标题 **title**

作者 *author*

2023 年 9 月 4 日

前言

目录

前言 i

第一部分　科学的逻辑 1

第一章　合情推理 2

§1.1 回顾：命题逻辑的演绎推理 2

§1.2 合情推理的数学模型 4

1.2.1　似然，合情推理的原则 4

1.2.2　似然与概率 6

§1.3 合情推理的归纳强论证 8

1.3.1　先验与基率谬误 8

1.3.2　归纳强论证 9

1.3.3　有效论证和归纳强论证的比较 12

§1.4 数据驱动的科学理论 14

第二章　Markov 链与决策 17

§2.1 Markov 链 17

§2.2 Markov 奖励过程（MRP） 21

§2.3 Markov 决策过程（MDP） 24

§2.4 隐 Markov 模型（HMM） 28

2.4.1　评估问题 29

2.4.2　解释问题 30

第二部分　信息与数据 **32**

第三章　信息论基础 **33**
§3.1 熵 33
3.1.1 概念的导出 33
3.1.2 概念与性质 36
3.1.3 熵与通信理论 41
§3.2 Kullback-Leibler 散度 44
3.2.1 定义 44
3.2.2 两个关于信息的不等式 46
3.2.3 在机器学习中的应用：语言生成模型 47
§3.3 附录：Shannon 定理的证明 48
§3.4 习题 49
§3.5 章末注记 51

第四章　Johnson-Lindenstrauss 引理 **53**
§4.1 机器学习中的数据 53
§4.2 矩法与集中不等式 54
§4.3 J-L 引理的陈述与证明 58
§4.4 J-L 引理的应用 62
§4.5 习题 63
§4.6 章末注记 63

第五章　差分隐私 **64**
§5.1 数据隐私问题 64
§5.2 差分隐私的定义与性质 66
§5.3 差分隐私的应用 70
5.3.1 随机反应算法 70
5.3.2 全局灵敏度与 Laplace 机制 71
5.3.3 DP 版本 Llyod 算法 73
§5.4 差分隐私与信息论 74
§5.5 习题 75
§5.6 章末注记 75

第三部分　决策与优化 **76**

第六章　凸分析 **77**
§6.1 决策与优化的基本原理 77
6.1.1 统计决策理论 77
6.1.2 优化问题 78
6.1.3 例子：网格搜索算法 81
§6.2 凸函数 83
§6.3 凸集 86
6.3.1 基本定义和性质 86
6.3.2 分离超平面定理 88

第七章　对偶理论 **90**
§7.1 条件极值与 Lagrange 乘子法 91
§7.2 Karush–Kuhn–Tucker 条件 94
§7.3 Lagrange 对偶 97
7.3.1 Lagrange 定理 97
7.3.2 弱对偶定理，强对偶定理 101
§7.4 应用：支持向量机（SVM） 105

第八章　不动点理论 **108**
§8.1 Banach 不动点定理 108
§8.2 Brouwer 不动点定理 111
§8.3 不动点的一般视角 114

第四部分　逻辑与博弈 **115**

第九章　动态博弈 **116**
§9.1 输赢博弈 116
§9.2 随机博弈（Markov 博弈） 121

第十章　静态博弈 **127**
§10.1 正则形式博弈 127
10.1.1 生成对抗网络 128
10.1.2 混合策略 130

§10.2 不完全信息博弈（Bayes 博弈） 131

第五部分　认知逻辑 **136**

第十一章　模态逻辑基础 **137**

§11.1 模态逻辑的起源 137

11.1.1 三段论 137

11.1.2 非经典逻辑 138

§11.2 模态语言 139

§11.3 Kripke 语义与框架语义 142

§11.4 模态可定义性 147

第十二章　认知逻辑与共同知识 **149**

§12.1 “泥泞的孩童”谜题 149

§12.2 认知逻辑的基本模型与性质 151

12.2.1 “泥泞的孩童”再回顾 155

12.2.2 Aumann 结构 156

§12.3 对不一致达成一致 157

§12.4 Rubinstein 电子邮件博弈 160

第六部分　附录 **164**

附录 A　线性代数基础 **165**

§A.1 线性空间 165

§A.2 线性映射 169

§A.3 矩阵 174

§A.4 双线性型与二次型 180

§A.5 带内积的线性空间 184

§A.6 行列式 190

§A.7 算子范数与谱理论 193

附录 B　微分学基础 **199**

§B.1 点集拓扑 199

B.1.1 度量空间，范数 199

B.1.2 开集与闭集 . . . 202
B.1.3 紧致性，收敛性，完备性 . . . 205
B.1.4 连续映射 . . . 208
B.1.5 与实数序有关的性质 . . . 211
§B.2 一元函数的微分学 . . . 213
B.2.1 导数与微分的定义 . . . 214
B.2.2 微分学基本定理 . . . 217
§B.3 多元函数的微分学 . . . 219
B.3.1 微分、偏导数与导数的定义 . . . 219
B.3.2 微分学基本定理 . . . 225
B.3.3 隐函数定理 . . . 227

附录 C 概率论基础 231
§C.1 从朴素概率论到公理化概率论 . . . 231
C.1.1 Kolmogorov 概率论 . . . 231
C.1.2 条件概率，独立性 . . . 235
§C.2 随机变量，分布函数 . . . 239
C.2.1 基本定义 . . . 239
C.2.2 离散型随机变量 . . . 243
C.2.3 连续型随机变量 . . . 243
C.2.4 随机向量，条件分布，独立性 . . . 247
C.2.5 随机变量（向量）的函数 . . . 251
§C.3 随机变量的数字特征，条件数学期望 . . . 254
C.3.1 数学期望，Lebesgue 积分 . . . 254
C.3.2 数学期望的性质 . . . 258
C.3.3 随机变量的内积空间 . . . 261
C.3.4 特征函数 . . . 263
C.3.5 条件数学期望 . . . 264
§C.4 多元正态分布（Gauss 向量） . . . 268

第一部分

科学的逻辑

第二部分

信息与数据

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

第六部分

附录

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）：

$$\ell^2 = \left\{ (x_1, x_2, \cdots) \in \mathbb{C}^\infty : \sum_{i=1}^{\infty} x_i^2 < \infty \right\}.$$

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 $ab + c = (ab) + c$.

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 $a \cdot x + y = (a \cdot x) + y$.

“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. □

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，$x_1, \cdots, x_n \in V$，$a_1, \cdots, a_n \in F$，则称 $a_1x_1 + \cdots + a_nx_n$ 是 $x_1, \cdots, x_n$ 的一个**线性组合**.

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**.

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 $V_1, \cdots, V_n$ 是 F-线性空间，则 $V_1 \times \cdots \times V_n$ 是一个 F-线性空间，其中加法和数乘分别定义为

$$(x_1, \cdots, x_n) + (y_1, \cdots, y_n) = (x_1 + y_1, \cdots, x_n + y_n),$$
$$a(x_1, \cdots, x_n) = (ax_1, \cdots, ax_n).$$

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 $\mathrm{Span}(S)$.

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 $x_1, \cdots, x_n \in S$，$a_1, \cdots, a_n \in F$，使得 $a_1x_1 + \cdots + a_nx_n = 0$，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.

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 $\mathrm{Span}(S) = V$，则称 S 是 V 的一个**基**.

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 $\{e_1, \cdots, e_n\}$，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 $\{1\}$. 但如果我们把 $\mathbb{C}$ 看成 $\mathbb{R}$ 的线性空间，那么 $\mathbb{C}$ 的基就是 $\{1, i\}$.
- $M_{m\times n}(F)$ 的标准基是 $\{E_{ij} : 1 \le i \le m, 1 \le j \le n\}$，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 $\{e_1, e_2, \cdots\}$，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 $x = a_1e_1 + a_2e_2 + a_3e_3$，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**.

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 $S_1 \cap S_2 = \varnothing$. 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

$$\dim V = |S_1 \cup S_2| = |S_1| + |S_2| = \dim U_1 + \dim U_2. \qquad \square$$

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 $f: V \to W$ 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**.

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：$f: V \to W$，$f(x) = 0$，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：$f: V \to V$，$f(x) = x$，这也是线性映射，我们通常记为 id.

线性映射有如下基本性质：

命题 A.1 *设 $f: V \to W$ 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 $g: W \to Z$ 是线性映射，则 $g \circ f: V \to Z$ 也是线性映射；*
5. *如果 $g: V \to W$ 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 $g: V \to W$ 是线性映射，$h: W \to Z$ 是线性映射，$k: Z \to V$ 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 $G(x) = a_0 + a_1 x + \cdots + a_n x^n$，我们可以定义一个新的线性映射 $G(f) = a_0 \mathrm{id} + a_1 f + \cdots + a_n f^n$.

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 $\mathrm{d}/\mathrm{d}x : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ 被称为**微分算子**. 容易验证，$\mathrm{d}/\mathrm{d}x$ 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

$$\pi_m : (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_m, 0, \ldots, 0)$$

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 $(x_1, \ldots, x_m, 0, \ldots, 0)$ 后面的 0 都丢掉，这就是一个 $\mathbb{R}^m$ 的元素.

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 $(x_1, \ldots, x_m, 0, \ldots, 0)$ 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 $f : V \to W$ 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**.

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.

$\Longleftarrow$：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 $\{v_1, \ldots, v_n\}$，W 的基是 $\{w_1, \ldots, w_n\}$，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.

$\Longrightarrow$：设两个线性空间由映射 $f : V \to W$ 给出同构. 假设 V 的基是 $\{v_1, \ldots, v_n\}$，我们证明 W 的基就是 $\{f(v_1), \ldots, f(v_n)\}$.

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 $\{f(v_1), \ldots, f(v_n)\}$，所以 $\dim V = \dim W$. □

此外，同构还有一个重要性质：

命题 A.2 *假设 $f : V \to W$ 和 $g : W \to U$ 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. □

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，$f : V \to W$ 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 $\operatorname{Im} f = \{f(x) : x \in V\}$.

“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，$f : V \to W$ 是一个线性映射，则 $\ker f$ 是 V 的线性子空间，$\operatorname{Im} f$ 是 W 的线性子空间. 另外，

$$\dim V = \dim \ker f + \dim \operatorname{Im} f.$$

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 $\operatorname{Im} f$.

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 $v_1, \ldots, v_k$，然后添加向量 $u_1, \ldots, u_l$ 扩充成 V 的基，然后证明 $f(u_1), \ldots, f(u_l)$ 是 $\operatorname{Im} f$ 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，$f : V \to W$ 是一个线性映射. 那么以下性质成立：

1. $\dim \operatorname{Im} f \leq \dim V$，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 $\dim \operatorname{Im} f = \dim W$；
4. f 是同构当且仅当 $\ker f = \{0\}$ 且 $\dim \operatorname{Im} f = \dim W$.

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，$f : V \to W$ 是一个线性映射. f 的**秩**定义为 $\operatorname{rank} f = \dim \operatorname{Im} f$.

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，$f : V \to W$ 和 $g : W \to U$ 是两个线性映射，则 $\operatorname{rank}(g \circ f) \leq \operatorname{rank} f$，等号成立当且仅当 $\operatorname{Im} f \cap \ker g = \{0\}$. 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，$\operatorname{rank}(g \circ f) \leq \operatorname{rank} g$，如果 f 是满射，那么等号成立.

