $X \implies Y$ 是一个合情推理, X 证据支持 Y : $\Pr(Y|X) > 0.5$.
3. $X \implies Y$ 是一个合情推理, X 正相关于 Y : $\Pr(Y|X) > \Pr(Y)$.

其中, 1 对应有效论证, 2 和 3 都是合情推理中的归纳强论证的必要条件. 我们将进一步讨论有效论证和归纳强论证的一些不同之处.

首先, 有效论证具有单调性: 论证的有效性随着前提的增加不会下降. 即: 对于任意 X, Y, Z , 若 $X \vdash Y$, 则 $X, Z \vdash Y$. 然而合情推理中, 单调性不再存在. 请看下面的例子.

例 1.11 (非单调性: 例子) 考虑命题 X, Y, Z 和对应的随机真值表.

Pr	X	Y	Z	$X \wedge Z$
0.1	⊤	⊤	⊤	⊤
0.2	⊤	⊤	⊥	⊥
0.2	⊤	⊥	⊤	⊤
0	⊤	⊥	⊥	⊥
0.1	⊥	⊤	⊤	⊥
0.1	⊥	⊤	⊥	⊥
0.1	⊥	⊥	⊤	⊥
0.2	⊥	⊥	⊥	⊥

$\Pr(Y|X) = (0.1 + 0.2)/(0.1 + 0.2 + 0.2 + 0) = 0.6 > 0.5$, 然而,
 $\Pr(Y|X \wedge Z) = (0.1)/(0.1 + 0.2) = 1/3 < 0.5$. 因此, X 证据支持 Y ,
但 $X \wedge Z$ 不支持 Y . □

接下来，我们用 Z 论证 Y ，将 X 看作某种附加的条件，我们考虑 X 对 Y 这一论据的影响。在演绎推理中，若 $Z, X \vdash Y$ 和 $Z, \neg X \vdash Y$ 都满足，则 $Z \vdash Y$ 。换言之，不论补充论据 X 还是 $\neg X$ ，从 Z 出发都可以论证出 Y 。

如果类比到合情推理中呢？这就涉及到确凿性原则：如果不论条件在 X 还是 $\neg X$ ， Z 都是 Y 的一个“好的论据”，那么 Z 就是 Y 的一个“好的论据”。与之相对应的是无条件确凿性原则：如果 $Z \wedge \neg X$ 和 $Z \wedge X$ 都是 Y 的一个“好的论据”，那么 Z 就是 Y 的一个“好的论据”。“好的论据”可以从证据支持和正相关性两方面考虑。

在任何随机真值表中，如果 $\Pr(Y|Z \wedge X) > 0.5$ 且 $\Pr(Y|Z \wedge \neg X) > 0.5$ ，那么根据全概率公式， $\Pr(Y|Z) > 0.5$ 。因此，从证据支持角度，确凿性原则是成立的。而同样的论证也说明，从证据支持角度，无条件确凿性原则也是成立的。

在任何随机真值表中，如果 $\Pr(Y|Z \wedge X) > \Pr(Y)$ 且 $\Pr(Y|Z \wedge \neg X) > \Pr(Y)$ ，那么同样根据全概率公式， $\Pr(Y|Z) > \Pr(Y)$ 。因此，从正相关性角度，无条件确凿性原则是成立的。那么，从正相关性角度，确凿性原则成立吗？

注. 在讨论正相关的时候，无条件确凿性原则和确凿性原则容易被混淆。在无条件确凿性原则中，我们要求 $Z \wedge X$ 和 $Z \wedge \neg X$ 都是 Y 的好论据，这表现为 $\Pr(Y|Z \wedge X) > \Pr(Y)$ 且 $\Pr(Y|Z \wedge \neg X) > \Pr(Y)$ 。而在确凿性原则中，我们要求不论条件在 X 还是 $\neg X$ ， Z 都是 Y 的好论据。这一定义其实引入了两个新的概率，即 $Q^+(\cdot) = \Pr(\cdot|X)$ 和 $Q^-(\cdot) = \Pr(\cdot|\neg X)$ 。他们的定义为

$$Q^+(\cdot) = \frac{\Pr(\cdot \wedge X)}{\Pr(X)}, \quad Q^-(\cdot) = \frac{\Pr(\cdot \wedge \neg X)}{\Pr(\neg X)}.$$

于是，确凿性原则中涉及到了三个正相关：在前提中，我们假设 $Q^+(Y|Z) >$

$Q^+(Y)$ 和 $Q^-(Y|Z) > Q^-(Y)$, 而在结论中, 我们希望有 $\Pr(Y|Z) > \Pr(Y)$. 这三个正相关的表述其实涉及了三个不同的概率. 第一个正相关表述等价于 $\Pr(Y|Z \wedge X) > \Pr(Y|X)$, 第二个正相关表述等价于 $\Pr(Y|Z \wedge \neg X) > \Pr(Y|\neg X)$, 因此, 他们是完全不同于第三个正相关的.

实际上, 并不一定成立! 有反直觉的例子: 存在 X, Y, Z 和对应的随机真值表, 使得

- $\Pr(Y|Z \wedge X) > \Pr(Y|X)$.
- $\Pr(Y|Z \wedge \neg X) > \Pr(Y|\neg X)$.
- 然而, $\Pr(Y|Z) \leq \Pr(Y)$.

这样的现象和 *Simpson 悖论* [Sim51, BHO75] 有关.

举个例子, 球员甲的两分球和三分球命中率均高于球员乙, 但是球员甲的总投篮命中率却可能低于乙. 具体来说, 有一个班中一半同学来自北京大学, 另一半来自清华大学, 我们抽出一名同学 Bob, 估计 Bob 投篮命中的概率. 我们给如下记号:

- 记 Y 为 Bob 投篮命中, 记 X 为 Bob 投出一个两分球, 则 $\neg X$ 为 Bob 投出一个三分球 (我们这里只考虑有两分和三分球) .
- 记 Z 为 Bob 来自清华大学.
- $\Pr(Y)$ 表示全班学生的投篮命中率.
- $\Pr(Y|Z)$ 表示全班来自清华大学的学生的投篮命中率.
- $\Pr(Y|X)$ 表示全班学生的两分命中率.
- $\Pr(Y|\neg X)$ 表示全班学生的三分命中率.

- $\Pr(Y|Z \wedge X)$ 表示这个班来自清华大学的学生的两分命中率.
- $\Pr(Y|Z \wedge \neg X)$ 表示这个班来自清华大学的学生的三分命中率.

考虑这样一个投篮数据的实例：

	全班同学	来自清华大学 Z
两分球 X	50/100	6/10
三分球 $\neg X$	1/101	1/100
总命中率	51/201	7/110

根据表格可以得到，投篮命中率为 $\Pr(Y) = 51/201$, $\Pr(Y|Z) = 7/110$.
 两分球的命中率为 $\Pr(Y|X) = 50/100 = 1/2$, $\Pr(Y|Z \wedge X) = 6/10 = 3/5$. 三分球的命中率为 $\Pr(Y|\neg X) = 1/101$, $\Pr(Y|Z \wedge \neg X) = 1/100$.

Simpson 悖论在这一实例下表现为：来自清华大学的学生两分命中率和三分命中率分别都比全班平均水平高，但总体投篮命中率反倒比全班水平低.

我们将这两个概率用全概公式展开来寻找原因：

$$\Pr(Y|Z) = \Pr(Y|Z \wedge X) \Pr(X|Z) + \Pr(Y|Z \wedge \neg X) \Pr(\neg X|Z),$$

$$\Pr(Y) = \Pr(Y|X) \Pr(X) + \Pr(Y|\neg X) \Pr(\neg X).$$

上下两行对应的项， $\Pr > \Pr$. 然而，关键是有可能发生 $\Pr(X|Z) \neq \Pr(X)$. 在上面篮球的例子中表现为清华大学的同学选择投两分球和三分球的比例和全班同学不同. 这表明，Simpson 悖论发生的核心原因与基率谬误类似，即没有正确区分先验概率和后验概率.

接下来我们考虑合取谬误（见本章开头）. 合取谬误是一种认知偏

差. 我们这里呈现一个简化的例子作为讨论. Linda 是一位单身、外向且年龄为 31 岁的女性. 在大学期间, 她主修哲学, 十分关注种族歧视和社会公正问题, 而且曾参加过反核游行. (记为 E) 请问以下哪一件事情更可能发生?

1. Linda 是一名银行出纳员 (记为 B) .
2. Linda 是一名银行出纳员, 还积极参加女权运动 (记为 $B \wedge F$) .

在调查实验中, 多数被试选择了 2. 但是, 我们可以肯定 $\Pr(B \wedge F|E) \leq \Pr(B|E)$. 如何理解这一现象?

为了理解这种谬误产生的原因, 考虑合取原则: 如果 E 是 $P \wedge Q$ 的“好论据”, 那么 E 也是 P 的“好论据”, 也就是说, 如果我可以论证两件事同时成立, 那么我也一定可以论证其中一件成立. 在演绎推理中, 因为 $E \rightarrow (P \wedge Q) \vdash E \rightarrow P$, 所以合取原则成立.

类似确凿性原则, 在合情推理中, 合取原则也可以从证据支持和正相关性两方面考虑. 从证据支持的角度, 合取原则成立, 这是因为, 如果 $\Pr(P \wedge Q|E) > 0.5$, 那么 $\Pr(P|E) \geq \Pr(P \wedge Q|E) > 0.5$.

然而, 从正相关性的角度, 合取原则未必成立. 也就是说, 假设 $\Pr(P \wedge Q|E) > \Pr(P \wedge Q)$, 不一定能推出 $\Pr(P|E) > \Pr(P)$. 当人们给定对 Linda 的描述 E 的时候, 很容易建立起 E 和 $B \wedge F$ 的正相关性. 然而这并不意味着 E 和 B 是正相关的! 因此发生了合取谬误. 从 Simpson 悖论和合取谬误可以看出, 只依靠正相关性进行推理很容易犯错误, 因此证据支持 (极大似然) 是归纳强论证不可缺少的要素.

总结一下, 我们讨论了单调性、确凿性原则和合取原则, 他们是演绎推理中有效论证成立的性质. 但是, 在合情推理中, 归纳强论证并不一定有这些性质. 具体结论总结如表 1.2 所示.

性质	证据支持	正相关
单调性	无	习题 10
无条件确凿性原则	有	有
确凿性原则	有	无
合取原则	有	无

表 1.2: 有效论证和归纳强论证的比较

注. 本节的内容主要讨论了归纳强论证，对本节感兴趣的读者，可以进一步参考教材 [Sky75] 以及课程讲义 [Pac]，了解更多这方面的内容。

§1.4 先验模型的存在性

我们现在讨论合情推理的另一个核心：先验的存在性。事实上，基于先验的推理不仅是 Bayes 概率论的基础，也是人类认知世界的基础。我们从出生的第一天就在建立和更新自己的先验模型：我们在蹒跚学步的过程中，逐渐理解了对于空间深度的感知 [WG61]；在学习语言的时候，我们通过眼睛观察、耳朵聆听、躯体触碰，不断建立对于物体和动作的概念 [VWOL24, SAN96]。大量发展心理学和认知科学的研究表明，人类通过无数与世界的交互，建立了对于整个世界的先验认知模型。

从更宏观的角度来看，历史上，绝大部分成功的科学理论也都是依赖大量实验与观察数据的。我们就以家喻户晓的 Newton 万有引力定律为例。这一定律表明，两个物体之间的引力与他们的质量分别成正比，与他们的距离的平方成反比。我们总是听到那个美妙的故事，说 Newton 因为被苹果砸到才发现了这一定律。然而，这个故事忽略了这

一定律背后的伟大科学家, Kepler, 他总结老师 Tycho 的海量肉眼观测数据, 提出了行星运动的三定律. 没有这些定律, Newton 也不可能准确写出平方反比定律.

总的来说, 数据驱动产生理论和先验模型的过程在人类个人和整体的发展上都至关重要. 几乎同样的技术和想法被广泛应用在深度学习中, 最大的区别是, 人类, 或者科学家, 被复杂的人工神经网络代替.

不论是人的认知过程、科学理论还是深度学习, 数据驱动方式的核心假设是, 确实存在一个世界的先验模型, 我们通过数据来逼近这个模型. 本节将要给出一个重要的定理, *de Finetti 定理* [DF80, KS06, Dur19], 它表明在一定合理的条件下, 无限多的数据本身一定蕴含了一个先验模型. 因此, 数据驱动的模式是合理的, 这也成为了 Bayes 概率论的基石之一.

接下来, 我们建模这一问题. 首先是数据, 观测数据需要遵循可重复性原则. 假如我们观察到了数据 x_1, x_2, \dots, x_n , 由于观测有不可控因素, 观测结果会有随机性, 所以我们假设 X_1, \dots, X_n 是观测结果 x_1, \dots, x_n 所对应的随机变量. 从合情推理的角度来说, 随机性更像是人对于观测不确定性的一种认识或信念. 所谓的可重复就是指, 这些实验不管以什么样的顺序完成, 他们都应该产生相同的结果. 假设 π 是一个 $\{1, 2, \dots, n\}$ 的排列, 我们要求

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}).$$

也就是说, 考虑了随机性之后, 打乱观测的顺序, 观测结果的概率分布应该是不变的. 随机变量列这样性质的被称可交换性.

那么, 什么是理论模型呢? 一个定量的理论模型可以看成某种对真实世界的参数化建模, 例如参数为 θ , 它能够解释 X_i 中的某些规律. 如

果我们知道了真实世界的 θ , 那么观测结果 X_i 之间不仅仅是可交换的, 而且是独立的. 这里的意思是, 科学理论所能解释的部分应该使得 X_i 的不确定性不再依赖于其他的 X_j , 而仅仅在于一次观测内部. 这样的性质被称为条件独立性.

接下来的问题是, 我们是否可以发现一个理论模型? 更直接的问题是, 如果给我们足够多的数据, 我们是否能够直接形成一个对于 θ 的先验信念 (即概率密度 p_θ), 而不需要再人为有某种特别的先验信念? 这就是 de Finetti 定理要回答的问题.

现在我们正式陈述这一定理. 首先, 让观测次数为无穷, 即一列 X_1, X_2, \dots . 此时, 可交换性指的是任意有限个 $X_{i_1}, X_{i_2}, \dots, X_{i_n}$ 以及任意一个集合 $\{i_1, \dots, i_n\}$ 上的排列 π 都有

$$(X_{i_1}, X_{i_2}, \dots, X_{i_n}) \stackrel{d}{=} (X_{\pi(i_1)}, X_{\pi(i_2)}, \dots, X_{\pi(i_n)}).$$

此时, 我们有 de Finetti 定理:

定理 1.6 (de Finetti 定理) 如果 X_1, X_2, \dots 是可交换的连续型随机变量列, 那么存在一个概率密度 p_θ , 使得对于任意的 n 和任意的 x_1, \dots, x_n , 都有

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \int p_\theta(\theta) \prod_{i=1}^n p_{X_i}(x_i | \theta) d\theta.$$

我们来解释这一定理与上面讨论的对应. 如果我们知道了 θ , 那么 X_1, \dots, X_n 条件在 θ 上的条件密度是 $\prod_{i=1}^n p_{X_i}(x_i | \theta)$, 因而是条件独立的, 这符合了我们对于理论模型的基本要求. 另外, 如果我们不知道 θ , 只知道无穷多的观测数据, 我们并不需要人为再给定一个关于 θ 的先验, 数据本身就已经足够形成先验了, 不需要人为的先验.

注. 如果我们把观测数据这件事看成人的主观概率（即似然），那么，de Finetti 定理说明，如果我们主观上认为一系列事件是可交换的（即事件的顺序不影响我们的似然判断），那么这等同于假设这些事件是独立的，并且存在一个客观的概率。我们对这个客观概率的不确定性可以通过主观的概率分布来描述。这意味着，即使概率是主观的，我们也可以像它们是由客观概率驱动的那样行动。这一理解为 de Finetti 著名的格言“概率是不存在的” [Spi24] 提供了实用的理解.

§1.5 习题

1. 请自己定义命题逻辑的推导规则，并在你的系统中证明以下命题：

- (1) $\phi \rightarrow (\psi \rightarrow \phi)$.
- (2) $(\phi \rightarrow \psi) \rightarrow (\neg\psi \rightarrow \neg\phi)$.
- (3) $(\neg\psi \rightarrow \neg\phi) \rightarrow (\phi \rightarrow \psi)$.

上面哪些推理需要使用 RAA?

2. 将习题 1 中的命题用真值表的方式证明他们是重言式.
3. 考虑两个事实，AI 中的数学考试只会做题的人必然考不了 70 分；潘数同学 AI 中的数学考了 100 分. 回答以下问题：
 - (1) 从中构造出一个强三段论的论证.
 - (2) 从中构造出一个弱三段论的论证.
4. [Cox46] 本题证明定理 1.2.

(1) 利用似然的等值原则和假设 1.1 证明: F 满足

$$F(F(x, y), z) = F(x, F(y, z)). \quad (1.1)$$

(2) 证明: 当 F 满足 $cf(F(p, q)) = f(p)f(q)$ 时, F 满足 (1.1).

(3) * 假设 F 是连续二阶可微的, 证明: 如果 F 满足 (1.1), 那么存在一个常数 c 和函数 f , 使得 F 满足 $cf(F(p, q)) = f(p)f(q)$.

5. [Cox46] 本题证明定理 1.3. 我们在推导否定规则的时候得到了如下的公式

$$S(x) = (1 - x^m)^{\frac{1}{m}}. \quad (1.2)$$

这个条件实际由下面的关系式联合推出

$$S[S(x)] = x. \quad (1.3)$$

$$xS[S(y)/x] = yS[S(x)/y]. \quad (1.4)$$

(1) 利用似然的等值原则, 证明: S 满足 (1.3).

(2) 利用似然的等值原则以及与规则, 证明: S 满足 (1.4).

(3) * 假设 S 是二阶可微的, 证明: S 满足 (1.2).

6. 利用似然的三个规则将似然 $[(A \vee \neg B) \rightarrow (\neg C \wedge D)]|E$ 写为只含 1 和与命题的似然 (形如 $A|B$, $AB|C$ 等, 但不能是 $\bar{A}B|C$ 这种) 的式子.

7. 甲袋中有 $N - 1$ 只白球和 1 只黑球, 乙袋中有 N 只白球, 每次从甲、乙两袋中分别取出一只球并交换放入另一袋中. 这样经过了 n 次,

- (1) 求黑球出现在甲袋中的概率 p_n .
- (2) 求 $\lim_{n \rightarrow \infty} p_n$.
8. 仿照第 1.3.1 节, 用似然解释另一个弱三段论:

$$\begin{array}{c} A \rightarrow B \\ A \text{ 是假的} \\ \hline B \text{ 的合理程度变小} \end{array}$$

9. 设 $I(A)$ 表示事件 A 的示性函数, 对于事件 X 和 Y , 证明: $I(X)$ 与 $I(Y)$ 相关系数为 0 等价于 X 与 Y 互相独立.
10. 从正相关的角度, 归纳强论证有单调性吗? 证明你的结论.
11. 回答以下问题.
- (1) 证明: 从正相关性角度, 无条件确凿性原则成立, 确凿性原则不一定成立.
- (2) 请构造一个“确凿性原则不一定成立”的例子并画出对应的随机真值表.
- (3) Simpson 悖论和上述哪一个原则有关?
12. 在合取谬误的部分, 我们提到 E 和 $P \wedge Q$ 正相关, 并不一定能推导出 E 和 P 正相关. 请构造出一个这样的例子并画出对应的随机真值表.
13. 设 $0 < \Pr(X), \Pr(Y) < 1$. 计算认可概率增量

$$d(X, Y) = \Pr(Y|X) - \Pr(Y) \tag{1.5}$$

与认可度似然比

$$\ell(X, Y) = \frac{\Pr(X|Y) - \Pr(X|\neg Y)}{\Pr(X|Y) + \Pr(X|\neg Y)} \quad (1.6)$$

的关系，并证明二者始终同号。

14. * 举一个可交换但是不相互独立的连续型随机变量列的例子，并构造出 de Finetti 定理中对应的概率密度 p_θ .

第二章 Markov 链与模型

我们都知道，人的推理和决策会受到时间的影响：我们宁愿要现在的十块钱也不要十年后的二十块钱。然而，人和人的耐心程度是不一样的。2010 年，来自意大利博洛尼亚大学 Manuela Sellitto 和她的同事进行了一项实验 [SCdP10]。他们的实验对象是一组大脑正常的人（对照组）、一组非眶额叶区受损的病人（实验组 1）和一组眶额叶周围受损的病人（实验组 2）。实验人员对被试进行提问，典型的问题如下：“你想要现在拿到 10 美元还是一个月后拿到 12 美元？”他们对食物和货币形式的奖励都进行了测试。

结果显示，眶额叶周围受损的病人更倾向于选择即时的奖励，而其他组的人更倾向于选择未来的奖励。如果货币奖励翻倍（例如，从 50 美元变为 100 美元），对照组和实验组 1 的人愿意等待 4 到 6 个月，而眶额叶周围受损的人（实验组 2）甚至不愿意等待 3 周。换句话说，人的耐心程度是大脑结构决定的，而不是自己可以轻易改变的！

以上例子不仅说明时间在决策中的重要性，更说明人脑中有特殊的结构和机制处理时间相关的决策问题。这样的观察对于人工智能也是同样重要的。本章我们将介绍 Markov 链，这是一种带有时间的概率模型，它是应用最为成功的含时决策模型之一。我们还将介绍人工智能领

域基于 Markov 链的各种应用，包括 Markov 决策模型与强化学习、隐 Markov 模型以及扩散模型。

§2.1 Markov 链

在第一章中，我们说明了合情推理是人和 AI 非常重要的推理方式，这一推理模式基于 Bayes 概率论和似然。然而，这一模型对推理的假设是逻辑的、静态的，时间的概念并不出现在似然里面。例如，考虑一个罐子，里面有除颜色之外不可区分的 N 个球，有 n 个白球，剩下的是黑球。顺序从中拿出 N 个球，第 k 次拿出的球颜色是 W_k 或 B_k 。

用 Bayes 定理很容易证明，对任意 $i \neq j$ ， $\Pr(W_i|W_j) = \Pr(W_j|W_i)$ 。也就是说， $\Pr(W_i|W_j)$ 和 $\Pr(W_j|W_i)$ 不仅是可计算的，而且是相等的。然而，如果从推理的角度来说，我们基于时间更晚的状态推理时间更早的状态，这样的推理需要我们能够有对未来的模型。因此，我们需要引入一个带有时间的推理模型，这就是 Markov 链。

注。从主观概率角度理解 Markov 链的这种方式，本书采纳了 Jaynes 著作 [Jay02] 中的观点。

定义 2.1 (Markov 链) **Markov 链**（马氏链）是一列随机变量 $\{X_t\}_{t=0}^{\infty}$ ，包含如下概念：

- 状态空间 \mathcal{S} : X_t 所有可能值构成的集合，有限或者可数。
- 转移矩阵 \mathcal{P} （转移核）：下一时刻系统状态之间转移的概率。 $\mathcal{P} = (p_{ij})_{i,j \in \mathcal{S}}$, p_{ij} 是从 i 状态转移到 j 状态的概率。
- *Markov 性*: 对任意时刻 $t = 0, 1, \dots$ 和任意状态 $j, k, j_0, \dots, j_{t-1} \in \mathcal{S}$

\mathcal{S} , 如下等式成立

$$\begin{aligned} & \Pr(X_{t+1} = j | X_t = k, X_{t-1} = j_{t-1}, \dots, X_0 = j_0) \\ &= \Pr(X_{t+1} = j | X_t = k) = p_{kj}. \end{aligned}$$

有时候也考虑带初态的 Markov 链, 此时 X_0 服从分布 $\lambda = (\lambda_s)_{s \in \mathcal{S}}$. \square

我们给出的定义是简化的 Markov 链, 每个时刻之间的转移都是一样的转移矩阵, 这样的 Markov 链被称为时齐的. 有时候也会考虑非时齐的 Markov 链 (例如扩散模型), 即每个时刻之间的转移矩阵不一样, 这样的 Markov 链被称为非时齐的, 此时 t 时刻的转移矩阵是 $\mathcal{P}^{(t)}$, 定义中 Markov 性的转移概率是 $p_{kj}^{(t)}$.

Markov 链是一种简化的带时间的概率模型, 它最重要的性质是 Markov 性, 即在固定现在的情况下, 过去与未来相互独立. 这一性质的数学表述为:

命题 2.1 (Markov 性) 条件在 $X_n = i$ 下, $\{Y_m\}_{m=0}^{\infty} := \{X_{m+n}\}_{m=0}^{\infty}$ 是一个转移矩阵为 P 的 Markov 链, 并且与 (X_0, \dots, X_{n-1}) 相互独立.

证明见习题 1.

我们考虑的 Markov 链还有时齐性, 即状态的转移不依赖当前时间, 只和当前的状态有关. 时齐性的数学表述为:

命题 2.2 设 $\{X_t\}_{t=0}^{\infty}$ 是一个 Markov 链, 那么对任意的 $t, m, n \in \mathbb{N}$ 和 $i, j, k \in \mathcal{S}$, 有 $\Pr(X_{m+n} = j | X_n = k) = \Pr(X_m = j | X_0 = k)$.

我们来看一个 Markov 链的例子.

例 2.1 (赌徒模型) 考虑公平对赌. 玩家 A 和 B 抛硬币来赌钱, A 赌正面, B 赌反面. 每一轮独立地抛硬币, 正面朝上的概率和反面朝上的概率相等, 都是 $1/2$. 赢的一方给输的一方一块钱. A 输 a 块钱破产, B 输 b 块钱破产, Z_i 是第 i 轮 A 的收入. $Z_0 = X_0 = 0$ 是 A 初始的收入. $X_n = Z_0 + \dots + Z_n$ 是 A 的累计收入. 那么, $\{X_n\}_{n \geq 0}$ 是一个 Markov 链.

- 状态空间: $\mathcal{S} = \{-a, -a+1, \dots, 0, 1, \dots, b\}$.
- 转移概率: 对 $-a < i < b-1$, $p_{i,i+1} = p_{i+1,i} = 1/2$; $p_{-a+1,-a} = p_{b-1,b} = 1/2$, $p_{-a,-a} = p_{b,b} = 1$; 其他值为 0.

转移矩阵可以画成图 2.1 所示的形式.

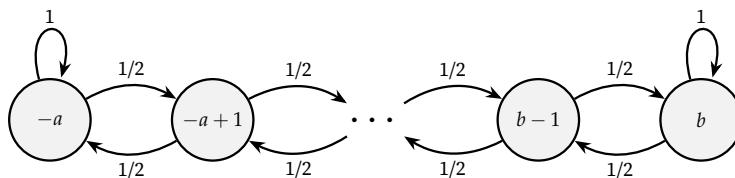


图 2.1: 赌徒模型的转移矩阵

一般地, 任何时齐可数状态的 Markov 链都可以用图表示. □

接下来, 我们来看一个非时齐的 Markov 链的例子.

例 2.2 (Pólya 的坛子) 想象现在有一个坛子. 开始的时候, 坛子里有 1 个白球和 1 个黑球. 每一轮, 我们从坛子里随机拿出一个球, 观察颜色, 然后放回去, 再加入一个和刚才拿出的球颜色相同的球. 例如, 如果第一轮我们拿出一个白球, 那么第二轮开始时坛子里有 2 个白球和 1 个黑球.

设 X_t 是第 t 轮之后白球的数量，我们会发现，它是一个非时齐的 Markov 链，转移概率为

- $\Pr(X_{t+1} = i+1 | X_t = i) = i/(t+2)$,
- $\Pr(X_{t+1} = i | X_t = i) = (t+1-i)/(t+2)$,
- 其他值的概率为 0.

这个 Markov 链的一个重要性质是当 $t \rightarrow \infty$ 时， X_t/t 趋于均匀分布 $\mathcal{U}[0, 1]$.

这个性质可以通过对 t 进行归纳证明，即 X_t 是 $\{1, \dots, t+1\}$ 上的均匀分布。 $t=0$ 时， $X_0=1$ ，显然成立。假设对 t 成立，即 X_t 是 $\{1, \dots, t+1\}$ 上的均匀分布。那么对 $t+1$ 以及 $1 \leq j \leq t+2$ ，根据全概率公式，我们有

$$\begin{aligned}\Pr(X_{t+1} = j) &= \sum_{k=1}^{t+1} \Pr(X_{t+1} = j | X_t = k) \Pr(X_t = k) \\ &= \frac{1}{t+1} (\Pr(X_{t+1} = j | X_t = j-1) + \Pr(X_{t+1} = j | X_t = j)) \\ &= \frac{1}{t+1} \left(\frac{j-1}{t+2} + \frac{t+2-j}{t+2} \right) = \frac{1}{t+2}.\end{aligned}$$

于是，归纳成立。

Pólya 的坛子还有一种有趣的解读：如果我们把人生的每一次选择都看成放球，我们经常会基于现在看到的东西（看到的黑球还是白球）投资自己的人生（放一个黑球还是白球），这样的结果是我们的人生会多姿多彩，百花齐放。□

我们回到赌徒模型. A 的累计收入 $\{X_n\}_{n \geq 0}$ 形成了 Markov 链. 根据 Markov 性, 未来双方的收入变化只取决于现在, 而和过去运气无关. 与之相关的一个现象是赌徒谬误, 即认为过去的运气会影响未来的运气. 例如, 如果一个人连续输了很多次, 那么他会认为自己未来运气会变好, 赢的概率更大. 但是, 根据 Markov 性, 过去的运气不会影响未来的运气, 因此这种想法是错误的. “风水轮流转” 在一场公平对赌中是不正确的认知. 那么, 如何评估赌局的公平性?

如果对赌是公平的, 那么我们应该认为两个人每一轮的累计收入分布都是一样的, 即

$$\Pr(X_n = i | X_0 = 0) = \Pr(X_n = -i | X_0 = 0).$$

因此, 我们需要能够计算多步转移的概率. 设 $p_{ij}^{(k)}$ 表示从状态 i 用 k 步转移到状态 j 的概率. k 步转移概率形成了一个矩阵 $\mathcal{P}^{(k)}$. 下面的定理给出了计算多步转移概率的方法.

定理 2.1 (Kolmogorov-Chapman 方程) $P^{(k+l)} = \mathcal{P}^{(k)} \mathcal{P}^{(l)}$.

证明. 由 Markov 性、时齐性和全概率公式,

$$\begin{aligned} p_{ij}^{(k+l)} &= \Pr(X_{k+l} = j | X_0 = i) \\ &= \sum_{\alpha} \Pr(X_{k+l} = j, X_k = \alpha | X_0 = i) \\ &= \sum_{\alpha} \Pr(X_k = \alpha | X_0 = i) \Pr(X_{k+l} = j | X_k = \alpha) \\ &= \sum_{\alpha} p_{i\alpha}^{(k)} p_{\alpha j}^{(l)}. \end{aligned} \quad \square$$

Kolmogorov-Chapman 方程有两个重要的特例，前向方程：

$$\mathcal{P}^{(k+1)} = \mathcal{P}^{(k)} \mathcal{P},$$

以及后向方程：

$$\mathcal{P}^{(l+1)} = \mathcal{P} \mathcal{P}^{(l)}.$$

这一过程示意见图 2.2 和图 2.3.

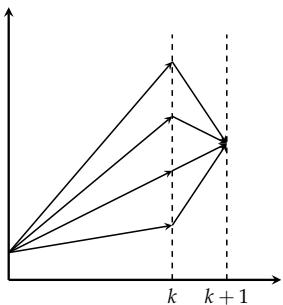


图 2.2: 前向方程 (往前一步)

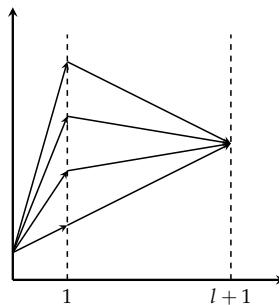


图 2.3: 后向方程 (往回一步)

此外，利用归纳法，我们还有如下推论：

推论 2.1 $\mathcal{P}^{(k)} = \mathcal{P}^k.$

若已知初始分布向量为 λ ，利用这一推论，我们可以计算 Markov 链分布随时间的演化：

$$\lambda^\top, \lambda^\top \mathcal{P}, \dots, \lambda^\top \mathcal{P}^n, \dots$$

回到赌徒模型，如何计算公平对赌中 X_n 的概率分布？我们先来看

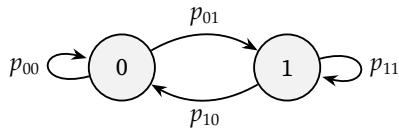


图 2.4: 只有两个状态的 Markov 链

一个简化的例子. 考虑只有两个状态 0, 1, 转移矩阵为

$$\mathcal{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

或者画成图 2.4 的形式.

可以归纳证明:

$$\begin{aligned} \mathcal{P}^n = & \frac{1}{2 - p_{00} - p_{11}} \begin{pmatrix} 1 - p_{11} & 1 - p_{00} \\ 1 - p_{11} & 1 - p_{00} \end{pmatrix} \\ & + \frac{(p_{00} + p_{11} - 1)^n}{2 - p_{00} - p_{11}} \begin{pmatrix} 1 - p_{00} & -(1 - p_{00}) \\ -(1 - p_{11}) & 1 - p_{11} \end{pmatrix}. \end{aligned}$$

假设 $|p_{00} + p_{11} - 1| < 1$, 等价地, p_{00} 和 p_{11} 不同时为 0 或同时为 1, 那么

- $\lim_{n \rightarrow \infty} p_{i0}^{(n)} = (1 - p_{11}) / (2 - p_{00} - p_{11})$,
- $\lim_{n \rightarrow \infty} p_{i1}^{(n)} = (1 - p_{00}) / (2 - p_{00} - p_{11})$.

因此, 无论初始分布是什么, 随着时间的推移, Markov 链的分布会收敛到一个同一个稳定的分布. 这个例子是否具有普遍性?

我们再考虑一个例子. 一个四元环 $\mathbb{Z}_4 = \{0, 1, 2, 3\}$ 上的 Markov

链 X_k , 转移矩阵为 $p(i, i+1) = p(i, i-1) = 1/2$, 这里的表达式将 -1 和 4 等同, 4 和 0 等同. 如果初始分布集中在 0 上, 那么我们会发现以概率 1 有 $X_{2k} \in \{0, 2\}$, $X_{2k+1} \in \{1, 3\}$. 这种情况下, 最终不会趋于一个稳定的分布! 如果初始分布等概率地分布在 $\{0, 1\}$ 上, 最终会趋于一个 \mathbb{Z}_4 上的均匀分布. 因此, 并不是所有 Markov 链都具有这样的性质. 然而, 对于相当广泛的一类 Markov 链, 这一结论成立, 这就是遍历定理.

定理 2.2 (遍历定理) 设 Markov 链的状态空间为 $S = \{1, \dots, N\}$, 转移矩阵为 $\mathcal{P} = (p_{ij})$. 如果存在某个 n_0 , 使得

$$\min_{ij} p_{ij}^{(n_0)} > 0, \quad (2.1)$$

那么存在分布 $\lambda = (\lambda_1, \dots, \lambda_N)$ 使得

$$\lambda_i > 0, \quad \sum_i \lambda_i = 1, \quad (2.2)$$

并且对于每一个 $j \in S$ 和任意 $i \in S$ 都有

$$p_{ij}^{(n)} \rightarrow \lambda_j, n \rightarrow \infty. \quad (2.3)$$

反之, 如果存在满足 (2.2) 和 (2.3) 的 λ , 则存在满足 (2.1) 的 n_0 .

最后, 在以上条件下, (2.2) 的 λ 满足

$$\lambda^T = \lambda^T \mathcal{P}. \quad (2.4)$$

条件 (2.1) 表明超过某个步数 n_0 之后, 从 i 出发到达 j 的概率总是正的, 这个条件被称为遍历. (见习题 2) 条件 (2.2) 表明每一个状态被访

问到的概率都是正的，没“死状态”。遍历定理表明遍历的 Markov 链从任何状态出发都是不可逆的，最终会把每个状态都走过一遍（遍历），变成一个混合均匀的状态。这可以用来解释物理学中的扩散现象，也是扩散模型的基础。

下面我们证明遍历定理。

证明。首先证明从 (2.1) 到 (2.2) 和 (2.3) 的过程。定义序列

$$m_j^{(n)} = \min_i p_{ij}^{(n)}, \quad M_j^{(n)} = \max_i p_{ij}^{(n)}.$$

我们先证明这两个序列是单调的。由于

$$p_{ij}^{(n+1)} = \sum_{\alpha} p_{i\alpha} p_{\alpha j}^{(n)},$$

可见

$$m_j^{(n+1)} = \min_i p_{ij}^{(n+1)} \geq \min_i \sum_{\alpha} p_{i\alpha} \min_{\alpha} p_{\alpha j}^{(n)} = m_j^{(n)},$$

因此 $m_j^{(n)} \leq m_j^{(n+1)}$ 。

类似地 $M_j^{(n)} \geq M_j^{(n+1)}$ 。

接下来我们说明， $M_j^{(n)} - m_j^{(n)}$ 会趋于零。设 $\varepsilon = \min_{i,j} p_{ij}^{(n_0)} > 0$ ，

由 Kolmogorov-Chapman 方程可得

$$\begin{aligned} p_{ij}^{(n_0+n)} &= \sum_{\alpha} p_{i\alpha}^{(n_0)} p_{\alpha j}^{(n)} = \sum_{\alpha} \left[p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \right] p_{\alpha j}^{(n)} + \varepsilon \sum_{\alpha} p_{j\alpha}^{(n)} p_{\alpha j}^{(n)} \\ &= \sum_{\alpha} \left[p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \right] p_{\alpha j}^{(n)} + \varepsilon p_{jj}^{(2n)}. \end{aligned}$$

而由于 $p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \geq p_{i\alpha}^{(n_0)} - \varepsilon \geq 0$, 可见

$$p_{ij}^{(n_0+n)} \geq m_j^{(n)} \sum_{\alpha} \left[p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \right] + \varepsilon p_{jj}^{(2n)} = m_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)},$$

最后一个等式是因为概率求和为 1. 由 i 的任意性, 左边的不等式对所有 i 都成立, 所以对最小的也成立:

$$m_j^{(n_0+n)} \geq m_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

同理, 考虑 $M_j^{(n)}$, 有

$$p_{ij}^{(n_0+n)} \leq M_j^{(n)} \sum_{\alpha} \left[p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \right] + \varepsilon p_{jj}^{(2n)} = M_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)},$$

类似可得

$$M_j^{(n_0+n)} \leq M_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

从而

$$M_j^{(n_0+n)} - m_j^{(n_0+n)} \leq (M_j^{(n)} - m_j^{(n)}) (1-\varepsilon),$$

说明当 $n \rightarrow \infty$, $M_j^{(n)} - m_j^{(n)} \rightarrow 0$, $M^{(n)}$ 和 $m^{(n)}$ 趋于同一个极限.

若记 $\pi_j = \lim_n m_j^{(n)}$, 则

$$\left| p_{ij}^{(n)} - \pi_j \right| \leq M_j^{(n)} - m_j^{(n)} \leq (1-\varepsilon)^{[n/n_0]-1},$$

即 $p_{ij}^{(n)}$ 以几何速度收敛于极限值 π_j .

因为 $m_j^{(n)} \geq m_j^{(n_0)} \geq \varepsilon > 0, n \geq n_0$, 所以 $\pi_j > 0$. 这就推出了 (2.2)

和 (2.3).

接下来，我们说明 (2.2) 和 (2.3) 可以推出 (2.1). 因为状态数有限，所以对任意充分小的 $\varepsilon > 0$ ，存在 n_0 使得对任意 i, j ，

$$\left| p_{ij}^{(n)} - \pi_j \right| \leq \varepsilon, n \geq n_0.$$

因此

$$p_{ij}^{(n)} \geq \pi_j - \varepsilon > 0, n \geq n_0.$$

最后我们说明 (2.3) 可以推出 (2.4). 注意到

$$\lim_{n \rightarrow \infty} \lambda^T \mathcal{P}^n = \left(\sum_i \lambda_i \pi_1, \sum_i \lambda_i \pi_2, \dots, \sum_i \lambda_i \pi_N \right) = \lambda^T.$$

等式两边同时右乘 \mathcal{P} ，左边的极限不变，右边变成 $\lambda^T \mathcal{P}$ ，所以 $\lambda^T \mathcal{P} = \lambda^T$. \square

满足条件 (2.4) 的分布被称为平稳分布. 用性质 $\lambda^T \mathcal{P} = \lambda^T$ 很容易说明，平稳分布为初始状态时，Markov 链的演化与时间无关：

命题 2.3 设 $\{X_n\}$ 是 Markov 链，如果 X_0 是平稳分布，那么随机变量列 (X_k, \dots, X_{k+l}) 的联合分布不依赖于 k .

如果 Markov 链是遍历的，那么平稳分布是唯一的：

命题 2.4 设 $\{X_n\}$ 是遍历的 Markov 链，那么它有唯一平稳分布 μ .

证明. 假设 μ 是另外一个平稳分布，那么 $\mu_j = \sum_\alpha \mu_\alpha p_{\alpha j} = \dots = \sum_\alpha \mu_\alpha p_{\alpha j}^{(n)}$. 因为 $p_{\alpha j}^{(n)} \rightarrow \lambda_j$ ，所以 $\mu_j = \sum_\alpha (\mu_\alpha \lambda_j) = \lambda_j$. \square

非遍历 Markov 链也可能存在（唯一）平稳分布，考虑如下转移矩

阵：

$$\mathcal{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

它有唯一平稳分布 $\lambda = (1/2, 1/2)^T$.

注. 关于 Markov 链，有很多非常好的教材可以作为参考。例如，本节的组织思路大量参考了 Shiryaev 的概率论教材 [Shi96]. 此外，Norris 的 Markov 链教材 [Nor97] 以及 Durrett 的概率论教材 [Dur19] 都是非常好的参考书。

§2.2 Markov 奖励过程 (MRP)

我们接下来的目标就是在 Markov 链上建立决策理论。很多认知科学的研究证明 [GIP18]，人和动物在做决策的时候，脑中会有一个价值系统，用来评估每一个行动的可能产生的价值/奖励。俗话说，最美味的食物是饿了一整天之后的白米饭；然而，如果我们每一顿饭都是山珍海味，就算是龙虾也会变得索然无味。这说明，我们的价值系统会随着时间和自身状态而发生改变。另一方面，本章开篇所讲的故事表明，我们的价值系统会对不同时期的奖励产生不同的反应，一般来说，我们会对即时奖励更加敏感。

总结上面的观察，我们可以在 Markov 链上引入类似的价值系统，这就是 Markov 奖励过程 (MRP)。在进入正式定义之前，我们先来看一个例子。

例 2.3 (李二的 MRP) 在一个学期中，学生李二可能处于几种状态：在教室 1 中、刷手机、在教室 2 中、约会、睡觉、考试通过、考试挂科。

学生在不同的状态下会有不同的奖励，例如，李二总是进入教室 1 逼迫自己学习，因为不情愿，所以奖励是 -2 ；但如果他被某个姑娘邀请去约会，他会很激动，所以奖励是 $+5$ 。

当处于某个状态时，李二会有一定的概率转移到另一个状态。例如，在教室 1 中，因为李二并不情愿学习，所以他会有 0.5 的概率开始刷手机，还有另外 0.5 的概率，他发现差不多该去上课了，于是进入了教室 2。简化起见，在状态转移中，我们考虑抽象的时间单位，对于李二来说，时刻只会有 $t = 0, 1, 2, \dots$

李二的人生就在这些状态之间循环往复，当他进入某个状态之后，他就会获得相应的奖励。这个过程可以被图 2.5 描述。

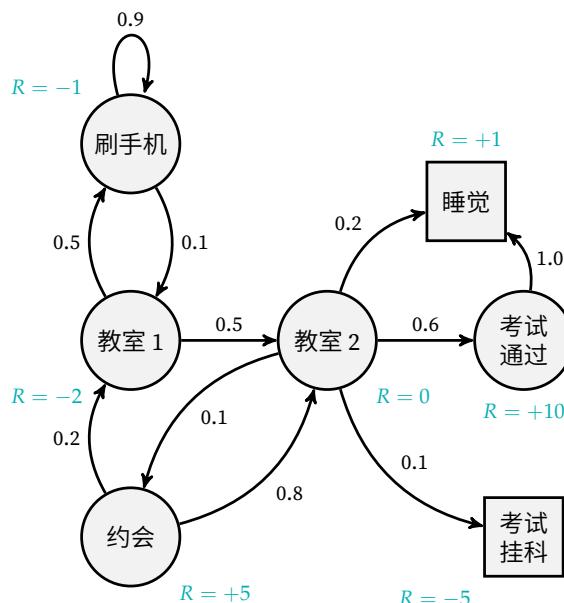


图 2.5: 学生 MRP

李二除了会获得即时奖励，他还会对未来有预期。例如，尽管李二不愿意学习，但是考试通过的奖励是 +10，为了这么大的奖励，现在遭罪一些是值得的，所以他会愿意坐在教室 1 里自习。假如下一刻李二就要到教室 1 考试，这一刻他开始在教室 1 中自习，他对于下一刻考试通过的奖励预期是 +9。也就是说，李二对未来的奖励会有一个折扣，这个折扣因子就是 $\gamma = 9/10 = 0.9$ 。□

更一般地，我们可以形式上定义 MRP。

定义 2.2 (Markov 奖励过程, MRP) 一个 **Markov 奖励过程 (Markov 奖励模型, MRP)** 由四元组 $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$ 构成：

- \mathcal{S} 是一个有穷的状态集合。
- \mathcal{P} 是一个状态转移矩阵，从 i 转移到 j 的概率记为 $P_{i,j}$ 。根据这一转移矩阵可以产生一个状态转移的 Markov 链 $\{S_t\}_{t \geq 0}$ 。
- \mathcal{R} 是（单步期望）奖励函数，定义为 $R_s = \mathbb{E}[R_{t+1} | S_t = s]$ ，其中，随机变量 R_{t+1} 表示下一阶段所处状态的奖励。也就是说，当 t 时刻位于状态 s 时， R_s 是下一时刻获得的奖励的期望。
- γ 是一个折扣因子， $\gamma \in [0, 1]$ 。□

在 MRP 中，我们最关心的并不是实时奖励，而是综合来看整个过程的奖励。为了描述这一点，我们引入回报的概念。

定义 2.3 (回报) MRP 中， t 时刻以后的总回报 G_t 定义为

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \quad \square$$

从定义中可以看到， $\gamma \in [0, 1]$ 衡量了未来下一时段单位奖励在当前时刻的价值， γ 越小，我们更偏好即时奖励，因而更“短视”； γ 越大，我们更偏好未来奖励，因而更有“远见”。

我们这里对折扣做了极大的简化。折扣因子和时间、状态都有可能有依赖关系。

- 在李二的例子中，他可能更愿意十天后有一次约会，而不是通过考试，换言之，约会对他的折扣因子可能是 0.99，而考试通过的折扣因子可能是 0.9.
- 如果老师给李二布置的作业难度波动很大，那么李二做完作业之后对通过考试的折扣因子就会发生变化：作业简单，李二觉得自己已经掌握了知识，对通过考试的折扣因子就会降低；作业困难，李二觉得自己可能很难通过考试，对通过考试的折扣因子就会提高.

然而，在 MRP 的定义中，我们假设了一个固定的折扣因子，并且规定了一个简单的方法计算折扣： t 时刻后的奖励是即时奖励的 γ^t 倍。这一定义体现了折衷的思想：如果折扣因子太符合实际，那么这个模型就不太实用。在我们的定义之下，随着时间改变，折扣因子会指数衰减，这不仅和实验结果比较符合，而且也使得 Markov 性能被很好利用。

除了回报之外，在 MRP 中，我们更关心的是价值函数，即处于某个状态时候预期的回报是多少：

定义 2.4 (价值函数) 在 MRP 中，状态-价值函数（或价值函数）定义为

$$v(s) = \mathbb{E}[G_t | S_t = s].$$

□

注意，等式右边有 t 但左边没有，所以我们需要说明这个定义对任意 t 都是成立的。我们只需要说明对于任意的 t , $t+1$ 和 t 定义了同一个 $v(s)$ 。注意，随机变量 R_{t+k} 只依赖于 S_{t+k} ，即 $R_{t+k} = R(S_{t+k})$ ，所以

$$\begin{aligned}\mathbb{E}[G_t | S_t = s] &= \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s \right] \\ &= \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k R(S_{t+k+1}) \middle| S_t = s \right] \\ &= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}[R(S_{t+k+1}) | S_t = s].\end{aligned}$$

注意，项 $\mathbb{E}[R(S_{t+k+1}) | S_t = s]$ 满足时齐性，即

$$\mathbb{E}[R(S_{t+k+1}) | S_t = s] = \mathbb{E}[R(S_{t+k+2}) | S_{t+1} = s].$$

因而对于任意的 t , $t+1$ 和 t 定义了同一个 $v(s)$ 。

因此，这一定义蕴含了 Markov 性：只从当前起考虑未来收益，不考虑历史收益（沉没成本）的影响；也蕴含了时齐性：价值函数的定义不依赖于时刻 t 。我们在后面要各种相关概念的定义都需要用到这个性质，证明都是类似的，不再赘述。

接下来我们展示价值函数的计算方法。直观上说，回报应该被分解为两部分：即时回报 R_{t+1} 以及未来的回报 $\gamma v(S_{t+1})$ ，也就是下一个状态期望回报再做折扣。具体来说，我们有

$$\begin{aligned}v(s) &= \mathbb{E}(G_t | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots) | S_t = s)\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}(R_{t+1} + \gamma G_{t+1} | S_t = s) \\
&= \mathbb{E}(R_{t+1} | S_t = s) + \gamma \mathbb{E}(G_{t+1} | S_t = s) \\
&= \mathbb{E}(\textcolor{red}{R}_{t+1} | S_t = s) + \mathbb{E}[\gamma v(S_{t+1}) | S_t = s] \\
&= \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'} v(s').
\end{aligned}$$

其中倒数第二行是因为

$$\begin{aligned}
\mathbb{E}(G_{t+1} | S_t = s) &= \sum_{s' \in \mathcal{S}} \mathbb{E}(G_{t+1} | S_{t+1} = s', S_t = s) \Pr(S_{t+1} = s' | S_t = s) \\
&= \sum_{s' \in \mathcal{S}} \mathbb{E}(G_{t+1} | S_{t+1} = s') \Pr(S_{t+1} = s' | S_t = s) \\
&= \sum_{s' \in \mathcal{S}} v(s') \Pr(S_{t+1} = s' | S_t = s) \\
&= \mathbb{E}[v(S_{t+1}) | S_t = s].
\end{aligned}$$

我们因此得到了 **Bellman 方程** [Bel52]:

定理 2.3 (Bellman 方程) $v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'} v(s')$.

在后面，我们将不断看到这样形式的方程，它将一个和 Markov 链有关的计算分解为当前的部分和未来的部分。我们将这样形式的方程统称 Bellman 方程。

Bellman 方程可以用矩阵形式表达：

$$v = \mathcal{R} + \gamma \mathcal{P}v.$$

这里 v 是列向量 $v = (v(s))_{s \in \mathcal{S}}$ 。

写成矩阵形式之后，我们可以看到，Bellman 方程其实是一个线性

方程，可以被直接解¹：

$$v = \mathcal{R} + \gamma \mathcal{P}v \implies (I - \gamma \mathcal{P})v = \mathcal{R} \implies v = (I - \gamma \mathcal{P})^{-1} \mathcal{R}.$$

我们还可以有另一种观点，考虑一个映射 $f : \mathcal{S} \rightarrow \mathcal{S}$, $f(v) = \mathcal{R} + \gamma \mathcal{P}v$, 那么 Bellman 方程可以被写作

$$v = f(v).$$

因此， v 是 f 的不动点. 在第 8 章，我们将系统地讨论不动点的性质以及它对于 Markov 链相关模型的重要性.

回到解 Bellman 方程，对于 n 个状态的 Markov 链，用线性方程组法的计算复杂度为 $\mathcal{O}(n^3)$. 对于较小的 MRP 可以直接解，太大的 MRP 开销太大. 对于大型 MRP，可以采用迭代算法，例如：动态规划、Monte-Carlo 法、时序差分学习等 [Wat89]. 动态规划法的思路我们将在第 2.4 节中介绍.

§2.3 Markov 决策过程 (MDP)

上一节中我们建模了 Markov 链上的价值系统，即 MRP，接下来我们进入建模决策的环节. 回顾李二的例子（例 2.3），我们在 MRP 中忽略了一个非常重要的考虑：当李二坐在教室 1 中的时候，李二不是随机地开始刷手机的，他需要选择开始刷手机. 换言之，在 MRP 中，在当前状态可以做什么行动是缺失的.

此外，李二的奖励其实不是由状态决定的，而是由他做了什么行动

¹这里我们假设 $\gamma \neq 1$ ，否则这个方程是奇异的.

决定的，这样，李二才能评估做什么行动可以最大化自己的奖励，然后选择奖励高的那个行动。实际上，认知神经科学的研究表明，人体的运动控制就是由类似的机制完成的：首先，大脑皮层会提出若干不同的运动计划，这些运动计划对应了不同的奖励（由多巴胺浓度来表征），人脑中有一个被称作基底神经节的结构，它负责“放行”奖励高于某个阈值的运动计划，于是人就可以产生这个动作了 [Cis07, DeL90]。

综合以上这些思考，我们就可以给出 Markov 决策过程 (MDP) 的模型。我们还是先看李二的例子，然后再推广到一般的情况。

例 2.4 (李二 MDP) 李二的 MDP 见图 2.6。在图中，状态依然是之前状

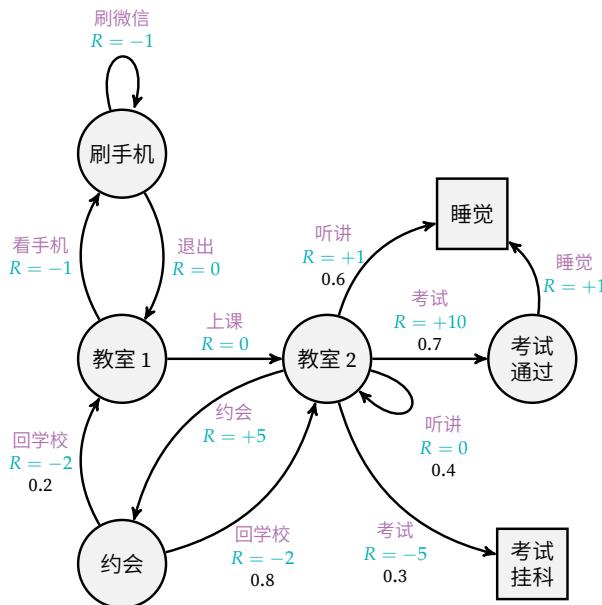


图 2.6: 学生 MDP

态，但是状态的转移被给上了两个标签：(1) 这个转移是什么动作引起

的 (紫色), (2) 这个动作可以带来多少的奖励 (蓝色). 例如, 当李二在教室 1 的时候, 她如果选择看手机, 那么就会有 -1 的奖励, 并且进入刷手机的状态. \square

接下来, 我们给出一般的 MDP 的定义.

定义 2.5 (Markov 决策过程, MDP) **Markov 决策过程 (MDP)** 是一个 MDP 是五元组 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, 其中

- \mathcal{S} 是一个有限的状态集合.
- \mathcal{A} 是一个有限的行动 (action) 集合.
- \mathcal{P} 是状态-行动转移概率矩阵,

$$\mathcal{P}_{ss'}^{\textcolor{violet}{a}} = \Pr(S_{t+1} = s' | S_t = s, A_t = \textcolor{violet}{a}).$$

- \mathcal{R} 是一个期望奖励函数, $\mathcal{R}_s^{\textcolor{violet}{a}} = \mathbb{E}(\textcolor{teal}{R}_{t+1} | S_t = s, A_t = \textcolor{violet}{a})$, 随机变量 $\textcolor{teal}{R}_{t+1}$ 是进行某一行动到达某一状态后的奖励.
- γ 是一个折扣因子 $\gamma \in [0, 1]$. \square

现在我们对比 MDP 和 MRP. MDP 中, 状态转移矩阵依赖动作, 奖励函数也依赖动作. 在李二的例子中, 在教室 2 如果选择听讲, 尽管都是在听讲, 但是并不一定会产生一个确定的结果: 他会有 0.4 的概率睡着, 0.6 的概率继续听讲. 自然, 睡着的奖励和继续听讲的奖励是不同的.

定义 2.6 (策略) 一个策略 π 是给定状态下行动的分布, 即对任意 $s \in \mathcal{S}$ 和 $a \in \mathcal{A}$, 有

$$\pi(a|s) = \Pr(A_t = a | S_t = s).$$

\square

一个策略完全决定了一个智能体在 MDP 环境中的行为。它的定义蕴含着 Markov 性：MDP 的策略取决于当前状态，而非历史状态；也蕴含着时齐性：MDP 的策略不依赖于时刻 t 。这样的定义会方便我们讨论价值函数以及决策的问题。

MDP 与 MRP 的关系由策略给出。

命题 2.5 给定一个 $MDP \mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ 和一个策略 $\pi, \langle \mathcal{S}, \mathcal{P}^\pi \rangle$ 是一个 Markov 链， $\langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$ 是一个 MRP，其中

$$\mathcal{P}_{s,s'}^\pi = \mathbb{E}_{a \sim \pi(\cdot|s)}(\mathcal{P}_{s,s'}^a) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}_{s,s'}^a,$$

$$\mathcal{R}_s^\pi = \mathbb{E}_{a \sim \pi(\cdot|s)}(\mathcal{R}_s^a) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}_s^a.$$

证明。对行动利用全概率公式。 □

现在我们有三个概念，MDP，MRP，以及 Markov 链，命题 2.6 给了他们三者的关系，我们可以总结到图 2.7.



图 2.7: MDP，MRP 和 Markov 链的关系

对于李二来说，选择什么样的策略很大程度取决于他能从中获得多少奖励。同样，在 MDP 中，我们可以定义回报的概念。

定义 2.7 (回报) 在 MDP 中， t 时刻以后的总回报 G_t 定义为

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$
□

类似 MRP，我们需要定义相应的价值函数。在 MDP 中，状态-价值函数和行动-价值函数是两个重要的价值函数，它们分别描述了从某一状态出发，遵从某一策略的期望回报。

定义 2.8 (价值函数) 状态-价值函数 $v_\pi(s)$ 是从状态 s 出发，遵从策略 π 的期望回报

$$v_\pi(s) = \mathbb{E}_\pi(G_t | S_t = s).$$

行动-价值函数 $q_\pi(s, a)$ 是从状态 s 出发，采取行动 a ，遵从策略 π 的期望回报

$$q_\pi(s, a) = \mathbb{E}_\pi(G_t | S_t = s, A_t = a).$$

□

注意，类似 MRP 中的价值函数，MDP 中的价值函数定义也不依赖 t 的选择，这是因为 MDP 以及其策略的 Markov 性和时齐性。

行动-价值函数比起状态-价值函数更加具体，它可以帮助李二评判在当前选择每一个行动的回报，从而选择最优的行动。而状态-价值函数则是对行动-价值函数的一个期望，因而是李二预期他从这个状态出发的回报。具体来说，这两个价值函数有如下关系：

命题 2.6 状态-价值函数 $v_\pi(s)$ 和行动-价值函数 $q_\pi(s, a)$ 之间有如下关系：

$$v_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}(q_\pi(s, a)) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a),$$

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s').$$

证明. 用 $q_\pi(s, a)$ 来表示 $v_\pi(s)$ ，对行动 a 用全概率公式即可得到。

另一方面，用 $v_\pi(s)$ 来表示 $q_\pi(s, a)$ 也是全概率公式。具体来说，在状态 s 采取行动 a 之后，期望上有 \mathcal{R}_s^a 的即时奖励，然后以 $P_{s,s'}^a$ 的概率

转移到状态 s' , 在状态 s' 的期望回报是 $\gamma v_\pi(s')$. 按照 s' 用全概率公式即可得到 $q_\pi(s, a)$. \square

下面我们给出状态-价值函数和行动-价值函数的 Bellman 方程. 首先, 价值函数可以被分解为即时回报加未来的折扣回报, 具体来说

- 状态-价值函数可以被分解为:

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s].$$

- 行动-价值函数可以被类似地分解,

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a].$$

他们的计算方法和定理 2.3 中的方法类似. 继续仿照定理 2.3 的证明, 我们可以得到 MDP 的 Bellman 方程 (见习题 5) .

定理 2.4 (Bellman 方程)

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a v_\pi(s') \right),$$

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s', a').$$

状态-价值函数的 Bellman 方程可以被写成矩阵形式:

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi.$$

注意, MDP 的 Bellman 方程只能告诉我们给定策略 π 的价值函数, 它并不能告诉智能体 (也就是李二) 要如何行动, 所以, 接下来我们要

讨论最优策略和最优价值函数.

定义 2.9 (最优价值函数) 最优状态-价值函数 $v_*(s)$ 是所有策略中最大的状态-价值函数

$$v_*(s) = \max_{\pi} v_{\pi}(s).$$

最优行动-价值函数 $q_*(s, a)$ 是所有策略中最大的行动-价值函数

$$q_*(s, a) = \max_{\pi} q_{\pi}(s, a).$$

□

相应地，最优价值函数确定了智能体在 MDP 中的最佳收益，解 MDP 即确定达到最优价值函数的策略.

在以上定义中可能会遇到这样的问题：对两个状态 s_1 和 s_2 ，存在两个不同的策略 π_1 和 π_2 ，使得 $v_*(s_1) = v_{\pi_1}(s_1)$, $v_*(s_2) = v_{\pi_2}(s_2)$. 此时，每个状态取到最大价值的策略 π 可能并不是同一个，因此 v_* 并不是某个特定策略可以实现的值. 所以，我们需要证明，存在一个策略 π_* , π_* 都取得最大价值函数.

我们有如下定理，说明了这样策略的存在性，因而也证明了 MDP 解的存在性.

定理 2.5 (MDP 解的存在性) 对任意 MDP，存在一个策略 π_* ,

- 对任意状态 s , $v_{\pi_*}(s) = v_*(s)$.
- 对任意状态 s 和行动 a , $q_{\pi_*}(s, a) = q_*(s, a)$.

证明. 我们给出一个构造性证明，即找出最优策略. 我们先找到一个 π_* 最大化 q ，然后说明这个 π_* 也可以最大化 v .

直观上说，我们只要对每个状态都选择最好的行动，这就是一个最优策略. 具体来说，我们可以通过如下步骤找到 π_* :

- 固定 s ,
- 找到一个 a_\star 最大化 $q_\star(s, \cdot)$, 即 $q_\star(s, a_\star) = \max_a q_\star(s, a)$, 令 $\pi_\star(a_\star|s) = 1$.
- 对其他 $a \neq a_\star$, 令 $\pi_\star(a|s) = 0$.

首先, 根据选法, π_\star 取得最优行动-价值函数. 接下来我们说明, 它也取得了最优状态-价值函数. 任意策略 π , 给定状态 s , 我们有如下计算:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)}[q_\pi(s, a)] \\ &\leq \mathbb{E}_{a \sim \pi(\cdot|s)}[q_\star(s, a)] \\ &\leq q_\star(s, a_\star) \\ &= v_{\pi_\star}(s). \end{aligned}$$
□

这个证明还有一个推论:

推论 2.2 对任意 MDP, 总存在一个非随机的最优策略, 即对任意状态 s , $\pi_\star(a|s) \in \{0, 1\}$.

两个最优价值函数之间有如下关系:

命题 2.7 以下等式成立:

$$v_\star(s) = \max_a q_\star(s, a),$$

$$q_\star(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a v_\star(s').$$

证明. 在定理 2.5 的证明中, 我们将 π 取为 π_* , 于是有

$$v_*(s) = v_{\pi_*}(s) = q_*(s, a_*) = \max_a q_*(s, a).$$

对于第二个等式, 根据定理 2.5,

$$v_*(s) = v_{\pi_*}(s), \quad q_*(s, a) = q_{\pi_*}(s, a),$$

在命题 2.6 取 $\pi = \pi_*$ 即可得到. \square

根据上面结论, 如果我们知道 $q_*(s, a)$ 或者 $v_*(s)$, 我们就能获得最优策略. 这一计算同样依赖 Bellman 方程²:

定理 2.6 (Bellman 方程)

$$v_*(s) = \max_a \left\{ \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a v_*(s') \right\},$$

$$q_*(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a \max_{a'} q_*(s', a').$$

证明. 对于第一个方程, 将命题 2.7 中第二个等式代入第一个等式即可得到.

对于第二个方程, 将命题 2.7 中第一个等式代入第二个等式即可得到. \square

Bellman 方程不是线性的. 因此很难有解析形式的解. 但是 MDP 的数值解是有 (多项式时间) 算法可以求出来的 [LDK95]. 我们一般采用迭代算法求解, 例如价值迭代、策略迭代等.

²在文献中, 这一 Bellman 方程被称为 Bellman 最优性方程, 而之前推导的 Bellman 方程被称为 Bellman 期望方程. 为了不过度引入术语, 我们这里不做这种区分.

求解 MDP 的过程实际上就是人工智能中强化学习的核心步骤。在本节的开头，我们说过人脑运动控制机制。实际上，绝大部分的哺乳动物都具备类似这样的学习机制：通过奖励的反馈，动物可以学会如何行动，从而获得最大的奖励。这样的学习机制被称为强化学习。

注. Bellman 方程是强化学习、经济学动态优化的核心。Bellman 方程的推导是 Markov 链中最为常用的技巧：考虑从当前状态转移到下一状态，利用全概率公式，一步转移会将两个状态之间的概率（期望）用递推公式联系起来。在随机过程中，有大量这样的例子：前向方程、Wald 等式、调和函数。后面的 HMM 也是类似的例子。

最后，我们谈谈深度强化学习。在一些非常复杂的情况下（例如下围棋），使用经典的迭代算法不容易求解 MDP。深度强化学习是一种结合了深度学习和强化学习的方法。在深度强化学习中，用神经网络来表示 π 和 v ，之后，用某种学习算法训练神经网络。

在很多深度强化学习模型中（例如 AlphaGo [SHM⁺16]），MDP 的策略是基于过去 k 期的状态做当前的决策，这样可以更好地利用状态的历史信息。这样的决策模型等价于一个利用一期信息决策的 MDP（见习题 3），因此依然可以用同样的深度学习算法来求解。

在一些深度强化学习模型中（例如 MuZero [SAH⁺20]），状态空间 S 也用一个神经网络表示。用神经网络来表示状态空间的好处是可以减少人类的特征工程，让神经网络充分发掘状态空间好的表示方法。

注. 1989 年，Christopher Watkins 在攻读他的心理学博士学位时，首次将基于动态规划的 MDP 作为强化学习的数学模型 [Wat89]。之后，强化学习在很多复杂交互的环境中都有广泛的应用，例如用于围棋的 Al-

phaGo [SHM⁺16]、星际争霸的 AlphaStar [VBC⁺19] 等。关于强化学习的详细讨论，可以参考 Richard Sutton 和 Andrew Barto 的经典教材 [SB18]。

§2.4 隐 *Markov* 模型 (HMM)

在本节，我们考虑 *Markov* 链上的另一种应用。在统计学和机器学习中，我们有时候要处理一类含时间的数据。例如，如果我们希望利用机器的力量帮助我们炒股赚钱，就要考虑如何预测股价然后做出相应的决策，这样的投资模式被称为量化投资。在 1989 年到 2009 年间，量化界的传奇人物 James Simons 操盘大奖章基金，平均年回报率高达 35%，即便是在次贷危机爆发的 2007 年，该基金的回报率仍高达 85%。据说，让 Simons 成功的秘诀是隐 *Markov* 模型，这正是本节的主题。

我们先给出隐 *Markov* 模型的定义。

定义 2.10 (隐 *Markov* 模型, HMM) 一个隐 *Markov* 模型 (HMM) 是两列随机变量（被称为观测序列） X_1, X_2, \dots 和 z_1, z_2, \dots ，（被称为隐状态序列）的序列，满足：

- $\{Z_t\}$ 构成一条 *Markov* 链。
- 对任意 t , X_t 的分布仅依赖于 Z_t 。
- 对于任意 t , $\Pr(X_t|Z_t)$ 服从分布 $F(Z_t)$ 。 □

示意见图 2.8。

为了理解这一概念，我们可以考虑炒股的例子。

例 2.5 (美为 HMM) 假设我们要投资美的股票，第 t 天的股票价格是 X_t 。我们很希望理解整个 X_t 的变化趋势。然而，美为股价的背后有一个

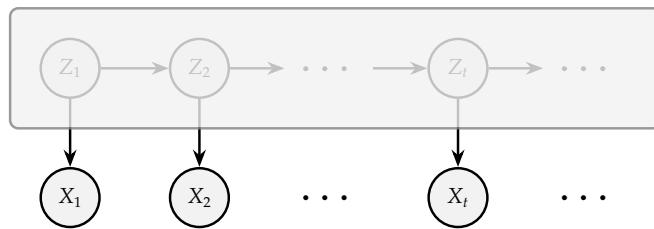


图 2.8: 隐 Markov 模型

神秘势力操控. 我们并不清楚这个神秘势力每天决策的具体细节, 只知道他们的决策非常健忘, 即他们的决策只依赖于前一天的决策. 他们会决定一个明天的预计股价 Z_t , 这构成了一个 Markov 链 $\{Z_t\}$.

因为市场和汇率的波动, 这个神秘势力无法完全决定股票的价格 X_t , 然而, 他们每一天所决定的预期价格 Z_t 会导致股票价格的变化, 我们可以认为 X_t 仅依赖于 Z_t , 但依然受到一些随机因素的影响. 作为简化, 我们假设每天的随机影响是独立且一致的, 服从分布 $F(Z_t)$. 这样的得到的模型就是一个 HMM. \square

接下来, 我们给 HMM 引入一些记号. 为简便起见, 我们只研究有限离散的 HMM. 设 HMM \mathcal{M} 所形成概率测度是 $\Pr_{\mathcal{M}}$. 我们假设模型中以下的量都是已知的:

- \mathcal{Z} : 有限的状态集合, 即 Z_t 的取值集合.
- \mathcal{X} : 有限的观测集合, 即 X_t 的取值集合.
- T : Markov 链 $\{Z_t\}$ 的转移矩阵, $T_{i,j} = \Pr_{\mathcal{M}}(Z_{t+1} = j | Z_t = i)$.
- M : 给定隐状态时的观测概率, $M_{i,k} = \Pr(X_t = k | Z_t = i)$.
- λ : 隐状态的初始分布.

我们会把序列 A_i, \dots, A_j 记作 $A_{i:j}$, 然后将它看作一个向量. 例如 $X = X_{1:t}$ 就是描述从时刻 1 到 t 的观测序列的随机向量, 而 $Z = Z_{1:t}$ 就是描述从时刻 1 到 t 的隐状态序列的随机向量.

为了理解 HMM 的任务, 我们接着看美的例子. 假设我们已经知道一个 HMM \mathcal{M} 可以预测美的股价 X_t , 作为一个使用者, 我们希望 \mathcal{M} 确实有用. 因此, 我们要评估这个模型的表现. 具体来说, 我们希望知道, 给定观测历史 $x = (x_1, x_2, \dots, x_t)$, 如何计算 $\Pr_{\mathcal{M}}(X = x)$. 这一方法也可以用来做预测: 利用全概率公式, 我们可以计算 $\Pr_{\mathcal{M}}(X_{t+1} = x_{t+1} | X = x)$, 从而预测未来的股价.

评估一个 HMM 固然重要, 但是评估依然是把 HMM 当成一个黑盒来使用. 我们还希望知道这个模型背后到底发生了什么, 这就是解释问题. 具体来说, 我们希望知道, 给定观测历史 $x = (x_1, x_2, \dots, x_t)$ 以及一个时刻 k , 如何计算 $\Pr_{\mathcal{M}}(Z_k | X = x)$. 这一分布表明了在 k 时刻神秘势力更有可能做了什么样的决策, 从而帮助我们更好理解股价的波动.

接下来我们将分别阐述如何解决评估问题和解释问题. 因为我们只讨论某一个具体的 HMM, 为简便起见, 我们此后都将概率测度 $\Pr_{\mathcal{M}}$ 简记为 \Pr .

§2.4.1 评估问题

我们引入记号随机向量 $X = (X_1, \dots, X_t)$, $Z = (Z_1, \dots, Z_t)$. 我们考虑 HMM 的评估问题: 给定一个 HMM \mathcal{M} , 以及它的观测历史 $x = (x_1, x_2, \dots, x_t)$, 计算 $\Pr(X = x)$.

关键困难是我们不知道隐状态历史 $Z = (z_1, z_2, \dots, z_t)$, 因此我们

需要利用全概率公式将隐状态消除掉，即：

$$\Pr(X = x) = \sum_{Z=(z_1, \dots, z_t) \in \mathcal{Z}^t} \Pr(X = x | Z = z) \Pr(Z = z).$$

接下来我们分别计算 $\Pr(X = x | Z = z)$ 和 $\Pr(Z = z)$. 对于前者，因为每一个观测值 X_i 仅依赖于 Z_i ，我们有

$$\Pr(X = x | Z = z) = \prod_{i=1}^t \Pr(X_i = x_i | Z_i = z_i) = M_{z_1, x_1} \cdot M_{z_2, x_2} \dots M_{z_t, x_t},$$

对于后者，因为 Z 是一个 *Markov* 链，我们有

$$\begin{aligned} \Pr(Z = z) &= \Pr(Z_1 = z_1) \prod_{i=2}^t \Pr(Z_i = z_i | Z_{i-1} = z_{i-1}) \\ &= \lambda_{z_1} \cdot T_{z_1, z_2} \cdot T_{z_2, z_3} \dots T_{z_{t-1}, z_t}. \end{aligned}$$

以上的量都是已知的，所以我们已经可以计算评估问题了. 然而，这一方法的需要计算的乘法次数是 $\mathcal{O}(t|\mathcal{Z}|^t)$ ，对于很大的 t 和 \mathcal{Z} ，这是不可接受的计算量. 我们需要更好的计算方法.

接下来，我们采用前向方程（见定理 2.1）的思路，从前 k 步的结果推出前 $k+1$ 步的结果，然后据此列出递推方程. 在第 $k+1$ 步，*Markov* 链的状态发生了转移，按照从哪个状态转移到了哪个状态，我们可以拆分概率：

$$\begin{aligned} &\Pr(X_{1:k+1} = x_{1:k+1}) \\ &= \sum_{z \in \mathcal{Z}} \Pr(X_{1:k} = x_{1:k}, Z_k = z) \Pr(X_{k+1} = x_{k+1} | Z_k = z) \end{aligned}$$

$$\begin{aligned}
&= \sum_{z \in \mathcal{Z}} \Pr(X_{1:k} = x_{1:k}, Z_k = z) \cdot \\
&\quad \sum_{z' \in \mathcal{Z}} \Pr(Z_{k+1} = z' | Z_k = z) \Pr(X_{k+1} = x_{k+1} | Z_{k+1} = z') \\
&= \sum_{z \in \mathcal{Z}} \Pr(X_{1:k} = x_{1:k}, Z_k = z) \sum_{z' \in \mathcal{Z}} T_{z,z'} M_{z',x_{k+1}}.
\end{aligned}$$

如果把左边按照 Z_{k+1} 拆分，我们有

$$\begin{aligned}
&\sum_{z \in \mathcal{Z}} \Pr(X_{1:k+1} = x_{1:k+1}, Z_{k+1} = z) \\
&= \sum_{z \in \mathcal{Z}} \Pr(X_{1:k} = x_{1:k}, Z_k = z) \sum_{z' \in \mathcal{Z}} T_{z,z'} M_{z',x_{k+1}}.
\end{aligned}$$

如果令 $\alpha_k(z) := \Pr(X_{1:k} = x_{1:k}, Z_k = z)$ ，用类似的计算，我们有递推：

- 当 $k = 1$, $\alpha_k(z) = \lambda(z) M_{z,x_k}$.
- 当 $k > 1$, $\alpha_{k+1}(z) = \sum_{z' \in \mathcal{Z}} \alpha_k(z') T_{z',z} M_{z,x_{k+1}}$.

最后，

$$\Pr(X = x) = \sum_{z \in \mathcal{Z}} \alpha_t(z).$$

这一方法需要算的乘法次数是 $\mathcal{O}(t|\mathcal{Z}|^2)$ ，比前一种计算方法要快很多。

镜像地，我们可以使用后向方程的思路，从前 $k + 1$ 步的结果推出前 k 步的结果。同样可以列出递推方程。定义 $\beta_k(z) := \Pr(X_{k+1:t} = x_{k+1:t} | Z_k = z)$ ，我们有递推：

- 当 $k = t$, $\beta_k(z) = 1$.
- 当 $1 \leq k < t$, $\beta_k(z) = \sum_{z' \in \mathcal{Z}} T_{z,z'} M_{z',x_{k+1}} \beta_{k+1}(z')$.

于是, $\Pr(X = x) = \sum_{z \in \mathcal{Z}} \lambda(z) M_{z,x_1} \beta_1(z)$. 这一方法需要算的乘法次数是 $\mathcal{O}(t|\mathcal{Z}|^2)$.

§2.4.2 解释问题

接下来我们讨论 HMM 的解释问题.

给定一个 HMM $\mathcal{M} = (\mathcal{Z}, \mathcal{X}, T, M, \lambda)$ 以及一列观测历史 $x = (x_1, x_2, \dots, x_t)$, 解释问题旨在寻找一个状态序列, 能最好地解释这些历史观察. 具体来说我们考虑如下四个问题:

1. 过滤: 计算 $\Pr(Z_k = s | X_{1:k} = x_{1:k})$.
2. 平滑: 计算 $\Pr(Z_k = s | X = x)$, $k < t$.
3. 预测: 计算 $\Pr(Z_k = s | X = x)$, $k > t$.
4. 解码: 找到最有可能的状态序列 $z = (z_1, z_2, \dots, z_t)$.

首先考虑过滤: $\Pr(Z_k = s | X_{1:k} = x_{1:k})$. 回忆, $\alpha_k(s) = \Pr(X_{1:k} = x_{1:k}, Z_k = s)$, 这其实已经足够我们计算过滤了. 根据条件概率的定义, 我们有

$$\begin{aligned}\Pr(Z_k = s | X_{1:k} = x_{1:k}) &= \frac{\Pr(X_{1:k} = x_{1:k}, Z_k = s)}{\Pr(X_{1:k} = x_{1:k})} \\ &= \frac{\alpha_k(s)}{\sum_{z \in \mathcal{Z}} \alpha_k(z)}.\end{aligned}$$

我们已经知道如何计算 $\alpha_k(s)$, 所以这就可以用来计算过滤了.

然后是平滑: $\Pr(Z_k = s | X = x)$, $k < t$. 回顾记号 $\alpha_k(s) = \Pr(X_{1:k} = x_{1:k}, Z_k = s)$, 以及 $\beta_k(s) = \Pr(X_{k+1:t} = x_{k+1:t} | Z_k = s)$. 可以证明 (见

习题 7) :

$$\Pr(z_k = s | X = x) = \frac{\beta_k(s) \alpha_k(s)}{\sum_{z \in \mathcal{Z}} \alpha_t(z)}. \quad (2.5)$$

同样，我们已经知道如何计算 $\alpha_k(s)$ 和 $\beta_k(s)$ ，所以这就可以用来计算平滑了。

之后是预测： $\Pr(Z_k = s | X = x)$, $k > t$. 首先用过滤计算 $\lambda = \Pr(Z_t = s | X = x)$. 从 t 之后的隐状态都只依赖于 t 时刻的隐状态 Z_t ，因此，条件在 $X = x$ 下， Z_t, Z_{t+1}, \dots, Z_k 构成了一个 *Markov* 链，它的初始分布为 λ ，转移矩阵为 T . 于是我们利用定理 2.1 来计算该 *Markov* 链第 $k - t$ 步的分布.

最后是解码：求 $z = (z_1, z_2, \dots, z_t)$ ，使得 $\Pr(Z = z | X = x)$ 最大. 注意，这一概率最大等价于 $\Pr(Z = z, X = x)$ 最大. 我们也使用递归的想法来解决这个问题. 同样，在前 $k - 1$ 个状态已经选好之后，考虑最后一个状态应该选哪个. 具体来说，定义

$$\delta_k(s) = \max_{z_{1:k-1}} \Pr(Z_{1:k} = (z_{1:k-1}, s), X_{1:k} = x_{1:k}).$$

于是

$$\begin{aligned} & \delta_{k+1}(s) \\ &= \max_{z_{1:k}} \Pr(Z_{1:k+1} = (z_{1:k}, s), X_{1:k+1} = x_{1:k+1}) \\ &= \max_{z_{1:k}} \{ \Pr(Z_{1:k} = z_{1:k}, X_{1:k} = x_{1:k}) \cdot \\ & \quad \Pr(Z_{k+1} = s | Z_k = z_k) \Pr(X_{k+1} = x_{k+1} | Z_{k+1} = s) \} \\ &= \max_q \{ \max_{z_{1:k-1}} \Pr(Z_{1:k} = (z_{1:k-1}, q), X_{1:k} = x_{1:k}) T_{q,s} \} M_{s,x_{k+1}} \end{aligned}$$

$$= \max_q \{ \delta_k(q) T_{q,s} \} M_{s,x_{k+1}}.$$

这就给出了从 k 推导到 $k+1$ 的递推方程. 这一递推的初始状态是

$$\begin{aligned} \delta_1(s) &= \Pr(Z_1 = s, X_1 = x_1) \\ &= \Pr(Z_1 = s) \Pr(X_1 = x_1 | Z_1 = s) = \lambda(s) M_{s,x_1}. \end{aligned}$$

利用这一递推，我们就可以解决解码问题了. 具体算法在下面给出.

1. 利用递推公式，逐层用第 k 层的 δ_k 计算第 $k+1$ 层的 δ_{k+1} ，最后得到 δ_t .
2. 求一个 z_t^* 使得 $\delta_t(z_t^*)$ 最大，根据定义，这个 z_t^* 也使得 $\Pr(Z = z, X = x)$ 最大，把这个最大值记为 δ_t^* .
3. 接下来，逐层用第 $k+1$ 层的 δ_{k+1}^* 计算第 k 层的 δ_k^* 和 z_k^* . 已知

$$\delta_{k+1}^* = \delta_{k+1}(z_{k+1}^*) = \max_q \{ \delta_k(q) T_{q,z_{k+1}^*} \} M_{z_{k+1}^*, x_{k+1}},$$

从中找到一个 q 使得 $\delta_k(q) T_{q,z_{k+1}^*}$ 最大，记此时的 $\delta_k(q)$ 为 δ_k^* , $z_k^* = q$.

4. 最后，我们就得到了最优的状态序列 $z = (z_1^*, z_2^*, \dots, z_t^*)$ ，使得 $\Pr(Z = z, X = x)$ 最大.

以上算法被称为 *Viterbi* 算法 [Vit67]，是解码问题的一个高效算法，它需要计算的乘法次数是 $\mathcal{O}(t|\mathcal{Z}|^2)$. 这一算法采用了动态规划的思想，实际上大部分和 Bellman 方程有关的问题（特别是最优化的问题）都可以用这一方法解决.

注. HMM 在诸多领域都有应用。最早是作为一个统计学模型出现 [BP66]，后来用来做语音识别 [JBM75, HAJ90]，也被用来做 DNA 序列分析 [BT86]。

§2.5 扩散模型

本节我们讨论基于 Markov 链的另一种模型，即扩散模型。不同于之前的模型，扩散模型的启发来自物理学。滴一滴墨水到水中，最后墨会均匀地在水中分布，这一过程不会反过来，即墨水不会自动聚集到一起。同样的现象在热力学中也有体现：如果把金属勺子的一端放在热水中，热量会从热水中通过勺子传导到勺子的另一端，最终整个勺子会均匀地变热，而不会反过来。这些现象都是扩散过程。本节的主要任务就是为这一过程建立数学模型，并讨论它的应用。

我们先看墨水的例子。

例 2.6 (墨水的扩散) 墨水的扩散有一个具体的数学模型，即 *Ehrenfest* 模型。

模型如图 2.9 所示。在这个模型中，一共有两个箱子 A 和 B，我们可以想象成这是水杯的上半部分和下半部分。两个箱子里一共 N 个球，我们可以想象成这是墨水分子。每个时刻，有很小的概率 $\alpha > 0$ ，盒子 A 和 B 中的球会保持不动；另外的 $1 - \alpha$ 的概率，我们均匀随机选择一个球，从一个箱子跳到另一个箱子，这就是扩散的过程。

设 X_k 是盒子 A 中球的数目，那么 $\{X_k\}$ 是一个 Markov 链，它的

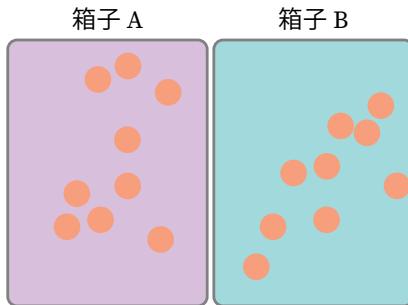


图 2.9: Ehrenfest 模型示意图

转移核³是

$$\Pr(X_{k+1} = i | X_k = j) = \begin{cases} \alpha, & i = j, \\ (1 - \alpha)j/N, & i = j - 1, \\ (1 - \alpha)(N - j)/N, & i = j + 1, \\ 0, & \text{其他.} \end{cases}$$

下面我们来解释为什么墨水最终会均匀分布在水中. 从遍历定理(定理 2.2)我们知道, 对于一个遍历的 Markov 链, 无论初始分布是什么, Markov 链最终都会收敛到唯一平稳分布. 可以证明(见习题 8), $\{X_k\}$ 是一个遍历的 Markov 链, 所以它会收敛到一个平稳分布 $\pi \sim B(N, 1/2)$.

根据二项式的性质, 该分布在 $N/2$ 附近取到最大值, 因此, 这是墨分子最有可能的分布情况. 更精细的结论是(见习题 8), 对任意 $\epsilon > 0$,

³我们在前面都说的是转移矩阵, 然而, 在扩散模型这一部分, 转移核是一个更恰当的表述. 这是因为, 我们之前处理的大部分 Markov 链都是离散的状态空间, 而这一部分处理的状态空间既有离散的也有连续的, 所以“矩阵”并不是一个恰当的词汇.

当 N 充分大的时候,

$$\Pr \left(\limsup_{k \rightarrow \infty} \left| \frac{X_k}{N} - \frac{1}{2} \right| < \epsilon \right) \approx 1.$$

换言之, 几乎以概率 1, 分子会均匀分布在两个箱子中, 即墨水会均匀地分布在水中. \square

以上例子告诉我们, 物理过程几乎是不可逆的. 然而, 如果我们仔细记录每一步扩散的过程, 我们可以尝试倒推 X_0 的分布. 这就是扩散模型的基本思想. 下面给出扩散模型的定义.

定义 2.11 (扩散模型) 扩散模型由随机变量 x_0, \dots, x_T 给出, 它包含两个过程:

- 扩散过程: 从 x_0 到 x_T 的 Markov 链, 它的核是 $q(x_{k+1}|x_k)$, 也叫正向过程. 通常, 扩散模型是连续型随机变量, 所以, 核是一个转移密度函数.
- 逆向过程: 从 x_T 到 x_0 的 Markov 链, 它的核是 $p(x_{k-1}|x_k)$. 通常, 扩散模型是连续型随机变量, 所以, 核是一个转移密度函数. \square

注意, 扩散过程和逆向过程都是非时齐的 Markov 链. 假设扩散过程是一个 Markov 链, 那么逆向过程必须是一个 Markov 链, 所以上述定义是良定义的 (见习题 9).

除了物理过程, 还有很多其他的场景也可以被看成扩散模型. 比如, 我们可以将一张图片看成刚滴了墨的水: 它有结构, 不是混乱无序的.⁴

⁴有趣的是, 古代中国和西方真的有在水面上撒颜料来作画的技法, 也就是在水上撒颜料, 然后用纸张把颜料印下来. 在中国, 这种技法叫湿拓画. 在西方, 这种技法叫 *marbling* 或者 *ebru* (来自土耳其语).

而水则可以看作一张由 Gauss 噪声生成的图片：它是随机的，缺乏结构。于是，墨分子扩散就可以被理解为一张有结构的图片逐渐变成噪声的过程。

在这种理解下，扩散过程就是图片加噪声的过程，逆向过程就是去噪的过程。尽管通过加噪，所有的图片都会变成噪声，然而，噪声与噪声之间细微的区别仍然是可以被区分的。于是，我们可以通过逆向过程，从噪声中恢复出原始的图片。

如果我们通过大量图片“图片加噪声”的过程得到一个扩散过程，那么我们可以反向学习“噪声去除”的过程，得到一个逆向过程，那么我们就可以用它来凭空生成图片。首先，我们随机生成一个噪声图片 x_T ，然后通过逆向过程，我们就可以回到有结构的原始图片 x_0 。因此，扩散模型可以被用作生成模型。

我们接下来都会以去噪扩散概率模型（DDPM）[HJA20]为例来讨论扩散模型，它是一个用来生成图片的神经网络模型。DDPM 的过程如图 2.10 所示。

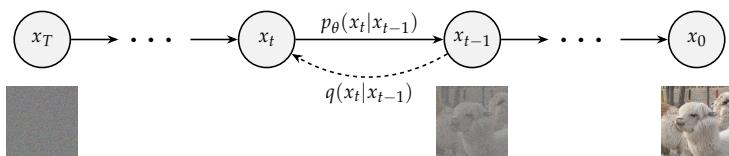


图 2.10: DDPM 示意图

DDPM 中的 Markov 核都是 Gauss 分布，具体如下：

- 扩散过程：假设真实图片的分布是 $q(x_0)$ ，这就应该 x_0 的初始分布。扩散过程从 x_{t-1} 到 x_t 就是在 x_{t-1} 上加上一个小的 Gauss 噪

声扰动，因此我们可以写出它的转移核和初始分布：

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad x_0 \sim q(x_0).$$

这里 β_t 是一个超参数， $\mathcal{N}(x; \mu, \Sigma)$ 是 Gauss 分布 $\mathcal{N}(\mu, \Sigma)$ 在点 x 的概率密度.

- 逆向过程：扩散的末端 x_T 是一个完全的 Gauss 噪声，我们不妨设它是标准 Gauss 分布，即 $x_T \sim \mathcal{N}(0, \mathbf{I})$. 逆向过程希望从 x_t 恢复到 x_{t-1} ，因此我们应该用 x_t 和 t 来预测 x_{t-1} 的期望，我们用一个神经网络 $\mu_\theta(x_t, t)$ 来表示这个期望. 于是，逆向过程的转移核和初始分布为：

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}), \quad p(x_T) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}).$$

这里 σ_t 是一个超参数.

根据我们的上面的讨论，扩散模型有两个任务：

- 训练一个逆向过程，使得 $p_\theta(x_0)$ 尽可能接近数据的分布 $q(x_0)$.
- 采样一个逆向过程，对一个训练好的扩散模型，这个步骤可以从噪声 x_T 出发生成一张有结构的图片 x_0 .

对于第一步，我们需要将训练的损失函数写出. 对于第二步，我们需要一个采样的方法. 由于采样比训练更简单，而且训练依赖采样，所以我们会先讨论采样.

§2.5.1 采样逆向过程

因为逆向过程是一个 Markov 链，我们需要能够采样一个 Markov 链的状态序列。因为随机变量之间有相互依赖关系，Markov 链的采样不是那么平凡的。但是，Markov 性给了我们一种采样的方法。

考虑 Markov 链 X_k ，转移核为 $p(i, j)$ ，初始分布为 λ 。我们可以通过以下方法采样 X_k ：

1. 从初始分布 λ 采样 X_0 。
2. 对于 $k = 1, 2, \dots$ ，从 $p(X_{k-1}, \cdot)$ 采样 X_k 。

根据 Markov 性，在给定 X_{k-1} 的情况下， X_k 的分布只依赖于 X_{k-1} ，所以这一采样方法可以正确地采样 X_k 。根据这一原则，我们可以很具体地将逆向过程的采样算法写出。

1. $x_T \sim \mathcal{N}(0, \mathbf{I})$
2. 对 $t = T, \dots, 1$ ，重复以下步骤采样 x_{t-1} ：
 - (1) 如果 $t \geq 1$ ，采样 $z \sim \mathcal{N}(0, \mathbf{I})$ ；否则， $z = 0$ 。
 - (2) $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$ 。
3. 输出 x_0 。

为了和后面训练部分对应，我们需要将 2.(2) 中的步骤重写为另一个等价的形式，现在可以先不管为什么要这么做。令 $\alpha_t = 1 - \beta_t$ ， $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ ， $\epsilon_\theta(x_t, t)$ 是 $\mu_\theta(x_t, t)$ 的重参数化：

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right).$$

于是，我们可以将 2.(2) 重写为：

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z.$$

§2.5.2 训练逆向过程

接下来我们讨论训练逆向过程的损失函数。我们需要衡量 $p(x_0)$ 和 $q(x_0)$ 的差异，一个常用的选择是交叉熵，我们将会在第三章中详细讨论这一概念。此时，我们先接受交叉熵的概念，将损失函数写为：

$$\mathbb{E}_q[-\log p_\theta(x_0)] = - \int q(x_0) \log p_\theta(x_0) dx_0.$$

为计算这一表达式，我们需要将 $p_\theta(x_0)$ 的表达式算出。类似 HMM，我们用 $x_{i:j}$ 表示 x_i, \dots, x_j 。

根据全概率公式：

$$\begin{aligned} p_\theta(x_0) &= \int p_\theta(x_{0:T}) dx_{1:T} \\ &= \int p_\theta(x_{0:T}) \frac{q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T} \\ &= \int q(x_{1:T}|x_0) \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} dx_{1:T} \\ &= \int q(x_{1:T}|x_0) p_\theta(x_T) \prod_{t=1}^T \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} dx_{1:T}. \end{aligned}$$

接下来，我们计算并放缩损失函数，最终得到容易利用梯度进行优

化⁵的形式：

$$\begin{aligned} K &= - \int q(x_0) \log p(x_0) dx_0 \\ &= - \int q(x_0) \cdot \log \left[\int q(x_{1:T}|x_0) \cdot p(x_T) \prod_{t=1}^T \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} dx_{1:T} \right] dx_0. \end{aligned}$$

利用 Jensen 不等式（定理 C.17）， $\log(\int q f dx) \geq \int q \log(f) dx$ ，这可以被放缩为：

$$K \leq - \int q(x_{0:T}) \log \left[p(x_T) \prod_{t=1}^T \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] dx_{0:T} := L.$$

可以证明（见第三章的习题 9）， L 可以被写作：

$$\begin{aligned} \mathbb{E}_q [\underbrace{D_{\text{KL}}(q(x_T|x_0) \| p(x_T))}_{L_T} + \\ \sum_{t>1} \underbrace{D_{\text{KL}}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{- \log p_\theta(x_0|x_1)}_{L_0}], \end{aligned}$$

其中 $D_{\text{KL}}(f\|g) = \int f \log(f/g) dx$ 是 K-L 散度.

接下来我们分别讨论每一部分的计算.

- L_T 不含参数，所以可以丢掉.
- L_0 和输出的数据格式有关（例如图片如何编码），需要具体问题具体处理，所以这里不讨论.
- 唯一需要处理的是 L_1 到 L_{T-1} 的计算.

⁵对于优化相关的详细讨论，见第三部分.

由于 $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I})$, 可以算得 (见习题 10):

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] + C, \quad (2.6)$$

其中 C 是一个不含 θ 的常数, 并且

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t.$$

因为 C 是一个常数, 我们可以丢弃它而不影响最优值. 接着, 利用重参数化

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right),$$

引入随机变量 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, 我们可以将 L_{t-1} 写作:

$$\mathbb{E}_{x_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon, t \right) \right\|^2 \right].$$

如果我们想用基于梯度方法对它进行优化, 我们需要计算它的梯度. 然而, 这是一个期望, 没有办法直接求梯度, 所以我们用采样到的 x_0 和 ϵ 代替期望中的 x_0 和 ϵ , 把它当成期望来求梯度.

另外, 如果把所有 L_{t-1} 都用来训练, 当总时长 T 非常大的时候, 训练也会非常困难, 所以在实际训练中, 我们会均匀随机选择一个 t , 对 L_{t-1} 进行优化.

最终, 我们得到了扩散模型的训练算法:

- 重复以下步骤直到收敛:

- 采样 $x_0 \sim q(x_0)$.
- 采样 $t \sim \mathcal{U}(\{1, \dots, T\})$.

- 采样 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
- 用下面的梯度做一次优化：

$$\nabla_{\theta} \left\| \epsilon - \epsilon_{\theta} \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2.$$

注. 最早将扩散模型应用于生成模型的工作是 [SWMG15]。后来, [HJA20] 将损失函数重新写为类似噪声得分匹配模型的形式 [SE19], 并提出了 DDPM. 这一模型在图像生成模型领域取得了很好的效果, 一举超越了 GAN [GPM⁺14], 成为了图像生成领域的宠儿。

另外, 关于 Markov 链收敛到平稳分布的一些性质, 比如混合时间, 可以参考教科书 [LPW08].

§2.6 习题

1. 设 $\{X_t\}_{t=0}^{\infty}$ 是一个时齐 Markov 链, 其转移矩阵为 $\mathcal{P} = (p(i, j))$, 初始分布为 λ .

(1) 证明:

$$\Pr(X_0 = i_0, X_1 = i_1, \dots, X_t = i_t) = \lambda_{i_0} \cdot p(i_0, i_1) \dots p(i_{t-1}, i_t).$$

- (2) 反之, 证明: 如果另外的随机变量 X'_0, \dots, X'_t 的联合概率分布满足上式, 则 $\{X'_t\}$ 是一个时齐 Markov 链, 转移矩阵为 \mathcal{P} , 初始分布为 λ .
- (3) 记 $\Pr(\cdot | X_n = s)$ 为 $Q(\cdot)$. 设 $Y_t = X_{t+n}$. 令 $X = (X_0, \dots, X_{n-1})$, $Y = (Y_0, \dots, Y_m)$, $i = (i_0, \dots, i_{n-1})$, $j = (j_0, \dots, j_m)$.

写出 X 和 Y 的联合概率分布 $Q(X = i, Y = j)$, 由此证明条件在 $X_n = s$, Y 与 X 相互独立, 且 $\{Y_t\}$ 是 Markov 链, 转移矩阵为 \mathcal{P} , 初始状态为 s .

2. 设 Markov 链的状态空间为 $\mathcal{S} = \{1, \dots, N\}$, 转移矩阵为 $\mathcal{P} = (p_{ij})$. 如果存在某个 n_0 , 使得

$$\min_{ij} p_{ij}^{(n_0)} > 0,$$

证明: 对任意 $i, j \in \mathcal{S}$, 任意 $n \geq n_0$, 都有 $p_{ij}^{(n)} > 0$.

3. 设 $\{X_n\}$ 是一个 Markov 链, 状态空间为 \mathcal{S} , 设

$$R_n = (X_n, X_{n+1}, \dots, X_{n+k-1}),$$

证明 $\{R_n\}$ 是一个 \mathcal{S}^k 上的 Markov 链.

4. 赌徒模型的输光时间. 假设甲乙两人参加公平对赌 (每局独立以 $1/2$ 概率赢或输), 甲最初有 a 元, 乙最初有 b 元 (a, b 都是正整数), 每局双方各下注 1 元, 赢家可以获得本局所有的下注. 当其中一方赌资为 0 时赌博结束. 设随机变量 τ 表示赌博结束时候经过的局数.

- (1) * 证明: $\mathbb{E}[\tau] < \infty$.

提示: 考虑每 $N = a + b$ 局为一个阶段. 观察以下两种极端路径:

- 甲连续赢 b 局 (概率为 $(1/2)^b$), 此时乙立即输光.
- 乙连续赢 a 局 (概率为 $(1/2)^a$), 此时甲立即输光.

因此在任意 N 局阶段中, 赌博结束的概率至少为 $p = 2^{-a} + 2^{-b}$.

利用几何分布的性质分析 $\Pr(\tau > kN) \leq (1 - p)^k$, 并由此推导 $\mathbb{E}[\tau]$ 的收敛性.

(2) 求 $E[\tau]$.

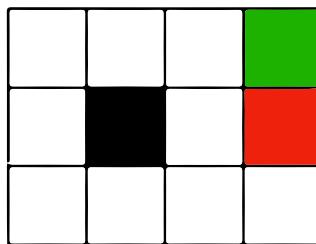
提示：按照一步转移分类讨论，写递推公式。注意，必须要先证明 (1) 才能写 (2) 的递推公式。如果直接从 (2) 的计算结果得出 $E[\tau] < \infty$ ，逻辑是完全不对的。考虑以下的例子：设 $S = 1 + 2 + 4 + \dots + 2^n + \dots$ ，那么 $S = 1 + 2S$ ，解得 $S = -1$ ，但是这个结果是荒谬的，因为 S 根本不收敛。所以，如果只是形式上算出来 $E[\tau]$ 是一个有限值，并不能说明 $E[\tau] < \infty$ 。

5. 证明定理 2.4.
6. 假设一个机器人身处于下图所示的方格环境中，我们为其定义一个 MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$:

- \mathcal{S} : 每个状态是一个方格。
- \mathcal{A} : 机器人可以采取的行动包含向上、向下、向左、向右。
- \mathcal{P} : 当机器人选择行动 a 时，其按照 a 所预设的行动转移的概率为
 1. 例如：机器人选择 $a = \text{向上}$ ，则机器人从当前方格转移到向上方向的方格概率为 1，转移到其他位置的概率为 0。
- \mathcal{R} : 假设机器人进入绿色格点的奖励为 $+1000000$ ，进入其他格点的奖励为 -1 。
- γ : 假设折扣因子 $\gamma = 0.99$.

我们额外假设机器人不会超出边界，若其在边界上选择了行动 a 使得下一个状态超出边界，则下一状态位于原地。假设机器人进入带有颜色的格点后不再运动，状态转移终止。

求解该 MDP 的一个最优决策 π_* .



