概率论讲义整理

by Chihao Lecture Notes

作者：PhantomPhoenix

组织：Labmem

时间：November 24, 2025

版本：1.0

邮箱：logic_1729@sjtu.edu.cn

既对沧海一声笑，何不潇洒走一回

目录

第 1 章 课程简介

概率在我们生活中无处不在。经常赌博的同学都知道，我们会谈论某个比赛结果或者某个事件发生的概率，这反映在赌博公司开出来的赔率上。比如约老师赢得 2025-2026 年 MVP 的赔率；在赌场的时候我们会计算扔三个骰子得到 666（即豹子）的概率；又或者说明天某股票会涨的概率。有很多不同的方式能够对这些事件进行概率建模，而每一种建模的方式都有其局限性。

在这门课里，我们会专注于所谓的Kolmogorov 的公理体系，它使得我们能够使用数学分析的工具来研究概率。一旦接受了这个体系，很多生活中的问题，比如扔一枚均匀硬币会有 50% 的概率出现正面，这句话就有了准确的数学的含义，尽管它可能和物理事实不一样（谁能否认当我们扔硬币的时候，每个硬币的结果都是由外星人控制的呢？想想三体里面的智子，我们没有办法否认它的存在），但物理事实如何，比如是否真的有外星人控制了硬币的结果，并不是我们讨论的范畴。有趣的是，通过这些数学理论我们能够预测出一些结果，它和现实符合的很好（比如扔一万次硬币里面大约有五千次正面）。

但某一些问题，比如说前面提到的 MVP 概率问题，并不适合在这个公理体系里进行建模。对于类似的事情，或者对于概率论的一些其它的解释，可以参考《概率论沉思录》。

接下来，我想通过一些例子，来说明一个严格的概率论是必须的，以及概率论是有用的。在这些例子中，我也许会用到一些在未来才严格介绍的记号。

1.1 圣彼得堡悖论 (St. Petersburg Paradox)

这是我最喜欢的例子之一。假设有一个基于掷硬币的赌博游戏。首先庄家扔一个公平硬币，如果结果是正面，则给玩家 2 元钱，游戏结束；如果结果是反面，庄家再扔一次硬币，如果结果是正面，则给玩家 4 元钱，游戏结束；否则按照同样的规则继续扔硬币，每一轮奖金翻倍。换句话说，庄家会生成一个无限长的投掷硬币的结果序列，如果这个序列里第一次出现正面是在第 k 次，则玩家获得 2^k 元的奖金。现在的问题是：你愿意花多少钱去购买一次玩这个游戏的机会？或者说，假设你可以无限次地玩这个游戏，但是每一次需要付门票 a 元，那你认为，a 设置成多少是合理的？

一个很自然的想法是计算每一轮游戏的平均收益。很显然，我们有 2^{-k} 的概率在第 k 轮拿钱走人。因此，平均收益是

$$\sum_{k\geq 1} 2^k \cdot 2^{-k} = 1 + 1 + 1 + \cdots = \infty.$$

也就是说，平均我们每一轮的收益是无穷大！

我们在生活中有一个常见的直观是，独立重复一个随机试验很多次，那么平均收益会趋近于这个实验的期

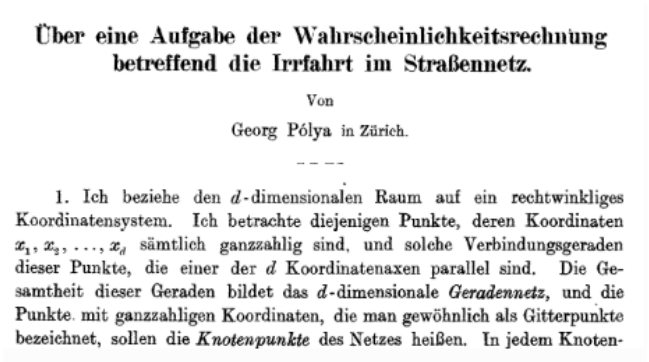
望收益，这个在概率论中叫做大数定律（Law of large numbers）。那这么说，如果允许我们不断的玩这个游戏，那每一轮门票的价格无论定多少钱，对于玩家来说都是赚的。但想想现实生活，你真的愿意花比如每轮一万元去玩这个游戏吗？我们可以用如下两种方式来问这个问题：

- 如果允许你花 a 元购买一次玩游戏的机会（可以重复任意次，每次 a 元），在 a 的定价是多少的时候对于玩家来说是合算的？
- 假设游戏门票是捆绑销售，即定价 $a \cdot n$ 元来购买 n 次游戏的机会，那 a 应该定价为多少呢？

我们会在这门课里发展足够的数学工具，来回答上面这些问题。

1.2 波利亚的困惑

有一个流传已广的关于数学家波利亚的小故事。他喜欢在公园里一边散步一边思考问题。有一次他在散步的时候，正好有一对夫妻也在公园里散步。他在散步的过程中好几次遇到了这对夫妻，导致对方怀疑波利亚是不是在猥琐的跟踪他们。波利亚知道自己并没有，并且非常好奇为什么总会遇到对方，因此，想从数学上来证明这件事情。



Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz.

Von

Georg Pólya in Zürich.

1. Ich beziehe den d-dimensionalen Raum auf ein rechtwinkliges Koordinatensystem. Ich betrachte diejenigen Punkte, deren Koordinaten $x_1, x_2, \ldots, x_d$ sämtlich ganzzahlig sind, und solche Verbindungsgeraden dieser Punkte, die einer der d Koordinatenaxen parallel sind. Die Gesamtheit dieser Geraden bildet das d-dimensionale *Geradennetz*, und die Punkte mit ganzzahligen Koordinaten, die man gewöhnlich als Gitterpunkte bezeichnet, sollen die *Knotenpunkte* des Netzes heißen. In jedem Knoten-

这个问题的一个简化建模是这样的：假设一个人在二维的 $\mathbb{Z}^2$ 的网格上随机游走。他从原点 $(0,0)$ 出发，每次分别以 $1/4$ 的概率往上下左右四个方向移动。我们现在想问，这个随机游走，是否会无数次回到原点？我们用 T 来表示第一次回到原点的时间，那么 T 的取值是随机的。可以证明，无数次回到原点等价于 $\mathbb{P}[T<\infty]=1$，即 T 以 1 的概率是有限的。当然，我们现在还没有定义这儿概率 $\mathbb{P}[T<\infty]$ 是什么意思，这是一个非平凡的事情，也是我们在未来的几节课里要做的事情。

当然，关于这个问题，数学家角谷静夫说过：

喝醉的人一定能够回家，而喝醉的鸟不一定能回家。

这也就是波利亚证明的，当考虑在 n-维格点 $\mathbb{Z}^n$ 上的随机游走的时候，对于 $n \le 2$，$\mathbb{P}[T<\infty]=1$，而对于 $n>2$，$\mathbb{P}[T<\infty]<1$。

1.3 投资策略问题

我们希望通过下面这个例子来说明，在计算机科学，或者说算法设计中，使用概率或者随机是不可或缺的。这个问题是在线优化（Online Learning）领域的经典问题，更多的相关资料可以查看这本专著。

我们考虑一个很简化的投资模型。假设现在有两只股票，我们进行 T 天的交易，每一天，玩家需要选择一只股票进行投资。假设当前是第 t 天，在这一天开始的时候，需要选定投资哪一只，在这一天结束的时候，可以看到收益。我们假设两只股票在第 t 天的收益是 $r_1^{(t)}, r_2^{(t)} \in [0,1]$。假设第 t 天玩家选择了投资股票 a_t，则玩家在 T 天的总收益是

$$R(T) := \sum_{t=1}^{T} r_{a_t}^{(t)}.$$

那么，我们应该如何选择一个好的投资策略呢?

首先，我们必须明确怎样衡量一个投资策略的好坏。一个很自然的假设是把 $R(T)$ 看成一个关于 T 的函数，我们当然是希望累计收益 $R(T)$ 越大越好。但是，这儿的 $L(T)$ 不仅与玩家的策略有关，还与两只股票每天的收益有关。假设因为大环境不好，两只股票的收益都很差，那自然不管投资策略如何聪明，都不可能有很高的收益。所以，一个很自然的想法是把玩家 T 天的累计收益 $R(T)$ 和表现最好的那只股票相比。这便是懊悔值（Regret）的定义：对于一个给定的投资策略，以及每天的收益情况 $\vec{r} = \left(\left(r_1^{(t)}, r_2^{(t)}\right)\right)_{1\le t\le T}$

$$\mathsf{Regret}(T) := \left(\max_{a\in\{1,2\}} \sum_{t=1}^{T} r_a^{(t)}\right) - R(T).$$

换句话说，$\mathsf{Regret}(T)$ 可以描述成因为没有事先知道哪只股票最好而产生的懊悔的程度。

我们希望一个好的投资策略是，不管两只股票每天的收益如何，即对于任意的 $\vec{r} = \left(\left(r_1^{(t)}, r_2^{(t)}\right)\right)_{1\le t\le T}$，$\mathsf{Regret}(T)$ 都比较小。注意到，$\mathsf{Regret}(T)$ 最大是 T，因此，我们希望我们的算法满足 $\mathsf{Regret}(T) = o(T)$，这表示当 T 足够大的时候，我们的投资策略事实上找到了最好的股票。

我们首先证明，任何确定性的策略，都不可能达到 $o(T)$ 的后悔值。首先明确，在第 t 天的时候，我们的策略可以看成前 $t-1$ 天我们对于股票的选择以及对应收益的情况到两个股票上的一个映射 f_t，即:

$$f_t : (a_1, a_2, \ldots, a_{t-1}, (r_1^{(1)}, r_2^{(1)}), (r_1^{(2)}, r_2^{(2)}), \ldots, (r_1^{(t-1)}, r_2^{(t-1)})) \mapsto a_t.$$

于是，我们想象有一个坏人可以针对这个策略来控制市场，即如果当前玩家选了股票 1，则让 $r_1^{(t)} = 0, r_2^{(t)} = 1$，如果当前玩家选了股票 2，则让 $r_2^{(t)} = 0, r_1^{(t)} = 1$。

我们来计算这个策略的懊悔值 $\mathsf{Regret}(T)$。容易看到，在这样针对性的设置下，$R(T) = 0$。并且，在每一天，两个股票的收益之和是 1。因此，一定有一个股票，它的 T 天累计收益之和 $\ge T/2$。所以，我们有后悔值 $\mathsf{Regret}(T) \ge T/2$。

可以看到，确定性算法之所以表现不好，在于对手可以进行针对性的设置。我们可以使用随机来避免这一点。这就是所谓的在线镜像下降（Online Mirror Descent）算法，它是一个在计算机科学非常著名的算法，在多个领域被重新发现过，因此，它也有很多其他的名字，比如 Multiplicative weight update method，Hedge 算法，EXP3 算法等。

简单来说，算法在每一轮会维护一个分布 D_t，然后玩家的决策来自于从这个分布中的采样。并且，玩家会根据每回合的反馈来更新这个分布。

1. 初始情况 $D_1 = (1/2, 1/2)$.
2. 对于 $t = 1, 2, \ldots T$
   - 玩家选择股票 $a_t \sim D_t$，并且观察到 $r_1^{(t)}, r_2^{(t)}$.
   - 更新 D_{t+1} 使得 $D_{t+1}(j) = \dfrac{D_t(j)\exp\left(-\eta(1-r_j^{(t)})\right)}{\sum_{k=1,2} D_t(k)\exp\left(-\eta(1-r_k^{(t)})\right)}$,

其中参数 $\eta = \sqrt{1/T}$。算法的想法很简单：这一轮哪个股票表现的好，就增加它在下一轮被选的概率。当然，为什么要像算法中这样增加，这样增加了能够达到什么样的效果，如果用别的方式或者程度增加概率行不行，这就是一个复杂而有点深刻的问题了。这一节最后给出的讲义上可以找到一些讨论。

可以证明，这个算法满足在期望上 $\mathsf{Regret}(T) = O(\sqrt{T})$。这个结论表示，使用随机，可以让算法的效果有质变。我们可以把一个问题的难度想象成算法设计者和给出环境数据的人（即这儿设计 $r^{(t)}$ 的人）的一个博弈结果。一旦允许算法设计者使用随机数，他便有了额外的手段，使得他所设计的算法不那么容易被对手针对了。

我们仅仅给出这个算法的描述，对分析感兴趣的同学，可以参考这个讲义（1,2）。对于这个算法在其他问题上的应用，可以参考这一篇 survey。

1.4 参考书

这儿，我列举一些这门课的参考书，它们会在课程的不同进度中有所帮助。值得注意的是，除了第一本 *Knowing the Odds* 之外，其它的都不太适合作为教材从头到尾读下去。

1. *Knowing the Odds*, John Walsh.
2. *Mathematics of Probability*, Daniel Stroock.
3. *Probability and Computing*, Michael Mitzenmacher and Eli Upfal.
4. An Introduction to Probability Theory and Its Applications (two volumes), William Feller.
5. *Probability Theory*, Terence Tao.

第 2 章 概率空间

我们今天开始介绍概率论的 Kolmogorov 公理体系。首先，我们要明确一点，就是严格的定义概率是必要的。我们可以看看著名的伯特兰悖论：“在一个圆内随机取一条弦，有多大概率它比内接三角形的边长要长?”这句用自然语言定义的话实际上是极不严谨的，取决于对“随机取一条弦”的理解方式，我们可能得到很多不同的答案。维基百科上关于这个悖论的解释说的非常清楚，我这儿就不再赘述了。

2.1 概率公理

概率公理都来自于对我们现实生活中随机试验的抽象。生活中常见的随机试验比如“投掷一枚硬币”，“投掷两个骰子“，或者“在平面上的单位圆内随机选一个点”等等。

在我们前两周的课程中，**我们会把注意力集中在所谓“离散概率空间”**，即那些随机实验的结果是有限个或者至多可数无限个的情况。因此，类似“在平面上的单位圆内随机选一个点”这样的随机试验（单位圆内的点的个数显然是不可数的）暂时还不在我们讨论的范围。

我们这样做的原因在于，我们需要先在离散的场合建立关于概率正确的直观，而不必过早的把精力花在由于不可数的概率空间引起的一些过于复杂的系统性问题上（比如我们在本次讲义最后一节所说的）。

当然，在这门课里，我们的目标是建立一个一般性的理论。我们将在对于离散的世界理解的足够好之后，再讨论一般的概率空间，并期待大家能够发现哪一些直观是能够直接类比过去，而哪一些是不能的。

因此，我们在介绍抽象的概率公理的时候，我们脑海中最好能有一个例子，从而明白每一条公理出现的理由。在这儿，我们不妨假设关心的随机试验是连续扔两枚 6 面的骰子。

一个概率空间是一个三元组 $(\Omega, \mathscr{F}, \mathbb{P})$。我们分别来解释其含义。

其中 Ω 是“样本集”，即所有随机实验的可能结果。在我们投掷两个骰子的例子里，我们可以让 $\Omega = [6]^2$（对于自然数 $n \in \mathbb{N}$，我们用记号 $[n]$ 表示集合 $\{1, 2, ..., n\}$）。

一个很自然的问题是，我们为什么不把 $\mathscr{F}$ 直接取成 2^Ω，而说它是 2^Ω 的一个子集呢? 这里我认为有好几个原因。首先一个比较重要的原因是，如果 Ω 是不可数的集合，那么把 $\mathscr{F}$ 定义成 2^Ω 会“太大了”，以至于没有办法定义合适的概率。这一点我们今天之后会解释。另外一个原因是，允许 $\mathscr{F}$ 是 2^Ω 的子集给我们提供了一些便利，这一点在未来学习条件期望、随机过程的时候就可以看到，我们更多的会把 $\mathscr{F}$ 解释成某种意义上的“信息”。

如果你同意 $\mathscr{F}$ 不是一定要取 2^Ω，那么我们就要给其加一些限制，因为并不是 2^Ω 的每一个子集都合适当成事件集的。在这儿，我们要求 $\mathscr{F}$ 构成一个 σ-代数，又称 σ-域。

定义 2.1 (σ-代数)

我们说集合族 $\mathscr{F} \subseteq 2^{\Omega}$ 是一个 σ-代数，如果其满足：

1. $\varnothing \in \mathscr{F}, \Omega \in \mathscr{F}$.
2. 如果事件 $A \in \mathscr{F}$，则它（在 Ω 下）的补集 $A^c \in \mathscr{F}$.
3. 如果可数（或者有限个）事件 $A_1, A_2, ..., A_n, \cdots \in \mathscr{F}$，则 $\bigcup_{n\geq 1} A_n \in \mathscr{F}$.

在上面的定义中，$\bigcup_{n\geq 1} A_n := \{\omega \in \Omega : \exists n \geq 1, \omega \in A_n\}$。

对于 σ-代数定义里的三条要求，实际上是与我们在生活中对于随机试验的直观理解是非常对应的。首先第一条告诉我们，我们需要有“不可能事件”和“必然事件”这两个事件。第二条告诉我们，如果 A 是一个合理的事件，那么“A 不发生”也应该是一个合理的事件。第三条是说，如果 $A_1, ..., A_n, ...$ 都是合理的事件，那么“$A_1, ..., A_n, ...$ 中至少有一个事情发生”也应该是一个合理的事件。

三元组中第三项 $\mathbb{P} : \mathscr{F} \to [0, 1]$ 给每一个事件赋予一个 $[0, 1]$ 之间的数，表示这个事件发生的概率，被称之为概率测度。由于 $\mathscr{F}$ 是一个 σ-代数，我们相对应的对于 $\mathbb{P}$ 也有要求。

定义 2.2 (概率测度)

函数 $\mathbb{P} : \mathscr{F} \to [0, 1]$ 被称为概率测度，如果它满足：

1. $\mathbb{P}(\varnothing) = 0, \mathbb{P}(\Omega) = 1$.
2. 对于任意 $A \in \mathscr{F}, \mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.
3. 对于任意的不相交的 $A_1, A_2, ..., A_n, \cdots \in \mathscr{F}, \mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$.