证明. 根据推论 A.1，我们有 $\operatorname{rank}(g\circ f)=\dim\operatorname{Im}(g\circ f)\leq\dim\operatorname{Im}f=\operatorname{rank}f$. 等号成立当且仅当在空间 $\operatorname{Im}f$ 中 g 的核是 $\{0\}$，换言之，$\operatorname{Im}f\cap\ker g=\{0\}$. 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 $\operatorname{Im}f\subseteq W$，所以 $\operatorname{Im}(g\circ f)\subseteq\operatorname{Im}g$，因此 $\operatorname{rank}(g\circ f)=\dim\operatorname{Im}(g\circ f)\leq\dim\operatorname{Im}g=\operatorname{rank}g$. 如果 f 是满射，那么 $\operatorname{Im}f=W$，所以 $\operatorname{Im}(g\circ f)=\operatorname{Im}g$，因此等号成立. □

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 $f:V\to W$，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

$$A=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&\ddots&\vdots\\a_{m1}&a_{m2}&\cdots&a_{mn}\end{pmatrix}.$$

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 $(a_{1j},\dots,a_{mj})^{\mathsf{T}}$. 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 $X=(x_1,\dots,x_n)^{\mathsf{T}}$. 我们现在来计算 $f(x)$ 在基 w_i 下

的坐标 $Y=(y_1,\dots,y_m)^{\mathsf{T}}$. 因为 $x=x_1v_1+\cdots+x_nv_n$，所以

$$\begin{aligned}f(x)&=f(x_1v_1+\cdots+x_nv_n)\\&=x_1f(v_1)+\cdots+x_nf(v_n)\\&=x_1(a_{11}w_1+\cdots+a_{m1}w_m)+\cdots+x_n(a_{1n}w_1+\cdots+a_{mn}w_m)\\&=(a_{11}x_1+\cdots+a_{1n}x_n)w_1+\cdots+(a_{m1}x_1+\cdots+a_{mn}x_n)w_m.\end{aligned}$$

因此，$f(x)$ 在基 w_i 下的坐标是 $(a_{11}x_1+\cdots+a_{1n}x_n,\dots,a_{m1}x_1+\cdots+a_{mn}x_n)^{\mathsf{T}}$，$y_i=a_{i1}x_1+\cdots+a_{in}x_n$. 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：

$$\begin{array}{ccc}x & \xmapsto{f} & f(x)\\ \big\downarrow & & \big\downarrow\\ X & \xmapsto{A} & Y\end{array}$$

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 $f_A:F^n\to F^m$：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，$f:V\to V$ 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 $X=(x_1,\dots,x_n)^{\mathsf{T}}$，在基 $\{v_i'\}$ 下的坐标是 $X'=(x_1',\dots,x_n')^{\mathsf{T}}$. 我们来计算 x 在基 $\{v_i'\}$ 下的坐标. 因为 $f(x)=f(x_1v_1+\cdots+x_nv_n)=x_1f(v_1)+\cdots+x_nf(v_n)$，而

$$f(v_i)=v_i'=\sum_{j=1}^{n}a_{ji}v_j,$$

所以

$$f(x)=\sum_{i=1}^{n}x_i\sum_{j=1}^{n}a_{ji}v_j=\sum_{j=1}^{n}\left(\sum_{i=1}^{n}a_{ji}x_i\right)v_j\implies$$
$$x=\sum_{j=1}^{n}\left(\sum_{i=1}^{n}a_{ji}x_i\right)f^{-1}(v_j)=\sum_{j=1}^{n}\left(\sum_{i=1}^{n}a_{ji}x_i\right)v'_j.$$

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，$f:V\to W$ 和 $g:V\to W$ 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，$f:V\to W$ 和 $g:W\to U$ 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 $\mathrm{id}:V\to V$，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

$$\begin{pmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{pmatrix}.$$

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 $f:V\to W$ 将 v_i 映到 $\lambda_i w_i$，那么 f 的矩阵就是

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

我们将这样的矩阵称为对角矩阵.

对于一般的同构映射 $f:V\to W$，它有一个同构逆映射 $f^{-1}:W\to V$. 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 $f^{-1}f=\mathrm{id}$，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 $A^{\mathsf{T}}=(a_{ji})$.

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 $A^{\mathsf{T}}=A$ 的矩阵，我们称之为**对称矩阵**. 对于满足 $A^{\mathsf{T}}=-A$ 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. $AI_n=I_nA=A$;
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. $(A^{\mathsf{T}})^{-1} = (A^{-1})^{\mathsf{T}}$;

13. $(A^{\mathsf{T}})^{\mathsf{T}} = A$;

14. $(A+B)^{\mathsf{T}} = A^{\mathsf{T}} + B^{\mathsf{T}}$;

15. $(AB)^{\mathsf{T}} = B^{\mathsf{T}}A^{\mathsf{T}}$;

16. $(\lambda A)^{\mathsf{T}} = \lambda A^{\mathsf{T}}$;

17. $(A^{-1})^{\mathsf{T}} = (A^{\mathsf{T}})^{-1}$;

18. $(A^{\mathsf{T}})^{-1} = (A^{-1})^{\mathsf{T}}$.

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 $f : V \to V$ 是一个线性算子，V 的两个基分别是 $\{v_1, \ldots, v_n\}$ 和 $\{v'_1, \ldots, v'_n\}$. 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 $\{v_1, \ldots, v_n\}$ 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，$f : V \to V$ 是一个线性算子，$\{v_1, \ldots, v_n\}$ 和 $\{w_1, \ldots, w_n\}$ 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 $\{v_1, \ldots, v_n\}$ 到 $\{w_1, \ldots, w_n\}$ 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，$f_A : F^n \to F^m$ 是它诱导的线性映射.

- 定义 A 的**秩**为 f_A 的秩，记为 $\operatorname{rank} A$.
- 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*

$$\operatorname{rank}(AB) \leq \min\{\operatorname{rank} A, \operatorname{rank} B\}.$$

如果 A 和 B 都是方阵，那么当 A 可逆时，$\operatorname{rank}(AB) = \operatorname{rank} B$；当 B 可逆时，$\operatorname{rank}(AB) = \operatorname{rank} A$.

§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 $f : V \times V \to F$，满足

1. 对于任意的 $v \in V$，$f(v, \cdot) : V \to F$ 是一个线性映射；
2. 对于任意的 $w \in V$，$f(\cdot, w) : V \to F$ 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

$$f(x, y) = X^{\mathsf{T}} A Y.$$

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

$$f(x, y) = X^{\mathsf{T}} A Y = X'^{\mathsf{T}} A' Y'.$$

根据坐标的基变换公式，$X = PX'$，$Y = PY'$，所以

$$X^{\mathsf{T}}AY = (AX')^{\mathsf{T}}P(AY') = X'^{\mathsf{T}}(A^{\mathsf{T}}PA)Y'.$$

由于 x, y 是任意的，X, Y 也是任意的，联立上面两式，$A' = A^{\mathsf{T}}PA$，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 $B = P^{\mathsf{T}}AP$，那么称 A 和 B 是**合同的**.

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.

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 $q : V \to F$ 为 $q(x) = f(x, x)$，称 q 是 f 诱导的**二次型**.

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.

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 $q(x) = X^{\mathsf{T}}AX$. 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**.

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是在一个 $n-1$ 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. □

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.

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. □

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 $A^{\mathsf{T}}A$ 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 $B = P^{\mathsf{T}}P$.*

§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 $\langle\cdot,\cdot\rangle : V \times V \to F$，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**.

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 $\|v\| = \sqrt{\langle v, v\rangle}$.

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：

$$\langle v, w\rangle = \frac{1}{4}\left(\|v+w\|^2 - \|v-w\|^2\right). \tag{A.1}$$

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 $v \perp w$.

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**.

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

$$4\langle v, w\rangle^2 - 4\|v\|^2 \|w\|^2 \leq 0 \iff |\langle v, w\rangle| \leq \|v\| \, \|w\| . \qquad \square$$

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

$$\begin{aligned} \|v + w\|^2 &= \langle v + w, v + w\rangle \\ &= \|v\|^2 + 2\langle v, w\rangle + \|w\|^2 \\ &\leq \|v\|^2 + 2\|v\| \, \|w\| + \|w\|^2 \\ &= (\|v\| + \|w\|)^2. \end{aligned} \qquad \square$$

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 $\lambda = \langle v, w\rangle / \|w\|^2$.

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 $v_1, \cdots, v_n \in V$ 是线性无关的.*

证明. 设 $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$ 满足 $\sum_{i=1}^n \lambda_i v_i = 0$，那么

$$0 = \left\langle \sum_{i=1}^n \lambda_i v_i, v_j \right\rangle = \sum_{i=1}^n \lambda_i \langle v_i, v_j \rangle = \lambda_j \left\| v_j \right\|^2.$$

因为 $v_j \neq 0$，所以 $\lambda_j = 0$，这就证明了线性无关性. □

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 $e_1, \cdots, e_n$，$v \in V$，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 $\langle x, y \rangle = X^\mathsf{T} Y$.

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 $\langle e_f, v \rangle = \left\langle e_f', v \right\rangle = f(v)$ 对任意的 v 都成立，那么取 $v = e_f - e_f'$ 得

$$\left\langle e_f - e_f', e_f - e_f' \right\rangle = 0 \iff e_f - e_f' = 0.$$

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 $v_1, \ldots, v_k$，定义他们的 **Gram 矩阵**为 $G = (\langle v_i, v_j \rangle)_{k \times k}$.

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，$A^\mathsf{T} A$ 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，$v_1, \ldots, v_k \in V$，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. *$v_1, \ldots, v_k$ 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 $f(x) = x^\mathsf{T} G x$. 对任意的 $x \in \mathbb{R}^k$，有

$$x^\mathsf{T} G x = \sum_{i,j=1}^{k} x_i x_j \langle v_i, v_j \rangle = \left\langle \sum_{i=1}^{k} x_i v_i, \sum_{j=1}^{k} x_j v_j \right\rangle = \left\| \sum_{i=1}^{k} x_i v_i \right\|^2 \geq 0.$$

   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 $x = 0 \iff v_1, \ldots, v_k$ 线性无关. □

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，$W \subset V$ 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，$e_1,\cdots,e_n$ 和 $e'_1,\cdots,e'_n$ 都是 V 的标准正交基. 设 $e'_j = a_{1j}e_1 + \cdots + a_{nj}e_n$，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，$e_1,\cdots,e_n$ 和 $e'_1,\cdots,e'_n$ 都是 V 的标准正交基. 设 e_i 到 e'_i 的过渡矩阵是 A，那么 $AA^\mathsf{T} = A^\mathsf{T}A = I_n$.

证明. 设 $A = (a_{ij})$，那么

$$\begin{aligned}\left\langle e'_i, e'_j\right\rangle &= \left\langle \sum_{k=1}^n a_{ki}e_k, \sum_{l=1}^n a_{lj}e_l \right\rangle \\ &= \sum_{k=1}^n \sum_{l=1}^n a_{ki}a_{lj}\left\langle e_k, e_l\right\rangle \\ &= \sum_{k=1}^n a_{ki}a_{kj}.\end{aligned}$$

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 $AA^\mathsf{T} = I_n$. 同理可证 $A^\mathsf{T}A = I_n$. □

我们将满足 $A^{\mathsf{T}}A = AA^{\mathsf{T}} = I_n$ 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. $A^{\mathsf{T}} = A^{-1}$.

证明. 按照定义写出即可. □

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，$T : V \to W$ 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，$f: V \to W$ 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 $f_A: \mathbb{R}^n \to \mathbb{R}^n$ 是一个等距同构.*

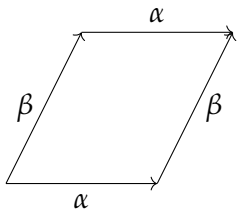
证明. 后半部分（“反之”之后）由定理 A.14 的第二点直接得出. 我们来证明前半部分.

设 $v_1, \cdots, v_n$ 是 V 的标准正交基，$w_1, \cdots, w_n$ 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. □

§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 $\alpha = (a_1, a_2)^{\mathsf{T}}$ 和 $\beta = (b_1, b_2)^{\mathsf{T}}$，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 $a_1b_2 - a_2b_1$，可以使用如下形象的记号表示：

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = \begin{vmatrix} \alpha & \beta \end{vmatrix}.$$

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 $A^1, \cdots, A^n$，我们考虑这些列向量张成的 n 维平行体：

$$\Pi(A) = \left\{ x_1 A^1 + \cdots + x_n A^n : 0 \le x_i \le 1 \right\}.$$

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 $\Pi(A^1, \ldots, A^{n-1})$. 为定义高，先将 A^n 投影到 $\mathrm{Span}(A^1, \ldots, A^{n-1})$，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 $A^1, \cdots, A^n$ 和 $\mathbb{R}^n$ 的标准正交基 $e_1, \cdots, e_n$ 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. $\det(A^\mathsf{T}) = \det A$；

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 $\det A$.