7. HMM 的平滑问题. 证明 (2.5).
8. 设 $\{X_k\}$ 是一个 Markov 链, 它的转移核是

$$\Pr(X_{k+1} = i | X_k = j) = \begin{cases} \alpha, & 0 \leq i = j \leq N, \\ (1 - \alpha)j/N, & 0 \leq i = j - 1 \leq N - 1, \\ (1 - \alpha)(N - j)/N, & 1 \leq i = j + 1 \leq N, \\ 0, & \text{其他.} \end{cases}$$

其中 $0 < \alpha < 1$, N 是一个大于 2 的正整数.

- (1) 证明: $\{X_k\}$ 是一个遍历的 Markov 链.
- (2) 证明: $\{X_k\}$ 的平稳分布是 $\pi \sim B(N, 1/2)$.
- (3) * 对任意 $\epsilon > 0$, 证明:

$$\lim_{N \rightarrow +\infty} \Pr \left(\limsup_{k \rightarrow \infty} \left| \frac{X_k}{N} - \frac{1}{2} \right| < \epsilon \right) = 1.$$

9. 证明: 假如扩散过程是一个 Markov 链, 那么逆向过程也必为一个 Markov 链.

10. 证明 (2.6).

第二部分

信息与数据

第三章 熵与 Kullback-Leibler 散度

人脑和机器的区别是什么？我们有可能模拟人脑的功能吗？这些问题根植于认知科学和人工智能领域。在 20 世纪 50 年代，计算机科学、认知科学和人工智能仍然处于萌芽状态。就是在这个时候，基于对人类如何解决问题和决策的研究，Herbert A. Simon 提出，人脑其实是一个“信息处理器”，也就是输入信息（视觉、听觉等），进行处理，然后输出信息（动作、语言等） [Sim81]。从这个观点上，人脑和机器并无区别。

基于这样的观点，Simon 和 Allen Newell、J. C. Shaw 一起合作，制造了逻辑理论家（Logic Theorist）、通用问题求解器（General Problem Solver，GPS）等计算机程序。逻辑理论家可以证明《数学原理》（作者是 Whitehead 和 Russell）第二章前 52 个定理中的 38 个，而 GPS 则可以解决汉诺塔问题。这展现出“信息处理器”观点的巨大潜力。

时至今日，“信息处理器”的观点已经深入认知科学和人工智能研究者的心中。然而，信息是一个特别抽象的概念。它不像重量，可以从沉甸甸的铅块中直观感受到。那么，信息到底是什么？本章将要讨论这

一问题，并给出它在人工智能领域的应用。

§3.1 熵

§3.1.1 概念的导出

我们常说“恐惧来源于未知”，信息似乎代表着某种确定的东西，某种知识，因而和不确定性有相反的关系。更精确地说，消除不确定性的
东西被称为信息。当然，这句话本身似乎是一种循环解释，它既没有回答信息是什么也没有回答不确定性是什么。所以我们进一步的问题是，
给定一个“对象”，如何定量衡量它不确定性（或信息量）？

我们先从一个例子看起。

例 3.1 (信息论读本) 假设我们有一个信息论的读本（例如本章就是），
我们想要衡量它的信息量。我们面临的第一个困难是，同样的内容对于
不同的人来说，信息量是完全不同的。已经学过信息论的读者再看这一
部分内容，他获得的信息会比没有学过的读者要少得多。因此，我们很
难直接给单个对象衡量它的信息量。

但是，信息论读本的读者背景是多样的、不确定的，可能学过信息论，
可能只学过概率论，也可能什么都没学过。要衡量这本书的信息量，
我们可以考虑所有可能的读者背景，然后给出一个信息的概率分析。例如，
读这本书的读者大概率不是信息论专家，但有一定概率论的背景，
他们可以获得很多信息；而还有很少部分读者精通信息论，因此这本书
给他们的信息量就很少。但综合来看，这本书的信息量依然是不少的。□

以上例子说明了这样一种思想：将世界视为不确定的，有多种可能
的结果，然后考虑这一堆结果所带来的平均信息量。

我们可以用数学来表述上面的考虑，假如我们进行一次试验，一共有 n 种可能的结果，第 i 种发生的概率为 p_i . 我们预测试验的结果，如果越能正确地预测，那么就说明我们对这个试验中包含的信息知道的越多.

- 假如 $p_1 = 1$ ，那么我们完全确定试验一定会产生结果 1；
- 如果 $p_i = 1/n$ ，那么我们完全无法预计试验的结果.

我们对试验结果的预期与试验结果的概率分布有密切联系. 因此概率分布给我们带来了信息，使得我们能够产生不同的判断. 另一方面，概率分布带来了不确定性，使我们不能总是确信预言会成真.

我们遵循“信息论之父”Shannon 的思路 [Sha48]，为信息提供一个严格的数学模型：熵. 假设随机变量 X 表示了所有可能的结果（编号为 1 到 n ）， $\Pr(X = i) = p_i$, $p = (p_1, \dots, p_n)$ ，有时候也把 p_i 写作 $p(i)$. 我们把不确定性度量记为 $H(p)$. Shannon 假设 H 满足以下三个性质：

1. H 是一个连续函数.
2. 事件结局可能数变多则不确定性增大： $p_i = 1/n$ 时， $H(p)$ 随 n 单调递增， n 是正整数.
3. 如果一个试验被分成了两个相继的试验，那么原来的 H 应该等于分开之后的 H 的加权和.

前两个假设都比较好理解，我们现在具体解释第三个假设.

如图图 3.1 所示，假设我们有一个试验，有三种可能的结果，1, 2, 3，概率分别为 $1/2, 1/3, 1/6$. 该试验的不确定性是 $H(1/2, 1/3, 1/6)$.

我们把试验分成两步相继的试验（右图）. 第一步试验有两种可能的结果，概率分别都是 $1/2$.

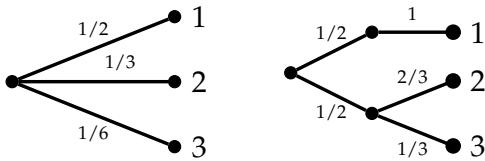


图 3.1: Shannon 的第三个假设

- 当第一步试验出现上面的结果时，第二步试验以概率 1 产生结果 1；
- 当第二步试验出现下面的结果时，第二步试验以概率 2/3 产生结果 2，以概率 1/3 产生结果 3.

我们可以看到，分成两步之后，第一步试验的不确定性是 $H(1/2, 1/2)$ ，第二步试验的不确定性有一半概率是 $H(1)$ （上面的分支），有一半概率是 $H(2/3, 1/3)$ （下面的分支），因而加权的不确定性是 $1/2 \cdot 0 + 1/2 \cdot H(2/3, 1/3)$. 因此第三个假设可以具体表述为

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \left[\frac{1}{2} \cdot H(1) + \frac{1}{2} \cdot H\left(\frac{2}{3}, \frac{1}{3}\right)\right].$$

这里，我们可以看出 Shannon 的哲学思想：不确定性只来自于概率分布而不是具体对象. 他的考虑具有浓厚的工程意味，正如他自己针对通信的数学理论所说：“消息是具有含义的……然而，通信的语义层面并不是工程问题所关心的.” 正是因为抽象掉了具体考虑的对象，信息论的应用才变得如此广泛.

基于上面三个假设，Shannon 证明了如下定理，这一定理直接给出了熵的概念.

定理 3.1 (Shannon 定理) H 满足三个假设当且仅当

$$H(p) = -C \sum_i p_i \log p_i,$$

其中 C 是正常数, $0 \log 0 = 0$.

这一定理的证明较长并且和后面的讨论关联较小, 所以我们在第 3.5 节中给出证明.

根据对数的换底公式, 可以将 $C \log p_i$ 写为 $\log_b p_i$, 这里 $C = 1 / \log b$. 于是, Shannon 定理直接给出了熵的如下定义:

定义 3.1 (熵) 分布列 $p = (p_1, \dots, p_n)$ 的熵定义为

$$H(p) = - \sum_{i=1}^n p_i \log_b p_i.$$

其中 $b = e$ (自然对数底数), $0 \log 0 = 0$. 当 $b = 2$ 时, 我们记熵为 $H_2(p)$. \square

通常来说, 使用 e 作为底数会使得数学推导简洁, 而用 2 为底数则常常是讨论信息量时的习惯. 在第 3.3 节中, 我们将讨论熵在通信中的含义, 以 2 为底的时候熵的实际意义会更清楚些. 如果没有特别强调, 我们在讨论时总是假设 $b = e$.

熵的定义还可以用数学期望的形式写出. 假设 X 的分布列是 p , $p(i) = \Pr(X = i)$, 那么我们也可以把熵写成期望的形式:

$$H(p) = -\mathbb{E}[\log p(X)].$$

每一个 (离散) 随机变量 X 会确定一个分布列 p_X , 因此我们也可以定

义随机变量的熵：

定义 3.2 (随机变量的熵) 随机变量 X 的熵定义为

$$H(X) = -\mathbb{E}[\log p_X(X)].$$

其中 p_X 是 X 的分布列, $0 \log 0 = 0$. □

尽管从信息论的角度我们可以唯一确定熵的定义, 但是熵的概念在物理学上早就已经存在. 下面我们给出统计力学中熵的推导过程, 更详细的讨论请参见统计物理教科书, 例如 [Tol10].

在经典力学中, 物理系统的状态由粒子的位置和动量 (也就是速度) 完全确定, 将粒子位置和动量可能的值集合称为相空间, 于是物理系统的演化就是相空间中的粒子状态的变化.

将相空间等分成 m 个单元, 编号 1 到 m . 假设相空间中有 N 个可区分的粒子, 相互独立, 没有相互作用, 每个粒子等可能出现在每一个单元中. 如果单元 i 中有 N_i 个粒子, 那么按照粒子在单元中的分布来看, 系统处于某个特定状态的概率为

$$P = \frac{N!}{N_1! \dots N_m!} \left(\frac{1}{m}\right)^N.$$

这是一个多项分布. 两边取对数, 得

$$\log P = \log(N!) - \sum_i \log(N_i!) - N \log m.$$

考虑充分大的 N_i , 由 Stirling 公式, 有

$$\log(N_i!) \sim \log \left(\sqrt{2\pi N_i} \left(\frac{N_i}{e} \right)^{N_i} \right) \sim N_i \log N_i.$$

因此,

$$\log P \sim N \log N - \sum_i N_i \log N_i - N \log m \sim N \log N - \sum_i N_i \log N_i. \quad (3.1)$$

假设 N_i 充分大的时候, N_i/N 呈现固定的比例 p_i , 那么

$$\begin{aligned} N \log N - \sum_i N_i \log N_i &\sim N \log N - \sum_i N p_i \log(N p_i) \\ &= -N \sum_i p_i \log p_i. \end{aligned}$$

$\log P \sim -N \sum_i p_i \log p_i$. 于是我们证明了:

$$\frac{1}{N} \log P \rightarrow H(p_1, \dots, p_m), \quad N \rightarrow \infty.$$

因此, 熵刻画了充分多粒子的物理系统某种特定状态出现概率! 熵越大的系统越有可能达到. 更进一步, 在统计力学中有 Boltzmann H -定理: 孤立的粒子系统会向着熵 (H) 增加的方向演化, 并最终达到熵最大的状态 [Uff22]. H -定理是热力学第二定律的微观解释, 熵越大的系统出现概率越大、越混乱、越接近均衡.

§3.1.2 概念与性质

现在, 我们将进一步探讨熵的若干拓展定义, 并讨论他们的性质.

首先，我们考虑最简单的情形，即分布列为 (p_1, p_2) ，此时，我们不妨设 $p_1 = p$, $p_2 = 1 - p$ ，那么熵就是

$$H(p_1, p_2) = H(p, 1 - p) = -p \log p - (1 - p) \log(1 - p).$$

H 是关于 p 的函数，作图如图 3.2 所示。

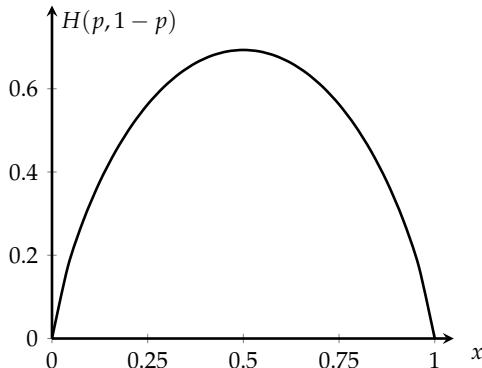


图 3.2: 熵 $H(p)$ 的图像。

利用导数的方法，很容易证明：

命题 3.1 $H(p)$ 在 $p \in (0, 1/2)$ 严格单调递增，在 $p \in (1/2, 1)$ 严格单调递减。它的最小值是 0，在 $p \in \{0, 1\}$ 取得；它的最大值是 $\log 2$ ，在 $p = 1/2$ 取得。

证明. 我们有

$$H'(p) = -\log p - 1 - \log(1 - p) + 1 = \log(p^{-1} - 1),$$

$H'(p)$ 是一个关于 p 单调递减的函数。我们有如下分类讨论：

- 当 $p \in (0, 1/2)$ 时, $p^{-1} - 1 > 1$, 所以 $H'(p) > 0$, $H(p)$ 是单调递增的;
- 当 $p \in (1/2, 1)$ 时, $p^{-1} - 1 < 1$, 所以 $H'(p) < 0$, $H(p)$ 是单调递减的;
- 当 $p = 1/2$ 时, $H'(p) = 0$, 结合前两点, $H(p)$ 取得最大值.

$H(0) = H(1) = 0$, $H(1/2) = \log 2$, 因此命题得证. \square

这与我们对于“不确定性”的直觉是相一致的: 当 p 接近 0 或 1 时, 我们对于 X 的取值几乎是确定的, 因此熵接近 0; 当 p 接近 $1/2$ 时, 我们对于 X 的取值几乎是完全不确定的, 因此熵接近最大值 $\log 2$.

实际上, 这样的性质对于一般的分布也是成立的, 我们分别将他们写在命题 3.2 和命题 3.4 中.

考虑一般分布的熵 $H(p) = H(p_1, \dots, p_n)$. 我们有如下性质:

命题 3.2 $H(p) \geq 0$, 等号成立当且仅当某个 $p_i = 1$.

证明. 这是一个典型的证明, 主要的技巧是使用熵的期望形式. 考虑随机变量 X , 其分布列为 p . 回忆 Jensen 不等式 (定理 C.17): 如果 f 是一个严格凸函数, 那么

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

等号成立当且仅当 X 是常数.

因为 $-\log(\cdot)$ 是严格凸函数, 所以根据 Jensen 不等式

$$H(X) = \mathbb{E}[-\log p(X)] \geq -\log \mathbb{E}[p(X)] \geq -\log 1 = 0.$$

等号成立当且仅当 X 是常数，即对某个 i , $p(i) = 1$. \square

命题 3.3 p_i 朝着相等方向改变的时候 H 增加. 也就是说，假设

$$p_i < p'_i \leq p'_j < p_j, \quad p_i + p_j = p'_i + p'_j,$$

那么，用 p'_i 和 p'_j 代替原来的 p_i 和 p_j , H 将会变大.

证明. 为简化符号，考虑 $i = 1$ 和 $j = 2$ ，一般情况是一样的证明. 利用假设三，第一步试验中，将试验的结果 1 和结果 2 合并，第二步试验再按照 $p_1/(p_1 + p_2)$ 和 $p_2/(p_1 + p_2)$ 的概率产生结果 1 和结果 2. 于是，

$$\begin{aligned} & H(p_1, p_2, \dots) \\ &= H(p_1 + p_2, p_3, \dots) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \quad (\text{假设三}) \\ &\leq H(p_1 + p_2, p_3, \dots) + (p_1 + p_2) \\ &< H\left(\frac{p'_1}{p'_1 + p'_2}, \frac{p'_2}{p'_1 + p'_2}\right) \quad (\text{命题 3.1}) \\ &= H(p'_1, p'_2, p_3, \dots). \quad (\text{假设三}) \end{aligned} \quad \square$$

命题 3.4 当且仅当 $p_1 = \dots = p_n = 1/n$ 时 H 取得最大值 $\log n$.

证明. 用反证法，假设达到最大值的时候存在 $p_k \neq 1/n$ ，那么，因为 $\sum_i p_i/n = 1/n$ ，根据鸽巢原理，则必有 i, j 满足

$$p_i < 1/n < p_j.$$

根据命题 3.3，我们可以将 p_i 和 p_j 替换为 $1/n$ 和 $p_i + p_j - 1/n$ ，而 H

增大，这与假设矛盾。因此， $p_1 = \dots = p_n = 1/n$ 是 H 取得最大值的必要条件。

另一方面，因为 H 连续，所以根据 Weierstrass 最值定理（定理 B.7）， H 一定有最大值，所以 $p_1 = \dots = p_n = 1/n$ 也是 H 取得最大值的充分条件。□

至此，命题 3.2 和命题 3.4 证明了一般情形的命题 3.1。在等可能的时候不确定性最大，熵最大；在确定事件的时候不确定性最小，熵最小。所以熵是符合直观的定义。

接下来，我们讨论熵的拓展形式。

在一次试验中，我们可以观察多个变量，比如说 X 和 Y 。等价地，我们其实只观察到了一个结果 (X, Y) ，只是这个结果是一个向量的形式，服从分布 $p(i, j)$ 。因此，这一向量也有对应的熵，这就是联合分布的熵：

$$H(X, Y) = -\mathbb{E}[\log p(X, Y)].$$

对应地，我们也可以写成和的形式：

$$H(p) = -\sum_{i,j} p(i, j) \log p(i, j).$$

自然，联合分布也可以引出边缘分布的熵：

$$H(X) = -\mathbb{E}[\log p_X(X)] = -\sum_i \sum_j p(i, j) \log \sum_j p(i, j).$$

$$H(Y) = -\mathbb{E}[\log p_Y(Y)] = -\sum_j \sum_i p(i, j) \log \sum_i p(i, j).$$

有了两个随机变量，我们就可以讨论“条件”的概念。具体来说，我

我们可以把试验分为两步，第一步观测 X ，第二步观测 Y ，那么，第二步所产生的熵就是已经知道第一步结果之后的熵，即：

$$\begin{aligned} H(Y|X=x) &= -\mathbb{E}[\log p_{Y|X=x}(Y)|X=x] \\ &= -\sum_j p_{Y|X=x}(j) \log p_{Y|X=x}(j), \end{aligned}$$

其中 $p_{Y|X=x}(j) = p(x, j)/p_X(x)$. 当我们知道了 $X = x$ 之后，对 Y 的观测就消除了部分的不确定性，因此根据我们对于不确定性和信息关系的讨论，从 $X = x$ 中获得的关于 Y 的信息是

$$I(X=x : Y) = H(Y) - H(Y|X=x).$$

考虑一个特殊情况， $Y = X$ ，那么刚刚的讨论就变成了自己从自己身上获得的信息，或者说知道 $X = x$ 带来的信息量. 首先有

$$p_{X|X=x}(i) = \begin{cases} 1, & i = x \\ 0, & i \neq x. \end{cases}$$

因此，

$$H(X|X=x) = -\sum_j p_{X|X=x}(j) \log p_{X|X=x}(j) = -1 \log 1 = 0.$$

于是，

$$I(X=x : X) = H(X) - H(X|X=x) = H(X).$$

这正是定量版本的“消除不确定性的信息”！此外，我们

之前说过，熵刻画的是一族可能对象的信息，这一点也反映在了这一公式中：只要知道了 X 的值，无论它具体是多少，我们得到的信息量是一样的！

再回到一般情况，还是同样的两步试验，我们定义给定 X 时 Y 的条件熵为

$$\begin{aligned} H(Y|X) &= \mathbb{E}[H(Y|X=x)] \\ &= -\mathbb{E}[\log p_{Y|X}(Y)] \\ &= -\sum_x p_X(x) \sum_j p_{Y|X=x}(j) \log p_{Y|X=x}(j) \\ &= -\sum_{x,j} p(x,j) \log p_{Y|X=x}(j). \end{aligned}$$

换言之，我们现在进一步假定 X 也是不知道的，于是 $H(Y|X)$ 就是平均上来说第二步中 Y 的不确定性。条件熵和熵有着类似的性质：

命题 3.5 $H(Y|X) \geq 0$ ，等号成立当且仅当 Y 是退化的，即 Y 概率 1 只取一个值。

证明. 仿照命题 3.2 的证明即可。 □

类似地，我们可以考虑平均上 Y 中包含的关于 X 的信息量：

$$\mathbb{E}[I(X=x : Y)] = H(Y) - H(Y|X).$$

与之相对应地，平均上 X 中包含的关于 Y 的信息量为

$$\mathbb{E}[I(Y=y : X)] = H(X) - H(X|Y).$$

一个自然的问题是，二者相互包含的信息量是什么关系？根据概率的链式法则， $p(x, y) = p_{X|Y}(x|y)p_Y(y)$ ，带入 $H(X, Y)$ 的定义得熵的链式法则：

命题 3.6 对任意离散随机变量 X, Y ， $H(X, Y) = H(Y) + H(X|Y)$.

利用链式法则，我们注意到，

$$\begin{aligned} H(X) - H(X|Y) &= H(X) - (H(X, Y) - H(Y)) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(Y) - (H(X, Y) - H(X)) \\ &= H(Y) - H(Y|X). \end{aligned}$$

所以， X 中包含的 Y 的信息和 Y 中包含的 X 的信息是一样多的！

此外，直观上我们还应该觉得，信息量不能是负的，实际上的确如此：

命题 3.7 $H(X) - H(X|Y) \geq 0$ ，等号成立当且仅当 X 和 Y 相互独立。

我们将在第 3.2 节看到，命题 3.7 就是 K-L 散度信息不等式的一个特例，所以我们就不在这里给出证明了。命题 3.7 表明知道任何信息都不会增加不确定性，这个原理被称为“Information doesn't hurt.”

根据以上讨论，我们可以自然地定义 X 和 Y 的互信息为

$$I(X; Y) = I(Y; X) = \mathbb{E}[I(X = x : Y)] = \mathbb{E}[I(Y = y : X)].$$

类似联合分布的熵，条件熵和互信息的概念也可以推广到多元情

形. 对于三个随机变量 X, Y, Z , 我们可以定义条件熵为

$$H(X, Y|Z) = H(X, Y, Z) - H(Z).$$

类似地, 我们可以定义互信息为

$$I(X, Y; Z) = H(X, Y) - H(X, Y; Z).$$

他们的含义以及性质和二元情形类似.

同样, 我们可以定义条件互信息为

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z),$$

表明 Z 已知时候 Y 给 X 带来的平均信息增益. 类似互信息, 我们如下性质:

命题 3.8 条件互信息满足以下性质 :

1. 非负性 : $I(X; Y|Z) \geq 0$, 等号成立当且仅当 X 和 Y 在给定 Z 的条件下相互独立.
2. 对称性 : $I(X; Y|Z) = I(Y; X|Z)$.
3. 链式法则 : $I(X, Y; Z) = I(X; Z|Y) + I(Y; Z)$.
4. 条件信息量 : $I(X : X|Y) = H(X|Y) - H(X|X, Y) = H(X|Y)$.

最后一条性质说的其实是, 在平均的意义下, 给定 Y 的时候, 知道 X 所能够得到的额外信息量就是 $H(X|Y)$. 这一命题的证明和前面都非常相似, 见习题 1.

最后，我们将各种熵以及信息量的关系总结为图 3.3. 在集合论中，这样的图被称为 Venn 图，所以我们可以用集合论来理解信息与熵. 对应关系可以总结为表 3.1.

信息论	集合论
$H(X)$	A
$H(Y)$	B
$H(X Y)$	$A \setminus B$
$H(X, Y)$	$A \cup B$
$I(X; Y)$	$A \cap B$

表 3.1: 信息论和集合论的对应关系.

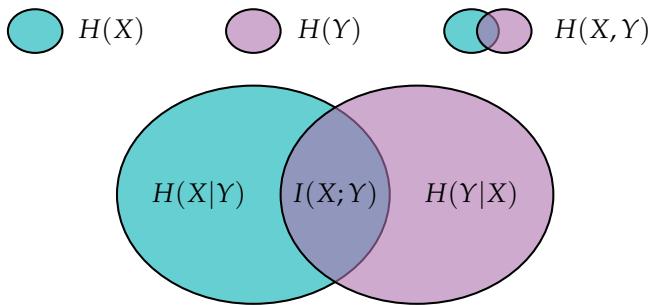


图 3.3: 熵和信息量的关系

§3.2 Kullback-Leibler 散度

§3.2.1 定义

为了引入 K-L 散度，我们从互信息出发。它的定义是：

$$\begin{aligned}
 I(X;Y) &= H(X) - H(X|Y) \\
 &= -\sum_x p_X(x) \log p_X(x) + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_Y(y)} \\
 &= -\sum_{x,y} p(x,y) \log p_X(x) + \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_Y(y)} \\
 &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)}.
 \end{aligned}$$

根据命题 3.7， $I(X;Y) \geq 0$ ，等号成立当且仅当 X, Y 相互独立，即 $p(x,y) = p_X(x)p_Y(y)$. X, Y 之间的互信息越大，说明他们之间的关联越强，分布越不独立， $p(x,y)$ 越不接近 $p_X(x)p_Y(y)$.

上面的推导说明，互信息其实在用分布列的比值比较两个分布的接近程度，这样的想法被 Kullback 和 Leibler [KL51] 推广到了一般分布上。考虑两个概率分布的似然函数 p_1 和 p_2 （也就是他们的分布列）。

抽取一个样本 X ，考虑如下两个假设：

H_1 : 样本 X 来自 p_1 的分布 vs. H_2 : 样本 X 来自 p_2 的分布

为了判断哪个假设是更有可能的，我们考虑两个假设分布的似然比 p_1/p_2 . 如果这个比值越大，就越说明 p_1 的值更大，因而更有可能，倾向于接受 H_1 ，反之则越倾向于接受 H_2 . 这种方法被称之为似然比检验法。

从上述讨论出发，我们定义区分 H_1 和 H_2 的检验量为对数似然比：

$$\log(p_1(x)/p_2(x)).$$

假设 H_1 是真的，那么在 H_1 成立的世界里，这个检验量的期望为

$$\mathbb{E}_{X \sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i) \log \frac{p_1(i)}{p_2(i)}.$$

期望越大，说明 H_1 越有可能成立。实际上，上面的期望就是 K-L 散度的定义。

定义 3.3 (Kullback-Leibler 散度，相对熵) 对于概率分布 p_1, p_2 ，他们的 **Kullback-Leibler 散度（相对熵）** 定义为

$$D_{\text{KL}}(p_1 \| p_2) = \mathbb{E}_{X \sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i) \log \frac{p_1(i)}{p_2(i)}.$$

其中规定 $0 \log(0/0) = 0$, $0 \log(0/a) = 0$, $a \log(a/0) = +\infty$. □

我们马上知道，互信息是 K-L 散度的一种特殊情况：

命题 3.9 对于两个随机变量 X, Y ，成立 $I(X; Y) = D_{\text{KL}}(p_{X,Y} \| p_X p_Y)$ ，其中 $p_{X,Y}$ 是 X, Y 的联合分布列， p_X, p_Y 分别是 X, Y 的边缘分布列。

K-L 散度可以看成两个分布之间的区分衡量标准，但他不是度量。一般来说，甚至连对称性都不成立。例如，设 p_1 和 p_2 都是定义在 $0, 1$ 上的 Bernoulli 分布，参数分别为 $1/2$ 和 $1/4$ 。于是

$$D_{\text{KL}}(p_1 \| p_2) = \frac{1}{2} \log \frac{1/2}{3/4} + \frac{1}{2} \log \frac{1/2}{1/4} = \frac{1}{2} \log \frac{4}{3}.$$

$$D_{\text{KL}}(p_2 \| p_1) = \frac{3}{4} \log \frac{3/4}{1/2} + \frac{1}{4} \log \frac{1/4}{1/2} = \frac{1}{2} \log \frac{3\sqrt{3}}{4}.$$

这两个值是不相等的. 进一步, 这个差值甚至可以到任意大 (见习题 5)

上面推导 K-L 散度的过程看起来有些捏造, 我们将在第 3.3 节中给出一个非常直观的理解方式. 现在, 我们先接受这个定义, 然后看一下 K-L 散度的一些性质.

§3.2.2 两个关于信息的不等式

利用 K-L 散度, 我们可以给出两个关于信息的不等式, 它们分别是信息不等式和数据处理不等式.

定理 3.2 (信息不等式) 对于两个概率分布列 p, q , 成立 $D_{\text{KL}}(p \| q) \geq 0$, 当且仅当 $p = q$ 时取等号.

证明. 由于 $\log x$ 是凸函数, 所以由 Jensen 不等式, 我们有

$$\begin{aligned} D_{\text{KL}}(p \| q) &= -\mathbb{E}_{X \sim p} \left[\log \frac{q(X)}{p(X)} \right] \\ &\geq -\log \mathbb{E}_{X \sim p} \left[\frac{q(X)}{p(X)} \right] \\ &= -\log \sum_i p(i) \cdot \frac{q(i)}{p(i)} = 0. \end{aligned}$$

因此, $D_{\text{KL}}(p \| q) \geq 0$, 当且仅当 $p = q$ 时取等号. \square

信息不等式表明, K-L 散度虽然不是度量, 但却是非负的, 因而确实可以被作为熵, 用来衡量“额外的不确定性”. 此外, 命题 3.7 是信息

不等式的直接推论. 利用类似的方法, 我们可以证明条件互信息的非负性 (即命题 3.8 中的第一条) .

接下来我们叙述并证明数据处理不等式.

定理 3.3 (数据处理不等式) 假设随机变量 X, Y, Z 形成了 *Markov 链*, 那么 $I(X; Y) \geq I(X; Z)$. 特别地, 对任意函数 f , 成立

$$I(X; Y) \geq I(X; f(Y)).$$

证明. 根据互信息链式法则,

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y). \end{aligned}$$

根据 *Markov 性*, 条件在 Y 上, X 和 Z 相互独立. 因此, $I(X; Z|Y) = 0$, 根据条件互信息的非负性, $I(X; Y|Z) \geq 0$, 所以 $I(X; Y) \geq I(X; Z)$.

显然, $X, Y, f(Y)$ 也形成了 *Markov 链*, 所以 $I(X; Y) \geq I(X; f(Y))$. \square

数据处理不等式表明, 无论我们对随机变量 Y 进行了何种处理, 甚至是允许带随机的处理, 它的信息量都不会增加. 这里举一个机器学习中的例子. 在大模型的时代, 人类的原生数据是远远不够用的, 因此, 人们会尝试使用合成数据的方法来增加数据量. 但是, 究竟哪一种合成数据才能真正让模型受益呢? 数据处理不等式可以给我们答案.

假设原生数据是 X , Y 是通过 X 训练得到的大模型的输出, Z 是对 Y 的任意一种处理, 例如添加一些随机的噪声, 或者不作任何处理. 数据处理不等式告诉我们, X 和 Z 之间的信息量不会比 X 和 Y 之间的信息量更大, 也就是说, Z 并不能带来更多的信息. 因此, 单纯使用大模

型的输出，而没有额外的信息注入，是不会带来任何帮助的。

相反，如果 Z 有额外的信息注入，例如 Z 是对 Y 进行的人工标注，那么 $I(X; Z)$ 就会比 $I(X; Y)$ 更大，这样的数据处理才是有意义的，典型的例子如人类反馈强化学习 (RLHF)，这是现在改善大模型表现的一种有效方法。

另一种带来额外信息的方式是，我们引入了一系列规则来计算 Z 的奖励，然后利用强化学习的方法来训练 Y ，这样的数据处理也是有意义的。例如，DeepSeek R1 模型训练的过程中，对于编程题和数学题有关的训练数据，都知道正确答案，因此可以直接通过大模型有没有输出正确答案来计算奖励，进行强化学习。通过这样的办法，DeepSeek R1 可以超越没有做强化学习的基础模型。

注。关于信息论的经典教科书，可以参见 [CT12]，此外，概率论的教材中也有很多很好的讨论，比如 [Dur19], [Shi96] 以及中文教科书 [李 10]。

§3.3 编码理论

最早的时候，Shannon 建立信息论，就是为了给通信和编码理论一个数学基础。从编码的角度出发，我们可以更本质地理解信息和熵。

§3.3.1 熵与编码

通信就是一个发射端和一个接收端，中间有信道传递消息。将所有可能要传递的消息集合记为 Ω （一个有限集），我们现在考虑 Ω 所蕴含的信息量是多少。注意，根据 Shannon 的思想， Ω 里面具体是什么并不重要，重要的是有多少个。我们可以用自然数 $1, 2, \dots$ 表示集合 Ω 里的

元素.

使用二进制编码，我们至少需要 $\log_2 |\Omega|$ 个比特来表示 Ω 里的元素. 于是，假如说随机变量 X 表示收到的消息，那么 X 的熵就定义为 $H(X) = \log_2 |\Omega|$ ，它衡量了接收端收到的消息的不确定性.

当我们选定了具体的消息 $m \in \Omega$ ， X 的不确定性被消除了，于是 $X = a$ 的过程产生了（或者说传递了） $\log_2 |\Omega|$ 比特的信息. 比如说，我们发送一个长为 n 的二进制序列，消息的集合大小就是 2^n ，发送任何一条具体的消息，我们就传递了 n 比特的信息.

有时候，我们会把消息看成一个序列. 具体来说，我们可以发送独立的 k 条消息，其中第 i 条 X_i 来自消息集合 Ω_i ， $|\Omega_i| = n_i$ ，那么 (X_1, \dots, X_k) 的熵就是

$$H(X_1, \dots, X_k) = \log_2 n_1 + \dots + \log_2 n_k,$$

它衡量了 k 条消息的不确定性.

在更常见的情况下，每次发送的其实不是一条消息，而是一个字母，所有的字母组成了一个字母表，我们用 $\Sigma = \{x_1, \dots, x_s\}$ 来表示. 于是， X_i 就是消息的第 i 个字母，于是，一条消息可以写作 $X_1 \dots X_k$ ，其中每一个 X_i 都来自 Σ .

我们现在考虑更简单的情形，即每个字母 X_i 其实是同一个随机变量 X 的独立采样. 如果具体知道某一个 x_i 出现的次数，我们其实可以有更高效的传递信息的方式. 比如说，在极端情况下，如果只有 x_1 和 x_2 会出现，那么我们其实只需要 $\log_2 2 = 1$ 比特就足够传递所有消息了：0 表示 x_1 ，1 表示 x_2 .

在一般情况下，考虑 Ω 中只包含长为 k 的消息，并且 x_i 在消息中

出现 k_i 次，那么所有可能的消息数量为

$$|\Omega| = N(k) = \frac{k!}{k_1! \cdots k_s!}.$$

假定我们需要 $h(\omega)$ 比特来具体确定发的消息是 ω ，我们来推导 $h(\omega)$ 的上下界。

- 无论如何编码，我们需要能区分 Ω 中的不同元素，这本身需要 $\log_2 |\Omega|$ 比特来表示。
- 在一些编码中，我们还需要确定 (k_1, \dots, k_s) 。确定它的一种方式是按照顺序给出每一个 k_i 。因为 $k_i \leq k$ ，每个 k_i 最多需要 $\log_2 k$ 比特来表示，所以按顺序表示所有的 k_i 至多需要 $s \log_2 k$ 比特。

于是，我们需要的比特数为

$$\log_2 \frac{k!}{k_1! \cdots k_s!} \leq h(m) \leq s \log_2 k + \log \frac{k!}{k_1! \cdots k_s!}.$$

这刚好和我们在统计力学中推导熵的过程是一致的！假设消息足够长， x_i 出现的频率逐渐接近 p_i ，那么同样的推理我们可以知道，

$$h(m) \sim -k \sum_i p_i \log_2 p_i = kH_2(p_1, \dots, p_s).$$

因此，如果知道字母的出现频率，我们传递单位长度的消息至少需要 $H(p_1, \dots, p_s)$ 比特，这完全给出了熵的具体含义，而且，我们现在也不难理解熵的形式为何会出现 \log 了：熵就是期望上编码一个字母需要的比特数（即 $\log(1/p(X))$ ）。Shannon 给出了一种渐进意义上的最优编码，它的编码长度达到了这个下界，后来独立地被 Fano [Rob49] 以一

种不同的形式发现，这种编码被称为 *Shannon-Fano* 编码。

那么，是否有一种编码确实达到了这个理论上的编码长度（非渐进）下界呢？答案是肯定的，它被称为 *Huffman* 编码。它的核心思想在于把出现频率高的字母用更短的编码表示。类似的思想被用在了机器学习的决策树中，作为选择节点非常常用的一种依据。

注. 决策树是一种常用的机器学习分类模型。假设数据有很多属性，记为 P_1, \dots, P_k ，这些属性共同决定了某一条数据的类别。比如，在银行的信用系统中，给定了一个人的性别、是否已婚、是否负债等信息，我们希望给他评估一个信用评级。

决策树的做法是，将决策过程写成一棵树，然后叶节点是决策类别的结果。比如说，我们会先看这个人是否负债，如果不负债，那么看是否已婚，如果已婚，那么我们信用评级就给 A。但如果负债，那么我们信用评级就给 B。

那么，每个节点应该判断什么属性呢？树本身其实是一种广义的消息，沿着树，从根节点走到叶节点得到的就是一条消息。于是，在这一观点下，我们可以用熵与编码的关系来选择属性。

如果我们选择编码最短的属性，这样我们的决策树就会更加简单。一种近似的做法是，对于每个节点，都优先选择信息增益最高的属性。这样的选择方式叫做 *ID3* 策略。

我们进一步的问题是，为什么我们知道了每个字母的频次就可以压缩编码？我们接下来将要说明，其实长为 k 的消息中的“典型消息”的数量远远少于所有 k 长消息的数目，因此我们实际上相当于只是针对一个子集进行编码。注意到，当 k 充分大的时候，

$$\log_2 N(k) \sim h(m) \sim kH_2(p_1, \dots, p_s).$$

因此,

$$N(k) \approx 2^{kH_2(p_1, \dots, p_s)} = e^{kH(p_1, \dots, p_s)}.$$

然而, 长为 k 的所有消息数目为

$$s^k = e^{k \log s}.$$

根据命题 3.4, 只有当所有 p_i 相等的时候 $N(k)$ 才会达到这一量级. 从这个意义上说, 熵所刻画的信息量定量刻画了数据压缩可能的极限.

以上关于信息编码下界以及数据压缩的讨论, 再更一般的情况下也成立, 此时这样的性质被称为渐近等分性. 而这一性质成立对应的结果被称为 *Shannon-McMillan-Breiman* 定理, 它的陈述以及证明都需要用到更多随机过程的知识, 这里就不再给出了, 感兴趣的读者请参阅 Shannon 的工作 [Sha48]、McMillan 的工作 [McM53] 和 Breiman 的工作 [Bre57].

§3.3.2 K-L 散度、交叉熵与编码

我们在 K-L 散度的定义中提到了它的另一个名字——相对熵. 实际上, 这可以从编码中看出来. 假设事实上消息中字母的分布是 p_1 , 那么期望上编码单位长度消息需要的比特数是

$$H(p_1) = \mathbb{E}_{X \sim p_1} [\log p_1(X)].$$

如果我们错误地认为消息中字母的分布是 p_2 并使用最优编码，那么实际上期望编码单位长度消息需要的比特数是

$$\mathbb{E}_{X \sim p_1} [\log p_2(X)].$$

由于错误的认识所产生的额外编码长度是

$$\mathbb{E}_{X \sim p_1} [\log p_1(X) - \log p_2(X)] = D_{\text{KL}}(p_1 \| p_2).$$

根据本节中的讨论，我们知道，额外的编码长度代表的是额外的不确定性，因而这一概念是某种“熵”的概念。这正是“相对熵”的由来， $D_{\text{KL}}(p_1 \| p_2)$ 表示了当我们错误地把 p_1 当成 p_2 时带来的额外的不确定性，或者说额外的信息损失。

在机器学习中，比起讨论 K-L 散度，更加常用的是直接讨论量 $\mathbb{E}_{X \sim p_1} [\log p_2(X)]$ 。从机器学习的观点来说， p_1 是真实的分布，而 p_2 是我们所学到的分布。根据刚刚的讨论，这个量越小越说明 p_2 接近真实的 p_1 ，因此这是一种衡量两个分布之间关系的量，我们称之为交叉熵：

定义 3.4 (交叉熵) 给两个随机变量 X, Y ， X 的分布为 p_X ， Y 的分布为 p_Y ，则 X 的分布 p_X 和 Y 的分布 p_Y 的交叉熵¹为

$$CH(p_X, p_Y) = -\mathbb{E}_{X \sim p_X} [\log p_Y(X)] = -\sum_i p_X(i) \log p_Y(i). \quad \square$$

在机器学习的分类问题中，我们希望学到的分布 p_Y 尽可能地接近真实的分布 p_X ，所以训练的目标经常是最小化交叉熵 $CH(p_X, p_Y)$ 。

¹文献中，经常会直接写为 $H(p_X, p_Y)$ ，但是在本书中为了区分熵，我们使用了符号 CH 。

有趣的是，从数理统计的角度来看，最小化交叉熵等价于进行最大似然估计（见习题 8），因此这为最大似然估计提供了一种信息论意义上的理解。

注. 现代的主流信息论都是从 Shannon 发展起来的。然而，这一信息论也有很多问题。

- 信息论使用了概率论进行建模。但我们可以看到，概率要么是作为频率的近似理论（频率学派），要么反映了人们对未知的信念（Bayes 学派）。无论哪种解释，都将问题简化了。正如 Kolmogorov 所说：“如果事情没有按照我们的预期发展，那么问题一定出在我们对于概率和真实世界的随机之间关系不清晰的认识上。”
- 这一信息论考虑的是一族对象的信息。我们是否能够用这样的方式来衡量单个对象的信息量呢？比如，我们要考虑这本书中包含的信息量，是它放在所有可能的书的集合中去考虑呢，还是把它的每一个章节分开考虑成一个随机序列呢？因此，信息论并不能很好地回答“单个对象”的信息量的问题。

现代概率论的奠基人 Kolmogorov 也非常严肃地考虑了这一问题。他提出了被后世称为 **Kolmogorov 复杂度** 的概念，旨在刻画一个随机字符串的随机程度。简单来说，一个字符串的 Kolmogorov 复杂度就是描述输出它所需要的最短代码长度。越随机的字符串就越需要更复杂的程序去描述它的输出方式。

例如，尽管字符串 $x = 0101010101$ 看起来非常长，但是我们可以用一个很短的程序来描述它：输出 5 次“01”。然而，尽管字符串 $y = 011001$ 比 x 短得多，我们却很难找到一个简短的程序来描述它。因此， y 的 Kolmogorov 复杂度要比 x 大，因而看起来更像是随机的。

利用这一概念，我们可以将信息的概念变成一个对象自己的属性，而不再需要把对象放在可能的一堆对象中去考虑。这是信息论的另一种构建

思路。

关于 Kolmogorov 复杂度的讨论，可以参见专著 [LV19]，这本书对于随机、信息、编码、复杂度，乃至归纳推理等概念都有非常独到的见解，值得一读。

§3.4 在机器学习中的应用：语言生成模型

现如今，机器学习中最为瞩目的成果之一就是大语言模型 (LLM)，它通过学习人类海量的高质量语料库来形成一个生成式的模型，其中最为典型的例子是 ChatGPT [Cha].

从思路上来说，大语言模型的核心思想非常简单：给一段话，将其中一些词掩盖掉，让模型填出这些词来。例如，给出

“我在[mask]面条，它真好吃。”

模型应该能够填出

“我在吃面条，它真好吃。”

对于 GPT 模型 [RNSS] 来说，这一思想更加简单：永远只预测下一个词。它的哲学是“通过预测下一个词，可以理解世界。”

这样的思想，对于更一般的数据也是成立的：用（修改过的）数据本身作为输入，训练一个编码器，然后将编码器的输出送入解码器，而解码器的输出具有原始数据的格式，我们希望这一输出能够尽量匹配原始的输入。这正是本章开头 Simon 所说的“信息处理器”的具体实现。

在语言模型中，一个生成模型往往同时有编码器和解码器。比如说，图 3.4 展示的就是 BART [LLG⁺19] 的结构。

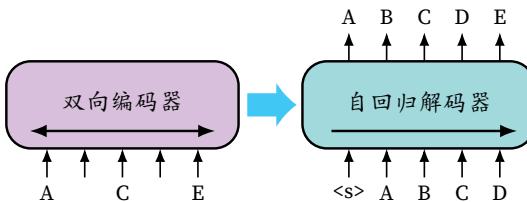


图 3.4: 生成式语言模型 BART 的示意图.

有的时候，编码器和解码器并不是显式给出的。例如，GPT 模型只有自回归解码器，没有编码器。然而，我们可以认为 GPT 的解码器实际上是一个编码器-解码器的结构：它总要处理输入的数据，因而需要编码器，同时也要输出数据，因而需要解码器。

在第 3.3 节我们指出，熵和编码有着密切的联系。从这个角度出发，我们很容易理解生成模型背后的思想：我们希望通过训练的方式得到一个由神经网络所表示的编码和解码规则，他要尽可能符合真实数据的分布。

我们可以用一种非常简单的模型去理解这一过程。假设所有的单词的集合为 Σ ，单词数为 k 的文本集合为 Ω 。我们希望训练一个生成模型 M ，给它输入 $k - 1$ 个单词，它可以给出第 k 个单词的概率分布，我们选择出现概率最大的那个词作为预测。

在训练的时候，对于一个句子 ω ，我们只保留前 $k - 1$ 个词，得到 $\omega[1 : k - 1]$ ，然后将它输入到生成模型 M 中，让它去预测第 k 个词。

对于这一个具体的句子 ω 来说，理想的分布应该是一个 Dirac 分布² $\delta_{\omega[k]}$ ，即以概率 1 取到 $\omega[k]$ 。假如说生成模型的输出是一个概率分布 $M(\omega[1 : k - 1]) = p$ ，那么，我们可以用 K-L 散度去衡量这两个分布

²Dirac 分布是一个数学物理中更加常用的名字。在概率论中，这也被称为退化分布；而在机器学习中，分布经常会表示为一个概率向量，文献中称为独热向量。

的差异.

因为 $H(\delta_{\omega[k]})$ 是固定的，所以我们只用考虑交叉熵 $CH(\delta_{\omega[k]}, p)$. 一次训练会给多个样本，所以我们的目标是同时最小化这些交叉熵的和. 假如训练集是 T ，我们的目标就是

$$\min_M \sum_{\omega \in T} CH(\delta_{\omega[k]}, M(\omega[1:k-1])).$$

实际上，这个例子是有普适性的，所有的监督训练的分类问题都可以用这种方式来建模. 而在第 6.1 节我们也会看到，此时交叉熵实际上被作为一种损失函数.

§3.5 附录：Shannon 定理的证明

我们在这一部分给出 Shannon 定理（定理 3.1）的证明. 整体上的思路是：

1. 证明如果 f 是单调函数，对正整数 m, n 成立 $f(mn) = f(m) + f(n)$ ，那么 $f(n) = C \log n$.
2. 求出 $H(1/n, \dots, 1/n)$ 的表达式.
3. 假设 p_i 是有理数，设 $p_i = n_i / \sum_j n_j$ ，考虑 $\sum_j n_j$ 个等可能试验结果，利用假设 3 推出 H 的表达式.
4. 利用有理数的稠密性和 H 的连续性推出一般情形.

最后一步是显然的，我们只需要证明前三步即可.

对第一步, 我们需要证明的是, 如果 f 是单调函数, 对正整数 m, n 成立 $f(mn) = f(m) + f(n)$, 那么 $f(n) = C \log n$. 首先, 利用数学归纳法容易看出, 对正整数 n, k , 成立

$$f(n^k) = kf(n). \quad (3.2)$$

设 m, n 是任意两个大于 1 的整数, 再选任意大的正整数 k , 从 m 进制数的性质可以看出, 总存在正整数 l 使得

$$m^l \leq n^k < m^{l+1}. \quad (3.3)$$

根据 f 的单调性, 我们有

$$f(m^l) \leq f(n^k) < f(m^{l+1}).$$

利用式 (3.2), 我们有

$$lf(m) \leq kf(n) < (l+1)f(m) \iff \frac{l}{k} \leq \frac{f(n)}{f(m)} < \frac{l+1}{k}.$$

将式 (3.3) 取对数, 得到

$$l \log m \leq k \log n < (l+1) \log m \iff \frac{l}{k} \leq \frac{\log n}{\log m} < \frac{l+1}{k}.$$

所以

$$\left| \frac{\log n}{\log m} - \frac{f(n)}{f(m)} \right| \leq \frac{1}{k}.$$

因为 k 可以是任意大的正整数, 取 $k \rightarrow \infty$, 我们就得到了

$$\frac{\log n}{\log m} = \frac{f(n)}{f(m)}.$$

由 m, n 的任意性, 取 $m = 2$, 我们就得到了 $f(n) = (f(2)/\log 2) \cdot \log n = C \log n$. 容易检验, $f(1) = 0 = C \log 1$, 因此这一等式对所有正整数 n 都成立.

对第二步, 我们需要求出 $f(n) = H(1/n, \dots, 1/n)$ 的表达式. 我们要利用第一步的结果, 首先, 根据假设二, $f(n)$ 是单调递增的函数. 然后, 考虑 mn 个等可能试验, 我们可以将它分成两步试验, 第一步有 m 中等可能的结果, 而在每一种结果之下, 第二步有 n 种等可能结果. 根据假设三,

$$f(mn) = f(m) + \frac{1}{n} \cdot nf(n) = f(m) + f(n).$$

所以 $f(n)$ 符合第一步的假设. 第二步就可以直接从第一步推出.

最后, 我们证明第三步. 设 p_1, \dots, p_n 都是有理数, 那么, 他们可以被写为

$$p_i = \frac{n_i}{\sum_{j=1}^n n_j}.$$

其中 n_i 是非负整数. 我们考虑 $\sum_j n_j$ 个等可能试验, 这个试验可以被看成两步的试验, 第一步有 n 种可能的结果, 第 i 种结果出现的概率是 p_i , 而在第 i 种结果之下, 第二步有 n_i 种等可能的结果. 根据假设三, 和证明的第三步, 我们有

$$C \log \sum_{j=1}^n n_j = H(p_1 + \dots + p_n) + \sum_{i=1}^n p_i \cdot C \log n_i.$$

因此,

$$\begin{aligned}
 H(p_1, \dots, p_n) &= C \left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log n_i \right) \\
 &= C \left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log \left(p_i \sum_{j=1}^n n_j \right) \right) \\
 &= C \left(\log \sum_{j=1}^n n_j - \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i \log \sum_{j=1}^n n_j \right) \\
 &= -C \sum_{i=1}^n p_i \log p_i.
 \end{aligned}$$

这正是我们要证明的. 于是, 我们证明了 Shannon 定理.

§3.6 习题

1. 证明命题 3.8.
2. [Tin62] 仿照集合论的思路, 我们可以定义三个随机变量的互信息为:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z).$$

- (1) 证明对称性: $I(X;Y;Z) = I(Y;X;Z) = I(X;Z;Y)$.
- (2) 举一个例子说明, 可能会有 $I(X;Y;Z) < 0$, 所以这样定义的互信息并不一定真的代表“信息量”.
3. 考虑关于 n 的正实数序列 $a_1(n), \dots, a_k(n)$ 以及 $b_1(n), \dots, b_k(n)$, 假

设对所有 i , 都成立 $\lim_{n \rightarrow \infty} a_i(n)/b_i(n) = 1$, 证明:

$$\lim_{n \rightarrow \infty} \frac{a_1(n) + \cdots + a_k(n)}{b_1(n) + \cdots + b_k(n)} = 1.$$

由此证明式 (3.1).

4. 我们在熵以及 K-L 散度的定义中, 都规定了一些无定义的量的值, 这些值并不是随便规定的, 他们实际上反映了熵或者 K-L 散度定义中的连续性.

- (1) 证明: 对给定的 $a > 0$, $\lim_{x \rightarrow 0^+} x \log(x/a) = 0$, 因此我们规定了 $0 \log 0 = 0$ 以及 $0 \log(0/a) = 0$.
- (2) 证明: 对给定的 $a > 0$, $\lim_{x \rightarrow 0^+} x \log(a/x) = +\infty$, 因此我们规定了 $0 \log(a/0) = +\infty$.
5. 证明: 对任意 $M > 0$, 存在概率分布 p_1 和 p_2 使得 $D_{\text{KL}}(p_1 \| p_2) < 1$ 但是 $D_{\text{KL}}(p_2 \| p_1) > M$.
6. 对任意离散随机变量 X 和函数 f , 证明: $H(X) \geq H(f(X))$.
7. **Pinsker 不等式.** 设 P 和 Q 是 Ω 上的两个离散概率分布, 证明

$$D_{\text{KL}}(P \| Q) \geq \frac{1}{2} \left(\sum_{x \in \Omega} |p(x) - q(x)| \right)^2.$$

提示: 令 $r(x) = \frac{p(x)}{q(x)} - 1$, 并有如下参考不等式:

$$(1+t) \log(1+t) - t \geq \frac{1}{2} \frac{t^2}{1 + \frac{t}{3}}, \quad \forall t \geq -1.$$

对于随机变量 X, Y , 成立 *Cauchy* 不等式:

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)E(Y^2).$$

8. 最大似然估计等价于最小化交叉熵. 考虑二分类的学习问题, 此时对单个样本我们观察到的结果是 0 或 1, 样本总体服从服从 Bernoulli 分布, 假设真实参数为 p , $\Pr(X=1) = 1 - \Pr(X=0) = p$, 模型参数为 θ , 假设数据集是 x_1, \dots, x_N , 他们是从总体中独立采样得到的.

(1) 将问题考虑成一个数理统计问题, 估计 θ . 写出似然函数

$$L(\theta; x_1, \dots, x_N).$$

(2) 再将问题考虑为一个信息论问题, 对第 k 个观测样本, 它的实际分布是 $\Pr(X_k = x_k) = 1$, 而写出每个样本的真实分布与估计分布之间的交叉熵的平均值

$$\frac{1}{N} S(\theta; y_1, \dots, y_N).$$

(3) 证明: 以概率 1 成立

$$\lim_{n \rightarrow \infty} \frac{1}{N} S(\theta; y_1, \dots, y_N) = CH(B(1, p) \| B(1, \theta)).$$

因此, 当样本数量足够大的时候, S 就是真实分布与模型分布的交叉熵.

(4) 证明: 对任意 N ,

$$\arg \max_{\theta} L(\theta; y_1, \dots, y_N) = \arg \min_{\theta} S(\theta; y_1, \dots, y_N),$$

也就是说, 最大似然估计等价于最小化交叉熵.

9. 在生成模型中, 我们假设数据 x 是由隐变量 z 生成的. 下面我们来推导 VAE (Variational Auto-encoder) 与扩散模型的损失函数.

(1) (VAE, [KW22]) 设 x 是由 z 生成的: 数据的生成分为两个步骤, 首先 z 由先验分布 $p_\alpha(z)$ 生成, 然后 x 由条件分布 $p_\beta(x|z)$ 生成. 为了最好的拟合数据, 我们希望最大化似然, 即 $\max p_\theta(x)$, 其中 $\theta = \{\alpha, \beta\}$, $p_\theta(x)$ 可以写成:

$$p_\theta(x) = \int p_\alpha(z)p_\beta(x|z)dz.$$

然而, 这个积分通常是不能直接求解的, 因为 z 的分布是未知的. 为了解决这个问题, 我们引入一个分布 $q_\phi(z|x)$ 来近似后验分布. 证明:

$$\log p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\beta(x|z)] - D_{\text{KL}}(q_\phi(z|x) \| p_\alpha(z)).$$

于是, 最大化似然可以通过优化上面的 ELBO (Evidence Lower Bound) 来实现.

(2) (扩散模型, [HJA20]) 我们假设数据 $x = x_0$ 是由 $z = (x_1, \dots, x_T)$ 生成的. 生成方式为逐渐对 x_t 去噪 (逆向过程), 去噪的过程是

Markov 链，起点为 x_T ，服从 $p(x_T) = \mathcal{N}(x_T; 0, I)$ ：

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t),$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

区别于 VAE，这里的后验概率 $q(x_{1:T}|x_0)$ 通过加噪（前向）过程来得到，即

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}),$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I).$$

扩散模型的思想就是学习后验概率中每一步加入的噪声，从而根据逆向过程生成数据。借助上一问的结论，证明损失函数可以写成：

$$\begin{aligned} \mathcal{L} = \mathbb{E}_q[D_{\text{KL}}(q(x_T|x_0)\|p(x_T)) + \\ \sum_{t>1} D_{\text{KL}}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1)], \end{aligned}$$

其中 $q(x_{t-1}|x_t, x_0)$ 是前向过程的后验。

10. 请查找文献回答以下问题：

- (1) Fisher 信息量是什么？它与 K-L 散度有什么样的关系？
- (2) 列举其他概率分布之间散度的概念，他们是否是度量？
- (3) 列举概率分布之间的度量，他们之间是否有关联？

第四章 高维几何， Johnson-Lindenstrauss 引 理

作为生活在三维空间中的人，一个令我们倍感着迷的问题是：如果我们的世界是四维、五维甚至更高维的，我们会看到什么样的景象？Christopher Nolan 导演的电影《星际穿越》(Interstellar) 给出了一个美妙的想象。

在近未来的地球上，资源匮乏和环境恶化使人类濒临灭绝。为了寻找新的生存空间，一群科学家和宇航员开始了一次前所未有的宇宙冒险。他们的目标是穿越一个神秘的虫洞，探索另一片星系中的适居行星。

然而，当主人公 Cooper 和他的团队成功穿越虫洞时，他们发现自己面对的并不仅仅是遥远的星系，还有更加不可思议的挑战。在探索的过程中，Cooper 最终进入了一个被称为 tesseract 的空间——一个超越我们三维世界的五维空间。在这个空间中，时间变得像空间一样可以自由导航，过去和未来不再是固定不变的线性进程，而是可以被观察和影

响的维度.

通过操控时间维度，库珀能够在这个超立方体般的结构中，跨越时间的界限，影响他女儿 Murph 的命运，从而拯救全人类. 这一情节不仅带给我们震撼的视觉体验，也揭示了一个深刻的物理与几何真理：在高维空间中，我们的直觉常常失效，必须依赖数学工具来理解和探索这些新维度的性质.

高维空间不仅是科学幻想中的概念，也是现实中的客观存在：当我们描述一个人的时候，我们会考虑他的年龄、身高、体重、学历、职业等多个属性，这些属性构成了一个多维的空间. 在这个空间中，每个人都是一个点，而这些点之间的距离和关系，构成了我们对这个人的认知和理解.

在计算机的世界中，世界的一切都被表示为了数据，而且往往是高维数据. 因此，如何理解和处理高维数据已经成为人工智能领域的一个重要问题.

本章将说明，高维空间有着令人着迷却看似矛盾的两种性质. 首先，与我们生活的三维空间相比，高维空间是极其反直觉的，这带来了望而生畏的复杂性，似乎“维数灾难”难以逾越. 然而，硬币的另一面是，高维空间中的随机数据几乎都是高度集中，因而他们其实可以被压缩到更低维度的空间中进行处理. 这一原理被广泛应用在机器学习中.

从技术上看，我们要证明 Johnson-Lindenstrauss 引理，它表明了高维随机变量的集中性. 证明这一引理所用到的概率论技术是矩法，这是机器学习理论中最为核心的几个技术之一. 因此本章也可以看做机器学习理论的一个引论.

§4.1 高维几何

§4.1.1 高维球体

首先，我们探讨高维空间中“球体”的特殊性。按照微积分的方式（见附录 C.1.1），我们可以定义 n 维空间中的体积和表面积。定义

$$B_n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}.$$

这是一个 n 维空间中的单位球体。

第一个反直觉的事实是，随着 n 趋于无穷， B_n 的体积和表面积都会趋于零！首先，我们给出 n 维单位球体的体积和表面积的计算公式：

定理 4.1 记 V_n 为 n 维单位球体的体积， S_n 为 n 维单位球体的表面积。那么，

$$\begin{aligned} V_n &= \frac{\pi^{n/2}}{\Gamma(1+n/2)}, \\ S_n &= \frac{2\pi^{n/2}}{\Gamma(n/2)}. \end{aligned}$$

其中 Γ 是 *Gamma* 函数，对自然数 n ，它定义为

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} (m-1)!, & n = 2m, \\ \frac{(2m)!}{4^m m!} \sqrt{\pi}, & n = 2m+1. \end{cases}$$

这一定理的证明对积分的技巧要求较高，我们这里不给出，感兴趣的读者见习题 1。

这一定理的推论是，随着 n 的增大， V_n 和 S_n 都会趋于零：

推论 4.1

$$\lim_{n \rightarrow \infty} V_n = 0,$$

$$\lim_{n \rightarrow \infty} S_n = 0.$$

证明. 当 n 趋于无穷时, 由 Stirling 公式, 我们有

$$\Gamma(n) \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

因此, 当 $n \rightarrow \infty$,

$$V_n \sim \frac{\pi^{n/2}}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} = \frac{1}{\sqrt{2\pi n}} \left(\frac{e}{n\sqrt{\pi}}\right)^n \rightarrow 0.$$

同理, S_n 的极限也为 0. □

接下来, 我们说明, 高维空间中的球体质量分布是极其不均匀的, 大部分质量都集中在球体的边界上. 定义一个半径为 r 的 n 维球为

$$B_n(r) = \{x \in \mathbb{R}^n : \|x\| \leq r\}.$$

那么我们有:

命题 4.1 对任意 $\epsilon \in (0, 1)$,

$$\frac{\lambda(B_n(1-\epsilon))}{\lambda(B_n(1))} = (1-\epsilon)^n.$$

其中 λ 是 \mathbb{R}^n 上的 Lebesgue 测度 (体积).

因此, 当 n 趋于无穷时, 这一比值以指数速度趋于零.

证明. 利用体积 (Lebesgue 测度) 的性质 (见附录 C.1.1) ,

$$\lambda(B_n(r)) = r^n \lambda(B_n(1)).$$

代入 $r = 1 - \epsilon$, 即可得证. \square

接下来, 我们再说明, 如果把 n 维球和 n 维超立方体放在一起看, 他们的质量分布也是非常反直觉的. 对于 $n \in \mathbb{N}$, $\epsilon > 0$, 定义我们定义一个 n 维的 “球壳” $A_{n,\epsilon}$:

$$A_{n,\epsilon} = \{x \in \mathbb{R}^n : (1 - \epsilon)\sqrt{n/3} < \|x\| < (1 + \epsilon)\sqrt{n/3}\}.$$

这是一个半径为 $\sqrt{n/3}$ 的 n 维 “球壳”, 厚度为 2ϵ .

我们有如下定理:

定理 4.2

$$\lim_{n \rightarrow \infty} \frac{\lambda(A_{n,\epsilon} \cap (-1,1)^n)}{\lambda((-1,1)^n)} = 1,$$

其中 λ 是 \mathbb{R}^n 上的 Lebesgue 测度 (体积) .

我们来解释这一定理反直觉的地方.

- 如果我们只看方向 $(1, 0, \dots, 0)$ (或者其他 “正方向”), 半径为 $\sqrt{n/3}$ 的球壳应该远远超出了 $(-1, 1)$ 的范围. 然而, 这一定理告诉我们, 球壳在超立方体内占据了几乎全部的体积, 这说明在其他的方向, 球体以不可思议的方式被 “压扁” 了.
- 当 n 很大时, 超立方体 $(-1, 1)^n$ 的绝大部分体积都是由一个厚度为 ϵ , 半径为 $\sqrt{n/3}$ 的 n 维球壳提供的, 当 ϵ 很小时, 这个球壳

非常薄. 一层薄球壳占据了一个实心立方体的绝大部分体积, 这个在二维和三维空间中也是难以想象的.

接下来, 我们证明这一定理.

证明. 为了证明这一定理, 我们可以考虑 n 维随机变量的分布.

设 X_1, X_2, \dots, X_n 是独立同分布 (i.i.d.) 的随机变量, 且服从均匀分布 $\mathcal{U}(-1, 1)$. $Z_n = (X_1, X_2, \dots, X_n)$ 服从均匀分布 $\mathcal{U}((-1, 1)^n)$. 对任意集合 $A \subseteq \mathbb{R}^n$, 有

$$\Pr(Z_n \in A) = \frac{\lambda(A \cap (-1, 1)^n)}{\lambda((-1, 1)^n)}.$$

我们来计算 $\Pr(Z_n \in A_{n,\epsilon})$.

$Y_i = X_i^2$ 也是 i.i.d. 的, 并且有

$$\mathbb{E}[Y_i] = \int_{-1}^1 \frac{1}{2} \cdot x^2 dx = \frac{1}{3}.$$

由弱大数定律 (定理 C.29), $\sum_{i=1}^n X_i^2 / n$ 偏离期望 $1/3$ 某个值的概率会随着 n 趋于无穷而趋于零. 更精确来说, 当 $n \rightarrow \infty$ 时,

$$\Pr \left[\left| \frac{\sum_{i=1}^n X_i^2}{n} - \frac{1}{3} \right| > \epsilon \right] \rightarrow 0.$$

变形得

$$\underbrace{\Pr \left[(1 - \epsilon) \sqrt{\frac{n}{3}} \leq \sqrt{\sum_{i=1}^n X_i^2} \leq (1 + \epsilon) \sqrt{\frac{n}{3}} \right]}_{=\Pr[Z_n \in A_{n,\epsilon}]} \rightarrow 1.$$

于是，

$$\frac{\lambda(A_{n,\epsilon} \cap (-1,1)^n)}{\lambda((-1,1)^n)} = \Pr[Z_n \in A_{n,\epsilon}] \rightarrow 1.$$

这就完成了证明. □

注. 关于高维几何的更多讨论，可以参考概率论教材 [Dur19]、[Shi96]，高维概率论的教材 [Ver18]，以及分析学的教材，例如 [Zor16].

§4.1.2 Stein 悖论

接下来，我们转向更加抽象的高维空间，先考虑一维空间. 假设 $X_1 \sim \mathcal{N}(\mu, 1)$ ，但我们并不知道 μ 是什么. 通过随机采样得到了一个样本 $x_1 = 7$ ，怎样合理地估计 μ ? 既然没有多余的信息，我们不妨就猜 $\hat{\mu} = 7$ ，这是一个符合直觉的估计.

然后转向二维空间，假设 $(X_1, X_2) \sim \mathcal{N}(\mu, \mathbf{1}_2)$, $\mu = (\mu_1, \mu_2)$ ，我们还是不知道 μ 是什么. 同样，随机采样得到样本 $x_1 = 7, x_2 = 6$ ，怎样合理地估计 μ ? 我们似乎依然没有多余的选择， $\hat{\mu}_1 = 7, \hat{\mu}_2 = 6$ 看起来也是一个“好的”估计.

现在，转向一般的 n 维空间， $n \geq 3$. 假设

$$(X_1, X_2, \dots, X_n) \sim \mathcal{N}(\mu, \mathbf{1}_n), \quad \mu = (\mu_1, \mu_2, \dots, \mu_n),$$

μ 未知. 随机采样得到样本 x_1, x_2, \dots, x_n ，怎样对 μ 进行估计?

直观上看，这似乎与一维、二维空间的情况并无区别. 我们除了直接取

$$\hat{\mu} = (x_1, x_2, \dots, x_n), \tag{4.1}$$

似乎没有更好的选择，我们把这种估计称为朴素估计。然而，在这部分我们将看到，在高维空间中，存在比朴素估计更好的估计方法！这就是 *Stein 悖论*。

为了衡量一个估计量方法的优劣程度，可以定义损失函数（均方误差）：

$$\ell = \mathbb{E} [\|\hat{\mu} - \mu\|^2].$$

均方误差越小，我们认为估计量越好。

然而，好坏这件事情似乎并没有那么简单，在一维的情况下，考虑如下估计量：

1. 一个估计量 $\hat{\mu}_1$ 是令 μ 等于得到的样本点，那么

$$\mathbb{E} [\|\hat{\mu}_1 - \mu\|^2] = \mathbb{E} [(x - \mu)^2] = \text{Var}[x] = 1.$$

2. 另一个估计量 $\hat{\mu}_2$ 是令 μ 等于一个固定的值，比如 $\mu = 7$ ，那么

$$\mathbb{E} [\|\hat{\mu}_2 - \mu\|^2] = \mathbb{E} [(7 - \mu)^2] = (7 - \mu)^2.$$

我们不能明确说明哪一种估计量更好，因为如果 μ 在 7 附近，第二种方法会更好；但是如果 μ 在 0 附近，第一种方法会更好。

上面的例子表明，如果一个模型的参数 μ ，我们很可能无法判断哪一种方法更好。但有一种情况，我们是可以明确说明一个方法 A 一定不好：有另外一个估计量在任何 μ 下都比它好。这就是如下定义：

定义 4.1 (可接受性) 考虑对参数 μ 的估计量方法 A ，如果存在估计量

方法 B , 在任意的 μ 下都成立

$$\ell_B > \ell_A,$$

我们就称 A 方法是不可接受的. 否则, 我们称 A 方法是可接受的. \square

有了评判估计量好坏的标准, 我们就可以来看 Stein 悖论了。首先, 对朴素估计来说, 我们可以计算它的均方误差:

引理 4.1 对于 $n \geq 3$, 朴素估计 (4.1) 的均方误差为

$$\mathbb{E} [\|\hat{\mu} - \mu\|^2] = n.$$

证明. 由于 $x_i - \mu_i$ 是独立同分布的标准正态分布, 我们有

$$\mathbb{E} [\|\hat{\mu} - \mu\|^2] = \sum_{i=1}^n \mathbb{E} [(x_i - \mu_i)^2] = n. \quad \square$$

接下来, 我们给出 James-Stein 估计量, 它是一个比朴素估计更好的估计量.

定义 4.2 (James-Stein 估计量) 假设采样得到数据点 x_1, x_2, \dots, x_n , 对参数

$$\mu = (\mu_1, \mu_2, \dots, \mu_n)^\top,$$

定义 **James-Stein 估计量**为：

$$\hat{\mu} = \left(1 - \frac{n-2}{x_1^2 + x_2^2 + \dots + x_n^2}\right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \quad (4.2)$$

□

定理 4.3 (Stein 悖论) 对于 $n \geq 3$, 朴素估计 (4.1) 是不可接受的, 具体来说, James-Stein 估计量 (4.2) 的均方误差为

$$\mathbb{E} [\|\hat{\mu} - \mu\|^2] = n - \mathbb{E} \left[\frac{(n-2)^2}{n-2+2K} \right],$$

其中 K 服从参数为 $\|mu\|^2/2$ 的 Poisson 分布。因此, 在任意 μ 下, James-Stein 估计量都比朴素估计更好。

这一定理的证明十分具有技巧性, 我们这里不给出证明, 感兴趣的读者见习题 2.

比起证明这个定理, 更重要的问题是, 为什么 James-Stein 估计量会比朴素估计更好? 答案正是在于高维空间的反直觉性。

如图 4.1 所示, 坐标轴上有一个圆心为 c 的单位圆, 圆内均匀随机选取一个点 x , 那么

$$\Pr(\|x\| > \|c\|) > \frac{1}{2}.$$

如果不是二维的圆而是高维的球, 这一不等式依然成立, 并且这种效应随维数的增加而变强 (见习题 3)。另一方面, 圆心 c 离中心越远, 这一概率越接近 $1/2$.

现在, 我们回到最早的估计问题, 我们把 μ 看作是圆心 c , 而随机

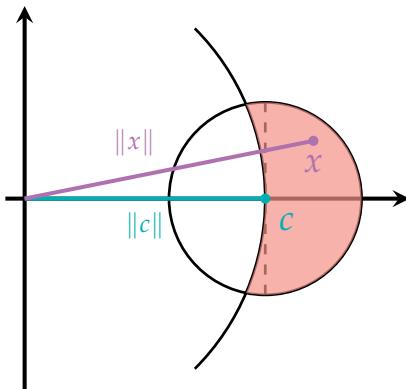


图 4.1: 高维空间中的采样点

采样的点就看作是样本点. 上面的概率不等式意味着, 随着维数变高, 朴素估计量 x 与真正的 μ 之间的差距会越来越大. 不仅如此, 朴素估计量对 μ 的估计会偏大.

直觉上, 要想更加精确估计 μ , 我们需要比样本点 x 更接近圆心 c . 仔细观察 (4.2), James-Stein 估计量的确是这样做的. 下面, 我们详细介绍这一估计量的几何推导.

由于坐标系的选取是随意的, 可以设一条坐标轴和 μ 的方向相同, 其他 $n - 2$ 根坐标轴方向随意, 但正交. 在新坐标系下,

$$\mu = (\|\mu\|, 0, \dots, 0)^\top.$$

设样本是 $x = (x_1, x_2, \dots, x_n)^\top$. 损失函数可以被分解成两个部分的和:

$$\ell = (x_1 - \|\mu\|)^2 + \sum_{i=2}^n x_i^2. \quad (4.3)$$

令

$$\rho = \sqrt{\sum_{i=2}^n x_i^2}.$$

因为 x 是分量相互独立的 Gauss 向量，坐标轴旋转不改变分量之间的独立性，因此 ρ 服从自由度为 $n - 1$ 的 χ 分布（可参见附录 C.4）。

假设样本点恰好满足 $x_1 = \|\mu\|$ ，而 x_i 以概率 1 都不为 0，可以画出图 4.2。

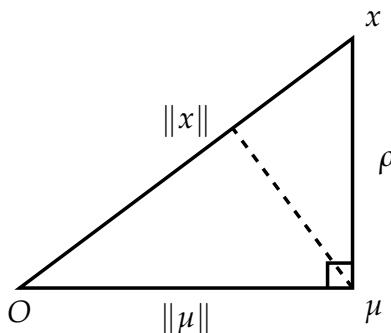


图 4.2: x 与 μ 形成的直角三角形

如果我们直接用样本点 x 作为估计量 $\hat{\mu}$ ，那么我们产生的偏差在于直边上，因此会有 ρ^2 的损失。现在，我们尝试移动 $\hat{\mu}$ ，来减少损失。

我们除了样本点 x 和原点 O 之外，其他任何信息都没有。因此，盲目的移动反而会带来更大的损失。结合前面的讨论，我们想要将 $\hat{\mu}$ 靠近原点，因此，一个合理的办法就是沿着斜边 Ox 向原点移动。

根据直角三角形的性质，当 $\hat{\mu} - \mu$ 与斜边垂直的时候，损失最小。我们来推导 $\hat{\mu}$ 的表达式。设 $\hat{\mu} = \alpha x$ 。根据三角形相似的原理，我们有

$$\frac{|O\hat{\mu}|}{|O\mu|} = \frac{|O\mu|}{|Ox|},$$

$$\begin{aligned} &\iff \frac{\alpha \|x\|}{\|\mu\|} = \frac{\|\mu\|}{\|x\|} \\ &\iff \alpha = \frac{\|\mu\|^2}{\|x\|^2} = 1 - \frac{\rho^2}{\|x\|^2}. \end{aligned}$$

因此，新的估计量是

$$\hat{\mu} = \left(1 - \frac{\rho^2}{\|x\|^2}\right)x = \left(1 - \frac{\rho^2}{x_1^2 + x_2^2 + \dots + x_n^2}\right)x.$$

然而，这一新估计量是不可计算的：因为我们只知道 O 和旋转之前的 x ，所以我们没有办法计算 ρ . 为了得到 James-Stein 估计量，我们用一些数字特征来代替 ρ . 对于自由度为 $k > 1$ 的 χ 分布，其众数是 $\sqrt{k-1}$ (见习题 4). 用众数来代替 ρ ，就得到了 James-Stein 估计量.

最后，我们给出一些关于 Stein 悖论的讨论.

- 存在比 James-Stein 估计量更好的估计量. 直观上，当样本 x 过于靠近原点的时候， $\|x\|$ 接近零，因此 James-Stein 估计量会穿过原点，往反方向跑到很远地方. 这自然会带来很大的损失. 因此，修正这一行为可以得到更好的估计量，比如

$$\hat{\mu} = \text{ReLU} \left(1 - \frac{n-2}{x_1^2 + x_2^2 + \dots + x_n^2}\right)x,$$

其中

$$\text{ReLU}(x) = \begin{cases} x, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

- 我们也可以从偏差-方差权衡的角度来理解 James-Stein 估计量.

“距离”这一概念，在代数上，可以做如下分解：

$$\begin{aligned}\mathbb{E} [(\hat{\mu} - \mu)^2] &= \text{Var} [\hat{\mu} - \mu] + (\mathbb{E} [\hat{\mu} - \mu])^2 \\ &= \text{Var} [\hat{\mu}] + (\mathbb{E} [\hat{\mu} - \mu])^2.\end{aligned}$$

前半部分是方差（距离差的平方的期望），后半部分是偏差（距离差的期望的平方）。朴素估计是无偏差估计，但会引入很大的方差。通过适当引入偏差可能会减小方差，从而减小总体的预测误差，这就是 James-Stein 估计量的原理。

不过，如果你仔细观察会发现，实际上 (4.3) 的第一项就是方差，第二项是偏差。因而，偏差-方差权衡（代数直观）和几何直观其实在说同一件事情。

- 为什么 James-Stein 估计量在一维和二维空间中不起效？Brown 证明了如下的定理： n 维空间中的朴素估计量是可接受的当且仅当 n 维空间中的简单对称随机游走以概率 1 无限次返回原点 [Bro71]。这一惊人的联系揭示了这一问题的答案。

一维和二维空间中的随机游走都会以概率 1 无限次返回原点，因此朴素估计量是可接受的，所以我们不可能找到更好的估计量。而更高维的空间中，随机游走以概率 1 无限次返回原点的概率是 0，因此朴素估计量是不可接受的，所以 James-Stein 估计量就起效了！

- Stein 悖论不意味着“中国茶叶的价格可以帮助预测墨尔本的降雨概率”。尽管我们总是可以把毫不相关的随机事件捆绑在一起，然后利用 James-Stein 估计量减少总体的预测误差，但是这不意味

着对其中任何一个事件的预测会更准确.

§4.1.3 为什么我们要正则化？远有潜龙，勿用

在前两节中，我们用了一些简单、理想化的模型说明了高维空间的一些奇异性. 尽管真实的机器学习问题远比这些讨论要复杂，但他们所带来的启示不容忽视.

在机器学习中，我们可以把问题都归结为参数估计问题. 虽然参数空间的“原点”看起来并没有什么不同，但原点这一概念本身确实带着人类先验的知识. 例如，如果一个神经网络的所有参数都是零，无论输入是什么，它都会输出全零. 同样，在 Stein 悖论中，靠近原点就是会产生更好的估计量.

在机器学习中，我们同样有偏好“原点”的倾向，例如使用 L^2 正则化这样的技术可以使模型变得更容易，因此不太可能过拟合. 本节的内容通过极端的例子展示了正则化背后的原理：在高维空间中，离原点较远的地方体积远大于靠近原点的地方. 因此，在高维空间中，向原点收缩一点就能减少大量的参数空间.

换句话说，对于一个大型机器学习模型来说，过拟合的方式远多于欠拟合的方式，所以我们倾向于让模型更偏向于欠拟合：欠拟合只会带来少部分问题，而过拟合带来的是数不胜数、千奇百怪的问题.

模型越远离原点，它的行为就越难以控制和解释. 高维几何与 Stein 悖论给我们的启示是，远离原点就会有危险，而在高维空间中，稍微远离原点就会引入大量危险. 化用《周易》的一句话：

“远有潜龙，勿用.”¹

¹原句出自《周易·乾卦》，“初九：潜龙勿用.”孔子对这句话的解释是如果身居下位，时机还没有成熟，应当像潜藏的龙一样不要施展你的才干. 这里，复杂的、远离原点模型

注. Stein 悖论的原始论文是 [Ste56]，后来，James 和 Stein 对估计量进行了简化 [JS61]，也就是本书中所讲的这个。然而，关于 Stein 悖论的讨论在各种领域都有，比如统计学和机器学习等。本文采取的几何视角，思路来自论文 [BZ12]，与正则化的关系则来自博客 [Und]。

§4.2 集中不等式

我们在前一节中阐述了高维空间中怪诞反直觉的性质。从本节开始，我们将阐述高维空间中随机变量的另一重属性：集中不等式。集中不等式说明的是，尽管整个空间非常庞大、难以理解，但如果随机变量具有某些性质，那么它们的取值就会集中在某个非常小的区域内，因而并没有我们所设想的那么复杂。利用这一原理，我们可以将非常高维的数据压缩到一个较为低维的空间中，从而可以驾驭他们。

接下来，我们先做一些准备工作，更详细的讨论参见附录 C。我们先引入示性函数的概念。

定义 4.3 (示性函数) 对事件 A ，定义 A 的示性函数为一个从样本空间 Ω 到 \mathbb{R} 的随机变量：

$$I(A)(\omega) := \begin{cases} 1, & \omega \in A. \\ 0, & \omega \notin A. \end{cases} \quad \square$$

从定义就可以得到如下基本性质：

命题 4.2 设 A, B 是两个事件，则

就像是潜龙，隐藏着巨大的力量，但是现在人类对他们的理解还远远不够，因此我们应当保持谨慎，不要轻易使用。

1. $I(AB) = I(A)I(B).$
2. $I(A)^2 = I(A).$
3. $I(A \cup B) = I(A) + I(B) - I(AB).$

证明. 这里只作为一个示意, 证明第三点, 其他都类似. 我们需要证明, 对任意样本点 $\omega \in \Omega$, 我们有

$$I(A \cup B)(\omega) = I(A)(\omega) + I(B)(\omega) - I(AB)(\omega).$$

假设 $\omega \in A \cup B$, 那么左边等于 1. 我们分类讨论:

- 如果 $\omega \in A$, 那么右边第一项为 1.
 - 如果 $\omega \in B$, 那么右边第二项为 1. 此时自然也有 $\omega \in AB$, 所以右边第三项为 1, 因此右边等于 1, 等于左边.
 - 如果 $\omega \notin B$, 那么右边第二项为 0. 此时自然也有 $\omega \notin AB$, 所以右边第三项为 0, 因此右边等于 1, 等于左边.
- 如果 $\omega \notin A$, 那么右边第一项为 0. 此时必须有 $\omega \in B$, 所以右边第二项为 1. 但是此时自然也有 $\omega \notin AB$, 所以右边第三项为 0, 因此右边等于 1, 等于左边.

如果 $\omega \notin A \cup B$, 讨论类似, 这里不再赘述. \square

示性函数之所以重要, 是因为它联系了期望与概率. 我们先来看一个显然的命题:

命题 4.3 设 A 是一个事件, 则

$$\mathbb{E}[I(A)] = \Pr(A).$$

示性函数可以把对概率的计算变成对期望的计算. 回忆期望的线性性 (见命题 C.10): 设 $a, b \in \mathbb{R}$, X, Y 是有期望的随机变量, 那么成立

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

利用期望的线性性, 示性函数可以导出很多概率恒等式与不等式. 例如:
容斥公式

$$\begin{aligned}\Pr(A \cup B) &= \mathbb{E}[I(A \cup B)] = \mathbb{E}[I(A) + I(B) - I(AB)] \\ &= \mathbb{E}[I(A)] + \mathbb{E}[I(B)] - \mathbb{E}[I(AB)] \\ &= \Pr(A) + \Pr(B) - \Pr(AB).\end{aligned}$$

对于概率论以及机器学习理论来说, 下面的这个不等式非常重要:

定理 4.4 (Markov 不等式) 如果 X 是非负有期望的随机变量, $a > 0$, 那么

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

证明. 直接利用示性函数, 我们有:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[XI(X \geq a) + XI(X < a)] \\ &= \underbrace{\mathbb{E}[XI(X \geq a)]}_{\geq aI(X \geq a)} + \underbrace{\mathbb{E}[XI(X < a)]}_{\geq 0} \\ &\geq a\mathbb{E}[I(X \geq a)] = a\Pr(X \geq a).\end{aligned} \quad \square$$

注意, 为了使得证明有效, 我们必须要假设上面的推导中出现的期望都是存在的, 当然这实际上很容易验证. 为了避免不必要的技术细节,

在后面的所有证明以及推导中，我们都会默认写出来的期望是存在的，不再赘述。

我们利用 Markov 不等式可以直接得到以下结果。

推论 4.2 (Chebyshev 不等式) 设 X 是任意有方差的随机变量，那么对任意 $a > 0$ ，成立

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

证明. 设 $Y = (X - \mathbb{E}[X])^2$, $t = a^2$, 那么 Y 是非负随机变量，且 $\mathbb{E}[Y] = \text{Var}(X)$, 于是由 Markov 不等式，我们有

$$\begin{aligned} \Pr(|X - \mathbb{E}[X]| \geq a) &= \Pr(|X - \mathbb{E}[X]|^2 \geq a^2) \\ &= \Pr(Y \geq t) \\ &\leq \frac{\mathbb{E}[Y]}{t} = \frac{\text{Var}(X)}{a^2}. \end{aligned} \quad \square$$

Chebyshev 不等式告诉我们采样到偏离其期望的概率有一个上界。像这样利用矩（即 $\mathbb{E}[f(X)]$ ）来估计概率上界的方法被称为矩法。

实际上，很多情况下，偏离期望是非常小概率的事件，远小于上面的估计值。为了得到更精确的上界，我们需要一些技巧。考虑任意随机变量 X ，对 $\lambda > 0$,

$$X \geq a \iff \lambda X \geq \lambda a \iff e^{\lambda X} \geq e^{\lambda a}.$$

由 Markov 不等式（如何得到?），

$$\Pr(X \geq a) = \Pr\left(e^{\lambda X} \geq e^{\lambda a}\right) \leq e^{-\lambda a} \cdot \mathbb{E}\left[e^{\lambda X}\right].$$

注意到这个不等式应该对任意 $\lambda > 0$ 成立，所以

$$\Pr(X \geq a) \leq \inf_{\lambda > 0} e^{-\lambda a} \cdot \mathbb{E}[e^{\lambda X}].$$

以上方法可以得到概率更精确的上界。这样用指数进行推导的方法称为指数矩或 *Cramér-Chernoff* 方法。

利用指数矩，我们可以更加精确地研究 Chebyshev 不等式中随机变量所表现出来的性质，这种性质被称为概率的集中性。我们可以用集中不等式来刻画这样的性质。这样的不等式描述随机变量 X 有多大概率偏离某个值 μ 多少值 (t)，它表现为

$$\Pr(|X - \mu| \geq t) \leq \text{小量}.$$

通常来说， μ 是随机变量的期望或者中位数，在这本书中，只会讨论关于期望的集中性。我们可以看到 Chebyshev 不等式就是一种特殊的集中不等式，但是它的界太松。利用指数矩，我们将证明更紧的 Hoeffding 不等式和 Chernoff 不等式。

定理 4.5 (Hoeffding 不等式) 设随机变量 X_1, \dots, X_n 相互独立且服从对称 Bernoulli 分布，即 X_i 满足 $\Pr(X_i = 1) = 1 - \Pr(X_i = -1) = 1/2$ 。考虑向量 $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ ，对任意 $t \geq 0$ ，我们有

$$\Pr\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

证明. 由指数矩, 我们有

$$\begin{aligned}\Pr\left(\sum_{i=1}^n a_i X_i \geq t\right) &= \Pr\left(\exp\left(\lambda \sum_{i=1}^n a_i X_i\right) \geq \exp(\lambda t)\right) \\ &\leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n a_i X_i\right)\right] \\ &= e^{-\lambda t} \prod_{i=1}^n \mathbb{E}[\exp(\lambda a_i X_i)].\end{aligned}$$

这个不等式对任意 $\lambda > 0$ 都成立.

利用 X_1, \dots, X_n 服从对称 Bernoulli 分布, 得到 (习题 5):

$$e^{-\lambda t} \prod_i \mathbb{E}[\exp(\lambda a_i X_i)] \leq \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i a_i^2\right). \quad (4.4)$$

由于这一不等式对任意 $\lambda > 0$ 都成立, 根据二次函数的性质, 取 $\lambda = t / \sum_i a_i^2$, 可得

$$\begin{aligned}\inf_{\lambda>0} \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i a_i^2\right) &= \exp\left(-\frac{t}{\sum_i a_i^2} t + \frac{1}{2} \left(\frac{t}{\sum_i a_i^2}\right)^2 \sum_i a_i^2\right) \\ &= \exp\left(-\frac{t^2}{2 \|a\|_2^2}\right).\end{aligned}$$

□

利用相同的证明技巧, 我们可以证明一般形式的 Hoeffding 不等式 (见习题 6).

定理 4.6 (Hoeffding 不等式, 一般情形) 设 X_1, \dots, X_n 是相互独立的随

机变量，对任意 i 都成立 $X_i \in [m_i, M_i]$. 那么对任意 $t \geq 0$, 我们有

$$\Pr \left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \right) \leq \exp \left(-\frac{2t^2}{\sum_{i=1}^n (M_i - m_i)^2} \right).$$

下面我们介绍 Chernoff 不等式.

定理 4.7 (Chernoff 不等式) 设 X_1, \dots, X_n 是相互独立的随机变量，分别服从于参数为 p_1, \dots, p_n 的 Bernoulli 分布. 记 $\sum_{i=1}^n X_i$ 的期望为 $\mu = \sum_{i=1}^n p_i$, 对于任意 $t > \mu$, 我们有

$$\Pr \left(\sum_{i=1}^n X_i \geq t \right) \leq e^{-\mu} \left(\frac{e^\mu}{t} \right)^t.$$

这里 e 是自然对数的底数.

证明. 和证明 Hoeffding 不等式的第一步相同，我们先利用指数矩，对任意 $\lambda > 0$ 有

$$\Pr \left(\sum_{i=1}^n X_i \geq t \right) \leq e^{-\lambda t} \prod_{i=1}^n \mathbb{E} [\exp(\lambda X_i)].$$

然后，将 $\prod_{i=1}^n \mathbb{E} [\exp(\lambda X_i)]$ 进一步放缩：

$$\begin{aligned} \prod_{i=1}^n \mathbb{E} [\exp(\lambda X_i)] &= \prod_{i=1}^n (e^\lambda p_i + (1 - p_i)) \\ &\leq \prod_{i=1}^n \exp ((e^\lambda - 1)p_i). \end{aligned}$$

因此

$$\begin{aligned}\Pr\left(\sum_{i=1}^n X_i \geq t\right) &\leq e^{-\lambda t} \prod_{i=1}^n \exp\left((e^\lambda - 1)p_i\right) \\ &= e^{-\lambda t} \exp\left((e^\lambda - 1) \sum_{i=1}^n p_i\right) \\ &= \exp\left(\mu e^\lambda - t\lambda - \mu\right).\end{aligned}$$

右边的最小值在 $\lambda = \log(t/\mu)$ 取得，代入得到：

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t.$$

□

注. 关于集中不等式的系统性论述，请参见高维概率论教材 [Ver18].

§4.3 J-L 引理的陈述与证明

有了上面矩法的准备，我们可以陈述并证明 J-L 引理了。

定理 4.8 (Johnson-Lindenstrauss 引理) 给定 N 个 \mathbb{R}^m 中的单位向量 v_1, \dots, v_N 和自然数 $n > 24 \log N / \epsilon^2$ ，随机矩阵 $A \in \mathbb{R}^{n \times m}$ 每个元素独立重复采样自 $\mathcal{N}(0, 1/n)$ ， $\epsilon \in (0, 1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i \neq j$ ，都成立

$$(1-\epsilon) \|v_i - v_j\|_2^2 < \|Av_i - Av_j\|_2^2 < (1+\epsilon) \|v_i - v_j\|_2^2.$$

我们可以把 n 理解成降维后的维度， Av_i 是降维后的向量。这个引理告诉我们只要 $n > 24 \log N / \epsilon^2$ ，我们就可以用变换 A 把原本 m 维的

向量映射到 n 维空间，并且保证它们相对距离的偏离不超过 ϵ .

通常来说，相对距离编码了很多重要的信息. 例如，如果两个人的年龄、身高、体重等属性相差很小，那么他们也应该更相似. 在这种观点下，我们可以把 A 看成一个损失率很低的压缩变换. 不严格地说，

塞下 N 个向量，只需要 $\mathcal{O}(\log N)$ 维空间.

值得注意的是，J-L 引理中压缩空间的维数并不依赖于原始空间的维数，而只依赖于数据的数量. 因此，这对于一些抽象空间中数据的降维是非常有用的，见第 4.4 节.

下面我们开始证明 J-L 引理. 为了看出来证明的思路，我们第一个任务是算出压缩后 Av_i 的分布. 我们首先回忆一些正态向量的基本性质（参考附录 C.4）.

命题 4.4 假设 $u \sim \mathcal{N}(\mu, \Sigma)$ 是一个 n 维正态向量， M 是一个 $m \times n$ 矩阵，那么 Mu 是一个 m 维正态向量，并且 $Au \sim \mathcal{N}(M\mu, M\Sigma M^T)$.

利用这一个命题，很容易可以得到 Av_i 的分布：

引理 4.2 假设 $u \in \mathbb{R}^m$ 是一个单位向量，那么 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

证明. 将 A 视作一个 mn 维的正态向量，注意到， $(Au)_i = \sum_{j=1}^m A_{ij}u_j$ ，所以 Au 是一个从向量 A 线性变换得到的向量. 根据命题 4.4， Au 是一个正态向量，只需计算它的期望和协方差矩阵.

注意到，对不同的 i ，向量 $(A_{ij})_j$ 相互是独立的，所以分量 $(Au)_i$ 相互也是独立的，因此只需要计算正态变量 $(Au)_i$ 的期望与方差. 其期望为 $\sum_{j=1}^m 0 \cdot u_j = 0$ ，方差为

$$\sum_{j=1}^m \left(\frac{1}{n} \cdot u_j^2 \right) = \frac{1}{n}.$$

所以 Au 的期望是 0, 协方差矩阵是 $n^{-1}I_n$. \square

然而, 我们关心的其实不单单是 Av_i 的分布, 更重要的其实是 $Av_i - Av_j$ 的分布, 即压缩后的向量之间的相对距离. 不过, 我们并不需要做额外的什么计算, 我们直接有如下结果:

引理 4.3 向量 $u = \frac{v_i - v_j}{\|v_i - v_j\|_2}$ 是一个单位向量, 因此 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

J-L 引理实际上在说, $\|Au\|_2$ 偏离 1 的一定程度的概率是非常小的. 于是, 为了证明 J-L 引理, 我们最重要的任务是给出 Au 这样向量模长的集中不等式:

引理 4.4 (单位模引理) 设 $u \sim \mathcal{N}(0, n^{-1}I_n)$, $\epsilon \in (0, 1)$ 是给定的常数, 那么我们有

$$\Pr(|\|u\|_2^2 - 1| \geq \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2 n}{8}\right).$$

注意到 $\mathbb{E}[\|u\|_2^2] = n \cdot (1/n) = 1$, 所以这个引理在说高维空间中, 如果正态向量具有单位模长平方期望, 那么它的模长就会集中在单位长度附近, 因此称为单位模引理.

证明. $|\|u\|_2^2 - 1| \geq \epsilon$ 发生有两种可能, $\|u\|_2^2 - 1 \geq \epsilon$ 和 $1 - \|u\|_2^2 \geq \epsilon$. 我们先来计算 $\|u\|_2^2 - 1 \geq \epsilon$ 的概率, 根据指数矩,

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq \inf_{\lambda > 0} \left\{ e^{-\lambda(\epsilon+1)} \mathbb{E} \left[e^{\lambda \|u\|_2^2} \right] \right\}.$$

因为 u 的各个分量是相互独立的, 所以我们可以把 $\|u\|_2^2$ 展开

$$\mathbb{E} \left[e^{\lambda \|u\|_2^2} \right] = \mathbb{E} \left[e^{\lambda \sum_i u_i^2} \right] = \mathbb{E} \left[\prod_i e^{\lambda u_i^2} \right] = \prod_i \mathbb{E} \left[e^{\lambda u_i^2} \right].$$

可以算得 $\mathbb{E} [e^{\lambda u_i^2}] = \sqrt{n/(n-2\lambda)}$ (见习题 7)，所以

$$\Pr \left(\|u\|_2^2 - 1 \geq \epsilon \right) \leq \inf_{\lambda > 0} \left\{ e^{-\lambda(\epsilon+1)} \left(\frac{n}{n-2\lambda} \right)^{n/2} \right\}.$$

可以验证最小值在 $\lambda = n\epsilon/(2(1+\epsilon))$ 处取到，代入可得

$$\Pr \left(\|u\|_2^2 - 1 \geq \epsilon \right) \leq e^{n(\log(1+\epsilon)-\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

这里最后一个不等号使用了不等式 $\log(1+\epsilon) \leq \epsilon - \epsilon^2/4$.

计算 $1 - \|u\|_2^2 \geq \epsilon$ 的概率的过程和 $\|u\|_2^2 - 1 \geq \epsilon$ 几乎完全相同的，可以得到

$$\Pr \left(1 - \|u\|_2^2 \geq \epsilon \right) \leq e^{n(\log(1-\epsilon)+\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

$$\begin{aligned} \Pr \left(|\|u\|_2^2 - 1| \geq \epsilon \right) &\leq \Pr \left(\|u\|_2^2 - 1 \geq \epsilon \right) + \Pr \left(1 - \|u\|_2^2 \geq \epsilon \right) \\ &\leq 2e^{-n\epsilon^2/8}. \end{aligned}$$
□

有了单位模引理，我们就可以很容易证明 J-L 引理了。将引理 4.3 中的 u 带入单位模引理，得到

$$\Pr \left(\left| \left\| \frac{A(v_i - v_j)}{\|v_i - v_j\|_2} \right\|_2^2 - 1 \right| \geq \epsilon \right) \leq 2 \exp \left(-\frac{\epsilon^2 n}{8} \right).$$

这个结论对任意 $i \neq j$ 成立，因此遍历所有 i, j 对，可得

$$\begin{aligned} \Pr \left(\exists (i, j) : \left| \left\| \frac{A(v_i - v_j)}{\|v_i - v_j\|_2} \right\|_2^2 - 1 \right| \geq \epsilon \right) &\leq 2 \sum_{i \neq j} \exp \left(-\frac{\epsilon^2 n}{8} \right) \\ &= 2 \binom{N}{2} \exp \left(-\frac{\epsilon^2 n}{8} \right). \end{aligned}$$

换言之，对任意 i, j ， $\left| \left\| \frac{A(v_i - v_j)}{\|v_i - v_j\|_2} \right\|_2^2 - 1 \right| < \epsilon$ 都成立的概率不小于

$$1 - 2 \binom{N}{2} \exp \left(-\frac{\epsilon^2 n}{8} \right) = 1 - N(N-1) \exp \left(-\frac{\epsilon^2 n}{8} \right).$$

代入 $n > \frac{24 \log N}{\epsilon^2}$ ，可得这一概率

$$1 - N(N-1) \exp \left(-\frac{\epsilon^2 n}{8} \right) \geq 1 - N(N-1)N^{-3} \geq 1 - N^{-1} = \frac{N-1}{N}.$$

很多时候，我们关心的并不是向量间的距离，而是向量的内积（比如使用余弦度量的时候），这时候我们可以使用内积版本的 J-L 的引理：

定理 4.9 (J-L 引理，内积形式) 给定 N 个单位向量 $v_1, \dots, v_N \in \mathbb{R}^m$ 和 $n > 24 \log N / \epsilon^2$ ，随机矩阵 $A \in \mathbb{R}^{n \times m}$ 每一个元素都独立重复采样自 $\mathcal{N}(0, 1/n)$ ， $\epsilon \in (0, 1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i \neq j$ ，都成立

$$|\langle Av_i, Av_j \rangle - \langle v_i, v_j \rangle| < \epsilon.$$

证明. 由原始 J-L 引理可知, 至少有 $\frac{N-1}{N}$ 的概率满足对于任意 $i \neq j$ 有:

$$(1 - \epsilon) \|v_i - v_j\|_2^2 < \|Av_i - Av_j\|_2^2 < (1 + \epsilon) \|v_i - v_j\|_2^2,$$

$$(1 - \epsilon) \|v_i + v_j\|_2^2 < \|Av_i + Av_j\|_2^2 < (1 + \epsilon) \|v_i + v_j\|_2^2.$$

我们将第一行乘 -1 加到第二行可以得到

$$4 \langle v_i, v_j \rangle - 2\epsilon(\|v_i\|_2^2 + \|v_j\|_2^2) < 4 \langle Av_i, Av_j \rangle < 4 \langle v_i, v_j \rangle + 2\epsilon(\|v_i\|_2^2 + \|v_j\|_2^2).$$

因为 v_i, v_j 是单位向量, 所以上式等价于 $|\langle Av_i, Av_j \rangle - \langle v_i, v_j \rangle| < \epsilon$. \square

§4.4 J-L 引理的应用

J-L 引理描述的是对于 N 个向量, 我们可以将它们降到 $\mathcal{O}(\log N)$ 维空间, 并将相对距离 (或内积) 的误差控制在一定范围内. 它的内容本身就和降维相关, 所以最基本的应用就是直接作为降维方法. 下面我们将介绍两个具体的应用案例说明 J-L 引理如何指导我们在深度学习中选择合适的维度.

例 4.1 (词向量维度) 在为自然语言建立深度学习模型的时候, 我们面对的首要问题就是如何在计算机中表示语言. 我们以中文为例. 如果我们看一句话, 中文组成的基本单元是词. 比如, 这句话就可以被分解为如下成分:

比如 / , / 这句话 / 就 / 可以 / 被 / 分解 / 为 / 如下 / 成分

任何的中文句子都可以被这样分解, 变成一个词的序列. 这一过程被称为分词. 于是, 为了表示一段中文, 我们只需要能够表示所有的词.