其中最后一条里面的 $\sum_{n\geq 1} \mathbb{P}(A_n)$ 意思是 $\lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(A_n)$。这个极限显然是存在的。当然，这里的几条定义是有冗余的，比如第二条是第一条和第三条的推论，我们这样写的目的是为了和 σ-代数的要求对应起来。如果你把样本集合 Ω 直观上解释成随机试验中的事件 A，那么，我想对于 $\mathbb{P}$ 的以上几条公理要求都是很自然的。

也许有人会问，为什么 $\mathbb{P}$ 要定义为在事件集 $\mathscr{F}$ 上的函数，而不是样本集 Ω 上的函数呢？事实上，在 Ω 是离散的（可数或者有限）时候，这两种定义并没有多大区别。我们假设 $\mathscr{F} = 2^{\Omega}$，那么根据我们的定义，对于每一个 $\omega \in \Omega$，在单点集 $\{\omega\}$ 上有定义，我们记 $p_\omega := \mathbb{P}(\{\omega\})$。那么，根据关于 $\mathbb{P}$ 的公理的第三条，我们有

$$\forall A \in \mathscr{F}, \mathbb{P}(A) = \sum_{\omega \in A} p_\omega.$$

也就是说，在单点集 $\{\omega\}$ 上的取值决定了它的全部取值。但值得注意的是，在 Ω 不可数时，这种做法是不一定行的通的，我们不得不按照公理的形式把 $\mathbb{P}$ 定义成事件上的函数。

所以我们可以这样给出扔两个独立六面骰子的概率空间。

例题 2.1. 独立六面骰子的概率空间

我们令 $\Omega = [6]^2, \mathscr{F} = 2^{\Omega}$，并且对于任何单点 $\omega \in \Omega, \mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{36}$。

2.2 基本性质

从上面关于概率公理的讨论可以看到，所谓概率，完全就是对于集合的操作。特别当 Ω 是离散的时候，计算概率无非就是组合计数。这是事实，但在很多时候，我们希望把纯粹对于概率空间的操作和实际背后的随机试验结合起来，这样能够给出我们一些重要的直观。这些直观能够帮助我们从“概率”的视角去看待问题。在未来，你一定会发现，有很多在概率视角看起来显然正确，无比简单的事情，如果你机械的把它翻译成概率空间上的对集合或者元素的数数问题，就会变的无比笨拙或者繁琐。

我们首先可以看一些简单的布尔代数和随机试验之间的对应。

集合视角	随机试验视角
A	事件 A 发生
$A \cup B$	事件 A 和 B 至少有一个发生
$A \cap B$	事件 A 和 B 同时发生
$A \setminus B$	事件 A 发生但是 B 没有发生
$A \subseteq B$	事件 A 蕴含事件 B
$A \cap B = \varnothing$	事件 A 和 B 不可能同时发生
$A \cup B = \Omega$	事件 A 和 B 必有一个发生

上面表格里其实用到了一些 σ-代数的性质，比如我们默认了如果 $A, B \in \mathscr{F}$，那么 $A \cap B$ 以及 $A \setminus B$ 均在 $\mathscr{F}$ 中。我们马上来验证这些基本性质。在下面的讨论中，我们均假设集合 $A, B, A_1, A_2, \ldots$ 都是 $\mathscr{F}$ 中的元素。

命题 2.1

$\bigcap_n A_n \in \mathscr{F}$。

证明 我们使用 De Morgan's law，有

$$\bigcap_n A_n = \left(\bigcup_n A_n^c \right)^c.$$

由于 $\mathscr{F}$ 是一个 σ-代数，$\bigcup_n A_n^c \in \mathscr{F}$，因此其补集也在 $\mathscr{F}$ 中。

命题 2.2

$A \setminus B \in \mathscr{F}$。

证明 $A \setminus B = A \cap B^c$。由于 $\mathscr{F}$ 是一个 σ-代数，$B^c \in \mathscr{F}$，因此 $A \cap B^c \in \mathscr{F}$。

命题 2.3

$A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$。

证明 $\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$。

命题 2.4

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$。

证明 我们同样把 $A \cup B$ 拆成若干不相交的集合的并。

$$\mathbb{P}(A \cup B) = \mathbb{P}((A \setminus B) \cup (A \cap B) \cup (B \setminus A)) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A).$$

我们又注意到有 $\mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) = \mathbb{P}(A)$ 以及 $\mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B) = \mathbb{P}(B)$。于是得证。

这一个结论有一个推论，即

推论 2.1 (Union Bound)

$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$，或者更一般的有 $\mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n)$。

这个被称之为 union bound，或者是 Boole's inequality，在概率分析中有很广泛的应用，因为它无条件的给出了若干个事件至少有一个发生的概率的上界。

接着我们来定义集合序列的极限。这个定义只在集合序列是单调的时候才有意义（我们暂时先不定义上极限和下极限）。设非降集合序列 $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$，则定义

$$\lim_{n \to \infty} A_n := \bigcup_{n \geq 1} A_n.$$

类似的，对于非增的集合序列 $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$，定义

$$\lim_{n\to\infty} A_n := \bigcap_{n\geq 1} A_n.$$

我们接下来想说明，概率测度 $\mathbb{P}$，作为一个定义在集合上的函数，是“连续”的，即它可以和求极限交换。

命题 2.5

设 $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$，则有 $\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$。

设 $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$，则有 $\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$。

证明 我们只证明非降的情况，非增的情况可以从它使用 De Morgan's law 得到。使用极限的定义以及单调性，我们有

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}(\bigcup_{n\geq 1} A_n) = \mathbb{P}(A_1 \cup \bigcup_{n\geq 2}(A_n \setminus A_{n-1})).$$

这样我们便把 $\lim_{n\to\infty} A_n$ 写成了一堆集合的不交并，于是使用 $\mathbb{P}$ 之公理第三条，可以得到

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}(A_1) + \sum_{n\geq 2} \mathbb{P}(A_n \setminus A_{n-1}) = \mathbb{P}(A_1) + \lim_{N\to\infty} \sum_{n=2}^{N} \mathbb{P}(A_n \setminus A_{n-1}).$$

由于 A_{n-1} 是 A_n 的子集，再次使用 $\mathbb{P}$ 之公理第三条，我们有

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}(A_1) + \lim_{N\to\infty} \sum_{n=2}^{N} (\mathbb{P}(A_n) - \mathbb{P}(A_{n-1})) = \lim_{N\to\infty} \mathbb{P}(A_N).$$

2.3 为什么 $\mathscr{F}$ 不能总取 2^Ω？

我们一开始说了，在 Ω 是不可数的时候，如果选 $\mathscr{F} = 2^\Omega$，也许没有办法定义出合适的概率测度。我们这儿给出一个证明。我们取 $\Omega = [0,1), \mathscr{F} = 2^\Omega$，并且试图在上面定义一个均匀分布 $\mathbb{P}$。那么 Ω 上的均匀分布应该满足对于每一个集合 $I \subseteq \Omega$，给出一个“长度”，而对于这个长度，我们有一些最基本的期待：首先是对于比如区间 $(a,b) \subset [0,1)$，这个长度应该就是 $b-a$。另外就是所谓的“平移不变性”，也就是说如果我们把某个集合 I 整体平移一个距离 r，那么它的长度应该是不变的，即 $\mathbb{P}(I) = \mathbb{P}(I+r)$，这里 $I+r$ 是把 I 平移了 r 之后的集合，即

$$I + r = \{(x+r) \bmod 1 : x \in I\},$$

其中记号 $k \bmod 1$ 的意思是如果 k 不在 $[0,1)$ 内的话，则把 k 加上或者减去整数使得其属于区间 $[0,1)$。

我们首先在 Ω 上定义一个等价关系：$x \sim y$ 当且仅当 $x - y \in \mathbb{Q}$。那么，根据等价关系的基本性质，这个等价关系诱导出的等价类构成了 Ω 的一个划分，即 $\Omega = \bigcup_{i\in I} \mathbb{P}_i$，满足对于 $i \neq j$，$\mathbb{P}_i \cap \mathbb{P}_j = \varnothing$ 并且任意 $x, y \in \mathbb{P}_i$，$x \sim y$。

我们现在从每一个 $\mathbb{P}_i$ 中选出一个元素 s_i 来，把它们放在一起，构成集合 N，即 $N = \{s_i : i \in I\}$（这一步需要选择公理保证）。我们对于每一个 $r \in \mathbb{Q} \cap [0,1)$，我们考虑集合 $N_r := N + r = \{(x+r) \bmod 1 : x \in N\}$。

我们首先证明 $\{N_r\}_{r\in\mathbb{Q}\cap[0,1)}$ 构成了 $[0,1)$ 的一个分划。首先，任意一个 $x \in [0,1)$ 一定属于某个 N_r。实际上，假设 $x \in \mathbb{P}_i$，那么 $x - s_i \in \mathbb{Q}$，因此 $x \in N + (x - s_i) = N_r$ 其中 $r = x - s_i \in \mathbb{Q} \cap [0,1)$。另一方面，如果同时 $x \in N_{r_1} \cap N_{r_2}$，说明我们可以找到两个不同的 $s, s' \in N$，使得 $s + r_1 = x = s' + r_2$。这样的话，$s - s'$ 就会是一个有理数，与我们对 N 的构造矛盾。

知道了 $\{N_r\}$ 是 Ω 的一个分划之后，我们的公理就保证了，一定有

$$\mathbb{P}(\Omega) = \sum_r \mathbb{P}(N_r).$$

我们知道上式的左手边等于 1，而且由于平移不变性，每一个 $\mathbb{P}(N_r)$ 都是相同的，并且都等于 $\mathbb{P}(N)$。因此，我们有可数个 $\mathbb{P}(N)$ 相加等于 1。但这显然是不可能的：如果 $\mathbb{P}(N) = 0$，则右手边等于 0；如果 $\mathbb{P}(N) > 0$，则

偶手边为无穷大。

这个证明告诉我们，2^Ω 里面的集合太多了，以至于我们没有办法给每个集合一个合法的长度。所以我们应该如何设定 $\mathscr{F}$ 从而定义出 $[0,1)$ 上的均匀分布呢？我们刚才说了，直观上对于区间 $(a,b)\in[0,1)$，我们一定要有 $\mathbb{P}((a,b))=b-a$。因此我们要把所有的区间都放到 $\mathscr{F}$ 里面去。我们考虑能省则省的原则，认为这就够了。由于我们要求 $\mathscr{F}$ 一定要是一个 σ-代数，我们可以取 $\mathscr{F}$ 为包含所有区间的“最小”的那个 σ-代数，这个被称之为 Borel 代数。我们接着定义这个最小的概念以及说明它总是存在的。

首先我们说明一下 σ-代数是对求交封闭的。固定样本空间 Ω。假设对于（可能不可数的）指标集 I 以及每一个 $\alpha\in I$，$\mathscr{F}_\alpha\subseteq 2^\Omega$ 均为 σ-代数，则有 $\mathscr{F}:=\bigcap_{\alpha\in I}\mathscr{F}_\alpha$ 也是 σ-代数。这件事情的证明非常简单，按照 σ-的定义逐条验证即可。值得说明的事情，如果把 $\bigcap$ 换成 $\bigcup$ 就不一定对了。

设 $\mathscr{G}\subseteq\Omega$ 是 Ω 的一些子集的集合（不一定是 σ-代数），我们用 $\sigma(\mathscr{G})$ 表示包含 $\mathscr{G}$ 的最小的 σ-代数。

- 首先，$\sigma(\mathscr{G})$ 是一个 σ-代数；
- 对于任何 $\mathscr{F}'\subsetneq\sigma(\mathscr{G})$，均不是 σ-代数。这便是“最小”的意思。

对于任何 $\mathscr{G}$，$\sigma(\mathscr{G})$ 总是存在的。这是由于首先 2^Ω 本身是包含了 $\mathscr{G}$ 的一个 σ-代数。因此，我们可以取 $\sigma(\mathscr{G})$ 为所有包含了 $\mathscr{G}$ 的 σ-代数的交。它一定是存在的，而且根据交的定义也一定是最小的。

第 3 章 事件的条件概率

上节课我们介绍了概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 的定义，以及它们需要满足的公理。很多关于概率空间的一些性质和操作，都可以解释成集合上对应的性质和操作。我们今天继续介绍一些性质，它们同样可以看成集合的性质，但本身也有很丰富的概率含义。可以看到，在实际应用中，我们实际思考的对象是这些集合后面的概率。

3.1 条件概率及性质

我们前面介绍过，概率空间，以及相关的记号，都是为了抽象出我们在日常做概率试验的时候的一些直观的事情。对于扔两个骰子的问题，我们可以问，在已知两个骰子的和是偶数的情况下，第一个骰子是 6 的概率。经验告诉我们，这个概率应该是用“第一个是 6 并且和为偶数”的概率，除上“和为偶数的概率”。数学上，我们把这个称之为 条件概率，是定义在两个事件 A 和 B 上，用来表示在已知 B 发生的情况下 A 发生的概率。我们用记号 $\mathbb{P}(A \mid B)$ 来表示，它仅在 $\mathbb{P}(B) > 0$ 的时候有定义。

定义 3.1 (事件的条件期望)

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

注意到这个是条件概率的定义式，我们可以把它写成

$$\mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A \mid B).$$

这个形式可以更方便的把它想象成“事件 A 和 B 同时发生的概率，等于 B 发生的概率，乘上在已知 B 发生的时候 A 发生的概率”。

对于 n 个事件 $A_1, \ldots, A_n$，我们不停地使用上式，可以得到

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{k=1}^{n} \mathbb{P}\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right).$$

这个式子被称之为条件概率的链式法则。他可以读作：

如果我们想计算 n 个事件同时发生的概率，我们先用第一个事件发生的概率，乘上在第一个事件发生的情况下第二个事件发生的概率，再乘上在第一第二个事件都发生的情况下，第三个事件发生的概率，...

我们可以通过一些简单的例子来验证我们对于这个概念的理解。

例题 3.1.

假设我有正好两个孩子，其中至少一个是女孩，那么两个都是女孩的概率是多大?

在计算一个概率问题的时候，总是要先对概率空间进行合适的建模。这个问题的概率空间我们可以取样本空间为四元组 $\Omega = \{FF, FM, MF, MM\}$，$\mathscr{F} = 2^{\Omega}$，$\mathbb{P}$ 为均匀分布。其中事件 $B :=$ “至少一个是女孩” $= \{FF, FM, MF\}$；事件 $A :=$ “两个都是女孩” $= \{FF\}$。

所以

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{FF\})}{\mathbb{P}(\{MF, FM, FF\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

例题 3.2.

假设面前有三个抽屉，每个抽屉里有两块硬币，分别是 “金币 + 金币”，“金币 + 银币”，“银币 + 银币”。现在我随机选一个抽屉，并且随机在这个抽屉里选一个硬币，请问在选出的硬币是金币的情况下，和选出的硬币同一个抽屉的另一个也是金币的概率是多大?

在这个问题里，我们可以用 $\Omega = \{一, 二, 三\} \times \{1, 2\}$，$\mathscr{F} = 2^{\Omega}$，$\mathbb{P}$ 为均匀分布来进行建模。对于样本点 (i, j)，我们想表达的意思是选出来的硬币是第 i 个抽屉里的第 j 个硬币（硬币的顺序就按照题干里说的那般）。那么定义事件

$$B = \text{“选出的硬币是金币”} = \{(一, 1), (一, 2), (二, 1)\};$$

$$A = \text{“和选出的硬币同一个抽屉的另一个也是金币”} = \{(一, 1), (一, 2), (二, 2)\}.$$

因此，我们有

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

3.2 独立性

如果对于事件 A 和 B，我们有 $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$，或者等价的 $\mathbb{P}(A \mid B) = \mathbb{P}(A)$，我们就称 A 和 B 是独立的，有时候会记作 $A \perp B$。直观上，事件 A 和 B 独立表示 A 或者 B 是否发生，对于对方是否发生没有影响。比如我们扔两次骰子，为事件 “第一次扔的是 3”，为事件 “第二次扔的是 4”，则这两件事情是独立的。独立性，我认为，是一个非常具有概率风味的概念。尽管到目前为止，如果 Ω 是有限集，我们之前说过，所有的这些概率都可以用组合计数来研究，但独立性提供的直观，用计数的语言来说，并不是特别方便。

我们可以把独立的概念推广到更多的事件上去。我们说事件 $A_1, A_2, \ldots, A_n$ 是**相互独立（mutually independent）**的，如果其满足对于任意的 $I \subseteq [n]$，

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

这个定义看起来非常强，因为它要求上述等式对于每一个 $I \subseteq [n]$ 都成立。

注解

1. 如果只要求 $I = [n]$ 是不够的。假设 $n = 3$，$A_1 = A_2$，$A_3 = \varnothing$，则 $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = 0$。但显然这仨不一定独立。事实上，我们可以把独立的定义写成

$$\mathbb{P}\left(\bigcap_{i=1}^{n} B_i\right) = \prod_{i=1}^{n} \mathbb{P}(B_i),$$

其中 $B_i = A_i$ 或者 Ω。

2. 如果只要求等式对于 $I \in \binom{[n]}{2}$ 成立是不够的。一个经典的例子是概率空间为 $\Omega = \{HH, HT, TH, TT\}$ 上的均匀分布，即投两次均匀硬币的结果（H=Head，T=Tail）。我们考虑 $A_1 =$ “第一次硬币出 H”，$A_2 =$ “第二次硬币出 H”，$A_3 =$ “两次硬币结果不一样”。可以验证，这三个事件对于任何 $I \in \binom{[3]}{2}$，是满足独立性定义的等式的，但对于 $I = [3]$ 并不满足。实际上，如果该等式仅对于 $I \in \binom{[n]}{2}$ 成立，我们称 $A_1, \dots, A_n$ 是**两两独立（pairwise independent）**的。这是一个在计算机科学里很重要的概念，原因在于，我们在设计算法的时候，独立的随机数本身是一个重要的资源，我们需要代价才能够产生它们。而很多问题里，对于独立的要求并没有像定义那么强，有的时候两两独立就足够满足我们的要求了。而生成两两独立的随机数，代价要小很多。

对于无穷多个事件 $\{A_j\}_{j\in J}$，我们说它们相互独立，但且仅当对于 J 的任何一个有限子集 I，$\{A_i\}_{i\in I}$ 相互独立。

3.3 全概率公式

假设事件 $B_1, B_2, \dots, B_n, \dots \in \mathscr{F}$ 构成了样本空间的一个分划，即 $\Omega = \bigcup_{i\geq 1} B_i$ 并且对于 $i \neq j$，$B_i \cap B_j = \varnothing$。根据集合论的知识我们知道，对于任何集合 A，$A \cap B_1, A \cap B_2, \dots, A \cap B_n, \dots$ 也构成了 A 的一个分划。如果我们取 $A \in \mathscr{F}$，则根据概率论的公理，我们有

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n\geq 1}(A \cap B_n)\right) = \sum_{n\geq 1}\mathbb{P}(A \cap B_n).$$

这便是全概率公式（Law of total probability）。我们可以把 $\mathbb{P}(A \cap B_n)$ 用条件概率写出来，即

$$\mathbb{P}(A) = \sum_{n\geq 1}\mathbb{P}(B_n)\cdot\mathbb{P}(A \mid B_n).$$

用人话说，就是想算 A 的概率，可以先按照 B_n 的情况进行分类再全部加起来，而每一类的概率是 B_n 发生的概率乘上在 B_n 发生的情况下 A 发生的概率。

3.4 一些例子

我们接下来使用上述工具，来计算一些例子。

例题 3.3. 抗原检测

假设人群中感染新冠的概率是 0.2%，而抗原检测的准确性是 99%。现在你做了个检测发现阳了，那真的感染新冠的概率是多大?