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 $X = (x_1,\dots,x_n)^{\mathsf{T}}$，那么

$$\begin{aligned}\|f(x)\|^2 &= \|AX\|^2 \\ &= \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{n} a_{ij}\right)x_j\right)^2 \\ &\le \max_k \sum_{i=1}^{n} |a_{ik}|^2 \cdot \sum_{j=1}^{n} x_j^2 \\ &= C\,\|X\| = C\,\|x\|\,.\end{aligned}$$

这里 $C = \max_k \sum_{i=1}^{n} |a_{ik}|^2$，这说明，对给定的算子 f，我们可以找到一个常数 C 使得 $\|f(x)\| \le C\,\|x\|$ 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

$$\|f\| = \inf\{C \ge 0 : \|f(x)\| \le C\,\|x\|, \forall x \in V\}\,.$$

或者等价地，

$$\|f\| = \sup_{x\in V} \frac{\|f(x)\|}{\|x\|} = \sup_{\|x\|=1} \|f(x)\|\,.$$

要保证定义出来的确实是范数，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，

$$\begin{aligned}\|f+g\| &= \sup_{\|x\|=1} \|(f+g)(x)\| \\ &= \sup_{\|x\|=1} \|f(x)+g(x)\| \\ &\le \sup_{\|x\|=1} (\|f(x)\| + \|g(x)\|) \\ &\le \sup_{\|x\|=1} \|f(x)\| + \sup_{\|x\|=1} \|g(x)\| \\ &= \|f\| + \|g\|\,.\end{aligned}$$

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，$\left\|f^k\right\| \leq \|f\|^k$，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 $\ker(f - \lambda \cdot \mathrm{id}) \neq \{0\}$. 因此，根据推论 A.1，$f - \lambda \cdot \mathrm{id}$ 不是满射，因此也不是双射，从而 $\det(f - \lambda \cdot \mathrm{id}) = 0$. 当选择一组基之后，$\det(f - \lambda \cdot \mathrm{id})$ 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：

$$\det(A - \lambda I_n) = 0. \tag{A.2}$$

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. $\det(P^{-1}AP - \lambda I_n) = \det(P^{-1}(A - \lambda I_n)P) = \det(P^{-1})\det(A - \lambda I_n)\det(P) = \det(A - \lambda I_n)$. □

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**.

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 $(AX)^{\mathsf{T}}Y = X^{\mathsf{T}}(AY)$，也就是 $X^{\mathsf{T}}A^{\mathsf{T}}Y = X^{\mathsf{T}}AY$. 将 X, Y 取遍所有可能的向量，我们得到 $A^{\mathsf{T}} = A$.

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 $P^{-1}AP = P^{\mathsf{T}}AP$ 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.

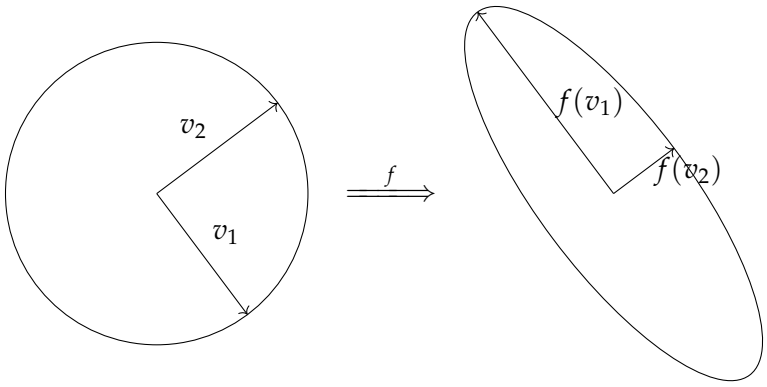


图 A.1: 自伴算子的作用

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

$$\|f(x)\|^2=\|AX\|^2=\sum_{i=1}^n\lambda_i^2x_i^2.$$

假设 $|\lambda_1|\geq|\lambda_2|\geq\cdots\geq|\lambda_n|$，那么

$$|\lambda_n|^2\sum_{i=1}^n x_i^2\leq\sum_{i=1}^n\lambda_i^2x_i^2\leq|\lambda_1|^2\sum_{i=1}^n x_i^2.$$

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　微分学基础

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学，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 $\mathbb{R}^n$ 相关的微分学.

§B.1　点集拓扑

本部分讨论极限、连续、紧致等概念，这些概念是微分学的基础.

B.1.1　度量空间，范数

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

定义 B.1 (度量空间) 设 X 是一个集合，$d : X \times X \to \mathbb{R}$ 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）：$d(x_1,x_2)=\max_i|x_1^i-x_2^i|=\lim_{p\to\infty}(\sum_{i=1}^n|x_1^i-x_2^i|^p)^{1/p}$.

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.

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，$\|\cdot\|:X\to\mathbb{R}$ 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**.

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：$d(\lambda x,\lambda y)=|\lambda|d(x,y)$.

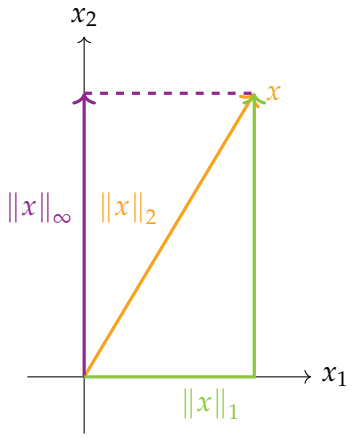
尽管都是 $\mathbb{R}^n$，但是不同 p 对应的 L^p 范数是不一样的. 他们之间有如下的关系：

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

$$\|x\|_p\ge\|x\|_q.$$

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. 绿色的是 $\|x\|_1=|x_1|+|x_2|$，相当于沿着坐标轴走；而橙色的是 $\|x\|_2=\sqrt{x_1^2+x_2^2}$，相当于沿着对角线走，肯定比沿着坐标轴走要快；紫色的是 $\|x\|_\infty=\max\{|x_1|,|x_2|\}$，相当于挑了较长的那条边走，仿佛虫洞一样，走完了就到了，所以甚至比对角线还快.



然而，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 $1\le p\le q\le\infty$，则存在正常数 $c_{p,q}$ 和 $C_{p,q}$，对任意 $x,y\in\mathbb{R}^n$，

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

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. $\Longrightarrow$：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 $B(x, r) \subseteq B(x_i, r_i) \subseteq U$.

$\Longleftarrow$：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. □

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，在拓扑意义下，$\mathbb{R}^n$ 空间到底装备了哪个 L^p 范数是不重要的，因此对于同一个数学对象（集合、序列、函数）来说，收敛性和连续性在 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 $c\|x\|_2 \le \|x\|_1 \le C\|x\|_2$，则 $B_2(x, r/c) \subseteq B_1(x, r) \subseteq U$，所以 U 是 $(X, \|\cdot\|_2)$ 中的开集. 反之亦然.

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

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：

[2]注意，对一般空间来说，这样的记号不意味着 L^1 或者 L^2 范数.

命题 B.4 设 (X,d) 是一个非空度量空间，则

1. X 和 $\varnothing$ 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. $\varnothing$ 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. □

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**.

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 $\varnothing$ 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 $n > N$，$d(x_n, x) < \epsilon$，则称 $\{x_n\}_{n=1}^\infty$ **收敛**到 x，记作 $\lim_{n\to\infty} x_n = x$ 或 $x_n \to x$，$n \to \infty$，x 称为 $\{x_n\}_{n=1}^\infty$ 的**极限**.

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 $n > N$，$x_n \in U$.*

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

证明. $\Longrightarrow$：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 $x_n\in B(x,r)\subseteq U$.

$\Longleftarrow$：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. □

在一般的拓扑空间中，甚至都没有度量的概念，然而，开集定义收敛依然是可以的. 这正是这一命题的意义.

下面给一些收敛的经典例子：

例 B.4

- 在 $\mathbb{R}$ 中，$\{1/n\}_{n=1}^{\infty}$ 收敛到 0，然而，序列 $\{n\}_{n=1}^{\infty}$ 则不收敛. 这个例子表明，极限未必需要在序列中出现，以及趋于无穷是一种特殊的不收敛.
- 在 $\mathbb{R}^n$ 和 L^p 范数下，$\{x_k\}_{k=1}^{\infty}$ 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，$f_n\to f$ 实际上是所谓**一致收敛**的概念，即对任意 $\epsilon>0$，存在不依赖 x 的 $N\in\mathbb{N}$，使得对任意 $n>N$，任意 $x\in[0,1]$，$|f_n(x)-f(x)|<\epsilon$. 在这一概念下，$\{x^n\}_{n=1}^{\infty}$ 就不收敛（尽管它逐点收敛）.

度量空间中紧集可以完全由收敛性来刻画：

定理 B.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**.

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 $m,n>N$，$d(x_m,x_n)\le d(x_m,x)+d(x,x_n)<\epsilon$，所以 $\{x_n\}_{n=1}^{\infty}$ 是 Cauchy 列. □

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**.

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.

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 $x\in F$，$\|x\|_p\le M$.*

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. □

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，$f : X \to Y$ 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 $\lim_{x\to x_0} f(x) = y$ 或 $f(x) \to y$，$x \to x_0$.

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：$\lim_{x\to\infty} f(x) = y$ 或 $f(x) \to y$，$x \to \infty$；
- 如果定义域是 $\mathbb{R}$，还可以定义趋于正、负无穷：$\lim_{x\to+\infty} f(x) = y$ 或 $f(x) \to y$，$x \to +\infty$，$\lim_{x\to-\infty} f(x) = y$ 或 $f(x) \to y$，$x \to -\infty$；
- 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 $n/(n+1) \uparrow 1$，$n \to \infty$.
- 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 $f(x) \to y$，$x \uparrow x_0$，以及从正向趋于某点（右极限）：$\lim_{x\downarrow x_0} f(x) = y$ 或 $f(x) \to y$，$x \downarrow x_0$.

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，$f : X \to Y$ 是一个映射，考虑点 $x \in X$，如果 $\lim_{x' \to x} f(x') = f(x)$，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**.

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 $f : X \to Y$ 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，$f : X \to Y$ 是一个映射，则*

1. *$\lim_{x \to x_0} f(x) = y$ 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，$f : X \to Y$ 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 $d:X\times X\to\mathbb{R}$ 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，

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

另一方面，

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

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，$f:X\to Y$ 是一个映射，$x_0\in X$，$y\in Y$，则下列表述等价：*

1. $\lim_{x\to x_0}f(x)=y$.
2. *对任意 $\{x_n\}_{n=1}^{\infty}$ 满足 $x_0\notin\{x_n\}$，如果 $x_n\to x_0$，则 $f(x_n)\to y$.*

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，$f:X\to Y$ 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 $x_n \to x$，则 $f(x_n) \to f(x)$.

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，$f: X \to Y$ 在 $x \in X$ 连续，$g: Y \to Z$ 在 $f(x) \in Y$ 连续，则 $g \circ f: X \to Z$ 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 $X \to X$ 的连续映射，向量加法是 $X \times X \to X$ 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 $\times: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ 和除法 $\div: \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$ 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，$f: X \to Y$ 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，$f: X \to Y$ 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 $f: X \to \mathbb{R}$. 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：

定义 B.13 (单调性) 设 $\{x_n\}_{n=1}^{\infty}$ 是一个实数列.

[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.

- 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.

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**.