于是，一个基本的任务就是如何表示一个词。词向量就是这样的一个表示方法。它的想法很简单：我们用一个向量空间来表示所有的词，每个词对应一个向量，我们用 $v(w)$ 来表示词 w 对应的向量。

很多时候，这些向量之间的关系可以反映出词之间的语义关系。例如，“男人”和“女人”之间差异应该和“国王”和“女王”之间差异是类似的。也就是说，我们希望

$$v(\text{男人}) - v(\text{女人}) \approx v(\text{国王}) - v(\text{女王}).$$

这里， \approx 表示两个向量之间比较相似。具体来说，我们通常会用向量的内积来衡量相似度。也就是说，给定一个词 w ，我们希望

$$\langle v(w), v(\text{男人}) - v(\text{女人}) \rangle \approx \langle v(w), v(\text{国王}) - v(\text{女王}) \rangle.$$

我们可以想象，在真实的世界中，所有的词也构成了一个抽象向量空间，以类似的方式来表示词之间的关系，但这个向量空间的维数我们完全不得而知。对于计算机来说，我们需要选择一个确定的、压缩过的空间来表示词向量。

为了表示这种相似性，在压缩过的词向量空间中，所有词之间的内积应该尽量保持词本身的相似度。这正是内积形式的 J-L 引理（定理 4.9）所描述的情况。于是，J-L 引理告诉我们， $\mathcal{O}(\log N)$ 维空间足以容纳下 N 个单词，还保持了单词之间的相似性。

到此时，我们要十分警惕“理论指导实践”这一表述的理解。J-L 引理（理论）所指导的结论，成立的前提是“正态随机矩阵”，然而我们完全不清楚单词的空间是否符合这一条件。所以，我们不能说 J-L 引理直接给了我们合适的词向量维度，我们只能说 J-L 引理给了词向量选择

的一个直觉。具体应该用多少维度的词向量，还需要实验来验证。□

例 4.2 (多头注意力) 注意力机制是现代深度学习架构中最核心的模块之一。注意力机制的一种理解方式是将其看作一个键-值存储。想象我们把数据都存储到了一个数据库中，它的存储方式是 $\{k_i : v_i\}$ ，其中 k_i 是键， v_i 是值，例如“性别：男”就是一种典型的键值对，其中“性别”是键，而“男”是值。

现在，假设数据库中的所有键、值对是 $\{k_i : v_i\}_{i=1}^N$ ，我们希望从中找到与某个查询 q 最接近的键对应的值。在深度学习中，键、值、查询都可以按照例 4.1 的方式表示成向量。类似地，我们可以用内积来衡量键和查询的相似度。

假设这些查询和键的向量都处在 \mathbb{R}^d 中，那么注意力的计算公式为

$$a_j = \frac{e^{\langle q, k_j \rangle}}{\sum_{j=1}^N e^{\langle q, k_j \rangle}}.$$

换言之，我们把相似性转化为了概率分布，相似度越高的键，被选中的概率越大。我们把 d 称为注意力头大小，在很多深度学习框架中，它被记为 `head_size`。

在很多场景下，一个数据库只能查询一种键-值对的关系，对于一些复杂的问题，我们可能需要查询多种不同的键-值对。例如，对于中文来说，两个词的意思是否相近可以形成一个数据库，而两个词的词性是否相近又可以形成另一个数据库。因此，我们需要多个注意力机制来查询不同的数据库，这就是多头注意力。

在简化的场景下，我们可以假设所有数据库里的 $k_i : v_i$ 对都是一致

的，只是查询 q_i 不同。那么，多头注意力就是要计算

$$a_{ij} = \frac{e^{\langle q_i, k_j \rangle}}{\sum_{j=1}^N e^{\langle q_i, k_j \rangle}}.$$

类似例 4.1 的问题，如果真实世界中 a_{ij} 是 p_{ij} ，我们应该如何选择向量维数 d 才能保证 a_{ij} 能够足够好地逼近 p_{ij} 呢？这个问题和例 4.1 是一样的。

在这个例子中，词向量的维度变成了 d ，词表大小变成了数据库中的键值对数量 N 。J-L 引理告诉我们的答案依然是只需要 $\mathcal{O}(\log N)$ 的空间就足以容纳下 N 个键值对，还能保持多组查询与键之间的相似性。

更为重要的是，这个压缩空间的维度 d 和查询的数量无关。这说明，如果有同样多的参数，头很大的单头注意力机制并不如头很小的多头注意力机制。此外，这也说明无论多少个头，多头注意力的 d 并不需要随着头的数量增加而显著增加。

同样地，J-L 引理只是给了我们一个直觉，具体的维度选择还需要实验来验证。□

注。J-L 引理最早出现在论文 [JL84] 中，最早考虑的是 Euclid 空间中点的嵌入问题。后来，这一引理为数据压缩提供很强的理论依据，也被广泛应用在深度学习中。关于 J-L 引理的应用，可以参见苏剑林的博客 [[苏 a](#), [苏 b](#), [苏 c](#)].

§4.5 习题

1. 证明定理 4.1.

2. * 本题证明定理 4.3. 首先, 定义 $X = (x_1, \dots, x_n)^\top$, 然后定义

$$\phi(X) = \left(1 - \frac{b}{\|X\|^2}\right) X.$$

(1) 证明, 均方误差具有如下分解:

$$\mathbb{E} [\|\phi(X) - n\|^2] = n - 2b\mathbb{E} \left[\frac{(X - \mu)^\top X}{\|X\|^2} \right] + b^2\mathbb{E} \left[\frac{1}{\|X\|^2} \right].$$

(2) 注意到, $\|X\|^2$ 是一个自由度为 n 、非中心参数为 $\|\mu\|^2$ 的 χ^2 分布。本问将它转化为中心 χ^2 分布。按如下方式定义随机变量 W : 首先, 生成一个随机变量 K , 服从参数为 $\|\mu\|^2 / 2$ 的 Poisson 分布, 然后, 条件在 $K = k$, W 服从自由度为 $n + 2k$ 的中心 χ^2 分布. 证明: W 和 $\|X\|^2$ 有相同的分布。

(3) 证明:

$$\mathbb{E} \left[\frac{1}{\|X\|^2} \right] = \mathbb{E} \left[\frac{1}{n - 2 + 2K} \right].$$

(4) 最后, 我们计算第二项期望。令

$$U = \frac{\mu^\top X}{\|X\|}, \quad V = \left\| X - \frac{\mu^\top X}{\|X\|^2} X \right\|^2.$$

证明: U 和 V 相互独立, 且 $U \sim \mathcal{N}(0, 1)$, $V \sim \chi_{n-1}^2$, 并且 W 与 $U^2 + V$ 同分布。

(5) 证明：

$$\mathbb{E} \left[\frac{U}{W} \right] = \frac{e^{-\|\mu\|^2/2}}{\sqrt{\pi}} \cdot \sum_{j=0}^{\infty} \frac{\Gamma(1/2) \|\mu\|^{2j+1}}{2^{j+1} \Gamma(j+1) \Gamma(n/2+j)}.$$

(6) 最后证明：

$$\mathbb{E} \left[\frac{(X - \mu)^T X}{\|X\|^2} \right] = 1 - \|\mu\| \mathbb{E} \left[\frac{U}{W} \right] = (n-2) \mathbb{E} \left[\frac{1}{n-2+2K} \right].$$

因此，当 $b = n - 2$ 时，我们就证明了定理 4.3.

3. 在 \mathbb{R}^n 中，坐标轴上有一个球心为 c 的单位球，球内均匀随机选取一个点 x ，记 $p_n = \Pr(\|x\| > \|c\|)$.
 - (1) 证明： $p_2 > 1/2$.
 - (2) 证明： $\lim_{n \rightarrow \infty} p_n = 1$.
4. 对于一个连续型随机变量 X ，如果 X 概率密度函数为 $p(x)$ ，那么 X 的众数是使得 $p(x)$ 取最大值的 x . 证明：自由度为 $k > 1$ 的 χ 分布，其众数是 $\sqrt{k-1}$.
5. 证明 (4.4).
6. 证明定理 4.6.
7. 证明：对于正态分布 $X \sim \mathcal{N}(0, 1/n)$ ，有

$$\mathbb{E} [e^{\lambda X^2}] = \sqrt{n/(n-2\lambda)}.$$

第五章 差分隐私

2006 年，Netflix 发起了一场全球竞赛，名为 Netflix Prize，旨在提升其电影推荐系统的精确度。当时，Netflix 还主要是一家邮寄 DVD 的租赁服务公司，推荐系统对于帮助用户发现可能喜欢的电影至关重要。为了激励全球数据科学家参与，Netflix 提供了 100 万美元的大奖，奖励给能将其推荐算法精度提高 10% 的团队。

Netflix 为比赛公开了一份庞大的数据集，包含超过一亿条匿名化的电影评分记录。这些数据来自近五十万名用户，涉及约一万八千部电影。为了保护隐私，Netflix 删除了所有直接的个人标识符，每个用户仅用一个随机的数字 ID 代替。同时，性别、年龄、邮政编码等信息也被去除。

然而，几周后，德克萨斯大学的博士生 Arvind Narayanan 和导师 Vitaly Shmatikov 证明，即使数据被匿名化，也能通过公开的互联网电影数据库（IMDb）进行去匿名化。他们发现，攻击者只需掌握用户在几部电影上的评分和时间，就能在 Netflix 数据集中重新识别出这个用户。这一发现揭示了严重的隐私风险。

这个隐私问题导致了一起重要的诉讼案件。一位未公开其性取向的同性恋母亲认为，Netflix 数据集的去匿名化可能会使她的性取向曝

光，进而影响她的职业生涯和家庭生活。她对 Netflix 提起了诉讼，要求 Netflix 根据《视频隐私保护法案》赔偿每位用户 2,500 美元，总计超过 5 亿美元。最终，Netflix 与她达成了和解，并决定取消计划中的第二届 Netflix Prize 竞赛。

直到 2024 年，Netflix 也没有举办过第二届 Netflix Prize 竞赛，它的服务也已经转变到互联网上的在线视频为主。然而，在当今世界，数据隐私的问题反而变得愈发严重。我们每天看的在线视频、购物、社交网络等行为都会产生大量的数据，根据 Statista 在 2020 年的一项估计，2024 年每人每天产生的数据量会达到 63.5GB！然而，这些数据都被私人公司所掌握！

本章我们关心数据的另一个维度：社会属性。尽管基于大数据和机器学习的技术的确改善了我们的生活，但是对数据的滥用也会带来严重的社会问题。如何保护个体的隐私，同时又能够让机器学习得到足够的数据？差分隐私就是解决这个问题的一个方法，本章将更详细地介绍差分隐私的概念和应用。

§5.1 数据隐私问题

本章开篇的 Netflix Price 案例展示了一个典型的数据隐私问题。实际上，不仅资本需要数据，科研也极其需要数据。以医学为例，我们需要大量真实的病人提供病情数据。但这些数据可能都涉及到病人的隐私信息，例如一些敏感数据。因此，我们必须找到一种方法，既可以收集到这些数据，又保护病人的隐私信息。

保护病人的隐私信息的一种理解是不会让数据获得者将数据和人对应起来。那么最直观的想法就是将每条数据匿名化。比如，每条数据

只包含患者的生日、性别、邮政编号（代表位置¹）和病情.

这样做依然有问题：同一天生日、同一性别、相同邮政编号的人很少，而这些信息很容易被找到，比如，在家过生日的时候，在微博上分享了带定位的生日照. 通过一条数据的各种属性可以轻易定位到这个人，这样的匿名化是不安全的.

以上例子展现的是一种去匿名化的现象，也就是说匿名的数据实际上揭示了数据对应的那个个体. 这去匿名化的出现都是因为多维数据具有稀疏性，也就是说，尽管维数很高，但是同时满足多个条件的数据却很少.

比如我们看表 5.1，这是医院甲的病人数据表. 56 岁的病人只有赵丽娜，所以假如我们知道赵丽娜的年龄并了解到她去过这家医院，便立即得知她患有艾滋病.

类似第四章，一种减少稀疏性的思想是降维，也就是我们用更粗糙的表示方法，但是让多维数据中同时满足多个条件的数据变得更多.

具体到隐私的情景下，我们称之为 k -匿名性：任何一个人的信息都不能和其他至少 $(k - 1)$ 人区分开. 比如，可以不明确写出姓名、年龄和邮编，只给出模糊的范围，于是数据变成了下面的表 5.2. 此时，有一个人的信息和赵丽娜完全相同，但是得了不同的病，这样赵丽娜的隐私就得到了保护.

这种方法仍然存在问题，因为我们不能把关键信息（病症信息）也模糊化. 如果我们还拿到了另一家医院乙模糊之后的病人数据表（表 5.3），那么依然有办法定位到赵丽娜. 这两张表上 50-60 岁的女性只有艾滋病是重合的，因此，如果我们知道赵丽娜的年龄并知道她同时去过两家医院，便立即得知她患有艾滋病.

¹如果研究的是传染病，那么位置的意义也是很大的.

姓名	年龄	性别	邮政编码	是否吸烟	诊断
李国强	64	男性	190146	是	心脏病
王秀英	61	女性	190118	否	关节炎
张建华	67	男性	190104	是	肺癌
陈玉兰	63	女性	190146	否	克罗恩病
刘志强	69	男性	190115	是	肺癌
赵丽娜	56	女性	190103	否	艾滋病
周志成	52	男性	190146	是	糖尿病
马文杰	59	男性	190130	是	过敏性鼻炎
李梅	55	女性	190146	否	溃疡性胃炎

表 5.1: 医院甲的病人数据表

除了使用匿名化的手段，还有一种方法可以保护隐私：不再提供单人的数据，而是直接公布将数据集的总体信息，比如平均值。但这种方法也不一定能保证不泄露单人数据，请看下面的例子。

在第 3.4 节中我们介绍过，要训练一个生成式语言模型（例如 Chat-GPT），需要大量的文本数据。这些文本数据通常都是公司的机密，我们不得而知。然而，我们可以用很简单的手段来推测这些数据。

2023 年 12 月 27 日，《纽约时报》因版权侵犯问题对 OpenAI 和微软提起诉讼。该报指控这两家公司在训练其自动聊天机器人（即 ChatGPT）时，未经许可使用了数百万篇《纽约时报》的文章。这些聊天机器人如今被视为可信的信息来源，并与《纽约时报》等新闻机构展开了直接竞争。

《纽约时报》在诉讼中提到几个实例，显示聊天机器人提供的内容

姓名	年龄	性别	邮政编码	是否吸烟	诊断
*	60-70	男	1901**	是	心脏病
*	60-70	女	1901**	否	关节炎
*	60-70	男	1901**	是	肺癌
*	60-70	女	1901**	否	克罗恩病
*	60-70	男	1901**	是	肺癌
*	50-60	女	1901**	否	艾滋病
*	50-60	男	1901**	是	糖尿病
*	50-60	男	1901**	是	过敏性鼻炎
*	50-60	女	1901**	否	溃疡性胃炎

表 5.2: 医院甲的病人数据表, 模糊了姓名、年龄和邮编

姓名	年龄	性别	邮政编码	诊断
*	50-60	女	1901**	艾滋病
*	50-60	女	1901**	系统性红斑狼疮
*	50-60	女	1901**	干燥综合征
*	60-70	男	1901**	胰腺癌
*	60-70	男	1901**	肝硬化
*	60-70	男	1901**	通风

表 5.3: 医院乙的病人数据表, 模糊了姓名、年龄和邮编

几乎与《纽约时报》的文章完全一致, 而这些文章在《纽约时报》的网站上需付费订阅才能阅读. 该报表示, OpenAI 和微软特别强调在训练过程中使用了《纽约时报》的报道, 因其认为这些材料具备可靠性和准

确性.

本案件如同罗生门，只要微软和 OpenAI 不主动承认，没有人可以证明他们确实使用了《纽约时报》的文章。然而，这一案例足够说明一个道理：即使不提供模型本身，只提供模型的输出，也可能泄露出原始数据。

注。以上这些内容都说明匿名化很难保护个人隐私。那么，密码学加密的手段是否能够保护隐私？比如说，为什么我们的所有数据不能需要账号密码才能被访问？其实，加密和隐私在出发点上完全不同。

加密的目的是为了不让别人获取到真实数据。而隐私不是简单地保护数据，它涉及更微妙的问题——我们希望利用这些数据，但是又不希望泄露某个个体的信息。因此我们这里讨论的其实是隐私保护问题而不是加密问题，密码学的方案并不适用于这里。

§5.2 差分隐私的定义与性质

我们上面探讨了隐私保护的必要性以及它的微妙之处，现在我们要给出一种合理的方案解决隐私保护的问题，这个方案就是差分隐私。要给出一个数学模型，不仅要知道什么情况下算是隐私泄漏，也需要知道什么情况下不算，所以我们再来看一个反面的例子。

李明是一位长期吸烟的男子，他参加了一项有关“吸烟与健康”的调查。这项调查在不久后发布了一项结果，表明长期吸烟的人患上肺癌的几率更大。伴随着这一结果的公布，保险公司在出售相同保险时会对长期吸烟者索要更高的价格。李明当然也受到了这一政策的影响。那我们是否可以认为这项研究泄露了李明的隐私（即更有可能患肺癌）呢？

直接告诉我们，这不应该算泄露了隐私。“长期吸烟的人患上肺癌

的几率更大”这项结论并不依赖于李明是否参加了调查。考虑这样的对照， x 代表原来参加调查的人的集合， x' 代表其他人不变，只是李明换成了另外一个人的集合。如果是 x' 这些人参与了调查，结论是否会发生变化？大概率不会！

李明的例子告诉我们，隐私应该有以下性质：

当数据集中包含李明的信息，相比数据集中不包含李明的信息，并不会显著增加损害李明的利益的概率。

这一思想引出了差分隐私的概念，我们将在下面给出数学形式的定义。

考虑数据的空间 \mathcal{X} ，其中的每一个元素都包含了个体的所有可能数据例如姓名、性别、年龄、国籍等。考虑 n 个人的数据，形成了有序的数据集

$$x = (x_1, \dots, x_n) \in \mathcal{X}^n.$$

我们希望对这些数据进行一定的加工，例如用这些数据来训练一个机器学习模型。抽象来说，加工的过程可以被看成一个随机算法 A ：在固定输入数据集 $x \in \mathcal{X}^n$ 下， $A(x)$ 是输出空间 \mathcal{Y} 上的一个随机变量。

当我们改变（增加或删除）一个人的数据时，我们希望结果分布的变化可以控制在一定范围内。为此，我们引入相邻数据集的概念：

定义 5.1 (k -相邻数据集) 设 $x, x' \in \mathcal{X}^n$ ，如果 x 和 x' 最多有 k 条数据不同，即至多存在 k 个不同的 $i_1, \dots, i_k \in [n]$ 使得 $x_{i_j} = x'_{i_j}$ 对 $j \in [k]$ 成立，那么称 x 和 x' 是 k -相邻的。

特别地，如果 $k = 1$ ，我们称 x 和 x' 是相邻的。 \square

在李明的例子中，我们把李明换掉之前和之后的数据集是 1-相邻的。

接下来我们给出差分隐私的定义。直观上，不论数据集中是否包含某个人的数据，算法的输出分布不会有太大的变化，因而我们有：

定义 5.2 (差分隐私, ϵ -DP) 考虑随机算法 $A : \mathcal{X}^n \rightarrow \mathcal{Y}$ ，如果对于任意一对 1-相邻数据集 x, x' ，对任意（可测）值集 $E \subseteq \mathcal{Y}$ ，有

$$\Pr(A(x) \in E) \leq e^\epsilon \cdot \Pr(A(x') \in E),$$

那么我们称 A 为数据集大小为 n 的 ϵ -DP 算法。 \square

这一定义看起来是不对称的，其实它是对称的，并且是用概率的比值衡量分布的差异。

命题 5.1 设 $A : \mathcal{X}^n \rightarrow \mathcal{Y}$ 是一个 ϵ -DP 算法，那么对于任意的 1-相邻数据集 x, x' 和任意（可测）值集 $E \subseteq \mathcal{Y}$ ，有

$$e^{-\epsilon} \leq \frac{\Pr(A(x) \in E)}{\Pr(A(x') \in E)} \leq e^\epsilon.$$

证明。 不等式右边就是定义，左边只需要把定义中 x 和 x' 的地位互换即可。 \square

从这一性质出发，如果 ϵ 越接近 0，那么 A 的输出分布在相邻数据集上的变化就越小，也就是说隐私保护效果越好。

注。 为什么选择用比值来衡量分布的差异，而不是直接用差值？注意，如果我们要讨论对数据的多重加工，那么加工之间的独立性就是一个很重要的性质。独立性的定义就使用了乘法（也就是比值），而不是加法。因此，我们将在后面看到，这样定义的差分隐私具有极其干净的数学性质。

以上定义需要对所有的（可测）值集 E 都成立，这给验证带来了极

大的困难，如果随机算法的输出分布是更加常规的，我们可以简化验证的过程。

对于离散型的输出，我们有如下等价定义：

命题 5.2 如果 $A(x)$ 对于任意 $x \in \mathcal{X}^n$ 都是离散型随机变量，那么随机算法 A 是数据集大小为 n 的 ϵ -DP 算法当且仅当对于任意一对 1-相邻数据集 x, x' 和所有的 $y \in \mathcal{Y}$ ，有

$$\Pr(A(x) = y) \leq e^\epsilon \cdot \Pr(A(x') = y).$$

证明. \implies ：取 $E = \{y\}$ 即可证明。

\impliedby ：假设 $E = \{y_1, \dots, y_k, \dots\} \subseteq \mathcal{Y}$. 对每一个 y_i ，都有

$$\Pr(A(x) = y_i) \leq e^\epsilon \cdot \Pr(A(x') = y_i).$$

因为 $A(\cdot) = y_i$ 对于不同的 i 是互斥事件，所以概率可以直接相加，于是：

$$\begin{aligned} \Pr(A(x) \in E) &= \sum_i \Pr(A(x) = y_i) \\ &\leq e^\epsilon \cdot \sum_i \Pr(A(x') = y_i) \\ &= e^\epsilon \cdot \Pr(A(x') \in E). \end{aligned} \quad \square$$

对连续型的输出，我们有如下等价定义：

命题 5.3 如果 $A(x)$ 对于任意 $x \in \mathcal{X}^n$ 都是连续型随机变量，那么它存在概率密度函数，记为 h_x . 此时，随机算法 A 是数据集大小为 n 的 ϵ -DP

算法当且仅当对于任意一对 1-相邻数据集 x, x' 和几乎所有的 $y \in \mathcal{Y}$, 有

$$h_x(y) \leq e^\epsilon \cdot h_{x'}(y).$$

证明. \implies : 这一部分的严格表述需要测度论的基础, 所以这一证明从直观上理解即可, 不需要考虑严格的定义. 如果需要看更严格的定义, 请参阅附录 C.

定义集合

$$E = \{y : h_x(y) > e^\epsilon \cdot h_{x'}(y)\}.$$

我们证明 $\lambda(E) = 0$, 其中 λ 是 Lebesgue 测度; 也就是说, 几乎所有的 y 都满足

$$h_x(y) \leq e^\epsilon \cdot h_{x'}(y).$$

为此, 我们定义集合

$$E_n = \left\{ y : h_x(y) > e^\epsilon \cdot h_{x'}(y) + \frac{1}{n} \right\}, n = 1, 2, \dots$$

因为

$$E = \bigcup_{n=1}^{\infty} E_n,$$

所以我们只需要证明 $\lambda(E_n) = 0$, 然后利用测度的次可加性即可证明 $\lambda(E) = 0$.

假设 $\lambda(E_n) > 0$, 那么, 在 E_n 上对密度函数积分, 有

$$\int_{E_n} (h_x(y) - e^\epsilon \cdot h_{x'}(y)) dy \geq \int_{E_n} \frac{1}{n} dy = \frac{\lambda(E_n)}{n} > 0.$$

根据概率密度的定义，

$$\Pr(A(x) \in E_n) = \int_{E_n} h_x(y) dy, \quad \Pr(A(x') \in E_n) = \int_{E_n} h_{x'}(y) dy.$$

所以

$$\Pr(A(x) \in E_n) > e^\epsilon \cdot \Pr(A(x') \in E_n).$$

这与 A 是 ϵ -DP 算法矛盾，所以 $\lambda(E_n) = 0$ ，这就完成了证明。

\Leftarrow ：依然根据概率密度函数的定义，考虑 $x, x' \in \mathcal{X}^n$ ，对任意可测 $E \subseteq \mathcal{Y}$ ，有

$$\begin{aligned} \Pr(A(x) \in E) &= \int_E h_x(y) dy \\ &\leq e^\epsilon \cdot \int_E h_{x'}(y) dy \\ &= e^\epsilon \cdot \Pr(A(x') \in E). \end{aligned} \quad \square$$

接下来我们给出差分隐私的基本性质。首先，如果我们对一组数据 x 依次独立使用两个差分隐私算法 A_1 和 A_2 进行处理（允许 A_2 使用 A_1 输出的结果），得到两组数据，那么总体上看，这整个过程还是一个差分隐私算法。

命题 5.4 (复合性，两个算法的情形) A_1 和 A_2 是相互独立的随机算法，其中

$$A_1 : \mathcal{X}^n \rightarrow \mathcal{Y}_1,$$

$$A_2 : \mathcal{Y}_1 \times \mathcal{X}^n \rightarrow \mathcal{Y}_2.$$

假设 A_1 是 ϵ_1 -DP 算法， A_2 是 ϵ_2 -DP 算法。

令 $A : \mathcal{X}^n \rightarrow \mathcal{Y}_1 \times \mathcal{Y}_2$ 是随机算法，输出为 $A(x) = (y_1, y_2)$ ，其中 $y_1 = A_1(x)$, $y_2 = A_2(y_1, x)$ ，那么 A 是 $(\epsilon_1 + \epsilon_2)$ -DP 算法。

证明. 为了简化记号，这里我们只证明离散的情况，一般情况的证明类似。令 x, x' 是 \mathcal{X}^n 中的两个 1-相邻数据集， A 输出为 $y = (y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ ，那么根据定义和独立性，

$$\Pr(A(x) = (y_1, y_2)) = \Pr(A_1(x) = y_1) \cdot \Pr(A_2(y_1, x) = y_2).$$

由于 A_1 是 ϵ_1 -DP 算法， A_2 是 ϵ_2 -DP 算法，得到

$$\begin{aligned} \Pr(A(x) = (y_1, y_2)) &= \Pr(A_1(x) = y_1) \cdot \Pr(A_2(y_1, x) = y_2) \\ &\leq e^{\epsilon_1} \Pr(A_1(x') = y_1) \cdot e^{\epsilon_2} \Pr(A_2(y_1, x') = y_2) \\ &= e^{\epsilon_1 + \epsilon_2} \cdot \Pr(A(x') = (y_1, y_2)). \end{aligned}$$

□

现在，我们已经看到为什么差分隐私的定义是用比值来衡量的：比值陈述的差分隐私，这一命题的结论和证明都非常干净。利用数学归纳法，很容易推广到多个随机算法的复合性：

命题 5.5 (复合性，多个算法的情形) 设 A_1, A_2, \dots, A_k 为一列相互独立的随机算法，

$$A_1 : \mathcal{X}^n \rightarrow \mathcal{Y}_1,$$

$$A_i : \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1} \times \mathcal{X}^n \rightarrow \mathcal{Y}_i, \quad i = 2, 3, \dots, k.$$

也就是 A_i 将 A_1, \dots, A_{i-1} 的输出和 \mathcal{X}^n 中的一个数据集作为输入元素。对 $i = 1, \dots, k$, A_i 是 ϵ_i -DP 算法。

依次运行算法 A_i 得到算法 $A : \mathcal{X}^n \rightarrow \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$, 那么 A 是 ϵ -DP, 其中 $\epsilon = \sum_{i=1}^n \epsilon_i$.

证明. 对 n 用数学归纳法.

- 当 $n = 1$ 时, 结论显然成立.
- 假设结论对 n 成立, 我们要证明对 $n + 1$ 结论成立.

考虑前 n 个算法的复合算法

$$A'(x) = (A_1(x), A_2(A_1(x), x), \dots, A_n(A_{n-1}(\dots A_1(x), \dots), x)).$$

由归纳假设, A' 是 $\sum_{i=1}^n \epsilon_i$ -DP 算法.

由于 A_{n+1} 是 ϵ_{n+1} -DP 算法, A' 与 A_{n+1} 的复合算法是 A , 所以利用命题 5.4, A 是 $\sum_{i=1}^{n+1} \epsilon_i$ -DP 算法. \square

接下来, 我们说明差分隐私最关键的性质: 一旦被差分隐私算法处理过, 无论后续如何处理, 都不会影响隐私保护的效果.

命题 5.6 (后处理) 令 $A : \mathcal{X}^n \rightarrow \mathcal{Y}$, $B : \mathcal{Y} \rightarrow \mathcal{Z}$ 为相互独立的随机算法, 其中 \mathcal{X} , \mathcal{Y} , \mathcal{Z} 是任意集合. 如果 A 是 ϵ -DP 算法, 那么复合算法 $B(A(\cdot))$ 也是 ϵ -DP 算法.

证明. 我们仍然只考虑离散情形, 采用定义的方法证明

$$\begin{aligned} \Pr(B(A(x)) = b) &= \sum_{y \in \mathcal{Y}} \Pr(A(x) = y) \Pr(B(y) = b) \\ &\leq e^\epsilon \sum_{y \in \mathcal{Y}} \Pr(A(x') = y) \Pr(B(y) = b) \\ &= e^\epsilon \Pr(B(A(x')) = b). \end{aligned} \quad \square$$

最后，我们讨论如果有多人的数据都发生变化的时候，隐私保护的性质会发生什么变化。

命题 5.7 (群体隐私) 令 $x, x' \in \mathcal{X}^n$ 是 k -相邻数据集， $1 \leq k \leq n$. 如果 A 是 ϵ -DP 算法，那么对所有的值集 E ，我们有

$$\Pr(A(x) \in E) \leq e^{k\epsilon} \Pr(A(x') \in E).$$

证明. 考虑数据集 x_0, x_1, \dots, x_k ，其中 $x_0 = x, x_k = x'$ ，且 x_i 和 x_{i+1} 是 1-相邻数据集， $i = 0, \dots, k - 1$. 那么

$$\begin{aligned} \Pr(A(x) \in E) &\leq e^\epsilon \Pr(A(x_1) \in E) \leq e^{2\epsilon} \Pr(A(x_2) \in E) \\ &\leq \cdots \leq e^{k\epsilon} \Pr(A(x') \in E). \end{aligned}$$
□

换言之， k -相邻数据集上 ϵ -DP 算法的表现仿佛一个 $k\epsilon$ -DP 算法。

这一性质还可以推出 ϵ 的含义。我们知道，数据集 $x, x' \in \mathcal{X}^n$ 最多在 n 个位置不同。所以对于一个 ϵ -DP 算法 A ，一定有

$$\Pr(A(x) \in E) \leq e^{n\epsilon} \Pr(A(x') \in E).$$

如果 ϵ 太小，这一算法对任何输入都有相似的输出。换句话说，算法压根没有输出任何有意义的内容。于是，我们定量说明了， ϵ 还代表信息的泄露量。因此，一个实用的 DP 算法不能让 ϵ 太小，否则输出没有意义；也不能让 ϵ 太大，否则隐私保护效果不好。

§5.3 差分隐私的应用

在这一部分，我们将会具体讨论三个差分隐私的算法或机制。

§5.3.1 随机反应算法

我们从一个具体场景开始。假设有一名老师想要调查班上的同学有多少人曾经在考试中作弊。设班上一共有 n 名同学，每个人的真实答案是一个数字 $x_i \in \{0, 1\}$ ，0 表示没有作弊，1 表示作弊。

没有人愿意主动承认自己作弊，所以，对于每个 i ，独立地按照以下规则根据 x_i 得到对应的 y_i ：

$$y_i = \begin{cases} x_i, & \text{以 } 2/3 \text{ 的概率,} \\ 1 - x_i, & \text{以 } 1/3 \text{ 的概率.} \end{cases}$$

然后，老师让每个同学回答 y_i ，并公布 $\sum_{i=1}^n y_i$ 。

这的确可以保护隐私：当 $y_i = 1$ 时，学生 i 可以声称这是由于算法的随机机制造成的，而并非自己真的作弊过，因而他愿意诚实回答。

我们称这一算法为随机反应 (RR) 算法，它是 S. L. Warner 在 1965 年提出的（远远早于差分隐私概念的提出！），被广泛用于调查敏感问题。

以下结论说明了 RR 算法的确是差分隐私算法。

定理 5.1 RR 算法是 log 2-DP 算法。

证明. 记 Y_i 是 y_i 对应的随机变量. 我们知道 y_i 之间相互独立, 所以

$$\Pr(A(x) = y) = \prod_{i=1}^n \Pr(Y_i = y_i | x_i).$$

对于 y_i , 我们有

$$\frac{\Pr(Y_i = y_i | x_i = y_i)}{\Pr(Y_i = y_i | x_i = 1 - y_i)} = \frac{2/3}{1/3} = 2.$$

类似地, 我们有

$$\frac{\Pr(Y_i = y_i | x_i = 1 - y_i)}{\Pr(Y_i = y_i | x_i = y_i)} = \frac{1/3}{2/3} = 1/2.$$

所以, 对于一对 1-相邻数据集 x, x' 和任意的 $y \in \mathcal{Y}$, 假设 $x_j \neq x'_j$, 不论 y_j , x_j 和 x'_j 的取值是什么, 我们都有

$$\begin{aligned} & \frac{\Pr(A(x) = y)}{\Pr(A(x') = y)} \\ &= \frac{\prod_{i=1}^n \Pr(Y_i = y_i | x_i)}{\prod_{i=1}^n \Pr(Y_i = y_i | x'_i)} \\ &= \frac{\Pr(Y_j = y_j | x_j)}{\Pr(Y_j = y_j | x'_j)} \leq 2. \end{aligned}$$

由命题 5.2, RR 算法是 log 2-DP 算法. □

另一方面, 我们真正关心的是 $\sum_{i=1}^n x_i$, 也就是有多少人作弊. 我们希望 RR 算法的输出能够很好地估计出这一值, 也就是 RR 算法得到的 $\sum_{i=1}^n Y_i$ 是否能很好的估计出 $\sum_{i=1}^n x_i$.

为此, 假设每个人独立地有 p 的概率作弊, 那么从期望上说, 作弊

的人大概有 np 个，但我们不知道 p . 假设每个人的真正回答是随机变量 Y_i , 我们来计算 $\mathbb{E}[\sum_{i=1}^n Y_i]$.

$$\begin{aligned}\mathbb{E}[Y_i] &= \Pr(Y_i = 1) = p \cdot \frac{2}{3} + (1 - p) \cdot \frac{1}{3} \\ &\implies p = 3\mathbb{E}[Y_i] - 1.\end{aligned}$$

对 i 累和，我们有

$$np = 3\mathbb{E}\left[\sum_{i=1}^n Y_i\right] - n.$$

所以，RR 算法的输出 $\sum_{i=1}^n y_i$ 在期望上很好地估计了 $\sum_{i=1}^n x_i$.

§5.3.2 全局灵敏度与 Laplace 机制

RR 算法是如何实现差分隐私的呢？最重要的一步是，它把真实数据加了噪声，但是这一噪声并不会过分地影响结果. 更一般地，我们似乎总是可以通过向输出添加噪声来实现差分隐私. 那么就引出了另一个问题，需要添加多大的噪声？

一个极端是，如果我们添加的噪声太大，那么输出就完全变成随机的了；另一个极端是，我们完全不加噪声，这样就没有隐私保护的效果. 我们希望能够找到一个合适的噪声水平，使得既能保护隐私，又能保持输出的有用性. 多大的噪声合适，这和算法本身的性质有关，比如，当输入只改变一点时，算法的输出会改变多大？我们定义全局灵敏度来衡量这一性质.

定义 5.3 (全局灵敏度) 给定算法 $f : \mathcal{X}^n \rightarrow \mathbb{R}$, 定义 f 的全局灵敏度为

$$\text{GS}_f = \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} |f(x) - f(x')|.$$

□

定义中的 1-相邻，可能会随着情景不同而改变。全局灵敏度的定义是很直观的，就是改变一条数据会对算法输出带来的最大可能变化。

我们来计算一个简单的例子。

例 5.1 设 $f(x) = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$, $\phi: \mathcal{X} \rightarrow [0, 1]$ 是满射。那么，

$$\begin{aligned} \text{GS}_f &= \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} |f(x) - f(x')| \\ &= \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 1-相邻}} \frac{1}{n} \left| \sum_{i=1}^n \phi(x_i) - \sum_{i=1}^n \phi(x'_i) \right| \\ &= \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 只在 } j \text{ 不同}} \frac{1}{n} |\phi(x_j) - \phi(x'_j)| = \frac{1}{n}. \quad \square \end{aligned}$$

接下来，我们试图给出一个一般的方法来构造差分隐私算法。给定一个数据集 $x \in \mathcal{X}^n$ 和参数 ϵ 。对于一个算法 $f: \mathcal{X}^n \rightarrow \mathbb{R}$ ，我们想要给 f 的输出加一个噪声 Z ，使它具有差分隐私的性质。新得到的算法 A 是一个随机算法，给定数据集 x ，它输出

$$A(x) = f(x) + Z.$$

设 x, x' 是两个 1-相邻数据集，记 $\mu = f(x)$, $\mu' = f(x')$ 。如果 Z 是连续型随机变量，我们希望密度函数 h_x 和 $h_{x'}$ 满足

$$\frac{h_x(y)}{h_{x'}(y)} \leq e^\epsilon.$$

如何选择这样的 Z 呢？注意， $A(x)$ 和 $A(x')$ 的只差一个常数 $\mu - \mu'$ ，

所以密度函数 h_x 和 $h_{x'}$ 应该是关于 y 的平移，也就是

$$h_x(y) = h_{x'}(y - (\mu - \mu')).$$

因为我们想要构造 e^ϵ ，所以不妨设想 Z 的密度函数具有指数 e^{-x} 的形式²，我们就会有

$$\frac{h_x(y)}{h_{x'}(y)} = \frac{h_{x'}(y - (\mu - \mu'))}{h_{x'}(y)} = e^{(\mu - \mu')}.$$

如果我们想要 ϵ -DP，那么我们就需要让 Z 的扰动不要太大。如果用 αZ 来替代 Z ，那么

$$\frac{h_x(y)}{h_{x'}(y)} = e^{\alpha(\mu - \mu')}.$$

注意到， $|\mu - \mu'| \leq GS_f$ ，所以我们需要 $\alpha \leq \epsilon/GS_f$.

以上推导过程虽然不是严格的，但是它给出了我们构造差分隐私算法的一个思路，这一思路的最终结果为 *Laplace* 机制。首先引入 Laplace 分布的概念。

定义 5.4 (Laplace 分布) 给定参数 $\mu \in \mathbb{R}$ 和 $\lambda > 0$ ，定义概率密度函数

$$h(x|\mu, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|x - \mu|}{\lambda}\right).$$

我们称具有这一密度的分布为 *Laplace* 分布，记为 $\text{Lap}(\mu, \lambda)$. □

Laplace 分布是服从双边指数分布的随机变量进行线性变换后服从的分布。具体来说，设 $X \sim DExp(1)$ ，那么 $\lambda X + \mu \sim \text{Lap}(\mu, \lambda)$ 。更多关于 *Laplace* 分布的性质，我们留做习题 1.

²在此时，我们先不管它是不是密度函数，而是看如何才能够造出来差分隐私算法。

下面我们来陈述 Laplace 机制.

- 给定算法 $f : \mathcal{X}^n \rightarrow \mathbb{R}$, 参数 $\epsilon > 0$.
- 计算全局灵敏度 GS_f .
- 构造随机算法 $A_{\text{Lap}}(x; \epsilon) = f(x) + Z$, 其中 $Z \sim \text{Lap}(0, GS_f / \epsilon)$.

如此, 我们就得到了一个随机算法 A_{Lap} .

定理 5.2 对于任意的 $\epsilon > 0$, A_{Lap} 是 ϵ -DP 算法.

证明. 设 x, x' 是两个 1-相邻数据集, 记 $\mu = f(x)$, $\mu' = f(x')$. 由 Laplace 分布的性质可知, $A_{\text{Lap}}(\epsilon, x) \sim \text{Lap}(\mu, GS_f / \epsilon)$, $A_{\text{Lap}}(\epsilon, x') \sim \text{Lap}(\mu', GS_f / \epsilon)$.

因此, 对于任意的 $y \in \mathcal{Y}$, 有

$$\begin{aligned} \frac{h_x(y)}{h_{x'}(y)} &= \exp\left(-\epsilon \frac{|\mu - y| - |\mu' - y|}{GS_f}\right) \\ &\leq \exp\left(\epsilon \frac{|\mu - \mu'|}{GS_f}\right) \leq \exp(\epsilon). \end{aligned}$$

根据命题 5.3, 命题得证. □

§5.3.3 DP 版本 Llyod 算法

作为一个 Laplace 机制的具体实例, 我们将 k -均值聚类问题的经典算法 Llyod 算法改造成一个差分隐私算法.

k -均值聚类问题指的是给定一个数据集 x , 找到 k 个点(中心) $\{c_i\} \subseteq \mathbb{R}^d$, 使得 $\sum_{i \in [n]} \min_{j \in [k]} \|x_i - c_j\|^2$ 最小. 通俗来说, 就是找到 k 个中心, 使得数据集中每个点到最近的中心的距离之和最小.

在我们开始讨论算法和改造之前，我们先要理解为什么 k -均值问题需要差分隐私。我们可以把每一个点看作一个人的数据，而中心看作是一个数据集的平均特征。我们只希望算法可以知道数据的平均特征，而不希望泄露某个个人的数据。因此，我们希望 k -均值算法是差分隐私的。

k -均值问题最常见的解决方法是使用迭代的启发式的 *Lloyd* 算法，其表述为下：

- 输入：数据集 $x \in \mathcal{X}^n$ ，这里 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$ ，参数 k ，最大迭代次数 T .
- 随机初始化 $c_1^{(0)}, c_2^{(0)}, \dots, c_k^{(0)} \in \mathcal{X}$.
- for $t = 1$ to T
 - for $j = 1$ to k
 - * 计算 $S_j = \{i : c_j^{(t-1)}$ 是 x_i 最近的中心 $\}$.
 - * 计算 $n_j = |S_j|$.
 - * 计算 $a_j = \sum_{i \in S_j} x_i$.
 - * 更新 $c_j^{(t)} = a_j / n_j$.
- 输出： $c_1^{(T)}, c_2^{(T)}, \dots, c_k^{(T)}$.

这一算法的思路大致是：首先随机初始化 k 个中心 c_i ，然后，对每一个 c_i ，找到离它最近的那些点，以这些点的平均值（也就是质心）作为新的 c_i ，重复这一过程直到收敛。

注意，如果初始化的中心就是各个质心，那么 Lloyd 算法就已经找到最优解了。事实上，可以证明，Lloyd 算法最终会收敛到最优解。

尽管 Lloyd 算法可以达到很好的效果，但它并不能保证 DP 性质。我们希望对这一算法进行小规模的修改，让它具有 ϵ -DP 的性质。我们给出如下的 DP 版本 Lloyd 算法。

- 输入：数据集 $x \in \mathcal{X}^n$ ，这里 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$ ，参数 k ，最大迭代次数 T ，参数 ϵ 。
- $\epsilon' = \frac{\epsilon}{2T}$ ，随机初始化 $c_1^{(0)}, c_2^{(0)}, \dots, c_k^{(0)} \in \mathcal{X}$ 。
- for $t = 1$ to T
 - for $j = 1$ to k
 - * 计算 $S_j = \{i : c_j^{(t-1)}$ 是 x_i 最近的中心 $\}$ 。
 - * $n_j = |S_j|$.
 - * $a_j = \sum_{i \in S_j} x_i$.
 - * 计算 $\hat{n}_j = n_j + Y$, $Y \sim \text{Lap}(0, 2/\epsilon')$.
 - * 计算 $\hat{a}_j = a_j + (Z_1, \dots, Z_d)$, Z_i i.i.d. $\sim \text{Lap}(0, 2/\epsilon')$.
 - * 更新 $c_j^{(t)} = \begin{cases} \hat{a}_j / \hat{n}_j, & \hat{n}_j \geq 1, \\ \mathcal{X} 上的一个随机均匀采样, & \hat{n}_j < 1. \end{cases}$
- 输出： $c_1^{(T)}, c_2^{(T)}, \dots, c_k^{(T)}$ 。

简单来说，我们对原来 Lloyd 算法中产生的中间值都是用 Laplace 机制来加噪声，最终得到的中心也是加了噪声的。以下定理表明上面的算法确实是一个 ϵ -DP 算法。

定理 5.3 DP 版本的 Lloyd 算法是 ϵ -DP 算法。

证明. (证明概要) 我们只在这里陈述证明的大致想法, 将细节留到习题 2.

首先, 作为一个迭代算法, 我们可以把每轮迭代看成一次算法 A_t 的执行, 而整个算法是这些算法的复合, 整个算法一共有 T 次复合. 由命题 5.5, 我们只需要证明每一轮迭代是 (ϵ/T) -DP 算法, 那么整个算法就是 ϵ -DP 算法.

进一步, 考虑证明每一个 A_t 是 (ϵ/T) -DP 算法. 每一个 A_t 内部都是 k 个独立的 Laplace 机制, 实际上是一个输入为 n_j, a_j , 输出为 \hat{n}_j, \hat{a}_j 的算法. 所以, 只需要证明: 每一个独立 Laplace 机制确实是 (ϵ/T) -DP, 然后再证明这些独立的 Laplace 机制总和的效果依然是 (ϵ/T) -DP, 我们就可以证明整个命题. \square

§5.4 习题

1. 关于双指教分布和 Laplace 分布, 请完成下面的习题.

- (1) 设 $X \sim \text{DExp}(1) = \text{Lap}(0, 1)$, 即它有密度函数 $e^{-|x|}/2$. 设 $\mu \in \mathbb{R}$, $\lambda > 0$. 设 $Y = \lambda X + \mu$, 证明 $Y \sim \text{Lap}(\mu, \lambda)$, 求 Y 的期望和方差.
- (2) 设 X, Y i.i.d $\sim U(0, 1)$, 证明: $\log(X/Y) \sim \text{DExp}(1) = \text{Lap}(0, 1)$.
如果 X, Y 不独立, 结论还成立吗?

2. 证明定理 5.3.

3. 现考虑输出为 d 维的算法 $f : \mathcal{X}^n \rightarrow \mathbb{R}^d$, 定义其全局灵敏度为

$$\text{GS}_f := \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 中的 } 1\text{-相邻}} \|f(x) - f(x')\|_1$$

$$= \sup_{x, x' \text{ 在 } \mathcal{X}^n \text{ 中的 } 1\text{-相邻}} \sum_{i=1}^d |f_i(x) - f_i(x')|.$$

请写出这种情况下的 Laplace 机制，使其成为 ϵ -DP，并证明。

4. 设输出为 d 维的算法 $f(x) = (\sum_{i=1}^n \Phi_1(x_i), \dots, \sum_{i=1}^n \Phi_d(x_i))$, $x_i \in \mathcal{X}$, 满足 $\Phi_j(\mathcal{X}) = \{0, 1\}$ 以及 $\sum_{j=1}^d \Phi_j(x_i) = 1$. 请计算 GS_f .
5. 差分隐私和基于信息论的度量有着密切的联系. 对于一个包含 n 个(带随机性的) 数据条目 $\mathcal{X}^n = (\mathcal{X}_1, \dots, \mathcal{X}_n)$ 的数据, 考虑一个随机算法 $M : \mathcal{X}^n \rightarrow \mathcal{Y}$, 输出 Y 的概率分布为 $P_{\mathcal{Y}|\mathcal{X}^n}$. 我们引入以下概念:

- **互信息差分隐私.** 如果算法 M 满足下面的不等式, 就称 M 是一个 ϵ -MI-DP:

$$\sup_{i, P_{\mathcal{X}^n}} I(\mathcal{X}_i; \mathcal{Y} | \mathcal{X}_{-i}) \leq \epsilon.$$

其中 $\mathcal{X}_{-i} = \{\mathcal{X}_1, \dots, \mathcal{X}_{i-1}, \mathcal{X}_{i+1}, \dots, \mathcal{X}_n\}$ 表示除了 \mathcal{X}_i 之外的所有数据条目.

- **K-L 散度差分隐私.** 如果对于所有相邻的数据集 x 和 \tilde{x} , 算法 M 满足下面的不等式, 就称 M 是一个 ϵ -KL-DP:

$$D_{\text{KL}}(P_{\mathcal{Y}|\mathcal{X}^n=x} \| P_{\mathcal{Y}|\mathcal{X}^n=\tilde{x}}) \leq e^\epsilon.$$

证明下列不等式链:

$$\epsilon\text{-DP} \succeq \epsilon\text{-KL-DP} \succeq \epsilon\text{-MI-DP}.$$

其中 \succeq 表示“不严格地更强”, 即满足左侧的差分隐私机制也满足右侧的差分隐私机制.

提示：条件互信息可以写成 K-L 散度的形式：

$$I(\mathcal{X}_i; \mathcal{Y} | \mathcal{X}_{-i}) = \mathbb{E} \left[D_{\text{KL}} \left(P(\mathcal{Y} | \mathcal{X}^n = \bar{x}^n) \| P(\mathcal{Y} | \mathcal{X}^{-i} = \bar{x}^{-i}) \right) \right],$$

其中 $\bar{x}^n \sim P_{\mathcal{X}^n}$.

6. 本题讨论差分隐私的其他形式。差分隐私实际上使用了如下不等式作为定义：

$$\Pr(A(x) \in E) \leq e^\epsilon \Pr(A(x') \in E).$$

即乘性不等式，而不是下面的加性不等式：

$$\Pr(A(x) \in E) \leq \Pr(A(x') \in E) + \epsilon.$$

除了数学上的简洁优美，我们没有理由说乘性不等式比加性不等式更好。因此，我们可以定义一种 (ϵ, δ) -DP，它要求

$$\Pr(A(x) \in E) \leq e^\epsilon \Pr(A(x') \in E) + \delta.$$

在这种定义中，乘性和加性都体现了出来。

对于两个 1-相邻数据集 X 和 X' ，输出 $y \in \mathcal{Y}$ 和随机算法 A ，我们定义随机算法 A 的隐私损失为：

$$c_A(y, X, X') := \log \frac{\Pr(A(X) = y)}{\Pr(A(X') = y)}.$$

证明：随机算法 A 是 (ϵ, δ) -DP 的一个充分条件是其隐私损失满足：

$$\Pr(c_A(y, X, X') > \epsilon) \leq \delta.$$

7. 本题讨论差分隐私对个体效用期望的影响. 设 $u_i : \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ 为算法输出对第 i 个个体的非负效用映射 (即对任意 $y \in \mathcal{Y}$, $u_i(y) \geq 0$) . 假设随机算法 A 满足 ϵ -差分隐私. 证明: 对于任意相邻数据集 x 和 x' , 有

$$\mathbb{E}_{y \sim A(x)}[u_i(y)] \leq e^\epsilon \mathbb{E}_{y \sim A(x')}[u_i(y)].$$

即, 差分隐私算法保证了个体效用的期望变动被限制在 e^ϵ 倍以内, 从而约束了通过操纵数据对自身效用的显著影响. 因此, 差分隐私是一个“几乎诚实”的机制.

8. 差分隐私通过对数据加入噪声来“牺牲准确性来换取隐私性”, 至于牺牲多少, 换取多少, 这是一个可调可配的过程, 因此差分隐私的理论与实际落地依然存在巨大的鸿沟. 请调研一下现实中的差分隐私:

- 2020 年美国首次使用差分隐私进行人口普查, 是否被发现隐患?
- 谷歌和苹果差分隐私算法中 ϵ 的取值是多少? 对哪些数据要用差分隐私, 是否有选择性?
- 常见的差分隐私库, 使用的时候容易犯错误, 为什么?

第三部分

决策与优化

第六章 凸分析

你上山采过蘑菇吗？尽管蘑菇是一种美味的食材，但是市场上迟到的大部分蘑菇都是人工养殖的。在中国云南，一片神奇的土地上，当地人非常喜欢食用野生生长的蘑菇，俗称“野生菌”。比起人工养殖的蘑菇，野生菌的风味要丰富和鲜美得多。

在一众野生菌中，一种叫“松茸”的种类最为珍惜名贵，松茸因其产量少、价值高而显得极其珍贵，被誉为“万菌之王”。迄今为止，松茸在世界上尚不能进行真正意义上的纯人工栽培，并且大多生长在高山林中，只能依靠人工寻找采摘。

如果让你上山去采松茸，但你对松茸一无所知，你会怎么做？你只能一点一点地摸索，爬遍整座山，对这图鉴寻找和它匹配的野生菌。这个过程非常耗时耗力，而且大概率无功而返。实际上，松茸价格昂贵就是因为它的稀有性。

然而，如果有一个向导，非常熟悉这些山，年年采摘松茸，虽然他也不知道松茸的确切位置，但是他可以凭借经验，大幅减少寻找的范围和时间。因此，在向导的帮助下，你采摘的成功率和效率都会大大提高。

乍一看，找松茸的故事也许与 AI 没有关系。然而，我们在现实中面对的很多问题都是在找松茸：我们希望找到一个预测准确率最高的机

器学习模型、一个最优的分配方案、一条最快的导航路线等等。这些问题都是本质上都是决策与优化的问题。

我们自然希望可以高效地求解这些问题，但是，正如松茸的稀缺性，我们希望找到的东西往往是“稀有”的。如果人们对问题的结构一无所知，我们只能非常低效地进行搜索，这就好比我们在山上摸索寻找松茸。然而，如果我们对问题的结构有所了解，我们就可以像拥有向导一样，利用这些信息来指导我们的搜索，从而提高效率。

本章将会建立关于决策与优化的基本理论，并指出其中一种极其重要的特例：凸优化，这一类存在良好结构的优化问题，因而可以被高效求解。特别地，凸优化关心的对象被称为凸函数和凸集，我们将初步探究凸函数和凸集的性质。

§6.1 决策与优化的基本原理

首先，我们介绍决策与优化的基本原理。我们会从统计决策理论出发，然后讨论优化是什么，在做什么，有什么最基本的事实。

§6.1.1 统计决策理论

我们考虑一个最简单的情形，统计中国大学生的平均身高。2023年，中国在校大学生有4700多万人，我们自然很难真的去测量每一个人的身高，然后算一个平均值。因此，我们只能通过抽样的方式来估计这个平均值。

我们抽取了1000名大学生，测量他们的身高，然后计算这1000名大学生的平均身高。这个平均身高就是我们的统计决策，我们的目标是希望这个平均身高能够尽可能地接近真实的平均身高。

上面简单的例子其实具有一般性. 在第三章我们讨论过, 我们可以从一堆对象中提取信息. 我们把我们所关心的集合对象称为 (随机) 总体 P . 从概率论角度看, 总体就是一个概率分布. 现在我们从总体 P 中抽取一个样本 X . 这件事情在概率论上意味着我们得到了一个随机变量 X 服从分布 P .

拿到样本之后, 我们的任务是做出好的决策.

- 决策 T 是一个依赖 X 的函数. 比如说, P 是所有大学生的身高, X 是随机抽选一个人测量的身高, 我们的决策 T 是估计大学生的平均身高.¹
- “好的决策” 指的是函数 T 能够具备某些量化指标. 其中非常常用的一个方法是通过损失函数来衡量, 它是总体 P 和决策 $T(X)$ 的函数, 即 $L(P, T(X))$. 例如, 我们可以用

$$L(P, T(X)) = |\mathbb{E}_{X \sim P}[X] - T(X)|$$

来衡量估计的平均身高和真实平均身高的差距.

注. 损失函数在不同语境下有不同称呼. 损失函数是机器学习和数理统计语境下常用的称呼. 在控制理论中以及机器学习中, 它被称为代价函数. 在经济学和金融学的风险理论中, 损失函数被称为风险函数, 它意味着个体在面对不确定的环境下所需要面对的风险. 而在优化理论中, 损失函数往往被称为目标函数, 表明所要优化的对象.

¹注意, 这里我们说的只是抽选一个人测量的身高, 而不是抽取 1000 人的身高. 如果要适配本节开头的例子, 那么总体是 1000 个大学生身高的分布, 而样本是 1000 个大学生的身高.

决策 T 的一种量化指标是最小化期望意义下的损失函数：

$$\underset{T}{\text{minimize}} \mathbb{E}_{X \sim P}[L(P, T(X))].$$

在经济学中，这一量化指标实际上是 von Neumann 和 Morgenstern 期望效用理论的具体体现，我们会在第 9.3.4 节中详细讨论。这一理论认为，个体在面对不确定的环境时，会选择最大化期望效用的决策；在这里，就是最小化期望损失的决策。

通常，我们不能知道总体 P 的分布是什么，只知道样本 X_1, \dots, X_n ，此时我们用平均损失来代替期望损失：

$$\underset{T}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n L(P, T(X_i)).$$

现在我们考虑一个非常一般的决策任务。假设我们的任务是估计函数 f ，但是我们只知道观测到的自变量 X （来自总体 P ）以及它的函数值 $Y = f(X)$ ，我们的决策是函数的估计值 \hat{f} 。在机器学习中， f 通常需要训练的模型。我们可以写出若干种损失函数：

- 平方 (L^2) 损失函数： $L(P, T(X)) = (Y - \hat{f}(X))^2$. 使用此损失函数的时候，我们要假定 f 在实数范围取值。
- L^1 损失函数： $L(P, T(X)) = |Y - \hat{f}(X)|$. 使用此损失函数的时候，我们要假定 f 在实数范围取值。
- SVM 损失函数 (hinge 损失函数)： $L(P, T(X)) = \max\{0, 1 - Y \cdot \hat{f}(X)\}$. 使用此损失函数的时候，我们一般要假定 $f(X) \in [-1, 1]$ 。
- 交叉熵损失函数： $L(P, T(X)) = CH(\hat{f}(X), Y)$. 在二分类问题中，

我们假设 Y 的取值为 0 或 1，分别代表负类和正类。函数的估计值 $\hat{f}(X)$ 输出的是样本 X 属于正类的概率，而我们把 Y 看成退化分布，此时交叉熵损失函数衡量的是 $\hat{f}(X)$ 和 Y 之间的差距。

这些损失函数会用在不同的场景之中。通常来说，机器学习中有两类问题：回归问题和分类问题。他们两个的区别主要在于：

- 回归问题中， f 取值为实数，而且通常随自变量连续变化；而分类问题中， f 只取有限多个值，他们通常被作为标签（比如这张图片是人还是青蛙）使用。
- 在回归问题中，我们通常使用平方损失函数或者 L^1 损失函数；在分类问题中，我们通常使用 SVM 损失函数或者交叉熵损失函数。

§6.1.2 优化问题与算法

现在我们从决策过渡到优化。在最简单的决策问题中，我们的目标就是找到某个 x 使得（期望）损失函数 f 最小。此时，问题的一般形式为：

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad f(x) \\ & \text{s.t.} \quad f_i(x) = 0, \quad i = 1, \dots, m, \\ & \quad g_j(x) \leq 0, \quad j = 1, \dots, n, \\ & \quad x \in \Omega. \end{aligned} \tag{6.1}$$

这里，s.t. (subject to) 之后的内容表明了 x 取值的限制，因此被称为约束。其中 $f_i(x) = 0$ 和 $g_j(x) \leq 0$ 被称为函数约束，而 $x \in \Omega$ 被称为集合约束。

通常，为了简化记号，我们会把函数约束用向量的形式表示，比如，用

$$f(x) = 0,$$

$$g(x) \leq 0.$$

来表示 (6.1) 中的函数约束。

如果和本章开头的故事对比，能最小化 f 的那个 x 就是我们要找的松茸，而我们要爬的山就是 (6.1). 不同的山自然有不同的特性，在优化这里，根据损失函数 f 、约束条件 f_i 和 g_j 的不同性质，我们可以对优化问题进行分类：

- 无约束优化：约束条件 f_i 和 g_j 实际上不存在，即 $m = n = 0$ ，并且 Ω 是全空间，比如 \mathbb{R}^n .
- 有约束优化：至少存在一个约束条件，即 $\min\{m, n\} \geq 1$ ，或者 Ω 不是全空间.
- 光滑优化：损失函数和约束条件都是可微函数.²
- 线性优化：损失函数和约束条件都是线性函数（形如 $a^\top x + b$ ）.

如此纷繁的分类，实际上是为了更好地理解优化问题的性质。不同的特殊结构会给优化带来特殊的性质，从而可以设计出更高效的算法。

下面我们看几个经典的优化例子。

²光滑这一词的含义在不同的文献中大相径庭，它可以指（连续）可微、连续可微、二次（连续）可微或者无穷次可微。

例 6.1 (最小二乘法) 给定矩阵 $A \in \mathbb{R}^{m \times n}$ 和向量 $b \in \mathbb{R}^m$, 考虑如下优化问题:

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_2^2$$

$$\text{s.t.} \quad x \in \mathbb{R}^n.$$

这个问题被称为**最小二乘法**. 目标函数可以被写为 $(Ax - b)^\top (Ax - b)$, 因此最小二乘法是一种典型的无约束光滑优化问题.

最小二乘法的解 x^* 实际上是投影解. 为了说明这一点, 我们需要基本的线性代数知识, 请参阅附录 A.

具体来说, 我们把 Ax 理解为 A 的列向量的线性组合, 于是整个优化问题可以被看成找一个向量到某个线性空间的最短距离, 即投影, 如图 6.1 所示.

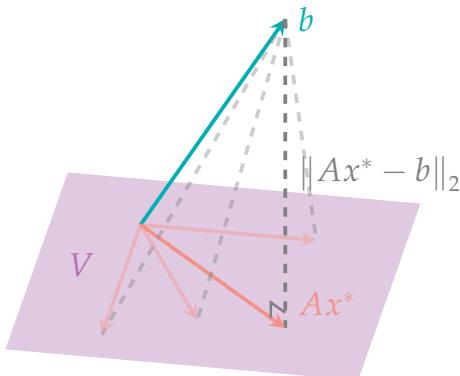


图 6.1: 最小二乘法的几何解法

我们现在从代数上解释这张图. 我们可以得到一个线性子空间

$$V = \{Ax : x \in \mathbb{R}^n\}.$$

全空间 \mathbb{R}^m 可以被分解为两个正交的子空间 V 和 V^\perp , 其中 V^\perp 是 V 的正交补:

$$\mathbb{R}^m = V \oplus V^\perp.$$

于是, b 可以被分解为 V 和 V^\perp 的分量:

$$b = b_V + b_{V^\perp}.$$

其中 b_V 可以被写作某个 Ax^* .

因此, 优化的目标可以重新写作

$$\begin{aligned}\|Ax - b\|_2^2 &= \|Ax - b_V - b_{V^\perp}\|_2^2 \\ &= \|Ax - b_V\|_2^2 + \|b_{V^\perp}\|_2^2 \\ &= \|A(x - x^*)\|_2^2 + \|b_{V^\perp}\|_2^2.\end{aligned}$$

当 $x = x^*$ 时这一目标函数取到最小值.

因此, 我们利用特殊的几何结构, 可以把最小二乘法转化为一个投影问题, 从而高效求解. 反之, 求投影也可以被写作一个优化问题. \square

例 6.2 (线性规划) 给定矩阵 $A \in \mathbb{R}^{m \times n}$ 和向量 $b \in \mathbb{R}^m$, 考虑如下优化问题:

$$\underset{x}{\text{minimize}} \quad c^T x$$

$$\text{s.t. } Ax \leq b,$$

$$x \geq 0.$$

这个问题被称为线性规划. 目标函数和约束条件都是线性的, 因此线性规划是一种典型的线性优化问题. 注意, 我们上面已经用了向量不等式的函数约束形式. \square

上面两个例子远远不能覆盖所有的优化问题, 实际上, 相当多的运筹学、机器学习和计算机科学中的问题都可以被视作(非线性)优化问题.

- **运筹学:** 线性规划、二次规划、整数规划、网络流问题、组合优化问题等.
- **金融学:** 投资组合优化、风险控制等.
- **机器学习:** 模型的训练.
- **计算机科学:** 图论中的极值问题, 例如最短路径问题、最小生成树问题等.

如果有能够解决一般优化问题的灵丹妙药, 那么将会有极其重大的意义. 然而我们后面将会看到, 一般的优化是一个难解的问题, 更严谨一点说, 不存在通用高效算法.

我们先需要明确什么是优化算法, 它应该要具备如下特征:

- 大部分优化算法都用了**迭代法**的思想: 算法 A 接受一个自变量 x , 输出一个自变量 $A(x)$, 并把它作为下一轮的输入.
- 此外, 一个算法还应该具有**通用性**, 即它必须要能解决一类优化问题 F .

- 然后，算法具备通用性就意味着它在进行**黑箱优化**： F 必须要给算法提供必要的信息来完成求解，我们将这样的提供机制抽象为**先知**，记为 \mathcal{O} . 具体来说算法输入 x 给 \mathcal{O} ， \mathcal{O} 返回一些信息给算法（例如 x 处的函数值、导数值、Hessian 矩阵）.

接下来的问题是衡量优化算法的性能好坏. 我们关注的是最坏情况，也就是说假如我们关注的是问题类 $P \subseteq F$ ，那么，我们要看的是优化算法在 P 中最差的表现如何. 衡量优化算法性能的指标有以下几个：

- 近似程度**: 我们需要求在允许误差 ϵ 的情况下的近似解. 例如，函数值不大于最优值的 ϵ ，或者离最优点距离不超过 ϵ . 考虑近似解是优化问题非常重要的一个想法，因为计算机的表示精度是有限的，我们不可能在所有情况下都求出精确解，所以求近似解是合理的要求.
- 运行时间 (收敛速度, 复杂度)**: 找到目标近似解需要调用先知的次数. 通常来说，运行时间会随近似度要求变高而变长，因此运行时间是一个关于近似程度的函数.

注. 通常来说，优化算法的执行过程中还会进行除了调用先知之外的操作，例如进行加减乘除. 然而，如果我们把所有这些操作都算入复杂度之中，算法的分析会变得非常困难，因此我们通常只考虑调用先知的次数.

这样做的合理性在于，每一次的加减乘除等额外操作，几乎都是因为调用一次先知所以才进行的，因此我们可以把这些额外操作的时间都算入先知调用的时间之中.

为了陈述“没有万能算法”这一事实，我们还需要引入一些记号.

设 F 是有限个优化问题的集合，优化的可行域是有限集，目标函数的到达域也是有限集， F 上有一个任意的概率分布.³

记号 d_t 表示 t 轮迭代之后算法产生的点列

$$(x_t(1), y_t(1)), \dots, (x_t(t), y_t(t)).$$

其中 $x_t(k)$ 是算法在第 k 轮迭代中产生的点， $y_t(k)$ 是这一步的函数值.

给定迭代轮数 t ，优化问题 f ，算法 A ，优化过程所产生的点列概率分布为

$$P(d_t | f, t, A).$$

有了上面这些记号，我们就可以将“没有万能算法”这一陈述写成定理了.

定理 6.1 (没有免费午餐定理) 对任意优化算法 A_1, A_2 ，

$$\sum_{f \in F} P(d_t | f, t, A_1) = \sum_{f \in F} P(d_t | f, t, A_2).$$

这一定理的证明可以用归纳法实现，见习题 2.

没有免费午餐定理意味着，对特定的点列，任何算法在所有实例上产生它的概率总和是一样的.

那么，点列和“没有万能算法”有什么样的关系呢？实际上，衡量算法性能的指标和点列有非常密切的联系. 比如说，算法花了 k 步找到一个 ϵ -近似解，用点列的语言来说就是，算法迭代产生的点列，长度至多是 k 并且最后一个点距离最优解距离不大于 ϵ .

³例如考虑 64 位浮点数，此时，尽管可行域和目标函数的到达域非常庞大，但都是有限集.

更一般地，算法具有某种指标的性能，就意味着算法可以产生某些特定的点列。因此，对于任何一类优化问题来说，不论以何种指标来衡量性能，优化算法在某些问题上表现出来的突出性能一定会在另一些问题上被抵消。没有一个万能的算法可以高效解决所有优化问题！

注。 在现实中，真的没有免费午餐吗？注意，没有免费午餐定理所谓的性能抵消，是条件在 f 意义上的。但是，如果我们将 f 本身的先验概率分布也纳入考虑，那么有一些算法平均的性能可能会更好。（回忆第一章中的基率谬误与 Simpson 悖论）

在现实生活中，我们需求解的优化问题都是天然有某种先验分布的，因此我们可以利用这个信息来设计更好的算法。这同样遵循着优化的基本原理：如果我们可以利用问题的结构，那么我们就可以更高效地求解问题。

§6.1.3 例子：网格搜索算法

前面对于概念的讨论可能比较抽象，所以下面我们看一个具体的例子，这个例子将会展示从算法分析的角度，优化所关注的主要问题。考虑如下优化问题：

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) \\ & \text{s.t.} && x \in [0, 1]^n. \end{aligned} \tag{6.2}$$

其中 $f(x)$ 是 Lipschitz 连续函数，即它满足

$$|f(x) - f(y)| \leq L \|x - y\|_\infty, \quad \forall x, y \in [0, 1]^n.$$

关于优化算法的假设如下. 首先, 我们可以访问零阶先知, 即 $\mathcal{O}(x) = f(x)$. 其次, 优化算法需要去找到 ϵ -近似解, 即函数值至多比最小值大 ϵ 的解.

注. 在优化中, 我们会经常使用词语“零阶”“一阶”等等, 所谓的“阶”指的是函数导数阶数, 零阶先知指的是我们可以访问函数值, 一阶先知指的是我们可以访问一阶导数, 以此类推. 后面还会有零阶条件、一阶条件等等, 他们的含义类似.

我们考虑一个非常简单的算法, 它被称为网格搜索:

- 将 $[0, 1]$ 等分成 p 份, $[0, 1] = [0, 1/p] \cup \dots \cup [(p-1)/p, 1]$.
- 遍历 $(p+1)^n$ 个格点:

$$x_{(i_1, \dots, i_n)} = \left(\frac{i_1}{p}, \dots, \frac{i_n}{p} \right)^\top,$$

$$i_k \in \{0, 1, \dots, p\}.$$

- 对每个格点询问先知得到其函数值, 输出函数值最小的一个 (记为 $(\bar{x}, f(\bar{x}))$).

这一算法和我们在山上找松茸的故事有异曲同工之妙: 我们爬山找松茸的时候, 也是将整座山分成很多小块, 然后在每个小块上找松茸, 最后找到最好的那个.

我们对于网格搜索算法问的问题是, 它的复杂度如何. 也就是说, 它需要调用先知多少次才能找到一个 ϵ -近似解? 我们从一个引理开始.

引理 6.1 设 (6.2) 的最优值为 f^* , 那么

$$f(\bar{x}) - f^* \leq \frac{L}{2p}.$$

证明. 设 x^* 是最优点, 存在一个方格包含 x^* :

$$x_{(i_1, \dots, i_n)} \leq x^* \leq x_{(i_1+1, \dots, i_n+1)}.$$

这个方格的长为 $1/p$, 所以我们可以选取方格的某个顶点 \hat{x} , 使得它的每一个轴离 x^* 的距离都不超过 $1/(2p)$, 如图 6.2 所示.

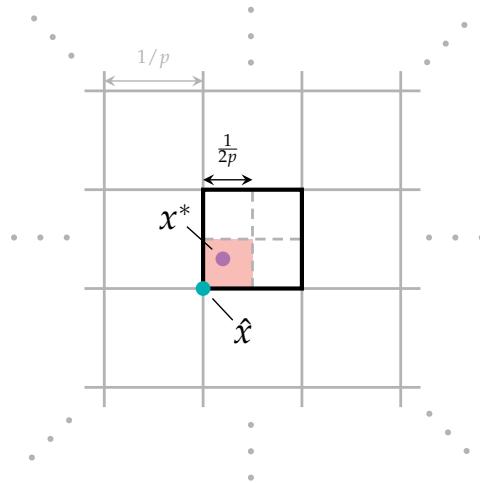


图 6.2: 网格示意图

于是根据 Lipschitz 条件,

$$f(\bar{x}) - f^* \leq f(\hat{x}) - f(x^*) \leq L \| \hat{x} - x^* \|_\infty \leq \frac{L}{2p}.$$

□

利用这个引理, 我们可以证明网格搜索算法的复杂度.

定理 6.2 网格搜索算法可以找到一个 ϵ -近似解, 其调用 \mathcal{O} 的次数

至多为

$$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2 \right)^n.$$

证明. 取 $p = \lfloor L/(2\epsilon) \rfloor + 1$, 代入引理 6.1 即可. \square

网格搜索法的运行时间给了优化问题 (6.2) 一个求解时间的上界. 然而这个上界和维数呈指数关系, 通常来说都是不可接受的复杂度.

自然地, 我们会问, (6.2) 会有更好的算法吗? 这就是下界问题. 令人惊讶的是, 对于这一个问题, 我们可以证明网格搜索法是渐近意义下最优的!

定理 6.3 设 $\epsilon < L/2$, 要想对任意 f 都能找到 (6.2) 的 ϵ -近似解, 访问 \mathcal{O} 的算法 (零阶算法) 调用 \mathcal{O} 的次数至少为

$$\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n.$$

证明. 设 $p = \lfloor L/(2\epsilon) \rfloor$, 对任意算法 A , 我们尝试构造一个函数, 使得 A 调用 \mathcal{O} 的次数小于 p^n 时最多找到一个 ϵ -近似解.

构造思路: 对任何测试点, 使得 \mathcal{O} 总是返回 0, 于是, 算法 A 只能找到 $f = 0$ 的解 \bar{x} . 然而, 这个函数的其他部分 A 都是访问不到的. 因为 A 最多可以访问 $p^n - 1$ 个点, 所以我们可以在他没有访问到的点上把这个函数往下“挖” ϵ , 这样函数的最小值就变成了 $-\epsilon$, 这就是一个 ϵ -近似解.

这一构造中, 最重要的是说明往下“挖” ϵ 是可行的. 我们来证明这一点.

同样, 我们把整个 $[0, 1]^n$ 分成 p^n 个小方格, 每个小方格的边长为 $1/p$.

因为只有 $p^n - 1$ 个测试点，所以根据鸽巢原理，网格中至少有一个长为 $1/p$ 的小方格 B 内部没有包含任何测试点。假设这个小方格的中心是 x^* ，构造

$$\bar{f}(x) = \min\{0, L \|x - x^*\|_\infty - \epsilon\}.$$

这一构造如图 6.3 所示，它就是以 x^* 为中心，往下“挖”了一个锥，最底部是 $(x^*, -\epsilon)$ 。

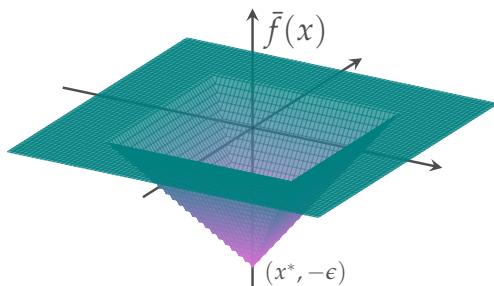


图 6.3: 函数构造示意图

容易看出， \bar{f} 是 L -Lipschitz 函数，并且最小值为 $-\epsilon$ 。

函数 \bar{f} 非零的点只在方格

$$B' = \{x \in [0, 1]^n : \|x - x^*\|_\infty \leq \epsilon/L\}$$

内部。因为 $1/(2p) \geq \epsilon/L$ ，所以 $B' \subseteq B$ 。所以所有测试点上 \mathcal{O} 都会返回 0，这是一个 ϵ -近似解。因此 A 通过小于 p^n 次对 \mathcal{O} 的调用最多只能找到 ϵ -近似解。□

以上两个结论分别给出了 (6.2) 问题的上下界，对比他们：

问题的上界：

$$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n$$

问题的下界：

$$\left\lfloor \frac{L}{2\epsilon} \right\rfloor^n$$

尽管网格搜索是一个很慢的算法，但是我们证明了，在渐近意义下，优化问题 (6.2) 的最优算法就是网格搜索！因此，我们可以说，一般的优化问题是难解的。

当我们聚焦在特定的问题类上，优化问题并不一定是难解的。因此，我们接下来的关键问题是识别出一类可以快速求解的优化问题，这就是凸函数的意义。

§6.2 凸函数

我们首先看无约束优化，看看什么样的损失函数可以快速求最小值。梯度下降方法是最古老也最常用的方法。为此，我们先回顾什么是梯度，更系统的讨论请参阅附录 B。

对于函数 $f(x)$, $x \in \mathbb{R}^n$. 梯度 $f'(x)$ ($\text{grad } f$ 或者 ∇f) 定义满足如下条件的线性函数：

$$f(x + h) - f(x) = \langle f'(x), h \rangle + o(\|h\|).$$

换言之，如果我们用一个线性函数来近似 f ，那么最优近似的系数就应该是 $f'(x)$ 。

梯度下降法的思路是非常直接的：我们每次都朝着可以让函数下降最快的方向移动一小步，这样就会逐渐找到函数的最小值。

那么，函数下降最快的是哪个方向？这等价于，我们要找一个单位

向量 h , 使得

$$\langle f'(x), h \rangle$$

最小. 不难看出, 这个方向就是梯度的反方向 $-f'(x)$. 因此, 负梯度是下降最快的方向.

梯度下降法的迭代公式如下:

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

其中 α_k 是第 k 步的步长.

与梯度下降算法相关的最小值必要条件是一阶条件.

定理 6.4 (一阶条件) 如果 x^* 是可微函数 f 的局部最小值, 那么

$$f'(x^*) = 0.$$

证明. 根据局部最小值的定义, 存在 $r > 0$, 对于任意 $\|y - x^*\| < r$, $f(y) \geq f(x^*)$. 因此

$$f(y) = f(x^*) + \langle f'(x^*), y - x^* \rangle + o(\|y - x^*\|) \geq f(x^*).$$

因此,

$$\langle f'(x^*), y - x^* \rangle + o(\|y - x^*\|) \geq 0.$$

注意, 根据内积的性质, 这等价于对任意 $s \in \mathbb{R}^n$,

$$\langle f'(x^*), s \rangle + o(\|s\|) \geq 0 \iff \left\langle f'(x^*), \frac{s}{\|s\|} \right\rangle + o(1) \geq 0.$$

沿着同一方向令 s 趋于 0, 我们得到

$$\langle f'(x^*), s \rangle \geq 0.$$

考虑方向 s 和 $-s$ 可得 $\langle f'(x^*), s \rangle = 0$. 由 s 的任意性, $f'(x^*) = 0$. \square

现在, 从一阶条件出发, 我们考虑如下优化函数类 \mathcal{F} , 满足如下三个假设:

- 假设 1: 对任意 $f \in \mathcal{F}$, 如果 x 满足一阶条件, 那么 x 是 f 的全局最小值点.
- 假设 2: 对任意 $f, g \in \mathcal{F}$, $\alpha, \beta \geq 0$, $\alpha f + \beta g \in \mathcal{F}$.
- 假设 3: 线性函数 $f(x) = \langle \alpha, x \rangle + b \in \mathcal{F}$.

假设 1 保证利用一阶条件的算法 (如梯度下降) 可以找到全局最优解.

假设 2 描述了对 \mathcal{F} 封闭的操作, 这样的操作实际上就是要求函数对线性组合封闭. 要求系数 α 和 β 非负是为了保证一阶条件得到的确实是最小值而不是最大值. 一个例子是, 如果 $x^2 \in \mathcal{F}$, 并且线性组合不限制非负系数, 那么 $-x^2 \in \mathcal{F}$, 但是后者一阶条件对应的是最大值而非最小值, 这就会与假设 1 矛盾.

假设 3 提供了 \mathcal{F} 的基本函数, 即线性函数, 这是除了常值函数之外最简单的函数, 我们应该要能够求解这一类函数.

从这三个假设出发, 我们可以给出函数类 \mathcal{F} 的刻画.

固定一个函数 $f \in \mathcal{F}$, 一个点 $x \in \mathbb{R}^n$, 定义 $\phi(y) = f(y) - \langle f'(x), y \rangle$. 根据假设 2 和假设 3, $\phi(y) \in \mathcal{F}$. $\phi'(y)|_{y=x} = f'(x) - f'(x) =$

0, 根据假设 1, x 是 ϕ 的全局最小值. 因此, $\phi(y) \geq \phi(x)$, 即

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle. \quad (6.3)$$

这一不等式给出了可微凸函数的定义:

定义 6.1 (可微凸函数) 如果可微函数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 对任意 $x, y \in \mathbb{R}^n$ 都满足 (6.3), 我们称 f 是凸函数. \square

实际上, 这一不等式有很强的几何直观, 如图 6.4 所示, 从 x 处做函数 f 的切线, 那么切线上的点都在函数下方. 从这个角度来看, 凸函数的定义是向下凸的函数.

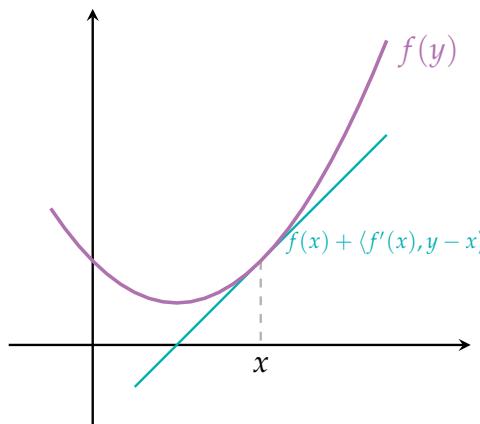


图 6.4: 凸函数的切线示意

非常有趣的是, \mathcal{F} 完全由可微凸函数组成, 这一点可以通过下面的定理得到证明.

定理 6.5 可微函数 $f \in \mathcal{F}$ 当且仅当 f 是凸函数.

证明. 只需验证满足 (6.3) 的函数属于 \mathcal{F} .

- 假设 1 令 $f'(x) = 0$ 即得任意 y 都有 $f(y) \geq f(x)$.
- 假设 2 利用内积的双线性性和导数加法公式.
- 假设 3 是平凡的. □

根据这一性质, 我们使用记号 $f \in \mathcal{F}$ 来表示 f 是可微凸函数.

下面, 我们举一些基本的可微凸函数的例子 (证明见习题 4).

例 6.3 • 对向量 x , $f(x) = \langle a, x \rangle + b$ 是凸函数.

- 对向量 x , $f(x) = \|x\|_2^2$ 是凸函数.
- 对实数 x , $f(x) = e^x$ 是凸函数.
- 对实数 x , $f(x) = \log(x)$ 是凸函数.
- 对正数 x , $f(x) = x^p$ ($p \in (-\infty, 0) \cup [1, +\infty)$) 是凸函数.
- 对正数 x , $f(x) = x^p$ ($p \in (0, 1)$) 是凸函数. □

现在, 我们给了凸性的定义, 下一步任务就是给出保持凸性不变的操作, 这样我们可以用基本函数构造出更多的函数.

假设 2 实际上已经给出了一种凸性不变的操作, 我们将它写成以下命题:

命题 6.1 对任意 $f, g \in \mathcal{F}$ 和实数 $\alpha, \beta \geq 0$, $\alpha f + \beta g \in \mathcal{F}$.

另一个可以保持凸性的操作是仿射变换可以保持凸性. 所谓仿射变换, 指的是向量空间 \mathbb{R}^n 到 \mathbb{R}^m 的映射 $x \mapsto Ax + b$, 其中 A 是 $m \times n$ 矩阵, $b \in \mathbb{R}^m$. 仿射变换实际上就是带平移的线性映射, 只是我们用变换的方式来表示它.

命题 6.2 假设函数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 属于 \mathcal{F} , 那么对任意仿射变换 $x \mapsto Ax + b$, $g(x) = f(Ax + b) \in \mathcal{F}$.

证明. $g'(x) = A^\top f'(Ax + b)$, 因此

$$\begin{aligned} g(y) &= f(Ay + b) \\ &\geq f(Ax + b) + \langle f'(Ax + b), (Ay + b) - (Ax + b) \rangle \\ &= f(Ax + b) + \langle f'(Ax + b), A(y - x) \rangle \\ &= g(x) + \langle A^\top f'(Ax + b), y - x \rangle \\ &= g(x) + \langle g'(x), y - x \rangle. \end{aligned} \quad \square$$

更多保持凸性不变的操作, 见习题 5.

凸函数的一个重要性质是 Jensen 不等式:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \quad (6.4)$$

Jensen 不等式具有很强的几何解释: 画一条 f 的割线, 那么 f 的函数图像位于割线上方, 如图 6.5 所示.

实际上, Jensen 不等式给了凸函数一种等价的定义:

定理 6.6 设 f 是可微的函数, 那么 f 满足 (6.3) 当且仅当 f 满足 (6.4).

证明. \implies : 我们需要在 (6.3) 中取恰当的 x, y , 然后试图把含梯度的项消掉. 为了消掉梯度, x 必须是固定值, 而 y 可以改变. 于是, 我们自然取 x 为 $\alpha x + (1 - \alpha)y$. 为了让 $f(x)$ 和 $f(y)$ 出现, y 的取值有如下两个考虑

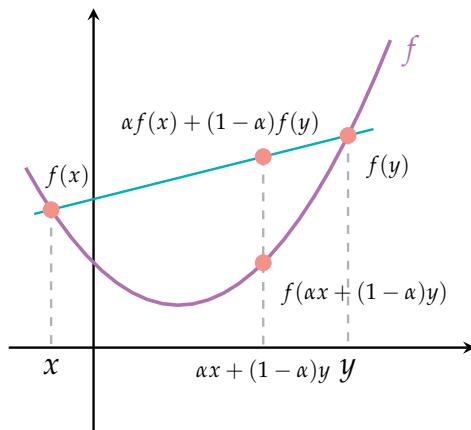


图 6.5: 凸函数的割线示意

- y 取为 x , 如此得到不等式

$$f(x) \geq f(\alpha x + (1 - \alpha)y) + \\ \langle f'(\alpha x + (1 - \alpha)y), (\alpha x + (1 - \alpha)y) - x \rangle.$$

- y 取为 y , 如此得到不等式

$$f(y) \geq f(\alpha x + (1 - \alpha)y) + \\ \langle f'(\alpha x + (1 - \alpha)y), (\alpha x + (1 - \alpha)y) - y \rangle.$$

第一个不等式内积第二项为 $(1 - \alpha)(y - x)$, 第二个不等式内积第二项为 $\alpha(x - y)$. 为把梯度消去, 把第一个不等式乘以 α , 第二个不等式乘以 $1 - \alpha$, 然后相加, 即得 (6.4).

\iff : 将 (6.4) 重写作

$$f(y) \geq \frac{1}{1-\alpha}(f(\alpha x + (1-\alpha)y) - \alpha f(x)).$$

为使不等式右边具备 (6.3) 的形式, 我们将它变形为

$$f(x) + \frac{1}{1-\alpha}(f(x + (1-\alpha)(y-x)) - f(x)).$$

令 $\alpha \rightarrow 1$, 结合梯度的定义, 即得 (6.3). □

如果函数 f 不是可微的, 那么定理 6.6 给了一个凸函数更加本质的定义:

定义 6.2 (凸函数) 函数 f 满足对任意 x, y 成立 (6.4), 那么称 f 是凸函数. □

容易证明, 之前陈述的可微凸函数的性质 (命题 6.1、命题 6.2) 对于这一更一般的凸函数定义也成立 (见习题 8).

扩展定义之后的凸函数包括了我们之前讲的 L^p ($p = 1, 2$) 损失和 SVM 损失, 以及机器学习中用到的大部分损失函数. 在实际情况中, 凸函数是一类存在快速收敛算法的函数, 例如梯度下降和 Newton 迭代法. 因此, 我们可以说, 凸函数类划定了优化问题中可以快速求解的函数类. 自此, 凸性成为了优化中的核心概念.

§6.3 凸集

接下来我们考虑约束优化问题：

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) \\ & \text{s.t.} && x \in \Omega. \end{aligned}$$

一个自然的问题是，什么样 Ω 会存在快速收敛的算法？我们将看到，凸集将会是这个问题的答案。

§6.3.1 基本定义和性质

回忆凸函数的一般定义：任意 $\alpha \in [0, 1]$ 和 $x, y \in \mathbb{R}^n$ ，

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

这里，我们隐含的要求是线段 xy 上的每一点都可以求函数值。因此，如果我们希望凸函数能够包含在带约束的优化中，一个自然的要求就是对任意 $x, y \in \Omega$ ，线段 $xy \subseteq \Omega$ 。这就是凸集的定义：

定义 6.3 (凸集) 集合 C 被称为凸集当且仅当对任意 $x, y \in C$ ，线段 xy 都在 C 内部，即

$$\{\alpha x + (1 - \alpha)y : \alpha \in [0, 1]\} \subseteq C.$$

□

我们来看一些凸集的例子：

例 6.4 • 超平面： $\{x \in \mathbb{R}^n : a^\top x = b\}$ ， $a \in \mathbb{R}^n$ ， $b \in \mathbb{R}$ 。

- 半空间: $\{x \in \mathbb{R}^n : a^\top x \geq b\}$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}$.
 - 球: $\{x \in \mathbb{R}^n : \|x - x_0\| \leq r\}$, 其中 $\|\cdot\|$ 是任意一种范数.
 - 锥: C 是一个锥指的是任意 $x, y \in C$ 和任意 $\alpha, \beta \geq 0$, $\alpha x + \beta y \in C$.
-

另外一些重要的例子是凸函数诱导的凸集. 首先是上图.

定义 6.4 (上图) 函数 f 的上图是指集合 $\text{epi}(f) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(x)\}$. 直观上说, $\text{epi}(f)$ 是位于函数 f 的图像上方的区域. □

例如, 对于函数 $f(x) = x^2$, 上图是 $\text{epi}(f) = \{(x, y) \in \mathbb{R}^2 : y \geq x^2\}$. 如图 6.6 所示.

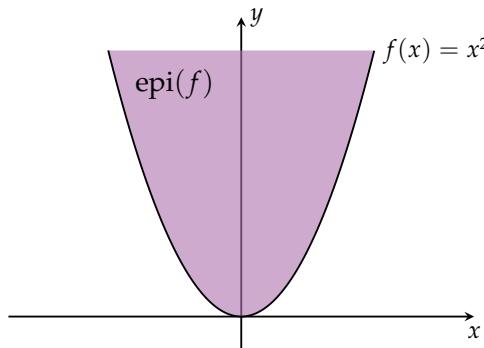


图 6.6: 函数 $f(x) = x^2$ 的上图

上图揭示了凸集与凸函数的关系:

定理 6.7 上图 $\text{epi}(f)$ 是凸集当且仅当 f 是凸函数.

证明. \implies : 只需要验证 (6.4). 取 $(x, f(x)), (y, f(y)) \in \text{epi}(f)$, 根据凸集的定义,

$$(\alpha x + (1 - \alpha)y, \alpha f(x) + (1 - \alpha)f(y)) \in \text{epi}(f),$$

所以根据上图的定义

$$\alpha f(x) + (1 - \alpha)f(y) \geq f(\alpha x + (1 - \alpha)y).$$

\Leftarrow : 取 $(x_1, y_1), (x_2, y_2) \in \text{epi}(f)$, 根据凸函数的定义,

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2).$$

根据上图的定义,

$$y_1 \geq f(x_1), \quad y_2 \geq f(x_2).$$

结合以上不等式,

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2) \leq \alpha y_1 + (1 - \alpha)y_2,$$

所以根据凸集的定义

$$(\alpha x_1 + (1 - \alpha)x_2, \alpha y_1 + (1 - \alpha)y_2) \in \text{epi}(f).$$

□

另一个联系凸函数与凸集的概念是下水平集.

定义 6.5 (下水平集) 给定 $t \in \mathbb{R}$, 函数 f 的下水平集是指集合

$$C_t(f) = \{x \in \mathbb{R}^n : f(x) \leq t\}.$$

□

直观上说, 下水平集是函数值小于 t 的区域.

命题 6.3 如果函数 f 是凸函数, 那么对任意 $t \in \mathbb{R}$, 下水平集 $C_t(f)$ 是凸集.

这个命题的证明是直接的, 见习题 9. 值得注意的是, 不同于上图, 这一命题的逆命题是不成立的, 我们在习题 9 中讨论.

接下来, 我们研究凸集的性质.

命题 6.4 凸集的任意交依然是凸集.

证明. 设 $\{C_\alpha\}_\alpha$ 是一族凸集,

$$C = \bigcap_{\alpha} C_\alpha.$$

取 $x, y \in C$, 那么对任意 α , $x, y \in C_\alpha$, 所以对任意 α , 线段 xy 都在 C_α 内部. 因此, 线段 xy 在 C 内部, 所以 C 是凸集. □

我们可以利用这个性质来构造新的凸集.

例 6.5 • 仿射空间: 有限个超平面的交, 等价地写作

$$\{x \in \mathbb{R}^n : Ax = b\},$$

其中 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

- 多面体：有限个半空间的交，等价地写作

$$\{x \in \mathbb{R}^n : Ax \leq b\},$$

其中 $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

- 单纯形：

$$\Delta_n = \{x \in \mathbb{R}^n : x_1 + \cdots + x_n = 1, \forall i, x_i \geq 0\},$$

是一种特殊的多面体.

- 凸包：给定任意集合 S , 可以定义包含它的最小凸集：

$$\bigcap_{S \subseteq C \text{ 是凸的}} C.$$

□

从优化的角度来看, 凸集本身具有最优近似性质. 我们之前在例 6.1 讨论过, 求点到线性空间的投影是一个优化问题. 任何一个点都可以唯一地投影到线性空间的某个点上, 因此整个空间通过投影就被近似到了一个线性子空间中.

现在我们来推广这一思考. 给定任意非空集合 $C \subseteq \mathbb{R}^n$, 我们尝试将整个空间近似到集合 C 中. 定义点 x 到 C 的距离为:

$$d(x, C) = \inf_{p \in C} \|x - p\|_2.$$

如果存在 $p \in C$ 达到了距离 $d(x, C)$, 我们就说 p 是 x 在 C 上的一个投影. 到当 C 就是线性空间的时候, 这个定义恰好也是原来投影的定义.

如果 \mathbb{R}^n 中的每个点都在 C 中有唯一的投影，那么就称 C 是 **Chebyshev 集**.

直观上，如果 C 是 Chebyshev 集，通过投影，我们可以用 C 中的点来近似整个空间，并且只可能有一种近似的方式.

定理 6.8 在 \mathbb{R}^n 中， C 是 Chebyshev 集当且仅当 C 是闭凸集.

这一定理的证明比较有技巧性，见习题 11.

因此，闭凸集是唯一具有良好近似性质的集合类，这又一次从优化角度说明了凸性的重要性.

§6.3.2 分离超平面定理

凸集还有一个不平凡且重要的性质：

定理 6.9 (分离超平面定理) 设 C, D 是两个非空不交凸集，也就是 $C \cap D = \emptyset$. 那么，存在 $a \neq 0$ 和 $b \in \mathbb{R}$ 使得

- 任意 $x \in C$, $a^T x \leq b$.
- 任意 $x \in D$, $a^T x \geq b$.

由 $a^T x = b$ 定义的超平面被称为 **分离超平面**.

如果两个凸集只有一个公共点，并且这个公共点是边界点，分离超平面定理依然成立，见习题 12.

注. 分离超平面定理是如此直观，以至于我们觉得他是一个显然的结论. 但是，这一定理的成立并不平凡. 比如说，如果不是在 \mathbb{R}^n 中，而是在更一般的实线性空间中，这一定理依然成立，我们称之为 *Hahn-Banach 定理*. 然而，Hahn-Banach 定理的证明需要使用选择公理（更准确说，Zorn 引

理), 这一公理的合理性至今依然是有争议的!

下面我们来证明定理 6.9.

证明. 我们有一个非常直观的几何证明. 如图 6.7 所示, 我们可以找到两个凸集离得最近的点, 然后作他们连线的中垂面, 这条中垂面就是分离超平面.

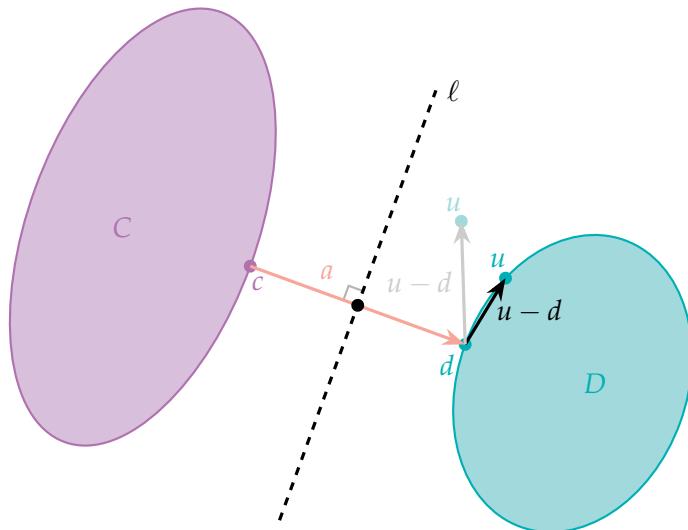


图 6.7: 分离超平面定理证明示意

下面我们严格叙述这一过程.

定义两个集合间的距离为:

$$d(\mathcal{C}, \mathcal{D}) = \inf_{x \in \mathcal{C}, y \in \mathcal{D}} \|x - y\|_2.$$

我们只证明 \mathcal{C} 和 \mathcal{D} 都是有界闭集的情况. 此时, 存在 $c \in \mathcal{C}, d \in \mathcal{D}$ 使得 $\|c - d\|_2 = d(\mathcal{C}, \mathcal{D})$. 这两个点就是 \mathcal{C} 和 \mathcal{D} 离得最近的点.

令 $\textcolor{red}{a} = \textcolor{teal}{d} - \textcolor{violet}{c}$, 这就是点 $\textcolor{violet}{c}$ 到 $\textcolor{teal}{d}$ 的向量. 接下来, 我们找中垂面的表达式.

中垂面的中点是 $(\textcolor{violet}{c} + \textcolor{teal}{d})/2$, 而它的法向量是 $\textcolor{red}{a}$, 所以中垂面的表达式是

$$\ell : \textcolor{red}{a}^\top (x - (\textcolor{violet}{c} + \textcolor{teal}{d})/2) = 0 \iff \textcolor{red}{a}^\top x - b = 0,$$

其中

$$b = (\|\textcolor{teal}{d}\|_2^2 - \|\textcolor{violet}{c}\|_2^2)/2.$$

为了说明 ℓ 是分离超平面, 只需证明 $f(x) = \textcolor{red}{a}^\top x - b$ 在 $\textcolor{violet}{C}$ 上非正, 在 $\textcolor{teal}{D}$ 上非负. 对称地, 我们只证明在 $\textcolor{teal}{D}$ 上非负.

假设对某个 $\textcolor{teal}{u} \in \textcolor{teal}{D}$, $f(\textcolor{teal}{u}) < 0$, 我们证明这将导致 $\textcolor{violet}{c}$ 和 $\textcolor{teal}{d}$ 不是离得最近的点.

注意到

$$f(x) = \langle \textcolor{red}{a}, x - \textcolor{teal}{d} \rangle + \frac{1}{2} \|\textcolor{red}{a}\|_2^2.$$

如果 $f(\textcolor{teal}{u}) < 0$, 那么

$$\langle \textcolor{red}{a}, \textcolor{teal}{u} - \textcolor{teal}{d} \rangle + \frac{1}{2} \|\textcolor{red}{a}\|_2^2 < 0 \implies \langle \textcolor{red}{a}, \textcolor{teal}{u} - \textcolor{teal}{d} \rangle < 0.$$

接下来我们说明这一不等式的几何意义. 向量 $\textcolor{red}{a}$ 和向量 $\textcolor{teal}{u} - \textcolor{teal}{d}$ 的夹角大于 90° , 因此, 如果沿着向量 $\textcolor{teal}{u} - \textcolor{teal}{d}$ 的方向从 $\textcolor{teal}{d}$ 出发, 我们可以得到一个离得更近的点, 并且根据凸集的性质, 这个点也在 $\textcolor{teal}{D}$ 中, 这与 $\textcolor{violet}{c}$ 和 $\textcolor{teal}{d}$ 是离得最近的点矛盾.

下面我们来进行具体计算说明这一几何意义. 令

$$g(t) = \|\textcolor{teal}{d} + t(\textcolor{teal}{u} - \textcolor{teal}{d}) - \textcolor{violet}{c}\|_2^2$$

$$= \|\textcolor{teal}{d} - \textcolor{violet}{c}\|_2^2 + 2t \underbrace{\langle \textcolor{teal}{u} - \textcolor{teal}{d}, \textcolor{violet}{a} \rangle}_{<0} + t^2 \|\textcolor{teal}{u} - \textcolor{teal}{d}\|_2^2.$$

根据二次函数的性质，对充分小的 t , $g(t) < g(0) = \|\textcolor{teal}{d} - \textcolor{violet}{c}\|_2^2$. 同时，因为 \mathcal{D} 是凸集， $\textcolor{teal}{d} + t(\textcolor{teal}{u} - \textcolor{teal}{d}) \in \mathcal{D}$. 以上两点说明 $\textcolor{teal}{d} + t(\textcolor{teal}{u} - \textcolor{teal}{d})$ 是离得更近的点，这与 $\textcolor{violet}{c}$ 和 $\textcolor{teal}{d}$ 是离得最近的点矛盾。 \square

§6.4 习题

1. **GPT 的损失函数.** 设 T 是所有词元的集合，GPT 的预训练过程本质上是 n 分类问题中 ($n = |T|$)，模型的输入空间是 $T^* = T \cup T^2 \cup \dots \cup T^K$, K 是模型允许的最长的输入长度. 输出（标签）空间则是所有可能的 T 上的概率分布，记为 $\Delta(T)$. 分类器 $f : T^* \rightarrow \Delta(T)$ 是一个输入到输出空间概率分布的映射，参数空间为 \mathcal{W} ，参数为 w ，输入为 x 的输出记为 $f(x; w)$.

与带有监督信号的分类任务不同，预训练过程是自监督训练：给定语料库中的一个词元序列 $x \in T^*$ ，模型会依次基于序列中的前 t 个词元来预测第 $t+1$ 个词元，并用交叉熵损失函数

$$CH(p, q) = - \sum_x p(x) \log q(x)$$

来衡量分类器的性能，其中 p 是真实标签的分布（一个以概率 1 取单点值的退化分布）， q 是分类器 f 输出的预测标签分布. 假设训练集为 $\{x^i\}_{i=1}^N \subseteq T^*$.

请将该分类器监督学习的过程写成优化问题.

2. * 证明定理 6.1.