这个问题的结论有时候会产生一些悖论。因为看起来抗原检测的准确性非常高，因此如果检测阳性，那应该大概率真的被感染了。但通过计算我们会发现不然。

首先我们用概率空间来进行建模。我们假设人群有 N 个人，因此样本空间可以取 $\Omega = [N] \times \{\pm 1\} \times \{\pm 1\}$。这里面的每一个三元组 (a, i, j)，其中 a 表示人的 id，i 表示是否被感染，j 表示是否抗原阳性。可以思考一下，我们如何设置 $\mathbb{P}$ 来满足题设的要求，这总是可以做到的。

我们来计算在已知抗原阳性的情况下，没有被感染的概率 $\mathbb{P}(\text{没感染} \mid \text{阳性})$。我们使用条件概率的定义，有

$$\mathbb{P}(\text{没感染} \mid \text{阳性}) = \frac{\mathbb{P}(\text{没感染} \cap \text{阳性})}{\mathbb{P}(\text{阳性})}.$$

我们希望把这个分式上下都写成我们题设里告诉能算的东西，因此需要用我们之前介绍的公式来改写一下。对于分子，我们有

$$\mathbb{P}(\text{没感染} \cap \text{阳性}) = \mathbb{P}(\text{阳性} \mid \text{没感染}) \cdot \mathbb{P}(\text{没感染}).$$

而对于分母，我们使用全概率公式，有

$$\mathbb{P}(\text{阳性}) = \mathbb{P}(\text{阳性} \cap \text{感染}) + \mathbb{P}(\text{阳性} \cap \text{没感染}) = \mathbb{P}(\text{阳性} \mid \text{感染}) \cdot \mathbb{P}(\text{感染}) + \mathbb{P}(\text{阳性} \mid \text{没感染}) \cdot \mathbb{P}(\text{没感染}).$$

这样一番操作之后，我们会发现每一个量我们都会计算了，即

$$\mathbb{P}(\text{阳性} \mid \text{没感染}) = 1\%, \quad \mathbb{P}(\text{没感染}) = 99.8\%, \quad \mathbb{P}(\text{阳性} \mid \text{感染}) = 99\%, \quad \mathbb{P}(\text{感染}) = 0.2\%,$$

把这些数字代进去，我们可以得到

$$\mathbb{P}(\text{没感染} \mid \text{阳性}) \approx 83.4\%.$$

也就是说，没事去测个抗原，即使阳性了，大概八成概率也没有被感染。

上面这个计算过程有时候也被称为贝叶斯公式（Bayes' formula）。

例题 3.4. 双胞胎

双胞胎有两种，同卵双生和异卵双生。一家医院想知道所有双胞胎中同卵双生的比例有多大，但是做这个检测的成本太高了。统计学家说，其实你只要统计一下每一对双胞胎的性别，就能推算出这个比例。我们知道，同卵双生子的性别一定是相同的，而异卵双生子的性别是独立的。因此，使用全概率公式，我们有

$$\mathbb{P}(\text{同性}) = \mathbb{P}(\text{同性} \mid \text{同卵}) \cdot \mathbb{P}(\text{同卵}) + \mathbb{P}(\text{同性} \mid \text{异卵}) \cdot \mathbb{P}(\text{异卵}) = 1 \cdot \mathbb{P}(\text{同卵}) + 0.5 \cdot (1 - \mathbb{P}(\text{同卵})).$$

因此，我们只要简单的统计出 $\mathbb{P}(\text{同性})$，即所有双胞胎中同性别的比例，便可以用上面的公式解出双胞胎中同卵双生的比例。

例题 3.5. 生日悖论

假设一个班级有 n 位同学（$n \geq 3$），每个人的生日都等可能地出现在一年中的 m 天里，并且每个人的生日是相互独立的。我们定义事件 $A_{i,j}$ 表示第 i 位同学和第 j 位同学的生日在同一天。

这个看似简单的问题，其结论常常令人惊讶。直觉上，如果一年有 $m = 365$ 天，那么需要很多人（比如超过 100 人）才能让“至少有两人同一天生日”的概率变得很大。但计算结果会告诉我们，其实只需要很少的人就能达到很高的概率。

我们先证明，任意两个不同的事件 $A_{i,j}$ 和 $A_{k,l}$ 是独立的。

对于任意 $i \neq j$，显然有 $\mathbb{P}(A_{i,j}) = \frac{1}{m}$，因为第 j 个人的生日与第 i 个人相同的概率是 $\frac{1}{m}$。

现在考虑两个事件 A_{i_1,j_1} 和 A_{i_2,j_2}，其中 $(i_1, j_1) \neq (i_2, j_2)$。我们需要分两种情况讨论：

情况一：四个下标 i_1, j_1, i_2, j_2 互不相同。此时，事件 A_{i_1,j_1} 和 A_{i_2,j_2} 涉及的是四组完全独立的人。因此，

$$\mathbb{P}(A_{i_1,j_1} \cap A_{i_2,j_2}) = \frac{m \cdot m \cdot m^{n-4}}{m^n} = \frac{1}{m^2} = \mathbb{P}(A_{i_1,j_1}) \cdot \mathbb{P}(A_{i_2,j_2}).$$

情况二：两个事件共享一个下标，例如 $i_1 = i_2$ 且 $j_1 \neq j_2$。此时，事件 A_{i_1,j_1} 和 A_{i_2,j_2} 意味着第 i_1 个人的生日同时等于第 j_1 个人和第 j_2 个人的生日，即三个人生日相同。因此，

$$\mathbb{P}(A_{i_1,j_1} \cap A_{i_2,j_2}) = \frac{m \cdot m^{n-3}}{m^n} = \frac{1}{m^2} = \mathbb{P}(A_{i_1,j_1}) \cdot \mathbb{P}(A_{i_2,j_2}).$$

综上所述，任意两个不同的事件 $A_{i,j}$ 和 $A_{k,l}$ 都满足独立性的定义，因此事件族 $\{A_{i,j}\}_{i,j \in [n], i<j}$ 是两两独立的。

然而，这组事件并非相互独立。要证明这一点，我们只需找到一个反例。

考虑所有事件的交集 $\bigcap_{i,j\in[n],i<j} A_{i,j}$。这个事件表示“班上所有 n 个人的生日都在同一天”。它的概率为：

$$\mathbb{P}\left(\bigcap_{i,j\in[n],i<j} A_{i,j}\right) = \frac{m}{m^n} = \frac{1}{m^{n-1}}.$$

另一方面，如果我们假设所有事件相互独立，那么它们的联合概率应该等于各自概率的乘积：

$$\prod_{i,j\in[n],i<j} \mathbb{P}(A_{i,j}) = \left(\frac{1}{m}\right)^{\binom{n}{2}} = \frac{1}{m^{\binom{n}{2}}}.$$

由于当 $n \geq 3$ 时，$\binom{n}{2} = \frac{n(n-1)}{2} > n-1$，所以

$$\frac{1}{m^{\binom{n}{2}}} \neq \frac{1}{m^{n-1}}.$$

因此，事件族 $\{A_{i,j}\}_{i,j\in[n],i<j}$ 不是相互独立的。

现在我们来解决最经典的问题：取 $m = 30$，请问至少需要多少人 (n)，才能使“班上存在两人同生日”的概率大于 $\frac{1}{2}$？

记事件 $B = \bigcup_{i,j\in[n],i<j} A_{i,j}$，即“至少有两人同生日”。我们通常用其补事件“所有人生日都不同”来计算：

$$\mathbb{P}(B) = 1 - \mathbb{P}(B^c) = 1 - \mathbb{P}\left(\bigcap_{i,j\in[n],i<j} A_{i,j}^c\right).$$

根据排列组合，所有人生日都不同的概率为：

$$\mathbb{P}(B^c) = \frac{m\cdot(m-1)\cdot(m-2)\cdots(m-n+1)}{m^n} = \frac{m!}{(m-n)!\cdot m^n}.$$

因此，

$$\mathbb{P}(B) = 1 - \frac{m!}{(m-n)!\cdot m^n}.$$

我们代入 $m = 30$ 进行计算：- 当 $n = 6$ 时，$\mathbb{P}(B) \approx 1 - \frac{30\cdot29\cdot28\cdot27\cdot26\cdot25}{30^6} \approx 0.4136$。- 当 $n = 7$ 时，$\mathbb{P}(B) \approx 1 - \frac{30\cdot29\cdot28\cdot27\cdot26\cdot25\cdot24}{30^7} \approx 0.5308$。

因此，当 n 至少为 7 时，班上存在两人同生日的概率才大于 $\frac{1}{2}$。

这个结果再次印证了“生日悖论”的惊人之处：即使在只有 30 天的“年份”里，也只需要 7 个人，就有超过一半的概率出现生日重复！

例题 3.6. 秘书问题

某公司需要招聘一名秘书，共有 n 位应聘者。为了节省时间和经济成本，公司决定采用一种“最优停止”策略：首先面试前 K 位应聘者，仅记录他们的表现作为参考标准，但不录用任何人；从第 $K+1$ 位应聘者开始，只要遇到一位比之前所有面试者都优秀的候选人，就立即录用此人，后续的应聘者不再考虑；如果直到最后都没有找到这样的人选，则招聘失败。

这个问题的结论常常令人拍案叫绝。直觉上，我们可能会认为应该在中间某个位置开始选择，或者干脆直接选最后一个。但通过严谨的概率计算，我们会发现存在一个最优的“观察期” K，使得成功选到最佳人选的概率最大化。

令事件 A 表示公司最终成功招聘到 n 人中的最佳人选。对于每个 $j \in [n]$，令事件 B_j 表示最佳人选恰好排在第 j 个面试的位置。

我们需要计算条件概率 $\mathbb{P}(A|B_j)$: 如果 $j \leq K$，即最佳人选出现在前 K 位面试者中，那么根据策略，公司不会录用他/她，因此 $\mathbb{P}(A|B_j) = 0$; 如果 $j > K$，即最佳人选出现在第 $K+1$ 位或之后。此时，事件 A 发生当且仅当在前 $j-1$ 位面试者中，实力最强的人恰好出现在前 K 位之中。因为只有这样，第 j 位的最佳人选才会被识别为“比之前所有人都好”。由于前 $j-1$ 位面试者的实力排序是随机的，最强者等可能地出现在这 $j-1$ 个位置中的任何一个，因此其落在前 K 个位置的概率是 $\frac{K}{j-1}$。

综上所述：

$$\mathbb{P}(A|B_j) = \begin{cases} 0, & \text{若} j \le K, \\ \frac{K}{j-1}, & \text{若} j > K. \end{cases}$$

根据全概率公式，我们可以计算出总体的成功概率 $\mathbb{P}(A)$：

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(A \cap B_j) = \sum_{j=1}^{n} \mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j).$$

由于 $\mathbb{P}(B_j) = \frac{1}{n}$ 对于所有 j 都成立，代入上式得：

$$\mathbb{P}(A) = \sum_{j=K+1}^{n} \frac{K}{j-1} \cdot \frac{1}{n} = \frac{K}{n} \sum_{j=K+1}^{n} \frac{1}{j-1} = \frac{K}{n} \sum_{j=K}^{n-1} \frac{1}{j}.$$

当 n 和 K 都足够大时，我们可以用积分来近似求和：

$$\sum_{j=K}^{n-1} \frac{1}{j} \approx \int_{K}^{n} \frac{1}{x} dx = \ln n - \ln K = \ln \frac{n}{K}.$$

因此，

$$\mathbb{P}(A) \approx \frac{K}{n} \ln \frac{n}{K} = -\frac{K}{n} \ln \frac{K}{n}.$$

令 $x = \frac{K}{n}$，则 $\mathbb{P}(A) \approx -x \ln x$。这是一个经典的优化问题，函数 $f(x) = -x \ln x$ 在区间 $(0,1)$ 上的最大值出现在 $x = \frac{1}{e}$ 处，最大值为 $\frac{1}{e} \approx 0.3679$。

因此，最优的策略是设置 $K \approx \frac{n}{e}$。在实际操作中，我们应取最接近 $\frac{n}{e}$ 的整数，即 $K = \lfloor \frac{n}{e} \rfloor$ 或 $K = \lceil \frac{n}{e} \rceil$。

此时，成功选到最佳人选的概率约为 36.79%。这个结果非常惊人——即使有成百上千的候选人，我们只需观察大约前 37% 的人，然后选择后面第一个比他们都优秀的人，就能以超过三分之一的概率获得最佳人选！

例题 3.7. 无限悖论

假设我们有可数无穷个球，用 $k = 1, 2, \ldots$ 来编号，并且有一个无限大的箱子。我们考察一个“放球”与“拿球”的过程。首先是放球：在 12 点前 1 分钟，我们把 $1, 2, \ldots, 10$ 号球放进去；在 12 点前 $\frac{1}{2}$ 分钟，我们把 $11, 12, \ldots, 20$ 号球放进去；在 12 点前 $\frac{1}{4}$ 分钟，我们把 $21, 22, \ldots, 30$ 号球放进去，以此类推。对于 $n = 0, 1, 2, \ldots$，我们在 12 点前 2^{-n} 分钟把 $10n+1, 10n+2, \ldots, 10(n+1)$ 号球放进去。

然后我们再使用不同的方式把球拿出来。

对于 $n = 0, 1, 2, \ldots$，我们在 12 点前 2^{-n} 分钟放完球后，把第 $10(n+1)$ 号球拿出来。

我们假设放球和拿球都是瞬时完成的，我们来计算一下，在 12 点的时候，箱子里有多少个球。显然，这种情况箱子里会有无穷多个球，因为所有编号不是 10 的倍数的球都被放了进去并且没有被拿出来。

我们换另外一种拿球的方式。

对于 $n = 0, 1, 2, \ldots$，我们在 12 点前 2^{-n} 分钟放完球后，把第 $(n+1)$ 号球拿出来。

使用这种拿球的策略，在 12 点的时候，箱子里有多少个球呢？一番思考之后，不难发现，箱子里一个球都没有！因为对于每一个 $k \in \mathbb{N}$，第 k 号球在 $n = k - 1$ 的时候被拿出来了。

和第一种情况一样，每次都是只拿了一个球出来，居然产生的结果截然不同。我现在想问，如果我随机的选一个球出来，又当如何呢？

对于 $n = 0, 1, 2, \ldots$，我们在 12 点前 2^{-n} 分钟放完球后，我们从箱子里均匀随机的拿一个球出来。

我们来计算每一个球在 12 点的时候留在箱子里的概率。我们这儿以 1 号球为例，其它的球的计算方式类似。对于每一个 $n = 0, 1, 2, \ldots$，我们用事件 A_n 来表示在第 n 轮操作之后（即 12 点前 2^{-n} 分钟放球拿球

的操作完成之后），1 号球还在箱子里这个事件。我们关注的是事件

$$A_\infty := \bigcap_{n\geq 0} A_n = \lim_{n\to\infty} A_n$$

发生的概率。根据我们上一节课介绍的概率测度的连续性，我们有 $\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$。因此，我们只需要计算 $\mathbb{P}(A_n)$，第 n 轮结束后 1 号球在箱子里的概率。对于 $n = 0, 1, \ldots$，我们再定义一个事件 B_n，用来表示，在第 n 轮的时候拿出的球不是 1 号球，则有 $A_n = B_0 \cap B_1 \cap \cdots \cap B_n$。
使用条件概率的链式法则，我们可以得到

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} B_i\right) = \prod_{k=0}^{n} \mathbb{P}\left(B_k \,\middle|\, \bigcap_{i<k} B_i\right).$$

我们把 $\mathbb{P}(A_n)$ 写成这样的目的是因为对于每一个 $k = 0, 1, \ldots, n$，条件概率 $\mathbb{P}(B_k \mid \bigcap_{i<k} B_i)$ 有着自然的组合意义，并非常好计算：在 $\bigcap_{i<k} B_i$ 的条件下，这对应于箱子里有 $9(k+1)+1$ 个球，其中一个 1 号球，我们拿了一个球，但它不是 1 号球的概率。显然，这个概率是 $1 - \frac{1}{9(k+1)+1}$。所以，我们有

$$\mathbb{P}(A_n) = \prod_{k=0}^{n} \left(1 - \frac{1}{9(k+1)+1}\right) \leq e^{-\sum_{k=0}^{n} \frac{1}{9k+10}}.$$

我们知道级数 $\sum_{k=0}^{n} \frac{1}{9k+10}$ 是发散的，因此，$\lim_{n\to\infty} \mathbb{P}(A_n) = 0$。
所以，我们知道了，在 12 点的时候，1 号球在箱子里的概率是 0。我们用 S_i 来表示第 i 号球在 12 点的时候还在箱子里的概率。我们刚才计算得知 $\mathbb{P}(S_1) = 0$。可以使用类似的方法，我们能够得到对于任意 $n \in \mathbb{N}$，$\mathbb{P}(S_n) = 0$。我们现在关心的是在 12 点的时候箱子里有球的概率，即 $\mathbb{P}(\exists n \in \mathbb{N}, S_n)$。我们使用上节课讲过的 union-bound，可以得到

$$\mathbb{P}(\exists n \in \mathbb{N}, S_n) \leq \sum_{n\geq 1} \mathbb{P}(S_n) = 0.$$

3.5 Example: 二分图完美匹配的概率与容斥原理

在本节中，我们将探讨一个有趣的概率问题：在一个给定的二分图上，随机选取一个映射，它恰好构成一个完美匹配的概率是多少？考虑一个无向且连通的二分图 $G = (V, E)$，其中顶点集被划分为两个大小相等的部分：$V = V_1 \cup V_2$，满足 $|V_1| = |V_2| = n$ 且 $V_1 \cap V_2 = \varnothing$。我们假设 V_1 和 V_2 内部均无边相连。

我们可以用一个 $n \times n$ 的邻接矩阵 $A \in \{0,1\}^{n\times n}$ 来描述这个图。矩阵元素 $A_{i,j}$ 定义如下：

$$A_{i,j} = \begin{cases} 1, & \text{如果} V_1 \text{ 中的第} i \text{ 个点与} V_2 \text{ 中的第} j \text{ 个点之间有边,} \\ 0, & \text{否则.} \end{cases}$$

我们的目标是：从所有可能的映射 $f : [n] \to [n]$ 中均匀随机地选取一个，计算它恰好对应图 G 的一个完美匹配的概率。

这里，样本空间 Ω 是所有映射的集合，记为 $T_n = \{f : [n] \to [n] \mid \forall i \in [n], A_{i,f(i)} = 1\}$。注意，这里的映射 f 不一定是置换，它只要求对于每个 i，其像 $f(i)$ 在 V_2 中对应的点必须与 i 相连。我们关心的事件 M 是“所选映射 f 正好是一个完美匹配”，即 f 必须是一个置换函数，并且对所有的 i 都有 $A_{i,f(i)} = 1$。

首先，我们直接写出事件 M 的概率。由于概率测度 $\mathbb{P}$ 是均匀分布，事件 M 的概率等于完美匹配的数量除以总的映射数量。令 S_n 表示 $[n]$ 上所有置换函数的集合。一个置换 $\sigma \in S_n$ 构成完美匹配当且仅当 $\prod_{i=1}^{n} A_{i,\sigma(i)} = 1$。因此，完美匹配的总数就是 $\sum_{\sigma\in S_n} \prod_{i=1}^{n} A_{i,\sigma(i)}$。而样本空间 T_n 的大小是 $\prod_{i=1}^{n} \left(\sum_{j\in[n]} A_{i,j}\right)$，因为对于每个

$i \in [n]$，都有 $\sum_{j\in[n]} A_{i,j}$ 种选择 $f(i)$ 的方式。

所以，我们得到第一个直观的概率表达式：

$$\mathbb{P}(M) = \frac{\sum_{\sigma\in S_n}\prod_{i=1}^{n} A_{i,\sigma(i)}}{|T_n|}.$$

这个表达式虽然清晰，但计算起来却非常困难，因为它需要枚举所有 $n!$ 个置换。为了找到一个更高效的计算方法，我们引入容斥原理。这是解决“至少有一个”或“恰好没有”这类计数问题的核心工具。我们定义一系列事件来刻画“映射的值域”。对于每个 $i \in [n]$，令事件 $E_i = \{f \in T_n \mid f^{-1}(i) = \varnothing\}$，表示“$V_2$ 中的第 i 个点没有原像”。