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 $\lim_{n\to\infty} a_n = \sup A$.*

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 $\lim_{n\to\infty} a_n = M$. □

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进大小，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 $\lim_{n\to\infty}\frac{x_n}{y_n}=0$，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 $|x_n|\le C|y_n|$，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 $\lim_{n\to\infty}\frac{x_n}{y_n}=1$，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
- $n\to\infty$ 时，$n^2=o(2^n)$，$n^{1/n}\sim 1$，$n^2\sim n^2+\log n$.
- $x\to 0$ 时，$\sin x\sim x$，$1/x=o(1/x^2)$.
- $n\to\infty$ 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 $f:F\to\mathbb{R}$ 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 $f:[a,b]\to\mathbb{R}$ 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 $\mathbb{R}\to\mathbb{R}$ 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 $f:\mathbb{R}\to\mathbb{R}$，和点 $x_0\in\mathbb{R}$，我们希望找到一个线性映射 $df_{x_0}:\mathbb{R}\to\mathbb{R}$，使得 $f(x)$ 在 x_0 附近的行为很接近这线性映射，即

$$f(x)\approx f(x_0)+\mathrm{d}f_{x_0}(x-x_0).$$

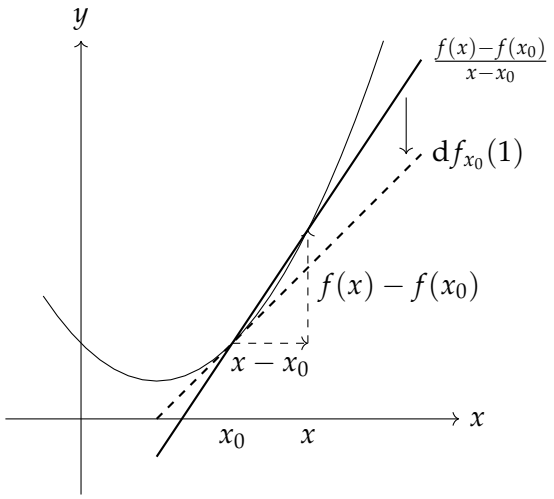
更精确来说，我们希望两边的误差是一个关于 $x-x_0$ 的高阶无穷小：

$$f(x)=f(x_0)+\mathrm{d}f_{x_0}(x-x_0)+o(x-x_0).$$

这一记号的含义可以通过一些变换看出来：

$$\mathrm{d}f_{x_0}(1)=\frac{f(x)-f(x_0)}{x-x_0}+o(1). \tag{B.1}$$

注意 $\mathrm{d}f_{x_0}(x)=kx$，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 $f:\mathbb{R}\to\mathbb{R}$，$x_0\in\mathbb{R}$，如果存在一个线性映射 $\mathrm{d}f_{x_0}:\mathbb{R}\to\mathbb{R}$，使得

$$f(x)=f(x_0)+\mathrm{d}f_{x_0}(x-x_0)+o(x-x_0),$$

则称 f 在 x_0 处**可微**或者**可导**，$\mathrm{d}f_{x_0}$ 是 f 在 x_0 处的**微分**. 微分具有形式 $\mathrm{d}f_{x_0}(x)=kx$，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，$\mathrm{d}f$ 是 f 的**微分**，f' 是 f 的**导（函）数**，也记作 $\frac{\mathrm{d}f}{\mathrm{d}x}$ 或 $\dot{f}$.

关于导数的符号有一些注. 最能体现几何意义的是 $\frac{\mathrm{d}f}{\mathrm{d}x}$，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 $\mathrm{d}x$ 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 $\mathrm{d}y$ 就是相应地沿着 y 轴的切向量. 从这个角度来说，$\frac{\mathrm{d}f}{\mathrm{d}x}$ 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 $f:\mathbb{R}\to\mathbb{R}$，$x_0\in\mathbb{R}$，那么 f 在 x_0 处可微当且仅当如下极限存在：*

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}.$$

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 $f,g:\mathbb{R}\to\mathbb{R}$ 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

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x}.$$

反函数的导数则可以写作

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^{-1}.$$

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 $f : \mathbb{R} \to \mathbb{R}$ 在 x_0 处可微，$g : \mathbb{R} \to \mathbb{R}$ 在 $f(x_0)$ 处可微，则

- $\mathrm{d}x = \mathrm{id}$;
- $\mathrm{d}(g \circ f)_{x_0} = \mathrm{d}g_{f(x_0)} \circ \mathrm{d}f_{x_0}$.
- 如果 f 存在反函数 f^{-1}，则 $\mathrm{d}(f^{-1})_{f(x_0)} = (\mathrm{d}f_{x_0})^{-1}$.

命题 B.18 提供了这样一种视角：微分号 d 相当于把 $\mathbb{R} \to \mathbb{R}$ 的函数变成了另外一个 $\mathbb{R} \to \mathbb{R}$ 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 $f : \mathbb{R} \to \mathbb{R}$，$x_0 \in \mathbb{R}$，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

$$\frac{\mathrm{d}^n y}{\mathrm{d}x^n} = \underbrace{\frac{\mathrm{d}}{\mathrm{d}x} \cdots \frac{\mathrm{d}}{\mathrm{d}x}}_{n\text{个}} y.$$

从这里我们可以看出，$\mathrm{d}/\mathrm{d}x$ 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 $f: X \to \mathbb{R}$，$x_0 \in X$，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**.

首先是 Fermat 引理，他其实就是极值的一阶必要条件:

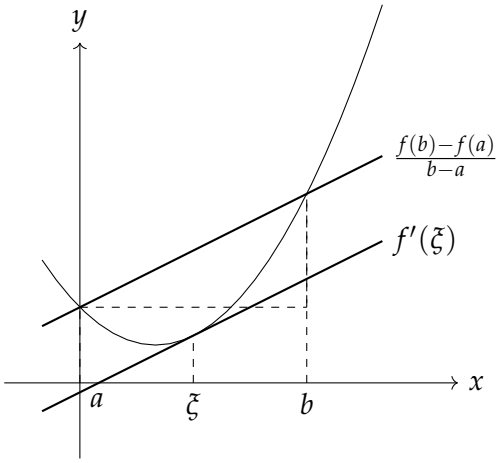
引理 B.1 (Fermat 引理) *设 $f: X \to \mathbb{R}$，$x_0 \in X$ 是 f 的一个极值点，且 f 在 x_0 处可微，则 $f'(x_0) = 0$.*

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 $f: [a,b] \to \mathbb{R}$ 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 $f:[a,b]\to\mathbb{R}$ 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 $f:\mathbb{R}\to\mathbb{R}$，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，$a_0=f(x_0)$，$a_1=f'(x_0)$，$a_2=\frac{f''(x_0)}{2}$，$a_3=\frac{f'''(x_0)}{6}$，$\cdots$，$a_k=\frac{f^{(k)}(x_0)}{k!}$，因此我们得到了 Taylor 公式：

定理 B.11 设 $f:\mathbb{R}\to\mathbb{R}$ 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 $f:(a,b)\to\mathbb{R}$ 在 x_0 处 n 次可微，且 $f'(x_0)=f''(x_0)=\cdots=f^{(n-1)}(x_0)=0$，$f^{(n)}(x_0)\neq0$，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 $\mathbb{R}^n \to \mathbb{R}^m$ 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 $df_x : \mathbb{R}^n \to \mathbb{R}^m$，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 $f : \mathbb{R}^n \to \mathbb{R}^m$，$x \in \mathbb{R}^n$，如果存在一个线性映射 $\mathrm{d}f_x : \mathbb{R}^n \to \mathbb{R}^m$，使得

$$f(x+h) = f(x) + \mathrm{d}f_x h + o(h),$$

则称 f 在 x_0 处**可微**或者**可导**，$\mathrm{d}f_{x_0}$ 是 f 在 x_0 处的**微分**. 这里 $o(h)$ 理解为一个向量值函数 $\alpha : \mathbb{R}^n \to \mathbb{R}^m$，它满足 $\lim_{h\to 0} \|\alpha(h)\| / \|h\| = 0$[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，$\mathrm{d}f$ 是 f 的**微分**.

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 $T\mathbb{R}^n_x$，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 $T\mathbb{R}^n_x = \mathbb{R}^n$. 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.

$$\begin{array}{ccc} f: & \mathbb{R}^n & \longrightarrow \mathbb{R}^m \\ \Big\downarrow \mathrm{d} & & \\ \mathrm{d}f_x: & T\mathbb{R}^n_x & \longrightarrow T\mathbb{R}^m_{f(x)} \end{array}$$

接下来的问题就是，如何表示线性映射 $\mathrm{d}f_x$？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 $T\mathbb{R}^n_x$ 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

$$\mathrm{d}f_x h = \langle g, h \rangle .$$

这个向量 g 被称为 f 在 x 处的**梯度**，记作 $\operatorname{grad} f(x)$.

我们需要进一步将梯度的坐标 $(g_1, \ldots, g_n)^{\mathsf{T}}$ 求出来. 考虑一个具体的分量 e_i，根据定义，

$$f(x + te_i) = f(x) + t\mathrm{d}f_x(te_i) + o(te_i) = f(x) + g_i t + o(t).$$

因此，

$$g_i = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}.$$

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：

$$\left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)^{\mathsf{T}}.$$

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**：

$$\frac{\partial f}{\partial u}(x) = \lim_{t \to 0} \frac{f(x + tu) - f(x)}{t}, \quad \|u\| = 1.$$

有了梯度，我们可以很快算出任意方向导数：

命题 B.19 *设 $f: \mathbb{R}^n \to \mathbb{R}$ 在 x 处可微，u 是单位向量，则*

$$\frac{\partial f}{\partial u}(x) = \langle \operatorname{grad} f(x), u \rangle .$$

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

$$\operatorname{grad} f(x) = \nabla f(x).$$

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 $(\partial_1, \dots, \partial_n)^\mathsf{T}$.

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 $T\mathbb{R}^n_x$）的标准正交基 e_i，选取 $\mathbb{R}^m$（也就是 $T\mathbb{R}^m_{f(x)}$）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 $\mathrm{d}f_x$，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 $(f_1(x), \dots, f_m(x))^\mathsf{T}$，考虑 $h \in T\mathbb{R}^n_x$，它的坐标是 $(h_1, \dots, h_n)^\mathsf{T}$，$\mathrm{d}f_x h$ 的坐标应该是

$$\begin{pmatrix} \mathrm{d}f_{1,x}h \\ \vdots \\ \mathrm{d}f_{m,x}h \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \partial_i f_1(x) h_i \\ \vdots \\ \sum_{i=1}^n \partial_i f_m(x) h_i \end{pmatrix} = \begin{pmatrix} \partial_1 f_1(x) & \cdots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(x) & \cdots & \partial_n f_m(x) \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}.$$

因此，我们得到了 Jacobi 矩阵：

$$J_f(x) = (\partial_j f_i) = \begin{pmatrix} \partial_1 f_1(x) & \cdots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(x) & \cdots & \partial_n f_m(x) \end{pmatrix}.$$

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 $\frac{\mathrm{d}f}{\mathrm{d}x}(x)$，于是，在坐标表示下，我们可以将微分简单写作 $\mathrm{d}f_x = f'(x)\mathrm{d}x$，这里我们将 $\mathrm{d}x$ 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 $f, g : \mathbb{R}^n \to \mathbb{R}^m$ 在 x 处可微，则*

- *f 在 x 处连续；*
- *对任意 $\lambda_1, \lambda_2 \in \mathbb{R}$, $\mathrm{d}(\lambda_1 f_x + \lambda_2 g_x) = \lambda_1 \mathrm{d}f_x + \lambda_2 \mathrm{d}g_x$；*

- 如果 $m=1$，那么 $\mathrm{d}(f\cdot g)_x = g(x)\mathrm{d}f_x + f(x)\mathrm{d}g_x$；
- 如果 $m=1$，并且 $g(x)\neq 0$，则
$$\mathrm{d}\left(\frac{f}{g}\right)_x = \frac{1}{g(x)^2}\left(g(x)\mathrm{d}f_x - f(x)\mathrm{d}g_x\right);$$
- 如果 $m=n$，那么 $\mathrm{d}x = \mathrm{id}$；
- 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 $\mathrm{d}(g\circ f)_x = \mathrm{d}g_{f(x)}\circ \mathrm{d}f_x$；
- 如果 f 存在反函数 f^{-1}，则 f^{-1} 在 $f(x)$ 处可微，且 $\mathrm{d}(f^{-1})_{f(x)} = (\mathrm{d}f_x)^{-1}$.

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（$\mathbb{R}^n \to \mathbb{R}^m$）到函数空间（$T\mathbb{R}^n_x \to T\mathbb{R}^m_{f(x)}$）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

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x}.$$

反函数的导数则可以写作

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^{-1}.$$

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：

$$\frac{\mathrm{d}(Ax)}{\mathrm{d}x} = A,\quad \frac{\mathrm{d}(b^{\mathsf{T}}x)}{\mathrm{d}x} = b^{\mathsf{T}}.$$

其中 A 是一个矩阵，b 是一个向量.