3. * 强化学习利用 MDP 框架来描述问题，这其中最重要的一步就是定义奖励函数。在最特殊的一种强化学习中，智能体只有在最后一轮才会获得奖励。此时，智能体最大化奖励就仿佛是在做一个优化问题。

反之，如果我们把一个离散空间中的优化问题想象成一个游戏，退出游戏时的函数值就是智能体所获得的奖励，最大化目标函数就是在最大化最终奖励。因此，智能体可以通过强化学习的方法来解决这个优化问题。

基于以上观察，建模并证明强化学习和离散优化问题的等价性。

4. 验证例 6.3 中的函数是凸函数。

5. 当 $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ 是凸函数， $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ 是不减的凸函数，证明：复合函数 $f(x) = g(h(x))$ 为凸函数。

6. 如果 f 是二次可微的，那么它的二阶导数（Hessian 矩阵） $f''(x)$ 和凸函数有何关系？

7. 神经网络的损失函数一般没有凸性。我们一步一步证明这个结论：

(1) 证明：如果 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 是凸的，那么对任意 $\mathbf{v} \in \mathbb{R}^n$ ， $g(t) = f(\mathbf{x} + t\mathbf{v})$ 关于 t 是凸的。也就是说，凸函数的任意方向上的截面函数依然是凸的。

(2) 考虑只有一个隐层 \mathbf{W}_1 和输出层 \mathbf{W}_2 的神经网络 $NN(\cdot)$ ：

$$NN(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}),$$

其中矩阵 $\mathbf{W}_1 = (w_{1,ij})_{ij}$ 大小为 $m \times n$, $\mathbf{W}_2 = (w_{2,i})_i$ 大小为 $1 \times m$, $\sigma(\cdot)$ 是激活函数, 本题考虑 ReLU 函数, 即对每一维 i 都有 $\sigma_i(\mathbf{x}) = \max\{0, x_i\}$. 假设训练集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N \subseteq \mathbb{R}_{>0}^n \times \mathbb{R}_{>0}$, 其中 $\mathbf{x}_i = (x_{i,k})_k$. 请写出训练该神经网络的均方损失函数 $J(\mathbf{W}_1, \mathbf{W}_2)$.

- (3) 使用 (1) 和 (2) 的结论证明, $J(\mathbf{W}_1, \mathbf{W}_2)$ 是非凸的.
8. 对一般定义的凸函数, 验证命题 6.1 和命题 6.2.
9. 对下水平集, 完成一下习题.
 - (1) 证明命题 6.3.
 - (2) 对函数 f , 如果对任意 $t \in \mathbb{R}$, 下水平集 $C_t(f)$ 是凸集, 那么 f 依然不一定是凸函数. 请给出一个 \mathbb{R} 上定义的反例.
10. * 我们在凸函数与凸集的关系中已经看到, 通过上图, 任何凸函数都可以用凸集来给出定义, 即“凸函数是上图是凸集的函数”. 那么反过来, 能否用凸函数给出凸集的定义?
 提示: 这一问题比反问题要复杂得多, 例如, 圆是一个凸集, 但是它不是任何凸函数的上图. 因此, 上图并不是这一问题的答案.
11. * 证明定理 6.8.
12. 设 C, D 是两个非空不交凸集, C 与 D 只有一个公共点 a , 且 a 是 C 和 D 的边界点. 证明定理 6.9 对这种情况依然成立.
 提示: 为了严格证明这一点, 请首先给出边界点的定义.
13. * 梯度下降与 Bayes 信念更新的关系. 考虑机器学习的分类问题, 如果我们使用交叉熵作为损失函数, 那么, 梯度下降的每一步都相当于对

目标分布进行的 Bayes 信念更新. 请严格叙述并证明这一点. 因此, 梯度下降相当于在做信念更新.

第七章 对偶理论

在经济社会中，通常会有买家和卖家两种角色。卖家要以尽可能高的售价卖出商品，而买家则希望以尽可能低的价格购买商品。因此，卖家和买家之间构成了相互矛盾的利益关系。下面我们来看一个具体的例子。

甲用三种纸浆混合生产两种抽纸。甲的目标是让总售价最大。表 7.1 描述了公司甲用纸浆生产抽纸的信息表。

	纸浆 1	纸浆 2	纸浆 3	售价 (万元/吨)
抽纸 A	0.25	0.50	0.25	12
抽纸 B	0.50	0.50		15
库存 (吨)	120	150	50	

表 7.1: 抽纸和纸浆信息表，其中，数据的第一（二）行表示生产一吨抽纸 A (B) 需要的纸浆吨数。

设抽纸 A 和 B 分别生产 x_1 和 x_2 吨，我们可以把甲的目标写成一个

优化问题：

$$\begin{aligned}
 & \underset{x_1, x_2}{\text{maximize}} \quad z = 12x_1 + 15x_2 \\
 \text{s.t.} \quad & 0.25x_1 + 0.50x_2 \leq 120, \\
 & 0.50x_1 + 0.50x_2 \leq 150, \\
 & 0.25x_1 \leq 50, \\
 & x_1 \geq 0, \\
 & x_2 \geq 0.
 \end{aligned}$$

当然，甲也有一种选择，自己不生产销售纸巾，而是直接售卖纸浆。此时，甲变成了卖家。现在有一个公司乙需要这三种纸浆，打算向甲购买，问甲应该如何定价纸浆？

假设三种纸浆的定价分别为每吨 y_1, y_2, y_3 万元。对于买家乙来说，它希望总价格尽量小，但不能低于甲用纸浆生产抽纸所产生的价值，因此，对于乙来说，优化问题为：

$$\begin{aligned}
 & \underset{y_1, y_2, y_3}{\text{minimize}} \quad w = 120y_1 + 150y_2 + 50y_3 \\
 \text{s.t.} \quad & 0.25y_1 + 0.50y_2 + 0.25y_3 \geq 12, \\
 & 0.50y_1 + 0.50y_2 \geq 15, \\
 & y_1 \geq 0, \\
 & y_2 \geq 0, \\
 & y_3 \geq 0.
 \end{aligned}$$

假设甲乙双方都知道表 7.1 的信息，如果甲对纸浆的定价高于上述乙优化问题的最优解，那么乙会选择不购买纸浆。此时，这一市场的资源配置发生了浪费：甲有多余的纸浆，乙没有得到所需的纸浆。

在上个世纪，苏联完全实行计划经济，一个东西的售价是多少，由国家计划决定，而不是由市场决定. 我们上面的小例子就是计划经济的一个缩影：如果没有合理的定价，社会资源的配置就会出现问题，想买的买不到，想卖的卖不出去.

1959 年，苏联经济学家 Kantorovich 出版了著作《经济资源的最佳利用》，第一次将上面线性规划的这种思路引入到资源配置中. 对于一个资源配置高效的经济社会，每一个产品的定价都应该接近于它对应优化问题的最优解，这样的定价被称为影子价格.

1965 年，因 Kantorovich 因为这一工作而获列宁奖金. 1975 年，Kantorovich 因此获得了诺贝尔经济学奖，成为第一个获得这一奖项的前苏联经济学家.

在我们上面纸浆定价的例子中，我们其实看到了两个优化问题之间非同寻常的联系：一个的目标函数是另一个的约束条件. 影子价格产生于两个最优解相等的情况，正是 Kantorovich 所观察到的核心现象.

这样的现象被称为对偶性，对偶性不仅仅是线性规划中的现象，它是优化问题中的一个普遍现象. 在本章中，我们考虑带约束的优化问题. 它的一般形式是

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) \\ & \text{s.t.} && h(x) = 0, \\ & && g(x) \leq 0, \\ & && x \in \Omega. \end{aligned} \tag{7.1}$$

函数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 是目标函数， $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 和 $g : \mathbb{R}^n \rightarrow \mathbb{R}^p$ 分别是等式约束和不等式约束. 我们假设 $, h, g$ 都是连续的，且通常假设它们拥有连续的二阶导数.

一个满足所有函数约束的点 $x \in \Omega$ 被称作可行解，而使得 f 取得最小值的可行解叫做最优解。有时候优化问题的目标可能是最大化 f ，此时相应的最优解就是使得 f 取得最大值的可行解。本章的任务是讨论各种情况下最优值的必要条件，这些必要条件最终推导出了对偶理论。

§7.1 约束的几何意义

我们首先指出，优化问题的函数约束其实有很强的几何意义，更偏微积分的讨论请参见附录 B。我们先只关注 (7.1) 中的等式约束 $h(x) = 0$ ，考虑如下例子。

例 7.1 (二维空间中的约束) 考虑二维空间中的如下约束：

$$h_1(x) = x_1^2 + x_2^2 - 1 = 0,$$

$$h_2(x) = x_1 + x_2 - 1 = 0.$$

第一个约束 $h_1(x) = 0$ 定义了一个圆环，它是一维曲面¹。第二个约束 $h_2(x) = 0$ 定义了一条直线，也是一维曲面。这两个约束的交集是两个点，即零维曲面。□

例 7.2 (三维空间中的约束) 考虑三维空间中的如下约束：

$$h_1(x) = x_1^2 + x_2^2 + x_3^2 - 1 = 0,$$

$$h_2(x) = x_1 + x_2 + x_3 - 1 = 0.$$

¹严格来说，一维空间应该叫曲线。不过，为了和后面高维空间的术语保持一致，我们都称之为曲面。

第一个约束 $h_1(x) = 0$ 定义了一个球面，它是一个二维曲面。第二个约束 $h_2(x) = 0$ 定义了一个平面，也是一个二维曲面。这两个约束的交集是一个圆环，即一维曲面。 \square

我们可以从另一角度来理解这两个例子。在例 7.1 中，原本 (x_1, x_2) 两个维度都是自由选择的，所以我们可以用两个互相独立的参数来描述这个点。当加入约束 $h_1(x) = 0$ 之后，给定一个 x_1 ，我们并不能自由选择 x_2 ，而是要满足约束 $h_1(x) = 0$ 。容易看出，我们只用一个参数 θ 就可以描述这个约束下的点：

$$(x_1, x_2) = (\cos \theta, \sin \theta), \quad \theta \in [0, 2\pi).$$

所以，约束 $h_1(x) = 0$ 将原本的二维空间约束到了一维空间。继续加入约束 $h_2(x) = 0$ ，我们已经不需要参数就可以描述这个约束下的点：

$$(x_1, x_2) \in \{(0, 1), (1, 0)\}.$$

因此，约束 $h_2(x) = 0$ 将原本的一维空间约束到了零维空间。

类似地，在例 7.2 中，原本 (x_1, x_2, x_3) 可以用三个互相独立的参数来描述，当加入约束 $h_1(x) = 0$ 之后，我们只能用两个独立的参数来描述，而加入约束 $h_2(x) = 0$ 之后，我们只需要一个参数就可以描述这个约束下的点。这对应的就是三维空间被约束到了二维空间，再被约束到了一维空间。

更一般地，如果 $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ，那么 h 的每一维都对 \mathbb{R}^n 增加了一个约束，最终 $h(x) = 0$ 定义了一个 $n - m$ 维的曲面（在通常情况下）。

不过，这一性质并不是绝对的，请看下面的例子。

例 7.3 在三维空间中，考虑如下约束：

$$h_1(x) = x_1^2 + x_2^2 + x_3^2 - 1 = 0,$$

$$h_2(x) = x_1 - 1 = 0.$$

容易看出，这一约束其实对应的是一个点 $(1, 0, 0)$ ，即零维曲面，而不是我们预期的一维曲面。

再考虑如下约束：

$$h_1(x) = x_1 + x_2 + x_3 - 1 = 0,$$

$$h_2(x) = x_1 + 2x_2 + 3x_3 - 1 = 0,$$

$$h_3(x) = x_2 + 2x_3 = 0.$$

这一约束对应的是一个直线，即一维曲面，而不是我们预期的零维曲面。 \square

上面的例子是很恼人的，因为我们无法通过直观的方式来判断曲面的维数。所以，我们需要一些更强的方法来判断曲面的维数。如果 h 是具有连续的一阶导数的函数，那么这个曲面是光滑²的。我们只考虑光滑曲面，因为它们是最常见的情况。

例 7.3 的第一个约束为什么不符合预期？在点 $(1, 0, 0)$ ，球面 $h_1(x) = 0$ 的切平面恰好是 $x_1 - 1 = 0$ ，这意味着在这个点， $h_1(x) = 0$ 和 $h_2(x) = 0$ 其实只产生了一个有效的约束！这说明，“切平面”这样的概念对于维数有着至关重要的作用。

²在文献中，“光滑”这一词的含义有多种多样，例如无穷次可微、具有连续二阶导数等等。因此，这里用光滑仅仅只是方便起见，在阅读文献时，需要根据具体的上下文来理解这一词的含义。

在一般空间中，我们可以通过切空间的概念来描述曲面在某个点的维数。切空间其实是所有过该点的切线的集合。为了引入切线，我们先介绍曲线，

定义 7.1 (曲线和切向量) 考虑曲面 S ，其上的一条曲线是一系列点的集合： $x(t) \in S$ ，它们以 $t \in [0, 1]$ 为参数且在该区间上连续。因为它只有一个参数，所以它是一维曲面。

如果曲线 $x(t)$ 在点 $x^* = x(t^*)$ 处可微，那么它在该点的导数被定义为

$$\dot{x}(t) = \left. \frac{dx(t)}{dt} \right|_{t=t^*}.$$

如果曲线处处可微，我们称它是可微的。

考虑向量 v ，如果存在一个可微曲线 $x(t)$ 和常数 $k > 0$ ，使得

$$\dot{x}(0) = kv, \quad x(0) = x^*,$$

那么我们称 v 是曲面 S 在点 x^* 处的切向量。 □

有了曲线和切向量的概念，我们可以引入切空间的概念。

定义 7.2 (切空间) 考虑曲面 S ，在点 $x^* \in S$ 处的切空间是所有在该点的切向量的集合，记作 $T_{x^*}(S)$ 。 □

下面我们看一个切空间的例子。

例 7.4 (三维球面的切空间) 考虑三维空间中的单位球面

$$S^2 = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}.$$

在点 $x^* = (1, 0, 0)$ 处，球面的切空间是什么？我们可以通过曲线来描述

切空间. 考虑过 x^* 的圆弧:

$$x_\theta(t) = (\cos t, \sin t \cos \theta, \sin t \sin \theta), \quad t \in [0, \pi],$$

其中 θ 是一个固定的参数, 它表示圆弧的方向. 那么,

$$\dot{x}_\theta(0) = (0, -\sin \theta, \cos \theta),$$

所以, x^* 处的切空间至少包含以下集合

$$\{(0, -k \sin \theta, k \cos \theta) : k \in \mathbb{R}, \theta \in [0, 2\pi)\} = \{(0, y, z) : y, z \in \mathbb{R}\}.$$

因为球面是一个二维曲面, 所以切空间不可能是整个三维空间. 因此, 过 x^* 的切空间就是

$$T_{x^*}(S^2) = \{(0, y, z) : y, z \in \mathbb{R}\}. \quad \square$$

例 7.4 中的切空间是一个二维的线性空间. 直观上, 任何切空间都是应该一个线性空间, 这也是它名字的来源.

引理 7.1 切空间是一个线性空间.

尽管引理 7.1 的直观是很明显的, 但是这一性质的证明需要一定程度的微积分知识, 所以我们这里略去. 我们也需要这一性质的直观理解, 而不需要深入的数学推导.

既然切空间是一个线性空间, 我们的一个主要目标就是给出切空间的显式表达. 这一部分需要一些基本的微积分和线性代数知识, 请参阅附录 B 和附录 A.

考虑一条曲线 $x(t)$, 如果它在 $h_i(x) = 0$ 形成的曲面上, 所以

$$\forall t, h_i(x(t)) = 0 \implies \forall t, \frac{d}{dt} h_i(x(t)) = 0.$$

那么, 根据复合函数的求导法则, 应该有

$$\frac{d}{dt} h_i(x(t)) = 0 \iff \nabla_x h_i(x(t)) \dot{x}(t) = 0.$$

因此 $x(t)$ 的切向量和该点处函数 $h_i(x(t))$ 的导数向量正交.

于是, 如果 $x(t)$ 在 $h(x) = 0$ 形成的曲面上, 那么 $x(t)$ 处的导数 $\nabla h(x(t))$ 是切空间的法向量. 这一数学推导的示意图见图 7.1.

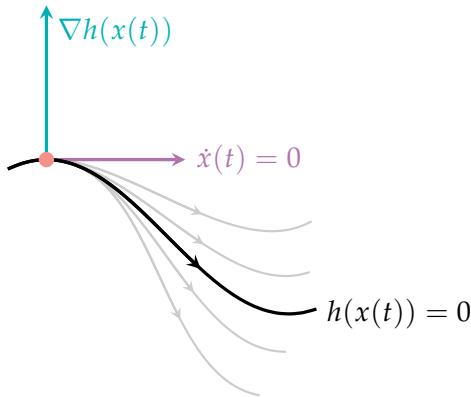


图 7.1: 切空间的法向量示意

对于例 7.4, 我们可以看到,

$$\nabla h_1(x) = (2x_1, 2x_2, 2x_3) \implies \nabla h_1(x^*) = (2, 0, 0).$$

因此, 切空间 $T_{x^*}(S^2)$ 的法向量是 $(2, 0, 0)$, 我们可以重新描述切空间为一个二维平面

$$T_{x^*}(S^2) : 2x_1 + 0x_2 + 0x_3 = 0 \iff x_1 = 0.$$

这和例 7.4 的结果是一致的.

记

$$M = \left\{ \sum_i \alpha_i \nabla h_i(x^*)^\top : \alpha_i \in \mathbb{R} \right\},$$

即 M 是 $\nabla h_i(x^*)^\top$ 张成的空间. 它的正交补是

$$M^\perp = \{y \in \mathbb{R}^n : \nabla h(x^*)y = 0\},$$

这里, $\nabla h(x^*)$ 是 h 在 x^* 处的 Jacobi 矩阵, 即对 h 的每一个分量求导得到的矩阵:

$$\nabla h(x^*) = \begin{pmatrix} \nabla h_1(x^*) \\ \nabla h_2(x^*) \\ \vdots \\ \nabla h_m(x^*) \end{pmatrix}.$$

我们已经证明 $T_{x^*}(S) \subseteq M^\perp$. 进一步, 例 7.4 的结果表明, $T_{x^*}(S) = M^\perp$, 即切空间和 M^\perp 是相等的. 然而, 如果对于例 7.3 中的第一个 h , 我们会发现切空间和 M^\perp 是不相等的: $h(x) = 0$ 对应的是单个点, 对于单个点的切空间自然是一个零维空间, 然而, 和 $\nabla h(x^*)$ 正交的空间是整个二维空间!

以上例子说明两件事, 首先, 切空间和 M^\perp 不一定相等; 其次, 切空间和 M^\perp 的关系和曲面的维数有关. 为了说明这一点, 我们引入正规

点的概念.

定义 7.3(正规点) 考虑优化问题 (7.2), 当一个点 $x^* \in \Omega$ 满足约束 $h(x^*) = 0$, 且梯度向量 $\nabla h_1(x^*), \nabla h_2(x^*), \dots, \nabla h_m(x^*)$ 线性无关时, 它被称作该约束的**正规点**. \square

直观上来说, 正规点上每一条约束都起到了实际的作用, 因此梯度向量 $\nabla h_i(x^*)^\top$ 形成了一个线性无关的集合, 张成了空间 M . 此时, 切空间恰好完全垂直于 M , 即 $T_{x^*}(S) = M^\perp$. 这一几何直观见图 7.2, 点 x^* 处的两个等式约束共同确定了该点的切空间.

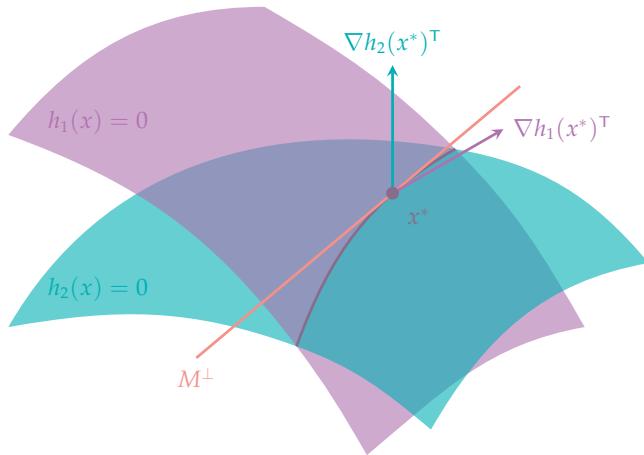


图 7.2: 正规点示意图

定理 7.1(正规点切空间刻画定理) 设曲面 $S \subseteq \mathbb{R}^n$ 由约束 $h(x) = 0$ 定义, $x^* \in S$ 是正规点, 那么,

$$T_{x^*}(S) = M^\perp = \{y : \nabla h(x^*)y = 0\}.$$

该定理的证明需要隐函数定理，对微积分要求较高，我们这里略去。

如此，针对正规点，我们找到了表达切空间的一种方法。这一方法还揭示了曲面维数和约束的梯度向量的关系。

注. 实际上，梯度向量 $\nabla h_i(x^*)$ 张成空间 M 的维数定义了曲面 S 在点 x^* 的维数。如果在点 $x^* \in S$ 一个邻域内维数都是 k ，那么，我们可以用一个 k 维的参数来描述这个邻域内的点。这一性质被称为秩定理。

§7.2 条件极值与 Lagrange 乘子法

有了切空间的准备，现在我们要对正规点推导带约束的优化问题的极值条件。我们首先考虑只有等式约束的情况：

$$\begin{aligned} &\underset{x}{\text{minimize}} && f(x) \\ &\text{s.t.} && h(x) = 0, \\ & && x \in \Omega. \end{aligned} \tag{7.2}$$

其中 f, h 都具有连续的一阶导数。

设 x^* 是一个约束 $h(x) = 0$ 一个正规点，同时也是函数 f 的一个在可行域中的极值点。这一部分的目标是得到条件极值的一阶必要条件：

定理 7.2 (条件极值的一阶必要条件) 令 x^* 是满足约束 $h(x) = 0$ 的正规点，如果 x^* 是 f 的极值点，那么存在一个 $\lambda \in \mathbb{R}^m$ 使得

$$\nabla f(x^*) + \lambda^\top \nabla h(x^*) = 0.$$

一阶必要条件 $\nabla f(x^*) + \lambda^\top \nabla h(x^*) = 0$ 以及约束 $h(x^*) = 0$ 给出

了 $n + m$ 个等式以及包含 x^*, λ 在内的 $n + m$ 个变量. 因此在非退化的
情况下，他们给出了一个唯一解.

引入与这个约束问题对应的 Lagrange 函数：

$$l(x, \lambda) = f(x) + \lambda^T h(x).$$

λ 被称为 *Lagrange 乘子*. 必要条件可以被写作：

$$\nabla_x l(x, \lambda) = 0,$$

$$\nabla_\lambda l(x, \lambda) = 0.$$

这一个求解条件极值的方法会在大部分微积分课程中给出，我们
这里的更重要的任务是给出这一方法的几何解释. 注意，定理 7.2 本
质上在说， $\nabla f(x^*)$ 是 $\nabla h_i(x^*)$ 的线性组合，所以我们的目标就是得到这
一事实.

假设 $h(x) = 0$ 形成的曲面是 S ，考虑正规极值点 $x^* \in S$. 我们任选
一条曲线 $x(t)$ 过 $x^* = x(0.5)$ ，那么， $f(x(t))$ 在 $t = 0.5$ 处取得了极小
值. 根据微积分的极值定理，我们有

$$\left. \frac{d}{dt} f(x(t)) \right|_{t=0.5} = 0 \iff \nabla f(x^*) \dot{x}(0.5) = 0.$$

因此， $\nabla f(x^*)$ 和切向量 $\dot{x}(0.5)$ 正交，因为曲线 $x(t)$ 是任意选取的，所
以 $\nabla f(x^*)$ 也和切空间 $T_{x^*}(S)$ 正交.

现在，回忆定理 7.1，我们知道切空间 $T_{x^*}(S) = M^\perp$ ，因此

$$\nabla f(x^*)^T \in (M^\perp)^\perp = M = \left\{ \sum_i \lambda_i \nabla h_i(x^*)^T : \lambda_i \in \mathbb{R} \right\}.$$

换言之， $\nabla f(x^*)$ 是 $\nabla h_i(x^*)$ 的线性组合，这就证明了定理 7.2.

最后，作为应用，我们考虑一个例子。

例 7.5 (最大熵) 考虑一个离散的概率分布，其分布列为 $p_i = \Pr(X = x_i), i = 1, \dots, n$. 该分布的熵为

$$\epsilon = - \sum_{i=1}^n p_i \log p_i.$$

该分布的均值为 $\sum_{i=1}^n x_i p_i$.

如果均值固定为 m ，求解使熵最大化的参数可以被转化成以下问题：

$$\begin{aligned} & \underset{p_i}{\text{maximize}} && - \sum_{i=1}^n p_i \log p_i \\ & \text{s.t.} && \sum_{i=1}^n p_i = 1, \\ & && \sum_{i=1}^n x_i p_i = m, \\ & && p_i \geq 0, \quad i = 1, 2, \dots, n. \end{aligned}$$

我们先忽略非负约束，假设这些约束不会被触发。引入两个 Lagrange 乘子， λ 和 μ ，则 Lagrange 函数为

$$l = \sum_{i=1}^n (-p_i \log p_i + \lambda p_i + \mu x_i p_i) - \lambda - \mu m.$$

由一阶必要条件， $-\log p_i - 1 + \lambda + \mu x_i = 0$, $i = 1, 2, \dots, n$. 因此，

$$p_i = \exp((\lambda - 1) + \mu x_i), \quad i = 1, 2, \dots, n.$$

注意 $p_i > 0$, 所以非负约束确实没有被触发. Lagrange 乘子 λ 和 μ 是两个用来保证等式约束被满足的参数. \square

§7.3 Karush-Kuhn-Tucker 条件

现在加入不等式约束, 考虑以下形式的问题:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \Omega. \end{aligned} \tag{7.3}$$

其中 f, h, g 具有连续的一阶导数.

我们将使用 Lagrange 乘子法来推导一阶必要条件. 现在, 最主要的问题在于多了不等式约束, 我们需要找到一种方法来处理这些约束.

假设 x^* 是一个极小值点, 那么, 我们可以将不等式约束 $g(x) \leq 0$ 分为两部分:

- $g_i(x^*) < 0$. 根据 g_i 的连续性, 在 x^* 的一个邻域内, 恒有 $g_i(x) < 0$, 因此这个约束在 x^* 附近一定不会违背. 我们称这样的约束为非激活约束.
- $g_i(x^*) = 0$. 如果稍微偏离 x^* , 那么 $g_i(x)$ 可能会变成正数, 因此, 这个约束在 x^* 附近是起作用的. 我们称这样的约束为激活约束.

因此, 在 x^* 的一个邻域 U 内, 如果激活的约束下标集是 J , 那么

(7.3) 可以被写作：

$$\begin{aligned}
 & \underset{x}{\text{minimize}} \quad f(x) \\
 \text{s.t.} \quad & h(x) = 0, \\
 & g_i(x) = 0, \quad i \in J, \\
 & x \in U.
 \end{aligned} \tag{7.4}$$

根据这一观察，我们可以自然地推广正规点 x^* 的定义：

定义 7.4 (正规点) 考虑优化问题 (7.3)，如果一个点 x^* 满足以下条件：

- 它是可行域中的点： $h(x^*) = 0, g(x^*) \leq 0, x \in \Omega$,
- 令 J 为满足 $g_j(x^*) = 0$ 的下标 j 的集合（激活的约束）。那么，梯度向量 $\nabla h_i(x^*), \nabla g_j(x^*), 1 \leq i \leq m, j \in J$ 是线性无关的，

那么， x^* 被称作该约束的正规点。 \square

换言之，此时的正规点不仅考虑等式约束，还要考虑起作用的（被激活的）不等式约束，这些不等式约束相当于等式约束。类似 Lagrange 乘子法，此时的一阶必要条件为：

定理 7.3 (Karush-Kuhn-Tucker 条件) 令 x^* 为优化问题 (7.3) 的正规极小值点，那么，存在向量 $\lambda \in \mathbb{R}^m$ 和向量 $\mu \in \mathbb{R}^p$ 且 $\mu \geq 0$ 使得

$$\nabla f(x^*) + \lambda^\top \nabla h(x^*) + \mu^\top \nabla g(x^*) = 0, \tag{7.5}$$

$$\mu^\top g(x^*) = 0. \tag{7.6}$$

证明. 考虑 x^* 的邻域 U , 在这个邻域内, 我们可以将问题 (7.3) 写作 (7.4), 即只考虑激活的约束. 由于 x^* 是一个极小值点, 根据定理 7.2, 存在 $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$ 使得

$$\nabla f(x^*) + \lambda^\top \nabla h(x^*) + \mu_J^\top \nabla g_J(x^*) = 0.$$

这里, $\nabla g_J(x^*) = (\nabla g_j(x^*))_{j \in J}$, 即只考虑激活的约束.

对于非激活的下标 i , 我们补充定义 $\mu_i = 0$, 于是, 上式可以被写作

$$\nabla f(x^*) + \lambda^\top \nabla h(x^*) + \mu^\top \nabla g(x^*) = 0.$$

这就得到了 (7.5).

对于被激活的下标 i , 我们有 $g_i(x^*) = 0$, 因此, $\mu_i g_i(x^*) = 0$; 对于非激活的下标 i , 我们有 $\mu_i = 0$, 因此, $\mu_i g_i(x^*) = 0$. 于是, (7.6) 得证.

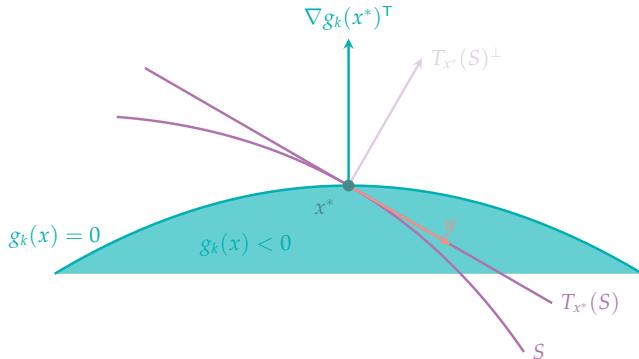
最后, 我们还需要证明 $\mu \geq 0$. 因为 x^* 是可行域内的极小值, 所以, 假设 x^* 沿着方向 y 使得恰好有一个 g_k 从激活变为非激活, 因为此时还在可行域, f 应该不变小.

下面我们来选取这样的 y . 考虑如下曲面:

$$S = \{x \in \mathbb{R}^n : h(x) = 0, g_j(x) = 0, j \in J \setminus \{k\}\}.$$

也就是除了 k 之外所有的等式约束形成的曲面. 我们从切空间 $T_{x^*}(S)$ 中选取一个 y , 使得 $\nabla g_k(x^*)y < 0$, 这样, g_k 会从激活变为非激活, 而其他约束依然得到满足. 这一选择的几何示意如图 7.3 所示.

下面我们说明为什么这样的 y 存在. 因为 x^* 是正规点, 所以根据

图 7.3: 选择 y 的几何示意

定理 7.1, $T_{x^*}(S) = M^\perp$, 其中

$$M = \left\{ \sum_i \alpha_i \nabla h_i(x^*)^\top + \sum_{j \in J \setminus \{k\}} \beta_j \nabla g_j(x^*)^\top : \alpha_i, \beta_j \in \mathbb{R} \right\}.$$

根据正规点的定义, $\nabla g_k(x^*)^\top$ 不在 M 中, 所以, $\nabla g_k(x^*)^\top$ 在 M^\perp 中的分量非零, 于是, 我们可以选择一个 $y \in M^\perp$ 使得 $\nabla g_k(x^*)y < 0$.

将 y 右乘 (7.5), 我们有

$$\nabla f(x^*)y + \lambda^\top \nabla h(x^*)y + \mu^\top \nabla g(x^*)y = 0.$$

由于 $y \in T_{x^*}(S) = M^\perp$, 我们有 $\nabla h(x^*)y = 0$, $\nabla g_j(x^*)y = 0$ ($j \in J \setminus \{k\}$), $\mu_j = 0$ ($j \notin J$) . 所以, 上式等价于

$$\nabla f(x^*)y + \mu_k \nabla g_k(x^*)y = 0.$$

令 $x(t)$ 为一条在 S 内且经过 x^* (此处 $t = 0$) 的曲线, 且有 $\dot{x}(0) = y$.
根据极小值的定义

$$0 \leq \frac{df(x(t))}{dt} \Big|_{t=0} = \nabla f(x^*)y = -\mu_k \underbrace{\nabla g_k(x^*)y}_{<0}.$$

$$\iff \mu_k \geq 0.$$

这一证明对所有激活约束的 k 都成立, 所以这就完成了证明. \square

条件 (7.5) 对应的就是 Lagrange 乘子, 而 (7.6) 则是互补松弛条件:

命题 7.1 (互补松弛条件) 对于一个优化问题 (7.3), 考虑一个正规极小值点 x^* 和对应的 Lagrange 乘子 λ, μ . 我们有以下结论:

- $\mu \geq 0$;
- 如果 $g_i(x^*) < 0$, 那么 $\mu_i = 0$;
- 如果 $\mu_i > 0$, 那么 $g_i(x^*) = 0$.

下面我们来看一个运用 KKT 条件的例子:

例 7.6 考虑问题

$$\begin{aligned} \underset{x_1, x_2}{\text{minimize}} \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 \leq 5, \\ & 3x_1 + x_2 \leq 6. \end{aligned}$$

KKT 条件为

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\mu_1 x_1 + 3\mu_2 &= 0, \\ 2x_1 + 2x_2 - 10 + 2\mu_1 x_2 + \mu_2 &= 0, \\ \mu_1(x_1^2 + x_2^2 - 5) &= 0, \\ \mu_2(3x_1 + x_2 - 6) &= 0, \\ \mu_i &\geq 0, \quad i = 1, 2. \end{aligned}$$

为了求解此类问题，我们假设一些约束被激活，然后检查所得出的 Lagrange 乘子的符号正负。在这个问题中，我们可以尝试假设有 0, 1, 2 个约束被激活。

假设第一个约束被激活，第二个约束没有被激活，得出等式

$$\begin{aligned} 4x_1 + 2x_2 - 10 + 2\mu_1 x_1 &= 0, \\ 2x_1 + 2x_2 - 10 + 2\mu_1 x_2 &= 0, \\ x_1^2 + x_2^2 &= 5. \end{aligned}$$

可得解 $x_1 = 1, x_2 = 2, \mu_1 = 1$ 。

由于 $3x_1 + x_2 = 5$ ，因此第二个约束也被满足了。因此，因为 $\mu_1 > 0$ ，我们得出结论，这个解满足一阶必要条件。□

§7.4 Lagrange 对偶

§7.4.1 原始规划与对偶规划

我们在推导 KKT 条件（定理 7.3）的时候，最终得到了如下的 Lagrange 函数：

$$L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x).$$

而 KKT 条件的第一条可以被写作

$$\nabla_x L(x, \lambda, \mu) = \nabla f(x) + \lambda^T \nabla h(x) + \mu^T \nabla g(x) = 0.$$

换言之，这是给定 λ, μ 之后 L 对 x 的一阶条件。

现在，我们不再假设 (7.3) 中的 f, h, g 具有一阶导数，只假定他们连续，此外，为简便起见，我们假设 $\Omega = \mathbb{R}^n$ ，即不考虑集合约束。我们的目标是求解

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) \\ & \text{s.t.} && h(x) = 0, \\ & && g(x) \leq 0, \\ & && x \in \mathbb{R}^n. \end{aligned} \tag{7.7}$$

我们先说明，求解这一问题可以用 Lagrange 函数重写。

命题 7.2 优化问题 (7.7) 可以被写作

$$\underset{x}{\text{minimize}} \quad \underset{\lambda, \mu \geq 0}{\sup} \quad L(x, \lambda, \mu).$$

假设它的最优值为 p^* , 那么我们有:

- 当 (7.7) 无可行解时, $p^* = +\infty$;
- 当 (7.7) 有可行解时, p^* 是 (7.7) 的最优值, 对应的 x^* 是 (7.7) 的最优解.

证明. 我们只需要证明

$$\sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) = \begin{cases} f(x), & \text{如果 } x \text{ 满足约束,} \\ +\infty, & \text{其他情况.} \end{cases}$$

当满足约束的时候,

$$h(x) = 0 \implies \lambda^T h(x) = 0,$$

$$g(x) \leq 0 \implies \mu^T g(x) \leq 0,$$

因此,

$$\sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) = L(x, \lambda, 0) = f(x).$$

当不满足约束的时候, 我们有两种情况:

- 有某个 $h_i(x) \neq 0$, 所以可以取 λ_i 使得 $\lambda_i h_i(x)$ 任意大, 于是

$$\sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) = +\infty.$$

- 有某个 $g_i(x) > 0$, 所以可以取 $\mu_i > 0$ 使得 $\mu_i g_i(x)$ 任意大, 于是

$$\sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) = +\infty.$$

这样, 我们就完成了证明. □

利用 Lagrange 函数, 我们其实将一个有约束的问题变成了无约束的问题. 特别地, 我们将优化问题转变为了原始规划的形式:

定义 7.5 (原始规划和原始函数) 优化问题 (7.7) 的原始规划是

$$\underset{x}{\text{minimize}} \quad \sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu).$$

其中,

$$\omega(x) = \sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu)$$

被称为原始函数. 原始规划的最优值记为 p^* . □

一个很自然的想法是, 我们可以把 min 和 max 的顺序交换, 这样我们就得到了对偶规划.

定义 7.6 (对偶规划和对偶函数) 优化问题 (7.7) 的对偶规划是

$$\underset{\lambda, \mu \geq 0}{\text{maximize}} \quad \inf_x L(x, \lambda, \mu).$$

其中,

$$\phi(\lambda, \mu) = \inf_x L(x, \lambda, \mu)$$

被称为对偶函数. 对偶规划的最优值记为 d^* . □

对偶函数并不是随手写出来的一个数学游戏，它有着很重要的意义。我们回到本章开头的买家卖家小问题，对于乙（即买家），我们可以把最小化买入价这个问题抽象为

$$\begin{aligned} & \underset{y}{\text{minimize}} \quad c^T y \\ & \text{s.t.} \quad Ay \geq b, \\ & \quad y \geq 0. \end{aligned} \tag{7.8}$$

它的 Lagrange 函数为

$$L(y, \mu) = c^T y - \mu_1^T (Ay - b) - \mu_2^T y,$$

它的对偶函数为

$$\phi(\mu) = \inf_y L(y, \mu) = \inf_y \{c^T y - \mu_1^T (Ay - b) - \mu_2^T y\}.$$

满足这一条件的 y 应该满足一阶条件：

$$\nabla_y L(y, \mu) = c - A^T \mu_1 - \mu_2 = 0,$$

只要确定了 μ_1 就能确定 μ_2 ，所以可以将 μ_2 消掉。将上式的 μ_2 代入 $\phi(\mu)$ ，我们有

$$\begin{aligned} \phi(\mu) &= \inf_y \{c^T y - \mu_1^T (Ay - b) - \mu_2^T y\} \\ &= c^T y - \mu_1^T (Ay - b) - \mu_2^T y \\ &= c^T y - \mu_1^T (Ay - b) - (c - A^T \mu_1)^T y \end{aligned}$$

$$= \mu_1^\top b.$$

此外，注意到 $\mu_2 \geq 0$ ，因此

$$0 \leq \mu_2 = c - A^\top \mu_1 \implies A^\top \mu_1 \leq c.$$

因此，对偶规划为

$$\begin{aligned} & \underset{\mu_1}{\text{maximize}} \quad b^\top \mu_1 \\ & \text{s.t.} \quad A^\top \mu_1 \leq c, \\ & \quad \mu_1 \geq 0. \end{aligned} \tag{7.9}$$

这正是我们在抽纸问题中甲的最大化自己纸浆售价的优化问题！

因此，我们可以想象，原始规划和对偶规划其实是买家和卖家的博弈，一个人希望最小化 L ，另一个人希望最大化 L ，这就是对偶性的一个体现。关于这一思路的详细讨论，请参阅第九章。

注。以上过程实际上给出了一个通用的方法，求一个线性规划的对偶规划。

§7.4.2 对偶的几何意义

除了从博弈角度理解对偶，类似前面几节的讨论，我们也可以从几何角度理解对偶，这一理解将最后给我们带来弱对偶定理和强对偶定理。

考虑方程

$$L(x, \lambda, \mu) = b \iff f(x) + \lambda^\top h(x) + \mu^\top g(x) = b.$$

我们暂且省略 x , 于是, 上面的式子可以被写作

$$\ell : f + \lambda^T h + \mu^T g = b.$$

如果我们固定 λ 和 μ , 那么 ℓ 定义了一个点 $(f, h^T, g^T)^T$ 形成的超平面, 这个超平面的法向量是 $(1, \lambda^T, \mu^T)^T$. 反之, 如果我们固定 $(f, h, g)^T$, 那么 ℓ 定义了一个点 $(1, \lambda^T, \mu^T)^T$ 形成的超平面, 这个超平面的法向量是 $(f, h^T, g^T)^T$.

因此, 对于 L 特定的取值, 我们有一个点-超平面对应关系: 我们可以把点重新看成超平面, 超平面重新看成点. 这就是对偶性的几何意义.

注. 实际上, 这样的想法构成了射影几何的核心. 在射影几何中, 我们把无穷远处的点加入到几何空间中, 然后研究这种几何的性质. 射影几何中, 对偶性体现如下: 任何命题, 将点和线在命题中的位置互换, 该命题仍然成立.

接下来, 我们要阐述这一几何性质如何与优化问题联系起来. 定义集合

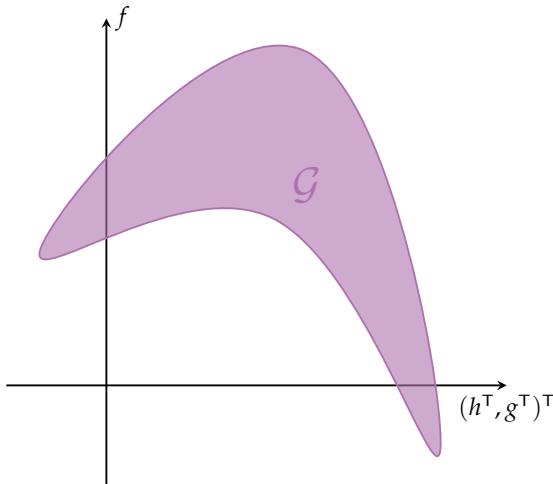
$$\mathcal{G} = \{(f(x), h(x)^T, g(x)^T)^T : x \in \mathbb{R}^n\},$$

这是所有可能的 $(f, h^T, g^T)^T$ 形成的集合. 我们可以在一个坐标系中画出这个集合, 如图 7.4 所示, 为简化起见, 我们总是忽略 $h(x)$ 的影响.

考虑 $(t, u^T, v^T)^T$ 形成的超平面

$$\alpha : t + \lambda^T u + \mu^T v = L(x, \lambda, \mu).$$

这个超平面过点 $(f(x), h(x)^T, g(x)^T)^T$, 并且法向量是 $(1, \lambda^T, \mu^T)^T$.

图 7.4: 集合 \mathcal{G} 的示意图

令 $u = v = 0$, 我们有 $t = L(x, \lambda, \mu)$, 这是 α 在 t 轴上的截距. 因此, 所有和 L 值相关的讨论都转变为了和 α 的截距相关的讨论.

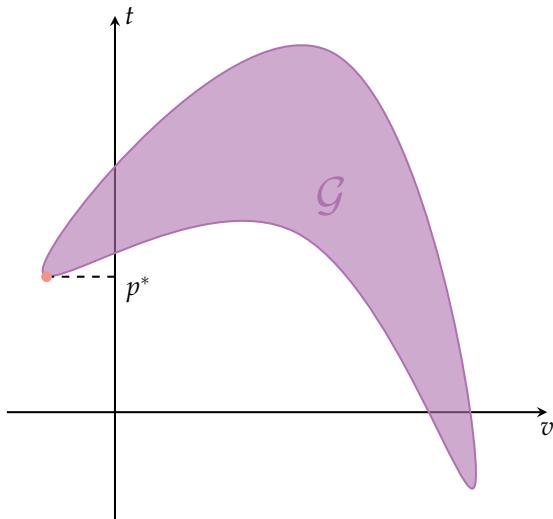
回忆原始函数的定义:

$$\omega(x) = \sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) = \sup_{\lambda, \mu \geq 0} \{f(x) + \lambda^T h(x) + \mu^T g(x)\}.$$

因此, 计算原始函数的过程其实就是, 给定一个点 $(t, u^T, v^T)^T \in \mathcal{G}$, 找到“斜率非负”的超平面, 使得截距尽可能大.

当 $h(x) = 0$ 且 $g(x) \leq 0$ 的时候, 这一截距一定在 $\mu = 0$ 的地方取到, 因此, 原始函数就是 \mathcal{G} 往 t 轴投影的值. 相应地, 原始规划的最优值 p^* 就是 \mathcal{G} 左半区域的最低点的投影, 如图 7.5 所示.

根据最低点的性质, 我们也可以把原始规划的最优值 p^* 看成是 \mathcal{G}

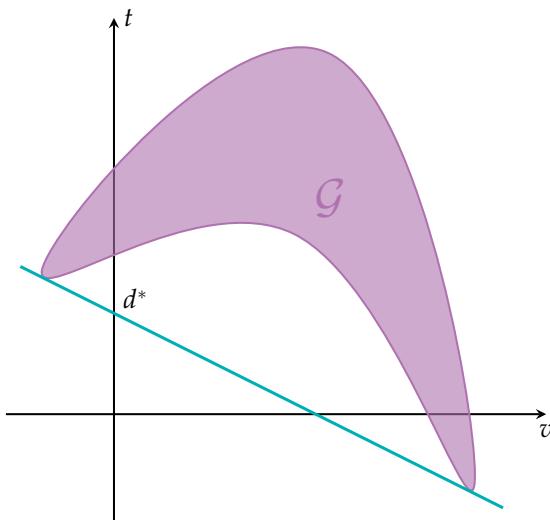
图 7.5: 原始规划的最优值 p^*

左半区域最低点切平面的截距.

那么，对偶规划是什么呢？回忆对偶函数的定义：

$$g(\lambda, \mu) = \inf_x L(x, \lambda, \mu) = \inf_{(t, u, v) \in \mathcal{G}} \{t + \lambda^\top u + \mu^\top v\}.$$

所以，这就是固定法向量 $(1, \lambda^\top, \mu^\top)^\top$ ，找到过 \mathcal{G} 且截距最小的超平面，几何上看，这一超平面是在 \mathcal{G} 最低边缘的切平面（也就是只有切点而不会“穿过” \mathcal{G} ）。相应地，对偶规划的最优值 d^* 就是这些切平面中最高的那个，如图 7.6 所示。

图 7.6: 对偶规划的最优值 d^*

§7.4.3 弱对偶定理

有了上述几何直观，我们可以阐述并证明弱对偶定理。

定理 7.4 (弱对偶定理) 对于任意优化问题 (7.7)，我们有

$$d^* \leq p^*.$$

证明. 直观上， p^* 对应的是左半区域的最低点的切平面截距，它有可能会“穿过” \mathcal{G} 的右半区域。为了让这一现象不发生，我们可以把 p^* 对应的切平面进行旋转和下移，直到它只和 \mathcal{G} 下边缘切点接触。这样，我们就得到了一个新的切平面，它的截距一定不会比 p^* 更大，因此， $d^* \leq p^*$ 。这一过程见图 7.7。

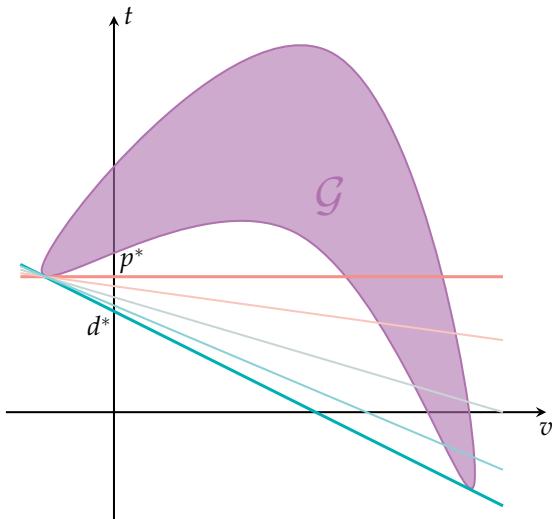


图 7.7: 弱对偶定理证明图示

下面我们来严格叙述这一点. 对 $\mu \geq 0$ 和 λ , 我们有

$$\begin{aligned} p^* &= \inf_x \sup_{\lambda, \mu \geq 0} L(x, \lambda, \mu) \\ &\geq \inf_x L(x, \lambda, \mu) \\ &= \phi(\lambda, \mu). \end{aligned}$$

这里, 第二个不等式就是在旋转切平面至法向量为 $(1, \lambda^\top, \mu^\top)^\top$, 然后使得它只和 \mathcal{G} 下边缘切点接触.

因此, $p^* \geq \phi(\lambda, \mu)$ 对所有 λ, μ 都成立, 于是也有 $p^* \geq d^*$. \square

§7.4.4 Slater 条件，强对偶定理

设原始规划对应的最低点为

$$K = (f(x^*), h(x^*)^\top, g(x^*)^\top)^\top = (p^*, u^*, v^*)^\top,$$

从图 7.7 看，如果 \mathcal{G} 完全位于过 K 的（水平）切平面上方并且 \mathcal{G} 在 t 轴左侧不为空，那么 p^* 和 d^* 一定是相等的，此时我们称之为强对偶定理。我们可以把这一条件形式化为凸规划和 *Slater* 条件。

定义 7.7 (凸规划) 对于优化问题 (7.7)，如果 $f, g_i (i = 1, \dots, p)$ 是凸函数， $h(x)$ 形如 $Ax + b$ ，那么这个问题被称为凸规划。这里， A 是一个 $m \times n$ 的矩阵， $b \in \mathbb{R}^m$. □

定义 7.8 (Slater 条件) 对于优化问题 (7.7)，如果存在一个 x 使得 $g(x) < 0$ 且 $h(x) = 0$ ，那么这个问题满足 **Slater 条件**. □

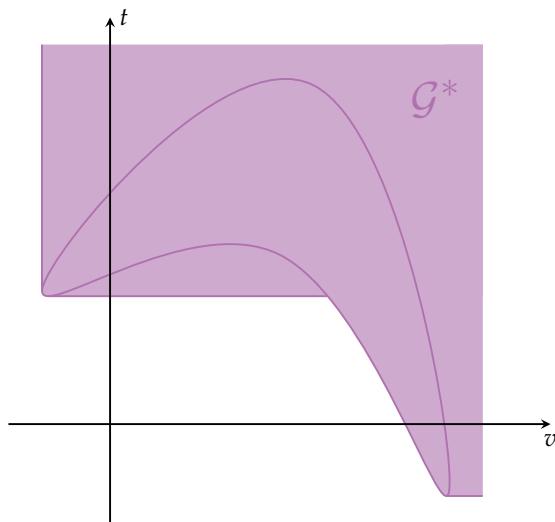
容易看出，Slater 条件意味着 \mathcal{G} 在 t 轴左侧不为空。接下来，我们要说明，满足 Slater 条件的凸规划， \mathcal{G} 完全位于过 K 的（水平）切平面上方。为此，我们定义如下集合：

$$\mathcal{G}^* = \{(t, u^\top, v^\top)^\top : \exists x \in \mathbb{R}^n, f(x) \leq t, h(x) = u, g(x) \leq v\}.$$

如图 7.8 所示， \mathcal{G}^* 是 \mathcal{G} 往右往上包络之后的集合。

我们只要说明，包络之后的集合 \mathcal{G}^* 也完全位于该切平面上方，就能说明 \mathcal{G} 完全位于过 K 的该切平面上方。下面，我们来证明这一点，从而证明强对偶定理。

首先我们证明 \mathcal{G}^* 是凸集。

图 7.8: 集合 \mathcal{G}^* 的示意图

引理 7.2 如果 (7.7) 是凸规划, 那么 \mathcal{G}^* 是凸集.

证明. 这一证明非常类似定理 6.7 的证明. 考虑系数 $\theta \in [0, 1]$ 和两个点 $(t_1, u_1^\top, v_1^\top)^\top, (t_2, u_2^\top, v_2^\top)^\top \in \mathcal{G}^*$, 我们要证明

$$\theta(t_1, u_1^\top, v_1^\top)^\top + (1 - \theta)(t_2, u_2^\top, v_2^\top)^\top \in \mathcal{G}^*. \quad (7.10)$$

根据定义, 存在 x_1, x_2 使得

$$f(x_1) \leq t_1, h(x_1) = u_1, g(x_1) \leq v_1,$$

$$f(x_2) \leq t_2, h(x_2) = u_2, g(x_2) \leq v_2.$$

令 $x = \theta x_1 + (1 - \theta)x_2$, 根据 f 和 g_i 的凸性, 我们有

$$f(x) \leq \theta f(x_1) + (1 - \theta)f(x_2) \leq \theta t_1 + (1 - \theta)t_2,$$

$$g_i(x) \leq \theta g_i(x_1) + (1 - \theta)g_i(x_2) \leq \theta v_{1,i} + (1 - \theta)v_{2,i}.$$

根据 h 的定义, 即 $h(x) = Ax + b$, 我们有

$$h(x) = \theta h(x_1) + (1 - \theta)h(x_2) = \theta u_1 + (1 - \theta)u_2.$$

这样, 我们就得到了 (7.10). 因此, \mathcal{G}^* 是凸集. \square

接下来, 我们说明 \mathcal{G}^* 完全位于过 K 的 (水平) 切平面上方, 从而证明强对偶定理.

定理 7.5 (强对偶定理) 如果 (7.7) 是凸规划并且满足 Slater 条件, 那么

$$d^* = p^*.$$

此外, 如果 p^* 有限, 那么存在 $x^* \in \mathbb{R}^n$, $\lambda^* \in \mathbb{R}^m$ 和 $\mu^* \in \mathbb{R}^p$ ($\mu^* \geq 0$) 使得

$$L(x^*, \lambda^*, \mu^*) = p^* = \sup_{\lambda, \mu \geq 0} L(x^*, \lambda, \mu) = \inf_x L(x, \lambda^*, \mu^*) = d^*. \quad (7.11)$$

证明. 如果 $p^* = -\infty$, 根据弱对偶定理, 我们有 $d^* \leq p^* = -\infty$, 所以 $d^* = -\infty = p^*$.

注意, $p^* = +\infty$ 的情况是不可能的, 因为 Slater 条件保证了至少有一个可行解.

现在假设 p^* 有限, 此时, 我们上面所描述的几何直观是有效的. 取 x^* 为原始规划的最优解, 并记

$$K = (f(x^*), h(x^*)^\top, g(x^*)^\top)^\top = (p^*, u^*, v^*)^\top.$$

设 $h(x) = Ax + b$, 我们不妨设 A 是满秩矩阵, 否则约束 $h(x) = 0$ 要么无法满足, 要么有冗余的约束.

根据引理 7.2, \mathcal{G}^* 是凸集. 我们需要选出来从 K 作出的切平面. 一个自然的选择是使用分离超平面定理 (定理 6.9). 定义另一个凸集为

$$\mathcal{C} = \{(t, 0, 0)^\top : t < p^*\}.$$

也就是一根恰好位于最低点下方的一个“杆”, 见图 7.9.

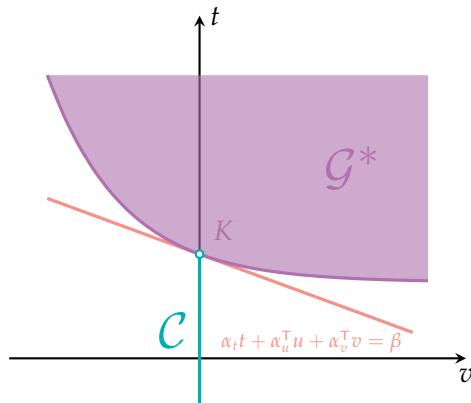


图 7.9: 分离超平面的图示

$K \in \mathcal{G}^*$, 因而不为空, \mathcal{C} 也不为空. 现在我们说明, \mathcal{G}^* 和 \mathcal{C} 是不相

交的. 假设 $(t, u^\top, v^\top)^\top \in \mathcal{G}^* \cap \mathcal{C}$, 那么存在 x 使得

$$f(x) \leq t < p^*, h(x) = u = 0, g(x) \leq v = 0,$$

这意味着 x 是一个可行解, 但是它的目标值小于 p^* , 这与 p^* 的定义矛盾. 所以, 没有这样的交点.

因此, 我们可以用分离超平面定理 (定理 6.9) 找到一个非零 $\alpha = (\alpha_t, \alpha_u^\top, \alpha_v^\top)^\top$ 和 β 使得

$$\begin{aligned}(t, u^\top, v^\top)^\top \in \mathcal{G}^* &\implies \alpha_t t + \alpha_u^\top u + \alpha_v^\top v \geq \beta, \\(t, 0, 0)^\top \in \mathcal{C} &\implies \alpha_t t \leq \beta.\end{aligned}$$

从几何上看, 这一超平面就是过 K 作的切平面.

现在我们说明 $\alpha_t \geq 0$ 并且 $\alpha_v \geq 0$. 如不然, $\alpha_t < 0$ (或者 $\alpha_v < 0$) 的话, 我们可以取一个足够大的 t (或者 v) 使得 $\alpha_t t + \alpha_u^\top u + \alpha_v^\top v$ 任意小, 这与第一个不等式矛盾.

接下来, 我们希望将这一分离超平面的系数与 Lagrange 函数对应起来, 从而和前面对偶的几何意义联系起来. 换言之, 我们希望取 $\alpha_t = 1$. 注意, α_t 是否为零决定了这一取法是否可行.

- $\alpha_t > 0$, 此时, 同时除以 α_t , 即可不妨设 $\alpha_t = 1$. 根据 \mathcal{G}^* 的定义,

$$(p^*, 0, 0)^\top \in \mathcal{G}^* \implies p^* \geq \beta.$$

另一方面, 第二个分离不等式直接得出

$$p^* \leq \beta.$$

因此, $p^* = \beta$. 所以 p^* 就是纵截距.

根据 \mathcal{G}^* 的定义, 对任意 x ,

$$\begin{aligned}(f(x), h(x)^\top, g(x)^\top)^\top \in \mathcal{G}^* &\implies f(x) + \alpha_u^\top h(x) + \alpha_v^\top g(x) \geq p^* \\ &\iff L(x, \alpha_u, \alpha_v) \geq p^*.\end{aligned}$$

因此,

$$\varphi(\alpha_u, \alpha_v) = \inf_x L(x, \alpha_u, \alpha_v) \geq p^*.$$

取 $\lambda^* = \alpha_u$ 和 $\mu^* = \alpha_v$,

$$d^* = \sup_{\lambda, \mu \geq 0} \varphi(\lambda, \mu) \geq \varphi(\lambda^*, \mu^*) \geq p^*.$$

根据弱对偶定理, $d^* \leq p^*$, 所以 $d^* = p^*$.

考虑 x^* , 因为它是原始规划的最优解, 所以

$$p^* = \sup_{\lambda, \mu \geq 0} L(x^*, \lambda, \mu) \geq L(x^*, \lambda^*, \mu^*) \geq \inf_x L(x, \lambda^*, \mu^*) \geq p^*.$$

结合 $d^* = p^*$, 我们就得到了 (7.11).

- $\alpha_t = 0$. 直观上, 此时超平面平行于 t 轴, 这意味着 \mathcal{G} 没有位于 t 轴左侧的点, 也就是 Slater 条件不成立, 这与我们的假设矛盾. 现在我们来严格说明这一点.

此时, 对任意 x ,

$$\alpha_u^\top h(x) + \alpha_v^\top g(x) \geq \beta \geq 0 \cdot t = 0,$$

选择满足 Slater 条件的 \tilde{x} , 我们有

$$\alpha_v^T g_i(\tilde{x}) \geq 0,$$

因为对所有 i , $g_i(\tilde{x}) < 0$, 同时又有 $\alpha_v \geq 0$, 所以 $\alpha_v = 0$. 因为 $\alpha \neq 0$, 所以 $\alpha_u \neq 0$. 于是, 对任意 $x \in \mathbb{R}^n$,

$$\alpha_u^T h(x) = \alpha_u^T (Ax + b) \geq 0. \quad (7.12)$$

$$\alpha_u^T (A\tilde{x} + b) = 0,$$

结合 A 是满秩矩阵, 存在 $x' \in \mathbb{R}^n$ 使得 $\alpha_u^T Ax' < 0$, 于是

$$\alpha_u^T (A(x' + \tilde{x}) + b) < 0,$$

这与 (7.12) 矛盾. 因此, 这种情况实际上是不可能的. □

注. 强对偶定理有一个特例, 即线性规划. 尽管一般而言, 线性规划不一定满足 Slater 条件, 但是只要线性规划有可行解, 它就成立强对偶定理. 线性规划的强对偶定理依赖于另外的几何直观, 利用的是 Farkas 引理, 这里就不展开了.

至此, 我们已经学习了约束优化极值的多种性质, 现总结如下:

- **Lagrange 乘子法** (定理 7.2): 适用于仅含等式约束的光滑优化问题, 给出了极值的一阶必要条件.
- **KKT 条件** (定理 7.3): 适用于同时包含等式和不等式约束的光滑优化问题, 给出了极值的一阶必要条件.

- **松弛互补条件**（命题 7.1）：适用于包含等式和不等式约束的优化问题，描述了不等式约束对应的 Lagrange 乘子的符号性质.
- **弱对偶定理**（定理 7.4）：对于任意优化问题，原始问题的最优值总是不小于对偶问题的最优值.
- **强对偶定理**（定理 7.5）：对于满足 Slater 条件的凸优化问题（或线性规划问题），原始问题的最优值等于对偶问题的最优值.

§7.5 应用：支持向量机（SVM）

作为前面极值必要条件的一个具体应用，我们考虑一个经典的机器学习分类器：支持向量机（SVM）.

考虑二分类问题，输入 $x \in \mathbb{R}^n$ ，函数 f 输出一个 $\{-1, 1\}$ 中的值. 二分类问题的学习问题指的是给定训练集 $\{(x_i, y_i)\}_{i=1}^N$ ，找到 f 使得 $f(x_i) = y_i$. 假设训练集是线性可分的，例如，存在某个 $w \in \mathbb{R}^n$ 和 $b \in \mathbb{R}$ 使得

$$f(x) = \begin{cases} 1, & w^\top x + b > 0, \\ -1, & w^\top x + b < 0. \end{cases}$$

学习问题的首要目标是找到正确的以及最优的 w 和 b . 本质上说，这就是一个找分离超平面的过程. 那么，什么才叫最优呢？从几何视角来看，一个自然的想法是最大化分离距离，即训练集中所有点到分离超平面的距离和的最小值，见图 7.10.

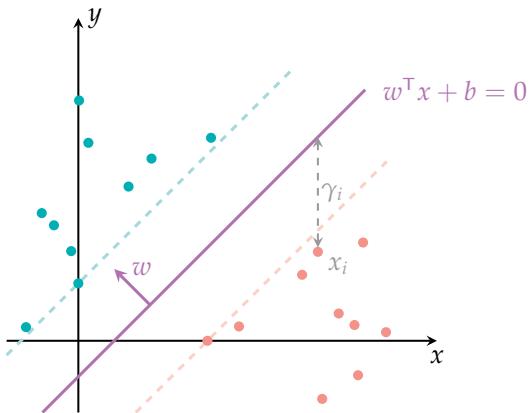


图 7.10: 分离距离示意图.

采样点 x_i 到分离超平面的归一化距离为

$$\gamma_i = y_i \left(\left(\frac{w}{\|w\|_2} \right)^\top x + \frac{b}{\|w\|_2} \right).$$

$\gamma = \min_i \gamma_i$ 是最小的归一化距离. 于是我们的任务变成了最大化 γ . 等价地, 我们求解如下优化问题

$$\begin{aligned} & \underset{w,b}{\text{maximize}} \quad \gamma \\ & \text{s.t.} \quad \gamma \leq \gamma_i, \quad i = 1, 2, \dots, N. \end{aligned}$$

$\gamma \leq \gamma_i$ 等价于

$$y_i \left(\left(\frac{w}{\gamma \|w\|_2} \right)^\top x + \frac{b}{\gamma \|w\|_2} \right) \geq 1.$$

简洁起见，把 w 替换成 $\frac{w}{\gamma\|w\|_2}$ ，把 b 替换成 $\frac{b}{\gamma\|w\|_2}$ ，我们有

$$y_i(w^\top x + b) \geq 1.$$

那么最大化 $\gamma = \frac{1}{\|w\|_2}$ 等价于最小化 $\|w\|_2^2$.

我们得到以下凸规划问题：

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq 1, \quad i = 1, 2, \dots, N. \end{aligned}$$

如何解决这个问题？利用对偶理论，我们有如下步骤（见习题 7）：

- 写出 Lagrange 乘子和对偶规划（max-min）.
- 验证 Slater 条件，于是只需要求解对偶规划.
- 利用 KKT 条件手动把对偶中的 min 消掉，得到一个二次规划.
- 用序列最小优化（SMO）等优化算法求解这个二次规划.

§7.6 习题

1. 用 Lagrange 乘子法重新证明命题 3.4.

2. 凸约束下的凸函数优化. 考虑以下问题：

- (1) 已知在无约束条件下，满足一阶条件的局部最小值点也是一阶可微凸函数的全局最小值点，即 $f'(x^*) = 0 \iff f(x) \geq f(x^*), \forall x \in$

\mathbb{R}^n . 当引入凸集约束时, 请举一个反例说明: 最小值点不一定满足一阶条件.

- (2) 假设 f 是 \mathbb{R}^n 上的可微凸函数, Q 是一个闭凸集. 证明: x^* 是 f 在 Q 上的最小值点当且仅当对任意 $x \in Q$, $\langle f'(x^*), x - x^* \rangle$.

3. 判断并说明下列情况的可行点是否都是一阶正视点.

- (1) $n = 3, m = 2, x \in \mathbb{R}^3$, 等式约束 $h_1(x) = x_1 = 0, h_2(x) = x_2 - x_3^2 = 0$.
- (2) $n = 3, m = 2, x \in \mathbb{R}^3$, 等式约束 $h_1(x) = x_1 - 2x_2 + 1 = 0, h_2(x) = -x_1 + x_2^2 = 0$.

4. 考虑以下线性约束二次规划问题:

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^\top Q x \\ \text{s.t.} \quad & Bx - b \leq 0. \end{aligned}$$

假设 Q 可逆.

- (1) 写出对偶规划, 要求形如

$$\begin{aligned} \max_\mu \quad & \varphi(\mu) \\ \text{s.t.} \quad & \mu \geq 0. \end{aligned}$$

- (2) 证明: 对偶规划的一阶条件蕴含了原始规划的松弛互补条件.

5. 一般空间中优化的弱对偶定理. 设 $f : X \times Y \rightarrow \mathbb{R}$ 有界, 记 $H = \sup_{x \in X} \inf_{y \in Y} f(x, y)$ 和 $L = \inf_{y \in Y} \sup_{x \in X} f(x, y)$. 本题证明 H 和 L 有

弱对偶的性质.

- (1) 验证“弱性”. 证明: $H \leq L$.

提示: 请直接证明, 因为此时没有写出原始规划与对偶规划.

- (2) 验证“对偶性”. 假设 X 是有限集, 把求 L 的过程写为一个优化问题, 并写出对偶规划问题, 证明 H 是对偶规划的一个可行解.

提示: 对原始规划和对偶规划的形式, 我们要求目标函数和约束中都不含 \inf , \sup , \max 或 \min .

6. 双人零和博弈的 **Nash 均衡**. 考虑一个双人零和博弈, 玩家 1 可选择 n 种行动, 玩家 2 可选择 m 种行动. 矩阵 A 中的元素 a_{ij} 表示玩家 1 选择行动 i , 玩家 2 选择行动 j 时, 玩家 1 的收益 $u_1(i, j)$. 对玩家 2 来说, 他的收益 $u_2(i, j) = -u_1(i, j)$. 玩家的混合策略是玩家行动的一个概率分布. 当玩家 1 和 2 分别采取行动 x 和 y 时, 他们的期望收益分别是 $x^T A y$ 和 $-x^T A y$. 两个玩家的目标是最大化自己的期望收益.

- (1) 玩家 1 选取策略的方式是最大化自己的最小收益, 即求 x^* 达到 $\max_x \min_y x^T A y$. 将该问题写为线性规划 P .
- (2) 玩家 2 选取策略的方式是最小化对方的最大收益, 即求 y^* 达到 $\min_y \max_x x^T A y$. 将该问题写为线性规划 P' . 证明 P' 是 P 的对偶问题.
- (3) 根据问题 (2) 的启发, 证明 minimax 定理:

$$\min_x \max_y x^T A y = \max_y \min_x x^T A y.$$

7. 补全 SVM 每一步的推导, 使之最终可以用 SMO 算法求解.

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 $f : X \rightarrow X$ ，如果元素 $a \in X$ 满足 $f(a) = a$ ，那么称 a 为映射 $f : X \rightarrow X$ 的不动点。

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 1.

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) 中的一点列 $\{x_n\}_{n \in \mathbb{N}}$, 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. \square

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 2.

例 8.2 • L^p 度量下下 \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

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

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

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

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. 注意到

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

从而

$$d(x_{n+k}, x_n) \leq d(x_n, x_{n+1}) + \cdots + d(x_{n+k-1}, x_{n+k})$$

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

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

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

□

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，这其实也就是说

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

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：

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

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 $a = f(a)$. □

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。

文献中，这一假设被称为强凸性。

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，根据定理 6.5，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) :

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

注意到 $\left\| I - \alpha \nabla^2 f(z) \right\|_2$ 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

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

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. \square

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 3).

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，且不是零矩阵，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

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

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. \square

接下来, 我们证明定理 8.6, 首先证明 A 存在一个正的特征值 $r > 0$.

引理 8.3 设 A 是一个 $n \times n$ 不可约实矩阵, 所有元素均非负, $a_{ij} \geq 0$, 则 A 存在一个正的特征值 $r > 0$.

证明. 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. □

接下来的内容主要依赖线性代数而非不动点定理, 不感兴趣的读者可以跳过.

证明. (定理 8.6 的剩余证明) 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 $\tilde{x}_0 = (\xi, 0)^T$, $\xi > 0$. 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^\top ，有

$$|\beta|x_1^\top y^* \leq x_1^\top Ay^* = r_1 x_1^\top y^*.$$

由 $x_1^\top y^* > 0$ 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

1. 验证例 8.1 中的都是度量.
2. 验证例 8.2 中的都是完备度量空间.
3. 考虑空间

$$\ell^2(\mathbb{R}) = \left\{ a = (a_1, a_2, \dots) \in \mathbb{R}^\infty : \|a\|_2 = \sqrt{\sum_i a_i^2} < +\infty \right\}.$$

定义向量 $e_i = (0, \dots, 0, 1_i, 0, \dots)$, 即第 i 位为 1、其他位为 0 的单位向量. 考虑集合 $C = \{e_1, e_2, \dots\}$ 是这些单位向量的集合. 证明:

- (1) C 是有界闭集.
- (2) 对任意不同的 $x, y \in C$, $\|x - y\|_2 = \sqrt{2}$, 因此 C 不存在任何收敛子列.
- (3) 证明: 度量空间中, 集合 F 是紧集等价于 F 中的任意无穷点列都有收敛子列. 由此证明, C 不是紧集.

4. **MDP 解的存在性.** 用 Banach 不动点定理证明 MDP 解的存在性.

提示：根据 *Bellman* 方程

$$V(s) = \max_a R(s, a) + \gamma \sum_{s' \in S} \Pr(s'|s, a) V(s')$$

构造压缩映射.

5. * 在本章引言中，曾经提到一个事实：地球上两个地方，它们分别位于地球的对径点，并且温度和湿度完全相同. 请查阅资料，找到这一现象背后的数学原理. 这一原理与不动点理论有何关系？

第四部分

博弈与逻辑

第九章 博弈与逻辑

2016 年 3 月，围棋界迎来了一场前所未有的挑战——谷歌 Deep-Mind 团队开发的人工智能 *AlphaGo* 挑战韩国围棋九段世界冠军李世乭。这场比赛不仅引发了全球的关注，更成为了人工智能领域的里程碑。

围棋被认为是极其复杂的游戏，其复杂性远远超越国际象棋，因此，许多人曾认为围棋是人工智能无法攻克的“堡垒”。然而，*AlphaGo* 改变了这一看法。

在首局比赛中，李世乭似乎还没有完全适应对手是一台超级计算机。起初，李世乭运用了传统的围棋策略，期待通过人类棋手的经验与直觉来击败 *AlphaGo*。然而，比赛后期，*AlphaGo* 展现出极其强大的计算能力，持续挖掘并利用局面中的潜在机会。最终，李世乭被逼至绝境，*AlphaGo* 成功赢得了第一局。

第二局比赛成为整个系列赛的关键点，也正是在这一局中，*AlphaGo* 下出了它最令人惊叹的一步——第 37 手。这一手棋打破了人们对围棋的传统理解，*AlphaGo* 将白子下在了一个似乎毫无意义的位置，许多围棋专家和职业棋手一度认为这是“臭棋”。李世乭一度陷入沉思，走出赛场短暂休息。

然而，随着局面的展开，这步棋逐渐展示出了它的深远战略意图，

它不仅打乱了李世乭的布局，还为 AlphaGo 赢得了巨大优势。最终，李世乭输掉了第二局，这一局被认为是 AlphaGo 表现出超越人类直觉的关键胜利。

第三局中，李世乭试图改变策略，以更加复杂、创新且进攻的方式应对 AlphaGo。然而，AlphaGo 表现得更加冷静和高效，它不仅成功化解了李世乭的进攻，还逐渐将局面转变为对自己有利的形式。在对局的后期，李世乭再次被迫认输。至此，AlphaGo 以 3:0 的比分提前赢得了这场五局比赛的胜利。

尽管前面三局失利，李世乭并没有放弃。在第四局中，他展示了超凡的创造力和直觉，走出了被称为“神之一手”的第 78 手。这一手棋打破了 AlphaGo 的计算预期，突然扭转了局面，让 AlphaGo 陷入困境。尽管 AlphaGo 做出了顽强的抵抗，但李世乭凭借这一步棋最终赢得了这一局胜利。这是人类在整个比赛中唯一的一胜。

在最后一局比赛中，李世乭保持了极高的斗志，但 AlphaGo 通过深度学习积累的经验和计算能力再次发挥作用。尽管李世乭尽力应对，但 AlphaGo 在关键时刻掌控了局面，最终赢得了第五局的胜利。整个比赛以 4:1 的结果结束，AlphaGo 取得了压倒性的胜利。

李世乭与 AlphaGo 的第四局对决，不仅是那一次比赛的唯一一次胜利，也是此后人类与顶尖围棋人工智能较量中的最后一次胜利。而第二局 AlphaGo 的神之一手，人类至今不能理解，只能效仿。AlphaGo 通过学习人类棋谱，再通过自我对弈，最终超越了人类的认知，成为了围棋的新王者。

毫无疑问，这一比赛彻底的改写了围棋的历史。过去，围棋被视为一种具有智慧和创造力的艺术；但现在，围棋选手获胜唯一的出路是模仿人工智能的策略。后来，*AlphaGo Zero* 横空出世，它完全不依赖人类

知识，但是完胜 AlphaGo. 人类积累了几千年的围棋经验，在人工智能面前显得如此渺小.

围棋代表了一种特别的决策与优化问题：我们的决策依赖于对手，而对手的决策又依赖于我们. 这样的决策问题形成了博弈论的研究对象. 博弈是如此复杂，以至于如何恰当地描述博弈的过程都是一个巨大的挑战. 本章的目标是给出博弈论的通用语言和基本概念，以及一些经典的博弈模型和他们在人工智能中的应用.

§9.1 博弈的基本语言：以井字棋为例

大家都玩过井字棋，这是一个简单的博弈. 如图 9.1 所示，在棋局中，两名玩家轮流在一个 3×3 的棋盘上放置自己的标记（X 或 O），直到有一方连成一条线（横、竖、斜）或者棋盘填满，在前一种情况下，这个玩家获胜，否则平局.

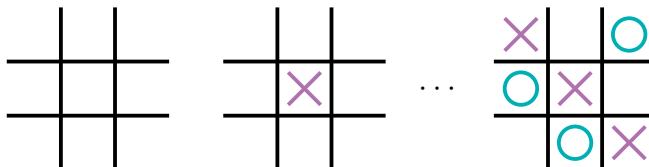


图 9.1: 井字棋示意图

上面的描述是自然语言，并不能被计算机直接理解. 我们需要将这个博弈的过程形式化，以便计算机能够理解和处理. 在井字棋中，有如下的基本概念：

- 玩家：两名玩家，一个执 X 子，另一个执 O 子.

- 棋盘局面：棋盘的当前状态，包括每个格子的占据情况（X，O 或 空）.
- 行动：每个玩家轮流在空格中放置自己的棋（X 或 O），直到出现 胜负或棋盘填满.
- 收益：游戏结束时，根据游戏的结果确定每个玩家的收益，胜者 为 +1，平局为 0，败者为 -1.

以上概念足够描述博弈是什么了. 然而，它不足以描述玩家是如何下棋的. 为此，我们需要引入策略的概念. 我们将在本章中看到，如何定义策略是博弈论中最为复杂的问题之一. 此刻，我们只关注井字棋这一场景中的策略.

我们假定玩家都有充分大的计算能力和记忆力. 于是，玩家可以记 住这一次游戏中所有的局面，以及每个轮次的自己和对手的行动. 玩家 可以知道自己的内心活动（也就是有内省的能力），但是，玩家绝不可能 知道对手的内心活动，更不知道他下一步会怎么走. 总而言之，玩家 只能知道对大家都是公开的这些信息以及自己独有的信息.

在知道所有的信息之后，玩家需要决定每一轮的走法. 或许他会猜 测对手的心理活动以及策略，并以此为根据做出自己的决策. 他也可能 完全不管对手的行动，而是我行我素. 无论如何，玩家的决策都是基于 他所知道的信息，因而我们可以认为玩家的决策是一个映射，将他知道 的信息映射到他的行动空间.

因此，玩家的策略，就是一个映射，给定当前处于哪个轮次、所有 历史局面和行动之后，它会输出下一步的行动.

需要注意的是，每个玩家在开局的时候就要选好自己的策略，此 后只能遵循这个策略进行行动. 初看之下，这一定义是极强的，我们似乎

无法在游戏中途做出调整. 然而, 这一定义其实是合理的, 因为 “调整” 本身也是策略的一部分.

例如, 一个策略可以是 “如果对手走了这一步, 那么我就走这一步; 否则, 我就走那一步”. 这其实就是调整. 策略也可以包括自我反省和对对手的猜测. 例如, 一个策略可以是 “我刚刚下的这几步棋不是很好, 我应该调整策略, 尽量避免这种情况再次发生”. 另一个策略可以是 “如果对手走了这一步, 那么我就认为他是这样的人” .

到此, 我们不仅定义了博奕的基本概念, 还定义了玩家的策略. 有了这两个概念, 我们就可以真正地让井字棋博奕进行起来了: 两名玩家根据自己的策略产生行动, 而棋盘则产生新的局面, 直到游戏结束, 然后获得收益.

接下来, 我们讨论不同类型的博奕, 以及他们对应的理论和应用.

§9.2 输赢博奕

输赢博奕指的是玩家的收益只能取两个值 (输或赢, -1 或 1) 的博奕. 输赢博奕中, 我们通常会有多轮博奕, 每轮博奕的结果会影响下一轮博奕的局面, 通常, 这种博奕被称为扩展式博奕. 围棋、象棋、斗地主都是输赢博奕.

§9.2.1 博奕的不同维度

输赢博奕有多种维度的分类方式, 见表 9.1. 这些分类都是比较直观的. 但是, 后面三个概念可能较为难以和形式化对应, 我们这里加以解释.

二人	多人
输赢	输赢平
有限深	无穷深
完全信息	非完全信息
确定性	非确定性
非合作	合作

表 9.1: 输赢博弈的分类.

- 完全信息与非完全信息：尽管这是一个直观的概念，但是如何在数学上区分完全信息与非完全信息确实极其困难的，我们这里给一种方法。

我们将博弈本身也看成一个玩家¹，那么，完全信息意味着，任何玩家可以不依赖其他玩家，自己模拟出整个博弈的进行过程。换句话说，他可以“扮演”其他任何角色。反之，非完全信息意味着，玩家不能模拟博弈，这实际上意味着他无法获取所有需要的信息来进行模拟。

- 确定性与非确定性：确定性的意思是，给定当前格局和所有玩家的行动，可以唯一确定下一回合的格局。例如，并字棋就是一个确定性博弈，因为每一步棋都会导致唯一的下一步棋局。

与之相对的概念是非确定性，比如，考虑一个非常简单的博弈。两名玩家轮流掷硬币，如果都是正面朝上，那么第一名玩家获胜，否则第二名玩家获胜。这个博弈是非确定性的，因为玩家的行动（掷

¹通常，在扩展式博弈中，我们将它称之为“天”(nature)。这里借用了中国传统文化的观念，“天”常被视为一种至高无上的力量或存在，例如“天命”和“无法无天”。

硬币) 会导致多种可能的结果.

- 非合作与合作: 在非合作博弈中, 每个玩家的决策不会被其他玩家的影响, 每个玩家都是在为自己的利益而行动. 在合作博弈中, 玩家之间可以合作, 共同制定策略, 共同获得收益. 因此, 合作博弈中的收益和策略都依赖于哪些玩家进行了合作.

注. 我们这里给出的关于完全信息的定义其实借鉴了密码学中的零知识证明的概念. 我们这里只给一个例子说明这个概念. 假设有甲乙两人, 甲宣称自己是一个硬币鉴定大师, 给任意两个硬币, 他可以判断出这两个硬币是不是一样的. 乙不确定甲是不是骗子, 所以想要验证这一能力. 而甲并不希望乙通过验证的过程学到他的鉴定方法.

于是, 我们可以这样做: 乙秘密随机准备两枚硬币, 一样或者不一样, 然后把这两枚硬币交给甲, 甲进行鉴定, 然后把硬币还给乙. 如此进行多次, 如果甲能够正确判断每一次, 那么乙就可以相信甲的能力.

如何判断是零知识? 直观上, 乙不知道除了硬币之外的任何信息, 所以他无法模拟出整个过程. 我们可以如下定义: 如果乙只知道甲有这个能力, 但是不知道甲的鉴定方法, 他依然可以把整个过程模拟出来, 那么这个过程就是零知识的.

我们给一个具体的例子.

例 9.1 如图 9.2 所示, 斗地主是一个多人有限轮非完全信息合作输赢博弈. 这个博弈有三个人, 两个农民和一个地主, 农民和地主是两个阵营. 三个人轮流出牌, 如果不能出牌, 要摸牌, 直到有一个人出完牌. 先出完牌的阵营获胜.

“多人” 是显然的, 有限轮是因为牌是有限多的, 非完全信息是因为有摸牌, 因此每个玩家只知道自己的牌, 不知道其他玩家的牌. 合作



图 9.2: 斗地主的游戏画面

是因为农民之间可以合作，地主是一个人。输赢是因为有且只有一个阵营先出完牌。 □

我们在本部分主要关注最简单的一种博弈，即完全信息确定性回合制博弈。这样的博弈可以用博弈树表示出来，例如，井字棋的博弈树可以画作图 9.3。

§9.2.2 Zermelo 定理与 AlphaGo Zero

输赢博弈一个自然的问题是：玩家是否总可以获胜？这就涉及到必胜策略的概念：无论对手如何进行行动，玩家都可以取得胜利的策略。必胜策略是一种解概念，即给定一个博弈，求解具有一定性质的玩家策略。如果某个玩家具有必胜策略，那么我们就说这个博弈是被决定的。

什么博弈是被决定的？这一问题的答案由 Zermelo 定理给出。

定理 9.1 (Zermelo 定理, Von Neumann) 如果一个博弈是双人的、有限深的、确定的、完全信息的、输赢的，那么这个博弈是被决定的。

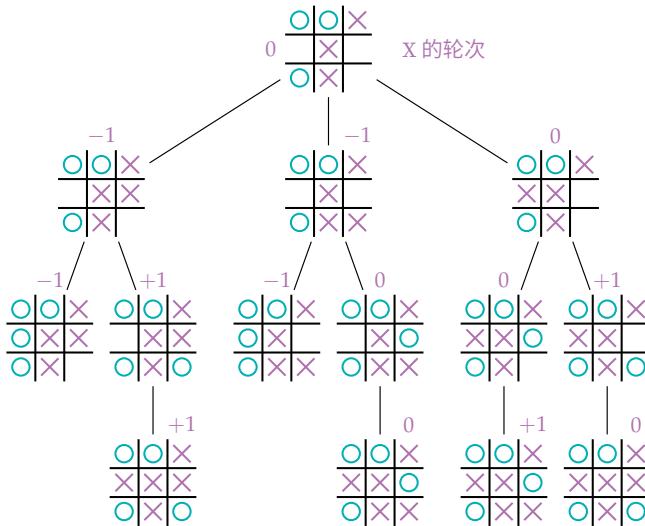


图 9.3: 井字棋的博弈树

以上限定词缺一不可，缺少了任何一个都导致结论不成立，见习题 1.

证明. (证明一：逻辑证明) 设 W_i 表示“玩家 i 获胜”， $i = 1, 2$. 于是
 $x \in W_1 \iff x \notin W_2$.

先手玩家有必胜策略当且仅当

$$\exists a_0 \forall b_0 \exists a_1 \forall b_1 \dots \exists a_n \forall b_n : (a_0 b_0 \dots a_n b_n) \in W_1.$$

后手玩家有必胜策略当且仅当

$$\forall a_0 \exists b_0 \forall a_1 \exists b_1 \dots \forall a_n \exists b_n : (a_0 b_0 \dots a_n b_n) \in W_2.$$

两个命题互为否定，因此二者恰有一个成立！

□

证明. (证明二：后向归纳法) 从博弈树的叶节点往根节点推理，见图 9.3 的收益部分。

如果此节点是玩家 i 的回合，那么往后一轮的局面已经完全确定。

- 如果有一种走法使得玩家 i 必胜，那么玩家 i 选择这种走法即可。
- 否则，玩家 i 无论如何也不可能获胜。

当到达根节点的时候，有一方有必胜策略，另一方必输。

这种证明方式被称为后向归纳法：从最后一期开始往前推理，最终确定策略。□

如果博弈的结局还有平局，我们有如下 Zermelo 定理：

定理 9.2 (有平局的 Zermelo 定理) 如果一个博弈是双人的、有限深的、确定的、完全信息的，博弈的结果有输赢平局三种，那么下面三条有且仅有一条成立：

- 第一个玩家有必胜策略。
- 第二个玩家有必胜策略。
- 双方都有不败策略。

证明见习题 2。

尽管 Zermelo 定理的第二个证明构造出了必胜策略，但是后向归纳法的搜索空间过于庞大。例如，充分大但有限的棋盘上，五子棋先手玩家存在不败策略（见习题 3），但是没有经过训练的人类或者简单的算法先手不一定会胜利。

究其原因，人的思考以及机器搜索的过程实际上是前向探索的过程。如何进行（启发式）搜索是取得胜利重要的因素。在本章开头，我

们讲述了 AlphaGo 的故事，这是正是一个很好的例子。下面我们就如何对围棋进行建模进行讨论。

由 Zermelo 定理可知，围棋也存在必胜策略。然而标准围棋棋盘大小为 19×19 ，状态空间量级为 10^{170} ，过大的状态空间使得我们无法使用后向归纳法求解出必胜策略。以下我们探讨 AlphaGo Zero（下面简称 Zero）如何通过神经网络建模博弈的过程。

首先，我们假定 Zero 可以记住的是从当前局开始往前 k 步的棋局信息（即落子方式）。我们假定这样的 k 步棋形成的棋局序列集合为 C 。于是，Zero 的策略是一个随机映射

$$\pi : C \rightarrow \Delta(\mathcal{A}),$$

其中 \mathcal{A} 是所有可能的落子方式的集合，而 $\Delta(\mathcal{A})$ 是 \mathcal{A} 上的概率分布。这里，我们假定 Zero 的策略是一个随机策略，而非确定性策略。此时，概率分布 $\pi(s)$ 表示在状态 s 下，Zero 选择对应的落子方式的似然（或者对胜利的自信程度）。

最后，当博弈结束时，Zero 会获得一定的收益，我们假定 Zero 赢的时候收益为 $+1$ ，输的时候收益为 -1 。

对于人类来说，我们的任务是让 Zero 的策略 π 尽可能地接近必胜策略，为此，我们需要用一个神经网络来拟合这个策略。此外，我们通常需要告诉 Zero 每一步棋获胜的概率（或者说期望收益），这也需要一个神经网络。具体来说，AlphaGo Zero 算法包含策略网络，价值网络和 Monte-Carlo 树搜索（MCTS）。

- 策略网络 p 和价值网络 v 的输入为当前状态 $s \in C$ ，即 $(P(s, \cdot), V(s)) = f_\theta(s)$ 。

- 策略网络 $P(s, \cdot)$ 的输出为下一步落子位置 $a \in \mathcal{A}$ 的概率分布.
- 价值网络 $V(s)$ 的输出为该状态的价值评估（期望收益、胜率）.
- MCTS 利用策略网络进行扩展，使用价值网络进行评估，利用 UCB 公式返回最优的搜索结果作为落子决策.

Zero 使用强化学习（自博弈，策略梯度）的方式训练策略网络，使用自我博弈过程中的数据监督训练价值网络. 这个过程如图 9.4 所示.

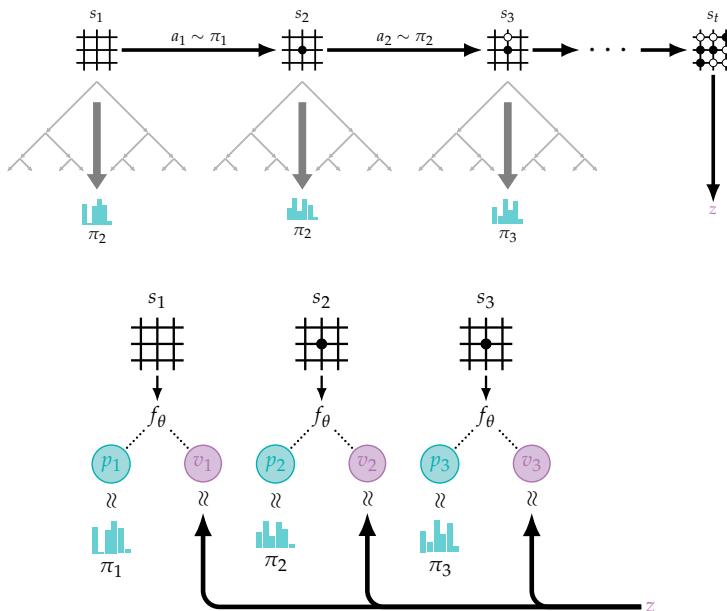


图 9.4: AlphaGo Zero 的训练过程，第一张是自我博弈的过程，第二张是神经网络训练的过程.

在开始细节之前，我们先给一个比喻，看看整个流程是如何模拟人类学习的. 假设 Zero 是一个人.

- 他可以按照他的技术水平（即启发式搜索函数）在脑中模拟一场围棋比赛，并且假定对手和自己一样聪明。最终，这一模拟会有一个结果（输或者赢），这一结果反映了他的策略的好坏。这就是用 MCTS 自博弈的过程。
- 假设他在脑中模拟了很多场比赛，他可以把这些比赛记录下来，然后根据这些记录来调整自己的策略，并调整自己对于胜率的评判。这就是策略网络和价值网络的训练过程。
- 他可以不断重复上面两个过程，来精进自己的技术水平。

下面，我们逐步解释 Zero 的训练过程。

- 自博弈过程 s_1, \dots, s_T : 在每个状态 s_t , 使用最近一次的网络 f_θ , 执行一次 MCTS α_θ (具体过程见后面)。下法根据 MCTS 计算的搜索概率 π_t 来选择, $a_t \sim \pi_t$. 最后, 依据围棋规则, 对终止状态 s_T 打分, 来计算胜利者 z .
- 神经网络训练: 使用原始的棋盘状态 s_t 作为输入, 输出 $(p_t, v_t) = f_\theta(s)$, 表示当前玩家在 s_t 的策略和胜率。训练时更新网络参数 θ , 以最大化策略 p_t 和搜索概率 π_t 的相似性, 并最小化预测赢家 v_t 与实际赢家 z 的误差。新参数将应用于下一次自博弈 a 的迭代。

MCTS 的过程较为复杂, 我们单独介绍。树的组成如下: 搜索节点是状态 s , 边是状态-行动对 (s, a) . 每条边需要存储以下信息:

- $N(s, a)$: 边的访问次数。
- $P(s, a)$: 策略网络在状态 s 中选择行动 a 的概率。

- $Q(s, a)$: 动作价值, $Q(s, a) = \frac{1}{N(s, a)} \sum_{s': s, a \rightarrow s'} V(s')$, 其中 V 是价值网络. 这一值反映了在状态 s 选择行动 a 的平均收益.

接下来, MTCS 要做如下迭代 (见图 9.5):

- 选择: 从根节点开始, 状态 s 固定, 选择具有最大的 $Q(s, a) + U(s, a)$ 的分支.
 - U 是上限置信度, $U(s, a) \propto P(s, a)/(1 + N(s, a))$.
 - $Q + U$ 是置信区间的上界, 称为 UCB 公式.

这一过程其实就是在模拟自己和对手的多轮行动, 其中选择 $Q + U$ 最大的分支即是启发式搜索的形式.

- 扩展与评估: 当选到叶节点时, 扩展叶节点. 使用神经网络 $f_\theta(s)$ 来计算新的 $P(s, a)$ 和 $V(s)$, 并把 P 存储到对应的边上. 只要还可以扩展, 就说明还有一方玩家可以继续行动, 所以这一过程可以持续到有一方获胜或者到达最大深度.
- 更新: 根据 V 更新动作价值 Q , 反映所有该动作的子树的平均值. 这反应了此次模拟的结果如何影响这一动作的评估: 赢了的话, 这一动作的评估会提高, 败了的话, 这一动作的评估会降低.
- 计算 π : 一旦搜索结束, 返回搜索概率 π , $\pi(a)$ 正比于 $N(s, a)^{1/\tau}$, τ 是一个参数, 控制着温度. 温度反映了 π 允许的随机性程度, 当 τ 趋于正无穷的时候, π 趋于均匀分布, 当 τ 趋于零的时候, π 趋于一个退化分布, 以概率 1 取最大 $N(s, a)$ 对应的 a .

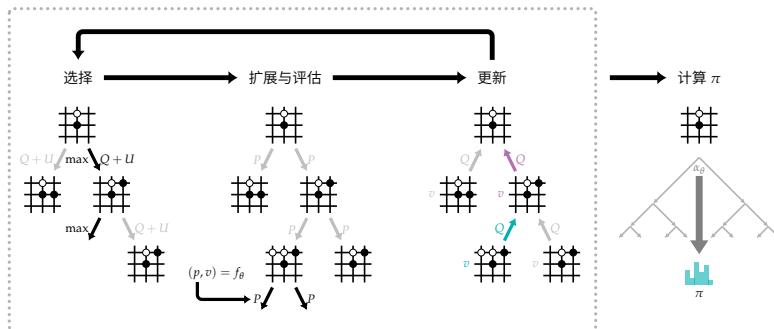


图 9.5: MCTS 的过程

§9.3 正则形式博弈

输赢博弈被建模为扩展形式博弈，它代表了多轮博弈的过程。与之相对的是正则形式博弈，即玩家只有一次行动的机会，所有玩家同时操作。正则形式博弈通常要求信息是完全的。那么，如何定义正则形式博弈和对应的解概念呢？

§9.3.1 定义

我们直接给出一个很直观的模型。

定义 9.1 (正则形式博弈) 一个正则形式博弈由以下要素构成：

- 玩家集合： I ，我们总是假设这是一个有限集合。
- 玩家的行动空间： $A_i, i \in I$.
- 玩家的收益： $u_i : \prod_j A_j \rightarrow \mathbb{R}$. □

有以下特殊的正则形式博弈：

- 当 A_i 有限，我们称之为矩阵博弈. 之所以称为矩阵博弈，是因为我们可以用一个矩阵来表示玩家的收益.

比如，考虑 $|I| = 2$ ，即有两个玩家， $|A_1| = m, |A_2| = n$ ，那么要确定 u_i 只需要确定 $u_i(a_1, a_2)$ 的值，这恰好就是一个 $m \times n$ 的矩阵. 此时，我们将第一个玩家称之为行玩家，第二个玩家称之为列玩家.

对于更多玩家的博弈，尽管 u_i 不再可以被矩阵表示，但上面的这一表示的思路依然适用，所得到的结果在线性代数上称之为张量. 但是遵循传统习惯，我们依然把这种博弈称之为矩阵博弈.

- 当 A_i 是 \mathbb{R}^n 的连通子集且和 u_i 都是连续的，我们称之为连续博弈.
- 当 $\sum_i u_i = 0$ ，我们称之为零和博弈. 一般地，如果存在常数 c 使得 $\sum_i u_i = c$ ，我们称之为常和博弈. 常和博弈和零和博弈通常具有一致的性质，所以我们也把常和博弈视为零和博弈. 如果 u_i 只有两个取值（例如 $-1, 1$ ），我们称之为输赢博弈.

下面，我们看两个典型的例子.

例 9.2 (囚徒困境) 在囚徒困境中，一共有两个玩家，行玩家和列玩家. 玩家的第一个选择是保持沉默，第二个选择是认罪并检举对方. 它有如下收益矩阵：

$$\begin{pmatrix} -1, -1 & -10, 0 \\ 0, -10 & -5, -5 \end{pmatrix}.$$

矩阵每一项第一个元素是行玩家的收益，第二个是列玩家的收益. 显然，这是一个非零和的矩阵博弈.

在这个例子中，博弈的属性非常鲜明：每个人的收益不仅仅取决于自己的选择，还取决于对方的选择。因此，玩家在做决策时，必须要考虑对方玩家可能会如何行动。□

例 9.3 (猜硬币游戏) 在猜硬币游戏中，行列玩家分别有一枚硬币，他们秘密地抛掷。如果两个玩家的硬币上面相同，行玩家获胜；否则列玩家获胜。收益矩阵为：

$$\begin{pmatrix} 1, 0 & 0, 1 \\ 0, 1 & 1, 0 \end{pmatrix}.$$

容易验证，这个博弈是一个零和博弈，进一步，是一个输赢博弈。□

至于连续博弈的例子，这里我们先略过。我们会在后面详细讨论生成对抗网络，它是一个连续博弈的例子。

现在，我们有了博弈，接下来的任务就是定义策略。假设所有人间是不能交流的，每个人独立做决策，因此玩家之间不能协调彼此的决策。所以，我们可以很简单地定义玩家 i 的策略 s_i 为行动空间 A_i 的一个元素。将所有玩家的策略收集起来，我们就得到了策略组合：

$$s = (s_i)_{i \in I}.$$

§9.3.2 理性与均衡

接下来，我们要定义正则形式博弈中的解概念，换言之，玩家最终会如何进行这场博弈并获得对应的收益。这里，我们将引入博弈论中两个重要的概念：理性和均衡。

尽管理性这一词已经被广泛使用，但是，究竟什么是理性依然是一个有巨大争议的问题。比如，我们考虑这些定义，以及让他们变得很微

妙的场景：

- 理性指的是无穷的信息获取能力. 我们把理性玩家处在一个到处都是抛硬币的世界，他怎么做才算理性？
- 理性是无穷的计算资源. 我们知道，这个世界上存在一些数学命题（记为 p ）虽然有真假，但是我们既不能证明它，也不能证伪它（Gödel 第二不完备性定理）. 现在，我们给理性玩家设计一个博弈，如果证明了 p ，他就赢，否则他就输. 这个玩家怎么做才算理性？

以上两种对于理性的理解只是冰山一角. 这里通过这些例子想说明，理性是一个非常复杂的概念，我们很难达成一个共识. 因此，更加务实的做法是，我们不去讨论理性的定义，而是直接给出一个理性的定义，然后看看这个定义下会发生什么.

均衡这一概念正是诞生于这样的背景下. 所谓均衡，指的是在某种理性的前提下，面对其他玩家，每个人都做出了他能做出的最优选择，这种最优选择被称为最优反应. 因此，所有人互相都是最优反应的时候，就构成了一个均衡.

上面的讨论总结起来就是如下的对应关系：

$$\text{理性} \iff \text{最优反应} \iff \text{均衡}.$$

博弈论的核心就在于研究以上三个概念之间的关系.

以上讨论比较抽象和一般，我们现在回到正则形式博弈，看看在这个背景下，这三个概念是如何定义的.

首先，理性的定义是非常简单的：在玩家知道其他玩家的信息（即策略）之后，他会最大化自己的收益. 然后，在这一定义下，最优反应

可以自然定义如下：

定义 9.2 (最优反应) 给定对手的策略组合 s_{-i} , 玩家 i 的最优反应是一个策略 s_i , 满足对任意 $a_i \in A_i$,

$$u_i(s_i, s_{-i}) \geq u_i(a_i, s_{-i}),$$

即玩家 i 的收益最大化. 最优反应用对应的策略集合记为 $BR_i(s_{-i})$. □

这里, s_{-i} 表示除了玩家 i 之外的所有玩家的策略组合, 我们将会频繁使用这个记号.

最后, 我们定义均衡的概念, 我们很自然有如下的定义:

定义 9.3 (纯策略 Nash 均衡) 纯策略 Nash 均衡指的是策略组合 s , 对任意玩家 i 和任意 $a_i \in A_i$, 有

$$u_i(s_i, s_{-i}) \geq u_i(a_i, s_{-i}).$$
□

我们也可以用不动点 (见第**八**章) 来理解纯策略 Nash 均衡. 纯策略 Nash 均衡的等价定义是每个人都达到了自己的最优反应, 即最优反应的不动点. 更一般地, 任何一个均衡的概念都可以理解为最优反应用对应的不动点.

作为一个例子, 我们继续考虑囚徒困境.

例 9.4 (囚徒困境的纯策略 Nash 均衡) 回忆例 9.2, 它有如下收益矩阵:

$$\begin{pmatrix} -1, -1 & -10, 0 \\ 0, -10 & -5, -5 \end{pmatrix}.$$

矩阵每一项第一个元素是行玩家的收益, 第二个是列玩家的收益.