对于任意子集 $J \subseteq [n]$，我们定义：

- 事件 $E_J = \{f \in T_n \mid \forall j \in J, f \in E_j; \forall j \notin J, f \notin E_j\}$，表示“映射的值域恰好是 $[n] \setminus J$”。
- 事件 $E_J^+ = \{f \in T_n \mid \forall j \in J, f \in E_j\}$，表示“映射的值域是 $[n] \setminus J$ 的子集”。

显然，事件 $E_\varnothing$ 就是我们关心的“映射的值域是整个 $[n]$”，即映射是一个满射。而在我们的问题中，由于 V_1 和 V_2 大小相同，一个满射映射就是一个双射，也就是一个置换。因此，事件 $E_\varnothing$ 等价于事件 M。

根据容斥原理，我们有：

$$\mathbb{P}(M) = \mathbb{P}(E_\varnothing) = \sum_{J\subseteq[n]} (-1)^{|J|}\mathbb{P}(E_J^+).$$

这个公式的证明可以通过分析每个映射 f 在等式两边被计算的次数来完成。对于一个属于 E_I 的映射 f，它在左边只被计算一次；在右边，它会被所有满足 $I \subseteq J$ 的项 $(-1)^{|J|}\mathbb{P}(E_J^+)$ 计算，其总贡献为 $\sum_{J\supseteq I}(-1)^{|J|} = (-1)^{|I|}\sum_{k=0}^{n-|I|}\binom{n-|I|}{k}(-1)^k = 0$（当 $I \neq \varnothing$ 时），只有当 $I = \varnothing$ 时，其贡献为 1。这证明了公式的正确性。

现在，我们只需要计算 $\mathbb{P}(E_J^+)$ 即可。事件 E_J^+ 表示映射的值域不包含 J 中的任何点，即所有 $f(i)$ 都必须落在 $[n] \setminus J$ 中。这意味着，对于每一个 $i \in [n]$，我们只能从 V_2 中那些与 i 相连且不属于 J 的点中进行选择。因此，对于固定的 i，可选的点的数量是 $\sum_{j\in[n]\setminus J} A_{i,j}$。

由于每个 i 的选择是独立的，事件 E_J^+ 包含的映射总数为：

$$|E_J^+| = \prod_{i=1}^{n}\left(\sum_{j\in[n]\setminus J} A_{i,j}\right).$$

于是，我们有：

$$\mathbb{P}(E_J^+) = \frac{|E_J^+|}{|T_n|} = \frac{\prod_{i=1}^{n}\left(\sum_{j\in[n]\setminus J} A_{i,j}\right)}{|T_n|}.$$

将此代入容斥原理的公式，我们得到了著名的 **Ryser 公式**：

$$\mathbb{P}(M) = \frac{\sum_{J\subseteq[n]}(-1)^{|J|}\prod_{i=1}^{n}\left(\sum_{j\in[n]\setminus J} A_{i,j}\right)}{|T_n|}.$$

让我们比较一下两种方法的计算效率：

- 直接法需要计算分子 $\sum_{\sigma\in S_n}\prod_{i=1}^{n} A_{i,\sigma(i)}$ 需要遍历所有 $n!$ 个置换，时间复杂度为 $O(n \cdot n!)$；
- 而 **Ryser 公式**只需要对 2^n 个子集 J 进行求和。对于每个 J，计算内层乘积需要 $O(n^2)$ 时间（因为需要对每个 i 计算一个和）。因此，总的时间复杂度为 $O(n^2 \cdot 2^n)$。虽然 $O(n^2 \cdot 2^n)$ 仍然是指数级的，但它比 $O(n!)$ 要好得多。

更重要的是，在计算机科学的理论框架下，如果假设“指数时间假设”（Exponential Time Hypothesis, ETH）成立——即不存在能在 $2^{O(n)}$ 时间内解决 n 个变量布尔公式的算法——那么 Ryser 公式在渐进意义上就是计算这个概率的最优算法。

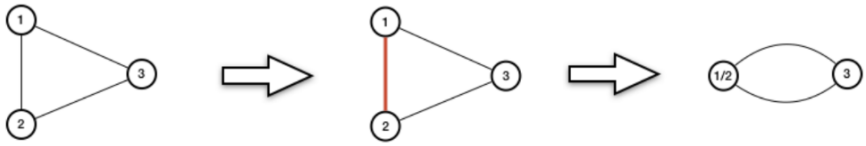
3.6 Example: Karger 的最小割算法

我们来看一个经典的算法问题，求图上的最小割。给定一个连通无向图 $G(V, E)$，我们说一个边集 $C \subseteq E$ 是一个割，当且仅当把 C 从图中删掉之后，得到的图 $G(V, E \setminus C)$ 是不连通的。最小割问题即寻找图上最小的

一个割。

学过算法的同学都知道，我们可以用最大流的方法来寻找最小割：固定一个源 s 并枚举所有可能的汇 t，对每一个源-汇对 (s,t) 求解最大流，其残量网络（residual network）给出了最小割。使用 2022 年最快的最大流算法，我们需要 $nm^{1+o(1)}$ 的时间来寻找这个最小割，其中 $n=|V|, m=|E|$。我们来介绍一个随机算法，实现非常简单，并且在有效的优化之后比基于最大流的算法要更快。

我们定义图上的所谓**缩边（contraction）**的操作。给定边 $e=\{u,v\}\in E$，我们缩掉 e 的操作指的是把 u 和 v 合并成一个点，并且删掉之前所有 u 与 v 之间的边（他们俩与别的点相连的边依然保留）。我们把缩完之后的图记作 G/e。如下图所示，我们缩掉了边 $\{1,2\}$。



Karger 算法非常简单，从 G 出发，每一次随机选择一条边，把它缩掉。重复执行 $n-2$ 次之后，图中只剩两个点了，然后输出所有剩下的边。

---

while *G contains more than two vertices* **do**
  Choose an edge e uniformly at random;
  Contract e in G;
return *remaining edges*;

这个算法“有可能”成功的原因也很简单，由于我们关心的是“最小的”割，那么算法的每一步，选到割中的边的概率都不会太大。当然了，要谈论这个概率，我们需要有合适的概率空间。一个自然的样本空间可以是算法执行过程中所有（有序的）删除的边的序列。我们可以在上面定义合适的概率测度。

我们设 C 是一个固定的最小割，并且其大小是 k。我们来计算算法最终输出 C 的概率。很容易发现，算法最终输出 C，当且仅当算法执行 while 的循环中，每一次都没有选到 C 中的边。因此，我们用 A_k 来表示上面代码里面第 k 次执行完 while 循环之后，还没有删过任何一个 C 边的这个事件。为了分析 $\mathbb{P}(A_k)$，我们对于每一个 $k=1,2,\dots,n-2$，定义事件 B_k 为“第 k 次 while 循环选择的不是 C 中边”这一事件。那么显然有 $A_k=\bigcap_{i=1}^{k}B_i$。因此，由条件概率的链式法则，

$$\mathbb{P}(\text{算法输出}C)=\mathbb{P}(A_{n-2})=\prod_{i=1}^{n-2}\mathbb{P}\left(B_i\middle|\bigcap_{j=1}^{i-1}B_j\right).$$

我们希望给上述概率一个下界，即每一轮均有比较大的概率不选到 C 中的边。由于我们每一轮是均匀的选，因此我们只需要证明，对于第 i 轮，在已知 $\bigcap_{j=1}^{i-1}B_j$，即前 $i-1$ 轮均没有选到 C 中边的情况下，图中剩余的边足够多即可。一个最重要的观察是 > 此时，图中每一个顶点的度数均不少于 k。

这件事情成立的原因是，如果有一个顶点的度数 $<k$，那么在当前的图中，这个顶点与其邻居相连的边，构成了该图中的一个割。同时，也不难发现，这个割也一定是原图 G 中的一个割（因为收缩这种操作不会破坏它是割的性质）。但这就说明，我们找到了一个大小比 k 更小的割，这与我们假设 C 是最小的割矛盾。

有了这个观察，因为我们知道算法在第 $i-1$ 轮后，还剩余 $n-i+1$ 个顶点。所以第 i 轮开始的时候，图中至少还有 $\frac{k}{2}\cdot(n-i+1)$ 条边。于是，我们有

$$\mathbb{P}\left(B_i\middle|\bigcap_{j=1}^{i-1}B_j\right)\geq 1-\frac{2k}{k(n-i+1)}=\frac{n-i-1}{n-i+1}.$$

这说明，

$$\mathbb{P}(A_{n-2})\geq\prod_{i=1}^{n-2}\frac{n-i-1}{n-i+1}=\frac{2}{n(n-1)}.$$

因此，我们的算法有 $\frac{2}{n(n-1)}$ 的概率能够输出最小割 C。我们可以重复这个算法 $N = 50n(n-1)$ 次，并且输出这 N 次中找到最小的那个割。那么，这个割是最小割的概率至少有

$$1 - \left(1 - \frac{2}{n(n-1)}\right)^N \geq 1 - e^{-100}.$$

最后我们来计算一下算法的复杂性。我们要重复算法 $N = O(n^2)$ 次，每一次要进行 $n-2$ 次收缩操作。如果我们使用并查集来维护的话，大约需要一共需要 $\tilde{O}(n^2 m)$ 的时间。这个时间是比前文提到的使用最大流的算法要慢的（都怪最大流算法发展的太快了，上一次讲的时候最大流还要 $\tilde{O}(nm)$ 呢），但好处在于实现非常简单。事实上，Karger 算法可以进一步改进，使得只需要 $\tilde{O}(n^2)$ 的时间即可。下面我们便来简要说明这一点。

Karger 算法的一个主要瓶颈在于，它需要将图一直收缩到只剩两个顶点。如果我们能在图规模还比较大的时候就停止收缩，转而用其他方法解决子问题，是否可以更快呢？答案是肯定的。这个改进的算法被称为 Karger-Stein 算法，其核心思想是：当图被收缩到一定规模时，不再继续收缩，而是递归地应用算法本身来解决子问题。

具体来说，我们定义一个函数 $\texttt{KS}(G)$，其工作流程如下：

Input: 一个无向图 $G = (V, E)$
Output: 图 G 的最小割
if $|V| \leq 6$ **then**
  使用暴力枚举法直接求出最小割;
else
  设 $t = \left\lceil 1 + \frac{|V|}{\sqrt{2}} \right\rceil$;
  $G_1 \leftarrow \texttt{contract}(G, t)$; // 将 G 收缩至 t 个顶点
  $G_2 \leftarrow \texttt{contract}(G, t)$; // 独立地再做一次
  return $\min(\mathit{KS}(G_1), \mathit{KS}(G_2))$;
end

这个算法的巧妙之处在于，它通过递归调用自身，将一个大问题分解为两个规模约为原图 $\frac{1}{\sqrt{2}}$ 倍的子问题。我们需要证明，这个递归算法以较高的概率能返回正确的最小割。

首先，我们分析 $\texttt{contract}(G, t)$ 这一步骤的成功概率。令 $p(G, t)$ 表示在将图 G 收缩至 t 个顶点的过程中，从未碰到最小割 C 中任何边的概率。根据我们在课堂上推导的结论，当 $t = \left\lceil 1 + \frac{n}{\sqrt{2}} \right\rceil$ 时，有：

$$p(G, t) = \mathbb{P}(A_{n-t}) \geq \prod_{i=1}^{n-t} \frac{n-i-1}{n-i+1} = \frac{t(t-1)}{n(n-1)} > \frac{1}{2}.$$

也就是说，单次收缩操作有超过一半的概率保留了最小割 C。

接下来，我们假设存在一个子程序 $\texttt{KS}$，对于规模为 t 的图 G'，它能以概率 $p(t) \geq \frac{c}{\log t}$ 输出正确的最小割（其中 c 是一个常数）。现在，我们从原图 G 出发，独立运行两次 $\texttt{contract}$ 得到 G_1 和 G_2，然后分别对它们调用 $\texttt{KS}$。我们关心的是，最终输出的两个结果中，较小的那个是原图最小割的概率。

记事件 B 为 “$\texttt{KS}(G_1)$ 和 $\texttt{KS}(G_2)$ 中较小的那个是原图 G 的最小割”。记事件 B_1 为 “$\texttt{KS}(G_1)$ 是 G_1 的最小割”，事件 B_2 为 “$\texttt{KS}(G_2)$ 是 G_2 的最小割”。

根据全概率公式，我们可以计算：

$$\mathbb{P}(B) = 1 - \mathbb{P}(B^c) = 1 - \left[\mathbb{P}(A_{n-t}^c) + \mathbb{P}(A_{n-t})\mathbb{P}(B_1^c)\right] \cdot \left[\mathbb{P}(A_{n-t}^c) + \mathbb{P}(A_{n-t})\mathbb{P}(B_2^c)\right].$$

代入已知的概率下界：

$$\mathbb{P}(B) > 1 - \left(1 - \frac{1}{2} \cdot \frac{c}{\log t}\right)^2 \geq \frac{c}{\log t} - \frac{c^2}{4\log^2 t} = \Omega\left(\frac{1}{\log n}\right).$$

这表明，即使在递归调用中，算法依然能保持一个与 $\frac{1}{\log n}$ 同阶的成功概率。

最后，我们分析算法的时间复杂度。设 $T(n)$ 为算法在处理 n 个顶点的图时所需的期望时间。

根据算法结构，我们有：

$$T(n) = 2T\left(\left\lceil 1 + \frac{n}{\sqrt{2}} \right\rceil\right) + O(n^2).$$

这是一个标准的分治递推式。我们可以使用主定理（Master Theorem）来求解。这里，子问题规模缩小因子为 $a = 2$，规模缩减比例为 $b = \sqrt{2}$，非递归部分为 $f(n) = O(n^2)$。由于 $\log_b a = \log_{\sqrt{2}} 2 = 2$，且 $f(n) = \Theta(n^{\log_b a})$，根据主定理的情况二，我们得到：

$$T(n) = \Theta(n^2 \log n).$$

然而，这还不是最终答案。因为我们只需要以常数概率（例如 $\frac{2}{3}$）成功即可，所以我们可以将整个算法独立重复 $O(\log n)$ 次，取其中最好的结果。这样，总的时间复杂度变为：

$$O(\log n) \cdot T(n) = O(n^2 \log^2 n) = \tilde{O}(n^2).$$

第 4 章 离散随机变量与期望

4.1 离散概率空间上的随机变量

在做随机试验的时候，我们经常会关注样本点的某些性质。比如，假设我们在班上随机选一个同学，我们会想知道该同学的身高。或者，我们随机投掷两个骰子，我们想知道两个骰子的点数和。这儿，同学的身高和骰子的点数和均是定义在样本空间 Ω 上的函数，这也是我们所谓随机变量的定义。

我们固定一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$。我们今天所有的讨论均假设 Ω 是离散的，即它是有限的或者可数的，并且 $\mathscr{F} = 2^{\Omega}$。这么做的目的是让大家尽快的接触到概率论里面的一些核心概念并建立相应的直观。事实上，本课程的主要目的之一便是在一般的样本空间上定义相关的概念，这也是我们之后会讨论的话题。

一个（实值）随机变量 X 指的是从 Ω 到 $\mathbb{R}$ 的函数，即

$$X : \omega \in \Omega \mapsto X(\omega) \in \mathbb{R}.$$

所以实际上，随机变量它既不随机，也不是变量，它仅仅是一个从样本集到实数集的函数而已。对于一般的样本空间，我们会要求 X 是可测（measurable）的。但我们这儿取的 $\mathscr{F} = 2^{\Omega}$，因此，任意一个函数均是随机变量。

比如说，在从班上随机选人的例子里，概率空间 Ω 是班上所有同学的集合，$\mathscr{F} = 2^{\Omega}$，而 $\mathbb{P}$ 是均匀测度。我们定义 $X : \omega \in \Omega \mapsto \omega$的身高, 这便是一个随机变量。

我们经常会关心一类特殊的随机变量，即所谓的指示变量。对于事件 $A \in \mathscr{F}$，我们定义

$$\mathbb{I}_A : \omega \in \Omega \mapsto \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

换句话说，$\mathbb{I}_A$ 用来指示样本点 $\omega \in \Omega$ 是否在集合 A 中。

4.2 一些新的记号

我们通常会关心关于随机变量的一些问题，比如 “随机选一个同学，身高不超过 170 的概率是多大”。我们因此需要引入一些新的记号，比如对于 $a \in \mathbb{R}$，定义记号形如 $\mathbb{P}[X \leq a]$。事实上，我们有

$$\mathbb{P}[X \leq a] := \mathbb{P}[\{X \leq a\}], \quad \text{其中} \{X \leq a\} := \{\omega \in \Omega : X(\omega) \leq a\}.$$

有了这个定义，我们便可以用 $\mathbb{P}[X \leq 170]$ 来表示随机选一个同学，身高不超过 170 的概率了。注意到，这儿的 $\mathbb{P}$ 就是概率空间里面的 $\mathbb{P}$，它的输入是 $\{X \leq a\}$，这是一个 $\mathscr{F}$ 中的集合。因此，这是良定义的。

类似的，对于任何一个集合 $A \in \mathscr{B}$，我们用 $\{X \in A\}$ 表示 $\{\omega \in \Omega : X(\omega) \in A\}$。因此，可以定义

$$\mathbb{P}[X \in A] := \mathbb{P}[\{X \in A\}].$$

我们同样可以类似的定义 $\mathbb{P}[X \geq a], \mathbb{P}[X = a]$ 等直观的记号，这儿就不再赘述了。

此外，我们有时候也用 $X^{-1}(A)$ 来表示 $\{X \in A\}$，即集合 A 在函数 X 下 Ω 中的原像。

4.3 随机变量的分布

我个人认为，关于随机变量的各种术语以及黑话里面，“分布”这个词经常被误用或者滥用，我们在这儿严格的把它定义清楚。考虑一个定义在离散概率空间 $(\Omega, \mathscr{F} = 2^{\Omega}, \mathbb{P})$ 上的随机变量 $X : \Omega \to \mathbb{R}$。设 $(\mathbb{R}, \mathscr{B})$ 是实数及其上面的 Borel 集的集合。我们可以定义出一个集合函数 $\mu_X : \mathscr{B} \to \mathbb{R}$，满足

$$\forall A \in \mathscr{B}, \quad \mu_X(A) = \mathbb{P}[X \in A].$$

那么，μ_X 被称为 X 的分布（distribution），或者是 X 的律（law）。我们接着验证，μ_X 是一个概率测度。

定理 4.1

$(\mathbb{R}, \mathscr{B}, \mu_X)$ 是一个概率空间。 ♡

证明 我们只需要使用概率空间的定义进行验证即可。

1. 首先 $\mu(\varnothing) = \mathbb{P}[X^{-1}(\varnothing)] = \mathbb{P}[\varnothing] = 0$。
2. 其次，对于任何 $A \in \mathscr{B}$，$\mu(A^c) = \mathbb{P}[X^{-1}(A^c)] = \mathbb{P}[\Omega \setminus X^{-1}(A)] = 1 - \mathbb{P}[X^{-1}(A)] = 1 - \mu(A)$。
3. 设不相交的集合 $A_1, A_2, \ldots, A_n, \cdots \in \mathscr{B}$，它们的原像 $X^{-1}(A_1), X^{-1}(A_2), \ldots, X^{-1}(A_n), \ldots$ 也是不相交的。因此