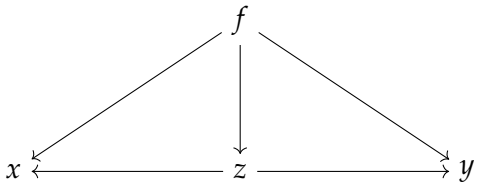
例 B.9 考虑一个多元实值函数 $g(x) = f(u_1(x),\ldots,u_k(x))$，先求 f 对 $u=(u_i)$ 的导数：

$$\frac{\mathrm{d}f}{\mathrm{d}u} = \left(\frac{\partial f}{\partial u_1},\ldots,\frac{\partial f}{\partial u_k}\right).$$

再求 u 对 x 的导数 $\mathrm{d}u/\mathrm{d}x = (u_1'(x), \ldots, u_k'(x))^\mathsf{T}$. 根据链式法则，

$$\frac{\mathrm{d}g}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x} = \left(\frac{\partial f}{\partial u_1}, \ldots, \frac{\partial f}{\partial u_k} \right) \begin{pmatrix} u_1'(x) \\ \vdots \\ u_k'(x) \end{pmatrix} = \sum_{i=1}^{k} \frac{\partial f}{\partial u_i} u_i'(x).$$

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 $z \to y$ 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 $f \to z \to x$ 和 $f \to x$，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.

例 B.11 考虑二次型 $f(x) = x^\mathsf{T} A x$（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 $g(x,y) = x^\mathsf{T} A y$，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 $\partial_1 g(x,x)$，$g(x,y) = (Ay)^\mathsf{T} x$，因此根据例 B.8，$\partial_1 g(x,y) = (Ay)^\mathsf{T}$，于是 $\partial_1 g(x,x) = x^\mathsf{T} A^\mathsf{T}$，同理 $\partial_2 g(x,x) = x^\mathsf{T} A$，因此 $f'(x) = x^\mathsf{T} A + x^\mathsf{T} A^\mathsf{T} = (2Ax)^\mathsf{T}$.

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 $2Axx$ 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 $f : \mathbb{R}^n \to \mathbb{R}$，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 $f : \mathbb{R}^n \to \mathbb{R}$ 具有下列两个二阶偏导数*

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x), \frac{\partial^2 f}{\partial x_i \partial x_j}(x),$$

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 $g'(t) = (h^\mathsf{T} \nabla) f$. 再求二阶导数，根据命题 B.21 和链式法则，

$$g''(t) = \sum_{i=1}^{n} h_i \frac{\mathrm{d}}{\mathrm{d}t}(\partial_i f(x + th)) = \sum_{i=1}^{n} h_i \sum_{j=1}^{n} \partial_j(\partial_i f(x + th)) h_j = \sum_{i,j=1}^{n} h_i h_j \partial_{j,i} f(x + th).$$

用 nabla 算子，我们可以写作 $g''(t) = (h^\mathsf{T} \nabla)^2 f(x + th)$. 一般地，我们有

$$g^{(m)}(t) = \sum_{i_1, \ldots, i_m} \partial_{i_m, \ldots, i_1} f(x + th) h_{i_1} \cdots h_{i_m} = (h^\mathsf{T} \nabla)^m f(x + th).$$

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 $\operatorname{grad} f$，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 $H_f(x)$. 很容易算出，Hessian 矩阵为

$$H_f(x) = \begin{pmatrix} \partial_{1,1} f(x) & \cdots & \partial_{1,n} f(x) \\ \vdots & \ddots & \vdots \\ \partial_{n,1} f(x) & \cdots & \partial_{n,n} f(x) \end{pmatrix}.$$

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：

$$g''(t) = h^{\mathsf{T}} H_f(x + th) h.$$

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 $f : \mathbb{R}^n \to \mathbb{R}$，$x_0 \in \mathbb{R}^n$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $r(2\pi) - r(0) = r'(\xi)(2\pi - 0)$.

[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.

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 $f:\mathbb{R}^n\to\mathbb{R}^m$ 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*

$$\|f(x+h)-f(x)\|\le\|h\|\sup_{\xi\in(x,x+h)}\left\|f'(\xi)\right\|.$$

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

$$\|\operatorname{grad}f(x)-\operatorname{grad}f(y)\|\le\|x-y\|\sup_{\xi\in(x,y)}\left\|H_f(\xi)\right\|.$$

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，$\left\|H_f(x)\right\|$ 连续. 由于 X 是紧集，根据 Weierstrass 最值定理（定理 B.7）$\left\|H_f(x)\right\|$ 在 X 上取到最大值 M，因此，我们有

$$\|\operatorname{grad}f(x)-\operatorname{grad}f(y)\|\le M\|x-y\|.$$

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 $\|F(x)-F(y)\|\le M\|x-y\|$.

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

$$\|\operatorname{grad}f(x)-\operatorname{grad}f(y)\|\le\sup_{\lambda\in\sigma(H_f(z)),z\in(x,y)}|\lambda|\cdot\|x-y\|.$$

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 8.5.

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

$$f(x+h) = \sum_{j=0}^{k} \frac{1}{j!}(h^{\mathsf{T}}\nabla)^j f(x) + o(\|h\|^k).$$

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 $f : \mathbb{R}^2 \to \mathbb{R}$，$f(x,y) = x^2 + y^2 - 1$，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 $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$，$F(x,y) = 0$，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：

$$\begin{aligned} &a_{11}x_1 + \cdots + a_{1n}x_n + b_{11}y_1 + \cdots + b_{1m}y_m - c_1 = 0, \\ &\cdots \\ &a_{k1}x_1 + \cdots + a_{kn}x_n + b_{k1}y_1 + \cdots + b_{km}y_m - c_k = 0. \end{aligned} \tag{B.2}$$

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：

$$\begin{aligned} x_1 + x_2 &= 0, \\ x_1 + x_2 &= 1. \end{aligned}$$

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 $F:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^m$，$F(x,y)=0$，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 8.1），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 $f:V\to W$ 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：

$$\begin{aligned}x&=r\cos\phi,\\ y&=r\sin\phi.\end{aligned}$$

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：

$$\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \end{pmatrix} = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix} \begin{pmatrix} \mathrm{d}r \\ \mathrm{d}\phi \end{pmatrix} = J_f \begin{pmatrix} \mathrm{d}r \\ \mathrm{d}\phi \end{pmatrix}.$$

这里对诸如 $\mathrm{d}x$ 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 $\Pi(\mathrm{d}r, \mathrm{d}\phi)$ 映射为平行体 $\Pi(\mathrm{d}x, \mathrm{d}y)$，而 Jacobi 行列式就是他们有向体积变化的比率. 如果我们把 $\Pi(\mathrm{d}r, \mathrm{d}\phi)$ 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 $\mathrm{d}x\mathrm{d}y$ 表示的正好就是长方形 $\Pi(\mathrm{d}x, \mathrm{d}y)$ 的有向面积（长乘宽），而 $\mathrm{d}r\mathrm{d}\phi$ 表示的正好就是长方形 $\Pi(\mathrm{d}r, \mathrm{d}\phi)$ 的有向面积，于是我们有

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}r\mathrm{d}\phi} = \frac{\partial(x, y)}{\partial(r, \phi)} \iff \mathrm{d}x\mathrm{d}y = \frac{\partial(x, y)}{\partial(r, \phi)} \mathrm{d}r\mathrm{d}\phi.$$

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 (古典概型) 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 $B_{ij} = \{(i,j)\}$.

例 C.2 (几何概型) 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.

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.

对于古典概型，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 $1/36$.

对于几何概型，不能再用古典概型的方式定义概率. 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，我们也不能用古典概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. 所以几何概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 $1/2$.

几何概型的定义看似合理，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？

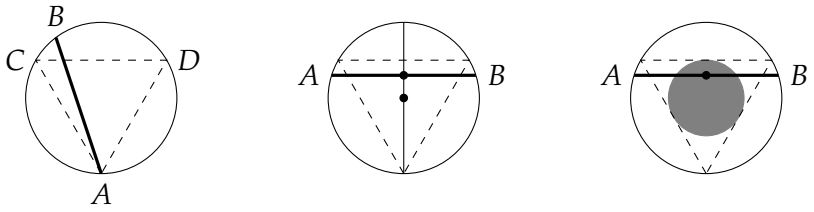
我们给出三种答案，这三种答案对应了我们对“随机”的不同理解.

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，$\mathscr{F}$ 是 Ω 的子集的集合. 如果 $\mathscr{F}$ 满足

1. $\Omega \in \mathscr{F}$;
2. 如果 $A \in \mathscr{F}$，则 A 的补集 $\Omega \setminus A \in \mathscr{F}$;
3. 如果 $A_1, A_2, \ldots \in \mathscr{F}$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

则称 $\mathscr{F}$ 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 $\mathscr{F}$. 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，$\mathscr{F}$ 是 Ω 上的一个 σ-代数. 则称 $(\Omega, \mathscr{F})$ 是一个**可测空间**.

设 $S \subseteq \Omega$，如果 $S \in \mathscr{F}$，则称 S 是 **$\mathscr{F}$-可测**的.

定义可测空间与 $\mathscr{F}$-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 $\mathscr{F}$-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 $(\Omega, \mathscr{F})$ 是一个可测空间. 如果函数 $\Pr : \mathscr{F} \to [0,1]$ 满足

1. 正则性：$\Pr(\Omega) = 1$;
2. 可列可加性：如果 $A_1, A_2, \cdots \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，则：

1. $\Pr(\varnothing) = 0$;
2. 单调性：对任意的 $A, B \in \mathscr{F}$，如果 $A \subseteq B$，则 $\Pr(A) \leq \Pr(B)$;
3. 有限可加性：对两两不相交的 $A_1, A_2, \ldots, A_n \in \mathscr{F}$，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

$$\prod_i (a_i, b_i) = \{x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : a_i < x_i < b_i\}.$$

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 $\mathscr{B}([0,1]^n)$. 换言之，如果还有一个 σ-代数 $\mathscr{F}$ 包含所有长方体，那么 $\mathscr{B}([0,1]^n) \subseteq \mathscr{F}$.

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 $A \in \mathscr{B}([0,1]^n)$，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 $A \in \mathscr{B}([0,1]^n)$，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 $\mathscr{B}(\mathbb{R}^n)$ 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 $\mathscr{B}(\mathbb{R}^n)$ 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 $\lambda(\varnothing) = 0$.

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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 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 $\mathscr{A} \subseteq \mathscr{F}$，如果对任意的有限个不同的 $A_1, A_2, \ldots, A_n \in \mathscr{A}$，都有

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr(A_i),$$

则称事件族 $\mathscr{A}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，那么

1. 对任意 $A \in \mathscr{F}$ 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 $A \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A_1, A_2, \ldots \in \mathscr{F}$ 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 $C \in \mathscr{F}$，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. □

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”.

定理 C.2 (贝叶斯公式) 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A, B \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$A_1, A_2, \ldots, A_n \in \mathscr{F}$，且 $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) > 0$，则

$$\begin{aligned}&\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)\\ =&\Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1 \cap A_2)\cdots\Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).\end{aligned}$$

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

$$\begin{aligned}\Pr(A_1) &= \frac{w}{w+b},\\ \Pr(A_2|A_1) &= \frac{w+c}{w+b+c},\\ \Pr(A_3|A_1 \cap A_2) &= \frac{w+2c}{w+b+2c},\\ &\cdots\\ \Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) &= \frac{w+nc}{w+b+nc}.\end{aligned}$$

所以

$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) = \frac{w}{w+b} \cdot \frac{w+c}{w+b+c} \cdots \frac{w+nc}{w+b+nc}.$$

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 $\Pr(A_1A_2)$ 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 $(\Omega, \mathscr{F}, \Pr)$，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. $\mathscr{F}$ 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.

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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 是一个函数. 如果对任意的 $x \in \mathbb{R}$，$\{X \in \mathscr{B}(\mathbb{R})\} \in \mathscr{F}$，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 $\mathscr{B}(\mathbb{R}^n)$，$X : \Omega \to \mathbb{R}^n$ 是一个映射. 如果对任意的 $A \in \mathscr{B}(\mathbb{R}^n)$，集合 $\{X \in A\}$ $\mathscr{F}$-可测，即 $\{X \in A\} \in \mathscr{F}$，则称 X 是一个 n **维随机向量**，简称**随机向量**. 如果 $(\Omega, \mathscr{F}) = (\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$，则称 X 是一个 **Borel 函数**.