这个博弈有唯一的纯策略 Nash 均衡：每个人都认罪（第二行），此时大家都会获得 -5 的收益。比如，行玩家选择第一行，那么，无论列玩家选了第一列还是第二列，行玩家都会选择第二行，因为这样他的收益更高。同理，列玩家也是如此。因此，这个策略组合是一个唯一的纯策略 Nash 均衡。

注意，如果两个选择都是保持沉默，那么他们的收益会更高 (-1) ，但他们却无法达到这个收益。这正是我们定义的理性所蕴含的意义：在博弈中，每个人都是自私的，如果知道了对方的选择，他们会选择最优的策略，而不会稍微放弃一点自己的利益，以换取博弈双方更大的利益。□

然而，我们继续猜硬币游戏，我们会发现这个博弈没有纯策略 Nash 均衡。回忆，这一博弈的收益矩阵为：

$$\begin{pmatrix} 1, 0 & 0, 1 \\ 0, 1 & 1, 0 \end{pmatrix}.$$

我们现在验证，这个博弈没有纯策略 Nash 均衡。如果行玩家选择第一行，那么列玩家的最优反应是第二列，然而，此时行玩家的最优反应是第二行；而行玩家选择第二行的情况类似。因此，无论如何选，最优反应都可能形成一个不动点（即均衡）。

更一般地，二人正则形式输赢博弈中纯策略 Nash 均衡往往不存在。我们有如下定理：

定理 9.3 设 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$ 是一个二人正则形式输赢博弈，其中 $I = \{1, 2\}$ 。那么， G 存在纯策略 Nash 均衡当且仅当其中一个玩家存在必胜策略。

证明见习题 7.

对比动态博弈中的 Zermelo 定理，静态的二人完全信息输赢博弈已经不能够保证必胜策略的存在性. 因此，静态输赢博弈的结局往往比动态输赢博弈更加不确定.

§9.3.3 生成对抗网络

接下来，我们给一个连续博弈的例子，即生成对抗网络 (GAN). 生成对抗网络由两个子模型组成，一个被称为生成模型，一个被称为判别模型. 生成模型的任务是生成看似真实的数据，而判别模型的任务是识别给定的数据是真实的还是伪造的. 这一过程的示意见图 9.6.

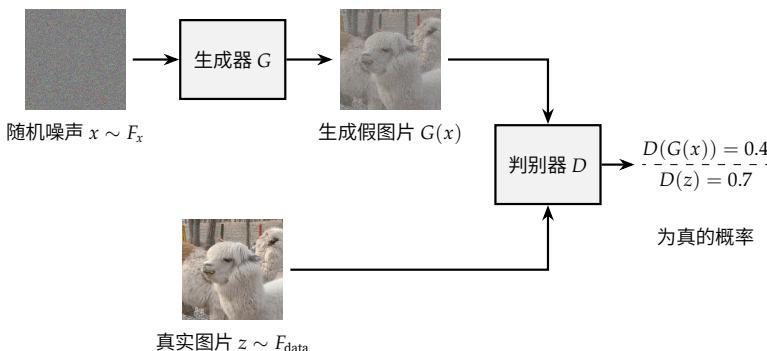


图 9.6: 生成对抗网络的示意图

假设真实数据的分布为 F_{data} .

- 生成模型为 $G(x; \theta_g)$, 参数为 θ_g , 输入随机向量 x , 输出数据向量 z . 当 x 服从分布 F_x , G 的输出会形成一个分布 F_g .
- 判别模型为 $D(z; \theta_d)$, 参数为 θ_d , 接受一个数据向量 z , 输出一个 $[0, 1]$ 中的实数, 表示 z 来自分布 F_{data} 的概率 (似然) .

我们假设 F_{data} 和 F_x 都是连续型分布，有密度函数 p_{data} 和 p_x . 我们再假设 D 和 G 都是连续的.

将 G 和 D 看成两个玩家，于是 GAN 可以被看成一个二人零和博弈，收益函数为：

$$V(G, D) = \mathbb{E}_{z \sim F_{\text{data}}} (\log D(z)) + \mathbb{E}_{x \sim F_x} (\log(1 - D(G(x)))).$$

D 最大化 V ， G 最小化 V .

接下来，我们要解释，为什么这个收益函数能够达到我们的目标. 我们有以下三点讨论：

- 一方面，如果 D 越厉害，那么 D 会尽可能地把真实数据和生成数据区分开， $1 - D(G(x))$ 也会更大，因此 V 会更大. 另一方面，如果 G 越厉害，那么 D 会更容易把真实数据和生成数据混淆，于是 $1 - D(G(x))$ 会更小， V 也会更小. 因此， D 和 G 之间是一个对抗的关系.
- 如果收益函数只有第二项，判别器 D 可以“作弊”，即无论输入什么都判定为假，这样他总是能得到最大的收益. 为了避免这种情况，我们引入了第一项，即真实数据的似然. 这样， D 还需要努力判断真实数据，而不是只判断生成数据.
- 为什么 V 是对数的形式？在第三章中我们解释过，这样形式的收益函数是交叉上损失函数的形式. 因此这一损失函数可以衡量两个分布之间的相似性，对于随机的数据来说，这是一个比较好的衡量方式.

从博弈论角度出发，一个基本的问题是 Nash 均衡是否存在？假设

D 和 G 都可以任意选择连续函数. 我们将展示一种通用的方式求解连续博弈的 Nash 均衡. 注意到 $G(x)$ 形成了一个连续分布, 密度记为 p_g . 首先证明密度函数存在性定理:

定理 9.4 设 $X \sim \mathcal{U}(0, 1)$. 对于任意密度函数 p , 存在一个连续函数 F 使得 $F(X)$ 具有密度 p .

证明. 设 F_p 是 p 对应的分布函数, 它是一个单调的连续函数. 取 $F(x) = \inf\{y \in \mathbb{R} : F_p(y) \geq x\}$ 即可. \square

因此, G 的行动等价于选择 p_g .

给定 G 的选择 p_g , 我们来求 D 的最优反应 D^* .

$$V(G, D) = \int (p_{\text{data}}(x) \log D(x) + p_g(x) \log(1 - D(x))) dx.$$

函数 $a \log x + b \log(1 - x)$ 最大值在 $x = a/(a + b)$ 的时候取得. 因此,

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}.$$

现在, 给定最优反应 $D^* = p_{\text{data}}(x)/(p_{\text{data}}(x) + p_g(x))$, 我们来求 G 的最优反应. 直观上, G 能做到的最好选择就是 $p_g = p_{\text{data}}$. 此时, $D^*(x) = 1/2$, 因此对任意 G , $V(G, D^*) = -\log 4$. G 选任何策略都是一样的收益, 因此这是一个 Nash 均衡. 我们证明了:

定理 9.5 (GAN 的 Nash 均衡存在性) 在 GAN 的博弈中, G 选择 p_{data} , D 选择 $1/2$ 是一个 Nash 均衡.

§9.3.4 混合策略

我们已经看到，在相当普遍的情况下，纯策略 Nash 均衡并不存在。所以我们需要允许玩家进行随机行动，这就是混合策略。

定义 9.4 (混合策略) 给定一个正则形式博弈 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$ ，其中 I 是玩家集合， A_i 是玩家 i 的行动空间， u_i 是玩家 i 的收益函数。玩家 i 的行动空间上的概率分布集合为 $\Delta(A_i)$ 。那么，玩家 i 的混合策略是一个概率分布 $\sigma_i \in \Delta(A_i)$ 。□

当 A_i 有 n 个元素（有限）， $\Delta(A_i)$ 可以被表示为标准的 n -单纯形：

$$\Delta(A_i) = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, \forall j, x_j \geq 0 \right\}.$$

特别地，纯策略可以被看成是一个特殊的混合策略，即退化的概率分布。

有了混合策略，我们还需要重新定义玩家的理性。一个非常标准的回答是期望效用理论，它由 Von Neumann 和 Morgenstern 提出。该理论认为，在面对不确定性时，人按照期望效用进行决策。

因此，我们需要计算玩家的期望效用。为此，引入混合策略组合： $\sigma = (\sigma_i)_{i \in I}$ ，其中 $\sigma_i \in \Delta(A_i)$ 。 σ 是一个 $(A_i)_{i \in I}$ 上的概率分布，每一维相互独立。当所有玩家选定策略之后，玩家 i 的期望收益是：

$$u_i(\sigma) = \mathbb{E}_{a \sim \sigma} u_i(a).$$

有了期望效用，我们可以重新定义最优反应：

定义 9.5 (最优反应) 给定对手的策略组合 σ_{-i} ，玩家 i 的最优反应是一

个策略 σ_i , 满足对任意 $\sigma'_i \in \Delta(A_i)$,

$$u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma'_i, \sigma_{-i}),$$

即玩家 i 的期望收益最大化. 最优反应对应的策略集合记为 $BR_i(\sigma_{-i})$. \square

最后, 我们可以重新定义均衡的概念:

定义 9.6 (Nash 均衡) 对于一个博弈 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$, 如果混合策略组合 σ 满足对于任意玩家 i 和任意 $\sigma'_i \in \Delta(A_i)$, 都有

$$u_i(\sigma_i, \sigma_{-i}) \geq u_i(\sigma'_i, \sigma_{-i}),$$

那么, σ 是一个 **Nash 均衡**.

\square

同样, Nash 均衡也可以理解为最优反应的不动点. 利用这一观点, Nash 得以使用不动点定理 (见习题8) 证明了 Nash 均衡的存在性:

定理 9.6 (Nash 均衡存在性定理) 任意有限正则形式博弈都存在一个 *Nash 均衡*.

注. 目前为止, 我们一共定义了两种均衡: 纯策略 Nash 均衡和 Nash 均衡. 后一种均衡实际上应该被称作混合策略 Nash 均衡, 但是为了和文献统一, 我们直接称之为 Nash 均衡, 而忽略前缀“混合策略”.

实际上, 按照我们理性-最优反应-均衡的思路, Nash 均衡和纯策略 Nash 均衡是本质不同的两种均衡: 前者引入了期望效用理论来体现玩家面对不确定性时候的行为, 而后者则是直接最大化, 没有不确定性. 因此, 我们并不能简单说前者是后者的推广.

我们来看一个例子.

例 9.5 (猜硬币游戏的 Nash 均衡) 继续考虑猜硬币游戏, 收益矩阵为

$$\begin{pmatrix} 1, 0 & 0, 1 \\ 0, 1 & 1, 0 \end{pmatrix}.$$

我们接下来证明, 唯一的 Nash 均衡是两个玩家都选择 $(1/2, 1/2)^\top$, 即两个玩家都以等概率来猜硬币的正反面. 我们有如下分类讨论:

- 两个玩家都是纯策略: 我们之前已经说明, 这个博弈没有纯策略 Nash 均衡.
- 行玩家是纯策略而列玩家的策略是 $(p, 1 - p)^\top$, 其中 $p \in (0, 1)$. 考虑行玩家的期望收益:

$$u_1(\sigma_1, \sigma_2) = \begin{cases} p, & \sigma_1 \text{ 是第一行,} \\ 1 - p, & \sigma_1 \text{ 是第二行.} \end{cases}$$

因此, 如果 $p > 1/2$, 行玩家会选择第一行, 但此时列玩家的最优反应是选择第二列, 因此这不是一个 Nash 均衡. 同理, 如果 $p \leq 1/2$, 也不是一个 Nash 均衡. 因此, 这种情况也不是一个 Nash 均衡.

- 列玩家是纯策略而行玩家的策略是 $(p, 1 - p)^\top$, 其中 $p \in (0, 1)$. 同理, 这种情况也不是一个 Nash 均衡.
- 行玩家的策略是 $\sigma_1^* = (p, 1 - p)^\top$, 列玩家的策略是 $\sigma_2^* = (q, 1 - q)^\top$, 其中 $p, q \in (0, 1)$. 这种情况下, 要确定 p, q 的值计算会很复杂. 所以我们这里介绍一个技巧.

注意到，行玩家的任意一个混合策略可以写作

$$\sigma_1 = p_1 a_1 + p_2 a_2,$$

其中 $p_1 + p_2 = 1$, a_1 和 a_2 是两个纯策略. 因此，在 Nash 均衡下，对任意 p_1, p_2 ,

$$\begin{aligned} u_1(\sigma_1^*, \sigma_2^*) &= p u_1(a_1, \sigma_2^*) + (1 - p) u_1(a_2, \sigma_2^*) \\ &\geq p_1 u_1(a_1, \sigma_2^*) + p_2 u_1(a_2, \sigma_2^*). \end{aligned}$$

取 $p_1 = 1$, $p_2 = 0$, 我们有

$$p u_1(a_1, \sigma_2^*) + (1 - p) u_1(a_2, \sigma_2^*) \geq u_1(a_1, \sigma_2^*).$$

同理，取 $p_1 = 0$, $p_2 = 1$, 我们有

$$p u_1(a_1, \sigma_2^*) + (1 - p) u_1(a_2, \sigma_2^*) \geq u_1(a_2, \sigma_2^*).$$

注意到，两个不等式的左边其实是右边的加权平均，平均值不小于任何一方，因此这两个不等式实际上是等号. 因此，我们有

$$\begin{aligned} u_1(a_1, \sigma_2^*) &= u_1(a_2, \sigma_2^*) \\ \iff q \cdot 1 + (1 - q) \cdot 0 &= q \cdot 0 + (1 - q) \cdot 1 \\ \iff q &= \frac{1}{2}. \end{aligned}$$

同理，我们有 $p = 1/2$. 因此，双方都选择 $(1/2, 1/2)^T$ 是唯一的 Nash 均衡.

如此, 我们证明了猜硬币游戏的唯一 Nash 均衡是双方都选择 $(1/2, 1/2)^\top$. \square

在上面的例子中, 我们实际上得到了一个非常重要的结论: Nash 均衡具有无差别原理, 它说明, Nash 均衡中出现的那些行动一定都是取到了最大收益的行动. 这一原理可以被用来求解任意二人博弈的 Nash 均衡. 这一原理的证明和应用见习题 10.

定理 9.7 (无差别原理) 考虑一个正则形式博弈 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$, 其中 I 是玩家集合, A_i 是玩家 i 的行动空间, u_i 是玩家 i 的收益函数. 如果 σ^* 是一个 Nash 均衡, 那么对于任意玩家 i 和任意 a_i, a'_i 满足

$$\sigma_i^*(a_i) > 0, \quad \sigma_i^*(a'_i) > 0,$$

有

$$u_i(a_i, \sigma_{-i}^*) = u_i(a'_i, \sigma_{-i}^*) = u_i(\sigma^*) = \max_{\sigma_i \in \Delta(A_i)} u_i(\sigma_i, \sigma_{-i}^*).$$

上面我们说明, 有限正则形式博弈一定有 Nash 均衡, 实际上, 类似的不动点技术可以被用来证明更一般的均衡存在定理:

定理 9.8 (Debreu-Glicksberg-Fan 定理) 一个正则形式博弈满足如下所有条件, 就会存在一个纯策略 Nash 均衡:

- 对任意玩家 i , 行动空间 A_i 是 Euclid 空间中的非空紧凸子集,
- 对任意玩家 i , 收益函数 u_i 连续, 对第 i 个分量上凹.

定理 9.9 (Glicksberg 定理) 一个正则形式博弈满足如下所有条件, 就会存在一个 Nash 均衡:

- 对任意玩家 i , 行动空间 A_i 是度量空间中的非空紧子集,

- 对任意玩家 i , 收益函数 u_i 连续.

尽管在数学上, 期望效用理论导出了漂亮的结果, 即 Nash 均衡的存在性, 但是 (期望) 效用理论非常不符合实验的观察. 比如说, 下面两个选项, 大部分人会更倾向于选择第一个:

- 你得到 1000 元.
- 你有 1% 的概率得到 100000 元, 99% 的概率得到 0 元.

然而, 如果我们套用期望效用理论, 两个选项的期望效用是一样的!

此外, 从某种角度来看, 价值也是相对的. 如果我们有 50% 的机会赢 100 块钱或 10 块钱, 拿到 10 块钱时可能会感到失望. 但如果改成我们只能赢得 10 块钱或 1 块钱, 那得到 10 块钱时我们就会感到更开心. 因此, 即便效用理论是对的, 在现实中我们完全不能知道真正的效用函数是什么, 也无法在现实中真的用来建模.

那么, 为什么我们还要研究 (期望) 效用理论, 甚至整个经济学和博弈论的体系都是基于 (期望) 效用理论的? 它之所以如此重要, 就是因为它很好地平衡了“可以写理论结果”和“可以解释现实现象”之间的关系. 正如 Robert Aumann 所说,

“另一种强调这一点的原因——即理论不应简单地被视
为真或假——是为了避免过分抠字眼带来的问题……有人反
对效用最大化的概念, 认为人并不真的在最大化效用. 对此,
有人提出了诸如‘满意准则’这样的替代方案……然而, 效
用最大化的有效性并不在于它能否精确描述个体行为, 而
在于它作为经济理论的基础假设, 能够整合大量经济学理
论……

“像‘满意准则’这样的替代方案虽然看起来有吸引力，但几乎没有什么实际作用，它们很少带来有趣的结果。在评价效用最大化时，我们不应问‘它是否合理？’，而是应该问‘它能整合什么？它能引导我们走向哪里？’”

在任何时候，作为一门语言，博弈论都应该具备 Aumann 所说的这种特性。我们不能为了数学的方便而过分简化理论，但更不能过分拟合现实，变得毫无指导和应用价值。

§9.4 随机博弈 (Markov 博弈)

本节我们讲述一种动态博弈和正则形式博弈的结合：随机博弈。为了引入随机博弈，我们需要 Markov 链和 Markov 决策过程 (MDP) 相关的知识，更详细的讨论请参阅第二章。一个 MDP 有如下的组成：

- 有限状态集合： $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$.
- 有限动作集合： \mathcal{A} .
- 每个状态有自己的动作空间： $\mathcal{A}_s = \{a_{s,1}, a_{s,2}, \dots, a_{s,N}\}, s \in S$.
- 每个动作空间有限： $|\mathcal{A}_{s,k}| = n_{s,k}, s \in S, k = 1, 2, \dots, N$.
- 在状态 s_k ，若选择第 i 个动作 $a_{k,i}$ ($1 \leq i \leq n_k$)，则可以定义
 - 状态转移概率： $\mathcal{P}(s'|s, a)$ ，表示在状态 s 选择动作 a 后，转移到状态 s' 的概率。
 - 即时奖励： $R(s, a, s')$ ，表示在状态 s 选择动作 a 后，转移到状态 s' 时获得的即时奖励。

- 折扣因子 $\gamma \in [0, 1]$, 用于计算远期收益.

MDP 的目标是找到一个策略 π , 使得在该策略下可以获得最大的期望累积奖励.

随机博弈可以看做 MDP 的多人扩展.

定义 9.7 (随机博弈, Markov 博弈) 随机博弈(又称为 **Markov 博弈**)有如下组成:

- N : 玩家的数量, $N = 1$ 退化为 MDP. 用 $1, \dots, N$ 表示玩家的编号.
- \mathcal{S} : 状态的集合.
- \mathcal{A} : 玩家的行动集合. $\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^N$. 设 $\mathcal{A}_i(s)$ 表示第 i 个玩家在状态 s 的行动空间.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$: 给定玩家的联合动作 $a \in \mathcal{A}$, 从状态 $s \in \mathcal{S}$ 转移到 $s' \in \mathcal{S}$ 的概率 $\mathcal{P}(s'|s, a)$.
- $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: 在状态 s , 当玩家的联合动作为 a 时, 玩家 i 的奖励值 $Q_i(a; s)$ (有界) .
- $\gamma \in [0, 1]$ 表示折扣因子. □

有了博弈的框架, 我们可以讨论随机博弈的过程.

- 首先, 博弈从某一个状态 s^0 开始, $s^0 \in \mathcal{S}$.
- 在每个阶段 t , 所有玩家同时选择自己的动作 a^t .
- 环境根据所有玩家的动作 a^t 和状态 s^t , 进行如下操作:
 - 给予每个玩家对应的收益 $q(a^t, s^t)$.

- 转移到新的状态 $s^{t+1} \in \mathcal{S}$.

接下来，我们可以定义玩家的策略。假设在阶段 T ，所有玩家可以观察到所有历史动作 $\{a^t\}_{t \leq T}$ 。和井字棋一样，我们可以定义每个玩家的策略 π ——基于历史信息（状态、行动）到当前状态的行动的映射。玩家在博弈的过程中，其实就是按照某个策略 π 进行行动的。

然而，每个玩家的策略可以依赖于历史信息，但这种策略通常很复杂，为了简化，我们考虑一些更简单的策略。简化的关键在于，这是一个无穷轮的博弈，所以玩家需要有无穷大的记忆来存储历史信息。

如果我们让玩家只有固定大小的记忆，那么玩家的策略就只能依赖于有限的历史信息。在最简化（健忘）的情况下，玩家的行动选择仅依赖于当前状态，而与之前的历史无关。这种策略被称为平稳策略。

平稳策略的正式定义如下：

定义 9.8 (平稳策略) 对于玩家 i ，平稳策略 π_i 是一个映射，从当前状态 $s \in \mathcal{S}$ 到行动空间 $\mathcal{A}_p(s)$ 的概率分布。

因而，平稳策略可以表示为

$$\pi_i : \mathcal{S} \rightarrow \Delta(\mathcal{A}_i),$$

其中 $\Delta(\mathcal{A}_i)$ 表示行动空间 \mathcal{A}_i 上的所有概率分布。

在每个状态 s 下，玩家 i 选择每个可能行动 $a \in \mathcal{A}_i(s)$ 的概率由 $\pi_i(s, a)$ 给出。

假设每个玩家 i 都采用平稳策略 π_i ，那么整个策略组合 $\pi = (\pi_1, \dots, \pi_N)$ 也被称为平稳策略组合。□

平稳这一词在概率论中有明确的含义。考虑一列随机变量 X_1, X_2, \dots ，

如果对于任意 n 和 k ,

$$(X_1, X_2, \dots, X_n) \stackrel{d}{=} (X_{k+1}, X_{k+2}, \dots, X_{k+n}),$$

这个等号的意思是两边的联合分布相同. 这时我们称 X_1, X_2, \dots 是一个平稳过程.

实际上, 平稳策略这一词正来源于此. 我们现在说明更强的结论, 即平稳策略组合其实诱导了一个 Markov 链. 假设玩家采用平稳策略 π , s^0, s^1, \dots 是一个随机博弈的状态序列, 那么容易证明 $\{s^t\}_{t=0}^\infty$ 是一个时齐 Markov 链 (见习题 11), 转移概率可以表示为

$$\mathcal{P}^{(\pi)}(s'|s) = \mathbb{E}_{a \sim \pi(s)}[\mathcal{P}(s'|s, a)].$$

因此, 从这个视角来看, 随机博弈就是在 Markov 链上进行的博弈, 因此有时被称为 *Markov* 博弈.

由于 Markov 博弈兼具正则形式博弈和 MDP 的特点, 所以我们接下来会将这两部分对应的概念和性质都引入到随机博弈中.

从博弈论的角度, 我们要讨论玩家的理性. 首先, 我们可以按照期望效用理论, 扩展收益函数 $Q_i(a; s)$ 为 $Q_i(\pi; s)$:

$$Q_i(\pi; s) = \mathbb{E}_{a \sim \pi(s)}[Q_i(a; s)].$$

接下来, 我们仿照 MDP, 定义玩家在整个博弈中的收益:

定义 9.9 (价值函数) 对于一个随机博弈, 玩家 i 的价值函数 $V_i^\pi(s)$ 定义

为

$$V_i^\pi(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t Q_i(\boldsymbol{a}; s^t) \middle| s^0 = s, \boldsymbol{a} \sim \pi(s^t), s^{t+1} \sim \mathcal{P}^{(\pi)}(\cdot | s^t) \right],$$

即从初始状态 s^0 开始，玩家 i 对每一期折现收益的期望。 \square

在随机博弈中，玩家的理性即是在给定其他玩家的策略 π_{-i} 的情况下，最大化其价值函数。

与 MDP 类似，价值函数也满足 Bellman 方程：

定理 9.10 (Bellman 方程)

$$V_i^\pi(s) = Q_i(\pi(s); s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^\pi(s'|s) V_i^\pi(s').$$

证明见习题 12.

有了理性的定义，我们可以自然地引入最优反应的概念：

定义 9.10 (Markov 最优反应) 对于一个随机博弈，给定其他玩家的平稳策略 π_{-i} ，玩家 i 的 **Markov 最优反应**是一个策略 π_i ，使得对于任意状态 s 和任意平稳策略 π'_i ，

$$V_i^{\pi_i, \pi_{-i}}(s) \geq V_i^{\pi'_i, \pi_{-i}}(s).$$

最优反应对应的策略集合记为 $BR_i(\pi_{-i})$. \square

最后，我们可以用最优反应来定义均衡：

定义 9.11 (Markov 完美均衡，MPE) 在所有玩家的平稳策略组合中，每一个玩家的策略都是 Markov 最优反应，那么这个策略组合被称为 **Markov 完美均衡 (MPE)** . \square

同样, MPE 也可以被看作是最优反应的不动点. 类似 Nash 均衡的存在性定理, 我们有如下的定理:

定理 9.11 (MPE 存在性定理) 对于 N 个玩家、有限状态、有限动作空间的随机博弈, MPE 存在.

下面我们介绍 Shapley 关于双人零和随机博弈情形的证明, 这一证明基于 Banach 不动点定理 (定理 8.2). 对于一般的情况, 可以类似 Nash 均衡存在性, 利用 Brouwer 不动点定理证明, 见习题 13.

在双人零和的语境下, 我们去掉收益函数 Q 的下标 i , 玩家 2 的收益函数是 $-Q$. 注意, 价值函数满足 Bellman 方程:

$$V^\pi(s) = Q(\pi(s); s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^\pi(s'|s) V^\pi(s').$$

我们的证明策略是用迭代的方法逼近这个方程的解 V . 为了利用 Banach 不动点定理, 我们需要定义一个迭代算子, 它是压缩映射, 并且迭代收敛到 V .

Bellman 方程的左右是同一个 V , 此时是不动点方程, 因此, 只要把右边看成迭代的过程, 左边看成迭代的结果, 我们就自然得到了一个迭代算子. 具体过程如下:

- 首先, 我们选择一个任意的函数 $\alpha : \mathcal{S} \rightarrow \mathbb{R}$, 称 α 为值函数. 这个函数与 V 属于同样的空间, 这是迭代的初始值 α_0 .
- 对任意 $s \in \mathcal{S}$, 定义矩阵 $R_s(\alpha)$ 为

$$R_s(\alpha)(a_1, a_2) = Q(a_1, a_2; s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|a_1, a_2, s) \alpha(s').$$

其中 $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$.

如此, 我们形成了一个矩阵零和博弈, 玩家 1 的收益矩阵是 $R_s(\alpha)$, 玩家 2 的收益矩阵是 $-R_s(\alpha)$. 注意, 根据 minimax 定理(见第七章习题 6), 玩家 1 在 Nash 均衡时候的收益是确定的一个值, 记为 $\text{val}(R_s(\alpha))$.

- 进行下一步迭代 $\alpha_k(s) = \text{val}(R_s(\alpha_{k-1}))$.

我们中间之所以先定义了一个矩阵 $R_s(\alpha)$, 是因为 Bellman 方程中 $\pi(s)$ 是不知道的, 因此 Q 是无法计算的. 根据 Q 的含义, 这其实是一期的收益, 所以我们可以用一个矩阵博弈来代替. 价值函数 V 满足

$$V(s) = \text{val}(R_s(V)),$$

因此 V 是 $\text{val}(R_s(\cdot))$ 的不动点, 这说明我们的矩阵博弈选择是合理的. 接下来, 只要证明 $\text{val}(R_s(\cdot))$ 是一个压缩映射, 因此根据 Banach 不动点定理, α_k 就会收敛到 V .

为方便, 我们定义迭代算子 $(T\alpha)(s) = \text{val}(R_s(\alpha))$. 我们需要有一个方法来衡量 $T\alpha$ 和 $T\alpha'$ 的差距, 因此, 我们给出如下引理:

引理 9.1 对任意 $m \times n$ 的矩阵 B, C , 成立:

$$|\text{val}(B) - \text{val}(C)| \leq \max_{i,j} |B_{ij} - C_{ij}|.$$

证明. 设 (s_1, s_2) 为矩阵博弈 B 的 Nash 均衡, (\bar{s}_1, \bar{s}_2) 为矩阵博弈 C 的 Nash 均衡. 于是由定义有: $s_1^\top B \bar{s}_2 \geq s_1^\top B s_2$, 且 $\bar{s}_1^\top C \bar{s}_2 \geq s_1^\top C \bar{s}_2$, 因此

$$\text{val}(B) - \text{val}(C) = s_1^\top B \bar{s}_2 - s_1^\top C \bar{s}_2$$

$$\begin{aligned} &\leq s_1^\top B \bar{s}_2 - s_1^\top B s_2 \\ &\leq \max_{i,j} |B_{ij} - C_{ij}|. \end{aligned}$$

根据 B 和 C 的对称性, 引理得证. \square

根据这个引理, 我们可以证明 T 是一个压缩映射:

引理 9.2 如果 $\gamma \in (0, 1)$, 那么 T 是一个压缩系数为 γ 的压缩映射.

证明.

$$\begin{aligned} \|T\alpha - T\alpha'\|_\infty &= \max_{s \in \mathcal{S}} |\text{val}(R_s(\alpha)) - \text{val}(R_s(\alpha'))| \\ &\leq \gamma \max_{s \in \mathcal{S}} \max_{a_1, a_2} \left| \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|a_1, a_2, s) (\alpha(s') - \alpha'(s')) \right| \\ &\leq \gamma \max_{s' \in \mathcal{S}} |\alpha(s') - \alpha'(s')| \\ &= \gamma \|\alpha - \alpha'\|_\infty. \end{aligned}$$

第一个不等式的成立使用了引理 9.1, 第二个不等式的成立使用了 $\mathcal{P} \in [0, 1]$ 的性质. \square

对于有折扣的博弈, $\gamma \in (0, 1)$, 因此 T 是一个压缩映射, 由 Banach 不动点定理可知, α_k 会收敛到 α^* 满足 $T\alpha^* = \alpha^*$, 这就是我们要求的 V .

为了证明定理, 我们还需要说明:

1. α^* 一定是均衡时候的玩家 1 的收益.
2. 存在一个策略组合达到均衡. 我们将要证明更强的结论: 玩家 1 有一个策略, 不论玩家 2 如何选择, 玩家 1 的收益至少是 α^* .

要想证明这两点, 我们需要理解 α 这一含义的直观. 选取 $\alpha_0(s) \equiv 0$, 则 $R_s(\alpha_0) = Q(a_1, a_2; s)$ 是从 s 出发、由 Q 定义的矩阵博弈. 于是, $\alpha_1(s)$ 就是这一矩阵博弈玩家 1 的收益.

$$\alpha_1(s) = \text{val}(R_s(\alpha_0)) = \text{val}(Q(\cdot, \cdot; s)).$$

为了方便起见, 我们总将 “玩家 1 的收益” 称之为 “值”. 再看下一轮迭代,

$$R_s(\alpha_1)(a_1, a_2) = Q(a_1, a_2; s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|a_1, a_2, s) \alpha_1(s').$$

这个公式可以被这样理解, 假想有一个被截断的两阶段随机博弈, 玩家的策略如下:

- 玩家在第一阶段从状态 s 出发, 行动 (a_1, a_2) 待定;
- 在第二阶段, 对于每个可能到达的状态 $s' \in \mathcal{S}$, 玩家采用矩阵博弈 $R_{s'}(\alpha_0)$ 的 Nash 均衡的行动.

注意, 此时我们没有要求玩家策略是平稳的, 也就是说, 第二阶段和第一阶段的策略不需要是一样的. 根据定义, 我们上面描述的这种策略达到了矩阵博弈 $R_s(\alpha_1)$ 的值, 因此这个两阶段随机博弈的值不小于 $R_s(\alpha_1)$ 的值. 更一般地, 被截断的 k 阶段随机博弈的值不小于 $\alpha_k(s)$.

接下来, 我们证明 α^* 一定是均衡时候的值.

引理 9.3 α^* 是一个 MPE 时候玩家 1 的收益.

证明. 考虑从 s 出发的双人零和随机博弈, 在前 k 局的博弈中, 玩家 1 采用前 k 局截断随机博弈的最优策略, 后续状态可选择任意动作. 由之

前的分析可知，前 k 局截断的随机博弈的值不小于 $\alpha_k(s)$. 而对于之后的博弈，玩家 1 损失的累积收益最差不超过

$$(\gamma^k + \gamma^{k+1} + \dots) \sup |Q| = \frac{\gamma^k}{1-\gamma} \cdot \sup |Q|.$$

因此，令 k 趋于无穷，我们得到玩家 1 的收益至少是 $\alpha^*(s)$. 另一方面，根据同样的论证，玩家 2 也可以确保自己的收益至少是 $-\alpha^*(s)$. 由零和的定义，均衡时玩家 1 的收益必定是 $\alpha^*(s)$. \square

最后，我们说明存在一个策略组合达到均衡. 如果让 R_s 作用在不动点 α^* 上，玩家 1 最大化 V 就是在选择 $R_s(\alpha^*)$ 的 Nash 均衡. 设 $R_s(\alpha^*)$ 的 Nash 均衡为 $(\pi_1(s), \pi_2(s))$ ，我们证明 (π_1, π_2) 是一个 MPE.

引理 9.4 设 $R_s(\alpha^*)$ 的 Nash 均衡为 $(\pi_1(s), \pi_2(s))$ ，则 (π_1, π_2) 是一个 MPE.

证明. 固定玩家 2 的一个任意策略 $\hat{\pi}_2$ (不一定是平稳策略). 玩家 1 在前 k 步选择 π_1 ，因为 π_1 是 $R_s(\alpha^*)$ 的 Nash 均衡，它也是前 k 步截断随机博弈的最优策略. 因此，根据之前的论证，无论玩家 2 选择何种行动，玩家 1 都能够至少拿到 α^* 的收益，即

$$\mathbb{E} \left[\sum_{t=0}^{k-1} \gamma^t Q(\pi_1(s^t), \hat{\pi}_2(s^t); s^t) + \gamma^k \alpha^*(s^k) \middle| s^0 = s \right] \geq \alpha^*(s).$$

整理得

$$\mathbb{E} \left[\sum_{t=0}^{k-1} \gamma^t Q(\pi_1(s^t), \hat{\pi}_2(s^t); s^t) \middle| s^0 = s \right] \geq \alpha^*(s) - \gamma^k \|\alpha^*\|_\infty.$$

因此，

$$V^{\pi_1, \hat{\pi}_2}(s) \geq \alpha^*(s) - \gamma^k \|\alpha^*\|_\infty - \frac{\gamma^k}{1-\gamma} \sup |Q|.$$

令 $k \rightarrow \infty$ 可得，上式右边趋于 $\alpha^*(s)$ ，因此我们为玩家 1 选择的策略达到了 α^* 的收益。

玩家 2 的 π_2 证明对称，因此 (π_1, π_2) 是一个 MPE. □

最后，我们指出随机博弈和人工智能的关系。正如 MDP 是强化学习的基础，随机博弈是多智能体强化学习的基础。在多智能体强化学习中，每个智能体都是一个独立的决策者，按照一个 MDP 来决策。但是，由于环境中有很多个智能体，每个智能体的奖励函数都会受到其他智能体的影响，这正是随机博弈的情形。

利用多智能体强化学习，我们可以训练出极其强大的人工智能，例如，DeepMind 在星际争霸 2 中训练出的 AlphaStar 就是一个例子。通过人类数据与多智能体强化学习的自我训练，AlphaStar 在星际争霸 2 的比赛中击败了世界冠军。它展示了博弈论如何为人工智能的发展提供语言和训练方法。

§9.5 习题

1. Zermelo 定理成立的条件是双人的、有限深的、确定的、完全信息的、输赢的博弈。当破坏其中任何一个条件时，博弈 Zermelo 定理不再成立。下面的题目都只破坏一个条件。

- (1) 写出一个非确定的但其他条件没有被破坏的博弈，要求是一个来自现实生活中的例子。

- (2) 写出一个不完全信息但其他条件没有被破坏的博弈，要求是一颗博弈树.
2. 证明定理 9.2.
3. 在充分大（但有限）的棋盘上进行五子棋，黑方先手，白方后手，对落子没有限制，只要有空格就可以落子，首先下出连续五个同色子的玩家获胜（横向、纵向或者两个斜对角线向都可以），如果棋盘被下满还没有人获胜，则双方平局. 证明：黑方有不败策略，即黑方有一种策略，不论对手如何行动，都使得要么黑方获胜要么平局.
4. 对于一个 MDP 问题，不同于围棋等游戏，它只有一个玩家，但是我们依然可以利用 AlphaGo Zero 的架构来解决这个问题，请简要说明如何实现这个方法. 进一步，这种实现和二人输赢博弈有什么异同？
5. * 请查阅资料，回答以下关于输赢博弈的问题.
- (1) 如何衡量一个输赢博弈比另一个输赢博弈更复杂？
- (2) 在输赢博弈中，表示完整策略需要多少比特？AlphaGo Zero 可以衡量多复杂的策略？
6. * 请查阅资料，了解对话逻辑（dialogical logic）的相关内容，回答以下问题.
- (1) 对话逻辑和博弈论有什么联系？Zermelo 定理能够得出什么样的结论？
- (2) 对话逻辑与现在的对话型生成式语言模型有什么联系？
7. 设 $G = (I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I})$ 是一个二人正则输赢博弈，其中 $I = \{1, 2\}$ ， A_i 是有限集合， $u_i(\cdot) \in \{0, 1\}$ ， $u_1 + u_2 = 1$.

- (1) 写出玩家 i 存在必胜策略的定义.
- (2) 证明: G 存在纯策略 Nash 均衡当且仅当其中一个玩家存在必胜策略, 从而证明定理 9.3.

8. 使用 Brouwer 不动点定理证明定理 9.6.

提示: 考虑混合策略空间 S 上的函数 $f(\sigma) = \sigma'$ 的不动点, 其中

$$\sigma'(a_i) = \frac{\sigma(a_i) + \phi_{i,a_i}(\sigma)}{1 + \sum_{b_i \in A_i} \phi_{i,b_i}(\sigma)},$$

这里 $\phi_{i,a_i}(\sigma) = \max\{0, u_i(a_i, \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i})\}$.

9. 考虑 A, B, C 玩家的三人正则形式博奕, 每位玩家有两种选择 X, Y . 构造一个三人博奕的有理收益矩阵, 使得它有一个 Nash 均衡中三人选择 X 的概率都是无理数. 并求出这个均衡.

10. 本题讨论无差别原理.

- (1) 证明定理 9.7.
- (2) 对有限行动空间的二人正则形式博奕, 利用无差别原理给出一个求解所有 Nash 均衡的算法.

11. 考虑一个随机博奕, 有 N 个玩家, 状态集合为 S , 玩家 i 的行动空间为 A_i , 收益函数为 $Q_i(a; s)$, 转移概率为 $P(s'|s, a)$. 假设玩家采用平稳策略 π , s^0, s^1, \dots 是一个随机博奕的状态序列, 证明: $\{s^t\}_{t=0}^\infty$ 是一个时齐 Markov 链.

12. 证明定理 9.10.
13. * 仿照习题 8, 使用 Brouwer 不动点定理证明定理 9.11.

第五部分

认知与逻辑

第十章 共同知识，Bayes 博弈，Aumann 知识算子

真实世界是一个巨大的游戏，参与者之间的信息不对称是一个普遍存在的现象，这些不对称往往形成了“丛林法则”。“丛林法则”从来不是一个书面的规则。刚来美国旅游的人可能会非常担心自己被抢劫；然而，“盗亦有道”，绝大部分时候，抢劫犯只会抢二十美元，以便吃一顿饭。如果坏了规矩，反而会被惩罚。

我们可以考虑一个看似非常疯狂的想法：既然有丛林法则，何不把“一次抢劫最多二十美元”写入法律，这会有什么区别呢？抛开道德和法律的问题，这样做似乎不无道理。

1982年，经济学家 Alvin E. Roth 和 J. Keith Murnighan 对讨价还价这一经典的市场博弈现象进行了同样的实验。实验中，有两个玩家，他们可能会收到特定价值的奖品，每个玩家的奖品价值可以不同。然而玩家并不总是能获得奖品，他有一个概率获得这一奖品。他们需要在规定的时间内就他们各自获得奖品的概率进行讨价还价。

具体来说，玩家其实在商讨如何分配一张100%概率的“彩票券”，

这个券决定了每个玩家赢得奖品的概率. 例如, 甲如果获得 40% 的彩票券, 就有 40% 的概率赢得奖品, 60% 的概率一无所获; 而乙则完全相反, 有 60% 的概率赢得奖品, 40% 的概率一无所获.

然而, 如果在规定时间内未能达成协议, 那么所有玩家都将一无所获. 因此, 只有在双方就彩票券的分配达成了协议, 并且该玩家在随后的抽奖中中奖时, 玩家才能获得相应的奖品. 否则, 他将得不到任何奖励. 我们将这种每个玩家只有两种可能金钱收益的游戏称为 “二元彩票游戏” .

在实际的实验中, 玩家甲的奖品价值 20 美元, 玩家乙的奖品价值为 5 美元, 谈判时间限制为 12 分钟. 在实验开始前, 实验人员会公开告知甲乙完整的游戏规则, 包括奖品价值和谈判时间限制. 实验人员还会公开告知甲乙关于他们私有信息的情况, 例如,

“在游戏开始后, 你们奖品的价值会被告诉对方.”

“在游戏开始后, 甲的奖品价值会被告诉乙, 但乙的不会被告诉甲.”

实验人员也可以选择不告知这些信息.

实验开始后, 实验人员会告知甲乙他们各自的奖品价值, 并且选择性地告知甲 (或乙) 乙 (或甲) 的奖品价值, 这些信息要和实验开始前的信息保持一致. 然后, 甲乙开始商讨如何分配彩票券. 如果在规定时间内达成了协议, 那么他们将按照协议分配彩票券. 如果未能达成协议, 那么他们将一无所获.

上面的实验设置中, 在游戏开始前是否公开宣告双方的信息结构, 和上面那个疯狂的想法是一样的. 然而, 在讨价还价的情景下, 我们似乎不会觉得这有什么区别. 然而, 实验结果却显示, 在信息不对称的情况下, 双方信息结构的公开宣告会对博弈结果产生显著的影响.

谈判过程展现出了明显的策略性行为。例如，甲（20 美元的玩家）通常不会提及自己奖品的价值；但如果游戏开始前宣布过“乙（5 美元的玩家）不清楚甲的奖品价值”，20 美元的玩家往往会虚报自己的奖品。一个典型的例子是：“我知道你的奖品是 5 美元。我的只有 2 美元。所以我应该得到超过 50% 的份额。”另一方面，当乙知道甲的奖品价值时，往往会透露信息。这两种策略通常都不被对方相信。

更重要的是，实验结果显示，当实验人员在游戏开始前宣布双方的信息结构时，玩家变得更加没有策略性，博弈的结果很可能没有达到 Nash 均衡：当只有 20 美元的玩家知道两个奖品时，其总体平均收益（包括达成协议和未达成协议的情况）为 34.9，显著低于 5 美元玩家的相应收益 53.6。

另一方面，当双方的信息结构在游戏开始前不被公开宣告时，玩家表现出更多的策略性。在这种情况下，因为他们不能确定对手是否知道自己自己的奖品，所以玩家无法像前一种情况那样自由地谎报自己的奖品价值。然而，因为玩家都不知道对方是否知道双方奖品的价值，所以，如果一个玩家知道两个奖品的价值，他完全可以假装自己只知道自己自己的奖品价值。由于更复杂的策略性行为，博弈的结果往往达到 Nash 均衡！

上面的故事告诉我们，信息结构（也就是我们知道什么）是否被公开宣告，对于博弈的结果有着重要的影响。实际上，这一问题是一个关于共同知识的问题。

在本章，我们将更系统地讨论共同知识，并定义 Bayes 博弈，它是我们研究博弈论中信息结构的基本语言。最后，我们将介绍 Aumann 知识算子，它是从 Bayes 博弈中起源的，一个关于“知识”的数学模型。

§10.1 “泥泞的孩童” 谜题

我们先从一个经典的谜题开始. 有 n 个孩子在玩泥巴, 他们互相泼泥巴. 母亲告诉孩子们, 如果他们脸上沾上了泥巴, 会受到严厉的惩罚. 孩子们不能看到自己的脸, 但是可以看到其他所有人的脸. 所有孩子都希望保持自己的脸干净, 但是弄脏别人的脸.

此时, 孩子的父亲出现了, 于是, 孩子们停止泼泥巴. 孩子们互相不说话. 父亲看到了 k ($k \geq 1$) 个人脸上有泥巴, 于是宣布: “你们至少有一个人脸上沾了泥巴.” 之后, 父亲会公开地问若干轮如下问题: “你们知道自己脸上有泥巴了吗?” 孩子们回答“知道”或者“不知道”.

假设孩子们观察力敏锐、聪慧且诚实, 并且每一轮他们都同时回答. 接下来会发生什么? 乍一看, 似乎大家每次都会回答“不知道”, 因为他们不能从这句话知道自己脸上有泥巴. 但是, 这个谜题远比想象的要复杂. 读者可以先不看后文, 先自己试一试.

假设有 k 个孩子脸上有泥巴. 这个问题的谜底是: 在前 $k - 1$ 轮中, 所有孩子都会说“不知道”, 在第 k 轮中, 所有脸上有泥巴的孩子都会说“知道”.

这一结论的论证来源于对较小 k 的归纳总结:

- 当 $k = 1$ 时, 脸上沾满泥巴的孩子看到其他人都没有泥巴. 既然他知道至少有一个孩子的脸上有泥巴, 他就能推出那个人肯定是他自己.
- 现在假设 $k = 2$, 脸上沾满泥巴的孩子是 a 和 b . 一开始, 因为他们分别看到了对方的脸上有泥巴, 所以他们每个人都回答“不知道”.

但是, 当 b 回答“不知道”时, a 可以代入 b 的角色, 意识到 b 看

到了 a 脸上有泥巴. 否则, b 在第一轮中就会知道泥巴在 b 的脸上, 并回答“知道”. 因此, a 在第二轮回答“知道”. b 也会通过同样的推理得出相同的结论.

- 现在假设 $k = 3$, 脸上沾满泥巴的孩子分别是 a , b 和 c . 孩子 a 的论证如下. 假设我没有泥巴落在脸上. 根据 $k = 2$ 的情况, b 和 c 在第二轮都会回答“是”. 他们没有这样做, 我意识到假设是错误的, 我的脸上也有泥巴. 因此在第三轮我会回答“知道”. b 和 c 的论证也是类似的.

容易看出, $k = 3$ 的论证具有一般性, 对一般的 k 也成立.

注. “泥泞的孩童”还有其他流行的陈述方式, 比如“蓝眼睛红眼睛”. 一个岛上有 100 个人, 其中有 5 个红眼睛, 95 个蓝眼睛. 这个岛有三个奇怪的宗教规则.

1. 他们不能照镜子, 不能看自己眼睛的颜色.
2. 他们不能告诉别人对方的眼睛是什么颜色.
3. 一旦有人知道了自己眼睛是红色, 他就必须在当天夜里自杀.

岛民不知道具体有几个红眼睛.

某天, 有个旅行者到了这个岛上. 由于不知道这里的规矩, 所以他在和全岛人一起狂欢的时候, 一不留神说了一句话: “你们这里有红眼睛的人.” 假设这个岛上的人足够聪明, 每个人都可以做出缜密的逻辑推理. 请问这个岛上将会发生什么?

谜题就解到这里. 然而, 一个好的谜题, 知道答案之后一定会带来更多的谜题. 为什么谜题的答案是这样的呢? 比如, 如果 $k > 1$, 那么所有人本来就都知道 p : “至少有一个人脸上有泥巴”. 那么父亲说这句话的意义是什么?

我们可以设想，如果父亲没有说 p ，会发生什么？容易发现，无论父亲问多少轮，所有孩子都只会回答“不知道”！（见习题 1）。因此，这里会产生极其反直觉的事实：即便大家都知道 p ，父亲说不说 p ，会导致完全不同的结果。为什么会这样？

我们重新审视这一问题的过程。

- 假设 $k = 2$ ，脸上沾满泥巴的孩子是 a 和 b 。在父亲宣布 p 之前， a 和 b 都知道 p 。然而，他们并不知道对方知道 p 。 a 可能会有两种想法：
 - 我的脸上有泥巴，所以 b 知道 p 。
 - 我的脸上没有泥巴， b 是唯一一个有泥巴的， b 看不到其他人脸上有泥巴，所以 b 不知道 p 。

当父亲宣布 p 之后， a 知道了 b 知道 p 。当第一轮 b 回答“不知道”之后， a 可以用“ b 知道 p ”这一事实排除第二种情况，从而推出自己脸上有泥巴。

- 假设 $k = 3$ ，脸上沾满泥巴的孩子是 a , b 和 c 。在父亲宣布 p 之前， a , b 和 c 不仅知道 p ，而且知道彼此知道 p 。比如说，以 a 的视角看， b 能看到 c 脸上有泥巴，所以 a 知道 b 知道 p 。

但是， a 不知道 b 知道 c 知道 p ，因为此时有两种情况：

- a 脸上有泥巴， b 能看到 c 和 a 脸上有泥巴，所以 b 知道 c 能看到 a 脸上有泥巴，从而知道 c 知道 p 。
- a 脸上没有泥巴， b 能看到 c 脸上有泥巴， a 脸上没有泥巴，但因为 b 不知道自己脸上有没有泥巴，所以 c 不一定知道 p ， b 不知道 c 知道 p 。

更一般地， a, b, c 都不知道所有人知道所有人知道 p ！然而，当父亲宣布 p 之后， a, b, c 都知道了所有人知道所有人知道 p .

用 $E^m p$ 表示所有人知道所有人知道……所有人知道 (m 次) p . 在一般情况下，父亲没有宣布 p 之前， $E^k p$ 并不成立. 父亲宣布了 p 之后，对任意 $m \geq 1$ ， $E^m p$ 都成立！因此，父亲宣布 p 带来了共同知识. 有了共同知识，这一谜题就可以按照我们所讨论的方式进行下去.

我们曾经假设过所有人“观察力敏锐、聪慧且诚实”. 然而，这一假设并不足够. 上面的论证其实暗含了，所有人都知道所有人“观察力敏锐、聪慧且诚实”，所有人都知道所有人都知道所有人“观察力敏锐、聪慧且诚实”，……换言之，我们需要假设“所有人观察力敏锐、聪慧且诚实”是共同知识.

如果没有这样的假设，上面的论证都将不成立. 例如，还是只有两个孩子 a, b 脸上有泥巴. 假如 a 不知道 b 是诚实的，即便 b 回答了“不知道”， a 也无法从 b 的回答中得到任何额外的知识！

除了假设“所有人观察力敏锐、聪慧且诚实”是共同知识，我们还需要假设以下陈述是共同知识：

- 每个人都能看到所有除自己外的人.
- 每个人都听到了父亲说的话.
- 父亲是诚实的.
- 每个人都在每一轮进行了充分的推理.
-

任何假设的破坏都会导致之前的讨论失效. 那么，为什么父亲宣布 p 就可以让 p 变成共同知识呢？

所有人都听到父亲说 p 并不能产生共同知识. 假如父亲只是对每一个孩子单独宣布 p . 所有人并不知道所有人都知道 p , 因而仅仅可以做到 $E p$, 没有共同知识.

那么, 所有人都知道所有人听到父亲说 p 会如何呢? 进一步假设每个孩子给每一个孩子都安装了窃听器, 每个人都能够偷听每个人与父亲的谈话内容. 所有人并不知道所有人都知道所有人都知道 p , 因而仅有 $E^2 p$.

所以关键在于, 父亲宣布 p 的过程是公开的, 每个人都可以仔细观察别人有没有听到父亲说 p , 也可以观察到别人有没有观察到别人有没有听到父亲说 p , 等等. 此时对每一个 m 都有 $E^m p$.

“泥泞的孩童”谜题足以表明, 关于“知道”的讨论远比想象的复杂. 关于“知道”和知识的研究在哲学中划归为知识论. 我们将介绍处理知识的两种数学模型:

- 一种源自 Aumann, Harsanyi 和 Rubinstein 等人, 以 Bayes 概率论为基础, 是偏经济学的学术风格, 在人工智能中对应了行为主义和连接主义的思路;
- 另一种源自 Kripke, Hintikka 和 Halpern 等人, 以模态逻辑为基础, 是偏计算机科学和哲学的学术风格, 在人工智能中对应了符号主义的思路.

在这一章, 我们主要讨论 Bayes 概率论的方法.

§10.2 不完全信息博弈 (Bayes 博弈)

接下来，我们介绍讨论“知识”的博弈论语言。我们先从正则形式博弈和 Nash 均衡开始说起。正则形式博弈隐含了一个重要的假设：所有玩家对整个世界有一致、完全的共同知识。Nash 均衡建立在这一假设之上：每个玩家可以在博弈结束后，根据其他玩家的策略，确定自己的最优反应。

然而，现实世界中，玩家对世界的认识是有限的，不能获得完全的信息。比如，我们可能不知道对手的收益函数，然而这在现实中极其普遍。

另一个问题是，Nash 均衡只有假设混合策略的情况下才能保证存在。我们在现实中并不真的在选择混合策略：所有的交易其实都是“一锤子买卖”，绝对不可能有人说“我今天以 0.5 的概率花一块钱买你的苹果，0.5 的概率花 5 块钱买你的苹果”。英语也有一句谚语：

“Decision makers do not flip coins in the real world.”

相比之下，纯策略 Nash 均衡更加符合实际。然而，即便是纯策略 Nash 均衡也可能是不合理的状态。考虑如下的二人博弈：

$$\begin{pmatrix} 1, 1 & 0, 0 \\ 0, 0 & 0, 0 \end{pmatrix}.$$

显然，两个人玩家都选择第二策略就达到了纯策略 Nash 均衡。

然而，当行玩家对列玩家的选择有任意小的不确定性时，他都更倾向于选择第一个策略。因此，我们给出的这个纯策略 Nash 均衡实际上描述了一种不太可能出现的状态。

因此，一种 Nash 均衡的修正概念被提出：颤抖的手完美化，它指的是， s 是一个纯策略 Nash 均衡，并且当对手玩家的策略有任何微小不确定性的时侯， s 中的策略依然是最优反应。

“颤抖的手”给了我们一个例子，说明对对手的不确定性会影响玩家的决策。因此，进一步的问题是，如何量化对对手的不确定性？Harsanyi 给出了一个现在已经是“标准答案”的解决方案：引入玩家的“类型”（可能世界）以及其他玩家的对此的先验的信念。他的采用了 Bayes 解释的概率论，信念在数学上被建模为对可能世界的概率分布。

我们先给出不完全信息博弈的定义：

定义 10.1 (不完全信息博弈, Bayes 博弈) 一个不完全信息博弈 (Bayes 博弈) 包含了以下组成部分：

- 玩家集合： I .
- 行动空间： $A = (A_i)_{i \in I}$, A_i 表示玩家 A_i 的所有可能行动.
- 类型空间： $\Theta = (\Theta_i)_{i \in I}$, Θ_i 表示玩家 i 的所有可能类型.
- 收益函数： $u_i : A \times \Theta \rightarrow \mathbb{R}$, 当所有人的行动和类型都确定的时候，玩家 i 能拿到的收益.

所有玩家的行动 $a = (a_i)_{i \in I}$ 形成了一个行动组合，所有玩家的类型 $\theta = (\theta_i)_{i \in I}$ 形成了一个类型组合。 \square

$P_i \in \Delta(\Theta_i)$ 是玩家 i 类型的概率分布。比较不直观的一点是， P_i 表示了其他玩家对玩家 i 类型的信念。因此，Bayes 博弈其实做了简化：

- 不论哪个玩家，对特定玩家 i 的信念是一致的.
- 所有 P_i 是相互独立的，因此玩家的类型之间不会有相互的关联.

因此，玩家 i 在博弈中的全部不确定性都来自其他玩家的类型，他对此的信念是

$$P_{-i} = \prod_{j \neq i} P_j.$$

最后，我们假设玩家 i 知道自己的类型。

更进一步，在一般情况下，玩家 i 对这个世界的信念应该包含：

- 博弈中其他玩家都有谁，
- 自己的可能行动，
- 自己的类型，
- 自己的收益函数，
-

在一个真实的博弈中，以上信息都会是不确定的：对手可以藏在暗处，我们会不知道自己本来拥有的选择，我们可以不了解自己的性格，我们甚至也不知道自己究竟在追求什么。尽管有很多的不确定性最终都可以归结为“类型”（见后面正文的若干例子和习题），Bayes 博弈依然是一个高度理想化的模型。

然而，Bayes 博弈的成功，正如 Robert Aumann 所说，不在于多么贴合实践，而是给了一种系统的方法，让我们可以建模不确定性和信念，从而理解这些概念在博弈中的作用：

“一个典型的例子是 Roth 和 Murnighan (1982) 在完全信息和不完全信息讨价还价方面的实验工作¹……他们将这

¹也就是本章开头讲的故事。

些结果与早期 Fouraker 和 Siegel (1960) 的实验进行了比较.

“Fouraker 和 Siegel 进行了类似的实验，但由于缺乏 Harsanyi 的模型，只能将不完全信息的情况描述为双方都没有被告知对方的收益.

“然而，Roth 和 Murnighan 则从类型的角度详细阐述了不完全信息，并明确考虑了博弈的共同知识方面.”

接下来，我们看一个 Bayes 博弈的例子.

例 10.1 (“工作还是偷懒” 博弈) 在这个博弈中，有两个玩家，他们共同完成一个项目. 两个玩家的行动都是“工作”(W) 或“偷懒”(S). 行玩家的类型集合是单点集，列玩家的类型是“勤奋”(D) 或“懒惰”(L). 收益矩阵为

$\theta_2 = D,$			
	W S		W S
W	3, 3 -1, 0	W	1, 1 -1, 2
S	2, 1 0, 0	S	2, -1 0, 0

换言之，我们其实不确定的只有玩家 2 是喜欢偷懒还是努力工作，但他具体是什么人又会影响双方的收益，从而影响双方的决策. \square

至此，我们已经建立了博弈的语言，下一任务就是定义一个玩家的策略. 和第九章的博弈极为不同，Bayes 博弈中的玩家要面对不确定性. 这给我们定义策略带来了一定的困难. 为此，我们先要讨论清楚，究竟是“不确定性”.

设想如下的情景：你选了一门课，期末要考试. 于是，你认真学习，

并且把往年题都找出来，认真做完，老师也划定了考纲和难度：你对这个考试胸有成竹。真正上场考试的时候，你的心理有一个对难度和考点的预期，因此，你在考试时候面对的是具有明确风险的不确定性。

然而，当你考的不是期末考，而是英语四级考试的时候，你可能就会变得非常随性：挂了还可以重新考，所以考前根本没有学习，甚至连题型都不知道有什么。这个时候，你甚至连四级总分是多少都不知道，所以，你甚至连考试结果的预期都没有。这个时候，你面对的是模糊的不确定性。

还有一种情景，你现在不是考四级，而是考 GRE，这是上机考试。GRE 的每道题的难度都不一样，题目是随机出现的，难度分布极其不均匀，并且他会根据你答题情况来自动调整题目的难度。此时，你面对的是不稳定的不确定性。

上面讲到了三种不确定性。那么，Bayes 博弈是他们中的是哪一种呢？显然，在 Bayes 博弈中，玩家依然只能做一次行动，所以不是不稳定的不确定性。此外，Bayes 博弈中，所有玩家的类型集是固定的，甚至连类型的似然也是确定的，因此，这是具有明确风险的不确定性。

接下来，我们可以定义玩家的策略和理性了。注意，玩家知道自己的类型，但不知道其他人的类型，所以，玩家只能根据自己的类型来决定自己的行为。因此，我们应该定义玩家的策略如下：

定义 10.2 (策略) 玩家 i 的策略是一个映射

$$s_i : \Theta_i \rightarrow A_i,$$

其中 Θ_i 是玩家 i 的类型集合， A_i 是玩家 i 的行动集合。 $s_i(\theta_i)$ 表示玩家 i 在类型 θ_i 下的行动。□

注. 自然，我们也可以定义混合策略，此时， s_i 是一个 Θ_i 到 $\Delta(A_i)$ 的映射。不过，在 Bayes 博弈中，混合策略会让情况（不论概念上还是计算上）变得复杂，所以我们避免混合策略的讨论。

当面对具有明确风险（似然）不确定性的时侯，我们可以遵循 von Neumann-Morgenstern 的期望效用理论，定义玩家的理性。首先，我们定义玩家的事中期望收益：

定义 10.3 (事中期望收益) 玩家 i 在类型 θ_i 下，采取策略 s_i ，对手的策略是 s_{-i} 时， i 的事中期望收益为

$$\tilde{u}_i(s_i, \theta_i, s_{-i}) = \mathbb{E}_{\theta_{-i} \sim P_{-i}}[u_i(s_i(\theta_i), s_{-i}(\theta_{-i}), \theta_i, \theta_{-i})]. \quad \square$$

于是，按照期望理论，玩家的事中理性就是在知道对手的策略之后，会最大化自己的事中期望收益。

据此，我们就可以定义最优反应：

定义 10.4 (事中最优反应) 考虑玩家 i 的策略 s_i ，对手的策略是 s_{-i} ，如果对任意的其他策略 s'_i 和任意类型 θ_i ，都有

$$\tilde{u}_i(s_i, \theta_i, s_{-i}) \geq \tilde{u}_i(s'_i, \theta_i, s_{-i}),$$

那么 s_i 是玩家 i 的事中最优反应。 \square

用最优反应，我们可以很容易定义 Bayes Nash 均衡：

定义 10.5 (Bayes Nash 均衡, BNE) 考虑策略组合 $s = (s_i)_{i \in I}$ ，如果对任意 i, θ_i, a_i 都有

$$\tilde{u}_i(s(\theta_i), \theta_i, s_{-i}) \geq \tilde{u}_i(a_i, \theta_i, s_{-i}),$$

那么 s 是一个 **Bayes Nash 均衡 (BNE)** . □

接下里，我们来看一个 BNE 的例子.

例 10.2 (猜硬币游戏) 考虑经典的猜硬币游戏（见例 9.3），这是一个正则形式博弈，收益矩阵为

	H	T
H	1, -1	-1, 1
T	-1, 1	1, -1

如果两个人都出 H 的时候收益有独立微小的扰动，我们就得到了一个 Bayes 博弈：

	H	T
H	1 + $\epsilon\theta_1$, -1 + $\epsilon\theta_2$	-1, 1
T	-1, 1	1, -1

其中 $\theta_i \sim \mathcal{U}[-1, 1]$ 且相互独立，这就是玩家 i 的类型.

那么，这个博弈的 BNE 会是什么呢？我们可以猜一个. 考虑策略：
 $s_i : [-1, 1] \rightarrow \{H, T\}$ 满足

$$s_i(\theta_i) = \begin{cases} H, & \theta_i \in [0, 1], \\ T, & \theta_i \in [-1, 0). \end{cases}$$

我们来验证，这的确是一个 BNE.

固定列玩家的策略 s_2 ，我们来计算行玩家的最优反应. 对于行玩家来说，不论他的类型是什么，他面对的是一个 50% 的概率选择 H 和 T

的对手, 因此, 如果选择 H , 行玩家的期望收益是

$$\frac{1}{2} \cdot (1 + \epsilon\theta_1) + \frac{1}{2} \cdot (-1) = \frac{\epsilon\theta_1}{2}.$$

如果选择 T , 行玩家的期望收益是

$$\frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot (1) = 0.$$

因此, 如果 $\theta_1 > 0$, 行玩家应该选择 H , 如果 $\theta_1 < 0$, 行玩家应该选择 T , 这是一个最优反应, 恰好是我们猜测的策略.

对于列玩家的最优反应, 我们可以做类似的计算, 也可以得到, s_2 是列玩家的最优反应. 因此, (s_1, s_2) 是一个 BNE. \square

以上例子有极其特殊的含义: 策略 (s_1, s_2) 导致的结果实际上是, 每个玩家计算最优反应的时候, 面对的对手其实仿佛是一个混合策略玩家, 他以等概率选择 H 和 T . 注意, 当 $\epsilon \rightarrow 0$, 这个博弈收益矩阵回到了原始博弈. 因此, Bayes 博弈里行动概率分布其实可以被视作原始博弈的混合策略.

猜硬币游戏的例子其实说明, 正则形式博弈的混合策略 Nash 均衡被理解为: 当不确定性趋于消失时候, BNE 形成的行动概率分布. 这不是偶然的, 实际上, 所有的正则形式博弈的混合策略均衡都可以用一系列 Bayes 博弈的 BNE 纯化.

下面我们把猜硬币游戏的过程一般化. 考虑一个正则形式博弈 (I, A, u) , 其中 I 是玩家集合, A 是行动空间, u 是收益函数. 我们可以定义一个 Bayes 博弈, 被称为扰动博弈:

- 给定一个扰动参数 $\epsilon > 0$, 定义类型组合为 $\theta = (\theta_i)_{i \in I}$, $\theta_i \in$

$[-1, 1]$, 然后, 将收益扰动为:

$$u'_i(s, \theta) = u_i(s) + \epsilon \theta_i.$$

- 假设 $\theta_i \sim F_i$, 相互独立, F_i 是具有连续可微密度的分布.

定义一个从 Bayes 博弈策略到正则形式博弈混合策略的映射 ϕ , 给定一个 Bayes 博弈的策略组合 s , $\phi(s)$ 是一个混合策略, 满足

$$\phi(s)_i(a_i) = \Pr_{\theta_i \sim F_i} [s_i(\theta_i) = a_i].$$

换言之, 在原本的 Bayes 博弈中, 概率定义在了玩家的类型上, 这个映射的作用是将这个概率转化为了玩家的行动上.

接下来, 我们可以正式给出纯化定理的表述:

定理 10.1 (Harsanyi 纯化定理) 给定玩家集 I 和行动空间 A . 对于一般的收益函数 u 和连续分布族 $\{F_i\}_{i \in I}$, 对任意完全信息正则形式博弈 (I, A, u) 的混合策略 Nash 均衡 σ , 存在一列扰动博弈纯策略 $BNE s_\epsilon$, 当扰动参数 $\epsilon \rightarrow 0$, $\phi(s_\epsilon) \rightarrow \sigma$.

这一定理的证明比较长, 并且需要用到较为复杂的数学技巧. 由于证明本身与本章的讨论无关, 所以这里略去.

这一定理给了 Nash 均衡 (也即混合策略) 一种新的解释: 混合策略定义的 Nash 均衡可以被看作不确定性趋于消失的时候的 (纯策略) BNE.

尽管我们说, “Decision makers do not flip coins in the real world.” 然而, 如果玩家对自己的收益有微小的不确定性, 他的行为就会仿佛在抛硬币. 这就是混合策略的似然解释.

注意，在 Bayes 博弈中，玩家对自己的类型是确定的，所以玩家在决策时不应该对自己的收益有微小的不确定，因而上面这一解释并不完全正确。然而，我们可以重新定义理性的概念，它会产生等价的 BNE 定义，但是玩家此时要面对自己类型的不确定性。

为了说明这一理性的概念，我们先看一个例子。假如你是一个顶尖斗地主玩家，知道自己拿到身份、拿到手牌之后要如何行动，这是你一贯的策略。当然，在游戏开始前，你其实并不知道自己的身份和手牌，这是你对自己的不确定性。不过，这并不影响你评估自己的胜率：你只需要知道其他玩家的水平，你就可以大概估计一个总体的胜率。

上面的例子里，玩家其实在博弈开始前就已经选好了一个类型到行动的策略，但是他要面临自己类型的不确定性。此时，我们计算的胜率其实是事前期望收益：

定义 10.6 (事前期望收益) 给定玩家集 I 、行动空间 A 和类型上的联合分布 P ，如果玩家 i 的策略是 s_i ，对手的策略是 s_{-i} ，那么 i 的事前期望收益为

$$\hat{u}_i(s_i, s_{-i}) = \mathbb{E}_{\theta \sim P}[u_i(s_i(\theta_i), s_{-i}(\theta_{-i}), \theta_i, \theta_{-i})].$$

□

在 Bayes 博弈中，事前理性指的就是，玩家在不知道自己的类型的情况下，最大化自己的事前期望收益。根据事前期望收益，我们可以定义事前最优反应：

定义 10.7 (事前最优反应) 考虑玩家 i 的策略 s_i ，对手的策略是 s_{-i} ，如果对任意的其他策略 s'_i 和任意类型 θ_i ，都有

$$\hat{u}_i(s_i, s_{-i}) \geq \hat{u}_i(s'_i, s_{-i}),$$

那么 s_i 是玩家 i 的事前最优反应。

□

注意，事前理性和事中理性考虑的都是同样的策略，但是玩家面对的不确定性是不同的。然而，事前理性和事中理性其实是等价的：

定理 10.2 给定玩家集 I 行动空间 A ，类型空间 Θ 、收益函数 u 和类型空间的联合分布 P ，考虑策略组合 $s = (s_i)_{i \in I}$ ，对任意玩家 i ， s_i 是 s_{-i} 的事前最优反应当且仅当 s_i 是 s_{-i} 的事中最优反应。

这一定理的证明类似正则形式博奕无差别原理（定理 9.7）的证明，见习题 4.

这一定理其实有一些反直觉：在 Bayes 博奕中，如果玩家面对的仅仅只是具有确定风险的不确定性，那么他是否不确定自己的类型并不会影响他的理性决策。这其实是因为，玩家面对的仅仅是具有确定风险的不确定性，而不是更加复杂的不确定性，因此，无论博奕如何进行，他在开始前就已经可以确定自己的最优策略。

这一定理的直接推论是，我们可以根据事前最优反应来定义 BNE：

$$\hat{u}_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$$

对任意 i 和任意策略 s'_i 成立。

当所有的不确定性都消失的时候，我们得到的收益是真实的，被称为事后收益，此时不再有任何的概率，因此博奕退化为了正则形式博奕。事前、事中、事后分别表明了信息的确定（披露）程度。

§10.3 电子邮件博弈

作为一个例子，接下来我们使用 Bayes 博奕来研究知识的性质。这个例子由 Ariel Rubinstein 给出，它说明在二人正则形式博奕中，共同

知识对到底实现哪一个 Nash 均衡非常关键.

考虑两个玩家和两个可能的收益矩阵:

	A	B		A	B
A	(0, 0)	(-10, 1)	A	(8, 8)	(-10, 1)
B	(1, -10)	(8, 8)	B	(1, -10)	(0, 0)

在左边的矩阵中, (B, B) 是唯一的 Nash 均衡. 在右边的矩阵中有多个 Nash 均衡: (A, A) 和 (B, B) . (A, A) 给出比 (B, B) 更高的收益, 但行动 A 比 B 更有风险.

左边矩阵是真实矩阵概率是 $p > 1/2$. 玩家 1 知道真实的矩阵, 而玩家 2 不知道. 如果选择了右边矩阵, 玩家 1 会给玩家 2 发送一条消息. 如果玩家 2 收到了消息, 他会回复. 如果玩家 1 收到了回复, 他会发送第二条消息来确认他收到了玩家 2 的回复. 以此类推. 每条消息都以 ϵ 的概率独立等可能丢失.²

以上传信的过程可以用 Bayes 博弈的类型来刻画. 具体来说, 两个玩家的类型集合为 $\Theta_i = \{\theta_i^0, \theta_i^1, \theta_i^2, \dots\}$. θ_i^m 表示玩家 i 发了 m 封邮件. θ_i^m 有直观的含义. 例如, 类型 θ_1^0 表示真实收益矩阵是左边的. 而类型 θ_1^1 表示真实收益矩阵是右边的, 1 发送了一封电子邮件, 但 2 没有收到.

实际上, θ 包含了所有可能的情况:

- (θ_1^0, θ_2^0) : 真实收益矩阵是左边的.
- (θ_1^1, θ_2^0) : 真实收益矩阵是右边的, 1 发送了一封电子邮件, 但 2 没有收到.

²注意: 发送电子邮件不是一个行动, 而是一个规则.

- (θ_1^1, θ_2^1) : 真实收益矩阵是右边的, 2 收到了第一封电子邮件, 但 1 没有收到 2 的回复.
-

我们可以算出来, 当真实矩阵为左边矩阵时, 每个类型出现的概率. 首先, 以概率 p 选择左边的矩阵, 而且没有人发送消息. 因此, (θ_1^0, θ_2^0) 的概率是 p , 其他项概率都是 0. 我们得到如下的概率表:

左	θ_2^0	θ_2^1	θ_2^2	...
θ_1^0	p	0	0	...
θ_1^1	0	0	0	...
θ_1^2	0	0	0	...
\vdots	\vdots	\vdots	\vdots	\ddots

同样可以算出来, 当真实矩阵为右边矩阵时, 每个类型出现的概率. 首先, 以概率 $1 - p$ 选择右边的矩阵, 玩家 1 发送一条消息, 它会以概率 ϵ 丢失. 因此, (θ_1^1, θ_2^0) 的概率是 $\epsilon(1 - p)$. 以此类推, 可以得到计算. 我们得到如下的概率表:

右	θ_2^0	θ_2^1	θ_2^2	...
θ_1^0	0	0	0	...
θ_1^1	$\epsilon(1 - p)$	$\epsilon(1 - \epsilon)(1 - p)$	0	...
θ_1^2	0	$\epsilon(1 - \epsilon)^2(1 - p)$	$\epsilon(1 - \epsilon)^3(1 - p)$...
θ_1^3	0	0	$\epsilon(1 - \epsilon)^4(1 - p)$...
\vdots	\vdots	\vdots	\vdots	\ddots

容易看出来, 当真实类型为 θ_i^m 时, 收益矩阵是到第 m 层的共同知

识, 即 E^m . 所以对于很大的 m , 收益矩阵是“几乎公共知识”. 所以, 这个模型在研究的问题是: 如果收益矩阵是“几乎共同知识”, 那么 Nash 均衡是什么? 为此, 我们需要求出来这个 Bayes 博弈的 BNE.

我们需要弄清楚对每个类型 θ_i^m , 玩家会做什么. 假设玩家 1 的类型为 θ_1^0 . 玩家 1 知道 (θ_1^0, θ_2^0) 是真实的类型, 所以左边的矩阵被选择. 据此推理: 玩家 1 选择占优策略 B .

假设玩家 2 的类型为 θ_2^0 . 我们对玩家 1 的所有可能类型分类讨论:

- 如果玩家 1 的类型为 θ_1^0 , 那么左边的矩阵被选择, 对于玩家 2 来说, 这种情况的概率为

$$\Pr(\theta_1^0 | \theta_2^0) = \frac{p}{p + \epsilon(1 - p)} := \mu_2^0.$$

- 如果玩家 1 的类型为 θ_1^1 , 那么右边的矩阵被选择, 对于玩家 2 来说, 这种情况的概率为

$$\Pr(\theta_1^1 | \theta_2^0) = 1 - \mu_2^0.$$

现在我们来分析玩家 2 的两种选择: A 和 B .

- 选择 B 的期望收益至少是 $8\mu_2^0$. 推理如下:

- 玩家 1 的类型是 θ_1^0 时, 这是左边的矩阵, 玩家 1 肯定选择 B , 此时玩家 2 选择 B 的收益是 8.
- 玩家 1 的类型是 θ_1^1 时, 这是右边的矩阵, 无论玩家 1 怎么选, 此时玩家 2 选择 B 的收益至少是 0.

因此, 按照全概率公式计算, B 的期望收益至少是 $8\mu_2^0$.

- 选择 A 的期望收益至多是 $-10\mu_2^0 + 8(1 - \mu_2^0)$. 推理如下:
 - 玩家 1 的类型是 θ_1^0 时, 这是左边的矩阵, 玩家 1 肯定选择 B , 此时玩家 2 选择 A 的收益是 -10 .
 - 玩家 1 的类型是 θ_1^1 时, 这是右边的矩阵, 无论玩家 1 怎么选, 此时玩家 2 选择 A 的收益至多是 8 .

因此, 按照全概率公式计算, A 的期望收益至多是 $-10\mu_2^0 + 8(1 - \mu_2^0)$.

注意,

$$\begin{aligned} & 8\mu_2^0 - (-10\mu_2^0 + 8(1 - \mu_2^0)) \\ &= 10\mu_2^0 - 8 \\ &= \frac{10p - 8(p + \epsilon(1 - p))}{p + \epsilon(1 - p)} \\ &= \frac{(2 + \epsilon)p - 8\epsilon}{p + \epsilon(1 - p)} \\ &> \frac{1 - 8\epsilon}{p + \epsilon(1 - p)}. \end{aligned}$$

对充分小的 $\epsilon > 0$, 这个值是正的. 因此, B 更好.

假设玩家 1 的类型为 θ_1^1 , 于是, 右边的矩阵被选择. 同样, 对玩家 2 的类型分类:

- 如果玩家 2 的类型为 θ_2^0 , 对于玩家 1 来说, 这种情况的概率为

$$\Pr(\theta_2^0 | \theta_1^1) = \frac{\epsilon(1 - p)}{\epsilon(1 - p) + \epsilon(1 - \epsilon)(1 - p)} := \mu_2^1.$$

- 如果玩家 2 的类型为 θ_2^1 , 对于玩家 1 来说, 这种情况的概率为

$$\Pr(\theta_2^1 | \theta_1^1) = 1 - \mu_2^1.$$

同样可以计算玩家 1 的两种选择的期望收益:

- 选择 B 的期望收益至少为 0. 推理如下: 玩家 2 是类型 θ_2^0 时肯定选择 B (上面已经证明), 因此最坏的情况是玩家一类型为 θ_2^1 .
- 选择 A 的期望收益至多为 $-10\mu_2^1 + 8(1 - \mu_2^1)$. 推理如下: 玩家 2 是类型 θ_2^0 时肯定选择 B (上面已经证明), 因此最好的情况是玩家一类型为 θ_1^2 .

综合两方面, B 更好, 因为对于所有 ϵ , $\mu_1^1 > 1/2$.

逐步迭代上述过程, 我们发现, 在唯一的 BNE 中, 所有玩家在所有类型下都选择 B .

然而, 如果右边的矩阵是共同知识, (A, A) 也是一个真正的 Nash 均衡, 然而, 因为对于收益矩阵的不确定性, 即便收益矩阵是“几乎共同知识”, 这个均衡也不会被实现! 这个例子说明了, 共同知识对于均衡的实现是非常关键的. 我们在习题 5 中会进一步讨论 Nash 均衡与共同知识的关系.

§10.4 Aumann 知识算子

在上一节中, 我们给了一个非常具体的例子探讨共同知识和 Nash 均衡的关系. 上面的例子看起来太特定, 那是不是说明, 用 Bayes 博弈研究知识, 就是得具体问题具体建模呢? 这个问题的答案, 是也不是.

一方面，用具体的 Bayes 博弈去说明具体的道理往往简洁且富有内涵；另一方面，我们也可以用更加抽象的方式去研究知识。这一节，我们将介绍 Aumann 的知识算子，它是一种抽象的方式去研究知识。

首先，我们需要第一章的观点：概率论（或者似然）理解世界的方式基于“事件”。我们只能感知事件的发生与否，而不能具体知道是哪个样本点。用事件的方式理解认知，得到的结构被称为 *Aumann* 结构。下面，我们具体介绍这一数学模型。

考虑全集 Ω ，它的含义有多种多样，比如可以理解为样本空间（概率论视角）、状态空间（动态博弈视角）或者可能世界（信念的视角出发）的全体。我们就具体考虑为，从一个罐子里抽球， Ω 是这次抽到球的颜色。

对于最后一种关于 Ω 的理解，我们可以再多解释一些。当我们面对不确定性的时侯，我们会有一些信念，比如“这个球是红色的”、“这个球是蓝色的”等等。实际上，这些不同信念背后对应了一个不同的“世界”。比如，我们考试的时候，我们可能会幻想考试通过时的场景，也可能幻想考试不通过时的场景。他们代表了这个现实世界在未来不同的走向，因而被称为可能世界。

注意，上面的解释容易引起误会，实际上，可能世界完全可以不是未来的世界。例如，我们常常会做反事实因果的推理，比如“如果我当时不天天吃油炸食品，那么我现在就不需要减肥了”。这个“如果”引导我们进入了一个可能世界，这个可能世界甚至存在于过去，而不是未来。

我们不再过多讨论 Ω 的具体含义，而是专注于罐子的例子。事件 $e \subseteq \Omega$ 是样本点的集合，它表示了某些性质的发生。例如， e 可能是“抽到了红球”，或者“抽到的不是白色”等等。一次观测会产生一个具体的

颜色，这个颜色对应了 Ω 中的一个元素 ω ，事件 e 发生当且仅当 $\omega \in e$.

接下来，我们进入博弈的部分. 我们会有集合 I 的玩家，每个玩家 i 都有一个关于 Ω 的知识. 例如，色觉正常的人会知道红色和绿色是不同的，而红绿色盲的人可能不知道. 因此，每个人会有不一样的知识. 如何用数学语言来描述这种知识呢？

实际上，知道一件事 e 发生与否意味着这个玩家有能力区分一个样本点 ω 是否属于 e . 所以，我们可以先定义玩家知道的“基础知识”，即他能够感知到的最基本（原子）事件. 数学上，我们给如下的定义：

定义 10.8 (信息集) 对于每一个玩家 i ，他的信息集是 Ω 的一个划分

$$\mathcal{P}_i = \{\Omega_j\}_j.$$

划分的含义是， Ω_j 满足

- $\Omega_j \neq \emptyset$,
- 对所有 $j \neq k$, $\Omega_j \cap \Omega_k = \emptyset$ ，并且
- $\bigcup_j \Omega_j = \Omega$.

此外， $\mathcal{P}_i(\omega)$ 被定义为 ω 所属于的那个信息集. □

对于玩家 i 来说，他无法区分 Ω_j 中元素，这就是原子性的含义. 换言之，如果他能区分，就说明这个 Ω_j 还不够小，因为它还是包含了可区分的不同元素.

接下来，我们来定义知识中最基本的陈述，即“玩家 i 知道事件 e ”. 我们还是用红绿色盲作为例子. 显然，红绿色盲不知道事件 e : “这个球是红色的”. 这是因为，他的信息集中没有单独的红色，只有“红色或绿色”这一集合.

上面的例子说明，玩家 i 知道事件 e ，说明对于这一个特定的观测 ω ，玩家 i 能够完全确定 e 发生了。注意，此时玩家 i 能够观察到的最基本事件是 $\mathcal{P}_i(\omega)$ ，因此，这说明

$$\mathcal{P}_i(\omega) \subseteq e.$$

我们可以把所有能够完全确定 e 发生的 ω 收集起来，在这些 ω 上，玩家 i 充分必要地知道 e 发生。因此，我们有如下的定义：

定义 10.9 (Aumann 知识算子) 对于每一个玩家 i ，我们定义 **Aumann 知识算子** $K_i : 2^\Omega \rightarrow 2^\Omega$ 为

$$K_i(e) := \{\omega \in \Omega : \mathcal{P}_i(\omega) \subseteq e\}.$$

因此 $K_i(e)$ 是一个事件，表示“个体 i 知道事件 e ”。不引起混乱的时候我们也省略括号，写作 $K_i e$. □

Aumann 知识算子的想法很清晰，它把所有关于知识的讨论转化为关于事件的讨论。在第一章似然的讨论中，我们建立了集合论和逻辑的对应关系，如表 10.1 所示。

更进一步，我们可以将多个玩家的知识结合起来，定义共同知识算子：

定义 10.10 (共同知识算子) 定义“所有人都知道”算子 $E : 2^\Omega \rightarrow 2^\Omega$ 为

$$E(e) = \bigcap_{i=1}^n K_i(e).$$

事件	命题
Ω	\top
\emptyset	\perp
$\sim A$	$\neg A$
$A \cap B$	$A \wedge B$
$A \cup B$	$A \vee B$
$A \subseteq B$	$A \rightarrow B$
$A = B$	$A \leftrightarrow B$

表 10.1: 事件和命题的对应关系

然后，我们可以定义共同知识算子 $C : 2^\Omega \rightarrow 2^\Omega$ 为

$$C(e) = \bigcap_{k=1}^{\infty} E^k(e).$$
□

接下来，我们讨论知识的一些基本性质。这里，我们只讨论知识本身的性质，而不关心不同玩家之间知识的交互，因此，我们都把 K_i 写作 K 。

命题 10.1 (意识公理)

$$K(\Omega) = \Omega. \tag{K0}$$

证明。显然成立。 □

意识公理意味着，每个人知道他在某一个状态（可能世界）中。尽管是一个显然的公理，我们依然可以考虑“意识不到自己处于某个状态”的情况，见习题 6。

命题 10.2

$$K(e \cap f) = K(e) \cap K(f). \quad (\text{K1})$$

证明. 我们证明两边相互包含.

- \subseteq : 设 $\omega \in K(e \cap f)$, 则

$$\mathcal{P}(\omega) \subseteq e \cap f.$$

因此, 同时成立

$$\mathcal{P}(\omega) \subseteq e, \quad \mathcal{P}(\omega) \subseteq f.$$

因此, 根据知识算子的定义,

$$\omega \in K(e), \quad \omega \in K(f).$$

因此,

$$\omega \in K(e) \cap K(f).$$

- \supseteq : 将上面的证明倒着写一遍即可. □

这条性质意味着, 一个人知道事件 e 和事件 f , 当且仅当他知道事件 $e \cap f$. 考虑 $e \cap f = f$ 的特殊情况, 我们可以得到

$$e \subseteq f \implies K(e) \subseteq K(f).$$

换言之, 如果客观事实上 e 可以推出 f (即 $e \subseteq f$), 那么个体知道 e 就可以推出他知道 f . 这条性质意味着玩家是逻辑全知的, 他可以对知识做

任意复杂的逻辑推理。特别地，他甚至可以做关于知识的逻辑推理！接下来，我们很快就会看到这会带来多么复杂的情况。

命题 10.3 (知识公理, 真理公理)

$$K(e) \subseteq e. \quad (\text{K2})$$

知识公理意味着，知道的一定是真的。在知识论中，这一要求实际上反映了“拥有知识”需要付出努力、值得一定的奖励。与此相对应地，信念则是更加主观、随意的，因而并不具有真理性。我们可以通过下面两句话来体会这两者的区别：

- 我考试挂了，但不知道我考试挂了。
- 我考试挂了，但我不相信我考试挂了。

命题 10.4 (正内省公理)

$$K(e) \subseteq K(K(e)). \quad (\text{K3})$$

命题 10.5 (负内省公理)

$$\sim K(e) \subseteq K(\sim K(e)). \quad (\text{K4})$$

这两条内省公理意味着，个体会通过内省来知道自己的处境，特别是“我知道什么”和“我不知道什么”，通过内省，个体可以产生更高层级的知识：“我知道自己知道什么”或者“我知道自己不知道什么”。

值得注意的是，内省这样的能力是灵长类动物区别于其他动物的重要特征。例如，实验人员给黑猩猩若干工具，只有其中一种工具可以解

决问题. 黑猩猩在没有任何提示的情况下, 第一次就能够选出正确的工具. 科学家认为, 这表明黑猩猩有能力内省, 即知道自己知道什么 (特别是物理世界的因果关系), 并且利用这些知识来解决问题. 然而, 诸如松鼠这样的哺乳动物, 就没有这样的能力.

相比正内省公理, 负内省公理可能会引起一些争议. 例如, 在数学界有一个 “著名” 的猜想, 叫几何 *Langlands* 猜想. 这个猜想是一个非常复杂的数学问题, 最近, 一个由 9 位数学家组成的团队成功证明了这个猜想. 这样专门的数学猜想, 对于非数学爱好者来说, 确实是不知道的. 然而, 如果负内省公理是成立的, 那么他们就应该知道自己不知道这个猜想. 这显然是不合理的.

上面的很多命题我们都会称呼为 “公理”, 这是因为他们反映了知识的基本性质, 是我们对知识的直觉认识. 实际上, 我们可以反过来, 从这些公理出发, 推导出 Aumann 知识算子的定义. 于知识的逻辑推理!

定理 10.3 考虑一个映射 $K : 2^\Omega \rightarrow 2^\Omega$, 那么, 以下两条等价:

- K 满足 (K0)–(K4).
- 存在一个信息集划分 \mathcal{P} , 使得 K 是由这个信息集划分定义的 *Aumann* 知识算子.

这一定理的证明类似于上面公理的证明, 见习题 7.

诚然, Aumann 知识算子的定义 (在接受了之后) 是非常直观且自然的. 然而, 定理 10.3 告诉我们, 这样定义的知识算子, 必须要承认很多不是特别合理的性质, 例如负内省公理或者逻辑全知. 这似乎是一个两难的选择.

因此, 回到本节开头的问题, 用 Bayes 博弈研究知识, 是不是得具体问题具体建模呢? Aumann 知识算子给了一个很好的回答: 如果过分

一般，就会过分简化知识的概念，引入很多不合理的性质³. 因此，既需要具体问题具体建模，也需要对知识的一般性质有所了解.

至此，在 Aumann 结构下，我们已经给知识一个清晰的定义，作为本节的结束，我们给出 Aumann 结构中信念的定义. 同样，回忆第一章的思想，在基于事件的认知理论中，我们很容易通过概率（似然）来定义信念.

玩家 i 可以对信息集 $\mathcal{P}_i(\omega)$ 中的状态形成信念. 设 ρ_i 为集合 Ω 上的概率分布，它代表 i 的先验信念，即在没有任何额外信息的情况下他会持有的信念. 如果 $\rho_i(\mathcal{P}_A(\omega)) > 0$ ，那么 i 在状态 ω 处对世界状态的信念由以下概率分布给出

$$\rho_i(e|\omega) = \rho_i(e|\mathcal{P}_A(\omega)) = \frac{\rho_i(e \cap \mathcal{P}_A(\omega))}{\rho_i(\mathcal{P}_A(\omega))}.$$

接下来，我们对这一定义做几点说明.

首先，这一定义是良定义的，因为对于所有的 $\omega' \in \mathcal{P}_A(\omega)$ ， $\rho_i(e|\omega')$ 都是一样的.

其次，直观上说，这一定义相当于后验信念：每个人根据自己知道的信息做了 Bayes 更新.

最后，这个定义与 Aumann 知识算子是相容的. 事实上，我们可以定义“知道”是“具有必然的信念”：

$$K_i(e) = \{\omega \in \Omega : \rho_i(e|\omega) = 1\}.$$

³实际上，直到今天，做知识论的哲学家普遍持有这样的观点：知识是一个无法被严格定义的哲学概念！因此，当知识被更进一步形式化成数学模型，我们更不能迷信它的普适性.

§10.5 习题

- 在泥泞的孩童中，如果父亲没有说 p ，请简要说明，无论经过多少轮询问，所有孩子都只会回答“不知道”。
- 在 Bayes 博弈的定义中，我们假设所有玩家的可能行动集和可能类型是所有玩家都是共同知识。假设玩家的真实行动集可能依赖玩家的类型，也就是说玩家 i 具有不同类型的时候真实行动集可能不同。如果玩家对于世界的不确定性只在于其他玩家的类型，而不在乎真实行动集，证明：每个玩家 i 的真实行动集 $A_i(\theta_i)$ 实际上不依赖它的类型，即对 i 的所有类型 θ_i ， $A_i(\theta_i)$ 是同一个集合。
- 考虑工作偷懒二人 Bayes 博弈。两个人的行动都是“工作”（W）或“偷懒”（S）。行玩家的类型集合是单点集，列玩家的类型是“勤奋”（D）或“懒惰”（L）。收益矩阵为

$\theta_2 = D,$	$\theta_2 = L,$
$\begin{array}{c cc} & W & S \\ \hline W & 3, 3 & -1, 0 \\ S & 2, 1 & 0, 0 \end{array}$	$\begin{array}{c cc} & W & S \\ \hline W & 1, 1 & -1, 2 \\ S & 2, -1 & 0, 0 \end{array}$

假设 $\Pr(\theta_2 = D) = p$, $0 < p < 1$, 计算它的所有纯策略 BNE. 这一结果给你带来了什么启示?

- 仿照正则形式博弈无差别原理（定理 9.7）的证明，证明定理 10.2.
- * 达成 Nash 均衡的知识条件. 请查阅资料，形式化建模并证明下面的事实：

- (1) 在二人博弈中，假设所进行的博弈（即双方的收益函数）、玩家的理性以及他们对别人策略的信念都是相互知识，那么这些信念就构成了一个 Nash 均衡。
- (2) 在多人博弈中，假设所进行的博弈（即所有玩家的收益函数）、玩家的理性以及他们对别人策略的信念都是共同知识，那么这些信念就构成了一个 Nash 均衡。
- (3) 但是，哪怕在三人博弈中，如果是上述事实只是任意有限阶的相互知识，这些信念都不一定构成一个 Nash 均衡。

因此，二人博弈和多人博弈达成 Nash 均衡的知识条件是本质不同的。

6. 如果我们写一个算子 $U(e)$ 表示“个体意识不到事件 e ”，它应该有性质如下两个性质。

- $\sim KU(e) = \Omega$: 他不知道他不能意识到他不能意识到的事件。
- $U(e) \subseteq \sim K(\sim K(U(e)))$: 如果他不能意识到事件 e ，那么他不知道他不知道他不能意识到事件 e 。

这样的算子 U 存在吗？

7. * 证明定理 10.3.

8. * 请查阅资料，回答以下关于 Nash 均衡的问题：

- (1) 用 $\text{Nash}(x)$ 表示“ x 是 Nash 均衡”，那么公式 $\exists x C(\text{Nash}(x))$ 和公式 $C(\exists x \text{Nash}(x))$ 的含义是否一样？
- (2) 如果玩家不是逻辑全知的，或者说他的推理、计算能力是有限的，那么 Nash 均衡还是会达到？是否可接近？

第十一章 模态逻辑，知识的逻辑

你听说过魔芋与高僧的故事吗？相传，这个有趣的故事发生在日本的一座深山之中。在那幽静的山谷中，有一座古老而荒废的寺庙，几百年来一直无人问津，只有密林与野兽相伴。

一天，一个以卖魔芋为生的流浪小贩推着他的木车，偶然来到这片偏僻的地方。他发现这座破旧的寺庙中竟然无人居住，便心想：“如此偏僻之地，不如就在此安顿下来吧。”于是，他便住了下来，每天挑着魔芋，走村串户，卖给山中偶尔路过的行人。

不久之后，一位云游四方的僧人也来到了这片深山。看到这座荒废的寺庙，他心生好奇，便前去查看。僧人惊讶地发现寺庙中竟然住着一位老人，便以为这是位得道高僧。僧人心中一动，想着既然能在如此偏远之地修行，必定有过人之处，于是便恭敬地上前，请教佛法。

在佛教传统中，有一种不动声色的辩经方式，那就是只用手势来比划，而不发一言。僧人遵循这一传统，开始用手势向卖魔芋的人比划起来。然而，卖魔芋的人哪里知道这些佛教传统？他误以为僧人是在挑剔

他卖的魔芋不好，来讨说法的。他心中不悦，但还是用手势回应僧人，心想着：“我卖的魔芋可是顶好的，若是敢挑衅，我便和你理论理论！”

于是，两人各怀心事，展开了一场激烈的手势比划。僧人暗自佩服卖魔芋的人的沉稳与深不可测的手势，心想自己在佛法修行上还有很大的差距；而卖魔芋的人则越比越得意，心中大呼：“这人还真有些手段，但我魔芋的品质岂是一般人能质疑的？”

比划结束后，僧人心生敬意，深感不如；卖魔芋的人则自信满满，觉得自己的魔芋赢得了这场对峙。因此，“卖魔芋的人获胜”这件事成了他们二人的共同知识。然而，两人心中所思所想却完全不同！僧人认为自己遇到了高人，而卖魔芋的人则确信自己击退了一个挑刺的顾客。

在第十章中，我们探讨了知识的一些基本性质。上面的魔芋故事实际上揭示了两件重要的事情：

- 一方面，知识（尤其是共同知识）要比我们想象的复杂得多。在这个例子中，两个人知道的“事实”究竟是不是同样的呢？这个问题并不容易回答。
- 另一方面，知识的传递与理解可以超越自然语言。在这个例子中，两人完全没有用语言交流，而是依靠手势。这说明知识可以有不同于自然语言的表达方式。

在本章中，第二点将是我们的讨论重点。它意味着我们可以用形式逻辑来描述知识的概念和性质。这种形式逻辑在现代逻辑学中被归类为模态逻辑。模态逻辑是一种扩展了经典形式逻辑（如命题逻辑和一阶逻辑）的逻辑，通过引入模态词——例如“必然”、“可能”、“知道”等——来描述命题的性质。最后，作为一个例子，我们将使用模态逻辑的语言表述并证明 Robert Aumann 的“对不一致达成一致”定理。

§11.1 知识逻辑的形式语言

模态词是指说话者对语句的限定. 不同的限定反映了不同的逻辑状态. 例如: 必然、可能、过去、未来、知识、信念和可证明等, 都是典型的模态概念. 比如, 我们可以说“明天可能会下雨”、“我知道明天会下雨”、“我相信明天会下雨”、“明天可能会下雨”等等, 他们要表达的意思并不是经典逻辑可以表述的.

在经典逻辑中, 我们只能表达“明天下雨”或者“明天不下雨”这样的命题. 因而, 模态逻辑提供了一套严密的数学工具, 可以把模态的概念从自然语言中“搬到”形式语言中.

因为模态词往往都与人的认知和思维有关, 因此, 通过模态逻辑, 我们可以算法化、自动化地模拟和推理人类的思维和认知过程. 这正是人工智能中符号主义的思想.

在研究模态逻辑时, 重要的一部分是研究它的模型论, 即如何定义模态逻辑的语言、然后赋予这些形式语言具体的含义与真假, 即语义.

模型论的角度看, 一个逻辑系统包括三个要素: 语言、模型、语义. 他们分别代表, 我怎么说话、我关心的对象是什么、我说的话和我关心的东西之间有什么关系. 接下来, 我们以命题逻辑为例, 介绍三要素.

例 11.1 (命题逻辑的逻辑三要素) 命题逻辑是由命题字母、逻辑联结词和括号组成的形式语言. 例如, $p \rightarrow q$ 这样的东西属于命题逻辑, 但是 $\forall x P(x)$ 这样的东西不属于命题逻辑 (而属于一阶逻辑). 关于命题逻辑更系统的讨论, 请参阅第一章.

命题逻辑的三要素是:

- 语言 (我怎么说话): 用规则写成的字符串, 例如我们可以用字符串 $p \rightarrow q$ 表示一个逻辑公式, 但是不能写 $\forall x P(x)$ 、 $2 + 2 = 4$ 或

者 $p \vee \vee q$. 我们可以简单理解成，我们说普通话的时候，不能说英语，也不能说语法错误的句子（如“我是是”）.

- 模型（我们关心的对象是什么）：我们关心的世界结构，在命题逻辑中就是“真”和“假”. 因此，在命题逻辑的世界里，没有实数，也没有人类，更没有诗和远方；只有冰冷（但是精确）的“真”和“假”.
- 语义（我说的话和我关心的东西之间有什么关系）：我们可以先给命题字母真假的概念，例如 p 赋值为“真”， q 赋值为“假”. 当每一个命题字母都被赋予了真假之后，我们就可以定义整个公式的真假. 例如， $p \rightarrow q$ 为真当且仅当 p 为假或者 q 为真. □

尽管在第一章中，我们也简要介绍了逻辑三要素. 但是，我们在这里才将这样的讨论展开. 这是因为，只有到了模态逻辑的世界，逻辑三要素才变得如此不平凡，我们才能真正体会到他们的意义. 接下来，我们将介绍模态逻辑的语言、模型和语义. 本节先介绍模态逻辑的语言.

首先，我们只考虑最简单的情况，基础语言是命题逻辑，然后在其 中加入一个模态算子.

定义 11.1 (基本模态语言) 给定命题字母表 \mathbf{P} ，我们定义它的**基本模态语言** L ，按照如下方式递归生成：

- 命题字母 $p \in \mathbf{P}$ 属于 L ， \top 属于 L .
- 如果 ϕ 属于 L ，那么 $\neg\phi$ 和 $\Box\phi$ 也属于 L .
- 如果 ϕ_1, ϕ_2 属于 L ，那么 $(\phi_1 \wedge \phi_2)$ 也属于 L . □

和命题逻辑比较，我们多了一个模态算子 \Box ，它读作“Box”，对应的自然语言可以读为“必然”（更多讨论见本部分后文）。

我们在后面将会频繁定义各种不同的模态语言，为了方便，我们引入一种更简洁的记号：

$$\phi ::= p \mid \top \mid \neg\phi \mid (\phi \wedge \phi) \mid \Box\phi.$$

这种记号被称为 *Backus-Naur 范式 (BNF)*，是一种用来描述形式语言的标准记号，在编程语言设计和编译器的实现中也经常使用。

类似命题逻辑，我们有如下缩写：

- $\phi \vee \psi \iff \neg(\neg\phi \wedge \neg\psi).$
- $\phi \rightarrow \psi \iff \neg\phi \vee \psi.$
- $\perp \iff \neg\top.$

引入模态算子 \Box 之后，我们还有它对应的对偶算子 \Diamond ，它是如下的缩写：

$$\Diamond\phi \iff \neg\Box\neg\phi.$$

\Diamond 读作“diamond”，对应的自然语言可以读为“可能”。为什么它的自然语言解释是“可能”呢？因为，“不是必然不”的意思就是“可能”。 \Box 和 \Diamond 的对偶性，我们可以类比 \exists 和 \forall 的对偶性，这是因为 \exists 也可以写成 $\neg\forall\neg$ 。我们将在第 11.2 节中看到，这一类比其实有极其自然的模型论解释。

既然可以加入一个模态算子，我们也可以加入更多的模态算子，一个模态算子也可以修饰多个公式。例如，我们可以引入一个新的模态算

子 \square_a , 表示个体 a 认为的必然性. 于是, 我们可以写出这样的公式:

$$\square_a \phi \wedge \square_b \psi \rightarrow \square_a (\phi \wedge \psi).$$

我们也可以引入模态算子 $\nabla(\phi, \psi)$, 表示 ϕ 成立的时候, ψ 必然成立. 于是, 我们可以写出这样的公式:

$$\nabla(\phi \rightarrow \psi, \square\phi) \rightarrow \square\psi.$$

通过引入不同的模态算子来描述不同的对象, 模态逻辑被赋予了这样的哲学: 多视角下看同一个数学概念. 比如, 我们可以把 \square 在自然语言中用不同的词来解释, 由此得到不同的模态逻辑:

- 基本模态逻辑: 可能/必然是
- 时序逻辑: 将会是
- 道义逻辑: 被允许是
- 知识逻辑: 被知道是
- 可证性逻辑: 可以被证明是
- 动态逻辑: (在经过某些程序步骤之后) 会是

接下来, 我们具体看两个例子, 这是本节最关心的两种解释: 基本模态逻辑和知识逻辑.