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \mathbb{P}\left[X^{-1}\left(\bigcup_{n\geq 1} A_n\right)\right] = \mathbb{P}\left[\bigcup_{n\geq 1} X^{-1}(A_n)\right] = \sum_{n\geq 1} \mathbb{P}[X^{-1}(A_n)] = \sum_{n\geq 1} \mu(A_n).$$

在我们讨论的场合里，由于 X 是定义在可数集上的函数，它的值域最多包含可数个点。我们用 $\mathsf{Im}(X) = \{x_1, x_2, \ldots\}$ 来表示。我们也因此称 X 为 离散随机变量。显然，X 的分布由 X 取 $\mathsf{Im}(X)$ 中值的概率唯一确定，即

$$\forall A \in \mathscr{B}, \quad \mu(A) = \sum_{a \in \mathsf{Im}(X) \cap A} \mathbb{P}[X = a].$$

因此，我们可以定义一个函数 $p_X : \mathbb{R} \to [0, 1]$，满足对于任何 $x \in \mathsf{Im}(X)$，$p_X(x) = \mathbb{P}[X = x]$，且在 $\mathbb{R} \setminus \mathsf{Im}(X)$ 上的定义都是零。这个被称之为概率质量函数（probability mass function, pmf）。概率质量函数唯一确定了随机变量的分布 μ_X。

4.4 分布的例子

我们来看几个分布的例子。

我们考察扔一个（不一定均匀）硬币的例子。对于给定的 $p \in [0, 1]$，定义样本空间 $\Omega = \{H, T\}$，$\mathscr{F} = 2^{\Omega}$，满足 $\mathbb{P}(\{H\}) = p$，$\mathbb{P}(\{T\}) = 1 - p$。我们考虑随机变量 $X : \Omega \to \mathbb{R}$ 满足 $X(H) = 1$，$X(T) = 0$。换句话说，随机变量把 H（表示正面）映射到了 1，把 T（表示反面）映射到了 0。

我们来看 X 定义出来的分布 μ_X。它的概率质量函数显然满足 $p_X(1) = p$，$p_X(0) = 1 - p$。我们把这样一个分布称之为参数为 p 的伯努利分布（Bernoulli distribution），记作 $\mathsf{Ber}(p)$。

我们考虑另外一个随机试验，即扔 $n \geq 1$ 个硬币，每个硬币都是以 p 的概率出现正面。这个随机试验对应的概率空间如下：$\Omega = \{H, T\}^n$ 为所有长度为 n 的 HT 串的集合；$\mathscr{F} = 2^{\Omega}$ 并且对于每一个 $s \in \Omega$，

$$\mathbb{P}(s) = p^{s \text{ 中 } H \text{ 的个数}} (1 - p)^{s \text{ 中 } T \text{ 的个数}}.$$

我们现在定义一个随机变量 Y，用来表示做了一次这样的随机试验后，得到了多少个正面朝上的硬币。即

$$Y : s \in \Omega \mapsto s \text{ 中 } H \text{ 的个数.}$$

我们来考虑 Y 的分布 μ_Y。显然 $\mathsf{Im}(Y) = \{0, 1, \dots, n\}$。由于 Y 也是离散随机变量，μ_Y 由 pmf p_Y 决定。容易计算得知

$$\forall k = 0, 1, \dots, n, \quad p_Y(k) = \mathbb{P}[Y = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

我们把这样一个分布称为参数为 n 和 p 的二项式分布（Binomial distribution），记作 $\mathsf{Bin}(n, p)$。

4.5 “分布”一词容易混淆的地方

我们前文出现“分布”这个词的时候，实际上指了两件事：

- 给定一个具体的概率空间，一个定义在这个概率空间上的随机变量 X，该随机变量诱导出的分布 μ_X；
- 一个具体的概率质量函数定义出来的分布，比如 $\mathsf{Ber}(p)$ 或者 $\mathsf{Bin}(n, p)$。

在教科书或者文献中，我们经常会看到形如“设 $X \sim \mathsf{Ber}(p)$”，或者等价的，“设 X 为满足参数为 p 的伯努利分布的随机变量”。这句话往往让人费解。按照定义 X 应该是一个从概率空间到实数的一个函数，如果轻易的就这么设出来了，那概率空间是啥？

对于这个问题，我是这么理解的。比如说，我们设 $Y \sim \mathsf{Bin}(n, p)$，实际上，我们理解成构造一个随机变量 Y，使得其诱导出来的分布 μ_Y 为 $\mathsf{Bin}(n, p)$。对于定义 Y 的方式，包括函数的形式以及对应的概率空间，它的选择并不是唯一的。比如说，我们可以选择上述引入二项式分布时候介绍的扔 n 个硬币所定义出来的概率空间和 Y，也可以选择下面这个看起来有点“平凡”的概率空间：

$$\Omega' = \{0, 1, \dots, n\}, \quad \mathscr{F}' = 2^{\Omega'},$$

$$\forall k \in \Omega', \quad \mathbb{P}'(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

然后，我们定义随机变量

$$\forall k \in \Omega', \quad Y' : k \mapsto k.$$

显然，Y' 的分布也是 $\mathsf{Bin}(n, p)$。

那么问题来了，我们究竟用哪个？这取决于应用。在很多应用中，我们可能只关心 pmf 的性质，那选择怎样定义的 Y 其实无所谓。并且，容易想到，我们定义 Y' 的方式，可以推广到任何分布上，但这个定义丧失了分布本身的“结构”，或者说“组合含义”。在有一些应用中，比如我们今天最后会讲到的使用期望的线性性来计算二项式分布的期望的时候，选择 $\Omega = \{H, T\}^n$ 这样有着更加丰富组合结构的概率空间，会更加方便。

4.6 随机变量的期望

随机变量的一个重要的“数字特征”，便是它的期望。它可以想象成当我们做随机试验的时候，$X(\omega)$ 的平均值。它的定义，并不是特别平凡的。我们今天先从定义在离散概率空间上的随机变量开始。

假设 X 是定义在离散概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的一个随机变量，它的值域 $\mathsf{Im}(X) = \{x_1, \dots, x_n, \dots\}$。我们尝试把它的期望定义为 $\sum_{x \in \mathsf{Im}(X)} x \cdot \mathbb{P}[X = x]$。但这个求和可能是个无穷级数，因此，我们要对其行为进行控制。因此，要进行更加细致的讨论。

我们首先设 $\{\Lambda_i\}_{i \geq 1}$ 是对样本空间的一个划分，并且对于任何 $i \geq 1$，在 Λ_i 上 X 是常数，即 $\forall \omega, \omega' \in \Lambda_i, X(\omega) = X(\omega')$，$\forall \omega \in \Lambda_i$。这样的一个划分肯定是存在的，比如我们可以设 $\Lambda_i = X^{-1}(x_i)$。但是，我们这儿给出的划分定义更加一般，因为我们允许对于 $i \neq j, \omega \in \Lambda_i, \omega' \in \Lambda_j$，有 $X(\omega) = X(\omega')$。我们设对于 $\omega \in \Lambda_i$，$X(\omega) = z_i$（刚才说了，我们允许 $i \neq j, z_i = z_j$）。

我们称随机变量 X 是可积（integrable）的，当且仅当级数 $\sum_{i=1}^{\infty} z_i \cdot \mathbb{P}[\Lambda_i]$ 是绝对收敛（converge absolutely）的，或者等价地，

$$\sum_{i=1}^{\infty} |z_i| \cdot \mathbb{P}[\Lambda_i] < \infty.$$

如果一个随机变量 X 是可积的，我们就把它的期望 $\mathbb{E}[X]$ 定义为

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} z_i \cdot \mathbb{P}[\Lambda_i].$$

关于这个定义，小朋友一定有很多问号，包括但可能不限于：

1. 这个定义依赖于一个不唯一的分划，它是良定义的吗？
2. 为什么要求级数收敛？
3. 为什么要求级数绝对收敛？

我们先回答后两个问题。首先，根据我们的定义，一定有 $\mathbb{E}[X] < \infty$。实际上，这个要求是为了我们目前理论开展的方便。我们在不久的未来就会把期望拓展到无穷的情况。所以，我们暂时只允许期望取有限值。其次，为什么要绝对收敛。原因很简单，如果一个级数只是条件收敛，那么变换求和的顺序就可能得到不同的极限值，我们不希望一个随机变量的“平均值”会随着求和的顺序不同而不一样。

回到第一个问题，这个定义是良定义的吗。事实上，我们可以把所有的 $\{\Lambda_i\}_{i\geq 1}$ 进行分类，如果 $z_i = z_j$，我们就认为其为一类。我们在计算级数 $\sum_{i=1}^{\infty} z_i \cdot \mathbb{P}[\Lambda_i]$ 的时候，可以先按照 z_i 所有可能的取值进行求和，这对应于对 x_i 进行求和，再对于同一类里面的 Λ_j 进行求和，他们一起构成了 $X^{-1}(x_i)$：

$$\sum_{i=1}^{\infty} z_i \cdot \mathbb{P}[\Lambda_i] = \sum_{i=1}^{\infty} x_i \sum_{j\geq 1 : z_j = x_i} \mathbb{P}[\Lambda_j] = \sum_{i=1}^{\infty} x_i \cdot \mathbb{P}[X = x_i].$$

这就说明了，这个求和与我们选择的分划无关（只需要满足在每一个 Λ_i 内 X 是常数）。注意到，所有这些求和可以随意交换的性质，是该级数绝对收敛所保证的。

我们使用这个分划的语言，而不是直接用 $\sum_{i=1}^{\infty} x_i \cdot \mathbb{P}[X = x_i]$ 来定义期望，是为了某些证明的方便。

比如说，我们假设 $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$，我们可以令 $\Lambda_i = \{\omega_i\}$，则我们得到期望的另一个表达式

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\{\omega\}].$$

这个和 $\sum_{i=1}^{\infty} x_i \cdot \mathbb{P}[X = x_i]$ 相比，我们分别对函数 X 的“左边”和“右边”加权求和，并得到了一样的值。这种 double counting 的技巧，在处理某些问题的时候会非常有用。

我们来计算一下二项式分布 $Y \sim \mathsf{Bin}(n, p)$ 的期望。根据定义，我们有

$$\mathbb{E}[Y] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np.$$

第 5 章 离散期望的基本性质

我们今天来讨论离散随机变量的期望的一些基本性质，这些性质在解决具体问题中起着非常重要的作用。

5.1 LOTUS

下面这个结论，被称之为 LOTUS （law of the unconcious statistician），原因是它是如此显然，以至于在很多书上被直接当成期望的定义。实际上，它是我们刚才定义的期望的一个推论，是需要证明的。我们说一个函数 $f:\mathbb{R}\to\mathbb{R}$ 是可测的，当且仅当对于任何 $A\in\mathscr{B}$，$f^{-1}(A)\in\mathscr{B}$。我们未来会仔细讨论“可测性”的问题，现在大家不用太在意这个条件。对于可测的 f，$f(X):\omega\mapsto f(X(\omega))$ 也是一个随机变量。

定理 5.1 (Law of the Unconscious Statistician (LOTUS))

对于可测函数 f，如果满足 $\sum_{i=1}^{\infty}|f(x_i)|\mathbb{P}[X=x_i]<\infty$，则随机变量 $f(X)$ 是可积的，并且

$$\mathbf{E}[f(X)]=\sum_{i=1}^{\infty}f(x_i)\cdot\mathbb{P}[X=x_i].$$

♡

证明 我们考虑一个分划 $\{\Lambda_i\}_{i\geq1}$，$\Lambda_i=X^{-1}(x_i)$。那么 $f(X)$ 在每个 Λ_i 上均为常数 $f(x_i)$。条件保证了可积性，因此，

$$\mathbf{E}[f(X)]=\sum_{i\geq1}f(x_i)\cdot\mathbb{P}[\Lambda_i]=\sum_{i\geq1}f(x_i)\cdot\mathbb{P}[X=x_i].$$

5.2 期望的线性性

下面一个结论，被称为期望的线性性（Linearity of expectation），是非常有用的性质。我们在未来真的用概率论解决一些问题的时候，会发现其妙用无穷。我们现在，先证明它。

定理 5.2 (期望的线性性)

如果定义在同一个概率空间上的随机变量 X 和 Y 都是可积的，那么对于 $a,b\in\mathbb{R}$，$aX+bY$ 也是可积的，并且有

$$\mathbf{E}[aX+bY]=a\mathbf{E}[X]+b\mathbf{E}[Y].$$

♡

证明 我们先证明 $aX+bY$ 是可积的。我们可以给概率空间找到一个划分 $\{\Lambda_i\}_{i\geq 1}$ 满足对于 X 和 Y 在每个 Λ_i 上都是常数。对于每一个 $i\geq 1$，我们记这个常数为 x_i 和 y_i。于是，

$$\sum_{i\geq 1}|ax_i+by_i|\mathbb{P}(\Lambda_i)\leq\sum_{i\geq 1}(|a||x_i|+|b||y_i|)\mathbb{P}[\Lambda_i]=|a|\mathbf{E}[|X|]+|b|\mathbf{E}[|Y|]<\infty.$$

对于期望的表达式，我们可以把重复上述计算，把所有的绝对值去掉，并且把不等号换成等号即可。

我们可以使用数学归纳法，把上述结论推广到任意 n 个随机变量。即如果定义在同一个概率空间上的随机变量 $X_1,\dots,X_n$ 均是可积的，那么 $\sum_{i=1}^n a_i\cdot X_i$ 也是可积的，并且

$$\mathbf{E}\left[\sum_{i=1}^n a_iX_i\right]=\sum_{i=1}^n a_i\cdot\mathbf{E}[X_i].$$

我们前面已经通过期望的定义，计算了对于 $Y\sim\mathsf{Bin}(n,p)$，其期望 $\mathbf{E}[Y]=np$。前面的定义实际上只用到了此分布概率质量函数的性质。事实上，如果回到一开始引入二项式分布的随机试验，即统计“扔 n 个硬币，正面朝上的个数”，我们可以使用期望的线性性更加方便的计算 Y 的期望。回顾我们的样本空间是 $\Omega=\{H,T\}^n$，并且对于每一个 $s\in\Omega$，我们有

$$Y:s\in\Omega\mapsto s\text{ 中 }\mathrm{H}\text{ 的个数}.$$

我们现在定义 n 个随机变量 $Y_1,\dots,Y_n$，满足

$$\forall i\in[n],\quad Y_i:s\in\Omega\mapsto\mathbb{I}_{s\text{ 的第 }i\text{ 位是 H}}.$$

换句话说 $Y_i(s)=1$ 当且仅当 s 的第 i 位是正面。那么显然 $Y=\sum_{i=1}^n Y_i$（我再强调一下，我们写这个等式的意思是，对于任何 $s\in\Omega$，$Y(s)=\sum_{i=1}^n Y_i(s)$ 成立）。由于 Y_i 表示的就是第 i 个硬币的结果，满足 $Y_i\sim\mathsf{Ber}(p)$。容易计算 $\mathbf{E}[Y_i]=p$。

因此，由期望的线性性，

$$\mathbf{E}[Y]=\mathbf{E}\left[\sum_{i=1}^n Y_i\right]=\sum_{i=1}^n\mathbf{E}[Y_i]=np.$$

那么，期望的线性性不能推广到无穷多个随机变量，即 $\mathbf{E}[\sum_{i=1}^\infty X_i]=\sum_{i=1}^\infty\mathbf{E}[X_i]$ 不一定成立（你能想到反例吗?）。对于无穷多个随机变量，求和和期望什么时候能交换，是我们未来会重点研究的一个问题。

5.3 期望线性性的一些应用

期望的线性性可以方便很多计算，我们来看几个例子。

例题 5.1.

考虑一个箱子里有 100 个球，其中 10 个是红球，剩下的是白球。现在无放回的随机摸 20 个球出来，请问平均会有多少个红球。

对于这个概率实验，我们先用概率空间建模。这儿 $\Omega=\binom{[100]}{20}$ 表示集合 $\{1,2,\dots,100\}$ 的所有大小为 20 的子集的集合。$\mathscr{F}=2^\Omega$，$\mathbb{P}$ 为 Ω 上的均匀分布。我们用随机变量 X 来表示抽出来的 20 个球里红球的个数。对于 $i=1,2,\dots,20$，我们定义 $X_i=\mathbb{I}_{[\text{第}i\text{ 个球是红球}]}$ 为事件“第 i 个球是红球”的指示变量。那么显然有 $X=\sum_{i=1}^{20}X_i$。使用期望的线性性，我们有 $\mathbf{E}[X]=\mathbf{E}[\sum_{i=1}^{20}X_i]=\sum_{i=1}^{20}\mathbf{E}[X_i]$。接下来，我们有两个观察。

1. 对于 X_1，$\mathbf{E}[X_1]=\mathbb{P}[X_1\text{ 是红球}]=0.1$。
2. 由对称性，对于每一个 i，它的分布和 X_1 是一样的。因此 $\mathbf{E}[X_i]=\mathbf{E}[X_1]$。

所以，$\mathbf{E}[X]=20\cdot\mathbf{E}[X_1]=2$。

大家可以看到，上面例子里 X_i 之间看起来是有一些“联系”的，但是期望的线性性依旧无条件成立。下面是一个类似的例子。

例题 5.2.