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，这一点经常会被人误解. 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 $\{X \in B\}$（$B \in \mathscr{B}(\mathbb{R})$）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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}^n$ 是一个随机向量，$g : \mathbb{R}^n \to \mathbb{R}^m$ 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 $A \in \mathscr{B}(\mathbb{R}^n)$，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 $(\Omega, \mathscr{F}, \Pr)$，以及一个随机变量 $X : \Omega \to \mathbb{R}$. 要刻画概率测度 $\Pr$，我们需要给出所有的事件 $A \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 是一个随机变量. 定义函数 $F_X : \mathbb{R} \to \mathbb{R}$ 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 $X : \Omega \to \mathbb{R}^n$ 是一个随机向量，定义函数 $F_X : \mathbb{R}^n \to \mathbb{R}$ 为 $F_X(x) = \Pr(X \leq x)$，这里 $X \leq x$ 是指对任意的 $i = 1, 2, \ldots, n$，都有 $X_i \leq x_i$. 我们称 F_X 是 X 的**分布函数**，记作 $X \sim F_X$.

容易验证，分布函数具有如下的性质：

命题 C.4 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}$ 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. $\lim_{x\to-\infty} F_X(x) = 0$，$\lim_{x\to+\infty} F_X(x) = 1$；
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 $\mathbb{R} \to \mathbb{R}$ 的函数，满足命题 C.4 的前三条性质.

在概率空间 $([0,1], \mathscr{B}([0,1]), \lambda)$ 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，

- $\Pr(X \le a) = F(a)$,
- $\Pr(X < a) = F(a-)$,
- $\Pr(X > a) = 1 - F(a)$,
- $\Pr(X \ge a) = 1 - F(a-)$,
- $\Pr(X = a) = F(a) - F(a-)$.

对于 $\mathbb{R}^n \to \mathbb{R}$ 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：

$$\begin{aligned}&\Delta_{a_i b_i} F(x_1, x_2, \ldots, x_n)\\ =&F(x_1, x_2, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) - F(x_1, x_2, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n).\end{aligned}$$

例如，对于区间 $(a, b] = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : a_i < x_i \le b_i\}$，我们有

$$\Pr(X \in (a, b]) = \Delta_{a_1 b_1} \Delta_{a_2 b_2} \cdots \Delta_{a_n b_n} F_X(x_1, x_2, \ldots, x_n).$$

容易证明，分布函数具有如下的性质：

命题 C.5 设 $(\Omega, \mathscr{F}, \Pr)$ 是一个概率空间，$X : \Omega \to \mathbb{R}^n$ 是一个随机向量，F_X 是它的分布函数，则

1. 对任意 $a_i \le b_i$，$i = 1, 2, \ldots, n$，都有 $\Delta_{a_i b_i} F_X(x_1, x_2, \ldots, x_n) \ge 0$；
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 $F_X(y) \to F_X(x)$.

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 $\mathbb{R}^n \to \mathbb{R}$ 的函数，满足命题 C.5 的三条性质.
在概率空间 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 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 $([0,1]^n, \mathscr{B}([0,1]^n), \lambda)$ 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 $X \overset{d}{=} Y$.

现在，我们将分布函数与概率测度联系在一起：

定理 C.6 设 $F : \mathbb{R}^n \to \mathbb{R}$ 是一个分布函数，则在可测空间 $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ 上，存在唯一的概率测度 $\Pr$，使得对任意 $a_i \le b_i$，

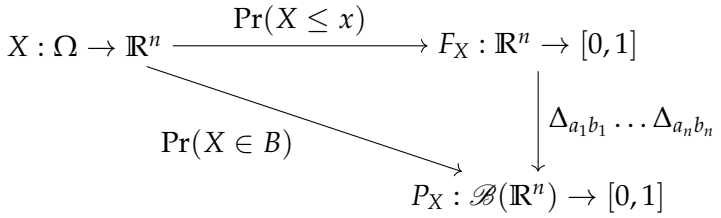
$$\Pr\left(\prod_{i=1}^{n}(a_i, b_i]\right) = \Delta_{a_1 b_1}\Delta_{a_2 b_2}\cdots\Delta_{a_n b_n} F(x_1, x_2, \ldots, x_n).$$

特别地，分布函数

$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

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：

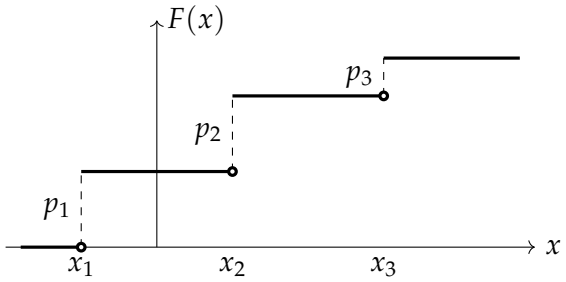


这张图的每一个箭头都可以反过来，但是反过来的这些关系都比较不直观，所以我们不再讨论.

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 $p_i = \Pr(X = i)$.

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.

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

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 $A \in \mathscr{B}(\mathbb{R})$ 满足 $\lambda(A) < \delta$，都有 $\Pr(A) < \epsilon$.

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 $F : \mathbb{R} \to \mathbb{R}$ 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

$$F(b) - F(a) = \int_a^b f(x)\mathrm{d}x. \tag{C.1}$$

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

$$\frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) = \lim_{\lambda(A)\to 0, x\in A} \frac{\Pr(A)}{\lambda(A)},$$

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：

$$\Pr(A) = \int_{x\in A} \frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x)\mathrm{d}\lambda(x).$$

我们可以把(C.1) 改写成如上的形式：

$$\Pr((a,b]) = \int_{x\in(a,b]} f(x)\mathrm{d}x.$$

在一维的情况下，x 的微小变化就是 $\lambda(x)$ 的微小变化，所以 $\mathrm{d}x = \mathrm{d}\lambda(x)$. 综合这两点，我们容易相信，

$$f(x) = \frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) \iff \mathrm{d}\Pr = f(x)\mathrm{d}\lambda.$$

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 $\int \Pr(X = x)\mathrm{d}x$ 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 $\mathrm{d}\Pr = p_X\mathrm{d}\lambda$；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 $\mathrm{d}\Pr/\mathrm{d}\lambda$ 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 $\mathrm{d}\Pr/\mathrm{d}\lambda$ 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 $B \in \mathscr{B}(\mathbb{R})$，都有

$$\Pr(X \in B) = \int_{x \in B} f(x)\mathrm{d}x.$$

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.

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	$\mathsf{Exp}(\lambda)$	$p(x) = \lambda \mathrm{e}^{-\lambda x},\ x \geq 0$	$\lambda > 0$
双指数	$\mathsf{DExp}(\lambda)$	$p(x) = \frac{\lambda}{2}\mathrm{e}^{-\lambda\lvert x\rvert},\ x \in \mathbb{R}$	$\lambda > 0$
Laplace	$\mathsf{Lap}(\mu,\lambda)$	$p(x) = \frac{\lambda}{2}\mathrm{e}^{-\lambda\lvert x-\mu\rvert},\ x \in \mathbb{R}$	$\mu \in \mathbb{R},\ \lambda > 0$
正态（Gauss）	$\mathcal{N}(\mu,\sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}$	$\mu \in \mathbb{R},\ \sigma > 0$

表 C.2: 本书中用到的连续型分布

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，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，$U(a,b)$ 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 $U(a,b)$ 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 $\Omega \to \mathbb{R}^n$ 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 $X = (X_1, \ldots, X_n)^\mathsf{T}$. 通常，我们将 X 的分布函数称为 $X_1, \ldots, X_n$ 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：$\Pr(X_1 = x_1, \ldots, X_n = x_n) = p_{x_1,\ldots,x_n}$，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 $PN(n, p_1, \ldots, p_k)$，分布列为

$$\Pr(X_1 = i_1, \ldots, X_n = i_n) = \frac{n!}{i_1! \cdots i_k!} p_1^{i_1} \cdots p_k^{i_k},$$

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 $F : \mathbb{R}^n \to \mathbb{R}$ 是绝对连续测度对应的分布函数，那么存在一个非负可积函数 $f : \mathbb{R}^n \to \mathbb{R}$，使得对任意的 $(x_1, \ldots, x_n) \in \mathbb{R}^n$，都有*

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

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

$$\frac{\mathrm{d}\Pr}{\mathrm{d}\lambda}(x) = p(x) \iff \mathrm{d}\Pr = p(x)\mathrm{d}\lambda.$$

这里，我们需要再给出一些 $\mathrm{d}\lambda$ 和 $\mathrm{d}x$ 关系的讨论. $\mathrm{d}\lambda$ 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. $\mathrm{d}x$ 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 $\lambda(\mathrm{d}x)$. 所以，在 x 处，$\mathrm{d}\lambda$ 和 $\mathrm{d}x$ 之间的关系应该是 $\mathrm{d}\lambda = \lambda(\mathrm{d}x)$，于是 $\mathrm{d}\lambda$ 应该理解为 $\mathrm{d}x$ 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 $B \in \mathscr{B}(\mathbb{R}^n)$，都有*

$$\Pr(X \in B) = \int_{x \in B} f(x)\mathrm{d}x.$$

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 $1\le i\le n$, 都有

$$p_{X_i}(x_i)=\int_{(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\in\mathbb{R}^{n-1}} p_X(x_1,\dots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_{i-1}\mathrm{d}x_{i+1}\cdots\mathrm{d}x_n.$$

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:

$$p_{X_1,X_2}(x_1,x_2)=\int_{x_3\in\mathbb{R}} p_X(x_1,x_2,x_3)\mathrm{d}x_3.$$

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

$$p(x)=\frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)},$$

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: $\Pr(X_1\le x_1|X_2=x_2)$. 考虑 $\Pr(X_1\le x_1|x_2<X_2\le x_2+\epsilon)$，再令 $\epsilon\downarrow 0$，我们有如下计算:

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\Pr(X_1\le x_1|x_2<X_2\le x_2+\epsilon)\\
=&\lim_{\epsilon\downarrow 0}\frac{\Pr(X_1\le x_1,X_2\le x_2+\epsilon)-\Pr(X_1\le x_1,X_2\le x_2)}{\Pr(x_2<X_2\le x_2+\epsilon)}\\
=&\lim_{\epsilon\downarrow 0}\frac{F_X(x_1,x_2+\epsilon)-F_X(x_1,x_2)}{F_{X_2}(x_2+\epsilon)-F_{X_2}(x_2)}.
\end{aligned}$$

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：

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\frac{F_X(x_1,x_2+\epsilon)-F_X(x_1,x_2)}{F_{X_2}(x_2+\epsilon)-F_{X_2}(x_2)}\\
=&\frac{\partial F_X(x_1,x_2)}{\partial x_2}\frac{1}{p_{X_2}(x_2)}\\
=&\int_{-\infty}^{x_1}\frac{\partial^2 F_X(y,x_2)}{\partial x_2\partial y}\frac{1}{p_{X_2}(x_2)}\mathrm{d}y\\
=&\int_{-\infty}^{x_1}\frac{p_{X_1,X_2}(y,x_2)}{p_{X_2}(x_2)}\mathrm{d}y.
\end{aligned}$$

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 $1\le i_1,\dots,i_k\le n$，以及任意的 $B_{i_1}\in\mathscr{B}(\mathbb{R}^{n_{i_1}}),\dots,B_{i_k}\in\mathscr{B}(\mathbb{R}^{n_{i_k}})$，都有

$$\Pr(X_{i_1}\in B_{i_1},\dots,X_{i_k}\in B_{i_k})=\Pr(X_{i_1}\in B_{i_1})\cdots\Pr(X_{i_k}\in B_{i_k}),$$

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.