例 11.2 (基本模态逻辑) 我们可以把模态算子 \square 读成“必然”. 于是,

- $\square\phi$ 表示“必然有 ϕ ”.

- $\Diamond\phi$ 表示“不是必然有非 ϕ ”，即“可能有 ϕ ”，所以 \Diamond 读作“可能” .
- 反之， $\Box\phi$ 也可以读作“不可能有非 ϕ ”，即“必然有 ϕ ” .
- 因此， \Diamond 和 \Box 确实是对偶的.

在这个读法下，我们可以用形式语言去表达一些自然语言中很拗口的句子，但更加清晰和精确. 比如：

- $\Box p \rightarrow \Diamond p$: 必然的事也是可能的.
- $p \rightarrow \Box p$: 真的事是必然的.
- $\Diamond p \rightarrow \Box\Diamond p$: 可能的事是必然可能的. □

例 11.3 (知识逻辑) 在知识逻辑中，我们可以把模态算子 \Box 读成“知道”，并写成 K (know) . K 表示某个特定的个体对世界的认知. 例如：

- $K\phi$ (即 $\Box\phi$): 我知道 ϕ .
- $K\phi \rightarrow \phi$: 如果我知道 ϕ ，那么 ϕ 是真的.
- $\phi \rightarrow K\phi$: 如果 ϕ 是真的，那么我知道 ϕ .
- $\neg K\phi$ vs. $K(\neg\phi)$: 我不知道上帝存在 vs. 我知道上帝不存在. 这两句话的含义是不同的，因此，模态词使得否定的含义变得复杂.

在更一般的情况下，我们会有很多个个体，于是可以用 K_a 表示“个体 a 知道”. 同样，我们可以用 B_a 表示“个体 a 相信”. 这里是一些例子：

- $K_a K_b \phi \leftrightarrow K_b K_a \phi$: 我知道你知道 ϕ 当且仅当你知道我知道 ϕ .
- $K_1 K_2 p \wedge \neg K_2 K_1 K_2 p$: 1 知道 2 知道 p , 但是 2 并不知道 1 知道 2 知道 p .
- $\neg K_i p \rightarrow K_i(\neg K_i p)$: 如果我不知道 p , 那么我知道我不知道 p .
- $K_i(p \wedge \neg K_i p)$: 我知道如下的陈述: p 是真的, 且我不知道 p . 一种类似的写法是, $K_i p \wedge K_i \neg K_i p$, 即我知道 p , 但是我又知道我不知道 p .
- 共同知识算子¹ C : $C\phi$ 当且仅当 $K_a(\phi \wedge C\phi)$ 对任意 a 成立. 注意, $C\phi$ 并不等价于对任意 a , $K_a\phi$ 成立. □

§11.2 Kripke 语义

接下来我们讨论模态语言对应的模型, 以及模态逻辑语言在这一模型上的语义. 从本节开始, 我们将模态算子限制为一元算子, 即它只能修饰一个公式.

我们考虑的模型被称为 *Kripke* (点) 模型. 它可以看作是一个带有标记的有向边和节点的图:

- 节点表示可能世界, 上面用命题字母标记, 表示这个可能世界上成立的原子命题;

¹这个算子的定义以及性质在第 11.4 节中会详细讨论, 这里仅作为一个例子, 读者不必理解具体的意义.

- 边表示节点之间的关系，用模态算子标记，表示这两个可能世界之间的关系。如果只有一个模态算子，我们省略模态算子标记，只写箭头。
- 我们有一个指定的节点，作为真实世界。

我们来看一个例子。

例 11.4 我们可以用下面的图 11.1 来表示一个 Kripke 模型。

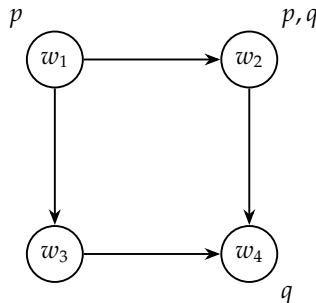


图 11.1: Kripke 模型的示例

模型中，一共有四个可能世界，用 w_1, w_2, w_3, w_4 表示。他们之间的关系用箭头表示，例如 $w_1 \rightarrow w_2$ 表示 w_1 到 w_2 有模态算子 \square 对应的关系。我们可以把箭头更直观地理解为，从 w_1 这个世界的视角看，他能够想象一个 w_2 的可能世界。

我们还要指定一个节点作为真实世界，这里我们指定 w_1 作为真实世界。 \square

我们将节点解读为可能世界。此时， \square 被理解为“必然”， \diamond 被理解为“可能”。接下来，我们把模态的语言和模型联系起来，定义模态逻辑的语义，完成模态逻辑三要素的定义。

设想，在真实世界 w_1 ，我们说 $\Box p$ 是真的，换言之，“必然有 p ” . 当我们在说这句话的时候，我们其实在脑中想象了所有能想到的可能世界，然后发现在所有可能世界中， p 都是真的，于是，我们才说，“必然有 p ” . 这就是模态逻辑的语义.

形式化地说， $\Box\phi$ 在世界 w 上成立当且仅当 ϕ 在 w 的所有后继上为真. 这种语义通常被称为 Kripke 语义或可能世界语义. 在一个世界讨论可能与必然的时候，会取决于它与其他世界的联系.

我们再来看一个例子.

例 11.5 考虑图 11.1 中的 Kripke 模型. 对哪些 i 来说，下式成立？

$$\Box(p \rightarrow \Box q).$$

我们以 w_1 为例进行讨论. 其他情况非常类似，读者可以自行验证. 如果 w_1 是真实世界，上要成立上式，必须要所有后继节点上成立

$$p \rightarrow \Box q.$$

p 的后继节点有 w_2, w_3 ，分别考虑这两个节点：

- 对于 w_2 ， p 为真，所以需要看 $\Box q$ 是否成立. w_2 的后继节点只有 w_4 . 注意到， w_4 上 q 为真，因此 w_2 上 $\Box q$ 成立，所以 w_2 上 $p \rightarrow \Box q$ 也成立.
- 对于 w_3 ， p 为假，所以 $p \rightarrow \Box q$ 自动成立.

因此， w_1 上 $\Box(p \rightarrow \Box q)$ 成立. 同理，我们可以验证其他节点上的情况，得到在 w_2, w_3, w_4 上， $\Box(p \rightarrow \Box q)$ 都成立.

特别注意，在验证 w_4 的时候，因为 w_4 没有后继节点，所以任何命题 ϕ , $\Box\phi$ 都是成立的. \square

游了对概念的自然语言描述和例子，接下来，我们形式上给出模态逻辑的模型和语义定义.

首先，定义一个 *Kripke* 框架，它单纯描述可能世界及其之间的关系，不考虑可能世界上有哪些命题成立.

定义 11.2 (Kripke 框架) 考虑基本命题模态逻辑 L . 一个 **Kripke** 框架是一个元组 $\mathcal{F} = (W, R)$, 其中：

- W 是非空集合（可能世界集）；
- $R \subseteq W \times W$ 是一个 W 上的二元关系（边）.

 \square

接下来，对每一个可能世界，我们赋予它为真的原子命题，如此得到了一个 *Kripke* 模型.

定义 11.3 (Kripke 模型) 一个 **Kripke** 模型 \mathcal{M} 是一个元组 (\mathcal{F}, V) , 其中 \mathcal{F} 是 *Kripke* 框架, $V : W \rightarrow 2^P$ 是赋值函数，表示每个可能世界上为真的那些命题字母（即原子命题）.

 \square

Kripke 模型并没有指定一个真实世界，接下来，我们引入一个指定的点，作为真实世界，得到 *Kripke* 点模型.

定义 11.4 (Kripke 点模型) 一个 **Kripke** 点模型 (\mathcal{M}, w) 是 *Kripke* 模型 \mathcal{M} 加上一个指定的点 $w \in W$.

 \square

现在，我们可以定义模态逻辑的语义，即 *Kripke* 语义.

定义 11.5 (Kripke 语义) 考虑基本命题模态逻辑 L . 符号 $\mathcal{M}, w \models \phi$ 表示 ϕ 在点模型 \mathcal{M}, w 是可满足的. 这一概念可以递归定义如下：

- $\mathcal{M}, w \models \top$ 永远成立.
- $\mathcal{M}, w \models p$ 当且仅当 $p \in V(w)$.
- $\mathcal{M}, w \models (\phi \wedge \psi)$ 当且仅当 $\mathcal{M}, w \models \phi$ 且 $\mathcal{M}, w \models \psi$.
- $\mathcal{M}, w \models \neg\phi$ 当且仅当 $\mathcal{M}, w \not\models \phi$.
- $\mathcal{M}, w \models \Box\phi$ 当且仅当对所有 v , 如果 wRv , 那么 $\mathcal{M}, v \models \phi$.

因此, $\mathcal{M}, w \models \Diamond\phi$ 当且仅当存在 v 满足 wRv 且 $\mathcal{M}, v \models \phi$. □

注意, 上面的定义都是在只有一个模态算子的情况下定义的. 不过, 我们可以很容易地推广到多个一元模态算子的情况. 假设模态算子是 \Box_a , 那么 Kripke 框架中的边需要附上标签 a , 表明这个边是 \Box_a 对应的关系, 即 $w \rightarrow_a v$. 在这种情况下, $\Box_a\phi$ 的语义定义为: 对所有 v , 如果 $wR_a v$, 那么 $\mathcal{M}, v \models \phi$.

多个一元模态算子在知识逻辑中是很常见的. 每个人都有自己对于世界的认知, 因此每个人 a 在模型中会有一个对应的关系 R_a , 以描述他所认为的可能世界架构, R_a 对应的 \Box 模态算子写作 K_a .

现在, 我们进一步讨论语义的性质. 逻辑公式的语义, 无论如何定义, 其实都在讨论“什么是真的”这一基本问题. 在模态逻辑中, 这一概念尤其复杂.

比如, 在例 11.5 中, 我们讨论了 $\phi : \Box(p \rightarrow \Box q)$ 在一个 Kripke 模型中每个可能世界上的真值. 在例子的计算中, 我们发现, 所有可能世界上 ϕ 都是真的. 所以, 我们其实可以说, ϕ 在整个 Kripke 模型上也是真的.

更进一步, 我们也想知道一些关于可能和必然的真理. 为了说明这一点, 先回到命题逻辑中, 考虑任意命题字母 p , 无论 p 是否是真的,

下面这个公式永远都是真的：

$$p \rightarrow p.$$

我们在命题逻辑中称这样的公式为重言式。重言式其实反映了命题逻辑的一些根本性质，这些性质独立于具体的命题而存在。比如排中律：对于任意命题 p , $p \vee \neg p$ 是重言式，再比如三段论： $((p \rightarrow q) \wedge p) \rightarrow q$ 是重言式。他们是我们做逻辑推理的时候自动会假设的性质。

同样的问题也可以对模态逻辑提出：有没有一些关于可能和必然的性质，是独立于具体的可能世界而存在的？比如，我们是否可以说，如果 p 是真的，那么 p 必然是真的？把它写成模态逻辑的形式就是：

$$p \rightarrow \Box p.$$

如果这是一个关于必然的真理，那么，无论 p 是否为真，无论我们处于什么样的可能世界，这个公式都应该是真的。

根据上面的启发，在模态逻辑的体系中，我们可以定义各种不同粒度的“真”的概念：

定义 11.6 (模态逻辑的真值) 给定模态逻辑公式 ϕ ，我们可以定义它的真值：

- ϕ 在点模型 \mathcal{M}, w 可满足指的是 $\mathcal{M}, w \models \phi$.
- ϕ 在模型 \mathcal{M} 有效，记为 $\mathcal{M} \models \phi$ ，指的是 $\mathcal{M}, w \models \phi$ 对所有 w 成立。
- ϕ 在点框架² \mathcal{F}, w 有效，记为 $\mathcal{F}, w \models \phi$ ，指的是 $\mathcal{M}, w \models \phi$ 对所有

²我们在前文中没有定义点框架，但是点框架的定义相当直接：它是一个 Kripke 框架

基于 \mathcal{F} 的模型 \mathcal{M} 成立.

- ϕ 在框架 \mathcal{F} 有效, 记为 $\mathcal{F} \models \phi$, 指的是 $\mathcal{M} \models \phi$ 对所有基于 \mathcal{F} 的模型 \mathcal{M} 成立.
- ϕ 对框架类 (即框架的一个集合) K 有效, 记为 $\models_K \phi$, 指的是 $\mathcal{F} \models \phi$ 对所有 $\mathcal{F} \in K$ 成立. \square

我们可以看到, 模态逻辑的真值有两个维度: 局部-全局, 模型-框架. 越靠左边的真值越具体, 越靠右边的真值越一般. 我们可以用下面的表格来总结这些概念:

	模型	框架
局部	$\mathcal{M}, w \models \phi$	$\mathcal{F}, w \models \phi$
全局	$\mathcal{M} \models \phi$	$\mathcal{F} \models \phi$

我们主要讨论高亮的两个部分.

左上角反映了某个具体可能世界结构中的某个具体世界上的真值. 例如, 假设 p 表示“地球是球形的”, 那么在我们所生活的这个世界上, p 是真的, 但是, 我们的确可以想象出来另一个可能世界, 地球是平的, 那么在这个可能世界上, p 是假的. 因此, 这是一个只适用于我们所生活的这个可能世界的真值.

右下角反映了某个具体框架中所有可能世界上的真值. 例如, 考虑命题 $\phi : p \rightarrow \Diamond p$ 表示“如果 p 是真的, 那么可能 p 是真的”. 这是一个关于可能的陈述, 看起来是非常合乎道理的. 我们无法想象出一个可能世界, p 是真的, 但不可能 p 是真的. 所以, 如果 \mathcal{F} 表示“可能”所

加上一个指定的点.

对应的框架，那么 ϕ 在 \mathcal{F} 上是有效的。然而，如果 \diamond 被理解为“相信”， ϕ 就完全没有任何道理。因而，在“相信”所对应不同的框架 \mathcal{F}' 上， ϕ 不再有效。

因此，从左上角到右下角的过程，是从具体、特定的真值到一般、普遍的真值的过程。

更抽象地来看，例 11.5 其实给出了左上角的一个例子，下面，我们可以演示右下角的一个例子。

例 11.6 考虑图 11.2 中的框架，它是图 11.1 中的 Kripke 模型去掉了命题字母的结果。

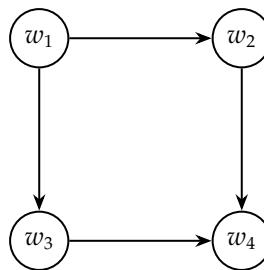


图 11.2: 框架语义的示例

我们问，对于这个框架，下面的公式是否有效？

$$\Box(p \rightarrow \Box p).$$

为此，我们需要考虑四个可能世界上的情况。同样，我们只对 w_1 进行讨论，其他情况类似。对于 w_1 ，有两个后继节点 w_2, w_3 ，我们需要验证 $p \rightarrow \Box p$ 在这两个节点上是否成立。

先考虑 w_2 ，假设 p 在 w_2 上为真，我们需要验证 $\Box p$ 是否成立。 w_2

的后继节点只有 w_4 , 但我们可以让 p 在 w_4 上为假, 这样 $\Box p$ 就不成立。因此, w_2 上 $p \rightarrow \Box p$ 也不成立, 从而 w_1 上 $\Box(p \rightarrow \Box p)$ 不成立。

于是, 我们找到了一种赋值的方案, 使得 $\Box(p \rightarrow \Box p)$ 在 w_1 上不成立。因此, 这个公式在整个框架上也不有效。□

最后, 我们说明, 上面的讨论在知识逻辑中都是自然成立的。此时, 模态公式 $K_i\phi$ 被读作 “ i 知道 ϕ ” 。从语义来说, K_i 是 \Box 算子, 即我知道 ϕ 意味着在我认为的所有可能世界中 ϕ 都是真的, 即:

$$\mathcal{M}, w \models K_i\phi \iff \text{对任意 } v, \text{ 如果 } w \rightarrow_i v, \text{ 那么 } \mathcal{M}, v \models \phi.$$

同样, 模态公式的真值有两个层面:

- 在点模型上可满足: $\mathcal{M}, w \models \phi$;
- 在框架上有效: $\mathcal{F} \models \phi$.

§11.3 模态可定义性

逻辑的意义在于把对事物的抽象认知用形式化的语言表述出来。我们已经看到, 我们对事物的认知可以被两种方式描述出来: 一是通过 Kripke 模型 (框架) 的特殊结构, 二是通过具体的模态公式。那么, 这两种方式之间有什么联系呢? 本节我们研究这个问题。

先看几个例子。

例 11.7 考虑知识逻辑, 假设只有一个知识算子 K 。设有命题 p , 如果 p 是真的, 那么我不知道非 p , 即 $\phi : p \rightarrow \neg K \neg p$ 。这是用逻辑表述了关于知道的一种性质。

另一方面，我们也可以将对于 K 的理解反映到 Kripke 模型中。一个合理的性质是：真实世界是我能够感知到的可能世界。换言之，对任何 x ，都有 xRx ，这是一个自反关系。在这种情况下，容易验证，对于自反点模型以及框架 \mathcal{F} ，我们有 $\mathcal{M}, v \models \phi$ 以及 $\mathcal{F} \models \phi$ 。 \square

上面的例子说明，对于特定的公式，它会在某类模型上成立。反过来，我们也可以讨论，成立特定公式的模型是什么样的，我们看下面的例子。

例 11.8 考虑基础模态逻辑和它的公式 $\diamond\top$ 。对于一个 Kripke 点模型 \mathcal{M}, w ，假设 $\mathcal{M}, w \models \diamond\top$ ，我们来研究 \mathcal{M} 的特殊性质。

根据定义，这意味着存在一个点 v ，使得 wRv 并且 $\mathcal{M}, v \models \top$ ，由于后一半永远成立，所以，这等价于 w 有一个后继。同样的，对于任意框架 \mathcal{F} ，如果 $\mathcal{F} \models \diamond\top$ ，则基于 \mathcal{F} 的每个点模型 \mathcal{M}, v 都满足 $\diamond\top$ ，即 \mathcal{F} 的每个点都有后继。 \square

上面两个例子说明了模型性质和模态公式之间单向的联系，下面的例子说明，这种联系也可以是双向的。

例 11.9 考虑公式 $\phi : p \rightarrow \diamond p$ 。我们研究和这个公式完全对应的框架 \mathcal{F} 的性质。

- 任给一个赋值 V 和点 w 都有 $\mathcal{M}, w \models p \rightarrow \diamond p$ 。考虑一个赋值 V ，使得只有 w 上有 p 。因为 w 上有 p ，为了使 ϕ 成立， w 必须要有一个后继上也有 p 。在我们的这个赋值下，这意味着 w 必须以自己为后继。因此，对任意 w ， wRw 成立，即 \mathcal{F} 是一个自反框架。
- 反过来，如果 \mathcal{F} 是一个自反框架，那么对于任意赋值 V 和点 w ，容易验证 $\mathcal{M}, w \models p \rightarrow \diamond p$ 。因此， $\mathcal{F} \models p \rightarrow \diamond p$ 。

结合这两点，我们可以说， $p \rightarrow \Diamond p$ 定义了自反框架.

此外，结合这个例子与例 11.8，我们会发现，一个框架的特定性质可以由多个不同的公式定义. \square

以上例子给我们启示，我们可以用模态公式去定义模型或框架的性质. 这种定义被称为模态可定义性，我们给出如下定义.

从点模型的角度，我们可以讨论模态公式定义了什么样的点模型.

定义 11.7 (点模型可定义性) 设 \mathcal{K} 是一些点模型的集合， Σ 是一些模态公式的集合. 我们说 \mathcal{K} 可由公式集 Σ 定义，指的是对于任意点模型 \mathcal{M}, w , $\mathcal{M}, w \in \mathcal{K}$ 当且仅当对任意 $\phi \in \Sigma$,

$$\mathcal{M}, w \vDash \phi.$$

如果 $\Sigma = \{\phi\}$ ，我们就说 \mathcal{K} 可以由公式 ϕ 定义. \square

我们也可以从框架的角度讨论模态可定义性，定义类似.

定义 11.8 (框架可定义性) 设 \mathcal{K} 是一些框架的集合， Σ 是一些模态公式的集合. 我们说 \mathcal{K} 可由公式集 Σ 定义，指的是对于任意框架 \mathcal{F} , $\mathcal{F} \in \mathcal{K}$ 当且仅当对任意 $\phi \in \Sigma$,

$$\mathcal{F} \vDash \phi.$$

如果 $\Sigma = \{\phi\}$ ，我们就说 \mathcal{K} 可以由公式 ϕ 定义. \square

§11.4 知识逻辑的基本模型与性质

接下来，我们将注意力放在知识逻辑上，将前面若干小节讨论的概念应用到知识逻辑中.

§11.4.1 知识逻辑的 Kripke 模型与公理

首先，从模型的角度看，一个自然的假设是，对每个个体 i 来说，他认为的可能世界之间应该是不可区分的。这意味着，对于每个个体 i ，他认为的可能世界之间的关系 R_i 应该是等价关系，我们记作 \sim_i ：

- 自反： $\forall x, x \sim_i x$. 我能够想象到真实的世界.
- 传递： $\forall x, y, z (x \sim_i y \wedge y \sim_i z) \rightarrow x \sim_i z$. 如果我在 x 世界能够想象到 y 世界，在 y 世界能够想象到 z 世界，那么我在 x 世界也能够想象到 z 世界.
- 对称： $\forall x, y (x \sim_i y \leftrightarrow y \sim_i x)$. 我能够在 x 世界想象到 y 世界，那么我也能够在 y 世界想象到 x 世界.

我们将 R_i 是等价关系的框架类记为 H .

从模态可定义性的角度来说， \sim_i 的特殊性质会对应 K_i 特殊的公式。这些公式就可以被看成关于“知道”的公理或推导规则。承认某一条公理或推导规则就必须承认可能世界具有某一种性质，反之亦然。

接下来，我们给出知识逻辑的基本性质，即公理。

命题 11.1 (分配公理)

$$\models (K_i(\phi \rightarrow \psi) \wedge K_i\phi) \rightarrow K_i\psi.$$

证明. 假设 $\mathcal{M}, w \models K_i(\phi \rightarrow \psi)$ 且 $\mathcal{M}, w \models K_i\phi$. 于是，对所有 R_i 后继 v 都有 $\mathcal{M}, v \models \phi \rightarrow \psi$ 和 $\mathcal{M}, v \models \phi$ ，因而 $\mathcal{M}, v \models \psi$. 根据定义， $\mathcal{M}, w \models K_i\psi$ ，因而对所有 \mathcal{F} ，分配公理有效。 \square

分配公理意味着拥有知识的个体可以对自己的知识做任意的演绎

推理，因而假设个体是逻辑全知的。注意，分配公理并没有用到框架类 H 的性质，所以，它其实是模态逻辑的一个普适公理。

命题 11.2 (泛化规则) 对所有 \mathcal{F} ，如果 $\mathcal{F} \models \phi$ ，那么 $\mathcal{F} \models K_i\phi$ 。

证明. 假设 $\mathcal{F} \models \phi$ ，这意味着对所有基于 \mathcal{F} 的点模型都有 $\mathcal{M}, w \models \phi$ 。因此，对任意 w 的 R_i 后继 v ，也有 $\mathcal{M}, v \models \phi$ 。所以也有 $\mathcal{M}, w \models K_i\phi$ 成立，因而 $\mathcal{F} \models K_i\phi$. \square

泛化规则可能不太容易理解，我们举一个例子。生活在中国，我们都知道农历五月初五是端午节，“五月五，迎端午”，按照传统习俗要吃粽子。这意味着，在我们所有对未来生活的设想中，“农历五月五要吃粽子”这一习俗是牢固存在的。因此，对于中国人来说，我们身处的世界以及幻想的可能世界组成了框架 \mathcal{F} ，而“农历五月五吃粽子”是这个框架中的一个普适规律。

然而，如果我们生活在冰岛，一个中国人非常少的国家，那么，“农历五月五要吃粽子”这一习俗可能就不再广为流传。在这种情况下，我们所处的世界及想象的可能世界组成了另一个框架 \mathcal{F}' ，在这个框架中，“农历五月五要吃粽子”不再是普适规律。

所以，泛化规则其实在说，如果 ϕ 是 \mathcal{F} 普遍适用的规律，不依赖具体的命题，那么我就知道 ϕ 。当然，这个规律 ϕ 可能只适用于 \mathcal{F} 。但无论如何，我清晰地知道这个世界的运作规律。特别地，我知道关于知识的这些规律。

同样，泛化规则也没有使用框架类 H 的性质，所以它也是模态逻辑的普适规则。

命题 11.3 (知识公理，真理公理)

$$\vDash_H K_i\phi \rightarrow \phi.$$

注意，从知识公理开始，我们就开始使用了框架类 H 的性质。从这里开始的性质，和第十章中讨论的知识的性质是一致的，所以这里不再赘述。他们的证明类似前述性质，这里不再给出，见习题 2。

命题 11.4 (内省公理) 正内省公理：

$$\models_H K_i\phi \rightarrow K_iK_i\phi.$$

负内省公理：

$$\models_H \neg K_i\phi \rightarrow K_i\neg K_i\phi.$$

以上五条性质（四条公理 + 一条推导规则）加上 MP 形成的推理系统称为 $S5$ 公理系统。需要注意的是，这些公理其实都是公理模式，对每一个具体的命题 ϕ ，都有一条公理，因此，它其实包含了无穷条公理。

从哲学的角度讨论，还有一些别的公理。

命题 11.5 (一致性公理)

$$\models_H \neg K_i\perp.$$

因此，个体不能够知道假的陈述，以此区别于信念。

我们基于框架类 H 给出了关于知识的公理。反过来，公理对应什么样的框架结构呢？我们总结于表 11.1。

表中有两个性质是我们没有讨论过的，这里给一个简短的解释。

- Euclid 性： $\forall x, y, z(xR_iy \wedge xR_iz \rightarrow yR_iz)$ ，也就是说，关系一定形成三角形。
- 序列性： $\forall x \exists y xR_iy$ ，所有点都有后继。

公理	R_i 的性质
$K_i\varphi \rightarrow \varphi$	自反性
$K_i\varphi \rightarrow K_iK_i\varphi$	传递性
$\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$	Euclid 性
$\neg K_i\perp$	序列性
$\varphi \rightarrow K_i\neg K_i\neg\varphi$	对称性

表 11.1: 知识公理对应的框架结构

以上的模态可定义性和第 11.3 节中的例子类似，这里不再给出，请见习题 ??.

可以看出来，以上关系其实并不是孤立的，我们有以下引理：

- 引理 11.1**
- 如果 R_i 是自反和 Euclid 的，那么 R_i 是对称和传递的.
 - 如果 R_i 是对称和传递的，那么 R_i 是 Euclid 的.
 - 以下命题等价：
 - R_i 是自反、对称和传递的.
 - R_i 是对称、传递和序列的.
 - R_i 是自反和 Euclid 的.

这一性质的验证根据定义即可，见习题 4.

下面我们向知识逻辑语言中加入共同知识算子和它的语义. 首先加入“所有人都知道”这个算子：

$$E\phi \leftrightarrow \bigwedge_i K_i\phi.$$

记 $E^k\phi$ 为 $\underbrace{E\dots E}_{k}\phi$. 于是, 我们可以给共同知识算子一个语义定义:

定义 11.9 (共同知识算子) 共同知识算子 C 的语义定义为:

$$\mathcal{M}, w \models C\phi \iff \mathcal{M}, w \models E^k\phi, \quad k = 1, 2, \dots$$

□

我们可以从图结构来理解共同知识算子. 把所有边上的标记忽略掉, Kripke 模型变成一个有向图, 节点上的标记是这个节点上成立的原子命题. 于是, 我们有如下解读:

- $\mathcal{M}, w \models E^k\phi$ 的含义是, 从 w 出发走恰好 k 步可到达的可能世界 v 上都有 $\mathcal{M}, v \models \phi$. 这一解读的示意见图 11.3,

在这个图中, $\mathcal{M}, w_1 \models E^3\phi$, 因为从 w_1 出发, 经过三步可以到达的可能世界上都有 $\mathcal{M}, v \models \phi$, 我们具体地标出了两条图路径: $w_1 \rightarrow_1 w_2 \rightarrow_3 w_3 \rightarrow_1 w_5$ 对应 $K_1K_3K_1\phi$, $w_1 \rightarrow_2 w_4 \rightarrow_2 w_3 \rightarrow_2 w_6$ 对应 $K_2K_2K_2\phi$.

- 因此, $\mathcal{M}, w \models C\phi$ 的含义便是, 从 w 出发经过有限步可到达的可能世界 v 上都有 $\mathcal{M}, v \models \phi$.

接下来, 我们给出共同知识算子的性质. 实际上, 共同知识算子可以充分必要地被下面两个公理所定义 (验证见习题 6):

命题 11.6 (不动点公理)

$$\models C\phi \leftrightarrow E(\phi \wedge C\phi).$$

命题 11.7 (归纳规则) 如果

$$\mathcal{F} \models \phi \rightarrow E(\phi \wedge \psi),$$

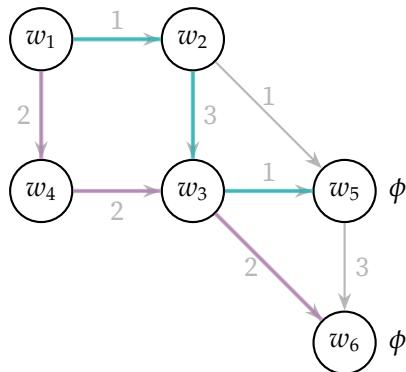


图 11.3: $\mathcal{M}, w_1 \models E^3\phi$ 的示意图, 我们标出了其中的两条路径.

那么

$$\mathcal{F} \models \phi \rightarrow C\psi.$$

现在我们来讲解上面两个公理的含义.

不动点公理是比较直接的. 什么是共同知识? 对任意 k , $E^k\phi$ 成立.

我们把不动点公理展开, 可以得到:

$$\begin{aligned}
 C\phi &\leftrightarrow E(\phi \wedge C\phi) \\
 &\leftrightarrow E\phi \wedge E(C\phi) \\
 &\leftrightarrow E\phi \wedge E(E\phi \wedge E(C\phi)) \\
 &\leftrightarrow E\phi \wedge E^2\phi \wedge E^3(C\phi) \\
 &\dots
 \end{aligned}$$

因为逻辑公式不允许写无穷长的公式, 所以, 不动点公理就是在用一个

算子的递归方程来定义共同知识：

$$\forall \phi, C(\phi) \leftrightarrow E(\phi \wedge C(\phi)).$$

然而，一个递归方程可能会有无穷多个解，并不是所有的解都是我们要的那个。所以，归纳规则给出了共同知识算子的另一个性质：这是这个递归方程的最小解。下面我们来说明这一点。

我们先从一个看似无关的例子讲起。用数学归纳法证明一个关于自然数的命题 $P(n)$ 时，我们需要两个步骤：

- 证明基础情形 $P(0)$ 成立；
- 证明归纳步骤：如果 $P(n)$ 成立，那么 $P(n+1)$ 也成立。

因此，归纳的思想是，我们先找到一个基础情形，然后一点一点把成立的集合扩大，直到包含所有的情况。

注意，在真正的证明中，我们不能真的把每一个 $P(n)$ 的正确性都从更小的情况推导出来，因为这是无穷多个情况。实际上，我们写归纳法证明的过程中，隐含地写了一个关于 P 的递归方程，左边是这个一点一点扩大的过程，右边是我们要证明的命题：

$$(P(0) \wedge \forall n, (P(n) \rightarrow P(n+1))) \leftrightarrow \forall n, P(n).$$

所以，数学归纳法就是一个公理，它说明这样一点一点扩大的过程（左边）的确可以用来证明所有的情况（右边）。注意，这个公理产生的 P 是递归方程的最小解，因为它是从最小情况逐步扩大得到的结果。

回到共同知识算子 C ，其实 C 和数学归纳法里的证明 P 的过程是非常相似的。它的定义也是从基础情形 ϕ 开始，然后一点一点扩大，每

次扩大都是用 E 算子，直到无穷层。因此，它也具有归纳的性质。

观察归纳规则，它的前提条件是：

$$\mathcal{F} \vDash \phi \rightarrow E(\phi \wedge \psi).$$

这一个步骤描述了，如果我们已经构造了 ϕ ，我们如何用 E 算子构造下一层的命题。这就相当于数学归纳法中从 $P(n)$ 到 $P(n+1)$ 的过程。

而归纳规则的结论是：

$$\mathcal{F} \vDash \phi \rightarrow C\psi.$$

这个结论说明，如果我们可以逐层加 E 算子，那么这个叠加可以无穷进行下去，最终得到的就是 C 算子。这就相当于数学归纳法中证明 $P(n)$ 对整个自然数集成立。

所以，上面的类比说明了归纳规则的含义：它确保了 C 算子是递归方程的最小解，也就是那个逐层叠加 E 所定义的算子。

最后，将 S5 公理系统中加入关于 E 和 C 的公理，我们就扩展了知识逻辑。因为 E 和 C 是用 K_i 定义的，因此它们本身并不会带来 Kripke 模型新的结构性质。

§11.4.2 Kripke 模型与 Aumann 结构

Kripke 模型与 Aumann 结构之间存在着非常本质的对应关系。

- 在 Kripke 模型中，我们将算子 K_i 对应的关系限制为了等价关系，因为我们希望 i 想象的可能世界之间是不可区分的。
- 在 Aumann 结构中，我们将 i 的信息集定义了全集的一个划分，

把每个划分作为他的原子信息.

上面的两个概念实际上是可以相互转换的. 数学上说, 等价关系充分必要地给了集合一个划分; 从概念上说, 所谓不可区分, 就是原子性. 因此, 他们之间其实存在一一对应的关系.

进一步, 如果回忆第一章的思想, 我们可以看到逻辑与集合之间的联系如表 11.2 所示.

Kripke 模型	Aumann 结构
可能世界	样本点
公式	事件
原子命题	基本事件
模态算子	集合-集合映射
i 的等价关系	i 的划分
逻辑连接词	集合操作

表 11.2: Kripke 模型与 Aumann 结构的对应关系

实际上, 在 S5 公理系统下, 这种对应关系可以被严格地叙述和证明, 见习题 7.

Kripke 模型和 Aumann 结构的一个重要区别在于它们各自的研究偏好: Kripke 语义偏重于逻辑, 而 Aumann 结构则偏重于 (Bayes) 概率论. 因此, 用数学来研究知识论可以呈现出两种风格: 一种是计算机科学、逻辑学和哲学的风格, 另一种则是经济学和信息论的风格.

然而, 这种对应关系完全依赖于我们对知识的基本假设:

- 对于 Kripke 模型, 我们假设了知识满足 S5 公理系统;

- 对于 Aumann 结构，我们假设了知识满足(K0)–(K4).

如果这些假设被打破，那么这样的对应关系就不再成立. 例如，如果我们移除负自省公理，Kripke 模型就不再具备等价关系的性质，因此不再能够对应于 Aumann 结构中的信息集.

从对知识刻画精细程度来说，两种研究方法也有区别. 通常来说，Aumann 结构要想破坏某一条性质是非常难的，因为它是基于概率论和事件的概念. 然而，Kripke 模型具有（模态）逻辑的风格，因此，研究加入或者去除某个公理带来的影响是它最“正统”的研究方法.

§11.4.3 “泥泞的孩童”再回顾：形式化解法

作为一个具体的例子，我们现在把第十章中的“泥泞的孩童”问题用知识逻辑的语言来重新讨论. 我们终于可以用形式化的方式来严格讨论这一问题了.

我们要给出“泥泞的孩童”的逻辑三要素. 首先，语言就是知识逻辑语言. 然后，我们给出 Kripke 模型以及对应的语义.

可能世界可以表示为 $\{0,1\}^n$ 的元素 $x = (x_1, \dots, x_n)$ ，其中 $x_i = 1$ 表示孩子 i 的脸上有泥巴， $x_i = 0$ 表示 i 的脸上没有泥巴. 我们假设每个孩童 i 的可达关系 R_i 都是一个等价关系. 在这种假设下，每个孩子唯一不是共同知识的事情就是他们脸上泥巴的状态，而其他所有事情都被隐含在了 Kripke 模型之中.

接下来，我们给出命题字母：原子命题 p_i 表示孩子 i 脸上有泥巴，命题 p 表示至少有一个孩子脸上有泥巴.

假设现在父亲还没有宣布 p . 对于孩子 i 来说，他的认知中只有两个可能世界：一个是他的脸上有泥巴，另一个是他的脸上没有泥巴. 其

对他来说都是确定的. 因此, 有 xR_iy 当且仅当 $x_j = y_j$ 对于任意 $j \neq i$ 成立. 在这种情况下, 框架 \mathcal{F} 对应于一个 n 维超立方体.

例如, 当 $n = 3$ 时, 框架 \mathcal{F} 对应的立方体结构如图 11.4 所示.

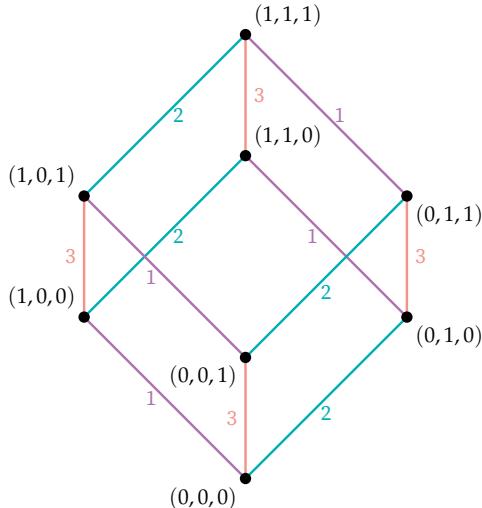


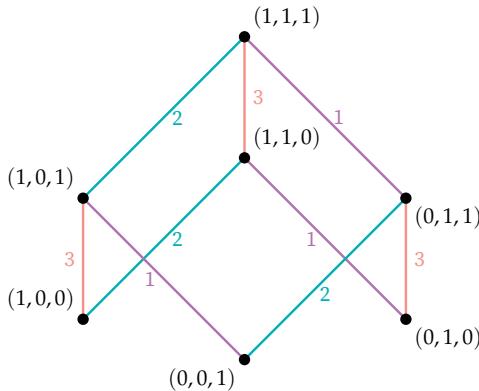
图 11.4: “泥泞的孩童”问题的 Kripke 模型

从框架 \mathcal{F} 到模型 \mathcal{M} , 我们还需要确定赋值 V .

- 对任意 p_i 和 w , 有 $w \in V(p_i)$ 当且仅当 $w_i = 1$;
- $p \in V(w)$ 当且仅当 w 的所有分量 w_j 不全为零.

从模型到点模型, 我们还需要确定我们所处的可能世界, 从而可以讨论模态公式的可满足性. 例如: $\mathcal{M}, (1, 0, 1) \models Ep$, 但是 $\mathcal{M}, (1, 0, 1) \models \neg E^2 p$.

假设父亲宣布了 p , 那么框架 \mathcal{F} 将会发生变化, 如图 11.5 所示.

图 11.5: 父亲宣布 p 之后的 Kripke 模型

在 i 的认知中，只有两个可能世界. 因此， i 回答“知道”意味着她能够确定只有一个世界；她回答“不知道”意味着还有两个可能世界.

假如现在是第一轮问答，如果所有人都回答了“不知道”，考虑状态 $s = (1, 0, 0, \dots)$. 如果真实世界是 s ，那么对于 1 来说，可能世界只有一个了，但她却说“不知道”，这说明真实世界不是 s . 同理，所有只有一个 1 的可能世界都会被排除. 由此可以归纳得出，第 k 轮的时候，所有只有 k 个 1 的可能世界都会被排除.

然而，如果父亲没有宣布 p ，那么 \mathcal{M} 是一个超立方体. 在任何一轮中，每个孩子都会认为有两个可能世界，因此不会有任何可能世界被排除.

因此，从结构上来说，父亲宣布 p 改变了每个孩子对应的 R_i 等价的可能世界，使得一些孩子可以确定自己所处的世界. 因而，父亲宣布 p 之后，孩子们的回答会有本质不同.

实际上，这一套方法可以将类似的智力谜题用算法化的方式得到解答，这是知识逻辑的一个有趣的应用. 这是早年的人工智能研究最为

热衷的几个方向之一（见第三章的章首引言）.

§11.5 对不一致达成一致

作为一个更正式的（也更复杂！）例子，本节将用模态逻辑的方式来探讨达成一致与共同知识的关系，这个问题最早由 Robert Aumann 提出。现代版本的问题是：两个拥有相同决策规则的 AI 是否会进行交易？如果交易发生，这意味着买家和卖家采取了不同的决策（一个购买，一个出售）；否则，他们将不会达成交易。

我们将证明，对于有相同决策方式的两个个体，他们不可能对采取不同行动这件事具有共同知识。因此，如果两个 AI 按照相同的规则行事，交易便不会发生。一句话概括，

玩家不能“对不一致达成一致”（agree to disagree）.

§11.5.1 模型

下面，我们开始介绍这个模型。

首先，我们介绍一下二人博弈的模型。这个模型是一个无随机性的 Markov 博弈，详细讨论见第 9.4 节。设有两个玩家 1 和 2，在任意时刻，每个玩家处于某个局部状态 s_i 之中，局部状态空间分别记为 S_1 和 S_2 。整个系统的全局状态为 $(s_1, s_2) \in S_1 \times S_2 = \mathcal{G}$ 。

接下来，我们描述博弈的整个过程。用非负整数 m 来表示当前时刻，初始时刻为 0。系统的一次运行是一个函数 $r : m \mapsto (s_1, s_2)$ ，描述了系统每一时刻的全局状态。 \mathcal{R} 是全局状态空间 \mathcal{G} 上的所有可能运行的集合。对于每个运行 $r \in \mathcal{R}$ ， (r, m) 被称为系统 \mathcal{R} 的一个点。

玩家处于某个状态时，可以采取某种行动。为了反映“玩家按照相同的规则行事”，我们规定两个玩家的行动集相同，记为 A ，且这一行动集不依赖于全局或局部状态。给定所有人的行动和一个全局状态，系统的转移函数 $\tau : A^2 \times \mathcal{G} \rightarrow \mathcal{G}$ 描述了如何从一个状态转移到另一个状态。

为了描述“按照规则行事”，我们引入策略的概念。玩家 i 的策略 P_i 是从局部状态 S_i 到行动集 A 的映射，即根据状态采取行动。两个玩家的策略组合记为 $P = (P_1, P_2)$ 。这是 Markov 博弈的一种限制，我们要求玩家的策略只依赖于自己的状态，并且不允许随机行动。

策略的执行需要初始状态。初始状态可能的集合记为 \mathcal{G}_0 。给定初始状态集 \mathcal{G}_0 和转移函数 τ ，我们就可以在系统上执行任何一种策略。我们称元组 $\gamma = (\mathcal{G}_0, \tau)$ 为系统的上下文。

给定上下文 $\gamma = (\mathcal{G}_0, \tau)$ 和一个策略组合 P ，整个系统就可以运行起来了，产生一个运行。不难看出，给定一个运行 r ，它与策略 P 相容的条件为：

$$r(0) \in \mathcal{G}_0, \quad \forall m, r(m+1) = \tau(P(s_1), P(s_2))(r(m)),$$

其中 $r(m) = (s_1, s_2)$ 。换言之， r 是从某个可能的初始状态开始执行策略 P 产生的运行。系统 \mathcal{R} 表示上下文 γ 和策略组合 P ，其包含的所有运行 r 都与 P 相容。我们用 $\mathcal{R}^{rep}(P, \gamma)$ 表示这样的系统。

以上我们就完成了博弈部分的建模。接下来，我们引入逻辑部分的建模，并把它与博弈的部分对应起来。

首先，我们引入 Kripke 模型。Kripke 模型的点是系统的点。然后是原子命题集 \mathbf{P} 。它的元素形如 $perf_i(a)$ ，表示玩家 i 采取行动 a 。

我们还需要定义赋值函数 V ，它将 Markov 博弈中的概念和原子命

题联系起来。具体来说，定义如下：原子命题 $perf_i(a) \in V(s)$ 当且仅当在状态 s 玩家 i 采取了行动 a 。

然而，赋值函数的定义域应该是一个 Kripke 模型中的点，这个点不同于 Markov 博弈中的状态。因此，我们需要将赋值函数扩展到系统的点集： $p \in V((r, m))$ 当且仅当 $p \in V(r(m))$ 。

随后，我们引入知识算子 K_i ，假设 K_i 对应的等价关系是 \sim_i 。玩家 i 只能区分自己的局部状态 s_i ，即

$$(r, m) \sim_i (r', m') \iff r(m)_i = r'(m')_i.$$

根据 Aumann 结构，局部状态 s_i 对应的信息集为

$$IS_i(s_i, \mathcal{R}) = \{(r, m) : r \in \mathcal{R}, r(m) = s_i\}.$$

如此一来，我们得到了 Kripke 点模型 $\mathcal{M}, (r, m)$ 以及知识算子 K_i 。

此外，除了知识算子，Markov 博弈还有时间的概念，因此我们引入时间模态算子 X ，表示“下一时刻”。其语义定义为：

$$\mathcal{M}, (r, m) \models X\phi \iff \mathcal{M}, (r, m + 1) \models \phi.$$

利用算子 X ，我们可以用公式 $act_i(a)$ 表示“ i 将要采取行动 a ”：

$$act_i(a) \leftrightarrow \neg perf_i(a) \wedge Xperf_i(a).$$

最后一步，为了在点模型中讨论 Markov 博弈的策略，我们还需要定义 Kripke 模型的决策函数。玩家 i 的策略是从局部状态 S_i 到行动集 A 的映射，因此，我们把这个定义自然地扩展到 Kripke 模型的点上，定

义决策函数 D 为从 Kripke 模型点集 S 的某个子集映射到某个行动 a 的函数. 为了表示两个玩家采取相同的决策策略, 决策函数没有下标.

我们还没有说“某个子集”的选法. 我们要求策略 P_i 和决策函数 D 相容, 也就是说, 如果玩家 i 在这个状态选择了行动 a , 那么决策函数也应该对应产生 a . 我们可以用信息集来描述这个条件: 决策函数在某个信息集上的行动恰好是策略在该状态的行动, 即

$$P_i(s_i) = D(IS_i(s_i, \mathcal{R})), \quad \forall s_i \in S_i.$$

策略和决策函数虽然有密切联系, 但易混淆. 策略描述的是在什么状态下采取什么行动, 而决策函数则是基于知识做出行动决策. 在我们的背景下,

玩家知道的信息 = 玩家处于的局部状态,

因此二者是从不同角度在描述同一概念.

我们还需要知道, 给定策略组合 P 以及博弈的上下文 γ , 我们能否找到一个相容的决策函数? 下面的条件给了我们一个答案:

我们对决策函数 D 要求并-一致. 即对于互不相交的子集 T_1, \dots, T_k , 如果 $D(T_i) = a$, 则 $D(\bigcup_i T_i) = a$. 如果我们把 T_i 理解成“知道的信息”, 那么这一条件并不难理解. 请看下面的例子:

- 假设我的决策函数是这样描述的: 如果今天下雨, 并且今天星期四, 那么我会去 KFC 疯狂星期四; 如果今天不下雨, 并且今天星期四, 那么我会去 KFC 疯狂星期四.
- 那么, 我的决策还应该有: 虽然我不知道今天下不下雨, 但是如果今天是星期四, 那么我会去 KFC 疯狂星期四.

命题 11.8 给定策略组合 P 和上下文 γ , 存在一个并-一致的决策函数 D , 使得 P 和 D 相容.

这一性质的证明并不难, 见习题 9.

现在, 我们完成了逻辑部分的建模, 我们来总结一下. 在这个模型中, 两个玩家处于同一系统中. 虽然他们的局部状态和信息集可能不同, 但他们的行动集和决策函数相同. 决策函数是并-一致的, 并由某个策略组合实现. 通过上下文中的初始状态和转移函数, 系统可以产生一系列可能的运行.

§11.5.2 定理及其证明

有了上面的模型, 接下来, 我们给出达成一致定理的表述和证明.

定理 11.1 (Aumann 达成一致定理) 给定策略组合 P 和上下文 γ , 由此产生 Kripke 框架 \mathcal{F} . 设 $a, b \in A$ 是两个不同的行动, 如果在上下文 γ 中 P 实现了某个并-一致决策函数, 那么

$$\mathcal{F} \models \neg C(act_1(a) \wedge act_2(b)).$$

如果两个玩家选择了同样的并-一致决策函数, 那么他们不可能对“我们采取不同行动”这件事形成共同知识. 因此, 他们不可能对不一致达成一致.

证明. 我们用反证法证明. 假设某个基于 \mathcal{F} 的点模型 $\mathcal{M}, (r, m)$ 使得

$$\mathcal{M}, (r, m) \models C(act_1(a) \wedge act_2(b)).$$

接下来我们证明 $a = b$.

首先介绍一下思路，共同知识对应了从 (r, m) 出发可到达的状态集 S' 的性质。从玩家 1 的视角来看，她在 S' 所关联的信息集上都要采取行动 a ，根据并-一致性，应该有 $D(S') = a$ 。从玩家 2 来看同理，因此也应该有 $D(S') = b$ 。因此， $a = b$ 。下面，我们把这一思路细化，得到完整的证明。

假设 S' 是从 (r, m) 出发，通过关系 \sim_1 或 \sim_2 可到达的点集。取一个点 $(r', m') \in S'$ ，设 $r'(m')_1 = s'_1$ 。假设 $(r'', m'') \sim_1 (r', m')$ ，那么 $(r'', m'') \in S'$ 。因此，

$$IS_1(s'_1, \mathcal{R}) \subseteq S'.$$

当 s'_1 取遍 S_1 ，根据信息集的性质， S' 是 $IS_1(s'_1, \mathcal{R})$ 的不交并。

因为 $\mathcal{M}, (r, m) \models C(act_1(a))$ ，所以有 $\mathcal{M}, (r', m') \models act_1(a)$ 。这一公式意味着 $P_1(s'_1) = a$ 。根据 P 和 D 的关系，这等价于 $D(IS_1(s'_1, \mathcal{R})) = a$ 。因为这件事对任意 s'_1 都成立，根据 D 的并-一致性， $D(S') = a$ 。同理，从玩家 2 的角度来说， $D(S') = b$ 。因此， $a = b$ 。 \square

下面是一些关于这个定理的讨论：

- 我们的定理是对于确定性的策略证明的。然而，一个策略可能是非确定的，也就是在一个状态可能会有多种行动的选择，比如选择带有随机性。此时，达成一致定理依然成立，但我们需要恰当地定义 Kripke 模型和决策函数以适应非确定性的策略。
- 当策略具有非确定性时，我们可以用这一模型来理解带有先验知识、风险或者不确定性下的达成一致定理。只要策略能够对应一个并-一致的决策函数，结论都有效。这实际上是 Aumann 最初考虑的版本，即共同信念，而不是共同知识。这个版本的定理见习题 10。

§11.6 习题

1. * 考虑基础模态逻辑, 证明 Löb 公式: $L = \square(\square p \rightarrow p) \rightarrow \square p$ 定义的恰好是关系 R 满足如下条件的框架类:

- (1) 传递, 即 $\forall x, y, z(xRy \wedge yRz) \rightarrow xRz$.
- (2) 不存在 R -无穷迹, 即没有一列 $\{x_n\}_n$ (允许重复) 使得 x_iRx_{i+1} , $i = 0, 1, \dots$.
2. 证明命题 11.3、命题 11.4 和命题 11.5.
3. 形式化叙述并证明表 11.1 中公理和 R_i 性质之间的对应关系.
4. 证明引理 11.1.
5. 给定任意 Kripke 框架 \mathcal{F} , 证明: 如果 $\mathcal{F} \models \phi$, 那么 $\mathcal{F} \models C\phi$.
6. 证明命题 11.6 和命题 11.7.

7. * **S5 公理系统和 Aumann 结构的等价性.** 我们将分两个方向证明这一等价性. 考虑可能世界 (样本点) 的集合 S . 在 Aumann 结构 A^M 中, 我们定义了玩家 i 的信息集 \mathcal{P}_i 为 S 的一个划分. 自然, 划分本身就定义了 S 上的一个等价关系 R_i : $\mathcal{P}_i(s) = \mathcal{P}_i(s')$ 当且仅当 sR_is' . 这一等价关系自然对应了 Kripke 模型中 K_i 算子的关系. 这样, 就自然定义了由可能世界集 S 和等价关系 R_i 组成的 Kripke 框架 \mathcal{F} .

接下来, 考虑 Kripke 模型 $M = (\mathcal{F}, V)$, 其中 V 是将原子公式 p, q, \dots 赋值到 S 的赋值函数. 为了说明等价性, 我们还需要说明如何构造 Kripke 模型中逻辑公式和 Aumann 结构中事件的对应. 我们分两个方向做这件事.

- (1) 给定一个 Kripke 模型的公式 ϕ , 定义 Aumann 结构中的事件 ϕ^M 为

$$\phi^M = \{s \in S : M, s \models \phi\}.$$

特别地, 对原子公式 p , 我们可以对应到基本事件 p^M . 为了验证对应性, 还需要验证这一对应保持逻辑连接词, 请完成这一步证明.

- (2) 反之, Aumann 结构中并没有真正意义上的“基本事件”, 但是我们关心的问题会对每一个 S 中的状态赋予若干基本事实, 例如“孩子 i 脸上没有泥巴”. 这样, 我们自然而然定义了赋值函数 V , 它给出了基本事件和原子命题之间的联系. 为了验证对应性, 还需要验证这一对应保持逻辑连接词, 请完成这一步证明.
8. 我们使用知识逻辑建模了“泥泞的孩童”谜题. 假设脸上有泥巴的孩子有 k 个. 证明: 对任意 $l < k$, 第 l 轮回答之后, Kripke 模型中所有那些恰好有 l 个 1 的可能世界会被消掉.
9. 证明命题 11.8.
10. * 查阅资料, 叙述并证明由 Aumann 结构给出的概率版本的 Aumann 达成一致定理.

第六部分

附录：预备知识

附录 A 线性代数基础

§A.1 线性空间

从动机上说，线性空间试图将 \mathbb{R}^n 或者 \mathbb{C}^n 这样的集合连同他们上面的代数结构抽象出来。除此之外，函数和无穷数列的集合也是非常重要的对象，比如说 \mathbb{R} 上的连续函数组成的集合 $C(\mathbb{R})$ ，或者具有“模长”的无穷复数列 (ℓ^2 空间)：

$$\ell^2 = \left\{ (x_1, x_2, \dots) \in \mathbb{C}^\infty : \sum_{i=1}^{\infty} x_i^2 < \infty \right\}.$$

我们将这些对象的共性抽象出来，得到线性空间的概念。线性空间都是基于某个域定义的，我们先给出域的定义。

定义 A.1 (域) 一个域是一个集合 F ，其上定义了两种二元运算：加法 $+$ 和乘法 \cdot ，他们都是 $F \times F$ 到 F 的映射，满足下面的公理：

1. (结合律) 对于任意的 $a, b, c \in F$ ，有 $(a + b) + c = a + (b + c)$ 和 $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ ；
2. (交换律) 对于任意的 $a, b \in F$ ，有 $a + b = b + a$ 和 $a \cdot b = b \cdot a$ ；

3. (分配律) 对于任意的 $a, b, c \in F$, 有 $a \cdot (b + c) = a \cdot b + a \cdot c$.
4. (单位元) 存在唯一的两个元素 $0, 1 \in F$, 使得对于任意的 $a \in F$, 有 $a + 0 = a$ 和 $a \cdot 1 = a$;
5. (加法逆元) 对于任意的 $a \in F$, 存在唯一 $b \in F$, 使得 $a + b = 0$, 记 b 作 $-a$;
6. (乘法逆元) 对于任意的 $a \in F$, 如果 $a \neq 0$, 则存在唯一 $b \in F$, 使得 $a \cdot b = 1$, 记 b 作 a^{-1} .

通常将 $a \cdot b$ 写作 ab , 并且乘法的优先级高于加法, 即 $ab + c = (ab) + c$. \square

域的重要例子包括有理数域 \mathbb{Q} , 实数域 \mathbb{R} 和复数域 \mathbb{C} , 他们都是无限域. 我们将在后面的内容中使用这些域. 接下来, 我们定义线性空间.

定义 A.2 (线性空间, 向量空间) 设 V 是一个集合, F 是一个域. 如果在 V 上定义了两种运算: 加法 $+$ 和数乘 \cdot , 使得 V 满足下面的公理:

1. (V 的结合律) 对于任意的 $x, y, z \in V$, 有 $(x + y) + z = x + (y + z)$;
2. (V 的交换律) 对于任意的 $x, y \in V$, 有 $x + y = y + x$;
3. (加法零元) 存在唯一的元素 $0 \in V$, 使得对于任意的 $x \in V$, 有 $x + 0 = x$;
4. (加法逆元) 对于任意的 $x \in V$, 存在唯一 $y \in V$, 使得 $x + y = 0$, 记 y 作 $-x$;
5. 对于任意的 $x \in V$, 有 $1 \cdot x = x$;
6. 对于任意的 $a, b \in F$ 和 $x \in V$, 有 $(ab) \cdot x = a \cdot (b \cdot x)$;

7. 对于任意的 $a \in F$ 和 $x, y \in V$, 有 $a \cdot (x + y) = a \cdot x + a \cdot y$;
8. 对于任意的 $a, b \in F$ 和 $x \in V$, 有 $(a + b) \cdot x = a \cdot x + b \cdot x$.

则称 V 是一个 F -线性空间, 简称线性空间, 也称向量空间. V 中的元素被称为向量. 通常将数乘 $a \cdot x$ 写作 ax , 并且乘法的优先级高于加法, 即 $a \cdot x + y = (a \cdot x) + y$. \square

“线性”一词的含义是指的 $ax + by$ 这种形式的数学对象, 线性代数就是研究这种对象的学科. 线性空间的典型例子包括:

- \mathbb{R}^n 和 \mathbb{C}^n .
- $M_{m \times n}(F)$, 即所有 $m \times n$ 矩阵组成的集合.
- $C(\mathbb{R})$, 即 \mathbb{R} 上的连续函数组成的集合.
- $C^k(\mathbb{R})$, 即 \mathbb{R} 上的 k 次连续可微函数组成的集合.
- ℓ^2 空间, 即所有二次可和的复数序列组成的集合.

如同所有其他的代数结构, 线性空间也有各式各样构造新的线性空间的方法. 为了看出来线性空间本质的特性, 我们有如下引理:

引理 A.1 设 V 是 F -线性空间, W 是 V 的一个子集. 则 W 是一个线性空间当且仅当对任意 $a, b \in F$ 和 $x, y \in W$, 有 $ax + by \in W$.

证明. 按照定义即可验证. \square

我们给 $ax + by$ 这样的对象一个正式的定义.

定义 A.3 (线性组合) 设 V 是 F -线性空间, $x_1, \dots, x_n \in V$, $a_1, \dots, a_n \in F$, 则称 $a_1x_1 + \dots + a_nx_n$ 是 x_1, \dots, x_n 的一个线性组合. \square

接下来，基于某些特定的线性空间，我们构造各种新的线性空间.

定义 A.4 (线性子空间) 设 V 是 F -线性空间， W 是 V 的一个子集. 如果 W 是一个线性空间，则称 W 是 V 的一个线性子空间. \square

例如， \mathbb{Q} 是 \mathbb{R} 的一个线性子空间，但 \mathbb{Z} 不是 \mathbb{R} 的一个线性子空间.
再比如，当 $k < l$, $C^k(\mathbb{R})$ 是 $C^l(\mathbb{R})$ 的一个线性子空间.

定义 A.5 (乘积空间) 设 V_1, \dots, V_n 是 F -线性空间，则 $V_1 \times \dots \times V_n$ 是一个 F -线性空间，其中加法和数乘分别定义为

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n),$$

$$a(x_1, \dots, x_n) = (ax_1, \dots, ax_n).$$

 \square

例如， \mathbb{R}^n 就是 n 个 \mathbb{R} 的乘积空间， $M_{m \times n}(F)$ 就是 $m \times n$ 个 F 的乘积空间.

接下来，我们按照表示论的观点，引入基的概念. 线性空间是抽象的数学概念，因此我们需要一些具体的元素去表示这整个空间.

定义 A.6 (生成集) 设 V 是 F -线性空间， $S \subseteq V$ ，如果 V 中的每一个元素都是 S 的线性组合，则称 S 是 V 的一个生成集.

更一般地，任意一个 $S \subseteq V$ ，我们可以定义 S 生成的线性子空间为所有 S 的线性组合的集合，记为 $\text{Span}(S)$. \square

我们希望用尽可能少的元素来表示整个线性空间，为此，我们需要把“可表示”这样的概念严格化.

定义 A.7 (线性相关) 设 V 是 F -线性空间， $S \subseteq V$ ，如果存在 $x_1, \dots, x_n \in S$, $a_1, \dots, a_n \in F$ ，使得 $a_1x_1 + \dots + a_nx_n = 0$ ，且至少有一个 $a_i \neq 0$ ，则称 S 是线性相关的，否则称 S 是线性无关的. \square

S 线性相关意味着 S 中的一些元素可以被另一些元素的线性组合表示出来，因而 S 中有一些冗余。线性无关意味着 S 中的元素都是必要的，没有冗余。由此，我们可以给出基的定义。

定义 A.8 (基) 设 V 是 F -线性空间， $S \subseteq V$ ，如果 S 是线性无关的，并且 $\text{Span}(S) = V$ ，则称 S 是 V 的一个基。 \square

线性空间的一个核心定理是基的存在性定理。

定理 A.1 (基的存在性定理) 设 V 是 F -线性空间，则 V 中存在一个基。

要注意，这一定理不是平凡的。首先，基是线性无关的集合，所以 V 本身通常就不是基。此外，这一定理要求有一个线性无关的集合 $S \subseteq V$ ，任意向量 $x \in V$ 都可以用 S 中有限个元素的线性组合来表示，这样的 S 并不容易找到。该定理的证明是构造性的，这一构造依赖于选择公理（或者 Zorn 引理），我们在此略去。

基的典型例子包括：

- \mathbb{R}^n 的标准基是 $\{e_1, \dots, e_n\}$ ，其中 e_i 是第 i 个分量为 1，其余分量为 0 的向量。
- 特别地， \mathbb{R} 的基就是 $\{1\}$ ，一般地，域 F 作为线性空间的时候，其基就是 $\{1\}$ 。但如果我们将 C 看成 \mathbb{R} 的线性空间，那么 C 的基就是 $\{1, i\}$ 。
- $M_{m \times n}(F)$ 的标准基是 $\{E_{ij} : 1 \leq i \leq m, 1 \leq j \leq n\}$ ，其中 E_{ij} 是第 i 行第 j 列为 1，其余元素为 0 的矩阵。

特别注意，无穷维空间经常违背直觉。例如，考虑 ℓ^2 空间和向量组 $\{e_1, e_2, \dots\}$ ，其中 e_i 是第 i 个分量为 1，其余分量为 0 的实数列。这个

向量组看上去非常像一个基，然而并非如此！比如说， $(1/n)_{n=1}^{\infty} \in \ell^2$ ，但是它不能写成有限个 e_i 的线性组合。实际上， ℓ^2 空间的基一定是不可数的。

给定一个基，我们可以用基来表示线性空间中的元素，容易证明，这一表示是唯一的。因此，我们可以把线性空间中的元素看成基的线性组合，因而有了下面的定义。

定义 A.9 (坐标) 设 V 是 F -线性空间， S 是 V 的一个基， $x \in V$ ，如果 $x = \sum_{v \in S} a_v v$ ，则称 $(a_v)_{v \in S}$ 是 x 在基 S 下的坐标。□

例如， \mathbb{R}^3 的标准基是 $\{e_1, e_2, e_3\}$ 。任意 $x \in \mathbb{R}^3$ 都可以表示为 $x = a_1 e_1 + a_2 e_2 + a_3 e_3$ ，其中 a_i 是 x 的第 i 个分量。因此，我们可以把 x 看成一个三元组 (a_1, a_2, a_3) ，这就是 x 在标准基下的坐标。这样的讨论也适用于 \mathbb{R}^n 或 \mathbb{C}^n 。另外，坐标本身的集合也可以被看作是一个线性空间，例如， \mathbb{R}^3 的坐标集合就是 \mathbb{R}^3 本身。

线性空间的基可以衡量线性空间的复杂程度，基元素越少，线性空间越简单。我们可以定义维数来衡量线性空间的复杂程度。

定义 A.10 (维数) 设 V 是 F -线性空间，如果 V 的一个基有限，则称 V 是有限维的，否则称 V 是无限维的。有限维线性空间的基的元素个数称为 V 的维数，记为 $\dim V$ 。□

这一定义隐含的事实是，如果 V 有有限基，那么所有基都是有限的，并且任意两个基的元素个数相同。我们这里略去证明。

例如， \mathbb{R}^n 的维数是 n ， $M_{m \times n}(F)$ 的维数是 mn ， $C^k(\mathbb{R})$ 和 ℓ^2 都是无穷维的。

线性空间可以按照维数递降进行分解，变成越来越简单的线性空间的组合。这种组合称为直和。

定义 A.11 (和空间与直和) 设 V 是 F -线性空间, $U_1, U_2 \subseteq V$ 是 V 的子空间, 定义他们的和空间为

$$U_1 + U_2 = \{u_1 + u_2 : u_1 \in U_1, u_2 \in U_2\}.$$

如果 $U_1 \cap U_2 = \{0\}$, 换句话说, U_1 与 U_2 线性无关, 则称 $U_1 + U_2$ 是直和, 记为 $U_1 \oplus U_2$. 如果 $V = U_1 \oplus U_2$, 则称 U_1 和 U_2 是 V 的直和分解. \square

例如, \mathbb{R}^3 可以分解为 $\mathbb{R}^3 = \mathbb{R}e_1 \oplus \mathbb{R}e_2 \oplus \mathbb{R}e_3$, 其中 $\mathbb{R}e_i = \{\alpha e_i : \alpha \in \mathbb{R}\}$ 是 \mathbb{R}^3 的一维子空间, 它们的直和就是 \mathbb{R}^3 . 注意到这个分解将三维线性空间分解成了三个一维线性空间, 这不是偶然的, 一般地, 我们有下面的定理.

定理 A.2 (维数定理) 设 V 是有限维 F -线性空间, $V = U_1 \oplus U_2$, 则

$$\dim V = \dim U_1 + \dim U_2.$$

证明. 设 S_1 是 U_1 的一个基, S_2 是 U_2 的一个基, 那么根据直和的定义, $S_1 \cup S_2$ 是 V 的一个基. 因为 $S_1 \cup S_2$ 是线性无关的, 所以必然有 $S_1 \cap S_2 = \emptyset$. 又由于 V 中的任意元素都可以写成 $S_1 \cup S_2$ 中元素的线性组合, 因此

$$\dim V = |S_1 \cup S_2| = |S_1| + |S_2| = \dim U_1 + \dim U_2.$$

\square

通过直和分解, 我们可以把线性空间分成越来越简单的部分.

§A.2 线性映射

接下来我们研究线性空间之间的关系. 并不是所有的关系都是重要的, 我们所关心的是保持线性空间代数结构的这种关系, 这种关系称为线性映射.

定义 A.12 (线性映射, 线性算子, 线性函数) 设 V 和 W 是 F -线性空间, 如果映射 $f : V \rightarrow W$ 满足:

1. 对任意 $x, y \in V$, $f(x + y) = f(x) + f(y)$;
2. 对任意 $x \in V$ 和 $a \in F$, $f(ax) = af(x)$,

则称 f 是 V 到 W 的一个线性映射. 如果 $V = W$, 则称 f 是 V 上的一个线性算子或线性变换. 如果 $W = F$, 则称 f 是 V 上的一个线性函数. \square

一个更简洁但也更本质的定义是, 线性映射是保持线性组合的映射.

例 A.1 一个平凡的例子是零映射: $f : V \rightarrow W$, $f(x) = 0$, 这显然是线性映射, 我们通常记为 O . 另一个平凡的例子是恒等映射: $f : V \rightarrow V$, $f(x) = x$, 这也是线性映射, 我们通常记为 id . \square

线性映射有如下基本性质:

命题 A.1 设 $f : V \rightarrow W$ 是域 F 上的线性映射, 那么

1. $f(0) = 0$;
2. $f(-x) = -f(x)$;
3. $f(\sum_{i=1}^n a_i x_i) = \sum_{i=1}^n a_i f(x_i)$;

4. 如果 $g: W \rightarrow Z$ 是线性映射, 则 $g \circ f: V \rightarrow Z$ 也是线性映射;
5. 如果 $g: V \rightarrow W$ 是线性映射, $a, b \in F$, 则 $af + bg: x \mapsto af(x) + bg(x)$ 也是线性映射; 也是线性映射;
6. 如果 $g: V \rightarrow W$ 是线性映射, $h: W \rightarrow Z$ 是线性映射, $k: Z \rightarrow V$ 是线性映射, 则 $h \circ (f + g) = h \circ f + h \circ g$, $(f + g) \circ k = f \circ k + g \circ k$;
7. 如果 f 是双射, 则 f^{-1} 也是线性映射.

证明. 按照定义验证即可. □

为了简化记号, 我们会将线性映射 f 的作用 $f(x)$ 简记为 fx , 线性映射的复合 $g \circ f$ 简记为 gf , 同一线性映射 f 的 n 次复合简记为 f^n . 对于多项式函数 $G(x) = a_0 + a_1x + \cdots + a_nx^n$, 我们可以定义一个新的线性映射 $G(f) = a_0\text{id} + a_1f + \cdots + a_nf^n$.

线性映射可以被看成一种滤镜, 它可以将原始的空间进行变形, 变成一个新的空间. 比如说, 海上的月亮, 就是将三维空间的太阳与空间映到了海面上. 而线性算子则是一种特殊的线性映射, 它将原始空间变形成自身. 如果我们把线性空间看成一块橡皮泥, 那么线性算子可以被看成某种拉伸, 橡皮泥这个整体没有变多或者变少, 但是橡皮泥的形状发生了改变.

下面我们考虑两个线性映射的例子.

例 A.2 (微分算子) 考虑 $C^\infty(\mathbb{R})$, 即任意次可微的实函数空间. 求导 $d/dx: C^\infty(\mathbb{R}) \rightarrow C^\infty(\mathbb{R})$ 被称为微分算子. 容易验证, d/dx 是线性算子. □

例 A.3 (投影变换) 这个例子实际上是海上升明月的一般化. 考虑 \mathbb{R}^n , 设 $m \leq n$. 映射

$$\pi_m : (x_1, \dots, x_n) \mapsto (x_1, \dots, x_m, 0, \dots, 0)$$

称为 \mathbb{R}^n 的投影变换. 容易验证, π_m 是线性算子. 此外, 实际上, π_m 也可以被看作是 \mathbb{R}^n 到 \mathbb{R}^m 的线性映射, 将 $(x_1, \dots, x_m, 0, \dots, 0)$ 后面的 0 都丢掉, 这就是一个 \mathbb{R}^m 的元素. \square

从投影变换的例子中, 我们可以体会到线性空间的微妙之处: 不同的线性空间可能有着完全相同的本质. 由 $(x_1, \dots, x_m, 0, \dots, 0)$ 形成的空间实际上就是 \mathbb{R}^m , 只是我们用了 \mathbb{R}^n 的元素来表示它. 这里引申出来了代数中两个重要的概念: 同态与同构.

定义 A.13 (同态与同构) 设 V 和 W 是 F -线性空间, 如果映射 $f : V \rightarrow W$ 满足:

1. 对任意 $x, y \in V$, $f(x + y) = f(x) + f(y)$;
2. 对任意 $x \in V$ 和 $a \in F$, $f(ax) = af(x)$,

则称 f 是 V 到 W 的一个同态. 如果 f 是一个满射, 那么称 f 是一个满同态; 如果 f 还是一个单射, 那么称 f 是一个同构. \square

线性空间之间的同态实际上就是线性空间之间的线性映射, 所以同态是平凡的概念. 同态这个词表明了两个线性空间的相似性, 一个空间丢掉一些东西之后就可以被看成另一个空间的子空间. 而满同态则是说, 丢掉一些东西之后, 这个空间就是另一个空间. 比如说, 如果 $m < n$, 丢掉 \mathbb{R}^n 中元素的后面 $n - m$ 个分量, 就得到了 \mathbb{R}^m , 这就是

一个满同态. 同构则是说, 这两个线性空间就是一样的, 没有谁比谁更复杂, 比如说, \mathbb{R}^n 和 \mathbb{R}^m 就是同构的, 只要 $n = m$.

刚刚讨论的 \mathbb{R}^m 与 \mathbb{R}^n 的同构是具有一般性的, 这就是有限维线性空间的同构定理:

定理 A.3 (有限维线性空间的同构定理) 设 V 和 W 是有限维 F -线性空间, 则 V 与 W 同构当且仅当 $\dim V = \dim W$.

这一定理充分说明了, 有限维线性空间中维数的意义: 维数刻画了线性空间.

证明. 这一证明的思路是典型的: 先定义一个 (与基相关的) 基本映射, 然后进行扩张.

\Leftarrow : 如果 $\dim V = \dim W = n$, 那么两个线性空间的基的元素个数是一样的, 我们可以将它们一一对应起来. 比方说 V 的基是 $\{v_1, \dots, v_n\}$, W 的基是 $\{w_1, \dots, w_n\}$, 那么我们可以定义两个基之间的映射 f , 使得 $f(v_i) = w_i$.

我们可以将 f 扩张成一个线性映射. 比如说, 对于任意的 $x \in V$, 它用基表示就是 $\sum_{i=1}^n a_i v_i$. 我们可以定义

$$f(x) = f\left(\sum_{i=1}^n a_i v_i\right) = \sum_{i=1}^n a_i f(v_i) = \sum_{i=1}^n a_i w_i.$$

接下来验证 f 是 V 到 W 同构. 首先, 按照定义就可以验证这是一个线性映射. 其次, 因为 a_i 是任意的, 所以这显然也是一个满射. 最后, 如果有两个不同的 x, y 对应相同的 $f(x) = f(y)$, 那么 $f(x)$ 和 $f(y)$ 的坐标是一样的, 所以 x 和 y 的坐标也是一样的, 所以 $x = y$, 所以 f 也是一个单射.

\implies ：设两个线性空间由映射 $f : V \rightarrow W$ 给出同构。假设 V 的基是 $\{v_1, \dots, v_n\}$ ，我们证明 W 的基就是 $\{f(v_1), \dots, f(v_n)\}$ 。

首先，因为 f 是满射，而 v_i 生成了整个 V ，所以 $f(v_i)$ 生成了整个 W 。

再说明 $f(v_i)$ 线性无关。假设 $\sum_{i=1}^n a_i f(v_i) = 0$ ，那么 $f(\sum_{i=1}^n a_i v_i) = 0$ ，由于 f 是单射，所以 $\sum_{i=1}^n a_i v_i = 0$ ，由于 v_i 线性无关，所以 $a_i = 0$ ，所以 $f(v_i)$ 线性无关。

以上两点证明了 W 的基是 $\{f(v_1), \dots, f(v_n)\}$ ，所以 $\dim V = \dim W$. \square

此外，同构还有一个重要性质：

命题 A.2 假设 $f : V \rightarrow W$ 和 $g : W \rightarrow U$ 是两个线性映射，如果 f 和 g 都是同构，那么 $g \circ f$ 也是同构。

证明。根据定义即可证明。 \square

接下来我们进一步研究线性映射所带来的结构。我们刚刚说过，同态就是说把一些东西丢掉，剩下的东西可以被看成另一个空间的子空间。丢掉的东西是和剩下的东西，就是线性映射的核与像。

定义 A.14 (核与像) 设 V 和 W 是 F -线性空间， $f : V \rightarrow W$ 是一个线性映射。 f 的核定义为 $\ker f = \{x \in V : f(x) = 0\}$ ， f 的像定义为 $\text{Im } f = \{f(x) : x \in V\}$ 。 \square

“把一些东西丢掉”这一表述可以精确地由以下定理给出：

定理 A.4 设 V 和 W 是 F -线性空间， $f : V \rightarrow W$ 是一个线性映射，则 $\ker f$ 是 V 的线性子空间， $\text{Im } f$ 是 W 的线性子空间。另外，

$$\dim V = \dim \ker f + \dim \text{Im } f.$$

直观来说，这一定理表明，线性映射 f 把 V 抹掉了子空间 $\ker f$ ，最终得到了空间 $\text{Im } f$.

证明. 这一证明类似于定理 A.3 的证明，这里只给出思路，细节留给读者。首先选出 $\ker f$ 的基 v_1, \dots, v_k ，然后添加向量 u_1, \dots, u_l 扩充成 V 的基，然后证明 $f(u_1), \dots, f(u_l)$ 是 $\text{Im } f$ 的基。□

一个直接但重要的推论是：

推论 A.1 设 V 和 W 是 F -线性空间， $f : V \rightarrow W$ 是一个线性映射。那么以下性质成立：

1. $\dim \text{Im } f \leq \dim V$ ，等号成立当且仅当 $\ker f = \{0\}$ ；
2. f 是单射当且仅当 $\ker f = \{0\}$ ；
3. f 是满射当且仅当 $\dim \text{Im } f = \dim W$ ；
4. f 是同构当且仅当 $\ker f = \{0\}$ 且 $\dim \text{Im } f = \dim W$ 。

这一推论给了我们判断一个线性映射是否是单射、满射或者同构的方法。

最后，我们引入线性映射的秩的概念：

定义 A.15 (线性映射的秩) 设 V 和 W 是 F -线性空间， $f : V \rightarrow W$ 是一个线性映射。 f 的秩定义为 $\text{rank } f = \dim \text{Im } f$ 。□

换言之，线性映射的秩就是它的像的维数。秩越高的线性映射说明它把空间“压缩”得越少，丢掉的东西越少。例如， \mathbb{R}^n 的投影变换 π_m 的秩为 m ，说明它丢掉的东西只有 $n - m$ 维，也就是后面的 $n - m$ 个坐标。

推论 A.1 给出了复合线性映射秩的性质：

推论 A.2 设 V 、 W 和 U 是 F -线性空间, $f : V \rightarrow W$ 和 $g : W \rightarrow U$ 是两个线性映射, 则 $\text{rank}(g \circ f) \leq \text{rank } f$, 等号成立当且仅当 $\text{Im } f \cap \ker g = \{0\}$. 特别地, 如果 f, g 都是满射, 那么等号成立当且仅当 g 是同构. 此外, $\text{rank}(g \circ f) \leq \text{rank } g$, 如果 f 是满射, 那么等号成立.

证明. 根据推论 A.1, 我们有 $\text{rank}(g \circ f) = \dim \text{Im}(g \circ f) \leq \dim \text{Im } f = \text{rank } f$. 等号成立当且仅当在空间 $\text{Im } f$ 中 g 的核是 $\{0\}$, 换言之, $\text{Im } f \cap \ker g = \{0\}$. 特别地, 如果 f 是满射, 那么这一条件变为 $\ker g = \{0\}$, 如果 g 也是满射, 那么 g 是同构.

此外, 因为 $\text{Im } f \subseteq W$, 所以 $\text{Im}(g \circ f) \subseteq \text{Im } g$, 因此 $\text{rank}(g \circ f) = \dim \text{Im}(g \circ f) \leq \dim \text{Im } g = \text{rank } g$. 如果 f 是满射, 那么 $\text{Im } f = W$, 所以 $\text{Im}(g \circ f) = \text{Im } g$, 因此等号成立. \square

这一推论有非常直观的含义: 线性映射是同态, 因此会丢东西, 所以复合映射会丢更多的东西.

§A.3 矩阵

我们已经用基与坐标表示了线性空间, 接下来, 我们引入矩阵的概念来表示线性映射, 我们这一节考虑的线性空间都是有限维的. 考虑一个线性映射 $f : V \rightarrow W$, 如果 V 的基是 $\{v_1, \dots, v_n\}$, W 的基是 $\{w_1, \dots, w_m\}$, 那么 $f(v_i)$ 可以用 w_1, \dots, w_m 的线性组合表示出来, 即

$$f(v_i) = a_{1i}w_1 + \cdots + a_{mi}w_m.$$

我们把这些系数 a_{ji} 排成如下形状

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

此时 A 被称为矩阵. $m = n$ 的矩阵被称为方阵. $A_i = (a_{i1}, \dots, a_{in})$ 被称为矩阵 A 的第 i 行, $A^j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}$ 被称为矩阵 A 的第 j 列. 他们分别被叫做行向量与列向量. 第 i 行第 j 列的元素, 也就是 a_{ij} , 记为 A_{ij} . 为了节约空间, 列向量通常被写成转置的形式, 即 $(a_{1j}, \dots, a_{mj})^\top$. 注意, 坐标向量都被视作列向量.

假设在基 v_i 下 $x \in V$ 的坐标是 $X = (x_1, \dots, x_n)^\top$. 我们现在来计算 $f(x)$ 在基 w_i 下的坐标 $Y = (y_1, \dots, y_m)^\top$. 因为 $x = x_1v_1 + \dots + x_nv_n$, 所以

$$\begin{aligned} f(x) &= f(x_1v_1 + \dots + x_nv_n) \\ &= x_1f(v_1) + \dots + x_nf(v_n) \\ &= x_1(a_{11}w_1 + \dots + a_{m1}w_m) + \dots + x_n(a_{1n}w_1 + \dots + a_{mn}w_m) \\ &= (a_{11}x_1 + \dots + a_{1n}x_n)w_1 + \dots + (a_{m1}x_1 + \dots + a_{mn}x_n)w_m. \end{aligned}$$

因此, $f(x)$ 在基 w_i 下的坐标是 $(a_{11}x_1 + \dots + a_{1n}x_n, \dots, a_{m1}x_1 +$

$\cdots + a_{mn}x_n)^T$, $y_i = a_{i1}x_1 + \cdots + a_{in}x_n$. 我们将这一计算结果写作

$$Y = AX.$$

这就是矩阵与向量的乘法.

通过矩阵, 线性映射作用在向量上的结果可以被具体算出来. 从这个意义上说, 矩阵表示了线性映射. 这一观点可以用下图来表示:

$$\begin{array}{ccc} x & \xrightarrow{f} & f(x) \\ \downarrow & & \downarrow \\ X & \xrightarrow{A} & Y \end{array}$$

反之, 给定一个域 F 上的 $m \times n$ 矩阵 A , 我们可以定义一个线性映射 $f_A : F^n \rightarrow F^m$: 对 $X \in F^n$, $f_A(X) = AX$. f_A 被称为 A 诱导的线性映射. 所以, 矩阵本身也可以看成是一个线性映射, 不仅仅只是线性映射的表示.

我们看一个平凡的例子. 考虑零映射 $O : x \mapsto 0$. 不管在什么基下, O 的矩阵都是全零矩阵, 我们称为零矩阵, 依然使用符号 O 表示. 反过来, 如果一个线性映射的矩阵是零矩阵, 那么这个线性映射也是零映射.

我们再看一个的例子, 这个例子说明利用矩阵如何给出不同基之下的坐标变换公式. 设 V 是一个 n 维线性空间, $f : V \rightarrow V$ 是一个线性算子. V 的一组基是 $\{v_1, \dots, v_n\}$, 另一组基是 $\{v'_1, \dots, v'_n\}$. 定义一个 V 的自同构满足 $f(v_i) = v'_i$. 假设 f 在基 $\{v_i\}$ 下的矩阵是 A , 这被称为基 $\{v_i\}$ 到基 $\{v'_i\}$ 的过渡矩阵.

考虑一个点 $x \in V$, 它在基 $\{v_i\}$ 下的坐标是 $X = (x_1, \dots, x_n)^T$, 在

基 $\{v'_i\}$ 下的坐标是 $X' = (x'_1, \dots, x'_n)^\top$. 我们来计算 x 在基 $\{v'_i\}$ 下的坐标. 因为 $f(x) = f(x_1v_1 + \dots + x_nv_n) = x_1f(v_1) + \dots + x_nf(v_n)$, 而

$$f(v_i) = v'_i = \sum_{j=1}^n a_{ji}v_j,$$

所以

$$\begin{aligned} f(x) &= \sum_{i=1}^n x_i \sum_{j=1}^n a_{ji}v_j = \sum_{j=1}^n \left(\sum_{i=1}^n a_{ji}x_i \right) v_j \implies \\ x &= \sum_{j=1}^n \left(\sum_{i=1}^n a_{ji}x_i \right) f^{-1}(v_j) = \sum_{j=1}^n \left(\sum_{i=1}^n a_{ji}x_i \right) v'_j. \end{aligned}$$

因此, $X = AX'$.

线性映射相关的概念就可以被迁移到矩阵中来.

首先我们考虑映射的线性组合. 设 V 是 F -线性空间, $f : V \rightarrow W$ 和 $g : V \rightarrow W$ 是两个线性映射, $\lambda, \mu \in F$, 那么 $\lambda f + \mu g$ 也是一个线性映射. 如果 V 的基是 $\{v_1, \dots, v_n\}$, W 的基是 $\{w_1, \dots, w_m\}$. 假设在这些基下, f 和 g 的矩阵分别是 A 和 B , 那么 $\lambda f + \mu g$ 的矩阵可以很自然地记作 $\lambda A + \mu B$. 容易验证, $\lambda A + \mu B$ 的第 i 行第 j 列的元素是 $\lambda a_{ij} + \mu b_{ij}$. 用这样的办法, 我们就定义了矩阵的数乘和加法.

然后再考虑映射的复合. 设 V 是 F -线性空间, $f : V \rightarrow W$ 和 $g : W \rightarrow U$ 是两个线性映射. 如果 V 的基是 $\{v_1, \dots, v_n\}$, W 的基是 $\{w_1, \dots, w_m\}$, U 的基是 $\{u_1, \dots, u_l\}$. 假设在这些基下, g 和 f 的矩阵分别是 A 和 B , 那么复合 gf 的矩阵可以很自然地记作 AB . 我们来计

算 AB 的第 i 行第 j 列的元素. 因为 $gf(v_i) = g(f(v_i))$, 所以

$$\begin{aligned} gf(v_i) &= g(f(v_i)) \\ &= g(a_{1i}w_1 + \cdots + a_{mi}w_m) \\ &= a_{1i}g(w_1) + \cdots + a_{mi}g(w_m) \\ &= a_{1i}(b_{11}u_1 + \cdots + b_{l1}u_l) + \cdots + a_{mi}(b_{1m}u_1 + \cdots + b_{lm}u_l) \\ &= (a_{1i}b_{11} + \cdots + a_{mi}b_{1m})u_1 + \cdots + (a_{1i}b_{l1} + \cdots + a_{mi}b_{lm})u_l. \end{aligned}$$

因此, AB 的第 i 行第 j 列的元素是 $a_{1i}b_{j1} + \cdots + a_{mi}b_{jm}$. 这就是矩阵乘法的定义. 当有多个相同矩阵相乘时, 我们可以写成幂的形式. 比如, $A^2 = AA$, $A^3 = AAA$ 等等.

接下来, 我们考虑同构对应的矩阵. 最简单的同构是恒等映射 $\text{id} : V \rightarrow V$, 它的矩阵是单位矩阵 I_n . 容易看出, 无论在什么基下, I_n 都等于

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

有了单位矩阵, 类似算子的多项式, 给定多项式 $G(x) = a_0 + a_1x + \cdots + a_nx^n$, 我们可以定义矩阵 $G(A) = a_0I_n + a_1A + \cdots + a_nA^n$.

在更一般的情况下, 考虑 V 和 W 是 n 维的 F -线性空间, 基分别是 $\{v_1, \dots, v_n\}$ 和 $\{w_1, \dots, w_n\}$. 如果线性映射 $f : V \rightarrow W$ 将 v_i 映到 $\lambda_i w_i$,

那么 f 的矩阵就是

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

我们将这样的矩阵称为对角矩阵.

对于一般的同构映射 $f : V \rightarrow W$, 它有一个同构逆映射 $f^{-1} : W \rightarrow V$. 假设 V 的基是 $\{v_1, \dots, v_n\}$, W 的基是 $\{w_1, \dots, w_n\}$, 那么 f 和 f^{-1} 的矩阵分别是 A 和 B . 我们来计算 AB 和 BA . 因为 $f^{-1}f = \text{id}$, 所以 $AB = I_n$. 同理, $BA = I_n$. B 可以被看成 A 的逆元, 我们记作 $B = A^{-1}$. 这就是矩阵的逆的定义.

接下来, 我们引入矩阵转置的概念.

定义 A.16 (矩阵转置) 设 $A = (a_{ij})$ 是一个 $m \times n$ 矩阵, 我们定义 A 的转置为一个 $n \times m$ 矩阵 $A^T = (a_{ji})$. □

实矩阵的转置在线性映射中对应的是对偶空间的对偶映射, 这里我们不展开讨论了.

对于满足 $A^T = A$ 的矩阵, 我们称之为对称矩阵. 对于满足 $A^T = -A$ 的矩阵, 我们称之为反对称矩阵.

现在, 矩阵作为一个代数结构所需要的要素都已经给出了. 我们来看看矩阵的一些基本性质. 这些性质大多是从线性映射继承过来的 (命题 A.1), 我们在此略去证明.

命题 A.3 设 A, B, C 都是域 F 上的 n 阶方阵, $\lambda, \mu \in F$, 那么

1. $(\lambda A + \mu B)C = \lambda AC + \mu BC$;

2. $A(\lambda B + \mu C) = \lambda AB + \mu AC ;$
3. $(AB)C = A(BC) ;$
4. $A(B + C) = AB + AC ;$
5. $(A + B)C = AC + BC ;$
6. $AI_n = I_n A = A ;$
7. $AO = OA = O ;$
8. $A(-B) = (-A)B = A(-B) = -(AB) ;$
9. $(AB)^{-1} = B^{-1}A^{-1} ;$
10. $(A^{-1})^{-1} = A ;$
11. $(\lambda A)^{-1} = \lambda^{-1}A^{-1} ;$
12. $(A^T)^{-1} = (A^{-1})^T ;$
13. $(A^T)^T = A ;$
14. $(A + B)^T = A^T + B^T ;$
15. $(AB)^T = B^T A^T ;$
16. $(\lambda A)^T = \lambda A^T ;$
17. $(A^{-1})^T = (A^T)^{-1} ;$
18. $(A^T)^{-1} = (A^{-1})^T .$

回顾前面关于矩阵的表示论观点，从线性映射 f 得到矩阵 A 需要基于给定的基。一个自然的问题是，如果我们换了基，那么矩阵 A 会怎么变化？下面我们来具体计算。

设 $f : V \rightarrow V$ 是一个线性算子， V 的两个基分别是 $\{v_1, \dots, v_n\}$ 和 $\{v'_1, \dots, v'_n\}$ 。假设在这两个基下， f 的矩阵分别是 $A = (a_{ij})$ 和 $A' = (a'_{ij})$ 。我们来计算 A' 和 A 的关系。

假设 g 是一个自同构，使得 $g(v_i) = v'_i$ ，在 $\{v_1, \dots, v_n\}$ 下对应的矩阵是 B 。我们下面证明 $A' = B^{-1}AB$ ，注意到

$$fg(v_i) = f(g(v_i)) = f(v'_i) = \sum_j a'_{ji} v'_j = \sum_j a'_{ji} g(v_j),$$

因为 g 可逆，所以

$$g^{-1}fg(v_i) = \sum_j a'_{ji} v_j.$$

左边是 $B^{-1}AB$ ，而右边对应的就是 A' 。所以我们证明了

定理 A.5 设 V 是 F -线性空间， $f : V \rightarrow V$ 是一个线性算子， $\{v_1, \dots, v_n\}$ 和 $\{w_1, \dots, w_n\}$ 是 V 的两个基。假设 A 和 A' 分别是 f 在这两个基下的矩阵， B 是从 $\{v_1, \dots, v_n\}$ 到 $\{w_1, \dots, w_n\}$ 的过渡矩阵，那么 $A' = B^{-1}AB$ 。

矩阵 A 和 A' 通过可逆矩阵 B 联系了起来： $A' = B^{-1}AB$ ，行如这样的矩阵关系被称为相似，记作 $A \sim A'$ 。容易验证，相似是一个等价关系，也就是说：

- $A \sim A$ ；
- 如果 $A \sim A'$ ，那么 $A' \sim A$ ；

- 如果 $A \sim A'$, $A' \sim A''$, 那么 $A \sim A''$.

根据上面的讨论, 矩阵的相似关系对应了基的变换: 如果我们把 V 的基换成 V 的另一个基, 那么线性算子 f 的矩阵也会变成相似的另一个矩阵.

最后, 我们讨论矩阵的秩. 同样, 这一定义来自线性映射的秩:

定义 A.17 (矩阵的秩, 行空间, 列空间) 设 A 是一个 $m \times n$ 矩阵, $f_A : F^n \rightarrow F^m$ 是它诱导的线性映射.

- 定义 A 的秩为 f_A 的秩, 记为 $\text{rank } A$.
- A 的 m 个行向量生成了一个线性空间, 称为矩阵 A 的行空间; 类似地, 所有的列向量生成了一个线性空间, 称为矩阵 A 的列空间.
- 行空间的维数称为矩阵 A 的行秩, 列空间的维数称为矩阵 A 的列秩.
- 如果行秩等于 m , 即行空间的一组基就是所有行向量, 那么我们称 A 是行满秩的; 如果列秩等于 n , 即列空间的一组基就是所有列向量, 那么我们称 A 是列满秩的.
- 对于方阵来说, 如果它同时是行满秩的和列满秩的, 那么我们称它是满秩的. □

矩阵的秩最核心的定理是:

定理 A.6 设 A 是一个 $m \times n$ 矩阵, A 的行秩、列秩与秩都相等.

这一定理的证明通常涉及到矩阵的初等行变换, 也可以用对偶空间理论的证明, 我们这里就不给出了.

从线性映射复合的秩关系 (推论 A.2)，我们直接得到了矩阵乘法秩的性质：

命题 A.4 设 A 是一个 $m \times n$ 矩阵， B 是一个 $n \times p$ 矩阵，那么

$$\text{rank}(AB) \leq \min\{\text{rank } A, \text{rank } B\}.$$

如果 A 和 B 都是方阵，那么当 A 可逆时， $\text{rank}(AB) = \text{rank } B$ ；当 B 可逆时， $\text{rank}(AB) = \text{rank } A$.

§A.4 双线性型与二次型

本节考虑线性函数的一种推广，即双线性函数。它是两个变量的函数，而且每个变量都是线性的。这种函数在几何上有很多应用，比如内积。我们先来给出它的定义。

定义 A.18 (双线性型) 设 V 是一个 F -线性空间，如果 V 上有一个映射 $f : V \times V \rightarrow F$ ，满足

1. 对于任意的 $v \in V$ ， $f(v, \cdot) : V \rightarrow F$ 是一个线性映射；
2. 对于任意的 $w \in V$ ， $f(\cdot, w) : V \rightarrow F$ 是一个线性映射。

那么称 f 是 V 上的一个双线性型。 □

类似线性映射，我们的首要任务是表示一个双线性型。实际上，双线性型也可以用矩阵来表示。选定一组 V 的基 $\{v_1, \dots, v_n\}$ ，任意给定

两个向量 $x = \sum_{i=1}^n x_i v_i$ 和 $y = \sum_{i=1}^n y_i v_i$, 我们有

$$\begin{aligned} f(x, y) &= f\left(\sum_{i=1}^n x_i v_i, \sum_{j=1}^n y_j v_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i y_j f(v_i, v_j). \end{aligned}$$

如果我们知道了 $f(v_i, v_j)$, 那么 $f(x, y)$ 就完全可以用 x, y 的坐标表示出来. 这就是说, 我们可以用矩阵来表示双线性型. 定义 $A_{ij} = f(v_i, v_j)$, 我们将矩阵 $A = (A_{ij})$ 称为 f 在基 $\{v_1, \dots, v_n\}$ 下的矩阵. 假设 x, y 的坐标是 X, Y , 那么我们有

$$f(x, y) = X^T A Y.$$

反过来, 如果给定了一个 n 阶方阵 A , 那么我们可以定义一个双线性型 f_A , 使得 $f_A(v_i, v_j) = a_{ij}$, 这样的双线性型称为由矩阵 A 诱导的双线性型. 于是, 双线性型就和矩阵一一对应了.

类似线性映射, 双线性映射关心的一个重要问题是基变换. 设 A 是一个双线性型 f 在基 $\{v_1, \dots, v_n\}$ 下的矩阵, 而 A' 是基 $\{v'_1, \dots, v'_n\}$ 下的矩阵, 假设 $\{v_i\}$ 到 $\{v'_i\}$ 的过渡矩阵是 P . 现在任取 $x, y \in V$, 他们在基 $\{v_i\}$ 下的坐标分别是 X, Y , 在基 $\{v'_i\}$ 下的坐标分别是 X', Y' , 那么我们有

$$f(x, y) = X^T A Y = X'^T A' Y'.$$

根据坐标的基变换公式, $X = P X'$, $Y = P Y'$, 所以

$$X^T A Y = (AX')^T P (AY') = X'^T (A^T P A) Y'.$$

由于 x, y 是任意的, X, Y 也是任意的, 联立上面两式, $A' = A^T P A$, 这就是基变换公式. 对应到矩阵中, 这被称为合同变换.

定义 A.19 (合同矩阵) 设 A, B 都是 n 阶方阵, 如果存在一个可逆方阵 P , 使得 $B = P^T A P$, 那么称 A 和 B 是合同的. \square

容易验证, 合同关系是一个等价关系, 这与相似关系是类似的. 根据命题 A.4, 可逆矩阵相乘不改变矩阵的秩, 所以合同矩阵的秩是相同的. 于是, 双线性型的任意矩阵表示都有相同的秩, 我们因此可以定义双线性型的秩:

定义 A.20 (双线性型的秩) 设 f 是 V 上的一个双线性型, 如果 f 在某个基下的矩阵的秩是 r , 那么称 f 的秩是 r . \square

接下来, 我们考虑一种特殊的双线性型: 对称双线性型, 即 $f(x, y) = f(y, x)$ 对任意 $x, y \in V$ 成立. 注意到, 对称双线性型对应的矩阵是对称矩阵. 我们现在令 $x = y$, 定义 $q(x) = f(x, x)$, 那么 q 是一个实值函数, 这样的函数便是二次型.

定义 A.21 (二次型) 设 V 是数域 F 上的线性空间, f 是 V 上的一个对称双线性型, 那么定义 $q : V \rightarrow F$ 为 $q(x) = f(x, x)$, 称 q 是 f 诱导的二次型. \square

自然, 秩的概念也可以被迁移到二次型上:

定义 A.22 (二次型的秩) 设 q 是 V 上的一个二次型, 如果定义 q 的秩为诱导它的双线性型 f 的秩. \square

实际上, 二次型本身也可以算出双线性型: 设 q 是二次型, 那么我们可以定义

$$f(x, y) = \frac{1}{2}(q(x + y) - q(x) - q(y)).$$

容易验证, f 是一个对称双线性型. 于是, 二次型和对称双线性型是一一对应的.

自然, 二次型也可以用坐标表示. 选定一组基, 假设二次型 q 对应的对称双线性型是 f , 那么 q 在这组基下的矩阵就是 f 在这组基下的矩阵. 这样, 二次型就和对称矩阵一一对应了. 如果再给定 $x \in V$ 的坐标 X , 那么 $q(x) = X^T A X$. 如果把它展开写, 就是

$$q(x) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j.$$

这是一个关于 x_i 的二次函数. 通常来说, 我们希望化简这一表示, 如果能写成 $\sum_{i=1}^n a_{ii} x_i^2$ 的形式, 那么计算都会变得非常容易. 此时, 二次型 q 的矩阵是对角矩阵.

定义 A.23 (规范型) 设 V 是数域 F 上的线性空间, q 是 V 上的一个二次型, 如果存在一组基, 使得 q 在这组基下的矩阵是对角矩阵, 那么称这个对角矩阵是 q 的规范型, 这组基是规范基. \square

二次型的一个核心定理是, 规范型总是存在:

定理 A.7 设 V 是 F -线性空间, q 是 V 上的一个二次型, 那么 q 存在规范型.

证明. 首先, 选定一组基 $\{v_1, \dots, v_n\}$, 保证 $q(v_1) \neq 0$ (这样的 v_1 一定存在, 否则 q 就是零映射, 自然有规范型) 我们想办法把它变成另外一组基, 使得二次型 q 的矩阵是对角矩阵. 假设 q 对应的双线性型是 f . 对维数 n 用归纳法.

如果 $n = 1$, 这是显然的.

现在考虑一般的 n , 我们想办法将矩阵的第一行和第一列非对角的元素都变成 0, 那么剩下的矩阵实际上就在一个 $n - 1$ 维空间上的双线性型, 于是就可以用归纳假设了. 注意到这些元素其实就是 $f(v_1, v_j)$, 所以只需要把他们变成 0 就可以了.

注意到 $f(v_1, \cdot)$ 是非零线性函数, 秩是 1, 所以 $\dim \ker f(v_1, \cdot) = n - 1$, 于是我们选出核的基 v'_2, \dots, v'_n . 另外 $v_1 \notin \ker f(v_1, \cdot)$, 所以 v_1 与 $\{v'_2, \dots, v'_n\}$ 线性无关, 于是 $\{v_1, v'_2, \dots, v'_n\}$ 是一组基. 根据核的定义, 此时 $f(v_1, v'_i) = 0$, 因此这样就把第一行非对角的元素都变成了 0. \square

相应地, 在矩阵上, 这一定理的表述为:

推论 A.3 任何对称矩阵 A 都合同于一个对角矩阵.