考虑一个抽屉里有 10 双袜子（每双都是不同的款式），我们现在随机的摸五只袜子出来，请问这五只里平均会有配成几双。

在很多时候，我们不再严格的指出概率空间，而是直接定义随机变量。我们用 X 来表示这五只袜子里有能配成多少双。我们用 $X_i := \mathbb{I}[\text{第}i\text{ 只袜子被配对了}]$ 来表示我们抽出来的第 i 只袜子在这五只里被凑成一对的这个事件的指示变量。那么显然，$X = \frac{1}{2}\sum_{i=1}^{5} X_i$。这儿 $\frac{1}{2}$ 是由于我们问的是“双数”。那么，由期望的线性性，

$$\mathbf{E}[X] = \mathbf{E}\left[\frac{1}{2}\sum_{i=1}^{5} X_i\right] = \frac{1}{2}\sum_{i=1}^{5}\mathbf{E}[X_i].$$

同样，我们有两个观察。

1. 对于摸出来的第一只袜子，$\mathbf{E}[X_1] = \mathbb{P}[\text{第一只袜子被配对}]$。而第一只袜子被配对，当且仅当我们摸出来的第二到第五只袜子里，有一只正好是抽屉里第一只袜子的孪生兄弟。这个概率，我们用简单的组合计数就可以算出来，是 $\frac{1}{19}$。
2. 对于摸出来的其它袜子，由于对称性，它被配对的分布和第一只袜子是一样的。因此 $\mathbf{E}[X_i] = \mathbf{E}[X_1]$。

所以，$\mathbf{E}[X] = \frac{1}{2}\cdot 5 \cdot \frac{1}{19} = \frac{5}{38}$。

注意到在刚才的例子里，我们并没有严格的给出 X 和 X_i 所存在的概率空间，就开始进行计算了。大家需要仔细的想明白这背后的概率空间是什么，它的存在性是显然的。在今天后面我们会讲一些例子，所用随机变量的概率空间存在性并不那么显然，需要一些额外的知识才能严格的证明其存在，我们会在下周的课程中来证明。但我们所用的式子的正确性在直观上是显然的，我们今天暂且相信直观，把证明和计算进行下去。这就好比牛顿和莱布尼茨的微积分提出一百多年后严格性才被真正解决，但这之前大家已经用它计算出不少天体运行的规律了。严格性警察请暂时不要上班。

5.4 随机变量的独立性

我们接下来介绍随机变量的独立性。同样的，今天我们只考虑离散概率空间上的随机变量。对于一般的情况，我们在未来还会重新审视这个定义。给定离散概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$，我们之前已经定义了两个事件独立的概念，即对于 $A, B \in \mathscr{F}$，我们说 A 与 B 独立，当且仅当 $\mathbb{P}[A \cap B] = \mathbb{P}[A]\cdot\mathbb{P}[B]$。我们用这个定义来给出随机变量 X 和 Y 独立的定义。对于两个定义在同一**离散**样本空间上的随机变量 $X, Y : \Omega \to \mathbb{R}$，我们说 X 和 Y 独立，记作 $X \perp Y$，当且仅当对于任何 $a, b \in \mathbb{R}$，事件 $[X = a]$ 和事件 $[Y = b]$ 独立，或者等价的

$$\mathbb{P}[X = a \wedge Y = b] = \mathbb{P}[X = a]\cdot\mathbb{P}[Y = b].$$

Remark

在谈论随机变量的取值定义的事件的时候，我们有时候会把 $\cap, \cup$ 写成 $\wedge, \vee$，意思是一样的。

同样的，我们说定义在同一样本空间上的随机变量 $X_1, X_2, \ldots, X_n$ 相互独立，当且仅当对于任何指标集 $I \subseteq [n]$，任何实数 $(a_i : i \in I)$，有

$$\mathbb{P}\left[\bigwedge_{i\in I}(X_i = a_i)\right] = \prod_{i\in I}\mathbb{P}[X_i = a_i].$$

而我们说无穷多个随机变量相互独立，当且仅当它的任意有限子集相互独立。我们也可以类似的定义两两独立的随机变量。

我们今天，希望大家把独立性的定义更多的停留在**直观**上，即随机变量 X 和 Y 是相互没有影响的。比如我们扔两个骰子，X 表示第一个骰子的点数，Y 表示第二个骰子的点数，那么 X 和 Y 是独立的。我们在未来，当我们学会了足够多的黑话之后，会回过头来说明独立性和乘积概率空间的等价性。于是，在我们给定了概率空

间后，独立性便容易严格的验证了。

我们可以容易验证，对于两个集合 A, B，$X \perp Y$ 可以蕴含

$$\mathbb{P}[X \in A \wedge Y \in B] = \mathbb{P}[X \in A] \cdot \mathbb{P}[Y \in B].$$

我们接下来考察独立性的一个应用，即对于独立的 X 和 Y，乘积的期望等于期望的乘积。假设 X 和 Y 可积，则

$$X \perp Y \implies \mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

我们接下来的证明稍微修改一下便是 XY 的可积性的证明，因此我们略过。我们直接来验证 $\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y]$。假设 $\mathsf{Im}(X) = \{x_1, x_2, \dots\}$，$\mathsf{Im}(Y) = \{y_1, y_2, \dots\}$。考虑分划 (Λ_{ij}) 满足 $\Lambda_{ij} = X^{-1}(x_i) \cap Y^{-1}(y_j)$。那么

$$\mathbf{E}[XY] = \sum_{i,j} x_i y_j \mathbb{P}[\Lambda_{ij}] = \sum_{i,j} x_i y_j \mathbb{P}[X = x_i \wedge Y = y_j].$$

使用独立性的定义，我们有

$$\mathbf{E}[XY] = \sum_{i,j} x_i y_j \mathbb{P}[X = x_i] \cdot \mathbb{P}[Y = y_j] = \left(\sum_i x_i \mathbb{P}[X = x_i]\right)\left(\sum_j y_j \mathbb{P}[Y = y_j]\right) = \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

注意到，我们这儿第二个等号使用了 X 和 Y 的可积性。

5.5 Markov 不等式

我们这儿提一个关于期望的不等式，它想描述如下一件显然的事情：如果一个非负的随机变量期望一定，那么它的取值特别大的概率就不能很大（否则期望就炸了）。用数学的语言说就是

定理 5.3 (Markov 不等式)

对于非负随机变量 $X \geq 0$，任意 $a > 0$，

$$\mathbb{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

♡

证明 不等式的证明很简单，仅仅就是把我们觉得它对的原因严格的说一下。我这儿写一下，对于初学者来说，值得好好看一下我们这儿使用的所谓“截断”的技巧，它在未来很多证明里要用到。对于一个固定的 a，我们用事件 $[X \geq a]$ 来截断随机变量 X，得到

$$X = X \cdot \mathbb{I}[X < a] + X \cdot \mathbb{I}[X \geq a].$$

对于初学者，我认为值得好好读一下这个等式。首先，这个等式左右两边均是随机变量，也就是说它均是函数。因此，它的实际意思是，对于每一个样本空间的中的样本 $\omega \in \Omega$，函数左右两边在 ω 上的取值均是相同的。其次，在这个等式右边，分别乘上了指示变量，指示的是 $[X < a]$ 和 $[X \geq a]$ 这两个互补的事件，所以，其中正好一项非零，等式也因此成立。我们对左右两边取期望，并使用期望的线性性，就有

$$\mathbf{E}[X] = \mathbf{E}[X \cdot \mathbb{I}[X < a]] + \mathbf{E}[X \cdot \mathbb{I}[X \geq a]].$$

我们做一些放缩。首先 $X \cdot \mathbb{I}[X < a]$ 作为随机变量肯定是非负的，因此 $\mathbf{E}[X \cdot \mathbb{I}[X < a]] \geq 0$。其次，我们简单验算就能知道 $X \cdot \mathbb{I}[X \geq a] \geq a \cdot \mathbb{I}[X \geq a]$ 在每一个样本点上均成立。所以，

$$\mathbf{E}[X] \geq \mathbf{E}[a \cdot \mathbb{I}[X \geq a]] = a \cdot \mathbb{P}[X \geq a].$$

从上面的证明可以看出来，马尔可夫不等式是可以取到等号的，只要构造随机变量使得我们证明中用到的放缩都是紧的就行。这个留给大家做练习。

马尔可夫不等式在概率论和计算机科学中特别有用，因为它是一个尾不等式（*Tail inequality*），即描述某个随机变量特别大（或特别小）的概率不大的不等式。这个可以用来证明某个随机算法，它的输出，大概率在我们想要的结果附近。关于这类应用感兴趣的同学可以参见我另外一门课的 notes。

5.6 方差

期望是随机变量的“数字特征”之一，用来描述它的平均值。但有的时候在应用中，期望所反映出来的关于随机变量的信息是不够的。比如说，假设 X 是我们班上随机取样一位同学的身高。我们知道 $\mathbf{E}[X]$ 即平均身高是 170，这有可能是大家身高都在 170 附近，也有可能有一部分同学身高极高，有一部分极低，导致平均是 170。又或者，两位 NBA 球员在一场比赛中平均得到 40.5 分，可能是因为两个人得分能力都很强，也可能是因为其中一位是科比...

方差（Variance）便是一个用来描述和随机变量的偏离程度的数字特征，对于一个可积的随机变量 X，它定义为

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2].$$

从定义便可以看出，方差指的是 $(X - \mathbf{E}[X])^2$ 这个描述 X 和它的期望的偏差的随机变量的平均值。

我们可以把这个定义展开，并由期望的线性性，有

$$\mathbf{Var}[X] = \mathbf{E}[X^2 - 2X \cdot \mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2,$$

即方差等于“平方的期望减去期望的平方”。在有的地方，我们也把这个写成方差的定义。

这个时候，敏感的严格性警察可能要出警了：你在方差的定义里出现了 $\mathbf{E}[X^2]$，那为啥不要求 X^2 可积啊。这儿我们确实稍微扩展了一下期望的定义，由于 X^2 是非负的，如果它不可积，那也一定是发散到正无穷。这个时候，我们稍稍拓展一下期望的定义并称此时的期望是正无穷。我们在未来会严格的说这件事情。

注意到上面方差的式子里出现的（唯一新）量 $\mathbf{E}[X^2]$，我们把它称作 X 的二阶矩。同理，对于任意自然数 $k \geq 1$，我们把 $\mathbf{E}[X^k]$（如果存在）称之为 X 的 k-阶矩。我们在未来，会看到这些矩在实质上刻画了这个随机变量，并且有着丰富的应用。

我们来给出几个方差的基本性质，这些性质根据定义验证即可，大家可以当作练习。

1. 对于 $a, b \in \mathbb{R}$,$\mathbf{Var}(aX + b) = a^2 \cdot \mathbf{Var}(X)$
2. 对于 $X \perp Y$，$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$。
3. 如果 $X_1, \ldots, X_n$ 是两两独立的，则 $\mathbf{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbf{Var}[X_i]$。
   这便是对于独立随机变量的所谓的方差的线性性。和期望的线性性相比，它额外需要这些随机变量是独立的，这个要求的原因来自于我们需要独立性才能够使得 $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$ 成立。

5.7 切比雪夫不等式（Chebyshev's inequality）

切比雪夫不等式定量上解释并描述了我们引入方差的动机。我们关心一个可积随机变量与它的期望的偏差程度，这件事情通常被使用形如

$$\mathbb{P}[|X - \mathbf{E}[X]| \geq a] \leq b$$

这样的不等式所描述。这种不等式被称为集中不等式（Concentration Inequality），其重要性，在概率论和计算机科学中是难以衡量的。

切比雪夫不等式指的是：

定理 5.4 (切比雪夫不等式)

$$\forall a > 0, \quad \mathbb{P}[|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{Var}[X]}{a^2}.$$

♡

证明 其证明，是对 Markov 不等式的直接使用。

$$\mathbb{P}[|X - \mathbf{E}[X]| \geq a] = \mathbb{P}[(X - \mathbf{E}[X])^2 \geq a^2] \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{a^2}.$$

第 6 章 离散期望的一些应用

我们上节课介绍了期望和方差的定义。我们今天使用它们来解决一些实际的问题。使用离散概率工具解决算法、组合数学问题是一个很深刻的主题，感兴趣的同学可以参考 Mitzenmacher and Upfal, Probability and Computing 以及 Alon and Spencer, The Probabilistic Method 两本名著。

6.1 几何分布（Geometric Distribution）

我们今天要介绍一类重要的分布，叫做几何分布。它有一个参数 $p \in [0,1]$，被记作 $\mathsf{Geom}(p)$。我们上一节课说过，描述一个（离散）分布，给出它的概率质量函数 $p_k = \mathbb{P}[X = k]$ 就好了。当然，更加“合适”的做法是描述出这个满足分布的随机变量背后的随机试验。几何分布对应了如下随机试验：

考虑不停地扔一枚 p-偏差的硬币（即每次 p 的概率出正面，$1-p$ 的概率出反面），X 表示第一次出现正面的时候扔的次数。根据这个定义，我们“显然”有

$$p_k = \mathbb{P}[X = k] = (1-p)^{k-1} \cdot p.$$

这便定义了几何分布 $\mathsf{Geom}(p)$。

我们这儿要十分小心。我们现在通过直接给出概率质量函数 $\forall k \geq 1,\ p_k = (1-p)^{k-1}p$ 的形式“定义”了几何分布 $\mathsf{Geom}(p)$。我们上节课也说过，一旦给出基于概率质量函数的定义，我们可以构造一个平凡的概率空间与随机变量，使得它的分布是这个给定分布。在我们这个例子里，我们让 $\Omega' = \mathbb{Z}_{\geq 1}$ 为所有正整数，$\mathscr{F}' = 2^{\Omega'}$，$\mathbb{P}'(\{k\}) = (1-p)^{k-1}p$，$X'(k) = k$。那么 $X' \sim \mathsf{Geom}(p)$。但这个概率空间并不是我们直观上引入几何分布的那个概率空间。

我们现在来看看直观上引入概率几何分布的那个随机试验，所对应的概率空间是什么。一番思索之后，大家应该能发现样本空间是 $\Omega = \{H,T\}^{\mathbb{N}}$，即所有“无限长”的 HT 串的集合（注意，这不是 $\{H,T\}^*$，$\{H,T\}^*$ 是所有“有限长”HT 串的集合）。因此 Ω 并不是一个可数集，我们目前还没有办法在上面定义概率测度。事实上，Ω 中的元素可以看成无限（二进制）小数，因此可以把 Ω 和 $[0,1]$ 对应起来。我们在未来会让大家在作业里证明，这儿的 σ-代数可以取 $[0,1]$ 上的所有 Borel 集，对于 $p = \frac{1}{2}$ 而言，我们可以使用在 $[0,1]$ 上定义均匀分布同样的方法定义 Ω 上的均匀分布。

但是无论如何，我们有了概率质量函数我们可以做计算了。使用高中曾经擅长的数列求和技巧，对于 $X \sim \mathsf{Geom}(p)$，我们有

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} \cdot p = \frac{1}{p}.$$

我们同样可以计算出

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{1-p}{p^2}.$$

因此我们可以说，扔一枚均匀硬币，平均两次，会出现一次正面。虽然我们目前还没有完全严格的把这个分布和扔硬币的随机试验对应起来。

6.2 冒泡排序交换次数的期望与方差

冒泡排序是一种直观的排序算法，其核心操作是不断交换相邻的逆序对，直到整个序列有序。一个关键的观察是：冒泡排序的总交换次数等于输入序列中逆序对的总数。

假设我们从 $n!$ 个不同的排列中均匀随机地选取一个作为输入。令随机变量 X 表示该排列的逆序对数（即冒泡排序的交换次数）。下面我们用两种不同的方法来计算 $\mathbb{E}[X]$ 和 $\mathrm{Var}(X)$。

对于任意一对索引 (i,j)，其中 $1 \le i < j \le n$，我们定义一个指示随机变量：

$$X_{i,j} = \begin{cases} 1, & \text{如果} a_i > a_j \text{ (即} (i,j) \text{ 是一个逆序对)}, \\ 0, & \text{否则}. \end{cases}$$

显然，总的逆序对数为：

$$X = \sum_{1 \le i < j \le n} X_{i,j}.$$

由于输入是均匀随机排列，对于任意一对 (i,j)，a_i 和 a_j 的大小关系是等可能的，即 $\mathbb{P}(a_i > a_j) = \frac{1}{2}$。因此，

$$\mathbb{E}[X_{i,j}] = \mathbb{P}(a_i > a_j) = \frac{1}{2}.$$

根据期望的线性性质，我们有：

$$\mathbb{E}[X] = \sum_{1 \le i < j \le n} \mathbb{E}[X_{i,j}] = \binom{n}{2} \cdot \frac{1}{2} = \frac{n(n-1)}{4}.$$

接下来，我们计算方差 $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$。为此，我们需要计算 $\mathbb{E}[X^2]$。

$$\mathbb{E}[X^2] = \mathbb{E}\left[\left(\sum_{1 \le i < j \le n} X_{i,j}\right)^2\right] = \mathbb{E}\left[\sum_{1 \le i_1 < j_1 \le n} \sum_{1 \le i_2 < j_2 \le n} X_{i_1,j_1} X_{i_2,j_2}\right].$$

这个双重求和可以按索引对 (i_1,j_1) 和 (i_2,j_2) 的重叠情况进行分类：

1. **情况 1:** $(i_1,j_1) = (i_2,j_2)$。此时 $X_{i_1,j_1} X_{i_2,j_2} = X_{i_1,j_1}^2 = X_{i_1,j_1}$，因为 $X_{i,j}$ 是 0-1 变量。共有 $\binom{n}{2}$ 项，每项期望为 $\frac{1}{2}$。
2. **情况 2:** (i_1,j_1) 和 (i_2,j_2) 共享一个索引（例如 $i_1 = i_2$ 但 $j_1 \ne j_2$）。此时，三个元素 $a_{i_1}, a_{j_1}, a_{j_2}$ 的相对顺序是随机的，$X_{i_1,j_1} X_{i_2,j_2} = 1$ 当且仅当这三个元素是严格递减的，概率为 $\frac{1}{3!} \times 2 = \frac{1}{3}$。此类项共有 $2\binom{n}{2}(n-2)$ 项（选择两个共享索引的对，再选第三个不同的索引）。
3. **情况 3:** (i_1,j_1) 和 (i_2,j_2) 共享两个索引（即 $j_1 = i_2$ 或 $j_2 = i_1$）。这相当于考察一个三元组 (i,k,j)，其中 $i < k < j$，要求 $a_i > a_k$ 且 $a_k > a_j$。三个元素的六种排列中只有一种满足，故概率为 $\frac{1}{6}$。此类项共有 $\binom{n}{2}(n-2)$ 项。
4. **情况 4:** (i_1,j_1) 和 (i_2,j_2) 完全不相交。四个元素的相对顺序是随机的，$X_{i_1,j_1} X_{i_2,j_2} = 1$ 当且仅当两对都是逆序，概率为 $\frac{1}{4}$。此类项共有 $\binom{n}{2}\binom{n-2}{2}$ 项。

将以上各项代入并化简，可得：

$$\mathbb{E}[X^2] = \frac{n(n-1)(9n^2 - 5n + 10)}{144}.$$

因此，方差为：

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{n(n-1)(2n+5)}{72}.$$

当然我们也可以通过构建递推关系，从规模为 $n-1$ 的问题推导出规模为 n 的问题。

考虑一个随机排列 π_n。我们可以将其视为：先生成一个 $[n-1]$ 的随机排列 π_{n-1}，然后将元素 n（最大值）以等概率 $\frac{1}{n}$ 插入到 π_{n-1} 的 n 个位置之一。

设 X_n 为 π_n 的逆序对数，X_{n-1} 为 π_{n-1} 的逆序对数。令 Y_n 为因插入元素 n 而新产生的逆序对数。显然，Y_n 是一个离散均匀随机变量，取值于 $\{0,1,\dots,n-1\}$，且 $X_{n-1} \perp\!\!\!\perp Y_n$。

于是，我们得到递推关系：

$$X_n = X_{n-1} + Y_n.$$

期望：

$$\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}] + \mathbb{E}[Y_n] = \mathbb{E}[X_{n-1}] + \frac{n-1}{2}.$$

由初始条件 $\mathbb{E}[X_1] = 0$，解得：

$$\mathbb{E}[X_n] = \sum_{k=2}^{n} \frac{k-1}{2} = \frac{n(n-1)}{4}.$$

方差：

$$\mathrm{Var}(X_n) = \mathrm{Var}(X_{n-1}) + \mathrm{Var}(Y_n).$$

其中，$\mathrm{Var}(Y_n) = \mathbb{E}[Y_n^2] - (\mathbb{E}[Y_n])^2 = \frac{(n-1)(2n-1)}{6} - \left(\frac{n-1}{2}\right)^2 = \frac{n^2-1}{12}$。

由初始条件 $\mathrm{Var}(X_1) = 0$，解得：

$$\mathrm{Var}(X_n) = \sum_{k=2}^{n} \frac{k^2-1}{12} = \frac{n(n-1)(2n+5)}{72}.$$

两种方法殊途同归，均得到了相同的结果：

$$\mathbb{E}[X] = \frac{n(n-1)}{4}, \quad \mathrm{Var}(X) = \frac{n(n-1)(2n+5)}{72}.$$