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*

$$F(x_1,\dots,x_n)=F_1(x_1)\cdots F_n(x_n),$$

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

$$p(x_1,\dots,x_n)=p_1(x_1)\cdots p_n(x_n),$$

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

$$p(x_1,\dots,x_n)=p_1(x_1)\cdots p_n(x_n),$$

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

$$p(x_1,\dots,x_n)=f_1(x_1)\cdots f_n(x_n),$$

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

$$p(x_1,x_2)=\begin{cases}8x_1x_2, & 0\le x_1\le x_2, 0\le x_2\le 1,\\ 0, & \text{其他.}\end{cases}$$

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：

$$p_1(x_1)=\int_{x_2\in\mathbb{R}}p(x_1,x_2)\mathrm{d}x_2=\begin{cases}4x_1(1-x_1^2), & 0\le x_1\le 1,\\ 0, & \text{其他.}\end{cases}$$

再计算 X_2 的边缘密度：

$$p_2(x_2) = \int_{x_1 \in \mathbb{R}} p(x_1, x_2) \mathrm{d}x_1 = \begin{cases} 4x_2^3, & 0 \leq x_2 \leq 1, \\ 0, & \text{其他}. \end{cases}$$

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，$p(x_1, x_2) = 8x_1x_2I_{0 \leq x_1 \leq x_2 \leq 1}(x_1, x_2)$，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 $\mathbb{R}^n \to \mathbb{R}^n$ 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 $\mathrm{d}y$，$\mathrm{d}P_Y = p_Y \lambda(\mathrm{d}y)$ 是 Y 在 $\mathrm{d}y$ 上的概率，同样区域的概率也可以用 X 去计算：

$$\mathrm{d}P_X = p_X \lambda(\mathrm{d}x), \quad Y \in \mathrm{d}y \iff X \in \mathrm{d}x,$$

当然，这里 $\mathrm{d}y$ 和 $\mathrm{d}x$ 由函数 g 联系在一起，因为 $Y = g(X)$，所以 $\mathrm{d}y/\mathrm{d}x = g'(X)$，注意，这相当于微元 $\mathrm{d}y$ 和微元 $\mathrm{d}x$ 的有向体积的比. 最后，根据概率相等，可以写出如下的等式：

$$\mathrm{d}P_Y = \mathrm{d}P_X \iff p_Y \lambda(\mathrm{d}y) = p_X \lambda(\mathrm{d}x). \tag{C.2}$$

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），

$$p_Y(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p_X(x) = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1} p_X(x) = |\det g'(x)|^{-1} p_X(x).$$

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

$$\mathrm{d}P_Y = \sum_{g(x)=y} \mathrm{d}P_X(x) \iff p_Y \lambda(\mathrm{d}y) = \sum_{g(x)=y} p_X(x) \lambda(\mathrm{d}x)(x). \tag{C.3}$$

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 $\mathbb{R}^n \to \mathbb{R}^n$ 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

$$\begin{aligned}&\sum_{g(x)=y} |\det g'(g^{-1}(y))|^{-1} p_X(g^{-1}(y))\\ =&\sum_{x^2=y} \frac{1}{2|x|} p_X(x)\\ =&\frac{1}{2\sqrt{y}} p_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} p_X(-\sqrt{y}).\end{aligned}$$

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

$$p_Y \lambda(\mathrm{d}y) = p_X(\sqrt{y})\lambda(\mathrm{d}x)(\sqrt{y}) + p_X(-\sqrt{y})\lambda(\mathrm{d}x)(-\sqrt{y}).$$

两边除以 $\lambda(\mathrm{d}y)$，再利用 $\mathrm{d}y/\mathrm{d}x = 1/(2x)$，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 $g : \mathbb{R}^n \to \mathbb{R}^m$，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 $G : \mathbb{R}^n \to \mathbb{R}^n$ 满足

$$G(x_1, \ldots, x_n) = (g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n), x_{m+1}, \ldots, x_n)^{\mathsf{T}}.$$

然后，利用这一函数计算出 $g(X)$ 和 $(X_{m+1}, \ldots, X_n)$ 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 $(x, y) \mapsto (x + y, y)$ 显然是双射，所以 (C.3) 退化为 (C.2)，我们有

$$p_{Z,Y}(z, y) = \left|\frac{\partial(z, y)}{\partial(x, y)}\right|^{-1} p_{X,Y}(x, y) = p_{X,Y}(z - y, y).$$

[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.

于是，Z 的边缘密度为

$$p_Z(z) = \int_{y\in\mathbb{R}} p_{X,Y}(z-y,y)\mathrm{d}y.$$

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 $\mathbb{R}^{n_i}\to\mathbb{R}^{m_i}$ 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.

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 $\mathbb{E}[I_A] = 1 \cdot \Pr(A) = \Pr(A)$.

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 $n \to \infty$ 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 $\lim_{n\to\infty} \mathbb{E}[X_n]$.

定义 C.10 (数学期望（Lebesgue 积分），非负情形) 称

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$$

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：

[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.

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

$$E[X] = \int_\Omega X \mathrm{d}\Pr.$$

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 $A \in \mathscr{B}(\mathbb{R}^n)$，定义

$$\int_{\mathbb{R}^n} I_A(\omega)\lambda(\mathrm{d}\omega) = \lambda(A).$$

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

$$\int_{\mathbb{R}^n} f(x)\mathrm{d}x.$$

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，这样就转化为了一个离散型随机变量：

$$X_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \le X < (k+1)/n\}}.$$

容易看出，$X_n(\omega) \uparrow X(\omega)$ 对任意 ω 成立. 于是，我们有

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

[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，$\mathrm{d}x$ 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，$\mathrm{d}x$ 被理解为所对应平行体的体积. 所以其实 $\lambda(\mathrm{d}x)$ 这一记号虽然复杂，但是含义更准确.

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

$$\begin{aligned}\mathbb{E}[X_n] &= \sum_{k=0}^{\infty} \frac{k}{n} \Pr\left(\frac{k}{n} \le X < \frac{k+1}{n}\right) \\ &= \sum_{k=0}^{\infty} \frac{k}{n} \left(F\left(\frac{k+1}{n}\right) - F\left(\frac{k}{n}\right)\right).\end{aligned}$$

按照极限的想法，当 $n \to \infty$ 时，上式的求和项相当于 $x\mathrm{d}F(x)$，这里 $\mathrm{d}F$ 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

$$\mathbb{E}[X] = \int_{x \ge 0} x\mathrm{d}F(x) = \int_{\mathbb{R}} x\mathrm{d}F(x),$$

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 $\mathrm{d}F(x) = 0$.

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：

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\mathrm{d}F(x).$$

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 $x_1, x_2, \ldots$ 会发生改变，其他地方都是常值，所以我们有

$$\int_{\mathbb{R}} x\mathrm{d}F(x) = \sum_{k=1}^{\infty} x_k(F(x_k) - F(x_k-)) = \sum_{k=1}^{\infty} x_k \Pr(X = x_k).$$

对于连续型随机变量来说，$\mathrm{d}F = p\mathrm{d}x$，这里 p 是对应的密度. 于是我们有

$$\int_{\mathbb{R}} x\mathrm{d}F(x) = \int_{\mathbb{R}} xp(x)\mathrm{d}x.$$

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：

$$\int_A X\mathrm{d}\Pr = \int_\Omega XI_A\mathrm{d}\Pr = \mathbb{E}[XI_A].$$

相应地，在 $\mathbb{R}^n$ 上，对我们也可以定义

$$\int_A f(x)\lambda(\mathrm{d}x) = \int_{\mathbb{R}^n} f(x)I_A(x)\lambda(\mathrm{d}x).$$

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 $\mathsf{Var}(X)$；
- 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 $f_X:\mathbb{R}^n\to\mathbb{C}$，$f_X(t)=\mathbb{E}[\exp(i\langle t,X\rangle)]$ 被称为 X 的**特征函数**.

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 $X\le Y$，那么

$$\mathbb{E}[X]\le\mathbb{E}[Y].$$

3. 绝对值不等式：设 X 是随机变量，那么

$$\mathbb{E}[|X|]\ge|\mathbb{E}[X]|.$$

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，$g:\mathbb{R}^n\to\mathbb{R}$ 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

$$p_Y(y)=\begin{cases}\frac{1}{2\sqrt{y}}, & 0\le y\le 1,\\ 0, & \text{其他}.\end{cases}$$

于是，

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} y p_Y(x) \mathrm{d}y = \int_0^1 \frac{y}{2\sqrt{y}} \mathrm{d}y = \frac{1}{3}.$$

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]

$$\mathbb{E}[g(X)] = \lim_{n \to \infty} \mathbb{E}[g(X_n)].$$

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：

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) \mathrm{d}F(x).$$

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，$g : \mathbb{R}^n \to \mathbb{R}$ 是一个 Borel 函数，那么 $\mathbb{E}[g(X)]$ 存在，且*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) \mathrm{d}F_X(x).$$

特别地，如果 X 是一个离散型随机变量，取值为 $x_1, x_2, \ldots$，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

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) p_X(x) \mathrm{d}x.$$

[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.

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

$$E[X^2] = \int_0^1 x^2 \mathrm{d}x = \frac{1}{3}.$$

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 $\mathbb{E}[I_A] = \Pr(A)$，$\mathsf{Var}(I_A) = \Pr(A)(1-\Pr(A))$.*

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*

$$\mathbb{E}[I_A I_B] = \mathbb{E}[I_A]\mathbb{E}[I_B].$$

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 $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

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 $g : \mathbb{R} \to \mathbb{R}$ 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，$g : \mathbb{R} \to \mathbb{R}$ 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

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

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 *$\mathrm{Cov}(\cdot, \cdot)$ 具有以下性质：*

1. *对称性：任意随机变量 X, Y，$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$；*
2. *单边线性性：任意随机变量 X, Y，$a, b \in \mathbb{R}$，$\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$；*
3. *正定性：任意随机变量 X，$\mathrm{Cov}(X, X) \geq 0$，且 $\mathrm{Cov}(X, X) = 0$ 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 $\mathrm{Cov}(X, Y) = 0$.*

我们称 $\mathrm{Cov}(X, Y) = 0$ 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

$$\Sigma_{ij} = \mathsf{Cov}(X_i, X_j).$$

如果 X 和 Y 分别是 m 维和 n 维随机向量，那么符号 $\mathsf{Cov}(X, Y)$ 表示的是 $m \times n$ 的矩阵 $(\mathsf{Cov}(X_i, X_j))_{ij}$，称为 X 和 Y 的**协方差矩阵.** 特别地，如果 $X = Y$，那么我们记 $\mathsf{Cov}(X, X)$ 为 $\mathsf{Var}(X)$，称为 X 的**协方差矩阵.**

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 $X = (X_1, \dots, X_n)^\mathsf{T}$ 是一个 n 维随机向量，称

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])^\mathsf{T}$$

为 X 的**数学期望.**

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*

$$\mathsf{Var}(AX) = A\mathsf{Var}(X)A^\mathsf{T}.$$

证明. 考虑向量 t，和 n 维随机向量 Y，$t^\mathsf{T}Y$ 是一个随机变量，我们可以得到一个二次型

$$g(t) = \mathsf{Var}(t^\mathsf{T}Y) = \mathsf{Cov}(t^\mathsf{T}Y, t^\mathsf{T}Y) = t^\mathsf{T}\mathsf{Var}(Y)t.$$

当 $Y = AX$ 时，我们有

$$g(t) = \mathsf{Var}(t^\mathsf{T}AX) = \mathsf{Var}((A^\mathsf{T}t)X) = t^\mathsf{T}A\mathsf{Var}(X)A^\mathsf{T}t.$$

所以，对任意 t 都有 $t^\mathsf{T}\mathsf{Var}(AX)t = t^\mathsf{T}A\mathsf{Var}(X)A^\mathsf{T}t$，所以 $\mathsf{Var}(AX) = A\mathsf{Var}(X)A^\mathsf{T}$. □

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 $t^\mathsf{T}X$ 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 $t^\mathsf{T}AX = (A^\mathsf{T}t)X$，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 $A\mathsf{Var}(X)A^\mathsf{T}$.