在实数域上, 这一定理还可以被加强:

定理 A.8 (惯性定理) 设 V 是 \mathbb{R} 上的 n 维线性空间, q 是 V 上的一个二次型, 那么 q 存在形如

$$q(x) = \sum_{i=1}^r \lambda_i x_i^2$$

的规范型 ((x_i) 是 x 的坐标), 其中 $\lambda_i \in \{1, -1\}$, r 是 q 的秩, 且 λ_i 中 1 的个数和 -1 的个数只依赖于 q , 不依赖于规范基的选取.

这一定理我们就不再给出证明了.

惯性定理给出了几类特殊的二次型:

定义 A.24 (正定, 半正定, 负定, 半负定) 设 V 是 \mathbb{R} 上的 n 维线性空间, q 是 V 上的一个二次型.

- 如果 q 的规范型是

$$q(x) = \sum_{i=1}^n x_i^2,$$

那么称 q 是正定的.

- 如果 q 的规范型是

$$q(x) = \sum_{i=1}^r x_i^2 \quad (r \leq n),$$

那么称 q 是半正定的.

- 如果 q 的规范型是

$$q(x) = -\sum_{i=1}^r x_i^2,$$

那么称 q 是负定的.

- 如果 q 的规范型是

$$q(x) = -\sum_{i=1}^r x_i^2 \quad (r \leq n),$$

那么称 q 是半负定的.

对于实对称矩阵 A , 如果它对应的二次型是正定/半正定/负定/半负定的, 那么称 A 是正定/半正定/负定/半负定的. \square

以上概念都可以直接用二次型的取值去等价定义:

命题 A.5 设 V 是 \mathbb{R} 上的 n 维线性空间, q 是 V 上的一个二次型.

- q 是正定的当且仅当对任意的非零向量 $x \in V$, 都有 $q(x) > 0$.

- q 是半正定的当且仅当对任意的非零向量 $x \in V$, 都有 $q(x) \geq 0$.
- q 是负定的当且仅当对任意的非零向量 $x \in V$, 都有 $q(x) < 0$.
- q 是半负定的当且仅当对任意的非零向量 $x \in V$, 都有 $q(x) \leq 0$.

证明. 选定一组规范基, 按照定义验证即可. \square

自然, 这一命题的矩阵版本也是成立的. 一个直接的推论是:

推论 A.4 设 A 是一个矩阵, 那么 $A^T A$ 和 AA^T 都是半正定的. 此外, B 是一个正定矩阵当且仅当存在可逆矩阵 P , 使得 $B = P^T P$.

§A.5 带内积的线性空间

内积的考虑是从几何中来的. 一个典型的例子是平面欧氏几何. 我们知道, 笛卡尔的平面解析几何等价于平面欧氏几何. 建立坐标系的过程实际上就是选定了一个基, 而坐标就是基的坐标. 在这个基下, 平面上的点可以用坐标表示. 然而, 并不是所有的坐标轴都是好计算的, 我们考虑的是互相垂直的坐标轴, 此时, 平面上点的坐标就完全可以用投影来表示了. 计算投影的过程实际上就是内积的过程. 将平面解析几何的内积定义一般化, 我们就得到了线性空间的内积.

定义 A.25 (内积) 设 V 是一个实线性空间, 如果 V 上有一个对称双线性型 $\langle \cdot, \cdot \rangle : V \times V \rightarrow F$, 它诱导的二次型是正定的, 那么称 $\langle \cdot, \cdot \rangle$ 是 V 上的一个内积, 称 V 是一个内积空间. \square

注意, 讨论内积的时候, 我们只考虑实线性空间, 这是因为实数可以比大小, 并且不会像有理数那样对根号不封闭, 所以可以定义模长. 自然, \mathbb{R}^n 是内积空间, 因为我们可以定义 $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$.

利用内积，我们可以定义模长。

定义 A.26 (模) 设 V 是一个内积空间， $v \in V$ ，定义 v 的模或内积诱导的范数为 $\|v\| = \sqrt{\langle v, v \rangle}$. \square

容易证明，向量的模等于零当且仅当它是零向量，此外，对任意的 $v \in V$ 和 $\lambda \in \mathbb{R}$ ，有 $\|\lambda v\| = |\lambda| \|v\|$. 模长为 1 的向量称为**单位向量**.

反过来，内积诱导的范数也可以表示内积：

$$\langle v, w \rangle = \frac{1}{4} \left(\|v + w\|^2 - \|v - w\|^2 \right). \quad (\text{A.1})$$

利用内积，我们可以推广平面几何中的各种概念。首先是垂直的概念。

定义 A.27 (正交) 设 V 是一个内积空间， $v, w \in V$ ，如果 $\langle v, w \rangle = 0$ ，那么称 v 与 w 正交，记作 $v \perp w$. \square

对于一般情况，两个向量会有夹角的概念，我们可以利用内积来定义。

定义 A.28 (夹角) 设 V 是一个内积空间， $v, w \in V$ ，如果 $\theta \in [0, \pi]$ 满足

$$\cos \theta = \frac{\langle v, w \rangle}{\|v\| \|w\|},$$

那么称 θ 是 v 与 w 的夹角。 \square

夹角对任意非零向量都可以定义，这是因为内积有 Cauchy 不等式：

定理 A.9 (Cauchy 不等式) 设 V 是一个内积空间， $v, w \in V$ ，那么有

$$|\langle v, w \rangle| \leq \|v\| \|w\|.$$

证明. 取 $\lambda \in \mathbb{R}$, 那么

$$0 \leq \langle v + \lambda w, v + \lambda w \rangle = \|v\|^2 + 2\lambda \langle v, w \rangle + \lambda^2 \|w\|^2.$$

将最右边看作是 λ 的函数, 这是一个二次函数, 因为它恒大于等于 0, 所以判别式 $\Delta \leq 0$, 即

$$4 \langle v, w \rangle^2 - 4 \|v\|^2 \|w\|^2 \leq 0 \iff |\langle v, w \rangle| \leq \|v\| \|w\|. \quad \square$$

利用 Cauchy 不等式, 我们可以证明模长满足三角不等式:

定理 A.10 (三角不等式) 设 V 是一个内积空间, $v, w \in V$, 那么有

$$\|v + w\| \leq \|v\| + \|w\|.$$

证明. 由 Cauchy 不等式, 我们有

$$\begin{aligned} \|v + w\|^2 &= \langle v + w, v + w \rangle \\ &= \|v\|^2 + 2 \langle v, w \rangle + \|w\|^2 \\ &\leq \|v\|^2 + 2 \|v\| \|w\| + \|w\|^2 \\ &= (\|v\| + \|w\|)^2. \end{aligned} \quad \square$$

利用以上性质, 容易验证, 模实际上给了 V 一个范数. 关于范数的详细讨论, 见附录 B.1.1. 对于一般的范数, 我们无法像 (A.1) 一样去定义内积, 所以内积有它独特的性质.

内积还给出了投影的概念:

定义 A.29 (投影) 设 V 是一个内积空间, $v, w \in V$, 如果 $\lambda \in \mathbb{R}$ 满足 $\langle v - \lambda w, w \rangle = 0$, 那么称 λw 是 v 在 w 上的投影, 其中 $\lambda = \langle v, w \rangle / \|w\|^2$. \square

接下来, 我们继续表示论的观点, 讨论内积空间中的基与坐标. 首先是正交与线性无关的关系.

命题 A.6 设 V 是一个内积空间, 两两正交的非零向量 $v_1, \dots, v_n \in V$ 是线性无关的.

证明. 设 $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ 满足 $\sum_{i=1}^n \lambda_i v_i = 0$, 那么

$$0 = \left\langle \sum_{i=1}^n \lambda_i v_i, v_j \right\rangle = \sum_{i=1}^n \lambda_i \langle v_i, v_j \rangle = \lambda_j \|v_j\|^2.$$

因为 $v_j \neq 0$, 所以 $\lambda_j = 0$, 这就证明了线性无关性. \square

两两正交的基被称为正交基, 如果正交基的每个向量都是单位向量, 那么称为标准正交基. 在内积空间中, 基存在性定理 (定理 A.1) 可以被加强为标准正交基存在性定理:

定理 A.11 (标准正交基存在性定理) 设 V 是一个有限维内积空间, 那么 V 中存在一个标准正交基.

我们只提示这一定理的证明思路. 首先选择一个基, 然后利用 Gram-Schmidt 正交化方法将它正交化, 再将它单位化.

我们来看看这一定理的用处. 首先, 它给出了计算坐标的简易方式.

命题 A.7 设 V 是一个内积空间, 它的标准正交基是 e_1, \dots, e_n , $v \in V$, 那么

$$v = \sum_{i=1}^n \langle v, e_i \rangle e_i.$$

另外，成立勾股定理：

$$\|v\|^2 = \sum_{i=1}^n \langle v, e_i \rangle^2.$$

在标准正交基下，投影的系数就是坐标。此时，内积也可以被写成矩阵的形式。假设 $x, y \in V$ ，他们的坐标分别是 X 和 Y ，那么 $\langle x, y \rangle = X^T Y$ 。

使用标准正交基的另一个好处是，线性函数的表示变得简单了。给定 V 的一个标准正交基 $\{e_i\}_{i=1}^n$ 和 V 上的线性函数 f ，对任意一个向量 $v = \sum_{i=1}^n v_i e_i$ ，

$$f(v) = \sum_{i=1}^n v_i f(e_i).$$

考虑向量 $e_f = \sum_{i=1}^n f(e_i) e_i$ ，容易验证 $f(v) = \langle e_f, v \rangle$ 。

另一方面，这样的 e_f 必定是唯一的。如果有两个 e_f, e'_f 使得 $\langle e_f, v \rangle = \langle e'_f, v \rangle = f(v)$ 对任意的 v 都成立，那么取 $v = e_f - e'_f$ 得

$$\langle e_f - e'_f, e_f - e'_f \rangle = 0 \iff e_f - e'_f = 0.$$

综上，我们可以用一个向量的内积来表示线性函数：

定理 A.12 (Riesz 表示定理) 设 V 是一个内积空间， f 是 V 上的一个线性函数，那么存在唯一的向量 $u \in V$ ，使得 $f(v) = \langle u, v \rangle$ 对任意的 $v \in V$ 成立。

反之，给定一个向量 u ，很容易验证， $f_u(\cdot) = \langle u, \cdot \rangle$ 就是一个线性函数。如此我们就给出了内积空间中的线性函数和向量的一一对应。

此外，利用标准正交基，内积空间中一组向量的线性无关性也可以用内积来判断，这就是 Gram 矩阵。

定义 A.30 (Gram 矩阵) 给定内积空间 V 的一组向量 v_1, \dots, v_k , 定义他们的 **Gram 矩阵** 为 $G = (\langle v_i, v_j \rangle)_{k \times k}$. \square

利用标准正交基, 很容易计算 Gram 矩阵. 我们其实已经见过这样的例子. 给定任意一个实矩阵 A , $A^T A$ 就是列向量的 Gram 矩阵, AA^T 就是行向量的 Gram 矩阵.

Gram 矩阵的基本性质是:

命题 A.8 设 V 是一个内积空间, $v_1, \dots, v_k \in V$, 他们的 Gram 矩阵为 G , 那么

1. G 是对称矩阵;
2. G 是半正定的;
3. v_1, \dots, v_k 线性无关当且仅当 G 正定.

证明. 1. 显然.

2. 考虑 \mathbb{R}^k 上的二次型 $f(x) = x^T Gx$. 对任意的 $x \in \mathbb{R}^k$, 有

$$x^T G x = \sum_{i,j=1}^k x_i x_j \langle v_i, v_j \rangle = \left\langle \sum_{i=1}^k x_i v_i, \sum_{j=1}^k x_j v_j \right\rangle = \left\| \sum_{i=1}^k x_i v_i \right\|^2 \geq 0.$$

因此 G 是半正定的.

3. 我们已经证明 $f(x) \geq 0$. 由命题 A.5, G 是正定的当且仅当等价式 $f(x) = 0 \iff x = 0$ 成立. 而 $f(x) = \left\| \sum_{i=1}^k x_i v_i \right\|^2$, 因此 $f(x) = 0 \iff \sum_{i=1}^k x_i v_i = 0$, 所以 $x = 0 \iff v_1, \dots, v_k$ 线性无关. \square

利用第三点，我们可以很容易地判断一组向量的线性无关性。

半正定和对称性还暗示着， G 可以形成某种半正定的二次型。我们在证明中已经给出这样的二次型 f 。在附录 C.3.3，我们会遇到这样的例子，即一族随机变量的协方差矩阵。

向量组之间的正交性也可以用 Gram 矩阵来刻画：

命题 A.9 设 V 是一个内积空间， $v_1, \dots, v_k \in V$ ，那么下列命题等价：

1. v_1, \dots, v_k 两两正交；
2. G 是对角矩阵。

特别地， v_1, \dots, v_k 是标准正交的当且仅当 $G = I_k$ 。

这一命题的证明是显然的。

内积空间中，直和分解也可以被加强。为此，我们先引入正交补的概念。

定义 A.31 (正交补) 设 V 是一个内积空间， $W \subseteq V$ ，那么 W 的正交补是

$$W^\perp = \{v \in V : \forall w \in W, \langle v, w \rangle = 0\}.$$

□

定理 A.13 设 V 是一个内积空间， $W \subseteq V$ 是一个有限维子空间，那么

$$V = W \oplus W^\perp, \quad (W^\perp)^\perp = W.$$

这一定理的证明思路类似定理 A.4 的证明，区别是这里需要扩充标准正交基。这里不再给出具体证明。

最后，我们考虑标准正交基之间的过渡矩阵。设 V 是一个内积空间， e_1, \dots, e_n 和 e'_1, \dots, e'_n 都是 V 的标准正交基。设 $e'_j = a_{1j}e_1 + \dots + a_{nj}e_n$ ，如此就得到了过渡矩阵 A 。我们来看看 A 的性质。

命题 A.10 设 V 是一个内积空间, e_1, \dots, e_n 和 e'_1, \dots, e'_n 都是 V 的标准正交基. 设 e_i 到 e'_i 的过渡矩阵是 A , 那么 $AA^\top = A^\top A = I_n$.

证明. 设 $A = (a_{ij})$, 那么

$$\begin{aligned}\langle e'_i, e'_j \rangle &= \left\langle \sum_{k=1}^n a_{ki} e_k, \sum_{l=1}^n a_{lj} e_l \right\rangle \\ &= \sum_{k=1}^n \sum_{l=1}^n a_{ki} a_{lj} \langle e_k, e_l \rangle \\ &= \sum_{k=1}^n a_{ki} a_{kj}.\end{aligned}$$

当 $i = j$, 上式就是 $\sum_{k=1}^n a_{ki}^2 = 1$, 当 $i \neq j$, 上式就是 $\sum_{k=1}^n a_{ki} a_{kj} = 0$. 这就证明了 $AA^\top = I_n$. 同理可证 $A^\top A = I_n$. \square

我们将满足 $A^\top A = AA^\top = I_n$ 的矩阵称为**正交矩阵**, 它的逆矩阵就是它的转置矩阵.

正交矩阵有很多等价的刻画:

定理 A.14 设 A 是一个 n 阶方阵. 下列陈述等价:

1. A 是一个正交矩阵.
2. 对任意 $v \in \mathbb{R}^n$, 都有 $\|Av\| = \|v\|$.
3. A 的行向量是两两正交的单位向量.
4. A 的列向量是两两正交的单位向量.
5. $A^\top = A^{-1}$.

证明. 按照定义写出即可. \square

利用基变换，我们马上可以得到以下正交矩阵的性质：

命题 A.11 设 A, B 是 n 阶正交矩阵，那么

1. I_n 是正交矩阵；
2. AB 是正交矩阵；
3. A^{-1} 是正交矩阵；

这些性质使得正交矩阵构成了一个群，称为**正交群**，记作 $O(n)$. 这超出了本书的范围，我们就不继续深入了.

接下来，我们讨论内积空间的同构.

定义 A.32 (等距映射与等距同构) 设 V 和 W 都是内积空间， $T : V \rightarrow W$ 是一个线性映射，如果对任意的 $v_1, v_2 \in V$ ，都有 $\langle Tv_1, Tv_2 \rangle = \langle v_1, v_2 \rangle$ ，那么称 T 是一个等距映射. 如果 T 是一个双射，那么称 T 是一个等距同构. □

内积空间的同构是线性空间同构的加强版，因为它还要求保持内积.

同样，内积空间的等距同构类只取决于维数；

定理 A.15 设 V 和 W 都是有限维内积空间，那么 V 与 W 等距同构当且仅当 $\dim V = \dim W$.

证明. 证明完全类似定理 A.3 的证明，此时将 V 的标准正交基一一对应到 W 的标准正交基上. □

等距同构对应的矩阵恰好就是正交矩阵.

定理 A.16 设 V 和 W 都是有限维内积空间, $f : V \rightarrow W$ 是一个等距同构, 那么在 V 和 W 的标准正交基之下 f 是一个正交矩阵. 反之, n 阶正交矩阵 A 诱导的线性映射 $f_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 是一个等距同构.

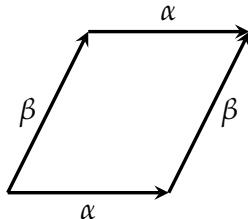
证明. 后半部分 (“反之” 之后) 由定理 A.14 的第二点直接得出. 我们来证明前半部分.

设 v_1, \dots, v_n 是 V 的标准正交基, w_1, \dots, w_n 是 W 的标准正交基, A 是 f 在这两组基下的矩阵. 考虑任意一个点 $x \in V$, 它的坐标是 X . 那么 $f(x)$ 的坐标是 AX . 由于 f 是等距同构, 所以 $\|f(x)\| = \|x\|$, 根据命题 A.7, 这等价于 $\|AX\| = \|X\|$, 由定理 A.14 的第二点, A 是一个正交矩阵. \square

§A.6 行列式

行列式可以进一步理解为矩阵的表示: 将很多个数的矩阵压缩到一个数. 我们将会从几何观点讨论, 先从平面开始.

考虑平面 \mathbb{R}^2 上的两个向量 $\alpha = (a_1, a_2)^\top$ 和 $\beta = (b_1, b_2)^\top$, 我们可以用这两个向量作为平行四边形的两条边, 构造一个平行四边形:



现在我们定义这个平行四边形的有向面积. 数值上, 有向面积就是我们通常理解的平行四边形面积. 有向面积的符号按照如下的规则给

出. 从 α 旋转到 β 所在的方向, 转动一个平角以内的角度. 如果这个角度是逆时针的, 面积就是正的, 否则就是负的.

容易算出, 这个有向面积是 $a_1 b_2 - a_2 b_1$, 可以使用如下形象的记号表示:

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = |\alpha \quad \beta|.$$

我们可以把这个面积的计算推广到 n 维空间中去. 设 A 是一个 n 阶方阵, 它的列向量是 A^1, \dots, A^n , 我们考虑这些列向量张成的 n 维平行体:

$$\Pi(A) = \left\{ x_1 A^1 + \cdots + x_n A^n : 0 \leq x_i \leq 1 \right\}.$$

平行体的体积可以归纳定义. 一维的情况下, 向量就是实数, 这个实数的绝对值就作为体积. 假设已经定义了 $n-1$ 维平行体的体积, 那么 n 维平行体的体积就是它的底的体积乘以高. 我们需要定义什么是底什么是高. 底是 $n-1$ 维平行体 $\Pi(A^1, \dots, A^{n-1})$. 为定义高, 先将 A^n 投影到 $\text{Span}(A^1, \dots, A^{n-1})$, 投影 A_*^n 就是垂足, 高就是 $A^n - A_*^n$ 的长度. 这样, 我们就得到了 n 维平行体的体积的定义.

行列式与有向体积有关. 我们这里不会专门定义有向体积的概念, 只给出一个二维情况下的直观. 通常 \mathbb{R}^2 的标准正交基 e_1, e_2 , 从 e_1 到 e_2 是逆时针的. 我们之前定义有向面积的时候也遵循了这样的原则: 与 e_1, e_2 形成相同定向 (即顺逆时针) 的面积是正的, 与 e_1, e_2 形成相反定向的面积是负的. 定向这一概念本质地反映了一组向量的顺序.

对于一般情况, 向量 A^1, \dots, A^n 和 \mathbb{R}^n 的标准正交基 e_1, \dots, e_n 之间同定向时 $\Pi(A)$ 的体积是正的, 否则就是负的. 这样, 我们就定义了有向体积. 我们指出, 行列式的定义恰好就给出了有向体积的计算公式. 下面我们给出行列式的定义.

定义 A.33 (行列式) 设 A 是一个 n 阶方阵, A 的行列式 $\det A$ 归纳定义为

1. $n = 1$ 时, $\det A = A_{11}$;
2. $\det A = A_{11} \det R_{11} - A_{12} \det R_{12} + \cdots + (-1)^{1+n} A_{1n} \det R_{1n}$,
其中 R_{ij} 是 A 是 A 去掉第 i 行第 j 列以后得到的矩阵. □

下面仅罗列一些行列式的性质, 但不给出证明.

命题 A.12 设 A, B 是两个 n 阶方阵, 则

1. $\det I_n = 1$;
2. $\det(AB) = (\det A)(\det B)$;
3. $\det(\lambda A) = \lambda^n \det A$;
4. $\det(A^{-1}) = (\det A)^{-1}$;
5. $\det A = 0$ 当且仅当 A 不可逆;
6. $\det(A^T) = \det A$;

矩阵的行列式自然地定义了线性算子的行列式. 这一定义是良定义的, 因为根据定理 A.5, 如果 A 和 A' 是同一个线性算子在不同基下的矩阵, 由过渡矩阵 B 给出, 那么 $A' = B^{-1}AB$, 因此根据命题 A.12,

$$\det A' = \det(B^{-1}AB) = (\det B^{-1})(\det A)(\det B) = (\det B^{-1})(\det B)(\det A) = \det A.$$

因而行列式也是刻画线性算子的一个不变量.

定义 A.34 (线性算子的行列式) 设 V 是一个 n 维内积空间, f 是 V 上的一个线性算子, A 是 f 在 V 的一个基下的矩阵, 则 f 的行列式 $\det f$ 定义为 $\det A$. \square

线性算子的行列式表明了行列式的几何意义. 一个线性算子将一个平行体每条边 (也就是基向量) 映射到另一个平行体的每条边, 如此就将一个平行体映射到了另一个. 原来平行体的体积在这个线性算子作用后会发生改变, 这个变化的比率就是行列式.

具体来说, 我们选择 V 的标准正交基 e_1, \dots, e_n , 他们的坐标是 E_1, \dots, E_n . 考虑线性算子 f , 它对应的矩阵是 A . 现在考虑单位立方体 (自然也是平行体) $\Pi(e_1, \dots, e_n)$, 那么在映射作用下平行体的边就变成了 $\Pi(f(e_1), \dots, f(e_n))$. 因为是 $\{e_i\}$ 是标准正交基, 所以这些平行体实际上也可以写作 $\Pi(E_1, \dots, E_n)$ 和 $\Pi(AE_1, \dots, AE_n)$, 注意到, 后者实际上就是 $\Pi(A^1, \dots, A^n)$, 考虑他们的有向体积, 这正好是行列式的定义.

作为一个注记, 从线性算子的角度来看, 行列式的性质 (命题 A.12) 就是非常自然的结果了. 我们将矩阵对应的线性映射写出. 那么, 除了最后一条, 命题 A.12 的性质都可以逐一解释:

1. 恒等算子不改变有向体积, 所以行列式为 1;
2. 两个算子 f_A 和 f_B 的复合对有向体积的改变是累乘, 即先按比例 $\det f_A$ 变, 再按比例 $\det f_B$ 变, 因此矩阵乘积的行列式等于行列式的乘积;
3. λA 对应的算子就是 A 算子作用后再按照 λ 的比率等比例伸缩, 对于一个 n 维图形来说, 这样的变化对有向体积的改变是 λ^n ;

4. 考虑可逆线性算子 f , 那么, 先作用 f 再作用 f^{-1} , 体积不变, 所以他们对体积变化的比率乘起来 1, 即逆的行列式等于行列式的逆;
 5. 最后, 如果线性算子不可逆, 那么像是更低维的, 比如在三维空间中的二维平面, 那么显然有向体积就是 0 了, 所以不可逆的行列式为 0.
- 行列式还有其他的一些性质, 这里不再讨论.

§A.7 算子范数与谱理论

本节讨论如何给算子定义范数, 以及如何利用范数来研究算子的性质, 特别是特征值相关的性质. 本节需要一些点集拓扑的知识, 请参阅附录 B.1.

考虑一个 n 维内积空间 V 以及其上的一个线性算子 f . 给定一个 V 的标准正交基 e_1, \dots, e_n , 我们可以用矩阵 A 来表示 f . 注意到对任意一个 $x \in V$, 假设它的坐标是 $X = (x_1, \dots, x_n)^\top$, 那么

$$\begin{aligned} \|f(x)\|^2 &= \|AX\|^2 \\ &= \left(\sum_{j=1}^n \left(\sum_{i=1}^n a_{ij} \right) x_j \right)^2 \\ &\leq \max_k \sum_{i=1}^n |a_{ik}|^2 \cdot \sum_{j=1}^n x_j^2 \\ &= C \|X\| = C \|x\|. \end{aligned}$$

这里 $C = \max_k \sum_{i=1}^n |a_{ik}|^2$, 这说明, 对给定的算子 f , 我们可以找到一

一个常数 C 使得 $\|f(x)\| \leq C \|x\|$ 对任意的 $x \in V$ 都成立，因此，我们可以定义算子 f 的范数为最小的这样的常数 C .

定义 A.35 (算子范数) 设 V 是一个 n 维内积空间， f 是 V 上的一个线性算子，那么 f 的范数定义为

$$\|f\| = \inf \{C \geq 0 : \|f(x)\| \leq C \|x\|, \forall x \in V\}.$$

或者等价地，

$$\|f\| = \sup_{x \in V} \frac{\|f(x)\|}{\|x\|} = \sup_{\|x\|=1} \|f(x)\|. \quad \square$$

要保证定义出来的是范数，我们需要验证它满足非负性、齐次性和三角不等式。我们只证明三角不等式，其他两个类似。考虑两个算子 f, g ，

$$\begin{aligned} \|f+g\| &= \sup_{\|x\|=1} \|(f+g)(x)\| \\ &= \sup_{\|x\|=1} \|f(x) + g(x)\| \\ &\leq \sup_{\|x\|=1} (\|f(x)\| + \|g(x)\|) \\ &\leq \sup_{\|x\|=1} \|f(x)\| + \sup_{\|x\|=1} \|g(x)\| \\ &= \|f\| + \|g\|. \end{aligned}$$

这里我们用到了 V 中向量的三角不等式。

需要注意的是，我们这里定义的算子范数是非常受限的一个定义：我们只考虑了有限维内积空间，由内积诱导的范数定义的算子范数。一

般地，任意两个线性赋范空间之间的线性映射都可以定义范数：

定义 A.36 (线性映射的范数) 设 V, W 是两个线性赋范空间， f 是 V 到 W 的一个线性映射，如果

$$\|f\| = \sup_{\|x\|=1} \|f(x)\|$$

不是无穷大，那么称 f 是有界的， $\|f\|$ 是 f 的算子范数. □

如果线性映射的定义域是有限维空间，那么范数一定存在. 无限维内积空间中算子范数不一定存在，所以我们接下来的讨论都默认有限维线性空间.

注意到，算子范数自然地诱导了矩阵的范数：

定义 A.37 (矩阵范数) 设 A 是一个 $m \times n$ 的矩阵，那么 A 的范数定义为

$$\|A\| = \sup_{\|x\|=1} \|Ax\|. □$$

给定标准正交基底，利用矩阵 A 表示线性映射 f . 假设此时 x 的坐标是 X ，那么 $\|f(x)\| = \|AX\|$ ，由例 B.6、命题 B.10 和推论 B.2， $\|AX\|$ 是 X 的连续函数，因此在紧集 $\{X : \|X\| = 1\}$ 上取到最大值. 因此定义中的 sup 实际上是一个 max. 这一结论对任意范数定义的算子范数都成立（而不仅仅只是内积诱导的范数）. 我们后面会利用内积的特性显式给出取到最大值的向量.

算子范数一个显然的性质是：

命题 A.13 对有限维内积空间 V 中的线性算子 f ， $\|f(x)\| \leq \|f\| \|x\|$ 对任意 $x \in V$ 成立.

利用这一条，我们马上得到

命题 A.14 对有限维内积空间 V 中的线性算子 f, g , $\|f \circ g\| \leq \|f\| \|g\|$.

命题 A.15 对有限维内积空间 V 中的线性算子 f , $\|f^k\| \leq \|f\|^k$, 这里 f^k 表示 f 的 k 次复合.

以上性质都可以迁移到一般的线性映射以及矩阵上，这里不再赘述.

上面的性质并不依赖于“内积”的性质，主要是依赖“范数”的性质. 接下来，我们将深入利用内积的性质来研究算子的性质. 这里面的关键概念是谱，或者特征值.

定义 A.38 (特征值, 特征向量, 谱) 设 V 是一个 n 维内积空间, f 是 V 上的一个线性算子, $\lambda \in \mathbb{R}$. 如果存在一个非零向量 $x \in V$ 使得 $f(x) = \lambda x$, 那么称 λ 是 f 的一个特征值, x 是 λ 对应的特征向量. f 的所有特征值的集合称为 f 的谱, 记作 $\sigma(f)$. □

我们这里限制特征值为实数. 实际上，一般的情况下，特征值应该为复数，而内积应该定义为复共轭定义的内积. 但是，我们不关心会出现复数的情况，因此这里只考虑实数的情况.

我们下面的任务是给出特征值的刻画. 首先定义特征子空间：

定义 A.39 (特征子空间) 设 V 是一个 n 维内积空间, f 是 V 上的一个线性算子, $\lambda \in \mathbb{R}$. 定义

$$V^\lambda = \{x \in V : f(x) = \lambda x\}.$$

称 V^λ 是 f 的特征子空间. □

显然, V^λ 是 V 的一个线性子空间. 我们下面的任务是刻画特征子空间.

特征向量存在的意味着 V^λ 非零, 也就是 $\ker(f - \lambda \cdot \text{id}) \neq \{0\}$. 因此, 根据推论 A.1, $f - \lambda \cdot \text{id}$ 不是满射, 因此也不是双射, 从而 $\det(f - \lambda \cdot \text{id}) = 0$. 当选择一组基之后, $\det(f - \lambda \cdot \text{id})$ 就可以写成 $\det(A - \lambda I_n)$ 的形式, 其中 A 是 f 在这组基下的矩阵. 因此, 我们得到了一个关于 λ 的方程:

$$\det(A - \lambda I_n) = 0. \quad (\text{A.2})$$

将 λ 展开, 我们得到一个关于 λ 的 n 次多项式, 称为 f 的 **特征多项式**. 这个多项式的根就是 f 的特征值.

我们需要验证特征多项式对于算子来说是良定义的, 也就是不管怎么选取基, 得到的特征多项式都是一样的.

命题 A.16 设 A 是一个 n 阶方阵, P 是一个可逆方阵, 那么 A 与 $P^{-1}AP$ 有相同的特征多项式.

证明. $\det(P^{-1}AP - \lambda I_n) = \det(P^{-1}(A - \lambda I_n)P) = \det(P^{-1}) \det(A - \lambda I_n) \det(P) = \det(A - \lambda I_n).$ \square

有了特征值, 其对应的特征向量完全由 V^λ 刻画. 在特定的基之下, 我们可以用线性方程组的方式求出这个子空间一组基的坐标.

至此, 我们有了求解特征值和特征向量的方法.

注. 根据上面的讨论, n 维内积空间的谱至多有 n 个元素, 因而是离散的. 这一点在无穷维内积空间中也不成立, 因而无穷维算子的谱要复杂得多. 无穷维谱理论在泛函分析、量子力学等领域有着广泛的应用. 然而遗憾的是, 这些都超出了本书的范围.

利用定义，很容易写出矩阵特征值的性质：

命题 A.17 设 A 是一个 n 阶方阵， λ 是 A 的一个特征值， v 是 λ 对应的特征向量，那么

- 如果 A 可逆，那么 λ^{-1} 是 A^{-1} 的一个特征值， v 是 λ^{-1} 对应的特征向量；
- 任意多项式 p ， $p(\lambda)$ 是 $p(A)$ 的一个特征值， v 是 $p(A)$ 的一个特征向量。

接下来，我们考虑一类特殊的线性算子，被称为自伴算子。

定义 A.40 (自伴算子) 设 V 是一个 n 维内积空间， f 是 V 上的一个线性算子。如果对任意 $x, y \in V$ ，都有 $\langle f(x), y \rangle = \langle x, f(y) \rangle$ ，那么称 f 是自伴算子。 \square

我们指出，自伴算子的矩阵在标准正交基下是实对称矩阵，这可以从内积的矩阵表示看出。假设 A 是 f 在标准正交基下的矩阵，对任意 $x, y \in V$ ，他们的坐标是 X, Y ，那么 $\langle f(x), y \rangle = \langle x, f(y) \rangle$ 等价于 $(AX)^T Y = X^T(AY)$ ，也就是 $X^T A^T Y = X^T AY$ 。将 X, Y 取遍所有可能的向量，我们得到 $A^T = A$ 。

自伴算子的谱可以有一个非常好的刻画：

定理 A.17 设 V 是一个 n 维内积空间， f 是 V 上的一个自伴算子， $\lambda \in \mathbb{R}$ 。那么 f 的特征多项式所有根都是实根，因此 $\sigma(f)$ 是一个非空实数集。此外，

$$V = \bigoplus_{\lambda \in \sigma(f)} V^\lambda.$$

因此，存在一组标准正交基，使得 f 在这组基下的矩阵是对角矩阵，这组基就是 f 的特征向量。此外，对角线上的元素恰好是 f 的特征值，对于 $\lambda \in \sigma(f)$ ，它在对角线上出现的次数就是 $\dim V^\lambda$ 。

这一证明非常类似定理 A.11 和定理 A.7 的证明，我们这里就不再赘述。

这一定理在矩阵上表述为：

推论 A.5 对任意实对称矩阵 A ，都存在一个正交矩阵 P 使得 $P^{-1}AP = P^TAP$ 是对角矩阵，这个对角矩阵的对角线上的元素就是 A 的特征值。

因此实对称矩阵可以通过一个正交矩阵相似并合同到对角矩阵。

这一结果在几何上有明确的意义。设 V 是一个 n 维内积空间， f 是 V 上的一个自伴算子， v_1, \dots, v_n 是 f 对应的标准正交的特征向量基，那么 f 的作用就是将这些坐标轴拉伸或者压缩，或者反转。例如，在 \mathbb{R}^2 上，假设 $f(v_1) = -1.5v_1$, $f(v_2) = 0.5v_2$ ，这一算子的效果可以用图 A.1 表示出来。

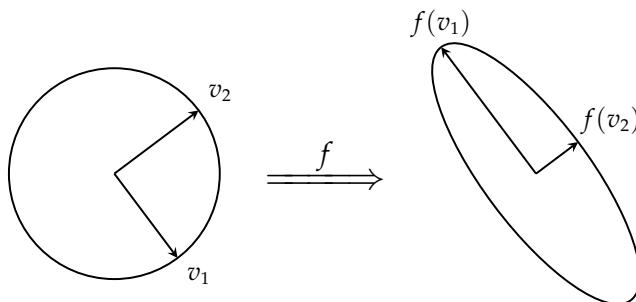


图 A.1: 自伴算子的作用

接下来，我们考虑自伴算子的范数与谱的关系。设 f 是 V 上的一个自伴算子，那么根据定理 A.17，存在一组标准正交基 e_1, \dots, e_n ，使得

$f(e_i) = \lambda_i e_i$, λ_i 是特征值 (可重复). 此时 f 对应的矩阵记为 A , 它是一个对角矩阵. 考虑任何一个向量 $x \in V$, 它的坐标是 X , 那么

$$\|f(x)\|^2 = \|AX\|^2 = \sum_{i=1}^n \lambda_i^2 x_i^2.$$

假设 $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, 那么

$$|\lambda_n|^2 \sum_{i=1}^n x_i^2 \leq \sum_{i=1}^n \lambda_i^2 x_i^2 \leq |\lambda_1|^2 \sum_{i=1}^n x_i^2.$$

因此

$$|\lambda_n| \|x\| \leq \|f(x)\| \leq |\lambda_1| \|x\|.$$

从左到右, 等号分别在 $x = e_n$ 和 $x = e_1$ 时取到. 因此, 我们得到了

定理 A.18 设 V 是一个 n 维内积空间, f 是 V 上的一个自伴算子, 那么

$$\|f\| = \max_{\lambda \in \sigma(f)} |\lambda|.$$

如果最大值在特征值 λ_0 取到, 其对应单位特征向量是 e_0 , 那么 $\|f(e_0)\| = \|f\| = |\lambda_0|$.

最大的特征值模称为算子的谱半径. 从几何意义来说, 谱半径就是算子对应的线性变换对应的线性变换对向量的最大拉伸率. 一个更直观的理解方式是, 将这些基向量画一个球包起来, 算子会将这个球映射到一个椭球, 这个椭球的最长轴就是谱半径. 例如, 对于图 A.1 中的算子, 谱半径就是 1.5.

附录 B 微分子基础

本书中涉及的积分学很少，并且集中在概率论部分，所以在本附录中我们只讨论微分子，积分学的内容在附录 C.3 中简单介绍。尽管我们的视角非常一般且抽象，但我们主要讨论的是 Euclid 空间 \mathbb{R}^n 相关的微分子。

§B.1 点集拓扑

本部分讨论极限、连续、紧致等概念，这些概念是微分子的基础。

§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 $x, y \in X$, $d(x, y) \geq 0$, $d(x, y) = 0$ 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 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：对 $x_1 \neq x_2$, $d(x_1, x_2) = 1$; 对 $x_1 = x_2$, $d(x_1, x_2) = 0$.
- 绝对值度量： $d(x_1, x_2) = |x_1 - x_2|$.

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. 容易验证这也是一个度量. \square

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

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. \square

容易验证, 例 B.1 中的度量都自然地导出了一个范数, 即 $\|x\| = d(x, 0)$. 我们可以沿袭度量的名字称呼这些范数, 例如 L^p 范数就是 L^p 度量所诱导的范数. 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 空间就是由 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:

定理 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: $d(\lambda x, \lambda y) = |\lambda|d(x, y).$

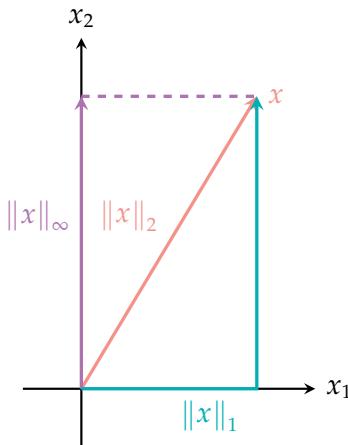
尽管都是 \mathbb{R}^n , 但是不同 p 对应的 L^p 范数是不一样的. 他们之间有如下的关系:

命题 B.1 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 $n = 2$ 以及 $p = 1, 2, \infty$ 的极端情形. 如下图所示, 我们要从原点到点 x . 青色的是 $\|x\|_1 = |x_1| + |x_2|$, 相当于沿着坐标轴走; 而粉色的是 $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$, 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，我们在后面会看到，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.2 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 $(a, b) = \{x \in \mathbb{R} : a < x < b\}$. \mathbb{R} 中的开集定义其实很简单, 就是若干开区间的并集. 在更一般的拓扑空间中, 开集的定义也是类似的. 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. 这样的定义有一般性:

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

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

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.3 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, 此时可以构造性地给出所有开集.

证明. \implies : 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

本书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，这一定义关心集合而不是具体的点。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。我们要证明，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，在拓扑意义上， \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，收敛性和连续性在 L^p 范数下都是完全一样的。

下面我们来证明这一点。我们用开集的点定义。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。 \square

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 $X_1 \times X_2$ 上的开集有两种自然的方式给出：

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。

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

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, 这两种方式给出了 $X_1 \times X_2$ 上相同的拓扑.

因此, 以后讨论 “拓扑空间 $X \times Y$ ” 的地方, 不管明里暗里, 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:

命题 B.4 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.

命题 B.4 其实就是一般拓扑空间中开集要满足的三条公理. 我们之所以将它写为命题, 是因为我们的开集定义基于度量空间, 而非一般的拓扑空间.

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. \square

闭集的定义是开集的对偶, 所以有如下性质:

命题 B.5 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.

开集和闭集的定义是对偶的, 但是性质却完全不同. 开集似乎可以简单理解为开区间的推广, 即把开区间拼起来, 它的构造是“把东西放进来”. 闭集是把若干开区间挖出来得到的集合, 它的构造方式是“把东西拿出去”, 这样的构造对我们来说是不够直观的. 我们可以构造非常奇怪的闭集, 例如 Cantor 集就是例子.

§B.1.3 紧致性，收敛性，完备性

接下来我们讨论一个更微妙的概念，紧致性或者紧集。紧致性与极限、收敛、连续等概念有着密切的联系，然而如何恰当的定义紧致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。我们的逻辑推理只能处理有限的东西，所以紧致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。□

第一次看到这样的定义大概率会不知所云。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.6 (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 \mathbb{R} 中紧集的存在性。

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 $n > N$ ，

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）。

$d(x_n, x) < \epsilon$, 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. \square

这一定义描绘了一幅图像: 一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.7 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 $n > N$, $x_n \in U$.

证明. \implies : 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 $x_n \in B(x, r) \subseteq U$.

\impliedby : 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

在一般的拓扑空间中, 甚至都没有度量的概念, 然而, 开集定义收敛依然可以. 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, $\{x_k\}_{k=1}^\infty$ 收敛到 x , 当且仅当对任意 $i = 1, \dots, n$, $\{x_k^i\}_{k=1}^\infty$ 收敛到 x^i , 其中 $x_k = (x_k^1, \dots, x_k^n)$, $x = (x^1, \dots, x^n)$.
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). \square

度量空间中紧集可以完全由收敛性来刻画:

定理 B.2 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.

定理 B.2 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. \square

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.

Cauchy 列与收敛列的关系如下. 首先, 收敛的点列是 Cauchy 列:

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, 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. 在 \mathbb{R} 上, 就如同有限覆盖原理, 这件事的成立性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. \square

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 \mathbb{R} 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. \square

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:

定理 B.3 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.9 设 ℓ^2 空间的标准正交向量组是 $\{e_i\}_{i=1}^\infty$, e_i 是第 i 个分量为 1, 其他分量为 0 的向量. 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.8 的逆否命题, $\{e_i\}$ 没有任何收敛子列, 因而根据定理 B.2, 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 δ - ϵ 语言的点定义, 然后再给出更像拓扑的定义.

δ - ϵ 语言的定义是从映射的极限这一概念出发的:

定义 B.10 (映射的极限) 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f : X \rightarrow Y$ 是一个映射, $x_0 \in X$, $y \in Y$, 如果对任意 $\epsilon > 0$, 存在 $\delta > 0$, 使得

对任意 $x \in X$, 如果 $0 < d_X(x, x_0) < \delta$, 则 $d_Y(f(x), y) < \epsilon$, 则称 y 是 f 在 x_0 处的极限, 记作 $\lim_{x \rightarrow x_0} f(x) = y$ 或 $f(x) \rightarrow y$, $x \rightarrow x_0$. \square

注意, 定义中我们划定了 x 的范围, 即 $x \neq x_0$. 此时极限的概念由去心邻域 $B(x_0, \delta) \setminus \{x_0\}$ 给出, 这样做允许极限并不等于 $f(x_0)$ 本身.

注. 映射的极限还可以定义自变量趋于无穷、单侧极限以及其他情况, 我们后面会使用这些概念, 他们的直观含义都是明确的, 这里我们不再给出正式定义, 我们只给出他们的记号:

- 趋于无穷: $\lim_{x \rightarrow \infty} f(x) = y$ 或 $f(x) \rightarrow y$, $x \rightarrow \infty$;
- 如果定义域是 \mathbb{R} , 还可以定义趋于正、负无穷: $\lim_{x \rightarrow +\infty} f(x) = y$ 或 $f(x) \rightarrow y$, $x \rightarrow +\infty$, $\lim_{x \rightarrow -\infty} f(x) = y$ 或 $f(x) \rightarrow y$, $x \rightarrow -\infty$;
- 单调递增趋于: $x \uparrow x_0$, 单调递减趋于: $x \downarrow x_0$. 这些记号既可以出现在自变量中, 也可以出现在函数值中, 例如我们可以写 $n/(n+1) \uparrow 1$, $n \rightarrow \infty$.
- 如果定义域是 \mathbb{R} , 还可以定义单侧极限, 从负向趋于某点 (左极限): $\lim_{x \uparrow x_0} f(x) = y$ 或 $f(x) \rightarrow y$, $x \uparrow x_0$, 以及从正向趋于某点 (右极限): $\lim_{x \downarrow x_0} f(x) = y$ 或 $f(x) \rightarrow y$, $x \downarrow x_0$.

由此, 我们可以定义连续映射:

定义 B.11 (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 $x \in X$, 如果 $\lim_{x' \rightarrow x} f(x') = f(x)$, 则称 f 在 x 处连续, 如果 f 在 X 的每一点都连续, 则称 f 是连续映射. \square

直观上说, 连续映射是指, x' 和 x 足够接近的时候 $f(x')$ 和 $f(x)$ 也足够接近. 不过, 数学定义其实是反过来的: 想让 $f(x')$ 和 $f(x)$ 足够接近, 我们只需要让 x' 和 x 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.12 (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极限和连续性了：

定理 B.4 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

1. $\lim_{x \rightarrow x_0} f(x) = y$ 当且仅当对任意包含 y 的开集 $U \subseteq Y$ ，存在包含 x_0 的开集 $V \subseteq X$ ，使得 $f(V \setminus \{x_0\}) \subseteq U$ ；
2. 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.3，我们这里就不给出了。注意，极限的开集定义所用的集合 $V \setminus \{x_0\}$ 也是一个开集，它是 x_0 的去心邻域，所以这一定义确实是纯拓扑的。

连续映射的定义也可以完全由拓扑给出：

定理 B.5 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.

利用定理 B.4、命题 B.5 以及开集的定义, 很容易证明这一命题, 这里不再赘述.

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映射的例子, 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 $(x_1, y_1) \in X \times X$. 我们利用定理 B.4 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,

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

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.

一个直接的推论是, 范数 $\|\cdot\|$ 也是连续函数. □

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, 极限也可以用收敛性刻画:

定理 B.6 设 (X, d_X) 和 (Y, d_Y) 是两个度量空间, $f : X \rightarrow Y$ 是一个映射, $x_0 \in X$, $y \in Y$, 则下列表述等价:

1. $\lim_{x \rightarrow x_0} f(x) = y$.
2. 对任意 $\{x_n\}_{n=1}^{\infty}$ 满足 $x_0 \notin \{x_n\}$, 如果 $x_n \rightarrow x_0$, 则 $f(x_n) \rightarrow y$.

利用这一条, 很快就可以得到连续的序列版本:

推论 B.1 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 在 $x \in X$ 连续.
2. 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.10 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 $x \in X$ 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.2 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.

根据推论 B.1, 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} . 将乘法 $\times : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 和除法 $\div : \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \rightarrow \mathbb{R}$ 看成向量空间 \mathbb{R} 上的数乘运算, 于是他们也都是连续映射⁴.

最后, 连续意味着有界:

命题 B.11 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.12 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.

其他性质将在下一节给出.

§B.1.5 与实数序有关的性质

本节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实值函数或简单称为函数. \mathbb{R} 与 \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可以被称为序的性质.

下面, 我们列出其中两个与实数公理等价的序性质. 这些性质需要用到单调性、界和确界的概念, 这些概念将会频繁出现在我们的讨论中, 所以这里单独给出:

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将类似的结论重复讲多次.

定义 B.13 (单调性) 设 $\{x_n\}_{n=1}^{\infty}$ 是一个实数列.

- 如果对任意 $n \in \mathbb{N}$, $x_n \leq x_{n+1}$, 则称 $\{x_n\}_{n=1}^{\infty}$ 是一个**单调递增的实数列**.
- 如果对任意 $n \in \mathbb{N}$, $x_n \geq x_{n+1}$, 则称 $\{x_n\}_{n=1}^{\infty}$ 是一个**单调递减的实数列**.
- 如果 $\{x_n\}_{n=1}^{\infty}$ 是单调递增的或单调递减的, 则称 $\{x_n\}_{n=1}^{\infty}$ 是一个**单调的实数列**. \square

定义 B.14 (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**. \square

上确界这个概念就是在说“最小可能的上界”, 下确界也有类似的解读.

现在我们可以阐述这两个实数序的性质了. 第一个是说单调有界的序列一定收敛.

命题 B.13 (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.

接下来一个是说有上 (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.14 (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.15 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 $\lim_{n \rightarrow \infty} a_n = M$. \square

对于实值函数来说, 我们还需要比较在极限情况下两个函数的渐近大小, 这就是 o 和 \mathcal{O} 符号. 我们先给出这一概念在序列上的定义:

定义 B.15 (阶, 无穷小, 等价) 设 $\{x_n\}_{n=1}^{\infty}$ 和 $\{y_n\}_{n=1}^{\infty}$ 是两个序列.

- 如果 $\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = 0$, 则称 $\{x_n\}_{n=1}^{\infty}$ 是 $\{y_n\}_{n=1}^{\infty}$ 的高阶无穷小, 记作 $x_n = o(y_n)$.
- 如果存在一个正常数 C 使得除去有限个 n 都有 $|x_n| \leq C|y_n|$, 则称 $\{x_n\}_{n=1}^{\infty}$ 的阶不高于 $\{y_n\}_{n=1}^{\infty}$, 记作 $x_n = \mathcal{O}(y_n)$.
- 如果 $x_n = \mathcal{O}(y_n)$ 且 $y_n = \mathcal{O}(x_n)$, 那么称 $\{x_n\}_{n=1}^{\infty}$ 和 $\{y_n\}_{n=1}^{\infty}$ 是同阶的.

- 如果进一步 $\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = 1$, 则称 $\{x_n\}_{n=1}^{\infty}$ 和 $\{y_n\}_{n=1}^{\infty}$ 是等价的, 记作 $x_n \sim y_n$. □

上述定义可以非常自然迁移到函数上, 我们不再赘述. 下面是一些例子:

例 B.7 • $n \rightarrow \infty$ 时, $n^2 = o(2^n)$, $n^{1/n} \sim 1$, $n^2 \sim n^2 + \log n$.

- $x \rightarrow 0$ 时, $\sin x \sim x$, $1/x = o(1/x^2)$.

- $n \rightarrow \infty$ 时, $\sum_{k=1}^n \frac{1}{n} \sim \ln n$. □

最后, 我们回到连续的整体性质上来. 首先是 Weierstrass 最值定理:

定理 B.7 (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.8 (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 $f(x) = y$.

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

接下来, 我们进入微分学的部分, 同样, 我们先从最基本的一元函数的情况 (即 $\mathbb{R} \rightarrow \mathbb{R}$ 的函数) 入手.

§B.2.1 导数与微分的定义

从近似的角度来说，微分或者导数的概念，本身在描述在某个点函数的线性近似，因此微分和导数本身也是一个（线性）映射。在一元函数中，我们或许无法看出来这一点，但是在更加一般的微分学中，这样的观点非常重要。因此，即便在一元部分，我们也尝试将这样的观点引入。

考虑一个函数 $f : \mathbb{R} \rightarrow \mathbb{R}$ ，和点 $x_0 \in \mathbb{R}$ ，我们希望找到一个线性映射 $df_{x_0} : \mathbb{R} \rightarrow \mathbb{R}$ ，使得 $f(x)$ 在 x_0 附近的行为很接近这线性映射，即

$$f(x) \approx f(x_0) + df_{x_0}(x - x_0).$$

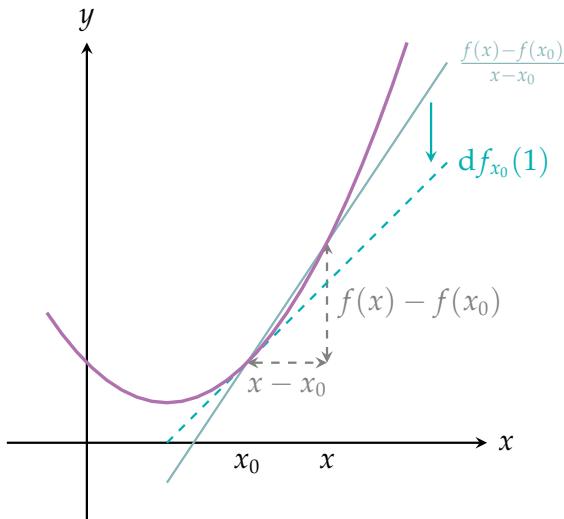
更精确来说，我们希望两边的误差是一个关于 $x - x_0$ 的高阶无穷小：

$$f(x) = f(x_0) + df_{x_0}(x - x_0) + o(x - x_0).$$

这一记号的含义可以通过一些变换看出来：

$$df_{x_0}(1) = \frac{f(x) - f(x_0)}{x - x_0} + o(1). \quad (\text{B.1})$$

注意 $df_{x_0}(x) = kx$ ，所以左边就是 k ，而右边是割线的斜率。式 (B.1) 的含义其实也就是说 k 就是割线斜率的极限，直观上这就是切线的斜率，这就是导数的几何含义。这一过程可以见下图：



我们将这些讨论整理为如下定义：

定义 B.16 (微分, 导数) 设 $f : \mathbb{R} \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}$, 如果存在一个线性映射 $df_{x_0} : \mathbb{R} \rightarrow \mathbb{R}$, 使得

$$f(x) = f(x_0) + df_{x_0}(x - x_0) + o(x - x_0),$$

则称 f 在 x_0 处可微或者可导, df_{x_0} 是 f 在 x_0 处的微分. 微分具有形式 $df_{x_0}(x) = kx$, 其中 k 称为 f 在 x_0 处的导数, 记作 $f'(x_0)$.

如果 f 在 \mathbb{R} 的每一点都可微, 则称 f 是可微的或者可导的, df 是 f 的微分, f' 是 f 的导(函)数, 也记作 $\frac{df}{dx}$ 或 \dot{f} . □

关于导数的符号有一些注. 最能体现几何意义的是 $\frac{df}{dx}$, 它是由 Leibniz 发明的. 符号 d 的意思就是“微”, 可以理解为无穷小的变化量, 所以导数就是自变量和函数值无穷小变化量的比值. 另一方面, 这个符

号也可以理解为“切”，表示切向量的意思，例如 dx 就是沿着 x 轴的任意切向量（实际上就是正方向或者负方向），而 dy 就是相应地沿着 y 轴的切向量。从这个角度来说， $\frac{df}{dx}$ 就是 x 轴切向量到 y 轴切向量的一个线性映射。因此，微分其实就是所谓的切映射，即切向量到切向量的映射。这一视角在更抽象的微分学中是更本质的。

导数的定义也可以用更常见的形式给出：

命题 B.16 设 $f : \mathbb{R} \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}$, 那么 f 在 x_0 处可微当且仅当如下极限存在：

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

这个极限就是 f 在 x_0 处的导数。因而，微分或者说导数是唯一的。

下面我们不加证明地列举导数的一些性质，这些性质自然也导出了微分的性质。

命题 B.17 设 $f, g : \mathbb{R} \rightarrow \mathbb{R}$ 在 x_0 处可微，则

- f 在 x_0 处连续；
- $(f + g)'(x_0) = f'(x_0) + g'(x_0)$ ；
- $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$ ；
- 如果 $g(x_0) \neq 0$, 则

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2},$$

- 链式法则：如果 f 在 x_0 处可微， g 在 $f(x_0)$ 处可微，则 $g \circ f$ 在 x_0 处可微，且 $(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$ ；

- 如果 f 存在反函数 f^{-1} , 则 f^{-1} 在 $f(x_0)$ 处可微, 且 $(f^{-1})'(f(x_0)) = \frac{1}{f'(x_0)}$.

在 Leibniz 记号下, 如果 $z = z(y)$, $y = y(x)$, 那么链式法则可以写作

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

反函数的导数则可以写作

$$\frac{dy}{dx} = \left(\frac{dx}{dy} \right)^{-1}.$$

我们再次看到这种记号的天才之处, 它将复杂的计算简化为了一种直观的形式.

我们指出, 链式法则和反函数求导法则在微分下有更加清晰的含义:

命题 B.18 设 $f : \mathbb{R} \rightarrow \mathbb{R}$ 在 x_0 处可微, $g : \mathbb{R} \rightarrow \mathbb{R}$ 在 $f(x_0)$ 处可微, 则

- $dx = \text{id}$;
- $d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$.
- 如果 f 存在反函数 f^{-1} , 则 $d(f^{-1})_{f(x_0)} = (df_{x_0})^{-1}$.

命题 B.18 提供了这样一种视角: 微分号 d 相当于把 $\mathbb{R} \rightarrow \mathbb{R}$ 的函数变成了另外一个 $\mathbb{R} \rightarrow \mathbb{R}$ 的函数 (即切映射), 同时保持函数复合运算的单位元 (id)、复合和逆元关系. 利用这一性质, 我们可以用更加代数的方法研究微分 (切函子), 但这超出了本书的范围, 我们就不再详细讨论了.

最后，我们讨论高阶导数的概念。注意，我们只关注导数而不关注微分，这是因为由于高阶微分是一个相当抽象的概念，所以就不深入讨论了。

定义 B.17 (高阶导数) 设 $f : \mathbb{R} \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}$, 如果 f 在 x_0 处可微，则 f 在 x_0 处的导数 $f'(x_0)$ 是一个实数。如果 f' 在 x_0 处可微，则称 f 在 x_0 处二阶可微，此时 $f''(x_0)$ 是 f' 在 x_0 处的导数，称为 f 在 x_0 处的二阶导数。一般地，如果 $f^{(n-1)}$ 在 x_0 处可微，则称 f 在 x_0 处 n 阶可微，此时 $f^{(n)}(x_0)$ 是 $f^{(n-1)}$ 在 x_0 处的导数，称为 f 在 x_0 处的 n 阶导数。□

在 Leibniz 的记号下， n 阶导数可以写作

$$\frac{d^n y}{dx^n} = \underbrace{\frac{d}{dx} \cdots \frac{d}{dx}}_{n\text{个}} y.$$

从这里我们可以看出， d/dx 这个记号又仿佛是一个算子，它作用在函数上，得到一个新的函数。这个视角在谱理论中得到发扬，继而成为了量子力学的数学基础，用它可以证明矩阵力学和波动力学的等价性。当然，这也不在本书的讨论范围之内了。

我们将在集合 X 上 n 次连续可微的函数（即 n 阶导数是连续函数）的集合记作 $C^n(X)$ ，任意次连续可微的函数的集合记作 $C^\infty(X)$ 。在后面更一般的微分学中， X 可以不是 \mathbb{R} 的子集，但我们依然沿用此记号，如果我们讨论的映射取值不在 \mathbb{R} 上，而是在抽象的集合 Y 上，我们将 $C^n(X, Y)$ 记作从 X 到 Y 的 n 次连续可微映射的集合， $C^\infty(X, Y)$ 记作任意次连续可微的从 X 到 Y 的映射的集合，这些概念的定义将在后面给出。

§B.2.2 微分学基本定理

微分学几乎都与极值联系在一起，刻画这些关系的定理就是微分学的基本定理。我们依然只罗列定理，不给出证明。首先我们给出极值的定义。

定义 B.18 (极大值, 严格极大值, 极小值, 严格极小值) 设 $f : X \rightarrow \mathbb{R}$, $x_0 \in X$, 如果存在包含 x_0 的开集 U , 使得对任意 $x \in U$, 有 $f(x) \leq f(x_0)$, 则称 $f(x_0)$ 是 f 在 x_0 处的一个极大值, x_0 是 f 的一个极大值点。如果 $f(x) = f(x_0)$ 只在 x_0 处成立, 则称 $f(x_0)$ 是 f 在 x_0 处的一个严格极大值, x_0 是 f 的一个严格极大值点。

如果不等式反向, 则称 $f(x_0)$ 是 f 在 x_0 处的一个极小值, x_0 是 f 的一个极小值点。如果 $f(x) = f(x_0)$ 只在 x_0 处成立, 则称 $f(x_0)$ 是 f 在 x_0 处的一个严格极小值, x_0 是 f 的一个严格极小值点。

如果 $f(x_0)$ 是 f 在 x_0 处的一个极大(小)值, 则称 $f(x_0)$ 是 f 在 x_0 处的一个极值, x_0 是 f 的一个极值点。 \square

首先是 Fermat 引理, 他其实也是极值的一阶必要条件:

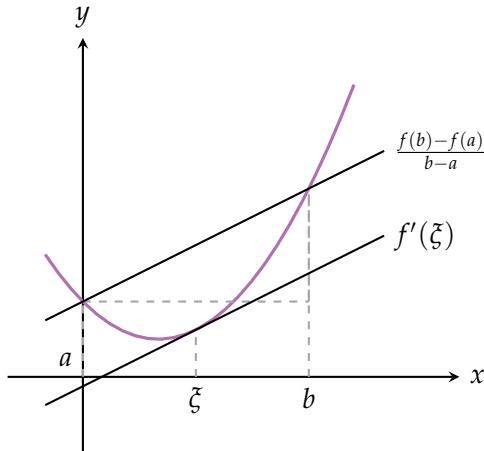
引理 B.1 (Fermat 引理) 设 $f : X \rightarrow \mathbb{R}$, $x_0 \in X$ 是 f 的一个极值点, 且 f 在 x_0 处可微, 则 $f'(x_0) = 0$.

接下来是一系列中值定理, 我们这里只给出 Lagrange 中值定理:

定理 B.9 (Lagrange 中值定理) 设 $f : [a, b] \rightarrow \mathbb{R}$ 是一个连续函数, 且在 (a, b) 内可微, 则存在 $\xi \in (a, b)$, 使得

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

这一定理给出了割线斜率和切线斜率的关系, 可以用下图来理解:



我们将在后面指出，这一定理只适用于实值函数，假如想对向量值函数使用，需要对其进行适当的修改：

定理 B.10 (Lagrange 有限增量定理) 设 $f : [a, b] \rightarrow \mathbb{R}$ 是一个连续函数，且在 (a, b) 内可微，则

$$|f(b) - f(a)| \leq |b - a| \sup_{\xi \in (a, b)} |f'(\xi)|.$$

接下来我们讨论高阶导数与极值的关系。这样的关系是由 Taylor 公式给出的。我们说过，微分是用线性函数去近似函数的过程，而 Taylor 公式则是用多项式去近似函数的过程。考虑函数 $f : \mathbb{R} \rightarrow \mathbb{R}$ ，如果 f 在 x_0 处 n 次可微，我们尝试用一个 n 次多项式去近似 f ，即

$$f(x) = a_0 + a_1(x - x_0) + \cdots + a_n(x - x_0)^n + o((x - x_0)^n).$$

容易求出， $a_0 = f(x_0)$ ， $a_1 = f'(x_0)$ ， $a_2 = \frac{f''(x_0)}{2}$ ， $a_3 = \frac{f'''(x_0)}{6}$ ， \dots ，

$a_k = \frac{f^{(k)}(x_0)}{k!}$, 因此我们得到了 Taylor 公式:

定理 B.11 设 $f : \mathbb{R} \rightarrow \mathbb{R}$ 在 x_0 处 n 次可微, 则

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + o((x - x_0)^n).$$

我们将这个 n 次多项式称为 **Taylor 展开**.

利用 Taylor 展开, 我们可以得到通过高阶导数判定极值的充分条件:

定理 B.12 设 $f : (a, b) \rightarrow \mathbb{R}$ 在 x_0 处 n 次可微, 且 $f'(x_0) = f''(x_0) = \dots = f^{(n-1)}(x_0) = 0$, $f^{(n)}(x_0) \neq 0$, 则

- 如果 n 为奇数, f 在 x_0 处没有极值;
- 如果 n 为偶数, f 在 x_0 处有极值, 且当 $f^{(n)}(x_0) > 0$ 时, f 在 x_0 处有严格极小值, 当 $f^{(n)}(x_0) < 0$ 时, f 在 x_0 处有严格极大值.

§B.3 多元函数的微分学

这一部分讨论 $\mathbb{R}^n \rightarrow \mathbb{R}^m$ 的微分学, 当 $m = 1$, 我们称之为实值函数; 对一般的 m , 我们称之为向量值函数. 这一部分需要很多线性代数的知识, 请参阅附录 A.

§B.3.1 微分、偏导数与导数的定义

沿着一元函数的思路, 我们希望找到一个线性映射 $df_x : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 使得 f 在 x 附近的行为很接近这个线性映射, 而这件事情本身就可以

作为微分的定义：

定义 B.19 (微分) 设 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $x \in \mathbb{R}^n$, 如果存在一个线性映射 $df_x : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 使得

$$f(x+h) = f(x) + df_x h + o(h),$$

则称 f 在 x_0 处可微或者可导, df_{x_0} 是 f 在 x_0 处的微分. 这里 $o(h)$ 理解为一个向量值函数 $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 它满足 $\lim_{h \rightarrow 0} \|\alpha(h)\| / \|h\| = 0$ ⁵.

如果 f 在 \mathbb{R}^n 的每一点都可微, 则称 f 是可微的或者可导的, df 是 f 的微分. \square

现在我们来解释这一定义的含义. 微分的定义依然是将一个关于 x 的函数转变到一个关于 h 的线性映射, 这个线性映射表明了函数在 x 处的线性近似, 而这个线性近似的误差是一个关于 h 的高阶无穷小. 从这个角度来说, 微分的定义和一元函数的情况是一样的, 只不过这里的误差是一个向量值函数而已.

h 被称为切向量 (回忆一维的情况), 所有允许 h 的集合记为 $T\mathbb{R}_x^n$, 称为在 x 点处的切空间. \mathbb{R}^n 上定义的函数, 切向量 h 自然可以取遍所有 \mathbb{R}^n 的向量, 所以其实 $T\mathbb{R}_x^n = \mathbb{R}^n$. 然而在第七章中, 我们会看到, 当定义域不是整个 \mathbb{R}^n , 而只是某个子集的时候, 切空间的定义就变得不平凡了. 从“切”的视角来看, 微分其实是一个切映射, 即从切向量到切向量的映射, 这可以用下图来表示:

⁵在例 B.2 中我们提到过, L^p 范数下的 \mathbb{R}^n 的拓扑都是一样的. 但是, 为了利用内积的性质, 之后我们写出符号 $\|\cdot\|$ 的时候, 都指 L^2 范数.

$$\begin{array}{ccc} f : & \mathbb{R}^n & \longrightarrow \mathbb{R}^m \\ \downarrow d & & \\ df_x : & T\mathbb{R}_x^n & \longrightarrow T\mathbb{R}_{f(x)}^m \end{array}$$

接下来的问题就是，如何表示线性映射 df_x ？我们先从实值函数开始。考虑 \mathbb{R}^n 的标准正交基 $\{e_i\}_{i=1}^n$ ，它也是切空间 $T\mathbb{R}_x^n$ 的标准正交基，根据 Riesz 表示定理（定理 A.12），存在一个向量 g ，使得

$$df_x h = \langle g, h \rangle.$$

这个向量 g 被称为 f 在 x 处的梯度，记作 $\text{grad } f(x)$ 。

我们需要进一步将梯度的坐标 $(g_1, \dots, g_n)^\top$ 求出来。考虑一个具体的分量 e_i ，根据定义，

$$f(x + te_i) = f(x) + t df_x(te_i) + o(te_i) = f(x) + g_i t + o(t).$$

因此，

$$g_i = \lim_{t \rightarrow 0} \frac{f(x + te_i) - f(x)}{t}.$$

我们给这样的导数一个名字，称为 f 在 x 对 x_i 的偏导数，记作 $\frac{\partial f}{\partial x_i}(x)$ 或 $\partial_i f(x)$ ，于是我们得到了梯度的坐标：

$$\left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^\top.$$

当然，我们不一定要沿着 e_i 去算导数，我们可以沿着任意单位向量

u 去算, 于是我们得到了 f 在 x 处沿着 u 的方向导数:

$$\frac{\partial f}{\partial u}(x) = \lim_{t \rightarrow 0} \frac{f(x + tu) - f(x)}{t}, \quad \|u\| = 1.$$

有了梯度, 我们可以很快算出任意方向导数:

命题 B.19 设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 在 x 处可微, u 是单位向量, 则

$$\frac{\partial f}{\partial u}(x) = \langle \text{grad } f(x), u \rangle.$$

在微积分中, 我们总是假设在标准正交基下进行计算, 在这种情况下, 我们有更简便的表示方式. 形式上, 记

$$\nabla = e_1 \frac{\partial}{\partial x_1} + \cdots + e_n \frac{\partial}{\partial x_n},$$

则

$$\text{grad } f(x) = \nabla f(x).$$

符号 ∇ 被称为 **nabla 算子**, 它就是标准正交基下梯度的具体表示. 通常, 我们会更简单地将 ∇ 记为 $(\partial_1, \dots, \partial_n)^\top$.

接下来我们讨论向量值函数微分的表示问题. 选取 \mathbb{R}^m (也就是 $T\mathbb{R}_x^n$) 的标准正交基 e_i , 选取 \mathbb{R}^m (也就是 $T\mathbb{R}_{f(x)}^m$) 的标准正交基 e_i , 则根据附录 A.3 的讨论, 我们可以用一个 $m \times n$ 的矩阵来表示 $d f_x$, 这个矩阵被称为 f 在 x 处的 **Jacobi 矩阵**, 记作 $J_f(x)$.

下面我们计算 $J_f(x)$ 的具体表示. 假设 $f(x)$ 的坐标是 $(f_1(x), \dots, f_m(x))^\top$,

考虑 $h \in T\mathbb{R}_x^n$, 它的坐标是 $(h_1, \dots, h_n)^\top$, $df_x h$ 的坐标应该是

$$\begin{pmatrix} df_{1,x} h \\ \vdots \\ df_{m,x} h \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \partial_i f_1(x) h_i \\ \vdots \\ \sum_{i=1}^n \partial_i f_m(x) h_i \end{pmatrix} = \begin{pmatrix} \partial_1 f_1(x) & \dots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(x) & \dots & \partial_n f_m(x) \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}.$$

因此, 我们得到了 Jacobi 矩阵:

$$J_f(x) = (\partial_j f_i) = \begin{pmatrix} \partial_1 f_1(x) & \dots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(x) & \dots & \partial_n f_m(x) \end{pmatrix}.$$

在 $m = n$ 的特殊情况下, $J_f(x)$ 的行列式被称为 f 在 x 处的 **Jacobi 行列式**, 记作

$$\frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)}(x).$$

在例 B.15 中我们会看到, Jacobi 行列式表明了坐标变换时相应体积变化的比率. 这一事实使得它在积分学的变量替换中有着核心作用.

总结来说, 实值函数的微分可以用行向量 $(\partial_1 f_1(x), \dots, \partial_n f_1(x))$ 和切向量相乘表示, 而向量值函数的微分可以用 Jacobi 矩阵 $J_f(x)$ 和切向量相乘来表示, 我们将这些符号统称 f 在 x 处的导数, 记为 $f'(x)$ 或 $\frac{df}{dx}(x)$, 于是, 在坐标表示下, 我们可以将微分简单写作 $df_x = f'(x)dx$, 这里我们将 dx 理解为一个切向量 (列向量).

接下来, 我们不加证明地列举微分的一些性质:

命题 B.20 设 $f, g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 在 x 处可微, 则

- f 在 x 处连续;

- 对任意 $\lambda_1, \lambda_2 \in \mathbb{R}$, $d(\lambda_1 f_x + \lambda_2 g_x) = \lambda_1 df_x + \lambda_2 dg_x$;
- 如果 $m = 1$, 那么 $d(f \cdot g)_x = g(x)df_x + f(x)dg_x$;
- 如果 $m = 1$, 并且 $g(x) \neq 0$, 则

$$d\left(\frac{f}{g}\right)_x = \frac{1}{g(x)^2} (g(x)df_x - f(x)dg_x);$$

- 如果 $m = n$, 那么 $dx = id$;
- 链式法则: 如果 f 在 x 处可微, g 在 $f(x)$ 处可微, 则 $g \circ f$ 在 x 处可微, 且 $d(g \circ f)_x = dg_{f(x)} \circ df_x$;
- 如果 f 存在反函数 f^{-1} , 则 f^{-1} 在 $f(x)$ 处可微, 且 $d(f^{-1})_{f(x)} = (df_x)^{-1}$.

同样, 最后三条说明了微分保持了复合单位元、复合和逆元关系. 而第二条则说明微分是一个函数空间 $(\mathbb{R}^n \rightarrow \mathbb{R}^m)$ 到函数空间 $(T\mathbb{R}_x^n \rightarrow T\mathbb{R}_{f(x)}^m)$ 的线性映射.

链式法则与反函数求导法则可以用导数写出:

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad (f^{-1})'(f(x)) = (f'(x))^{-1}.$$

这里我们都将导数理解为矩阵. 同样, 在 Leibniz 记号下, 如果 $z = z(y)$, $y = y(x)$, 那么链式法则可以写作

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

反函数的导数则可以写作

$$\frac{dy}{dx} = \left(\frac{dx}{dy} \right)^{-1}.$$

他们的含义都非常清晰.

我们看几个重要的例子.

例 B.8 线性映射和线性函数本身的导数也非常简单:

$$\frac{d(Ax)}{dx} = A, \quad \frac{d(b^T x)}{dx} = b^T.$$

其中 A 是一个矩阵, b 是一个向量. □

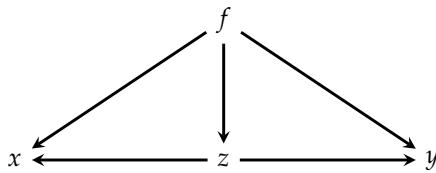
例 B.9 考虑一个多元实值函数 $g(x) = f(u_1(x), \dots, u_k(x))$, 先求 f 对 $u = (u_i)$ 的导数:

$$\frac{df}{du} = \left(\frac{\partial f}{\partial u_1}, \dots, \frac{\partial f}{\partial u_k} \right).$$

再求 u 对 x 的导数 $du/dx = (u'_1(x), \dots, u'_k(x))^T$. 根据链式法则,

$$\frac{dg}{dx} = \frac{df}{du} \cdot \frac{du}{dx} = \left(\frac{\partial f}{\partial u_1}, \dots, \frac{\partial f}{\partial u_k} \right) \begin{pmatrix} u'_1(x) \\ \vdots \\ u'_k(x) \end{pmatrix} = \sum_{i=1}^k \frac{\partial f}{\partial u_i} u'_i(x). \quad \square$$

例 B.10 我们来看一个更复杂的例子, 这个例子可以表明所谓求导链的意义. 考虑函数 $f(x, y, z) = z \exp(x + y)$, 其中 $z(x, y) = x + y$. 我们来求 f 对 x 的偏导数. 首先, 我们把变量之间的依赖关系用如下图表示出来, 其中 $z \rightarrow y$ 表示 z 依赖 y .



我们要求 f 对 x 的偏导数，首先找到从 f 出发可以到达 x 的全部路径，即 $f \rightarrow z \rightarrow x$ 和 $f \rightarrow x$ ，然后将路径上的相邻变量的偏导数相乘再相加，即

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial x} + \frac{\partial f}{\partial x} = \exp(x+y) \cdot 1 + z \exp(x+y).$$

这里两边都出现了 $\partial f / \partial x$ ，但他们的含义是不同的，左边的 $\partial f / \partial x$ 是 f 对 x 这个变量的偏导数，右边的 $\partial f / \partial x$ 是 f 对第一个位置偏导数，即 $\partial_1 f$. 一个不容易引起困惑但不太直观的写法是

$$\frac{\partial f}{\partial x} = \partial_3 f \cdot \partial_1 z + \partial_1 f.$$

这就是著名的反向传播算法的一个简单例子，它是神经网络训练中最重要的算法之一，也是很多神经网络框架优化的重点. \square

例 B.11 考虑二次型 $f(x) = x^T Ax$ (因此假设 A 是对称矩阵)，我们来求 $f'(x)$. 为此，考虑一个新函数 $g(x, y) = x^T Ay$ ，则 $f(x) = g(x, x)$ ，于是

$$f'(x) = g'(x, x) = \partial_1 g(x, x) \cdot \frac{\partial x}{\partial x} + \partial_2 g(x, x) \cdot \frac{\partial x}{\partial x} = \partial_1 g(x, x) + \partial_2 g(x, x).$$

我们来计算 $\partial_1 g(x, x), g(x, y) = (Ay)^T x$ ，因此根据例 B.8, $\partial_1 g(x, y) =$

$(Ay)^T$, 于是 $\partial_1 g(x, x) = x^T A^T$, 同理 $\partial_2 g(x, x) = x^T A$, 因此 $f'(x) = x^T A + x^T A^T = (2Ax)^T$. \square

注. 在求向量对向量的导数的时候, 很容易搞不清楚 Jacobi 矩阵的行列顺序, 一个简单的检查方法是看看导数的维度是否正确, 例如我们可以试试这一个矩阵是否可以乘自变量, 然后得到因变量的维数, 例如例 B.11 中, 如果我们求出来导数是 $2Ax$, 那么 $2Ax$ 是一个无意义的量, 说明我们的导数求错了, 应该要进行转置.

矩阵行列如何排列其实不影响导数值, 但是在进行链式法则的时候, 正确的排列可以机械地写出链式法则的结果, 这样才能实现自动求导器.

最后, 我们讨论高阶导数的概念. 对于向量值函数来说, 高阶导数是一个非常难以理解的概念, 所以我们只局限在实值函数讨论这一问题.

考虑一个实值函数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$, 它对第 i 个坐标的偏导数是 $\partial_i f$, 注意到这本身又是一个 \mathbb{R}^n 到 \mathbb{R} 实值函数, 我们可以继续讨论它的偏导数性质, 于是我们得到了二阶偏导数: $\partial_j(\partial_i f)$, 也记为

$$\frac{\partial^2 f}{\partial x_j \partial x_i}, \quad \partial_{j,i} f(x).$$

一般地, 我们也可以归纳定义 k 阶偏导数, 这里就不再赘述.

二阶偏导数一个重要的性质是, 一般情况下它可交换求偏导的顺序:

命题 B.21 设 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 具有下列两个二阶偏导数

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x), \quad \frac{\partial^2 f}{\partial x_i \partial x_j}(x),$$

并且在 x 处他们都连续，则两个偏导数相等。

一个直接的推论是，对于 $C^k(X)$ 的函数来说，求 k 阶偏导数不依赖于求导的顺序。

我们来看一个重要的例子。

例 B.12 设函数 $f \in C^k(\mathbb{R}^n)$, $x, h \in \mathbb{R}^n$, 考虑 $g(t) = f(x + th)$, $t \in [0, 1]$, 我们来求 $g^{(m)}(t)$, 其中 $m \leq k$. 先求一阶导数，根据链式法则，

$$g'(t) = \sum_{i=1}^n \partial_i f(x + th) h_i.$$

利用 nabla 算子，我们可以写作 $g'(t) = (h^\top \nabla) f$. 再求二阶导数，根据命题 B.21 和链式法则，

$$g''(t) = \sum_{i=1}^n h_i \frac{d}{dt} (\partial_i f(x + th)) = \sum_{i=1}^n h_i \sum_{j=1}^n \partial_j (\partial_i f(x + th)) h_j = \sum_{i,j=1}^n h_i h_j \partial_{j,i} f(x + th).$$

用 nabla 算子，我们可以写作 $g''(t) = (h^\top \nabla)^2 f(x + th)$. 一般地，我们有

$$g^{(m)}(t) = \sum_{i_1, \dots, i_m} \partial_{i_m, \dots, i_1} f(x + th) h_{i_1} \dots h_{i_m} = (h^\top \nabla)^m f(x + th). \quad \square$$

接下来，我们定义二阶导数⁶. 注意到，一个实值函数的一阶导数可以表示成一个向量值函数，即 $\text{grad } f$ ，因此，这个向量值函数的导数就是一个矩阵。我们将这个矩阵称为 f 的 **Hessian** 矩阵，记作 $H_f(x)$. 很

⁶更高阶的导数定义需要更加复杂的线性代数概念，我们这里就不引入了。

容易算出, Hessian 矩阵为

$$H_f(x) = \begin{pmatrix} \partial_{1,1}f(x) & \dots & \partial_{1,n}f(x) \\ \vdots & \ddots & \vdots \\ \partial_{n,1}f(x) & \dots & \partial_{n,n}f(x) \end{pmatrix}.$$

显然, Hessian 矩阵是一个对称矩阵, 因而可以构成某个二次型. 例 B.12 中二阶导数其实已经暗示了这一点, 我们可以将二阶导数写成一个二次型的形式:

$$g''(t) = h^\top H_f(x + th)h.$$

§B.3.2 微分学基本定理

类似一元函数, 我们讨论极值与导数的关系, 我们也不给出具体证明. 注意, 一元函数的极值的定义 (定义 B.18) 已经包含了多元函数情形, 所以这里就不再重复.

首先是 Fermat 引理的推广:

引理 B.2 (Fermat 引理) 设 $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}^n$ 是 f 的一个极值点, 且 f 在 x_0 处可微, 则 $f'(x_0) = 0$.

接下来是一系列中值定理, 我们这里依然只给出 Lagrange 中值定理:

定理 B.13 (Lagrange 中值定理) 设 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 是一个实值函数, 在闭区间 $[x, x+h] = \{x+th : t \in [0, 1]\}$ 上连续, 开区间 $(x, x+h) =$

$\{x + th : t \in (0, 1)\}$ 上可微, 则存在 $\xi \in (x, x + h)$, 使得

$$f(x + h) - f(x) = f'(\xi)h.$$

用参数的形式, ξ 可以写作 $\xi = x + \theta h$, 其中 $\theta \in (0, 1)$.

接下来我们讨论向量值函数的中值定理. 下面的例子表明, 向量值函数上中值定理不一定成立:

例 B.13 考虑匀速圆周运动, $r(t) = (\cos t, \sin t)$, 它的速度向量是 $r'(t) = (-\sin t, \cos t)$. 当绕一个周期之后, 位置又回到了原点, 于是 $r(2\pi) - r(0) = 0$, 然而 $r'(t)$ 恒不为 0, 因此不存在 $\xi \in (0, 2\pi)$ 使得 $r(2\pi) - r(0) = r'(\xi)(2\pi - 0)$. \square

不过, 中值定理的弱化版本, 有限增量定理, 是成立的:

定理 B.14 (Lagrange 有限增量定理) 设 $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 是一个函数, 在闭区间 $[x, x + h]$ 上连续, 开区间 $(x, x + h)$ 上可微, 则

$$\|f(x + h) - f(x)\| \leq \|h\| \sup_{\xi \in (x, x+h)} \|f'(\xi)\|.$$

注意, 这里的 $f'(\xi)$ 可能是一个矩阵, 此时 $\|f'(\xi)\|$ 的含义是矩阵范数, 具体的讨论参见附录 A.7.

例 B.14 这个例子探讨如何用二阶导数控制一阶导数的变化. 这一部分需要算子范数和谱理论的知识, 请参阅附录 A.7.

假设 $f \in C^2(X)$, $X \subseteq \mathbb{R}^n$, 那么对任意 $x, y \in X$ 满足 $[x, y] \in X$,

根据定理 B.14 和 Hessian 矩阵的定义，我们有

$$\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| \leq \|x - y\| \sup_{\xi \in (x,y)} \|H_f(\xi)\|.$$

我们讨论两种情况，首先假设 X 是紧集。因为 $H_f(x)$ 连续，根据例 B.6， $\|\cdot\|$ 连续，再根据命题 B.10， $\|H_f(x)\|$ 连续。由于 X 是紧集，根据 Weierstrass 最值定理（定理 B.7） $\|H_f(x)\|$ 在 X 上取到最大值 M ，因此，我们有

$$\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| \leq M \|x - y\|.$$

这一推导表明紧集上的 C^2 函数的梯度是 Lipschitz 连续的。所谓 F 是 Lipschitz 连续的，指的是存在一个常数 M ，使得对于任意定义域里的 x, y ，我们都有 $\|F(x) - F(y)\| \leq M \|x - y\|$ 。

在第二种情况下，假设 $X = \mathbb{R}^n$ ，我们使用 L^2 范数，于是根据定理 A.18，我们有

$$\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| \leq \sup_{\lambda \in \sigma(H_f(z)), z \in (x,y)} |\lambda| \cdot \|x - y\|.$$

这里 $\sigma(A)$ 是矩阵 A 的谱。

因此，只要知道了 Hessian 矩阵的谱，我们就可以控制梯度的变化，这一点对于凸优化算法非常重要，具体讨论见例 8.4。□

最后，我们要讨论高阶导数与极值的关系。首先，利用例 B.12 和一元的 Taylor 公式，我们可以得到多元函数的 Taylor 公式：

定理 B.15 (Taylor 展开) 设 $f \in C^k(U)$, $[x, x+h] \subseteq U$, 那么

$$f(x+h) = \sum_{j=0}^k \frac{1}{j!} (h^\top \nabla)^j f(x) + o(\|h\|^k).$$

根据这一定理, 我们可以得到二阶导数判定极值的充分条件:

定理 B.16 设 $f \in C^2(U)$, U 是开集, $x_0 \in U$ 且 $f'(x_0) = 0$, 则

- 如果 $H_f(x_0)$ 是正定的, 则 f 在 x_0 处取极小值;
- 如果 $H_f(x_0)$ 是负定的, 则 f 在 x_0 处取极大值;
- 如果 $H_f(x_0)$ 不定 (既非半正定也非半负定), 则 f 在 x_0 处不取极值.

§B.3.3 隐函数定理

微积分中, 还有一类非常重要的问题, 那就是解方程, 我们看一个非常简单的例子. 设 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $f(x, y) = x^2 + y^2 - 1$, 我们来求解方程 $f(x, y) = 0$, 也就是求单位圆的方程. 由于 f 是一个二次型, 因此我们可以直接求出它的根:

$$y = \pm \sqrt{1 - x^2}.$$

比如考虑圆周上的点 $(0.6, 0.8)$ 的附近, x 就可以把 y 表示出来: $y = \sqrt{1 - x^2}$. 如果考虑点 $(0.6, -0.8)$ 的附近, 我们也可以写出 $y = -\sqrt{1 - x^2}$.

总而言之, 只要给定圆周上一个点 (x_0, y_0) ($y_0 \neq 0$), 我们就可以找到一个邻域, 在这个邻域上确认一个 y 和 x 的函数关系 $y = y(x)$.

更一般地，给定函数方程 $F(x, y) = 0$ ，它确定了一个平面上的曲线 $C = \{(x, y) \in \mathbb{R}^2 : F(x, y) = 0\}$. 任取一点 $(x_0, y_0) \in C$ ，如果在 (x_0, y_0) 的某个邻域 U 上，我们可以确认一个 y 和 x 的函数关系 $y = y(x)$ ，使得 $U \cap C$ 中的所有点都可以用这个关系表示，那么我们其实就把一个隐藏在 $F(x, y) = 0$ 中的函数 y 解出来了，这就是隐函数的概念.

下面，我们考虑维数更高的情况. 设 $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^k$, $F(x, y) = 0$, 其中 $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. 任取一点 (x_0, y_0) 满足 $F(x_0, y_0) = 0$ ，同样，我们希望在 (x_0, y_0) 的某个邻域 U 上，将 $F(x, y) = 0$ 转化为一个等价的函数关系 $y = y(x)$.

首先我们指出，在一般情况下， $k = m$ 的时候讨论才有意义，这可以从线性方程组的理论看出，相关的线性代数理论可以参见附录 A. 假如说 $F(x, y) = 0$ 就是一个线性方程组：

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n + b_{11}y_1 + \cdots + b_{1m}y_m - c_1 &= 0, \\ &\dots \\ a_{k1}x_1 + \cdots + a_{kn}x_n + b_{k1}y_1 + \cdots + b_{km}y_m - c_k &= 0. \end{aligned} \tag{B.2}$$

我们也可以写成矩阵形式：

$$Ax + By = c.$$

其中 A 是一个 $k \times n$ 的矩阵， B 是一个 $k \times m$ 的矩阵， c 是一个 k 维向量.

如果 $k < m$ ，那么线性映射 $y \mapsto By$ 的秩是 $k < m$ ，根据推论 A.1，这个映射的核不是 $\{0\}$ ，所以对于任意一个满足 $Ax_0 + By_0 = c$ 的 (x_0, y_0) 来说，总可以再加上一个 $y \neq 0$ 使得 $Ax_0 + B(y + y_0) = c$ 且 $\|y\|$ 充

分小. 因此对任何 x_0 , 都找不到一个 y 的邻域, 其中有唯一的 y 使得 $Ax_0 + By = c$, 所以我们也不能解出 $y = y(x)$.

如果 $k > m$, 那么 $F(x, y) = 0$ 很有可能是空集. 比如, 下列线性方程组就没有解:

$$x_1 + x_2 = 0,$$

$$x_1 + x_2 = 1.$$

对于非线性方程组的情况, 如果 $F(x, y)$ 是可微的, 那么在一个点的局部函数的性质可以用线性映射近似, 于是也应该有 $k = m$. 这一事实可以简单归结为: 要解 m 个未知数 (即 y), 应该恰好有 m 个方程 (即 $F(x, y) = 0$) .

我们继续来看线性方程组的情况, 即 (B.2), 如果 $k = m$, $c = 0$, B 可逆, 很快就可以解出

$$y = -B^{-1}Ax.$$

对于一般的 F , 在它的每一个局部, 我们都可以用一个线性映射近似, 根据微分的偏导数表示, 这一线性映射恰好形如

$$\underbrace{\frac{\partial F}{\partial x}}_A h_x + \underbrace{\frac{\partial F}{\partial y}}_B h_y.$$

这里 h_y 和 h_x 就应该理解为映射在这一点的切向量. 于是, 假如 $\frac{\partial F}{\partial y}$ 可逆, 那么我们就可以解出

$$h_y = -\left(\frac{\partial F}{\partial y}\right)^{-1} \frac{\partial F}{\partial x} h_x.$$

假设我们可以解出函数关系 $y = f(x)$, 由于 h_x 和 h_y 是切向量, 根据导数的定义,

$$f'(x) = - \left(\frac{\partial F}{\partial y} \right)^{-1} \frac{\partial F}{\partial x}.$$

确定了导数就可以确定这个函数本身, 这就是隐函数定理的内容. 下面, 我们正式给出它的陈述.

定理 B.17 (隐函数定理) 设 $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, $F(x, y) = 0$, 其中 $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, 考虑点 (x_0, y_0) 的邻域 U , 如果:

- $F \in C^p(U; \mathbb{R}^m)$, $p \geq 1$;
- $F(x_0, y_0) = 0$;
- $\frac{\partial F}{\partial y}(x_0, y_0)$ 可逆,

那么存在开球 $B(x_0, r) \subseteq \mathbb{R}^n$, 开球 $B(y_0, s) \subseteq \mathbb{R}^m$, 以及一个函数 $f \in C^p(B(x_0, r); \mathbb{R}^m)$, 使得对任意 $x \in B(x_0, r)$, $y \in B(y_0, s)$, 都有

$$F(x, y) = 0 \iff y = f(x).$$

此外, f 的导数可以用 F 的偏导数表示:

$$f'(x) = - \left(\frac{\partial F}{\partial y} \right)^{-1} \frac{\partial F}{\partial x}.$$

用 Banach 不动点定理 (定理 8.2), 我们可以给一个该定理的简洁证明, 具体内容请参见第八章的习题??.

隐函数定理的一个特例是向量值函数的反函数的存在性定理:

定理 B.18 (反函数定理) 设 $f \in C^p(\mathbb{R}^n; \mathbb{R}^n)$, $p \geq 1$, $f'(x_0)$ 可逆, 那么存在 x_0 的邻域 V , 以及 $f(x_0)$ 的邻域 W , 使得 $f: V \rightarrow W$ 是一个双射, 且 $f^{-1} \in C^p(W; \mathbb{R}^n)$, 此外, f^{-1} 的导数可以用 f 的导数表示, 对于 $x \in V, y = f(x) \in W$, 我们有

$$f^{-1}(y)' = (f'(x))^{-1}.$$

作为一个注, 反函数定理中可逆性的判断可以利用 Jacobi 行列式是否非零来判断.

反函数定理最重要的用途是坐标变换. 我们之前的坐标变换都是线性的基到线性的基, 然而在微积分中非线性的坐标也非常常用, 比如极坐标、球坐标等. 这些坐标变换都是非线性的, 因此我们需要反函数定理来处理这些坐标变换. 我们考虑极坐标的例子.

例 B.15 (极坐标) 考虑一个半平面 $\mathbb{R}_{\geq 0} \times \mathbb{R} = \{(r, \phi) \in \mathbb{R} \times \mathbb{R} : r \geq 0\}$. 我们将它映射到 \mathbb{R}^2 平面上, 映射 f 定义为 $f(r, \phi) = (r \cos \phi, r \sin \phi)$. 我们也可以写得更像坐标变换一些:

$$x = r \cos \phi,$$

$$y = r \sin \phi.$$

这个变换的 Jacobi 行列式是 r , 因此除了 $r = 0$ 的点, 这个变换都是可逆的, 于是在 \mathbb{R}^2 的任何局部上, 我们都可以用极坐标来表示平面上的点.

我们借着这个例子来看一下 Jacobi 行列式的几何意义. 上述坐标

变换的微分是：

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \cos \phi & -r \sin \phi \\ \sin \phi & r \cos \phi \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix} = J_f \begin{pmatrix} dr \\ d\phi \end{pmatrix}.$$

这里对诸如 dx 的符号，我们有两种理解方式，一种是把他们理解为切向量，另一种是把他们理解为一段微小位移。不论哪一种，最终结果都是将 f 在局部近似为了一个线性映射。根据附录 A.6 的讨论，这一线性映射将平行体 $\Pi(dr, d\phi)$ 映射为平行体 $\Pi(dx, dy)$ ，而 Jacobi 行列式就是他们有向体积变化的比率。如果我们把 $\Pi(dr, d\phi)$ 的有向体积记为 $\partial(r, \phi)$ ，那么我们有

$$\frac{\partial(x, y)}{\partial(r, \phi)} = \det J_f.$$

这正是 Jacobi 行列式这一符号的意义。

我们还可以用 Leibniz 的记号更加形象地表达这件事情。作为坐标，我们认为 $dxdy$ 表示的正好就是长方形 $\Pi(dx, dy)$ 的有向面积（长乘宽），而 $drd\phi$ 表示的正好就是长方形 $\Pi(dr, d\phi)$ 的有向面积，于是我们有

$$\frac{dxdy}{drd\phi} = \frac{\partial(x, y)}{\partial(r, \phi)} \iff dxdy = \frac{\partial(x, y)}{\partial(r, \phi)} drd\phi.$$

在积分学中，这一符号（再加上绝对值）实际上直接给出了变量替换的公式。□

附录 C 概率论基础

本附录主要介绍 Kolmogorov 概率论，讨论只局限在数学层面，不涉及概率论的哲学讨论。本附录的连续型随机变量（向量）的讨论需要微积分的基本知识，关于微分学的部分，可以参见附录 B；积分学（主要是 Lebesgue 积分）我们会在附录 C.3 以数学期望的形式介绍。

§C.1 从朴素概率论到公理化概率论

§C.1.1 Kolmogorov 概率论

朴素的概率论通常讨论两种极端的情况，一个是可以用数数的方式来计算概率的情况，比如说掷骰子，另一个是用面积的方式来计算概率的情况，比如在随机选一个圆周上的点。这两个情况分别对应了古典概型和几何概型。

我们先给一些术语。考虑一个随机试验，它的所有可能结果组成的集合称为**样本空间**，记为 Ω 。样本空间的元素称为**样本点**，通常记为 ω 。样本空间的某些子集被称为**事件**。我们来看看这些概念在朴素的概率论中都具体是什么。

例 C.1 (古典模型) 先后掷两个骰子，样本空间为

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}.$$

样本点 (i, j) 表示第一个骰子掷出 i 点，第二个骰子掷出 j 点。

“第一个骰子掷出 i 点”这个事件可以表示为 $A_i = \{(i, j) : 1 \leq j \leq 6\}$ 。 “第一个骰子掷出 i 点，第二个骰子掷出 j 点”这个事件可以表示为 $B_{ij} = \{(i, j)\}$ 。 \square

例 C.2 (几何模型) 在圆周上随机选点。如果用弧度来表示圆周上的点，那么样本空间为

$$\Omega = [0, 2\pi).$$

样本点为 ω ，表示选出点的弧度。

事件 $A = [0, \pi)$ 表示选出了上半圆周，事件 $B = [0, \pi/2) \cup [\pi, 3\pi/2)$ 表示选出了右上或左下的 $1/4$ 圆周。 \square

那么，如何定义概率呢？朴素地说，概率是某个事件出现的可能性占总可能的比例。

对于古典模型，我们简单认为每个样本点出现的概率都是相同的，也就是说，如果用 p_ω 表示样本点 ω 出现的概率，那么对任意 $\omega \in \Omega$ ，都有 $p_\omega = 1/|\Omega|$ 。于是，对于任意事件 A ，它发生的概率为

$$\sum_{\omega \in A} p_\omega = \frac{|A|}{|\Omega|}.$$

例如在上面掷骰子的例子中， $p_\omega = 1/36$ ， A 发生的概率为 $1/6$ ， B 发生的概率为 $1/36$ 。

对于几何模型，不能再用古典模型的方式定义概率。一段长为 2π 的圆弧上，有不可数个点。如果选到每个点的概率相等，那么这个概率必须是 0，否则所有点的概率和是无穷大。更麻烦的是，我们也不能用古典模型的方式计算某个事件的概率！例如，选到上半圆周的概率，就是把所有上半圆周上的点的概率加起来，任意多个 0 相加依然还是 0，所以这样的定义出来的概率永远是零，这样是不可行的。

朴素的直觉告诉我们，选到上半圆周的概率是 $1/2$ ，因为上半圆周刚好占了半个圆周。所以几何模型的概率定义利用了体积的概念。事件 A 的概率定义为

$$\frac{\text{事件 } A \text{ 对应的体积}}{\text{样本空间 } \Omega \text{ 对应的体积}}.$$

这里体积是广义上的，一维集合的体积就是长度，二维集合的体积就是面积，三维集合的体积就是体积，以此类推。

例如在上面圆周的例子中， A 对应的体积（长度）为 π ， Ω 对应的体积（长度）为 2π ，所以 A 发生的概率为 $1/2$ 。同理， B 的概率也是 $1/2$ 。

几何模型的定义看似合理，却并不严谨：我们并不知道如何定义“体积”。我们来看一个有趣的例子。

例 C.3 (Bertrand 悖论) 考虑一个圆，它的半径为 1。现在我们随机地在圆上取一个弦，那么这个弦的长度超过 $\sqrt{3}$ （即圆内接正三角形的边长）的概率是多少？

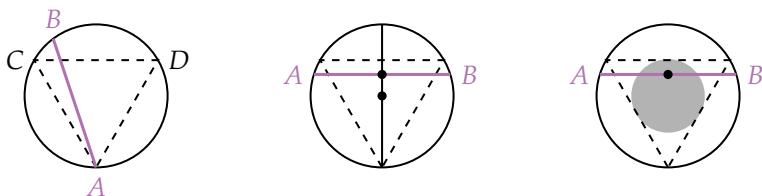
我们给出三种答案，这三种答案对应了我们对“随机”的不同理解。

解答 1. 不妨固定弦的其中一个点 A ，另一个点 B 在圆上等可能选取。以 A 为顶点作圆内接正三角形 ACD ，弦的长度超过 $\sqrt{3}$ 等价于 B 在弧 CD 上，所以概率为 $1/3$ 。

解答 2. 弦长只与它到圆心的距离有关系，与方向无关。弦长超过 $\sqrt{3}$ 等价于它到圆心的距离小于 $1/2$ ，所以概率为 $1/2$ 。

解答 3. 弦被它的中点唯一确定，弦长大于 $\sqrt{3}$ 等价于中点落在一个半径为 $1/2$ 的同心小圆内，所以概率为同心小圆面积比上大圆面积，即 $(1/2)^2 = 1/4$ 。

三种解答的示意图见下（从左到右分别是解答 1 到 3）：



□

同样的事件因为我们对“随机”的理解不同，得到了不同的概率！因此，我们需要一个更加严格的定义来描述概率。

首先注意到，概率应该是一个函数，它的值域是 $[0, 1]$ 。那么，它的定义域应该是什么呢？我们已经看到，概率应该定义在事件上，而非样本点上。那么，概率可以定义在任意事件上吗？这个问题很微妙，我们不在这里讨论。这里只是指出，我们关心的并不总是任意事件，而是一类被 σ -代数所刻画的事件。

定义 C.1 (σ -代数) 设 Ω 是一个集合， \mathcal{F} 是 Ω 的子集的集合。如果 \mathcal{F} 满足

1. $\Omega \in \mathcal{F}$ ；
2. 如果 $A \in \mathcal{F}$ ，则 A 的补集 $\Omega \setminus A \in \mathcal{F}$ ；

3. 如果 $A_1, A_2, \dots \in \mathcal{F}$, 则 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

则称 \mathcal{F} 是 Ω 上的一个 σ -代数. □

在样本空间中, 我们要求事件也形成一个 σ -代数, 这样的 σ -代数称为事件域, 记为 \mathcal{F} . 在数学上, σ -代数包括了绝大部分我们可以构造的事件, 这是因为, 容易验证, σ -代数中的事件对可数交、可数并和补运算都是封闭的, 并且包含了样本空间和空集. 关于这一定义的哲学讨论, 可以见第一章.

样本空间连同它的事件域, 被称为可测空间.

定义 C.2 (可测空间) 设 Ω 是一个集合, \mathcal{F} 是 Ω 上的一个 σ -代数. 则称 (Ω, \mathcal{F}) 是一个可测空间.

设 $S \subseteq \Omega$, 如果 $S \in \mathcal{F}$, 则称 S 是 \mathcal{F} -可测的. □

定义可测空间与 \mathcal{F} -可测的概念, 主要是为了区分一个集合到底是不是我们所关心的事件, 我们只关心 \mathcal{F} -可测的集合.

接下来, 我们给出 Kolmogorov 概率论的公理化定义.