6.3 奖券收集问题（Coupon Collector's Problem）

奖券收集问题是概率分析里面的一个重要的概率模型。我们使用今天介绍的技巧和不等式来研究这个模型。

考虑玩一个抽卡手游。现在总共有 n 种不同类型的卡，每一抽可以均匀的得到其中一种。现在想问平均要抽多少次，可以集齐一套，即 n 种卡每种至少一张。

我们考虑这个问题的概率空间建模。这儿的一个方便的样本空间和几何分布很类似，是 $\Omega = [n]^{\mathbb{N}}$，因此 Ω 是不可数的。我们不妨假设上面可以合理的定义均匀测度 $(\Omega, \mathscr{F}, \mathbb{P})$。定义随机变量 X 为第一次集齐一套的抽卡次数。我们关心的是 $\mathbf{E}[X]$ 即 X 的期望。由于 X 的取值还是离散的，我们可以把之前对于期望的定义稍微扩展一下，即定义 $\mathbf{E}[X] = \sum_{k=1}^{\infty} k \cdot \mathbb{P}[X=k]$。

> **再次警告**
>
> 我们接下来的计算是基于随机变量、随机试验、期望、独立性的直观进行的，它们是对的，但正确性的严格证明需要我们未来学习了更多的语言之后才能进行。现在我们暂时贷款一下。

直接通过 $\mathbf{E}[X]$ 的定义进行计算显然是困难的。我们又再次使用期望的线性性技巧。下面这种构造，第一次见是非常巧妙的，但它也是非常常用的构造，请大家务必理解。对于 $i = 1, 2, \dots, n$，我们定义随机变量 X_i 表示，"在当且已经有了 $i-1$ 种不同类型的卡之后，要获得另外一种新的类型的卡，还要抽几次" 这个随机变量。那么，我们有如下几个不言自明的观察：

1. $X = \sum_{i=1}^{n} X_i$；
2. $X_1, X_2, \dots, X_n$ 是相互独立的；
3. $X_i \sim \mathsf{Geom}\left(\frac{n-i+1}{n}\right)$。

于是，我们便可以使用期望的线性性和几何分布的性质得到

$$\mathbf{E}[X]=\sum_{i=1}^{n}\mathbf{E}[X_i]=\sum_{i=1}^{n}\frac{n}{n-i+1}=\sum_{i=1}^{n}\frac{n}{i}=nH_n,$$

其中 $H_n=\sum_{i=1}^{n}\frac{1}{i}$ 是调和级数。我们知道 $\lim_{n\to\infty}H_n=\log n+\gamma$，其中 $\gamma\approx 0.5772$ 是欧拉常数。

以上的计算告诉我们，如果 $n=1000$，那么要收集一套卡，平均需要 $nH_n=7485.47$ 次。但实际上，因为存在欧皇和非酋的缘故，我们往往并不关心平均数，我们关心，抽多少卡，可以“保证”凑齐一套。当然了，由于是随机问题，不可能 100% 保证，因此，我们想计算，假设希望以 99% 的概率收集齐一套，至少需要抽多少次？

我们用三种方法来估计这个上界。

1. 首先尝试用 Markov 不等式来估算。回忆 Markov 不等式它表达的是一个随机变量不太可能以特别大的概率特别大。这正好满足我们的要求。使用 Markov 不等式，我们有

   $$\mathbb{P}[X\geq a]\leq\frac{\mathbf{E}[X]}{a}=\frac{nH_n}{a}.$$

   如果我们想让上面的概率不超过 1%，我们需要取 $a=100nH_n$。即在我们上面的例子里，Markov 不等式告诉我们说，如果抽了 $100nH_n=748547$ 次卡，那么以 99% 的概率，能凑齐一套。
2. 当然游戏公司没有这么黑。上面估算的数值看起来很坏的原因在于，Markov 不等式在我们的例子上太松了，它并没有用到随机变量足够多的信息。我们这儿用上方差的信息试一试。由 Chebyshev's 不等式，

   $$\mathbb{P}[|X-\mathbf{E}[X]|\geq a]\leq\frac{\mathbf{Var}[X]}{a^2}.$$

   如果让上式不超过 1%，我们需要取 $a=10\sqrt{\mathbf{Var}[X]}$。因此，我们来计算一下 $\mathbf{Var}[X]$。我们注意到 $X=\sum_{i=1}^{n}X_i$，并且这些 X_i 是相互独立的。因此，我们可以用方差的线性性，得到

   $$\mathbf{Var}[X]=\sum_{i=1}^{n}\mathbf{Var}[X_i]=\sum_{i=1}^{n}\frac{1-\frac{n-i+1}{n}}{\left(\frac{n-i+1}{n}\right)^2}=\sum_{i=1}^{n}\frac{(i-1)n}{(n-i+1)^2}\leq\sum_{i=1}^{n}\frac{n^2}{(n-i+1)^2}.$$

   我们注意到

   $$\sum_{i=1}^{n}\frac{1}{(n-i+1)^2}=\sum_{i=1}^{n}\frac{1}{i^2}\leq\int_1^{\infty}\frac{dx}{x^2}=1.$$

   因此 $\mathbf{Var}[X]\leq n^2$。我们取 $a=10\sqrt{2}n$，即如果我们抽了 $H_n+10\sqrt{2}n\approx 21628$ 次卡，便可以以 99% 的概率收集一套了。
3. 实际上，我们可以直接计算抽了 m 张卡后还没有收集一套的概率。我们有

   $$\mathbb{P}[\text{抽了}m\text{ 张卡还没集齐}]=\mathbb{P}[\exists i\in[n]\text{抽了}m\text{ 次之后都没有抽到它}]\leq\sum_{i=1}^{n}\mathbb{P}[\text{抽了}m\text{ 次都没有抽到}i].$$

   上面这个小于等于号使用的是 union-bound。对于固定的卡 i，抽了 m 轮之后没有抽到它的概率是

   $$\left(1-\frac{1}{n}\right)^m\leq e^{-m/n}.$$

   因此，我们令 $ne^{-m/n}\leq 1\%$，可以得到 $m\geq n\log(100n)\approx 11513$。也就是说，抽了 11513 次之后，有超过 99% 的概率已经收集全一套了，这比之前计算的，又要好了一些。

上面几个分析可以看出，如果我们对于随机变量有更多的信息，可以让我们的估算更加准确。

6.4 随机图上的相变

考虑 **Erdős–Rényi** 随机图 $G(n,p(n))$，其中 $p(n):\mathbb{N}\to[0,1]$ 是一个关于顶点个数的函数。我们称一个图性质 $\mathcal{P}$ 具有相变性，如果 $\exists r:\mathbb{N}\to[0,1]$ 使得

1. 如果 $p(n)\ll r(n)$，$\lim_{n\to\infty}\mathbb{P}_{G\sim G(n,p(n))}[G\text{ satisfies }\mathcal{P}]=0$；
2. 如果 $p(n)\gg r(n)$，$\lim_{n\to\infty}\mathbb{P}_{G\sim G(n,p(n))}[G\text{ satisfies }\mathcal{P}]=1$。

Remark

图的性质 $\mathcal{P}: G \mapsto 0$ or 1 指的是从图到 0 或者 1 的一个映射.

这儿 r 被称之为 $\mathcal{P}$ 的阈值函数。下面我们将用二阶矩方法证明性质“一个图包含一个 K_4”具有相变性，并且其阈值函数是 $n^{-2/3}$。

定理 6.1

图性质“包含一个 K_4”具有阈值函数 $n^{-2/3}$。

令随机变量 X 表示 G 中 K_4 的数量。当 $p(n) \ll n^{-2/3}$ 时，根据马尔可夫不等式，我们有

$$\mathbb{P}_{G\sim G(n,p(n))}[G \text{ contains a } K_4] = \mathbb{P}[X \geq 1] \leq \mathbf{E}[X].$$

对于任意的图中的 4 个顶点构成的集合 $S \in \binom{[n]}{4}$，令 $X_S = \mathbf{1}[G[S] \text{ is a clique}]$。那么

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{S\in\binom{[n]}{4}} X_S\right] = \binom{n}{4} \cdot p^6 \leq n^4 p^6 = o(1).$$

另一方面，当 $p(n) \gg n^{-2/3}$，使用切比雪夫不等式

$$\mathbb{P}[X = 0] \leq \mathbb{P}[|X - \mathbf{E}[X]| \geq \mathbf{E}[X]] \leq \frac{\mathbf{Var}[X]}{(\mathbf{E}[X])^2} = \frac{\mathbf{E}[X^2] - (\mathbf{E}[X])^2}{(\mathbf{E}[X])^2}.$$

注意到

$$\begin{aligned}
&\mathbf{E}\left[X^2\right] - (\mathbf{E}\left[X\right])^2 \\
&= \mathbf{E}\left[\left(\sum_{S\in\binom{[n]}{4}} X_S\right)^2\right] - \left(\mathbf{E}\left[\sum_{S\in\binom{[n]}{4}} X_S\right]\right)^2 \\
&= 2\sum_{S\neq T} \mathbf{E}\left[X_S X_T\right] + \sum_S \mathbf{E}\left[X_S^2\right] - 2\sum_{S\neq T} \mathbf{E}\left[X_S\right]\mathbf{E}\left[X_T\right] - \sum_S (\mathbf{E}\left[X_S\right])^2 \\
&= 2\sum_{|S\cap T|=2} (\mathbf{E}\left[X_S X_T\right] - \mathbf{E}\left[X_S\right]\mathbf{E}\left[X_T\right]) + 2\sum_{S\cap T=3} (\mathbf{E}\left[X_S X_T\right] - \mathbf{E}\left[X_S\right]\mathbf{E}\left[X_T\right]) \\
&\quad + \sum_S \left(\mathbf{E}\left[X_S^2\right] - (\mathbf{E}\left[X_S\right])^2\right) \\
&\leq 2\sum_{|S\cap T|=2} \mathbf{E}\left[X_S X_T\right] + 2\sum_{S\cap T=3} \mathbf{E}\left[X_S X_T\right] + \sum_S \mathbf{E}\left[X_S^2\right].
\end{aligned}$$

正如 Figure 1 所示，当 $|S \cap T| = 2$ 时，$X_S = X_T = 1$ 当且仅当这 11 条边都被包含。因此

$$\mathbf{E}[X_S X_T] = \mathbb{P}[X_S = 1 \wedge X_T = 1] = p^{11}.$$

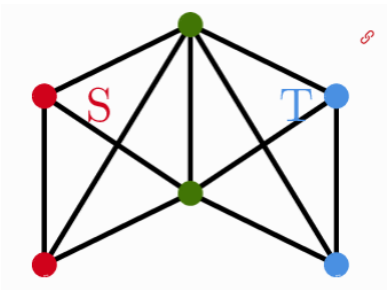


图 6.1: Figure1

类似地，当 $|S \cap T| = 3$（如 Figure 2 所示），

$$\mathbf{E}[X_S X_T] = \mathbb{P}[X_S = 1 \wedge X_T = 1] = p^9.$$

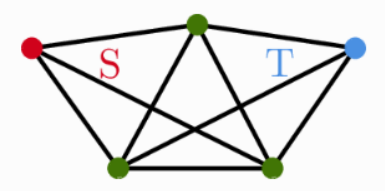


图 6.2: Figure2

因此，

$$\begin{aligned}\mathbf{E}\left[X^2\right]-(\mathbf{E}\left[X\right])^2 &\le 2\sum_{|S\cap T|=2}\mathbf{E}\left[X_SX_T\right]+2\sum_{S\cap T=3}\mathbf{E}\left[X_SX_T\right]+\sum_{S}\mathbf{E}\left[X_S^2\right]\\ &=2\binom{n}{2}\binom{n-2}{2}\binom{n-4}{2}p^{11}+2\binom{n}{3}\binom{n-3}{1}\binom{n-4}{1}p^9+\binom{n}{4}p^6\\ &\le n^6p^{11}+n^5p^9+n^4p^6=o\big((\mathbf{E}\left[X\right])^2\big).\end{aligned}$$

这说明了当 $p(n)\gg n^{-2/3}$ 时，$\mathbb{P}[G \text{ contains a } K_4]\to 1$。

事实上，我们可以更进一步拓展这个问题。

定理 6.2

对于任意一个固定的、边数为正的图 H，其“包含 H 作为子图”这一性质在 Erdős–Rényi 随机图 $G(n,p(n))$ 中的阈值函数为 $n^{-1/r(H)}$，其中 $r(H)=\max_{H_0\subseteq H}\frac{|E(H_0)|}{|V(H_0)|}$。 ♡

证明

令 $S_1,S_2,\dots,S_{m_n}$ 为在 n 个顶点的完全图中所有与 H 同构的子图。注意，m_n 是这些子图的数量，它是一个关于 n 的函数。对于每个 $i\in[m_n]$，我们定义一个指示随机变量：

$$X_i:=\mathbf{1}\{G \text{ 包含} S_i\}.$$

那么，随机变量 $X:=\sum_{i=1}^{m_n}X_i$ 就表示图 G 中包含的与 H 同构的子图的总数。显然，事件“G 包含 H 作为子图”等价于事件“$X\ge 1$”。我们的目标是分析 $\mathbb{P}(X\ge 1)$ 的渐进行为。为此，我们首先计算 X 的期望和方差。

根据期望的线性性质，我们有：

$$\mathbb{E}[X]=\sum_{i=1}^{m_n}\mathbb{E}[X_i].$$

由于 $G\sim G(n,p(n))$，一个特定的子图 S_i 出现在 G 中的概率等于其所有边都存在的概率，即 $p(n)^{|E(H)|}$。因此，

$$\mathbb{E}[X_i]=p(n)^{|E(H)|},\quad \forall i.$$

于是，

$$\mathbb{E}[X]=m_n\cdot p(n)^{|E(H)|}.$$

由于 $m_n=\Theta(n^{|V(H)|})$（因为我们需要从 n 个顶点中选择 $|V(H)|$ 个，并考虑它们之间的同构映射），所以

$$\mathbb{E}[X]=\Theta(n^{|V(H)|}p(n)^{|E(H)|}).$$

方差的计算更为复杂，因为它涉及不同子图之间的相关性。我们有：

$$\mathrm{Var}(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2=\sum_{i,j=1}^{m_n}\left(\mathbb{E}[X_iX_j]-\mathbb{E}[X_i]\mathbb{E}[X_j]\right).$$

关键观察是：当两个子图 S_i 和 S_j 没有公共边时，X_i 和 X_j 是独立的，此时协方差项为零。只有当它们共

享至少一条边时，协方差才可能非零。因此，我们可以将方差的上界写为：

$$\mathrm{Var}(X) \le \sum_{\substack{i,j\in[m_n]\\ E(S_i\cap S_j)\neq\varnothing}} \mathbb{P}(G\text{ 包含}S_i\cup S_j).$$

现在，我们来正式推导“包含 H 作为子图”这一性质的阈值函数。我们定义 $r(H)$ 为：

$$r(H) := \max_{H_0\subseteq H}\frac{|E(H_0)|}{|V(H_0)|},$$

其中 H_0 取遍 H 的所有子图。假设 H^* 是达到上述最大值的子图，即 $r(H)=\frac{|E(H^*)|}{|V(H^*)|}$。我们考虑性质 $\mathcal{P}'$：“包含 H^* 作为子图”。显然，如果 G 包含 H，则它必然包含 H^*。因此，

$$\mathbb{P}(G\text{ 满足}\mathcal{P}) \le \mathbb{P}(G\text{ 满足}\mathcal{P}').$$

令 X' 为 G 中包含 H^* 的数量。与之前类似，我们有：

$$\mathbb{E}[X'] = \Theta(n^{|V(H^*)|}p(n)^{|E(H^*)|}) = \Theta((n\cdot p(n)^{r(H)})^{|V(H^*)|}).$$

如果 $p(n)\ll n^{-1/r(H)}$，那么 $n\cdot p(n)^{r(H)}\to 0$，从而 $\mathbb{E}[X']\to 0$。根据马尔可夫不等式，$\mathbb{P}(X'\ge 1)\le\mathbb{E}[X']\to 0$。因此，$\mathbb{P}(G\text{ 满足}\mathcal{P})\to 0$。

对于另一边，我们使用切比雪夫不等式：

$$\mathbb{P}(X=0)\le\mathbb{P}(|X-\mathbb{E}[X]|\ge\mathbb{E}[X])\le\frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2}.$$

因此，要证明 $\mathbb{P}(X\ge 1)\to 1$，我们只需证明 $\frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2}\to 0$。

根据前面的结论，我们有：

$$\mathrm{Var}(X) \le \sum_{\substack{i,j\in[m_n]\\ E(S_i\cap S_j)\neq\varnothing}} \mathbb{P}(G\text{ 包含}S_i\cup S_j).$$

对于任意一对 (i,j)，设 $S_i\cap S_j$ 恰好包含 v 个顶点和 e 条边。那么 $S_i\cup S_j$ 包含 $2|V(H)|-v$ 个顶点和 $2|E(H)|-e$ 条边。因此，

$$\mathbb{P}(G\text{ 包含}S_i\cup S_j) = p(n)^{2|E(H)|-e}.$$

满足条件的 (i,j) 对的数量可以通过组合计数得到，其阶为 $O(n^{2|V(H)|-v})$。

于是，整个求和可以按 v 和 e 分组：

$$\mathrm{Var}(X) = O\left(\sum_{v,e} n^{2|V(H)|-v}p(n)^{2|E(H)|-e}\right).$$

另一方面，$(\mathbb{E}[X])^2=\Theta(n^{2|V(H)|}p(n)^{2|E(H)|})$。

因此，比值为：

$$\frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2} = O\left(\sum_{v,e} n^{-v}p(n)^{-e}\right).$$

由于 $S_i\cap S_j$ 是 H 的一个子图，我们有 $\frac{e}{v}\le r(H)$，即 $e\le r(H)\cdot v$。因此，$n^{-v}p(n)^{-e}\le(n\cdot p(n)^{r(H)})^{-v}$。

如果 $p(n)\gg n^{-1/r(H)}$，那么 $n\cdot p(n)^{r(H)}\to\infty$，从而 $(n\cdot p(n)^{r(H)})^{-v}\to 0$ 对于所有 $v\ge 1$ 成立。因此，整个求和趋于 0。

所以定理得证。

6.5 Weierstrass 近似定理

我们在数学分析中曾经学过，在一个闭区间上的任意一个连续的函数都可以被一个多项式函数任意地近似。我们现在使用二阶矩方法来证明这个定理。

定理 6.3 (Weierstrass 近似定理)

给定一个连续函数 $f:[0,1]\to[-1,1]$。对于任意的 $\varepsilon>0$ 都存在一个多项式 p 满足 $\forall x\in[0,1],\ |p(x)-f(x)|\le\varepsilon$。