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 $a < b$，

$$F_X(b) - F_X(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-ita} - \mathrm{e}^{-itb}}{it} f_X(t)\mathrm{d}t.$$

2. 如果 $\int_{\mathbb{R}} |f_X(t)|\mathrm{d}t < +\infty$，那么 X 具有密度 p_X，且

$$p_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-itx} f_X(t)\mathrm{d}t.$$

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

$$\Pr(X_1 = c_1, \ldots, X_{k-1} = c_{k-1}, X_k \in (a_k, b_k], X_{k+1} = c_{k+1}, \ldots, X_n = c_n) = 0,$$

其中 $c_i \in \{a_i, b_i\}$，我们有

$$\begin{aligned} &F_X(b) - F_X(a) \\ =& \lim_{T_1,\ldots,T_n\to\infty} \frac{1}{(2\pi)^n} \int_{-T_1}^{T_1} \cdots \int_{-T_n}^{T_n} \prod_{k=1}^{n} \frac{\exp(-it_k a_k) - \exp(-it_k b_k)}{it_k} f_X(t)\mathrm{d}t. \end{aligned}$$

2. 如果 $\int_{\mathbb{R}^n} |f_X(t)|\mathrm{d}t < +\infty$，那么 X 具有密度 p_X，且

$$p_X(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \mathrm{e}^{-it^{\mathrm{T}}x} f_X(t)\mathrm{d}t.$$

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 $X_1+\cdots+X_n$ 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 $X=(X_1,\dots,X_n)^\mathsf{T}$，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

$$f(t_1,\dots,t_n)=f_1(t_1)\cdots f_n(t_n).$$

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 $A_1, \ldots, A_n$，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 $\mathscr{A}$. 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，$\mathscr{A} = \{A_1, \ldots, A_n\}$ 是 Ω 的一个分割，满足 $\Pr(A) > 0$ 对任意 $A \in \mathscr{A}$ 成立. X 在 $\mathscr{A}$ 上的**条件数学期望**是一个随机变量，记为 $\mathbb{E}[X|\mathscr{A}]$，它的定义为

$$\mathbb{E}[X|\mathscr{A}](\omega) = \sum_{i=1}^n \frac{\mathbb{E}[XI_{A_i}]}{\Pr(A_i)} I_{A_i}(\omega).$$

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 $\mathscr{A}$ 其实生成了一个 Ω 的 σ-代数，即 $\sigma(\mathscr{A})$，它是包含 $A_1, \ldots, A_n$ 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 $A \in \sigma(\mathscr{A})$，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 $\mathbb{E}[XI_A|\mathscr{A}]$ 的数学期望 $\mathbb{E}[\mathbb{E}[XI_A|\mathscr{A}]]$.

对任意 $A \in \sigma(\mathscr{A})$，这两种计算方式都应该相等：

$$\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[XI_A|\mathscr{A}]]. \tag{C.4}$$

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，$\mathscr{G}$ 是 Ω 的一个 σ-代数，随机变量 $\mathbb{E}[X|\mathscr{G}]$ 被称为 X 关于 $\mathscr{G}$ 的**条件数学期望**，如果它满足

1. 对任意 $B \in \mathscr{B}(\mathbb{R})$，$\{\mathbb{E}[X|\mathscr{G}] \in B\}$ 是 $\mathscr{G}$-可测的；

2. 对任意 $A \in \mathscr{G}$，$\mathbb{E}[XI_A] = \mathbb{E}[\mathbb{E}[X|\mathscr{G}]I_A]$.

设 X 是一个一般的随机变量，如果

$$\min\{\mathbb{E}[X^+|\mathscr{G}], \mathbb{E}[X^-|\mathscr{G}]\} < +\infty,$$

那么 X 关于 $\mathscr{G}$ 的**条件数学期望**是一个随机变量，记为 $\mathbb{E}[X|\mathscr{G}]$，定义为

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X^+|\mathscr{G}] - \mathbb{E}[X^-|\mathscr{G}].$$

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，“$\mathbb{E}[X|\mathscr{G}]$ 落在合理的值集上”这件事情是可以用 $\mathscr{G}$ 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 $\mathscr{G}$ 可以理解为我们观测样本的能力. $\mathscr{G}$ 越大，则越能确定 ω 具体的范围，所以条件期望就越像 $X(\omega)$；$\mathscr{G}$ 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 $B \in \mathscr{B}(\mathbb{R}^n)$.

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 $\mathbb{E}[X|\sigma(Y)]$.

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 $g: \mathbb{R}^n \to \mathbb{R}$，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 $\mathbb{E}[X|Y]$，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 $(\Omega, \mathscr{F}, \Pr)$ 是概率空间，$\mathscr{G} \subseteq \mathscr{F}$ 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 $\mathbb{E}[X|\mathscr{G}]$ 和 $\mathbb{E}[Y|\mathscr{G}]$ 都存在，那么 $\mathbb{E}[aX + bY|\mathscr{F}]$ 存在，且

$$\mathbb{E}[aX + bY|\mathscr{G}] = a\mathbb{E}[X|\mathscr{G}] + b\mathbb{E}[Y|\mathscr{G}].$$

2. 单调性：设 X, Y 是随机变量，如果 $X \leq Y$，那么

$$\mathbb{E}[X|\mathscr{G}] \leq \mathbb{E}[Y|\mathscr{G}].$$

3. 绝对值不等式：设 X 是随机变量，那么

$$\mathbb{E}[|X||\mathscr{G}] \geq |\mathbb{E}[X]|\mathscr{G}|.$$

4. 如果 $\mathscr{G} = \{\varnothing, \Omega\}$，那么

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X].$$

5. 望远性：设 X 是随机变量，如果 $\mathscr{G}_1, \mathscr{G}_2 \subseteq \mathscr{F}$ 都是 Ω 的 σ-代数，且 $\mathscr{G}_1 \subseteq \mathscr{G}_2$，那么

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}_2]|\mathscr{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathscr{G}_1]|\mathscr{G}_2] = \mathbb{E}[X|\mathscr{G}_1].$$

6. 重期望公式：设 X 是随机变量，那么

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}]] = \mathbb{E}[X].$$

7. 设 X, Y 是随机变量，如果 $\sigma(Y) \subseteq \mathscr{G}$，那么

$$E[XY|\mathscr{G}] = Y\mathbb{E}[X|\mathscr{G}].$$

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 $\mathscr{A} = \{A_1, \dots, A_n\}$，并且 $\Pr(A_i) > 0$，那么

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathscr{A}]] = \sum_{i=1}^{n} \mathbb{E}[X|A_i] \Pr(A_i).$$

最后，性质 7 是在说，如果 Y 是 $\mathscr{G}$-可测的（也就是我们用 $\mathscr{G}$ 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

$$f_X(t) = \exp\left(it^\mathsf{T}\mu - \frac{1}{2}t^\mathsf{T}\Sigma t\right).$$

利用 (4.1)，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

$$\mathbb{E}[X] = \mu, \quad \mathsf{Var}(X) = \Sigma.$$

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

$$f_X(t) = \exp\left(it^{\mathsf{T}}\mu - \frac{1}{2}t^{\mathsf{T}}\Sigma t\right),$$

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 $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \Sigma)$，那么 $X_1, \ldots, X_n$ 相互独立的充分必要条件是 $X_1, \ldots, X_n$ 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 $(X_1, \ldots, X_n)$ 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 $X_1, \ldots, X_n$ 是 n 个 Gauss 向量，它们相互独立的充分必要条件是 $X_1, \ldots, X_n$ 两两不相关，即协方差矩阵 $\mathrm{Cov}(X_i, X_j) = O$，$i \neq j$.*

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 $AX \sim \mathcal{N}(A\mu, A\Sigma A^{\mathsf{T}})$.*

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 $X = (X_1, \ldots, X_n)^{\mathsf{T}} \sim \mathcal{N}(\mu, \Sigma)$，那么对任意的 $1 \leq k \leq n$，$i_1, \ldots, i_k \in \{1, \ldots, n\}$，$(X_{i_1}, \ldots, X_{i_k})^{\mathsf{T}}$ 也是 Gauss 向量.

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引

ℓ^2 空间, 165, 166, 168, 208
ℓ^p 空间, 200, 207
σ-代数, 233, 265, 266

Bayes 概率论, 237
Bayes 网络, 238
Bertrand 悖论, 232
Borel 代数, 235
Borel 函数, 239

Cauchy 不等式, 185
Cauchy 列, 206

Gauss 向量, 248, 268, 269
Gram 矩阵, 187

Hessian 矩阵, 225

Jacobi 矩阵, 221
Jacobi 行列式, 221, 252

Lebesgue 测度, 235, 242
Lebesgue 积分, 255, 256

nabla 算子, 221

Radon-Nikodym 定理, 245
Radon-Nikodym 导数, 245
Riesz 表示定理, 187

Taylor 展开, 218, 227

三角不等式, 185
上界, 212
上确界, 212
下界, 212
下确界, 212
乘积空间, 167
事件域, 233
二次型, 181
互斥, 238

偏导数, 220, 224
像, 172, 209
全概率公式, 237
内积, 184
内积空间, 184, 261
凸函数, 260
分布
    Bernoulli ~ , 244
    Gauss ~ , 246
    Laplace ~ , 246
    二项 ~ , 244
    双指数 ~ , 246
    多元正态 ~ , 248
    多项 ~ , 247
    对称 Bernoulli ~ , 244
    指数 ~ , 246
    条件 ~ , 248
    正态 ~ , 246

离散均匀 ~, 244
离散型 ~, 243
联合 ~, 247
边缘 ~, 247
连续均匀 ~, 246
连续型 ~, 243
分布函数, 240
切向量, 219
切映射, 215, 219
切空间, 219
列满秩, 179
列秩, 179
列空间, 179
勾股定理, 186
半正定, 182
半负定, 182
协方差, 261
协方差矩阵, 262
单位向量, 184
单位矩阵, 176
单调性, 211
卷积, 254, 264
原像, 209
双线性型, 180
反函数定理, 253
反向传播算法, 223
反对称矩阵, 177
可测空间, 233
合同矩阵, 181
同态, 171
同构, 171
同构定理, 172
向量, 166
向量空间, 166
和空间, 169
坐标, 168
域, 165
基, 168
多元正态分布, 268
夹角, 184
完备度量空间, 207
对称矩阵, 177
对角矩阵, 177
导数, 214–216, 221, 225
度量, 199
L^1 ~, 200
L^2 ~, 200
L^∞ ~, 200
L^p ~, 200
Chebyshev ~, 200
Euclid ~, 200
Manhattan ~, 200
Minkowski ~, 200
离散 ~, 199
绝对值 ~, 199
度量空间, 199
开球, 202
开覆盖, 205
开集, 202
微分, 214, 215, 219
微分算子, 171
恒等映射, 170
惯性定理, 182
投影, 185
投影变换, 171
拓扑空间, 202

收敛, 205
    一致 ~, 206
数学期望, 255, 256, 262
方向导数, 220
方差, 257
无穷小, 213
条件数学期望, 264–266
条件概率, 235
极值, 217
极限, 205, 208
标准正交基, 186
标准正交基存在性定理, 186
样本点, 231
样本空间, 231
核, 172
梯度, 220
概率, 234
概率密度函数, 245, 247
概率测度, 234
概率空间, 234
模, 184
正交, 184
正交基, 186
正交矩阵, 189
正交补, 188
正定, 182
测度
    离散型 ~, 243
    绝对连续 ~, 243, 244
满秩, 179

特征值, 195
特征函数, 257, 263
特征向量, 195
特征多项式, 195
特征子空间, 195
独立性, 236, 249
生成集, 167
直和, 169
直和分解, 169
相似, 179
矩, 257
矩阵, 174
示性函数, 240, 246, 251, 255, 265
秩, 173, 179, 181
等价, 213
等距同构, 189
等距映射, 189
紧集, 205
线性函数, 170
线性变换, 170
线性子空间, 167
线性映射, 170
线性相关, 167
线性空间, 166
线性算子, 170
线性组合, 167
绝对连续函数, 244
维数, 169
维数定理, 169

自伴算子, 196
范数, 184, 185, 200
    矩阵 ~, 194
    等价 ~, 202
    算子 ~, 193
    线性映射的 ~, 194
行列式, 191, 192

行满秩, 179
行秩, 179
行空间, 179
表示论, 167
规范型, 182
规范基, 182
谱, 195
谱半径, 198
贝叶斯公式, 238
负定, 182
赋范空间, 200
转置, 177
过渡矩阵, 175
连续映射, 209

逆转公式, 263
链式法则, 215, 222, 238
闭集, 204
阶, 213
随机变量, 239
    离散型 ~ , 243
    连续型 ~ , 243
随机向量, 239
随机过程, 237, 267
隐函数, 227
隐函数定理, 228
零映射, 170