定义 C.3 (概率空间, 概率测度) 设 (Ω, \mathcal{F}) 是一个可测空间. 如果函数 $\Pr : \mathcal{F} \rightarrow [0, 1]$ 满足

1. 正则性: $\Pr(\Omega) = 1$;

2. 可列可加性: 如果 $A_1, A_2, \dots \in \mathcal{F}$ 是两两不相交的事件, 则

$$\Pr \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \Pr(A_i),$$

则称 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, \Pr 称为概率测度或概率. □

容易证明，概率有如下性质：

命题 C.1 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间，则：

1. $\Pr(\emptyset) = 0$ ；
2. 单调性：对任意的 $A, B \in \mathcal{F}$ ，如果 $A \subseteq B$ ，则 $\Pr(A) \leq \Pr(B)$ ；
3. 有限可加性：对两两不相交的 $A_1, A_2, \dots, A_n \in \mathcal{F}$ ，有

$$\Pr\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \Pr(A_i).$$

他们的证明都不困难，我们略去.

下面，我们回到古典概型与几何概型，看看如何对他们构造概率空间.

对于古典概型来说，我们容易写出它的概率空间. 此时事件域恰好为所有 Ω 的子集的集合，概率测度的定义也就是我们之前的定义：
 $\Pr(A) = |A|/|\Omega|$.

对于几何概型来说，概率空间最大的困难在于事件域和概率测度的定义. 为了简化讨论，我们集中在 $\Omega = [0, 1]^n$ ，也就是 n 维立方体的情况.

先考虑事件域. 首先，事件域至少要包含长方体

$$\prod_i (a_i, b_i) = \{x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n : a_i < x_i < b_i\}.$$

这是我们可以构造的最基本的集合了. 我们就定义事件域为包含所有长方体的最小 σ -代数 $\mathcal{B}([0, 1]^n)$. 换言之，如果还有一个 σ -代数 \mathcal{F} 包含所有长方体，那么 $\mathcal{B}([0, 1]^n) \subseteq \mathcal{F}$. 我们将这一 σ -代数称为 **Borel 代数**.

Borel 代数包含了绝大部分我们要讨论的集合，例如开集、闭集、单点集、有限集、可数集等，可以简单归纳为“合理的集合”。

事件域的定义已经给出，我们还需要定义概率测度 \Pr ，它应该满足以下两个要求：

- 让正方体的概率等于它的体积。按照朴素的直觉，长方体 $\prod_i(a_i, b_i)$ 的体积应该是 $\prod_i(b_i - a_i)$ ，也就是

$$\Pr\left(\prod_i(a_i, b_i)\right) = \prod_i(b_i - a_i).$$

- 平移不变性。也就是说，如果 $A \in \mathcal{B}([0, 1]^n)$ ，那么对任意的 $x \in \mathbb{R}^n$ ，定义 $A + x = \{y \in \mathbb{R}^n : y = x + z, z \in A\}$ ，只要 $A \in \mathcal{B}([0, 1]^n)$ ，就有 $\Pr(A + x) = \Pr(A)$ 。

一个惊人的事实是，这样的概率测度是存在且唯一的，我们称之为 **Lebesgue 测度**，常记为 λ 。

注意，Borel 代数和 Lebesgue 测度的定义可以不局限在 $[0, 1]^n$ ，他们可以定义在与实数相关的各种集合上。在本附录中，我们最主要是用的是 \mathbb{R}^n 上的相关定义，例如 $\mathcal{B}(\mathbb{R}^n)$ 就是包含所有 n 维开长方体（每条边是开区间）的最小 σ -代数， λ 就是定义在 $\mathcal{B}(\mathbb{R}^n)$ 上的 Lebesgue 测度。 \mathbb{R}^n 上的 Lebesgue 测度其实是概率测度的扩展（而非概率测度），因为此时不再要求有正则性（即 $\lambda(\Omega) = 1$ ），但额外要求 $\lambda(\emptyset) = 0$ 。

§C.1.2 条件概率，独立性

接下来，我们讨论条件概率与独立性。我们还是看掷两个骰子的例子。掷完第一个骰子，我们马上观察结果，然后再掷第二个骰子。问第一

个骰子是 i , 第二个是 j 的概率是多少? 如果继续套用原来的概率空间, 很快就会觉得不对劲. 此时, 第一个骰子的结果完全没有随机性! 所以朴素的直觉告诉我们, 这里的概率应该有另一个依赖于第一次投骰子结果的定义, 这样的概率就是条件概率.

我们直接给出一般情况下条件概率的定义.

定义 C.4 (条件概率) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $A, B \in \mathcal{F}$ 是两个事件, 且 $\Pr(A) > 0$. 则称

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

是事件 B 在事件 A 发生的条件下发生的条件概率. □

以上定义要求 A 发生概率为正, 然而, A 是零概率的时候也是可能有条件概率的. 例如, 从 $[0, 1] \times [0, 1]$ 中均匀地随机选一个点 (X, Y) , 观察它的横坐标 X . 不管什么样的 x , $X = x$ 的概率都是 0. 然而, 直觉上, 条件在 $X = x$ 上, $Y > 1/2$ 的概率不仅存在, 而且应该是 $1/2$. 在附录 C.2 中, 我们会针对一类特殊的事件, 给出此时条件概率的定义.

我们继续看投两个骰子的例子. 假设事件 A 是“第一个骰子是 i ”, 事件 B 是“第二个骰子是 j ”. 我们可以计算出 $\Pr(B|A) = \Pr(B) = 1/6$. 如果单看计算的结果, 这是一个非常神奇的式子: 条件在 A 上和不条件在 A 上概率是一样的! 从直觉来说, 这件事情却并不神秘, 因为第一个骰子的结果和第二个骰子的结果是不应该有关系的. 我们把这种现象称为独立性. 更一般地, 对任意事件 A, B , 如果 $\Pr(A) > 0$, 那么

$$\Pr(B|A) = \Pr(B) \iff \frac{\Pr(A \cap B)}{\Pr(A)} = \Pr(B) \iff \Pr(A \cap B) = \Pr(A) \Pr(B).$$

最后一个式子并不要求 $\Pr(A) > 0$, 因此我们用它作为独立性的定义, 这样定义可以不依赖条件概率.

定义 C.5 (独立性) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $A, B \in \mathcal{F}$ 是两个事件. 如果 $\Pr(A \cap B) = \Pr(A)\Pr(B)$, 则称事件 A 和 B 相互独立.

一般地, 给定一个事件族 $\mathcal{A} \subseteq \mathcal{F}$, 如果对任意的有限个不同的 $A_1, A_2, \dots, A_n \in \mathcal{A}$, 都有

$$\Pr\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \Pr(A_i),$$

则称事件族 \mathcal{A} 中的事件是相互独立的. □

我们在定义中还给出了多个事件相互独立的定义, 这一定义是说不管挑出其中多少有限个事件, 他们都应该满足交的概率等于概率的积. 这并不等价于任意两个事件都相互独立, 我们看下面的例子.

例 C.4 两个人进行石头剪刀布游戏, 每个人独立等概率地出剪刀石头布.

考虑下面三个事件: $A = \{\text{甲出了石头}\}$, $B = \{\text{乙出了剪刀}\}$, $C = \{\text{甲赢}\}$.

容易算出, $\Pr(A \cap B) = \Pr(A)\Pr(B) = 1/9$, $\Pr(A \cap C) = \Pr(A)\Pr(C) = 1/9$, $\Pr(B \cap C) = \Pr(B)\Pr(C) = 1/9$, 所以 A, B, C 两两独立.

但是 A, B, C 不是相互独立的: $\Pr(A \cap B \cap C) = 1/9 \neq 1/27 = \Pr(A)\Pr(B)\Pr(C)$. □

这个例子说明, 三个事件的独立性远比他们任意两个之间的独立性要复杂, 三个事件放在一起可能才会出现不独立的情况. 对于一般情

况，这样的现象更加普遍，所以我们多个事件的独立性定义是要求任意有限个事件都独立，而不是任意两个事件都独立。

最后，我们给出条件概率的一些性质。

命题 C.2 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间，那么

1. 对任意 $A \in \mathcal{F}$ 满足 $\Pr(A) > 0$, $\Pr(\cdot|A)$ 也是一个概率测度；
2. $\Pr(\cdot|\Omega) = \Pr(\cdot)$ ，
3. 对任意 $A \in \mathcal{F}$ 满足 $\Pr(A) > 0$, $\Pr(A|A) = 1$.

以上性质的证明都很简单，我们就不给出了。

接下来我们给两个在 Bayes 概率论以及随机过程中很重要的性质。

定理 C.1 (全概率公式) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间， $A_1, A_2, \dots \in \mathcal{F}$ 是一列两两不相交的事件，且 $\Pr(A_i) > 0$, $\bigcup_{i=1}^{\infty} A_i = B$, 则对任意的 $C \in \mathcal{F}$, 有

$$\Pr(C|B) = \sum_{i=1}^{\infty} \Pr(C|A_i) \Pr(A_i).$$

特别地，对于有限个 A_i , 这一定理也成立。

证明. 注意到

$$\Pr(C) = \Pr(C \cap B) = \Pr\left(C \cap \bigcup_{i=1}^{\infty} A_i\right) = \Pr\left(\bigcup_{i=1}^{\infty} (C \cap A_i)\right) = \sum_{i=1}^{\infty} \Pr(C \cap A_i).$$

最后一个等号是因为 $C \cap A_i$ 两两不相交。另一方面，

$$\Pr(C \cap A_i) = \Pr(C|A_i) \Pr(A_i),$$

所以

$$\Pr(C) = \sum_{i=1}^{\infty} \Pr(C|A_i) \Pr(A_i).$$

对于有限个 A_i , 只需要把无穷求和改成有限求和, 利用有限可加性即可. \square

全概率公式是一种分而治之的思想, 它把一个复杂的事件分解成若干个简单的事件, 然后再把简单的事件的概率加起来. 我们来看一个例子.

例 C.5 从装有 w 个白球和 b 个黑球的盒子中随机地取出一个球, 不放回, 再取出一个球. 问第二个球是白球的概率是多少?

设事件 A 是“第一个球是白球”, 事件 B 是“第二个球是白球”. 我们有

$$\begin{aligned}\Pr(B) &= \Pr(B|A) \Pr(A) + \Pr(B|\bar{A}) \Pr(\bar{A}) \\ &= \frac{w-1}{w+b-1} \cdot \frac{w}{w+b} + \frac{w}{w+b-1} \cdot \frac{b}{w+b} \\ &= \frac{w}{w+b}.\end{aligned}$$

这里 \bar{A} 指的是 A 的补集, 即“第一个球是黑球”. \square

定理 C.2 (贝叶斯公式) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $A, B \in \mathcal{F}$ 且 $\Pr(A), \Pr(B) > 0$, 则

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}.$$

这一公式的证明几乎是显然的, 我们略去.

一个特别重要的推论被称为链式法则, 它是 Bayes 网络的基础.

推论 C.1 (链式法则) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $A_1, A_2, \dots, A_n \in \mathcal{F}$, 且 $\Pr(A_1 \cap A_2 \cap \dots \cap A_n) > 0$, 则

$$\begin{aligned} & \Pr(A_1 \cap A_2 \cap \dots \cap A_n) \\ &= \Pr(A_1) \Pr(A_2 | A_1) \Pr(A_3 | A_1 \cap A_2) \dots \Pr(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned}$$

我们也看一个例子.

例 C.6 (Pólya 的罐子) 一个罐子装有 w 个白球和 b 个黑球, 随机取出一个, 观察它的颜色, 放回, 再放回相同颜色的 c 个球, 再随机取一次, 重复上述操作, 如此反复 n 次, 问每一次都取到白球的概率是多少?

设事件 A_i 是“第 i 次取出的球是白球”. 我们有

$$\begin{aligned} \Pr(A_1) &= \frac{w}{w+b}, \\ \Pr(A_2 | A_1) &= \frac{w+c}{w+b+c}, \\ \Pr(A_3 | A_1 \cap A_2) &= \frac{w+2c}{w+b+2c}, \\ &\dots \\ \Pr(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}) &= \frac{w+nc}{w+b+nc}. \end{aligned}$$

所以

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \frac{w}{w+b} \cdot \frac{w+c}{w+b+c} \cdots \frac{w+nc}{w+b+nc}. \quad \square$$

注. 在概率论中, 我们经常要讨论事件的交, 所以我们通常会把 $A \cap B$ 简记为 AB . 此外, 事件不相交我们也称之为互斥. 事件 A 的补事件, 即 $\Omega \setminus A$. 我们会记为 \bar{A} 或 A^c .

另外, 我们也经常要讨论一个关于 ω 的陈述 $Q(\omega)$ 定义的事件 $\{\omega \in \Omega : Q(\omega)\}$, 在 Pólya 的罐子的例子中, 事件 A_1 其实就是由陈述 $Q(\omega)$: “ ω 中第一次取出的球是白球” 定义的事件. 在这种情况下, 我们将这一事件简记为 $\{Q\}$, 它的概率就是 $\Pr(\{Q\})$ 或者简记为 $\Pr(Q)$.

最后, 事件交的概率也经常以逗号的形式写出. 例如, 在 Pólya 的罐子的例子中, 我们会把概率 $\Pr(A_1 A_2)$ 记为

$\Pr(\text{第一次取出的球是白球}, \text{第二次取出的球是白球}).$

这样的记号更直观, 并且在随机变量部分会经常使用.

§C.2 随机变量, 分布函数

接下来, 我们讨论随机变量. 从某种意义上说, 随机变量是另一种刻画概率测度的手段. 不过, 随机变量能够更加直观、定量描述概率空间中的事件, 所以这是一个更加容易使用的概念.

§C.2.1 基本定义

为了理解随机变量的概念, 我们依然从古典概型入手.

例 C.7 继续考虑先后投两个骰子的情况, 假设它的概率空间是 $(\Omega, \mathcal{F}, \Pr)$, 他们的定义我们在附录 C.1.1 的末尾已经讨论过了.

我们可以定义一个从样本空间 Ω 到 \mathbb{N} 的函数 $S(i, j) = i + j$, 也就

是两个点数的和. 我们来看看 S 与事件域的关系. $\{S = s\} = \{(i, j) \in \Omega : i + j = s\}$, 所以 S 将原本的事件精简成了一个数字. 这个过程丢弃了一些事件, 例如 S 无法表达事件 $\{(1, 2)\}$, 实际上, 它无法区分 $(1, 2)$ 和 $(2, 1)$, 它把这两个样本点都看成了 3. 但是, S 仍然保留了很多信息, 例如, S 可以区分事件 $\{(1, 1)\}$ 和 $\{(2, 2)\}$, 它们分别对应 2 和 4. 总结来说, S 将原本更精细的事件域压缩成了更粗糙的事件域.

有了上面的感觉, 我们可以看一个更抽象的函数. 定义一个从样本空间 Ω 到 \mathbb{N}^2 的函数 X , 它的定义为 $X(i, j) = (i, j)$. 换句话说, 它把样本点看成一个 \mathbb{N}^2 的元素. \mathcal{F} 中的所有事件都可以表达为 $\{X \in B\}$, $B \subseteq \mathbb{N}^2$. 所以 X 完全刻画了整个事件域. \square

上面例子中的 S 和 X 都是随机变量的例子. 我们给出随机变量的定义.

定义 C.6 (随机变量, 随机向量, Borel 函数) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $X : \Omega \rightarrow \mathbb{R}$ 是一个函数. 如果对任意的 $x \in \mathbb{R}$, $\{X \in \mathcal{B}(\mathbb{R})\} \in \mathcal{F}$, 则称 X 是一个随机变量.

一般地, 考虑一个集合 \mathbb{R}^n 以及其上的 Borel 代数 $\mathcal{B}(\mathbb{R}^n)$, $X : \Omega \rightarrow \mathbb{R}^n$ 是一个映射. 如果对任意的 $A \in \mathcal{B}(\mathbb{R}^n)$, 集合 $\{X \in A\} \in \mathcal{F}$ -可测, 即 $\{X \in A\} \in \mathcal{F}$, 则称 X 是一个 n 维随机向量, 简称随机向量. 如果 $(\Omega, \mathcal{F}) = (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, 则称 X 是一个 **Borel 函数**. \square

下面对这个定义做一些说明. 首先, 随机变量是一个映射, 而不是一个数字, 这一点经常会被误解. 直观上说, 随机变量的值是随机的, 这个随机性是因为背后有一个未知的力量在抛硬币, 我们把从抛硬币到观测值这一整个东西称之为随机变量.

定义的后面还涉及了 σ -代数相关的东西, 我们也给一个简要说明.

Borel 代数包含了“合理的集合”，所以 $\{X \in B\}$ ($B \in \mathcal{B}(\mathbb{R})$) 表示事件“ X 落在合理的值集上”. 随机变量的要求其实就是，“ X 落在合理的值集上”是一个我们可以定义概率（即可测）的事件.

我们下面讨论一些随机变量的基本性质.

定理 C.3 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $X : \Omega \rightarrow \mathbb{R}^n$ 是一个随机向量, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 是一个 Borel 函数, 则 $g(X) = g \circ X$ 也是一个随机向量.

这一性质告诉我们了一种构造随机变量的方式：对一个随机变量进行一些 Borel 函数的操作. 下面的性质告诉我们，Borel 函数包含了绝大部分我们关心的函数，因此在实际中，我们不需要担心一个映射作用完之后是否还是随机变量.

命题 C.3 下面函数是 Borel 函数：

1. 所有的连续函数；
2. 给定 $A \in \mathcal{B}(\mathbb{R}^n)$, 示性函数 $I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$
3. 两个 Borel 函数的复合函数.

接下来，我们进入分布函数的讨论. 我们说过，随机变量某种意义上给出了概率测度的另一种刻画方式，而这一桥梁就是由分布函数给出的.

考虑概率空间 $(\Omega, \mathcal{F}, \Pr)$, 以及一个随机变量 $X : \Omega \rightarrow \mathbb{R}$. 要刻画概率测度 \Pr , 我们需要给出所有的事件 $A \in \mathcal{F}$ 的概率 $\Pr(A)$. 如果 A 可以被写成 $\{X \in B\}$ 的形式，那么我们可以用 $\Pr(X \in B)$

来刻画 $\Pr(A)$. 而我们之前说过, 要确定 $\Pr(X \in B)$, 至少要先确定 $\Pr(X \in (a, b))$. 这一概率还是有两个未定元 a, b , 所以更简便的方式是确定 $F_X(b) = \Pr(X \in (-\infty, b])$, 容易证明, 开区间的概率完全可以由 $F_X(b)$ 给出, 所以 F_X 完全刻画了 \Pr . 更一般地, 我们有如下定义.

定义 C.7 (分布函数) 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $X : \Omega \rightarrow \mathbb{R}$ 是一个随机变量. 定义函数 $F_X : \mathbb{R} \rightarrow \mathbb{R}$ 为 $F_X(x) = \Pr(X \leq x)$. 我们称 F_X 是 X 的分布函数, 记作 $X \sim F_X$.

如果 $X : \Omega \rightarrow \mathbb{R}^n$ 是一个随机向量, 定义函数 $F_X : \mathbb{R}^n \rightarrow \mathbb{R}$ 为 $F_X(x) = \Pr(X \leq x)$, 这里 $X \leq x$ 是指对任意的 $i = 1, 2, \dots, n$, 都有 $X_i \leq x_i$. 我们称 F_X 是 X 的分布函数, 记作 $X \sim F_X$. \square

容易验证, 分布函数具有如下的性质:

命题 C.4 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $X : \Omega \rightarrow \mathbb{R}$ 是一个随机变量, F_X 是它的分布函数, 则

1. F_X 是一个非减函数;
2. $\lim_{x \rightarrow -\infty} F_X(x) = 0$, $\lim_{x \rightarrow +\infty} F_X(x) = 1$;
3. F_X 是右连续的, 即对任意 $x \in \mathbb{R}$, 都有 $\lim_{y \downarrow x} F_X(y) = F_X(x)$;
4. F_X 在每一点处的左极限存在, 即对任意 $x \in \mathbb{R}$, 都有 $F(x-) = \lim_{y \uparrow x} F_X(y)$ 存在.

实际上, 分布函数也可以由命题 C.4 的前三条性质给出定义, 这是因为, 满足前三条性质的函数恰好是某个随机变量的分布函数:

定理 C.4 设 F 是 $\mathbb{R} \rightarrow \mathbb{R}$ 的函数, 满足命题 C.4 的前三条性质.

在概率空间 $([0, 1], \mathcal{B}([0, 1]), \lambda)$ 上, 存在一个随机变量 X , 使得 $F_X = F$.

所以, 我们今后也称呼满足命题 C.4 四条性质的函数为分布函数.

我们看一个分布函数计算概率的简单例子.

例 C.8 考虑 \mathbb{R} 上的分布函数 F , 它由随机变量 X 定义. 那么,

- $\Pr(X \leq a) = F(a),$
- $\Pr(X < a) = F(a-),$
- $\Pr(X > a) = 1 - F(a),$
- $\Pr(X \geq a) = 1 - F(a-),$
- $\Pr(X = a) = F(a) - F(a-).$

□

对于 $\mathbb{R}^n \rightarrow \mathbb{R}$ 型的分布函数 F , 我们也有类似的讨论. 此时有多个维度, 所以我们需要引入一个差分算子 $\Delta_{a_i b_i}$, 它的作用是对第 i 维作差:

$$\begin{aligned} & \Delta_{a_i b_i} F(x_1, x_2, \dots, x_n) \\ &= F(x_1, x_2, \dots, x_{i-1}, b_i, x_{i+1}, \dots, x_n) - F(x_1, x_2, \dots, x_{i-1}, a_i, x_{i+1}, \dots, x_n). \end{aligned}$$

例如, 对于区间 $(a, b] = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n : a_i < x_i \leq b_i\}$, 我们有

$$\Pr(X \in (a, b]) = \Delta_{a_1 b_1} \Delta_{a_2 b_2} \dots \Delta_{a_n b_n} F_X(x_1, x_2, \dots, x_n).$$

容易证明, 分布函数具有如下的性质:

命题 C.5 设 $(\Omega, \mathcal{F}, \Pr)$ 是一个概率空间, $X : \Omega \rightarrow \mathbb{R}^n$ 是一个随机向量, F_X 是它的分布函数, 则

1. 对任意 $a_i \leq b_i$, $i = 1, 2, \dots, n$, 都有 $\Delta_{a_i b_i} F_X(x_1, x_2, \dots, x_n) \geq 0$;
2. 所有 x_i 趋于正无穷时, F_X 趋于 1; 任意一个 x_i 趋于负无穷时, F_X 趋于 0;
3. F_X 对所有的 x_i 都是右连续的, 即当 $y \downarrow x$ (即对所有分量都有 $y_i \downarrow x_i$) 时, 都有 $F_X(y) \rightarrow F_X(x)$.

同样, 以上三条性质就决定了一个分布函数. 我们有如下的定理:

定理 C.5 设 F 是 $\mathbb{R}^n \rightarrow \mathbb{R}$ 的函数, 满足命题 C.5 的三条性质.

在概率空间 $([0, 1]^n, \mathcal{B}([0, 1]^n), \lambda)$ 上, 存在一个随机向量 X , 使得 $F_X = F$.

因此, 我们今后也称呼满足命题 C.5 三条性质的函数为分布函数.

注. 定理 C.4 和定理 C.5 其实还发挥着另一个重要的作用. 随机变量和随机向量的定义是非常抽象的, 所以我们并不能很直接验证随机变量的存在性. 然而, 分布函数却是极其容易构造的. 所以利用分布函数的存在性我们可以确保随机变量的存在性.

如果我们就限制在空间 $([0, 1]^n, \mathcal{B}([0, 1]^n), \lambda)$ 上, 随机向量几乎就等同于分布函数. 在更一般的情况下, 两个随机向量 X, Y 的分布函数相同时, 我们称 X, Y 同分布, 记为 $X \stackrel{d}{=} Y$.

现在, 我们将分布函数与概率测度联系在一起:

定理 C.6 设 $F : \mathbb{R}^n \rightarrow \mathbb{R}$ 是一个分布函数, 则在可测空间 $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ 上, 存在唯一的概率测度 \Pr , 使得对任意 $a_i \leq b_i$,

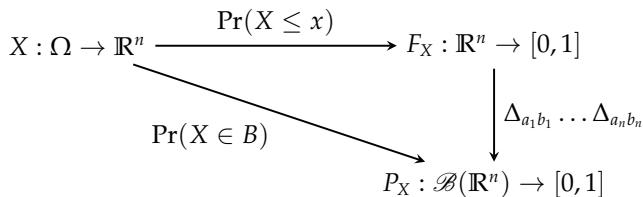
$$\Pr \left(\prod_{i=1}^n (a_i, b_i] \right) = \Delta_{a_1 b_1} \Delta_{a_2 b_2} \dots \Delta_{a_n b_n} F(x_1, x_2, \dots, x_n).$$

特别地, 分布函数

$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

对应的概率测度就是我们之前讨论的 $[0, 1]$ 上的 Lebesgue 测度.

总结来说, 随机向量 X 、概率测度 P_X 和分布函数 F_X 的关系如图:

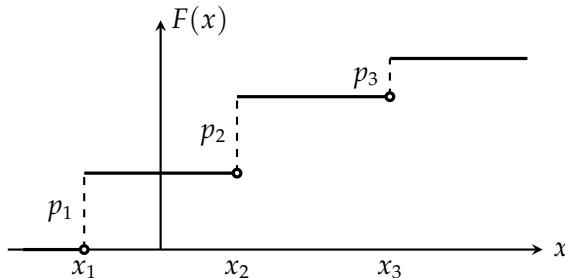


这张图的每一个箭头都可以反过来, 但是反过来的这些关系都比较不直观, 所以我们不再讨论.

根据上面的讨论, 分布函数的特性决定了随机变量的特性. 根据分布函数的不同性质, 我们可以将随机变量分为不同的类型. 下面我们将讨论一些重要的类别.

§C.2.2 离散型随机变量

我们首先讨论离散型随机变量. 离散型随机变量的分布函数 F 称之为离散型分布, 它是一个阶梯函数, 它的函数值只在有限或者可数个点 x_1, x_2, \dots 上发生跳变, 在 x_i 的跳变为 $p_i = F(x_i) - F(x_i^-)$. 这一分布函数对应的概率测度 \Pr 我们称之为离散型测度, 这种测度集中在 x_i 上, 即 $\Pr(X = x_i) = p_i$. 分布函数形如下图:



离散型分布可以由分布列给出, 分布列是一个序列 p_1, p_2, \dots , 其中 $p_i = \Pr(X = x_i)$, 且 $\sum_{i=1}^{\infty} p_i = 1$.

表 C.1 列举了一些本书中用到的离散型分布, 他们都是整数取值, 所以我们记 $p_i = \Pr(X = i)$.

名称	符号	分布列	参数
离散均匀	$\mathcal{U}[n]$	$p_i = 1/n, i = 1, \dots, n$	$n \in \mathbb{N}$
Bernoulli	$B(1, p)$	$p_1 = p, p_0 = 1 - p$	$p \in [0, 1]$
对称 Bernoulli	—	$p_1 = p_{-1} = 1/2$	—
二项	$B(n, p)$	$p_k = \binom{n}{k} p^k (1-p)^{n-k}$	$n \in \mathbb{N}, p \in [0, 1]$

表 C.1: 本书中用到的离散型分布

§C.2.3 连续型随机变量

我们再来讨论连续型随机变量, 连续型随机变量的分布函数 F 称为连续型分布, 对应的概率测度 \Pr 称之为绝对连续测度. 从名字上就可以看出, 测度才是定义连续型随机变量的关键. 我们给出绝对连续测度的定义.

定义 C.8 (绝对连续测度) \mathbb{R} 上的测度 \Pr 称为绝对连续测度, 如果对任意 $\epsilon > 0$, 存在 $\delta > 0$ 使得任意 $A \in \mathcal{B}(\mathbb{R})$ 满足 $\lambda(A) < \delta$, 都有 $\Pr(A) < \epsilon$. \square

直观上说, 绝对连续测度的意思是当体积 $\lambda(\cdot)$ 发生微小变化的时候(变化量为 $\lambda(A)$), 测度 $\Pr(\cdot)$ 也只发生微小的变化(变化量为 $\Pr(A)$), 这和通常函数连续的定义并没有太大的区别.

那么, 绝对连续测度对应的是连续分布函数吗? 并非如此! 不过, 绝对连续测度对应的分布函数有相当漂亮的一种刻画方式:

定理 C.7 (微积分基本定理) 设 $F : \mathbb{R} \rightarrow \mathbb{R}$ 是绝对连续测度对应的分布函数, 那么

$$\lambda(\{x \in \mathbb{R} : F'(x) \text{ 不存在}\}) = 0.$$

定义函数:

$$f(x) = \begin{cases} F'(x), & F'(x) \text{ 存在}, \\ 0, & \text{其他.} \end{cases}$$

则 f 是一个非负可积函数, 且对任意的 $a < b$, 都有

$$F(b) - F(a) = \int_a^b f(x) dx. \quad (\text{C.1})$$

此处的积分可以理解为 Riemann 积分或者后面附录 C.3 中的 Lebesgue 积分.

定理 C.7 意味着, 绝对连续测度对应的分布函数几乎处处可以求导, 并且所得到的导函数积分回去还是原来的分布函数, 也就是微积分基本定理成立. 这样的函数我们称之为绝对连续函数.

那么, 这个 f 应该如何理解呢? 先不管定理 C.7, 回到绝对连续测

度, 仿照导数的定义, 考虑极限

$$\frac{d\Pr}{d\lambda}(x) = \lim_{\lambda(A) \rightarrow 0, x \in A} \frac{\Pr(A)}{\lambda(A)},$$

也就是点 x 附近 $\Pr(\cdot)$ 的微小变化相对于 $\lambda(\cdot)$ 的微小变化.

那么, 给定一个集合 A , 要如何求 $\Pr(A)$? 按照微积分的朴素直观, 我们应该将 \Pr 微小的变化转变为 λ 微小的变化, 也就是积分:

$$\Pr(A) = \int_{x \in A} \frac{d\Pr}{d\lambda}(x) d\lambda(x).$$

我们可以把(C.1) 改写成如上的形式:

$$\Pr((a, b]) = \int_{x \in (a, b]} f(x) dx.$$

在一维的情况下, x 的微小变化就是 $\lambda(x)$ 的微小变化, 所以 $dx = d\lambda(x)$.

综合这两点, 我们容易相信,

$$f(x) = \frac{d\Pr}{d\lambda}(x) \iff d\Pr = f(x)d\lambda.$$

所以, f 应该理解为“密度”. 打个比方, λ 是物体的体积, \Pr 是物体的质量, 那么 f 就是这个物体每个很小的部分上的体积质量除以体积, 也就是密度. 所以, 我们将 f 称之为概率密度函数, 或者简称密度. 通常, X 的密度记作 p_X .

那么, 概率测度和密度的区别是什么呢? 对于刚接触概率论的人来说, 似乎很难理解他们之间的区别. 比如说, 有时候会写 $p(X = x)$ 甚至 $\Pr(X = x)$ 来表示密度在 x 处的值 $p(x)$, 又或者, 用 $\int \Pr(X = x) dx$

来表示对密度的积分. 这些当然都是不对的, 我们下面慢慢论述.

首先, 根据定理 C.7, F 是连续函数, 所以根据例 C.8, $\Pr(X = x) = F(x) - F(x-) = F(x) - F(x) = 0$. 所以 $\Pr(X = x)$ 根本就是零, 它和密度函数没有任何关系, 所以上面这些写法都是错的.

那么, 要怎么理解密度 $p(\cdot)$ 和概率测度 $\Pr(\cdot)$ 的区别呢? 当然, 从定义的角度他们就完全不同: 一个是从实数到实数的映射, 一个是从实数的集合到实数的映射. 但是这样的区别对于初学者来说并不直观. 最直观的区别就在于密度这一词: 虽然铅很重 (密度大), 但是几亿倍于铅体积的棉花却应该比铅重. 所以, 密度是微观的, 刻画很小部分集合的概率值, 也就是 $d\Pr = p_X d\lambda$; 而概率刻画的是宏观的, 计算任何一个集合的概率, 也就是 $\Pr(X \in A)$.

注. 上面的记号 $d\Pr / d\lambda$ 并不是随意写出来的, 我们叫它导数也不是随意的. 在测度论中, 定理 C.7 可以被推广为 **Radon-Nikodym 定理**, 这一定理直接保证了形如 $d\Pr / d\lambda$ 的函数的存在性, 这一函数被称之为 **Radon-Nikodym 导数**.

利用密度, 我们可以很容易计算概率:

定理 C.8 设 X 是一个连续型随机变量, f 是它的密度函数, 则对任意的 $B \in \mathcal{B}(\mathbb{R})$, 都有

$$\Pr(X \in B) = \int_{x \in B} f(x) dx.$$

在表 C.2 中, 我们给出本书中用到的一些连续型分布的密度函数.

注. 从定理 C.7 来看, 密度函数的定义似乎是唯一的, 但是从积分的角度, 如果密度函数在几个点上的值发生了变化, 并不影响整个积分的值,

名称	符号	密度函数	参数
连续均匀	$\mathcal{U}(a, b)$	$p(x) = \frac{1}{b-a}, x \in [a, b]$	$a < b$
指数	$\text{Exp}(\lambda)$	$p(x) = \lambda e^{-\lambda x}, x \geq 0$	$\lambda > 0$
双指数	$\text{DExp}(\lambda)$	$p(x) = \frac{\lambda}{2} e^{-\lambda x }, x \in \mathbb{R}$	$\lambda > 0$
Laplace	$\text{Lap}(\mu, \lambda)$	$p(x) = \frac{\lambda}{2} e^{-\lambda x-\mu }, x \in \mathbb{R}$	$\mu \in \mathbb{R}, \lambda > 0$
正态 (Gauss)	$\mathcal{N}(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$	$\mu \in \mathbb{R}, \sigma > 0$

表 C.2: 本书中用到的连续型分布

从而也不影响求概率. 比如均匀分布 $\mathcal{U}(a, b)$, 端点 a, b 的值到底是 0 还是 $1/(b-a)$ 并不重要, 取任何一个值都是可以的.

注. 密度函数通常是需要分段写出的, 比如, $\mathcal{U}(a, b)$ 的密度函数, 严格来说应该写为

$$p(x) = \begin{cases} 0, & x < a, \\ \frac{1}{b-a}, & x \in [a, b], \\ 0, & x > b. \end{cases}$$

为了简化记号, 我们可以用示性函数来表示这一分类. 设 $A \subseteq \mathbb{R}$, 定义函数

$$I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

则 $\mathcal{U}(a, b)$ 的密度函数可以写为

$$p(x) = \frac{1}{b-a} I_{[a,b]}(x).$$

更一般地, 示性函数中的字母 A 可以是任意一个事件, 而关于事件的那些记号都可以在 A 这里写出. 示性函数在概率论中有着核心的作用, 我们在后面将会经常用到示性函数.

§C.2.4 随机向量, 条件分布, 独立性

我们前面已经说过, 随机向量就是 $\Omega \rightarrow \mathbb{R}^n$ 的映射. n 维的随机向量可以看成 n 个随机变量的组合, 可以写作 $X = (X_1, \dots, X_n)^\top$. 通常, 我们将 X 的分布函数称为 X_1, \dots, X_n 的联合分布, 将 X_i 的分布函数称为 X 的边缘分布.

关于随机变量的分类可以完全平行移植到随机向量上. 下面我们分别讨论.

离散型随机向量指的是它对应的概率测度集中在有限或可数个点上. 这样的分布依然可以用分布列给出: $\Pr(X_1 = x_1, \dots, X_n = x_n) = p_{x_1, \dots, x_n}$, 其中 x_i 取遍所有可能的值.

本书中使用的离散型随机向量只有多项分布, 符号为 $PN(n, p_1, \dots, p_k)$, 分布列为

$$\Pr(X_1 = i_1, \dots, X_n = i_n) = \frac{n!}{i_1! \dots i_k!} p_1^{i_1} \dots p_k^{i_k},$$

其中 $n \in \mathbb{N}$, $p_i \geq 0$, $\sum_{i=1}^k p_i = 1$.

连续型随机向量指的是它对应的概率测度是绝对连续的. 连续型随机向量的分布函数依然由绝对连续函数刻画:

定理 C.9 设 $F : \mathbb{R}^n \rightarrow \mathbb{R}$ 是绝对连续测度对应的分布函数, 那么存在一

一个非负可积函数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$, 使得对任意的 $(x_1, \dots, x_n) \in \mathbb{R}^n$, 都有

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \dots, y_n) dy_1 \dots dy_n.$$

此时, f 称为 X 的概率密度函数, 或者简称密度. 通常, X 的密度记作 p_X .

类似随机变量的讨论, 密度函数依然可以被写做导数的形式. 假设 \Pr 是绝对连续测度, 它对应的密度是 p , 那么

$$\frac{d\Pr}{d\lambda}(x) = p(x) \iff d\Pr = p(x)d\lambda.$$

这里, 我们需要再给出一些 $d\lambda$ 和 dx 关系的讨论. $d\lambda$ 应该理解为 Lebesgue 测度的微小变化, 然而我们并不假定这一变化是如何产生的. dx 理解为 x 的微小变化. x 的微小变化自然就产生了 λ 的微小变化, 即 $\lambda(dx)$. 所以, 在 x 处, $d\lambda$ 和 dx 之间的关系应该是 $d\lambda = \lambda(dx)$, 于是 $d\lambda$ 应该理解为 dx 形成的长方体的体积.

同样, 密度给出了概率计算的一个重要工具:

定理 C.10 设 X 是一个 n 维连续型随机向量, f 是它的密度函数, 则对任意的 $B \in \mathcal{B}(\mathbb{R}^n)$, 都有

$$\Pr(X \in B) = \int_{x \in B} f(x) dx.$$

利用联合密度, 可以计算边缘密度:

定理 C.11 设 $X = (X_1, \dots, X_n)$ 是一个 n 维连续型随机向量, , 则对任

意的 $1 \leq i \leq n$, 都有

$$p_{X_i}(x_i) = \int_{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \mathbb{R}^{n-1}} p_X(x_1, \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n.$$

这一命题当然也可以自然推广到求随机向量的边缘密度, 例如利用 $X = (X_1, X_2, X_3)$ 的联合密度计算 (X_1, X_2) 的边缘密度:

$$p_{X_1, X_2}(x_1, x_2) = \int_{x_3 \in \mathbb{R}} p_X(x_1, x_2, x_3) dx_3.$$

连续型随机变量的一个重要的例子是多元正态分布, 或者 (非退化) Gauss 向量. 它的密度函数为

$$p(x) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)},$$

其中 $\mu \in \mathbb{R}^n$, Σ 是一个 $n \times n$ 的正定矩阵. 这一分布的符号是 $\mathcal{N}(\mu, \Sigma)$.

关于 Gauss 向量的性质, 我们将在附录 C.4 中讨论.

接下来, 我们讨论条件分布.

对于离散型随机向量 $X = (X_1, X_2)$, 它的分布完全由分布列给出. 我们可以定义 X_1 在给定 X_2 的条件下的分布列:

$$\Pr(X_1 = x_1 | X_2 = x_2) = \frac{\Pr(X_1 = x_1, X_2 = x_2)}{\Pr(X_2 = x_2)}.$$

由此给出了随机变量 X_1 在给定 X_2 的条件下的条件分布列, 继而给出了条件分布. 这一定义也可以推广到 X_i 是随机向量的情况.

然而, 对于一般的随机向量, 特别是连续型随机向量, 这一定义是行不通的. 比如, 如果 $X = (X_1, X_2)$ 是连续型随机向量, 那么 $\Pr(X_2 =$

$x_2) = \Pr(X_1 = x_1, X_2 = x_2) = 0$, 所以条件概率的分子和分布概率都是零, 这样的定义是没有意义的.

转换思路, 去尝试定义所谓的条件分布函数: $\Pr(X_1 \leq x_1 | X_2 = x_2)$.

考虑 $\Pr(X_1 \leq x_1 | x_2 < X_2 \leq x_2 + \epsilon)$, 再令 $\epsilon \downarrow 0$, 我们有如下计算:

$$\begin{aligned} & \lim_{\epsilon \downarrow 0} \Pr(X_1 \leq x_1 | x_2 < X_2 \leq x_2 + \epsilon) \\ &= \lim_{\epsilon \downarrow 0} \frac{\Pr(X_1 \leq x_1, X_2 \leq x_2 + \epsilon) - \Pr(X_1 \leq x_1, X_2 \leq x_2)}{\Pr(x_2 < X_2 \leq x_2 + \epsilon)} \\ &= \lim_{\epsilon \downarrow 0} \frac{F_X(x_1, x_2 + \epsilon) - F_X(x_1, x_2)}{F_{X_2}(x_2 + \epsilon) - F_{X_2}(x_2)}. \end{aligned}$$

如果上面的极限存在, 我们就定义它是 X_1 在给定 X_2 的条件下的条件分布.

如果 X 是连续性随机变量, 我们还可以继续算下去:

$$\begin{aligned} & \lim_{\epsilon \downarrow 0} \frac{F_X(x_1, x_2 + \epsilon) - F_X(x_1, x_2)}{F_{X_2}(x_2 + \epsilon) - F_{X_2}(x_2)} \\ &= \frac{\partial F_X(x_1, x_2)}{\partial x_2} \frac{1}{p_{X_2}(x_2)} \\ &= \int_{-\infty}^{x_1} \frac{\partial^2 F_X(y, x_2)}{\partial x_2 \partial y} \frac{1}{p_{X_2}(x_2)} dy \\ &= \int_{-\infty}^{x_1} \frac{p_{X_1, X_2}(y, x_2)}{p_{X_2}(x_2)} dy. \end{aligned}$$

对照定理 C.7, 我们知道 $p_{X_1, X_2} / p_{X_2}$ 具有密度函数的形式, 所以连续性随机向量所定义的条件分布也是连续型分布, 密度函数被 $p_{X_1, X_2} / p_{X_2}$ 通常记作 $p_{X_1 | X_2}$, 称为 X_1 在给定 X_2 的条件下的条件密度.

以上讨论也可以自然推广到 X_i 是随机向量的情况, 我们就不给出了.

最后, 我们讨论随机向量之间的独立性. 随机向量之间的独立性完全由事件的独立性刻画, 所以我们有如下定义:

定义 C.9 (随机向量的独立性) 设 X_1, \dots, X_n 是 n 个随机向量, 第 i 个的维数是 n_i . 如果对任意的 $1 \leq i_1, \dots, i_k \leq n$, 以及任意的 $B_{i_1} \in \mathcal{B}(\mathbb{R}^{n_{i_1}}), \dots, B_{i_k} \in \mathcal{B}(\mathbb{R}^{n_{i_k}})$, 都有

$$\Pr(X_{i_1} \in B_{i_1}, \dots, X_{i_k} \in B_{i_k}) = \Pr(X_{i_1} \in B_{i_1}) \dots \Pr(X_{i_k} \in B_{i_k}),$$

则称 X_1, \dots, X_n 是独立的.

特别地, 如果 X_1, \dots, X_n 是一维的, 那么这定义了随机变量之间的独立性. \square

这一定义中包含了无穷多个需要验证的等式, 利用分布函数, 我们可以将独立性的验证转化为一个等式的验证:

定理 C.12 设 X_1, \dots, X_n 是 n 个随机向量, 第 i 个的维数是 n_i , F_i 是 X_i 的分布函数, F 是 (X_1, \dots, X_n) 的联合分布函数. X_1, \dots, X_n 独立的充分必要条件是

$$F(x_1, \dots, x_n) = F_1(x_1) \dots F_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

对于离散型随机向量, 它的分布函数完全由分布列决定, 所以定理 C.12 等价于如下命题:

命题 C.6 设 X_1, \dots, X_n 是 n 个离散型随机向量, 第 i 个的维数是 n_i , p_i 是 X_i 的分布列, p 是 (X_1, \dots, X_n) 的联合分布列. X_1, \dots, X_n 独立的充分必要条件是

$$p(x_1, \dots, x_n) = p_1(x_1) \dots p_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

对于连续型随机向量, 它的分布函数完全由密度决定, 所以定理 C.12 等价于如下命题:

命题 C.7 设 X_1, \dots, X_n 是 n 个连续型随机向量, 第 i 个的维数是 n_i , p_i 是 X_i 的密度函数. 假设他们的联合分布具有密度函数 p . X_1, \dots, X_n 独立的充分必要条件是

$$p(x_1, \dots, x_n) = p_1(x_1) \dots p_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

上面两个命题都有更简单的形式:

推论 C.2 设 X_1, \dots, X_n 是 n 个连续型 (离散型) 随机向量, 第 i 个的维数是 n_i , 假设他们的联合分布具有密度函数 (分布列) p . X_1, \dots, X_n 独立的充分必要条件存在函数 f_1, \dots, f_n 使得

$$p(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n),$$

其中 $x_i \in \mathbb{R}^{n_i}$.

利用这一命题, 判断独立性的时候, 我们只要尝试将联合密度 (分布列) 分解成若干个函数的乘积即可.

对于连续型随机向量, 这一判据特别要注意密度函数的分段情况.

比如, 考虑 $X = (X_1, X_2)$, 其密度函数为

$$p(x_1, x_2) = \begin{cases} 8x_1x_2, & 0 \leq x_1 \leq x_2, 0 \leq x_2 \leq 1, \\ 0, & \text{其他.} \end{cases}$$

如果忽略了 x_i 的取值范围, 我们很容易以为 $p(x_1, x_2)$ 可以写成 $f(x_1)f(x_2)$, 所以他们独立. 然而事实并不是这样的! 计算 X_1 的边缘密度:

$$p_1(x_1) = \int_{x_2 \in \mathbb{R}} p(x_1, x_2) dx_2 = \begin{cases} 4x_1(1 - x_1^2), & 0 \leq x_1 \leq 1, \\ 0, & \text{其他.} \end{cases}$$

再计算 X_2 的边缘密度:

$$p_2(x_2) = \int_{x_1 \in \mathbb{R}} p(x_1, x_2) dx_1 = \begin{cases} 4x_2^3, & 0 \leq x_2 \leq 1, \\ 0, & \text{其他.} \end{cases}$$

显然, $p_1(x_1) \cdot p_2(x_2) \neq p(x_1, x_2)$, 所以 X_1, X_2 不独立.

如果使用示性函数来书写密度函数, 这一问题更不容易被忽视, 在上面的例子中, $p(x_1, x_2) = 8x_1x_2 I_{0 \leq x_1 \leq x_2 \leq 1}(x_1, x_2)$, 示性函数显然是拆不成分别只关于 x_1 和 x_2 的函数乘积的.

自然, 利用条件分布, 我们可以给出独立的另一种刻画:

命题 C.8 设 X_1, X_2 是两个随机变量, 他们的联合分布是离散型或连续型的. X_1, X_2 独立的充分必要条件是对任意的 x_1, x_2 , 都有

$$\Pr(X_1 \leq x_1 | X_2 = x_2) = \Pr(X_1 \leq x_1).$$

如果 X_1, X_2 是离散型的, 那么这一条件可以改写为

$$\Pr(X_1 = x_1 | X_2 = x_2) = \Pr(X_1 = x_1).$$

如果 X_1, X_2 是连续型的, 那么这一条件可以改写为

$$p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1).$$

注意, 上述判据并不需要真的把等式右边的量算出来, 我们只需要判断刻画条件分布的量 (条件分布函数、条件分布列或条件密度) 中, 是不是只出现了 x_1 而没有出现 x_2 .

§C.2.5 随机变量 (向量) 的函数

我们前面说过, 如果 X 是随机向量, g 是一个 Borel 函数, 那么 $g(X) = g \circ X$ 也是一个随机向量. 这里, 记号 $g \circ X$ 将 X 看成一个映射, 于是得到的是一个复合函数; 而记号 $g(X)$ 则更直观, 它表示把 X 看成一个数学对象 (随机向量), 然后对它进行函数运算, 得到另一个同类型的数学对象 (随机向量). 我们将始终采取后者的记号, 但请务必注意, 符号 $g(X)$ 中的 X 绝对不应该理解为一个数, 而应该理解为一个随机向量.

随机变量的函数最直接的问题就是, 它的分布是什么? 我们只关注离散型和连续型随机向量的情况.

对于离散型随机向量, 它的分布完全由分布列决定, 很容易得到如下命题:

定理 C.13 设 X 是一个离散型随机向量, g 是一个函数, 那么 $Y = g(X)$

也是一个离散型随机向量, 它的分布列为

$$\Pr(Y = y) = \sum_{x \in g^{-1}(y)} \Pr(X = x).$$

对于连续型随机向量, 它的分布完全由密度决定. 我们现在来推导连续型随机向量的函数的密度.

设 X 是一个 n 维连续型随机向量, $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, 即 $\mathbb{R}^n \rightarrow \mathbb{R}^n$ 的连续可微函数. 为了简便起见, 我们假设 g 是单射, 并且反函数也连续可微. 设 $Y = g(X)$, 可以证明, Y 是一个连续型随机向量.

我们现在来计算 Y 的密度. 考虑 Y 取值的一个微小的区域 dy , $dP_Y = p_Y \lambda(dy)$ 是 Y 在 dy 上的概率, 同样区域的概率也可以用 X 去计算:

$$dP_X = p_X \lambda(dx), \quad Y \in dy \iff X \in dx,$$

当然, 这里 dy 和 dx 由函数 g 联系在一起, 因为 $Y = g(X)$, 所以 $dy/dx = g'(X)$, 注意, 这相当于微元 dy 和微元 dx 的有向体积的比. 最后, 根据概率相等, 可以写出如下的等式:

$$dP_Y = dP_X \iff p_Y \lambda(dy) = p_X \lambda(dx). \quad (\text{C.2})$$

考虑到密度是计算体积而非有向体积, 根据 Jacobi 行列式的几何意义(见附录 B.3.1),

$$p_Y(y) = \left| \frac{dx}{dy} \right| p_X(x) = \left| \frac{dy}{dx} \right|^{-1} p_X(x) = |\det g'(x)|^{-1} p_X(x).$$

这就得到了 Y 的密度函数.

如果 g 不是单射, 那么上面的 (C.2) 需要考虑 g 每一个单射的局部. 例如, 如果 $g(x) = x^2$, 那么 g 在 $(0, +\infty)$ 上和 $(-\infty, 0)$ 上都是单射, 一个 y 对应了两个 x . 在这种情况下, 每一个 y 所对应的 x 都贡献了概率, 所以 (C.2) 需要写成

$$dP_Y = \sum_{g(x)=y} dP_X(x) \iff p_Y(\lambda(dy)) = \sum_{g(x)=y} p_X(x)\lambda(dx)(x). \quad (\text{C.3})$$

总结以上讨论, 我们得到连续型随机向量的函数的密度的计算公式:

定理 C.14 设 X 是一个连续型随机向量, $g \in C(\mathbb{R}^n, \mathbb{R}^n)$, 即 $\mathbb{R}^n \rightarrow \mathbb{R}^n$ 的连续函数, 假设 $\lambda(\{x \in \mathbb{R}^n : \det g'(x) \neq 0\}) = 0$, 则 $Y = g(X)$ 也是一个连续型随机向量, 它的密度函数为

$$p_Y(y) = \begin{cases} \sum_{g(x)=y} |\det g'(g^{-1}(y))|^{-1} p_X(g^{-1}(y)), & \det g'(g^{-1}(y)) \neq 0, \\ 0, & \text{其他.} \end{cases}$$

其中求和号中 $g^{-1}(y)$ 是根据相应的 x , 用反函数定理 (定理 B.18) 求出局部反函数.

这一定理的表述比较宽泛, 我们可以给一个具体的例子来理解.

例 C.9 设 X 是一个连续型随机变量, $g(x) = x^2$, 我们来计算 $Y = X^2 = g(X)$ 的密度. 直接计算定理 C.14 中的公式, 我们有

$$\begin{aligned} & \sum_{g(x)=y} |\det g'(g^{-1}(y))|^{-1} p_X(g^{-1}(y)) \\ &= \sum_{x^2=y} \frac{1}{2|x|} p_X(x) \end{aligned}$$

$$= \frac{1}{2\sqrt{y}} p_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} p_X(-\sqrt{y}).$$

这就给出了 Y 的密度. □

一般来说, 定理 C.14 中的公式并不好记, 最实用的还是根据 X 和 Y 在算相同的概率这一事实直接写出 (C.3), 然后根据具体的 g 来计算. 比如上面的例子, 我们可以直接写出

$$p_Y dy = p_X(\sqrt{y}) \lambda(dx)(\sqrt{y}) + p_X(-\sqrt{y}) \lambda(dx)(-\sqrt{y}).$$

两边除以 $\lambda(dy)$, 再利用 $dy/dx = 1/(2x)$, 就得到了 Y 的密度.

最后, 如果映射 g 并不是保持维度的, 例如 $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 但 $m < n$ ¹, 那么我们可以将 g 补全到 n 维映射, 也就是说, 我们可以定义一个新的函数 $G : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 满足

$$G(x_1, \dots, x_n) = (g_1(x_1, \dots, x_n), \dots, g_m(x_1, \dots, x_n), x_{m+1}, \dots, x_n)^T.$$

然后, 利用这一函数计算出 $g(X)$ 和 (X_{m+1}, \dots, X_n) 的联合概率密度, 再求出 $g(X)$ 的边缘密度.

我们看一个简单的例子.

例 C.10 (卷积) 设 X, Y 是随机变量, 我们来计算 $Z = X + Y$ 的密度. 我们可以将 Z 看成是 $g(X, Y) = (X + Y, Y)$ 的第一个维度. 映射 $(x, y) \mapsto$

¹如果 $m > n$, 那么 $g(X)$ 一定不会是连续型随机变量, 因为它的每个维之间一定会产生相互的关联, 所以我们不讨论这种情况.

$(x+y, y)$ 显然是双射，所以 (C.3) 退化为 (C.2)，我们有

$$p_{Z,Y}(z, y) = \left| \frac{\partial(z, y)}{\partial(x, y)} \right|^{-1} p_{X,Y}(x, y) = p_{X,Y}(z - y, y).$$

于是， Z 的边缘密度为

$$p_Z(z) = \int_{y \in \mathbb{R}} p_{X,Y}(z - y, y) dy.$$

这被称为 X 和 Y 的卷积. □

最后，对随机向量作用函数是不会影响独立性的：

命题 C.9 设 X_1, \dots, X_n 是 n 个随机向量，第 i 个的维数是 n_i ， g_i 是 $\mathbb{R}^{n_i} \rightarrow \mathbb{R}^{m_i}$ 的 Borel 函数， $Y_i = g_i(X_i)$ ，如果 X_1, \dots, X_n 相互独立，那么 Y_1, \dots, Y_n 也相互独立。

§C.3 随机变量的数字特征，条件数学期望

分布函数或者随机变量依然是一个映射，研究起来还是会比较复杂。我们希望能够用一些数字来刻画随机变量的特征，这样可以进一步简化问题。在这一节中，我们将介绍随机变量的数字特征，以及条件数学期望。

§C.3.1 数学期望, Lebesgue 积分

数学期望在数学上是很直观的, 我们可以从一个赌博的例子入手来找一些感觉.

例 C.11 在一个地下赌场, 有赌徒甲乙两人. 这是一个公平的赌局, 每局甲乙获胜概率都是 $1/2$, 每局各出赌注 50 块. 谁先赢到三局, 就可以赢得全部的赌注. 赌博进行了三轮, 甲赢了两局, 乙赢了一局. 这时, 突然有消息说警察马上就要来查封赌场, 甲乙于是决定将目前的所有赌资进行分割. 他们应该如何分割呢?

再赌两盘就会决出胜负, 赌博一共会有三种可能:

1. 第四盘甲赢, 于是甲赢的所有赌注, 这样的概率是 $1/2$;
2. 第四盘乙赢, 第五盘甲赢, 于是甲赢的所有赌注, 这样的概率是 $(1/2) \times (1/2) = 1/4$;
3. 乙连赢两盘, 于是乙赢的所有赌注, 这样的概率是 $(1/2) \times (1/2) = 1/4$.

现在的赌资是 $100 \times 3 = 600$ 块, 甲有 $1/2 + 1/4 = 3/4$ 的概率会拿到全部, 乙有 $1/4$ 的概率会拿到全部. 于是, 按照概率去平分的话, 甲应该拿走 450 块, 乙应该拿走 150 块. \square

这个例子说明了期望的一种理解方式: 在面对随机性的时候, 我们按照概率的权重分配. 比如, 上面的例子中, 设 X 是甲赢的赌注, 那么 X 的分布列为 $\Pr(X = 0) = 1/4$, $\Pr(X = 600) = 3/4$, 所以 $\mathbb{E}[X] = 0 \times 1/4 + 600 \times 3/4 = 450$.

以上的例子给了我们定义随机变量期望的基础: 定义示性函数的数学期望. 设 A 是一个事件, 那么 I_A 是一个随机变量:

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

我们称之为事件 A 的示性函数. 示性函数的分布列是

$$\Pr(I_A = 1) = \Pr(A), \quad \Pr(I_A = 0) = \Pr(A^c) = 1 - \Pr(A).$$

所以, 示性函数的数学期望, 按照上面的逻辑, 应该是

$$\mathbb{E}[I_A] = 1 \times \Pr(A) + 0 \times \Pr(A^c) = \Pr(A).$$

示性函数建立了概率和数学期望的联系. 下面, 我们来定义一般随机变量的数学期望, 这一定义的过程反映了一种数学的思想: 用简单东西的极限去研究复杂的东西.

第一步, 定义示性函数的数学期望². $\mathbb{E}[I_A] = \Pr(A)$.

第二步, 定义简单随机变量的数学期望. 简单随机变量是形如 $X = \sum_{k=1}^n x_k I_{A_k}$ 的随机变量, 其中 $x_k \in \mathbb{R}$, A_k 是事件. 定义

$$\mathbb{E}[X] = \sum_{k=1}^n x_k \Pr(A_k).$$

这一定义与第一步是相容的: 因为 $I_A = 1 \cdot I_A$, 所以 $\mathbb{E}[I_A] = 1 \cdot \Pr(A) = \Pr(A)$.

²从逻辑上说, 示性函数的数学期望是被定义出来的, 而不是被算出来的, 因为此时我们还完全没有定义什么是数学期望.

第三步, 定义非负随机变量的数学期望. 非负随机变量是指 $X(\omega) \geq 0$ 对任意 ω 成立的随机变量 X . 考虑一列简单随机变量 $\{X_n\}_{n=1}^{\infty}$, 它满足对于每一个 $\omega \in \Omega$ 都有当 $n \rightarrow \infty$ 时 $X_n(\omega) \uparrow X(\omega)$. 容易验证, $\mathbb{E}[X_n]$ 也是单调递增的, 所以根据命题 B.13, $\mathbb{E}[X_n]$ 有有限的极限或者趋于正无穷, 我们都记为 $\lim_{n \rightarrow \infty} \mathbb{E}[X_n]$.

定义 C.10 (数学期望 (Lebesgue 积分), 非负情形) 称

$$\mathbb{E}[X] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n]$$

为随机变量 X 的数学期望或 **Lebesgue 积分**.

□

可以证明, 这一定义不依赖于 $\{X_n\}_{n=1}^{\infty}$ 的选取, 因而是良定义的. 此外, 容易看出, 这一定义与第二步是相容的, 所以第三步扩展了第二步的定义.

第四步, 定义一般随机变量的数学期望. 考虑随机变量 X , 定义 $X^+ = \max\{X, 0\}$, $X^- = -\min\{X, 0\}$, 也就是 X 的正数部分和负数部分, 那么 $X = X^+ - X^-$. 我们有如下定义:

定义 C.11 (数学期望 (Lebesgue 积分), 一般情形) 称随机变量 X 的数学期望存在, 如果 $\mathbb{E}[X^+]$ 和 $\mathbb{E}[X^-]$ 至少有一个有限. 此时, 定义

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

为随机变量 X 的数学期望或 **Lebesgue 积分**.

如果 $\mathbb{E}[X^+]$ 和 $\mathbb{E}[X^-]$ 都是有限的, 那么称 X 有有限期望或可积的.

□

当我们强调积分的时候, $\mathbb{E}[X]$ 也会写为

$$E[X] = \int_{\Omega} X d\Pr.$$

以上定义适用于任何一种概率空间和概率测度. 容易看出来, 这一定义也适用于 \mathbb{R}^n 上的 Lebesgue 测度, 我们唯一需要改变的就是示性函数的 Lebesgue 积分的定义: 对任意 $A \in \mathcal{B}(\mathbb{R}^n)$, 定义

$$\int_{\mathbb{R}^n} I_A(\omega) \lambda(d\omega) = \lambda(A).$$

然后对简单函数定义积分, 再对非负函数定义积分, 最后对一般函数定义积分.

对于 \mathbb{R}^n 上的 Lebesgue 积分, 我们一般省略 λ^3 , 直接写成

$$\int_{\mathbb{R}^n} f(x) dx.$$

这与我们所熟知的积分符号就完全一致了.

上面定义随机变量期望的过程中, 最难以理解的是第三步, 也就是非负随机变量的数学期望. 我们来具体算一下它的表达式.

设 X 是一个非负随机变量, 分布为 F . 我们来计算 $\mathbb{E}[X]$, 与其说是计算, 不如说重新推导一遍第三步的过程. 首先, 我们将 X 取值范围离散化, 每 $1/n$ 一段, X 的值都压到形如 k/n 的点上, 这样就转化为了

³对于一维的情况, 见附录 C.2.3 的讨论. 在高维空间中, 这样的记号其实是相当糟糕的: 在微分学中, 求导数时, dx 被理解为切空间的向量, 或者一个微小的位移; 求然而在积分学中, dx 被理解为所对应平行体的体积. 所以其实 $\lambda(dx)$ 这一记号虽然复杂, 但是含义更准确.

一个离散型随机变量:

$$X_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \leq X < (k+1)/n\}}.$$

容易看出, $X_n(\omega) \uparrow X(\omega)$ 对任意 ω 成立. 于是, 我们有

$$\mathbb{E}[X] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n].$$

我们来计算 $\mathbb{E}[X_n]$, 注意到 X_n 是一个简单随机变量, 我们有

$$\begin{aligned} \mathbb{E}[X_n] &= \sum_{k=0}^{\infty} \frac{k}{n} \Pr\left(\frac{k}{n} \leq X < \frac{k+1}{n}\right) \\ &= \sum_{k=0}^{\infty} \frac{k}{n} \left(F\left(\frac{k+1}{n}\right) - F\left(\frac{k}{n}\right)\right). \end{aligned}$$

按照极限的想法, 当 $n \rightarrow \infty$ 时, 上式的求和项相当于 $x dF(x)$, 这里 dF 表示 x 微小变化时对应的 F 的微小变化. 所以形式上我们有

$$\mathbb{E}[X] = \int_{x \geq 0} x dF(x) = \int_{\mathbb{R}} x dF(x),$$

这里第二个等式是因为在 $x < 0$ 的时候 F 恒等于 0, 因而可以理解为 $dF(x) = 0$.

如果 X 不是非负的, 那么对 X^+ 和 X^- 分别计算数学期望, 然后相减, 就得到了一般随机变量的数学期望, 它依然满足:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x dF(x).$$

所以, 随机变量的数学期望完全取决于它的分布函数.

对于离散型随机变量来说, F 只在点 x_1, x_2, \dots 会发生改变, 其他地方都是常值, 所以我们有

$$\int_{\mathbb{R}} x dF(x) = \sum_{k=1}^{\infty} x_k (F(x_k) - F(x_k-)) = \sum_{k=1}^{\infty} x_k \Pr(X = x_k).$$

对于连续型随机变量来说, $dF = p dx$, 这里 p 是对应的密度. 于是我们有

$$\int_{\mathbb{R}} x dF(x) = \int_{\mathbb{R}} x p(x) dx.$$

以上就是概率论中常见的求期望的形式.

我们再介绍一个非常有用的符号, 它允许我们在某个事件 A 上求积分:

$$\int_A X d\Pr = \int_{\Omega} X I_A d\Pr = \mathbb{E}[X I_A].$$

相应地, 在 \mathbb{R}^n 上, 对我们也可以定义

$$\int_A f(x) \lambda(dx) = \int_{\mathbb{R}^n} f(x) I_A(x) \lambda(dx).$$

刻画随机变量的数字特征, 除了可以用随机变量的期望, 还可以用随机变量的函数的期望, 我们列举一个重要的概念.

定义 C.12 (矩, 方差, 特征函数) 设 X 是一个随机变量, 我们有如下定义:

- k 是一个正整数, 称 $\mathbb{E}[X^k]$ 为 X 的 k 阶矩; 称 $\mathbb{E}[(X - \mathbb{E}[X])^k]$ 为 X 的 k 阶中心矩;
- 称 $\mathbb{E}[(X - \mathbb{E}[X])^2]$ 为 X 的方差, 记为 $\text{Var}(X)$;

- 称 $f_X(t) = \mathbb{E}[\exp(itX)]$ 为 X 的**特征函数**. 一般地, 如果 X 是 n 维随机向量, 那么 $f_X : \mathbb{R}^n \rightarrow \mathbb{C}$, $f_X(t) = \mathbb{E}[\exp(i\langle t, X \rangle)]$ 被称为 X 的**特征函数**. \square

我们将会在后面讨论他们的性质.

§C.3.2 数学期望的性质

我们已经给出了数学期望的定义, 下面我们罗列一些数学期望的性质, 但都不给出证明.

命题 C.10 1. 期望的线性性: 设 X, Y 是随机变量, $a, b \in \mathbb{R}$, 如果 $\mathbb{E}[X]$ 和 $\mathbb{E}[Y]$ 都存在, 那么 $\mathbb{E}[aX + bY]$ 存在, 且

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

2. 单调性: 设 X, Y 是随机变量, 如果 $X \leq Y$, 那么

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

3. 绝对值不等式: 设 X 是随机变量, 那么

$$\mathbb{E}[|X|] \geq |\mathbb{E}[X]|.$$

4. 局部可积性: 设 X 是随机变量, 并且 $\mathbb{E}[X]$ 存在, 那么对任意事件 A , $\mathbb{E}[XI_A]$ 也存在; 如果 $\mathbb{E}[X]$ 有限, 那么 $\mathbb{E}[XI_A]$ 也有限.

接下来, 我们讨论随机变量函数的期望的求法. 假设 X 是一个 n 维

随机向量, $g : \mathbb{R}^n \rightarrow \mathbb{R}$ 是一个 Borel 函数, 那么 $g(X)$ 也是一个随机变量 (定理 C.3) . 计算 $\mathbb{E}[g(X)]$ 有以下两种方式, 我们下面分别讨论.

第一种, 利用附录 C.2.5 中的方法, 我们可以将 $g(X)$ 的分布写出来, 然后计算期望. 我们来看一个例子.

例 C.12 设 $X \sim \mathcal{U}(0, 1)$, 计算 $\mathbb{E}[X^2]$. 直接算出 $Y = X^2$ 的密度函数为

$$p_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

于是,

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} y p_Y(y) dy = \int_0^1 \frac{y}{2\sqrt{y}} dy = \frac{1}{3}. \quad \square$$

第二种, 我们从定义出发, 直接计算 $\mathbb{E}[g(X)]$. 我们先考虑最简单的情况, 即 g 连续并且 $0 \leq g \leq C$ 的情况, 这里 C 是一个正常数. 我们还是试图使用第三步, 用简单随机变量去逼近 $g(X)$. 我们选择离散化 X , 还是一样定义

$$X_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \leq X < (k+1)/n\}}.$$

可以证明⁴

$$\mathbb{E}[g(X)] = \lim_{n \rightarrow \infty} \mathbb{E}[g(X_n)].$$

⁴注意, 这里 $g(X_n)$ 未必单调趋于 $g(X)$ 了, 所以这里我们其实跳了一个比较重要的步骤, 即不单调趋于的时候极限也可以拿到期望外面. 由于这一步的证明比较技术, 而且对本书的讨论不是特别重要, 所以这里略去.

用 X 的分布函数 F 写出来 $\mathbb{E}[X_n]$ 就是

$$\mathbb{E}[X_n] = \sum_{k=0}^{\infty} g\left(\frac{k}{n}\right) \left(F\left(\frac{k+1}{n}\right) - F\left(\frac{k}{n}\right)\right).$$

取极限，写成积分的形式，我们有：

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) dF(x).$$

利用逼近的思想，我们可以将上述结论推广到 g 是任意的 Borel 函数的情况，于是我们有：

定理 C.15 设 X 是一个 n 维随机向量，每一维都可积， $g : \mathbb{R}^n \rightarrow \mathbb{R}$ 是一个 Borel 函数，那么 $\mathbb{E}[g(X)]$ 存在，且

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) dF_X(x).$$

特别地，如果 X 是一个离散型随机变量，取值为 x_1, x_2, \dots ，那么

$$\mathbb{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k) \Pr(X = x_k).$$

如果 X 是一个连续型随机变量，密度为 p_X ，那么

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) p_X(x) dx.$$

例 C.13 我们重新算一次上面的例 C.12，这次我们用定理 C.15 来计算。

设 $X \sim \mathcal{U}(0,1)$, 我们有

$$E[X^2] = \int_0^1 x^2 dx = \frac{1}{3}. \quad \square$$

从这两个例子就可以看出, 以上两种方法, 通常来说第二种会更加容易计算一些, 因为它只需要做一次积分, 而第一种方法还需要算变量替换的 Jacobi 行列式.

接下来, 我们讨论示性函数的性质.

命题 C.11 1. 设 A 是一个事件, 那么 $E[I_A] = \Pr(A)$, $\text{Var}(I_A) = \Pr(A)(1 - \Pr(A))$.

2. 设 A, B 是两个事件, 那么 $I_A I_B = I_{AB}$, 特别地, $I_A^2 = I_A$.

这些性质的证明都比较容易, 这里就不给出了.

利用示性函数, 我们可以重写事件独立性的定义:

命题 C.12 设 A, B 是两个事件, 那么 A 和 B 独立的充分必要条件是

$$E[I_A I_B] = E[I_A]E[I_B].$$

如果我们还记得随机变量的期望是如何定义的, 那么我们可以发现, 命题 C.12 的结论可以推广到随机变量的情形:

定理 C.16 设 X, Y 是两个相互独立的随机变量, 那么 $E[XY] = E[X]E[Y]$.

需要注意的是, 这一命题的逆命题不一定成立.

最后, 我们给一个重要的不等式. 我们说函数 $g : \mathbb{R} \rightarrow \mathbb{R}$ 是凸函

数, 如果对任意 $x, y \in \mathbb{R}$, $t \in [0, 1]$, 都有

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y).$$

关于凸函数的更多讨论, 见第 6.2 节. 我们有如下不等式:

定理 C.17 (Jensen 不等式) 设 X 是一个随机变量, $g : \mathbb{R} \rightarrow \mathbb{R}$ 是一个凸函数, 那么

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

§C.3.3 随机变量的内积空间

我们指出, 随机变量利用期望可以定义内积, 从而定义内积空间, 关于内积空间的讨论, 见附录 A.5. 在附录 C.4 中, 这一事实非常重要.

我们定义内积如下:

定义 C.13 (协方差) 设 X, Y 是两个随机变量, 称

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

为 X 和 Y 的协方差. □

容易验证, 在差一个常数的意义下, 协方差是一个对称正定的双线性型:

命题 C.13 $\text{Cov}(\cdot, \cdot)$ 具有以下性质:

1. 对称性: 任意随机变量 X, Y , $\text{Cov}(X, Y) = \text{Cov}(Y, X)$;
2. 单边线性性: 任意随机变量 X, Y , $a, b \in \mathbb{R}$, $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$;

3. 正定性: 任意随机变量 X , $\text{Cov}(X, X) \geq 0$, 且 $\text{Cov}(X, X) = 0$ 当且仅当存在常数 C 使得 $\Pr(X = C) = 1$.

于是, 在差一个常数的意义下, 协方差是一个随机变量空间的内积. 按照内积空间的性质, 随机变量的范数自然就是它的方差.

注. 在命题 C.13 中, 我们使用了 $\Pr(X = C) = 1$ 这样的表达. 在概率论中, 如果一个事件是概率 1 发生的, 我们称之为几乎必然发生. 在涉及与数学期望有关的性质的时候, 我们通常只能在几乎必然的意义下成立, 而不能在一般意义下成立. 比如说, “在差一个常数的意义下, 协方差是一个随机变量空间的内积” 这句话其实并不准确, 严格来说, 应该是“在差一个常数和几乎必然相等的意义下, 协方差是一个随机变量空间的内积”. 也就是说, 如果 $\|X\| = 0$, 那么 X 几乎必然为常数.

协方差与独立性密切关联:

命题 C.14 设 X, Y 是两个随机变量, 如果 X 和 Y 相互独立, 那么 $\text{Cov}(X, Y) = 0$.

我们称 $\text{Cov}(X, Y) = 0$ 的两个随机变量是不相关的, 用内积空间的术语, 不相关的意思就是随机变量正交. 不相关的随机变量不一定是独立的, 但是独立的随机变量一定是不相关的.

协方差的概念可以推广到多个随机变量上:

定义 C.14 (协方差矩阵) 设 X_1, \dots, X_n 是 n 个随机变量, 称他们的 Gram 矩阵为协方差矩阵, 记为 Σ , 其中

$$\Sigma_{ij} = \text{Cov}(X_i, X_j).$$

如果 X 和 Y 分别是 m 维和 n 维随机向量, 那么符号 $\text{Cov}(X, Y)$ 表示的是 $m \times n$ 的矩阵 $(\text{Cov}(X_i, X_j))_{ij}$, 称为 X 和 Y 的协方差矩阵. 特别

地, 如果 $X = Y$, 那么我们记 $\text{Cov}(X, X)$ 为 $\text{Var}(X)$, 称为 X 的协方差矩阵. \square

根据 Gram 矩阵的性质 (命题 A.8), X 的协方差矩阵是一个对称半正定矩阵.

类似地, 我们也可以定义随机向量的数学期望:

定义 C.15 (随机向量的数学期望) 设 $X = (X_1, \dots, X_n)^\top$ 是一个 n 维随机向量, 称

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])^\top$$

为 X 的数学期望. \square

接下来, 我们按照线性代数的思路, 研究线性变换对于期望以及协方差矩阵的影响.

首先是期望, 很容易证明如下的结论:

定理 C.18 设 X 是一个 n 维随机向量, A 是一个 $m \times n$ 的矩阵, 那么 $\mathbb{E}[AX] = A\mathbb{E}[X]$.

接下来是协方差矩阵. 利用 Gram 矩阵与二次型的关系, 我们容易写出如下的结论:

定理 C.19 设 X 是一个 n 维随机向量, A 是一个 $m \times n$ 的矩阵, 那么

$$\text{Var}(AX) = A\text{Var}(X)A^\top.$$

证明. 考虑向量 t , 和 n 维随机向量 Y , $t^\top Y$ 是一个随机变量, 我们可以得到一个二次型

$$g(t) = \text{Var}(t^\top Y) = \text{Cov}(t^\top Y, t^\top Y) = t^\top \text{Var}(Y)t.$$

当 $Y = AX$ 时, 我们有

$$g(t) = \text{Var}(t^T AX) = \text{Var}((A^T t)X) = t^T A \text{Var}(X) A^T t.$$

所以, 对任意 t 都有 $t^T \text{Var}(AX)t = t^T A \text{Var}(X) A^T t$, 所以 $\text{Var}(AX) = A \text{Var}(X) A^T$. \square

上面的计算可以有一个线性代数的理解. 假如说 X_1, \dots, X_n 是线性无关的, 那么 $t^T X$ 可以理解为某个向量在 X_1, \dots, X_n 下的基表示, 于是 t 是坐标. 而 $t^T AX = (A^T t)X$, 因此 A^T 应该理解为某个线性映射 F 在 X_1, \dots, X_n 下的矩阵. Gram 矩阵是二次型 $f(x) = \|x\|^2$ 在 X_1, \dots, X_n 下的矩阵, 因此在 F 的作用下, 二次型的矩阵表示会做一个相应的合同变换, 即 $A \text{Var}(X) A^T$.

§C.3.4 特征函数

在这一部分, 我们讲述随机变量的特征函数, 它是分布的另一种刻画方式.

显然, 特征函数由分布函数决定. 反过来, 特征函数也可以唯一决定分布!

定理 C.20 具有相同特征函数的随机变量(向量)具有相同的分布函数.

特征函数其实可以求出随机变量的分布函数:

定理 C.21 (逆转公式) 设 X 是随机变量, 它的特征函数为 f_X , 分布函数为 F_X , 那么

1. 对于 F 的任意两个连续点 $a < b$,

$$F_X(b) - F_X(a) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} f_X(t) dt.$$

2. 如果 $\int_{\mathbb{R}} |f_X(t)| dt < +\infty$, 那么 X 具有密度 p_X , 且

$$p_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} f_X(t) dt.$$

这一公式也有随机向量的版本:

定理 C.22 (逆转公式, 随机向量版本) 设 X 是 n 维随机向量, 它的特征函数为 f_X , 分布函数为 F_X , 那么

1. 对于 F 的两个点 $a < b$, 满足

$$\Pr(X_1 = c_1, \dots, X_{k-1} = c_{k-1}, X_k \in (a_k, b_k], X_{k+1} = c_{k+1}, \dots, X_n = c_n) = 0,$$

其中 $c_i \in \{a_i, b_i\}$, 我们有

$$\begin{aligned} F_X(b) - F_X(a) \\ = \lim_{T_1, \dots, T_n \rightarrow \infty} \frac{1}{(2\pi)^n} \int_{-T_1}^{T_1} \dots \int_{-T_n}^{T_n} \prod_{k=1}^n \frac{\exp(-it_k a_k) - \exp(-it_k b_k)}{it_k} f_X(t) dt. \end{aligned}$$

2. 如果 $\int_{\mathbb{R}^n} |f_X(t)| dt < +\infty$, 那么 X 具有密度 p_X , 且

$$p_X(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-it^\top x} f_X(t) dt.$$

特征函数特别适合处理独立随机变量的和:

定理 C.23 设 X_1, \dots, X_n 是 n 个相互独立的随机变量, 它们的特征函数分别为 f_1, \dots, f_n , 那么 $X_1 + \dots + X_n$ 的特征函数为 $f_1 \dots f_n$.

比起卷积, 用特征函数来算独立随机变量的和, 方便得多.

特征函数也可以用来判定随机变量的独立性:

定理 C.24 设 X_1, \dots, X_n 是 n 个随机变量, 它们的特征函数分别为 f_1, \dots, f_n , 随机向量 $X = (X_1, \dots, X_n)^\top$, 它的特征函数为 f , 那么 X_1, \dots, X_n 相互独立的充分必要条件是

$$f(t_1, \dots, t_n) = f_1(t_1) \dots f_n(t_n).$$

特征函数的导数可以用来计算随机变量的矩:

定理 C.25 设 X 是一个随机变量, 它的特征函数为 f_X , 那么对任意正整数 k ,

$$\mathbb{E}[X^k] = \frac{f_X^{(k)}(0)}{i^k}.$$

总结起来, 我们之前可以用分布列和密度函数来计算或者判定随机变量的各种性质和特征, 现在都可以用特征函数来处理了.

§C.3.5 条件数学期望

数学期望的定义, 从本质上说, 就是对所有的取值做加权平均. 但是, 有时候我们并不需要对所有的取值做加权平均, 而只需要对某些取值做加权平均. 这时候, 我们就需要引入条件数学期望的概念. 我们从一个直观的例子出发.

例 C.14 一个罐子里有 4 个红球, 2 个灰球, 4 个白球. 红球, 灰球和白球的分数分别是 4, 2, 1. 现在随机抽一个球, 抽球人戴着黑白滤镜的眼

镜观察球的颜色, 他不能分辨红球和灰球, 但是可以区分这两种球和白球. 那么, 在他观察过这个球之后, 期望上得到的分数是多少?

和条件概率有类似的情况, 此时并不完全是纯随机的, 因为抽球人可以区分一些东西. 于是, 样本空间可以分成两个部分, 一个是 $A_1 = \{r, g\}$, 即抽到的球是红球或灰球; 另一个是 $A_2 = \{w\}$, 即抽到的球是白球. 在第一种情况下, 期望上的分数是

$$4 \cdot \Pr(\{s\}|A_1) + 2 \cdot \Pr(\{g\}|A_1) = 3.$$

在第二种情况下, 期望上的分数是

$$1 \cdot \Pr(\{w\}|A_2) = 1.$$

□

更一般地, 考虑样本空间 Ω , 事件 A_1, \dots, A_n , 它们两两互斥, 且 $\bigcup_{i=1}^n A_i = \Omega$, 这形成了 Ω 的一个分割, 记为 \mathcal{A} . 我们再假设 $\Pr(A_i) > 0$, 我们有如下定义:

定义 C.16 (基于分割的条件数学期望) 设 X 是一个随机变量, $\mathcal{A} = \{A_1, \dots, A_n\}$ 是 Ω 的一个分割, 满足 $\Pr(A) > 0$ 对任意 $A \in \mathcal{A}$ 成立. X 在 \mathcal{A} 上的条件数学期望是一个随机变量, 记为 $\mathbb{E}[X|\mathcal{A}]$, 它的定义为

$$\mathbb{E}[X|\mathcal{A}](\omega) = \sum_{i=1}^n \frac{\mathbb{E}[XI_{A_i}]}{\Pr(A_i)} I_{A_i}(\omega).$$

□

这个定义就是在说, 当 ω 落在分割的某个集合 A_i 上时, 我们按照 A_i 上的条件概率算期望. 记号 $\frac{\mathbb{E}[XI_{A_i}]}{\Pr(A_i)}$ 也记为 $\mathbb{E}[X|A_i]$, 它的含义可以从

$X = I_B$ 来理解:

$$\frac{\mathbb{E}[I_B I_{A_i}]}{\Pr(A_i)} = \frac{\Pr(A_i B)}{\Pr(A_i)} = \Pr(B|A_i).$$

这一计算对示性函数解释了“按照 A_i 上的条件概率算期望”. 按照随机变量数学期望的定义, 这一理解可以推广到一般随机变量.

条件数学期望是一个随机变量, 意思就是我们能够消除某些不确定性. 在求数学期望的时候, 我们完全不知道样本 ω 落在哪里, 所以只能对整个 Ω 有一个预期. 在求对分割的条件数学期望的时候, 我们能够知道 ω 落在了某个 A_i 中, 因此我们的不确定性只在于 A_i 上, 所以我们可以只对 A_i 中的 ω 有一个预期.

下面我们推广这一定义. 注意到, 分割 \mathcal{A} 其实生成了一个 Ω 的 σ -代数, 即 $\sigma(\mathcal{A})$, 它是包含 A_1, \dots, A_n 的最小 σ -代数. 容易验证, 这一 σ -代数里的集合都是若干个 A_i 的并形成的. 分割里的事件代表了我们可以感知到的最小事件, 而 σ -代数里的事件代表了我们可以感知到的事件的集合.

取 $A \in \sigma(\mathcal{A})$, 要如何计算 X 在 A 上的期望呢? 我们有两种方式, 第一种, 直接计算: $\mathbb{E}[XI_A]$. 第二种, 我们将 A 写成 $A = \bigcup_{i=1}^k A_{n_i}$. 在每个 A_i 上, 我们知道期望是 $\mathbb{E}[XI_A|A_i]$. 而落到 A_i 上的概率是 $\Pr(A_i)$, 于是, 按照数学期望加权平均的直觉, X 在 A 上的期望应该是

$$\sum_{i=1}^k \mathbb{E}[XI_A|A_i] \Pr(A_i).$$

这正好就是随机变量 $\mathbb{E}[XI_A|\mathcal{A}]$ 的数学期望 $\mathbb{E}[\mathbb{E}[XI_A|\mathcal{A}]]$.

对任意 $A \in \sigma(\mathcal{A})$, 这两种计算方式都应该相等:

$$\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[XI_A|\mathcal{A}]]. \quad (\text{C.4})$$

这给了我们一般情况下的条件数学期望的定义:

定义 C.17 (基于 σ -代数的条件数学期望) 设 X 是一个非负随机变量, \mathcal{G} 是 Ω 的一个 σ -代数, 随机变量 $\mathbb{E}[X|\mathcal{G}]$ 被称为 X 关于 \mathcal{G} 的条件数学期望, 如果它满足

1. 对任意 $B \in \mathcal{B}(\mathbb{R})$, $\{\mathbb{E}[X|\mathcal{G}] \in B\}$ 是 \mathcal{G} -可测的;
2. 对任意 $A \in \mathcal{G}$, $\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]I_A]$.

设 X 是一个一般的随机变量, 如果

$$\min\{\mathbb{E}[X^+|\mathcal{G}], \mathbb{E}[X^-|\mathcal{G}]\} < +\infty,$$

那么 X 关于 \mathcal{G} 的条件数学期望是一个随机变量, 记为 $\mathbb{E}[X|\mathcal{G}]$, 定义为

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X^+|\mathcal{G}] - \mathbb{E}[X^-|\mathcal{G}].$$
□

这一定义分成两部分, 这类似于我们在定义数学期望时候做的事情: 先定义非负的情况, 再定义一般情况. 对于非负随机变量的定义, 第一条要求是说, “ $\mathbb{E}[X|\mathcal{G}]$ 落在合理的值集上” 这件事情是可以用 \mathcal{G} 中事件描述的, 这和随机变量的定义是类似的; 而第二条则反映了“条件”的性质, 也就是我们刚刚讨论的 (C.4) 式.

定义中的 \mathcal{G} 可以理解为我们观测样本的能力. \mathcal{G} 越大, 则越能确定 ω 具体的范围, 所以条件期望就越像 $X(\omega)$; \mathcal{G} 越小, 则越不能确定 ω

具体的范围, 所以条件期望就越像 $\mathbb{E}[X]$.

注意, 基于 σ -代数的条件数学期望和基于分割的条件数学期望是一致的, 所以这一定义是合理的.

最后, 随机向量也是可以诱导条件数学期望的:

定义 C.18 (随机向量诱导的 σ -代数) 设 X 是一个 n 维随机向量, 那么 X 诱导的 σ -代数是 Ω 的一个 σ -代数, 记为 $\sigma(X)$, 它的元素为 $\{X \in B\}$, 其中 $B \in \mathcal{B}(\mathbb{R}^n)$. \square

我们说过, $\{X \in B\}$ 表示 “ X 落在合理的值集上”. 在之前定义随机变量的时候, 我们要求取合理的值集是一个事件, 这里则是更加简单粗暴, 我们直接定义 $\{X \in B\}$ 是一个事件. 接下来, 我们可以定义随机向量诱导的条件数学期望:

定义 C.19 (随机向量诱导的条件数学期望) 设 X 是一个随机变量, Y 是一个随机向量, 那么 X 关于 Y 的条件数学期望是一个随机变量, 记为 $\mathbb{E}[X|Y]$, 定义为 $\mathbb{E}[X|\sigma(Y)]$. \square

我们之前定义过条件分布 $\Pr(X \leq x|Y = y)$, 利用这一分布, 我们可以求出一个条件数学期望 $\mathbb{E}[X|Y = y]$. 下面的命题表明, 这一定义和定义 C.19 是相容的:

命题 C.15 设 X 是一个随机变量, Y 是一个 n 维随机向量, 那么存在一个 Borel 函数 $g : \mathbb{R}^n \rightarrow \mathbb{R}$, 使得对任意 $\omega \in \Omega$, 有

$$\mathbb{E}[X|Y](\omega) = g(Y(\omega))$$

并且

$$\mathbb{E}[X|Y = y] = g(y).$$

我们不满足于 $\mathbb{E}[X|Y = y]$, 而是费尽周章定义条件期望 $\mathbb{E}[X|\mathcal{Y}]$, 是因为他通常来说更好用, 特别是在随机过程中, 它能给出很多公式直观上的含义. 这一点在第二章中会有很多体现.

接下来我们讨论条件数学期望的性质, 我们依然只列举而不证明.

命题 C.16 设 $(\Omega, \mathcal{F}, \Pr)$ 是概率空间, $\mathcal{G} \subseteq \mathcal{F}$ 是 Ω 的一个 σ -代数, 那么

1. 期望的线性性: 设 X, Y 是随机变量, $a, b \in \mathbb{R}$, 如果 $\mathbb{E}[X|\mathcal{G}]$ 和 $\mathbb{E}[Y|\mathcal{G}]$ 都存在, 那么 $\mathbb{E}[aX + bY|\mathcal{F}]$ 存在, 且

$$\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}].$$

2. 单调性: 设 X, Y 是随机变量, 如果 $X \leq Y$, 那么

$$\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}].$$

3. 绝对值不等式: 设 X 是随机变量, 那么

$$\mathbb{E}[|X||\mathcal{G}] \geq |\mathbb{E}[X]|\mathcal{G}|.$$

4. 如果 $\mathcal{G} = \{\emptyset, \Omega\}$, 那么

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X].$$

5. 望远性: 设 X 是随机变量, 如果 $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{F}$ 都是 Ω 的 σ -代数,

且 $\mathcal{G}_1 \subseteq \mathcal{G}_2$, 那么

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2] = \mathbb{E}[X|\mathcal{G}_1].$$

6. 重期望公式: 设 X 是随机变量, 那么

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X].$$

7. 设 X, Y 是随机变量, 如果 $\sigma(Y) \subseteq \mathcal{G}$, 那么

$$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}].$$

我们主要需要解释的是望远性. 可以把 σ -代数理解成观测的能力, 这一代数越大, 观测的越细致. 望远性的意思就是, 如果我们用两次观测能力强弱不同的 σ -代数观测 X , 那么最终的结果只取决于最粗糙的那个 σ -代数.

另外, 重期望公式本质上就是期望版本的全概率公式 (定理 C.1). 从基于分割的条件数学期望的角度来看, 这件事会更明显. 假设我们有一个分割 $\mathcal{A} = \{A_1, \dots, A_n\}$, 并且 $\Pr(A_i) > 0$, 那么

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{A}]] = \sum_{i=1}^n \mathbb{E}[X|A_i] \Pr(A_i).$$

最后, 性质 7 是在说, 如果 Y 是 \mathcal{G} -可测的 (也就是我们用 \mathcal{G} 可以完全确定 Y), 那么求条件期望的时候 Y 就相当于一个常数, 可以提到期望的外面.

§C.4 多元正态分布 (Gauss 向量)

在这一节中，我们利用附录 C.3.3 和附录 C.3.4 中的工具，来研究多元正态分布。

多元正态分布的定义在附录 C.2.4 中已经给出，首先，我们不加证明地给出它的特征函数：

定理 C.26 设 $\mu \in \mathbb{R}^n$, Σ 是一个 $n \times n$ 的对称正定矩阵，那么随机向量 $X \sim \mathcal{N}(\mu, \Sigma)$ 的特征函数为

$$f_X(t) = \exp\left(it^\top \mu - \frac{1}{2}t^\top \Sigma t\right).$$

利用 (4.4)，我们可以计算出多元正态分布的期望和协方差矩阵：

命题 C.17 设 $X \sim \mathcal{N}(\mu, \Sigma)$, 那么

$$\mathbb{E}[X] = \mu, \quad \text{Var}(X) = \Sigma.$$

现在我们将这一定义推广。注意到， Σ 就是 X 的协方差矩阵，所以定理 C.26 中的 Σ 并不要求正定，只要半正定就可以定义一个特征函数了。我们将这一定义推广到半正定矩阵的情形：

定义 C.20 (Gauss 向量) 设 $\mu \in \mathbb{R}^n$, Σ 是一个 $n \times n$ 的对称半正定矩阵，如果随机向量 X 的特征函数为

$$f_X(t) = \exp\left(it^\top \mu - \frac{1}{2}t^\top \Sigma t\right),$$

那么称 X 是一个 **Gauss 向量**，记为 $X \sim \mathcal{N}(\mu, \Sigma)$. □

如果 Σ 退化, X 不能写出密度, 所以也不是连续型随机向量. 但是, 利用特征函数, 我们依然可以研究 X 的性质. 特别是命题 C.17, 对于 Gauss 向量仍然成立.

Gauss 向量可以完全由它的期望和协方差矩阵刻画. 首先, Gauss 向量的独立性等价于不相关性:

定理 C.27 设 $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu, \Sigma)$, 那么 X_1, \dots, X_n 相互独立的充分必要条件是 X_1, \dots, X_n 两两不相关, 即 Σ 是一个对角矩阵.

需要注意的是, 如果 X 是正态分布, Y 是正态分布, 这并不意味着 (X, Y) 是 Gauss 向量, 因而并不能用不相关来作为独立性的判据. 因此, 在一般情况下, 我们必须要验证 (X_1, \dots, X_n) 是 Gauss 向量, 然后才能断言不相关等价于独立.

当然, 这一判据可以推广到多个 Gauss 向量的情形:

推论 C.3 设 X_1, \dots, X_n 是 n 个 Gauss 向量, 它们相互独立的充分必要条件是 X_1, \dots, X_n 两两不相关, 即协方差矩阵 $\text{Cov}(X_i, X_j) = O$, $i \neq j$.

其次, 利用定理 C.18 和定理 C.19, 我们可以得到如下的结论:

定理 C.28 设 $X \sim \mathcal{N}(\mu, \Sigma)$, A 是一个 $m \times n$ 的矩阵, 那么 $AX \sim \mathcal{N}(A\mu, A\Sigma A^\top)$.

取特定的 A , 我们可以得到一个实用的推论: Gauss 向量的子向量仍然是 Gauss 向量, 也就是说, 取 $X = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$, 那么对任意的 $1 \leq k \leq n$, $i_1, \dots, i_k \in \{1, \dots, n\}$, $(X_{i_1}, \dots, X_{i_k})^\top$ 也是 Gauss 向量.

§C.5 大数定律

接下来我们讨论大数定律. 大数定律来自人类对随机现象的直观认识, 它表明, 如果重复抛一枚公平的硬币, 那么正面朝上的次数会趋于总次数的一半.

首先, 我们要定义, 对于一系列随机的实验, 什么叫做“趋于”. 比如说, 抛一万次硬币, 当然有正的概率会出现一万次正面朝上, 但是这件事情在现实中是不太可能的. 我们要定义一个概率的概念, 来描述这种“趋于”.

定义 C.21 (依概率收敛) 设 X_1, X_2, \dots 是一系列随机变量, X 是一个随机变量. 如果对任意 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \varepsilon) = 0,$$

那么称 X_n 依概率收敛到 X , 记为 $X_n \xrightarrow{P} X$. □

换句话说, 依概率收敛的意思就是, 当 n 趋于无穷, X_n 和 X 有任何固定偏差的概率趋于 0.

接下来, 我们给出大数定律的一个形式化描述:

定理 C.29 (Khinchin 大数定律) 设 X_1, X_2, \dots 是一系列独立同分布的随机变量, $\mathbb{E}[X_i]$ 存在, 那么

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X_1].$$

这一定律说明, 独立重复试验的结果会趋于它们的期望.

证明. 由于 X_1, X_2, \dots 是独立同分布的, 所以 $\mathbb{E}[X_1], \mathbb{E}[X_2], \dots$ 都是相等的, 记为 μ . 令

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

我们计算 S_n 的特征函数:

$$\begin{aligned} f_{S_n}(t) &= \mathbb{E}[e^{itS_n}] \\ &= \mathbb{E}[e^{it/n(X_1 + \dots + X_n)}] \\ &= \mathbb{E}[e^{it/nX_1} \dots e^{it/nX_n}] \\ &= \mathbb{E}[e^{it/nX_1}] \dots \mathbb{E}[e^{it/nX_n}] \\ &= \left(\mathbb{E}[e^{it/nX_1}] \right)^n \\ &= \left(f_{X_1} \left(\frac{t}{n} \right) \right)^n. \end{aligned}$$

根据定理 C.25, 我们有

$$f_{X_1}(t) = 1 + it\mu + o(t), \quad t \rightarrow 0.$$

于是

$$f_{S_n}(t) = \left(1 + \frac{it\mu}{n} + o\left(\frac{t}{n}\right) \right)^n \rightarrow e^{it\mu}, \quad n \rightarrow \infty.$$

因此, 根据定理 C.20, S_n 的分布趋于一个在 μ 处集中的退化分布, 根据定义, 这意味着对任意 x , 当 $n \rightarrow \infty$ 时,

$$\Pr(S_n \leq x) \rightarrow \begin{cases} 0, & x < \mu, \\ 1, & x \geq \mu. \end{cases}$$

于是，对于任意 $\varepsilon > 0$ ，当 $n \rightarrow \infty$ 时，

$$\begin{aligned}\Pr(|S_n - \mu| > \varepsilon) &\leq \Pr(S_n < \mu - \varepsilon) + \Pr(S_n > \mu + \varepsilon) \\&= \Pr(S_n < \mu - \varepsilon) + 1 - \Pr(S_n \leq \mu + \varepsilon) \\&\rightarrow 0.\end{aligned}$$

这就完成了证明. □

参考文献

- [Bar80] Maya Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, May 1980.
- [Bel52] Richard Bellman. On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719, August 1952.
- [BHO75] P. J. Bickel, E. A. Hammel, and J. W. O’Connell. Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187(4175):398–404, February 1975.
- [BP66] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966.
- [Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

- [Bro71] L. D. Brown. Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *The Annals of Mathematical Statistics*, 42(3):855–903, June 1971.
- [BT86] M. J. Bishop and E. A. Thompson. Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology*, 190(2):159–165, July 1986.
- [BZ12] Lawrence D. Brown and Linda H. Zhao. A Geometrical Explanation of Stein Shrinkage. *Statistical Science*, 27(1):24–30, 2012.
- [Cha] ChatGPT | OpenAI. <https://openai.com/chatgpt/overview/>. (accessed 2025-02-13).
- [Cis07] Paul Cisek. Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1485):1585–1599, September 2007.
- [Cox46] R. T. Cox. Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14(1):1–13, January 1946.
- [CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.
- [de 17] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. Wiley Series in Probability and Statistics. Wiley, 1 edition, February 2017.

- [DeL90] Mahlon R. DeLong. Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, 13(7):281–285, July 1990.
- [DF80] P. Diaconis and D. Freedman. Finite Exchangeable Sequences. *The Annals of Probability*, 8(4):745–764, August 1980.
- [Dur19] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 5 edition, 2019.
- [GIP18] Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun Ph.D. *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton & Company, New York, 5th ed. edition, 2018.
- [GPM⁺14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, volume 2 of *NIPS’14*, pages 2672–2680, Cambridge, MA, USA, December 2014. MIT Press.
- [H AJ90] Xuedong Huang, Yasuo Ariki, and Mervyn Jack. *Hidden Markov Models for Speech Recognition*. Columbia University Press, USA, 1990.

- [Ham88] A. G. Hamilton. *Logic for Mathematicians*. Cambridge University Press, October 1988.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
- [Há23] Alan Hájek. Interpretations of Probability. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2023 edition, 2023.
- [Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.
- [JBM75] F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256, May 1975.
- [JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [JS61] W. James and Charles Stein. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contribu-*

- tions to the Theory of Statistics*, volume 4.1, pages 361–380. University of California Press, January 1961.
- [KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [KS06] G. Jay. Kerns and Gábor J. Székely. Definetti’s Theorem for Abstract Finite Exchangeable Sequences. *Journal of Theoretical Probability*, 19(3):589–608, December 2006.
- [KT73] Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237–251, 1973.
- [KW22] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022.
- [LDK95] Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, pages 394–402, San Francisco, CA, USA, August 1995. Morgan Kaufmann Publishers Inc.
- [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.

- [LPW08] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island, 第 1st 版 edition, 2008.
- [LV19] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2019.
- [McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.
- [Nor97] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997.
- [NS97] Anil Nerode and Richard A. Shore. *Logic for Applications*. Springer, New York, NY, 1997.
- [Pac] Eric Pacuit. Reasoning for Humans: Clear Thinking in an Uncertain World. <https://phil1171.org/>. (accessed 2025-02-13).
- [RNSS] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

- [Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.
- [SAH⁺20] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020.
- [SAN96] Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928, December 1996.
- [SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018.
- [SCdP10] Manuela Sellitto, Elisa Ciaramelli, and Giuseppe di Pellegrino. Myopic Discounting of Future Rewards after Medial Orbitofrontal Damage in Humans. *Journal of Neuroscience*, 30(49):16429–16436, December 2010.
- [SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 1067, pages 11918–11930. Curran Associates Inc., Red Hook, NY, USA, December 2019.

- [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.
- [SHM⁺16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.
- [Sim51] E. H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951.
- [Sim81] Herbert A. Simon. Information-processing models of cognition. *Journal of the American Society for Information Science*, 32(5):364–377, 1981.
- [Sky75] Brian Skyrms. *Choice and Chance : An Introduction to Inductive Logic*. Encino, Calif. : Dickenson Pub. Co., 1975.

- [Spi24] David Spiegelhalter. Why probability probably doesn't exist (but it is useful to act like it does). *Nature*, 636(8043):560–563, December 2024.
- [Ste56] Charles Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3.1, pages 197–207. University of California Press, January 1956.
- [SWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015.
- [Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.
- [TK74] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974.
- [TK83] Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

- [Tol10] Richard C. Tolman. *The Principles of Statistical Mechanics.* Dover Publications, New York, NY, 2010.
- [Tom22] Michael Tomasello. *The Evolution of Agency.* The MIT Press, September 2022.
- [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.
- [Und] Understanding Stein’s paradox. <https://joe-antognini.github.io/machine-learning/steins-paradox>. (accessed 2025-02-14).
- [VBC⁺19] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019.
- [VD13] Dirk Van Dalen. *Logic and Structure*. Universitext. Springer, London, 2013.
- [Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge

- Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.
- [Vit67] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.
- [VWOL24] Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, February 2024.
- [Wat89] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.
- [WG61] Richard D. Walk and Eleanor J. Gibson. A comparative and analytical study of visual depth perception. *Psychological Monographs: General and Applied*, 75(15):1–44, 1961.
- [Zor16] Vladimir A. Zorich. *Mathematical Analysis II*. Universitext. Springer, Berlin, Heidelberg, 2016.
- [李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.
- [苏 a] 剑林苏. 最小熵原理 (六) : 词向量的维度应该怎么选择? . <https://kexue.fm/archives/7695>. (accessed 2025-02-14).

- [苏 b] 剑林苏. 让人惊叹的 Johnson-Lindenstrauss 引理: 应用篇. <https://kexue.fm/archives/8706>. (accessed 2025-02-14).
- [苏 c] 剑林苏. 让人惊叹的 Johnson-Lindenstrauss 引理: 理论篇. <https://kexue.fm/archives/8679>. (accessed 2025-02-14).