我们可以用以下的观点来看待这个问题：

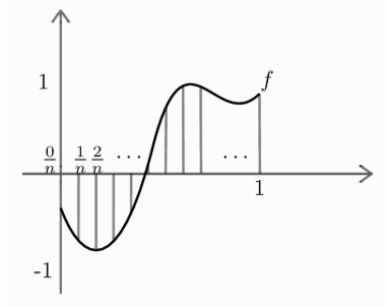


图 6.3: Figure3

证明

如 Figure 3 所示，我们从函数中选定 $n+1$ 个点，第 i 个点的取值为 $f(\frac{i}{n})$。然后我们定义

$$\mathbb{P}_n(x)=\sum_{i=0}^{n}E_i(x)\cdot f\left(\frac{i}{n}\right).$$

我们希望 $E_i(x)$ 满足下列条件：

1. 对于任意的 x，$\sum_i E_i(x)=1$。也就是说，系数是一个依赖于 x 的概率分布；
2. 对于任意的 i,x，$E_i(x)$ 是一个多项式；
3. 对于任意的 x，系数比较集中在离 x 最近的点 $\frac{i^*}{n}$ 上 $(i^*=\arg\min_i|x-i/n|)$。

我们可以定义出某些随机变量来表达 $E_i(x)$，也即令 $E_i(x)=\mathbb{P}[Y=i]$。注意到我们希望当 x 接近于 $\frac{i^*}{n}$ 的时候 $E_i(x)$ 很大，并且对于其他离 x 比较远的点我们希望他们的系数之和比较小。考虑随机变量 $Y\sim\mathsf{Bin}(n,x)$。我们有 $\mathbf{E}[Y]=nx$ 并且 $\mathbf{Var}[Y]=x(1-x)n\le n/4$。根据切比雪夫不等式，

$$\mathbb{P}\left[\left|\frac{Y}{n}-x\right|\ge n^{-1/3}\right]=\mathbb{P}[|Y-nx|\ge n^{2/3}]\le\frac{n/4}{n^{4/3}}=\frac{1}{4n^{1/3}}.$$

令 $E_i(x)=\mathbb{P}[Y=i]=\binom{n}{i}x^i(1-x)^{n-i}$。对于任意 $x\in[0,1]$，

$$\begin{aligned}|\mathbb{P}_n(x)-f(x)|&\le\sum_{i=0}^{n}E_i(x)\left|f\left(\frac{i}{n}\right)-f(x)\right|\\&=\sum_{i:|i-nx|\le n^{2/3}}E_i(x)\left|f\left(\frac{i}{n}\right)-f(x)\right|+\sum_{i:|i-nx|>n^{2/3}}E_i(x)\left|f\left(\frac{i}{n}\right)-f(x)\right|.\end{aligned}$$

由于函数 f 是连续的，那么存在 δ 使得 $\forall|x-y|<\delta,\ |f(x)-f(y)|<\varepsilon/2$。当 $n^{-1/3}<\delta$，我们有第一项 $\le\varepsilon/2$。同时，当 $n^{-1/3}<\epsilon$ 时，第二项 $\le 2\sum_{i:\|i-nx\|>n^{\frac{2}{3}}}E_i(x)\le\frac{n^{-\frac{1}{3}}}{2}\le\frac{\epsilon}{2}$。因此，选定 $n\ge\max\{\varepsilon^{-3},\delta^{-3}\}$，对于所有 $x\in[0,1]$，我们有 $|\mathbb{P}_n(x)-f(x)|\le\varepsilon$。

第 7 章 测度与单调类定理

今天我们正式开始关于概率论严格理论的学习。我们会用到不少分析和测度论的结论。有一些结论非常深刻，无法在课堂上证明，我会给出相应的参考。

7.1 $(0,1]$ 上的均匀概率测度

我们首先还是从如何定义 $(0,1]$ 上的均匀分布说起。我们之前提到，我们需要对所有 Borel 集 $\mathscr{B}$，也就是那些包含了开区间的最小的 σ-代数里的集合定义概率。这件事情是比较困难的，因为 $\mathscr{B}$ 里集合的形态比较复杂。于是，我们从更简单的集合，也就是那些区间开始，一步一步的构造概率。我们将要构造的这个概率测度用 λ 表示。

7.1.1 区间与有限区间的并

直观上来说，我们就是要给 $(0,1]$ 的子集定义出它的大小/长度。最简单的自然是区间。对于 $a \le b$，一个左开右闭区间 $I=(a,b]$ 指的是 $\{x \in \mathbb{R} : a < x \le b\}$。对于 $\Omega=(0,1]$ 上的区间 $I=(a,b]$，我们定义 $\lambda(I)=b-a$ 为 I 的长度。

区间的长度定义好了，接下来要考虑的集合就是那些**有限**个区间的并。我们说一个集合 $I \subseteq (0,1]$ 是有限个区间的非交并，当且仅当 $I=\bigsqcup_{i=1}^{n} I_i$，其中 $I_i=(a_i,b_i]$ 是一个区间。

> **记号惯例**
>
> 为了省略墨水，从今以后，我们用 $A \sqcup B$ 或者 $\bigsqcup A_n$ 表示非交并（disjoint union），并不再额外说明涉及的集合是非交的了。

对于 $I=\bigsqcup_{i=1}^{n} I_i$，我们定义 $\lambda(I)=\sum_{i=1}^{n}\lambda(I_i)=\sum_{i=1}^{n}(b_i-a_i)$。我们用 $\mathscr{B}_0(\Omega)$ 来表示 Ω 上所有的可以写成有限个区间的并的集合的集合。

我们可以验证，$\mathscr{B}_0(\Omega)$ 对于两个集合的并，以至于任意有限个集合的并是封闭的：因为有限个可以写成有限个区间的并的集合的并可以写成有限个区间的并（好绕口）。那我们的构造是不是到 $\mathscr{B}_0(\Omega)$ 就可以结束了呢？并不是，因为 $\mathscr{B}_0(\Omega)$ 还不是一个 σ-代数，它对于**可数并**并不封闭。比如说，只包含一个点的集合 $\{0.5\} \notin \mathscr{B}_0(\Omega)$，但显然 $\{0.5\}=\bigcap_{n\ge 1}\left(0.5-\frac{1}{2^{n+1}},0.5\right]$（由于对补封闭，由 De-Morgan 律，对可数并封闭等价于对可数交封闭）。

事实上，$\mathscr{B}_0(\Omega)$ 是一个代数（algebra）。

定义 7.1 (代数)

对于一个集合 Ω 和定义在上面的子集集合 $\mathscr{F}$，如果它满足如下条件，我们就称其为代数：

1. $\varnothing \in \mathscr{F}$；
2. 如果 $A \in \mathscr{F}$，则 $A^c \in \mathscr{F}$；
3. 如果 $A, B \in \mathscr{F}$，则 $A \cup B \in \mathscr{F}$。

第三条可以推出对于任何有限个 $A_1, \ldots, A_n \in \mathscr{F}$，我们有 $\bigcup_{i=1}^n A_i \in \mathscr{F}$。这个定义和我们之前学过的 σ-代数唯一的区别就是第三条，我们把可数并的要求弱化成了有限并（σ-代数名字里面的 σ 便是可数并的意思）。

我们同样也可以在代数上定义概率测度。

定义 7.2 (代数上的测度)

我们说非负函数 $\mathbb{P} : (\Omega, \mathscr{F}_0) \to [0,1]$（其中 $\mathscr{F}_0$ 是 Ω 上的一个代数）上的一个（σ-可加）概率测度，当且仅当

1. $\mathbb{P}(\Omega) = 1$；
2. 如果 $A \in \mathscr{F}_0$，那么 $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega)$；
3. 如果不相交的 $A_1, A_2, \cdots \in \mathscr{F}_0$ 并且 $\bigsqcup_{n\geq 1} A_n \in \mathscr{F}_0$，则 $\mathbb{P}(\bigsqcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$。

我们注意到，这儿和 σ-代数 $\mathscr{F}$ 上定义的概率测度唯一的区别是，我们第三条只需要在“如果 $\bigsqcup_{n\geq 1} A_n \in \mathscr{F}_0$”成立的情况下对。因为根据定义，$\mathscr{F}_0$ 对可数并不一定封闭。这一条性质被称为 σ-可加性。

如果我们把第一条里的 $\mathbb{P}(\Omega) = 1$ 去掉，则是测度的定义。有的时候，我们允许 $\mathbb{P}$ 在某些集合上的取值是 ∞，因此，我们会说 $\mathbb{P}$ 的值域是 $[0, \infty]$。这儿，∞ 可以想象成我们在 $\mathbb{R}$ 里添加了一个称为无穷大的数，它的一些运算法则包括：

1. $\forall a \in [0, \infty]$, $a + \infty = \infty$;
2. $\forall a \in (0, \infty]$, $a \cdot \infty = \infty$;
3. $\forall a \in [0, \infty)$, $\infty - a = \infty$;
4. $0 \cdot \infty = 0$;
5. $\infty - \infty$ 无定义.

我们称测度 $\mathbb{P}$ 是有限的，当且仅当 $\mathbb{P}(\Omega) < \infty$，否则称之为无限的。

7.1.2 测度的扩张

我们希望我们的概率论是定义在 σ-代数上的（比如我们希望上面提到的 $\lambda(\{0.5\})$ 是有定义的），因此，我们要扩展 $\mathscr{B}_0(\Omega)$。我们定义 $\mathscr{B}(\Omega) = \sigma(\mathscr{B}_0(\Omega))$ 为包含 $\mathscr{B}_0(\Omega)$ 的最小的 σ-代数。我们之前已经说过这是良定义的，并且也说明了它严格比 $\mathscr{B}_0(\Omega)$ 大。下面这个定理是测度论中的重要定理，它说明，对于代数上的一个（σ-可加）测度，它可以被扩张到更大的 σ-代数上去。

定理 7.1 (Carathéodory 扩张定理)

设 Σ_0 是集合 Ω 上的一个代数（algebra）。假设集合函数 $\mu_0 : \Sigma_0 \to [0, \infty]$ 是 σ-可加的，即对于不相交的 $A_1, A_2, \cdots \in \Sigma_0$，如果 $\bigcup_{n=1}^\infty A_n \in \Sigma_0$，则有

$$\mu_0\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu_0(A_n).$$

那么，μ_0 可以被扩张成 $\Sigma = \sigma(\Sigma_0)$ 上的一个集合函数 $\mu : \Sigma \to [0, \infty]$，并且 μ_0 和 μ 在 Σ_0 上是一致的。

为了应用这个定理，我们必须首先证明 λ 在 $\mathscr{B}_0(\Omega)$ 上是 σ-可加的。证明这件事需要利用到实数集的一些不平凡的性质。我们把它的证明放在本次讲义的最后。

引理 7.1

λ 在 $\mathscr{B}_0(\Omega)$ 上是 σ-可加的。

上述引理和 Carathéodory 扩张定理保证了我们可以把 λ 的定义域从 $\mathscr{B}_0(\Omega)$ 扩张到 $\mathscr{B}(\Omega)$。λ 被称为勒贝格（Lebesgue）测度。

7.2 $\mathbb{R}$ 与 $\mathbb{R}^d$ 上的勒贝格测度

由于我们引入了 ∞，我们前面在 $(0,1]$ 构造勒贝格测度的方法可以无痛推广到 $\mathbb{R}$ 上。也就是说，对于 $a \leq b \in \mathbb{R}$，我们定义 $\lambda((a,b]) = b - a$，然后使用同样的方法把测度扩张到 $\mathbb{R}$ 上所有的 Borel 集 $\mathscr{B}(\mathbb{R})$ 上。我们未来会用 $\mathscr{B}$ 来直接代表 $\mathscr{B}(\mathbb{R})$。注意到，这样定义的勒贝格测度可能会取无穷大，比如 $\lambda(\mathbb{R}) = \infty$。

同样，对于整数 $d \geq 2$，我们可以类似的定义 $\mathbb{R}^d$ 上的勒贝格测度。仅仅需要把区间 $(a,b]$ 换成矩形 $\prod_{i=1}^d (a_i, b_i]$，并定义其体积为 $\prod_{i=1}^d (b_i - a_i)$。

7.3 单调类定理（Monotone Class Theorem）

在未来，我们通常会想做如下的事情：已知某个性质在一个代数 $\mathscr{F}_0$ 上是成立，现在需要证明该性质在生成的 σ-代数，也就是 $\sigma(\mathscr{F}_0)$ 上也是成立的。从代数到 σ-代数，我们需要该验证性质对于可数并是封闭的，这有时候会比较困难，而单调类定理就给了我们处理类似问题的一个工具。

现在固定一个集合 Ω。我们为了表述方便，我们再回顾一下 Ω 上代数的定义。我们说 $\mathscr{F} \subseteq 2^\Omega$ 是一个代数，当且仅当：

1. $\varnothing \in \mathscr{F}$；
2. $A \in \mathscr{F} \implies A^c \in \mathscr{F}$；
3. $A, B \in \mathscr{F} \implies A \cup B \in \mathscr{F}$。

我们定义一类新的集合类 $\mathscr{M} \subseteq 2^\Omega$，叫单调类（Monotone Class），如果其满足：

1. 如果 $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{M}$，则 $\bigcup_{n\geq 1} A_n \in \mathscr{M}$；
2. 如果 $A_1 \supseteq A_2 \supseteq \cdots \in \mathscr{M}$，则 $\bigcap_{n\geq 1} A_n \in \mathscr{M}$。

也就是说 $\mathscr{M}$ 对集合取极限封闭。很显然，一个 σ-代数是一个单调类。对于一个集族 $\mathscr{G}$，我们也会用 $\mathscr{M}(\mathscr{G})$ 来表示包含 $\mathscr{G}$ 的最小的单调类（容易验证这是良定义的）。

单调类定理是说的下面这件事情：

定理 7.2 (单调类定理)

设 $\mathscr{F}_0$ 是一个代数，$\mathscr{F}_0 \subseteq \mathscr{M}$，并且 $\mathscr{M}$ 是一个单调类。那么 $\sigma(\mathscr{F}_0) \subseteq \mathscr{M}$。特别的，$\mathscr{M}(\mathscr{F}_0) = \sigma(\mathscr{F}_0)$。

在证明单调类定理之前，我们先来看一个应用，来解决本节一开始提出的那类问题。这个解决方法也是非常典型的，我们在未来也会多次用到。

7.3.1 单调类定理应用：关于 σ-代数上测度的两个小结论

我们前面构造 Borel 集的方法是先考虑所有的可以写成有限个区间的并的集合的集合 $\mathscr{B}_0$，这是一个代数，在取 $\mathscr{B} = \sigma(\mathscr{B}_0)$。$\mathscr{B}$ 里面的集合的结构是非常复杂的。但下面这个定理说，$\mathscr{B}$ 中每一个集合，均可以被 $\mathscr{B}_0$ 中的某个集合很好的“逼近”。

我们先定义什么叫逼近。

给定两个集合 A, B，我们定义它的对称差 $A \Delta B$ 为 $(A \setminus B) \cup (B \setminus A)$。

定理 7.3

固定 $(\Omega,\mathscr{F},\mathbb{P})$ 为一个概率空间。设 $\mathscr{F}_0\subseteq\mathscr{F}$ 是一个代数。对于任何 $\varepsilon>0$，给定集合 $A\in\sigma(\mathscr{F}_0)$，存在集合 $B_\varepsilon\in\mathscr{F}_0$，满足 $\mathbb{P}[A\Delta B_\varepsilon]\leq\varepsilon$。

证明

我们这儿应用一个典型技巧。我们设 $\mathscr{G}=\{A\subseteq\Omega:\exists B_\varepsilon\in\mathscr{F}_0,\ \mathbb{P}[A\Delta B_\varepsilon]\leq\varepsilon\}$，即所有那些满足我们想要性质的集合的集合。我们只需要证明，$\sigma(\mathscr{F}_0)\subseteq\mathscr{G}$ 就行了。显然，$\mathscr{F}_0\subseteq\mathscr{G}$，由单调类定理，我们只需要证明 $\mathscr{G}$ 是一个单调类即可。

我们现在证明，对于 $A_1\subseteq A_2\subseteq\cdots\in\mathscr{G}$，有 $A=\bigcup_{n\geq1}A_n\in\mathscr{G}$。证明的思路也是比较直接的：由于概率函数的连续性，我们首先可以选取足够大的 N，使得 $\bigcup_{n=1}^N A_n$ 足够接近 A。然后，由于每个 $A_n\in\mathscr{G}$，它们均能够被 $\mathscr{F}_0$ 中集合很好的逼近，因此 $\bigcup_{n=1}^N A_n$ 也能够被 $\mathscr{F}_0$ 中集合很好的逼近。

我们来执行上述证明计划。我们首先选取 N，使得 $\mathbb{P}[A\Delta\bigcup_{n=1}^N A_n]\leq\varepsilon/2$。然后，对于每一个 $n=1,2,\ldots,N$，我们选取 $B_n\in\mathscr{F}_0$，满足 $\mathbb{P}[A_n\Delta B_n]\leq\varepsilon/2^{n+1}$。最后，我们定义 $B_\varepsilon:=\bigcup_{n=1}^N B_n$。我们现在来说明 B_ε 是对 A 的一个足够好的逼近。

我们首先可以用定义验证，对于任意两个集合 X,Y，如果 $X'\subseteq X$，那么 $\mathbb{P}[X\Delta Y]\leq\mathbb{P}[X\setminus X']+\mathbb{P}[X'\Delta Y]$。于是，

$$\mathbb{P}[A\Delta B_\varepsilon]\leq\mathbb{P}\left[A\setminus\left(\bigcup_{n=1}^N A_n\right)\right]+\mathbb{P}\left[\left(\bigcup_{n=1}^N A_n\right)\Delta B_\varepsilon\right]\leq\frac{\varepsilon}{2}+\mathbb{P}\left[\left(\bigcup_{n=1}^N A_n\right)\Delta\left(\bigcup_{n=1}^N B_n\right)\right].$$

根据定义，我们又可以验证

$$\left(\bigcup_{n=1}^N A_n\right)\Delta\left(\bigcup_{n=1}^N B_n\right)\subseteq\bigcup_{n=1}^N(A_n\Delta B_n).$$

因此，由 union-bound，我们有

$$\mathbb{P}\left[\left(\bigcup_{n=1}^N A_n\right)\Delta\left(\bigcup_{n=1}^N B_n\right)\right]\leq\mathbb{P}\left[\bigcup_{n=1}^N(A_n\Delta B_n)\right]\leq\sum_{n=1}^\infty\frac{\varepsilon}{2^{n+1}}\leq\frac{\varepsilon}{2}.$$

这说明，$\mathbb{P}[A\Delta B_\varepsilon]\leq\varepsilon$ 成立。

我们可以类似的证明，对于 $A_1\supseteq A_2\supseteq\cdots\in\mathscr{G}$，$\bigcap_{n\geq1}A_n\in\mathscr{G}$ 也成立。这儿就不再赘述了。

我们用类似的技巧来说明，对于一个使用 Carathéodory 扩张定理得到的 σ-代数上的概率测度是唯一的。具体来说，我们假设可测空间 $(\Omega,\mathscr{F})$ 上的概率测度 $\mathbb{P}$ 是由某个代数 $\mathscr{F}_0\subseteq\mathscr{F}$ 上的测度扩张而得并且 $\mathscr{F}=\sigma(\mathscr{F}_0)$。那么，这个扩张是唯一的。

假设存在两个概率测度 $\mathbb{P}$ 和 $\mathbb{Q}$，满足在 $\mathscr{F}_0$ 上一致。我们现在定义 $\mathscr{G}=\{A\in\mathscr{F}:\mathbb{P}(A)=\mathbb{Q}(A)\}$，即所有在测度 $\mathbb{P}$ 和测度 $\mathbb{Q}$ 上一致的那些集合的集合。显然 $\mathscr{F}_0\subseteq\mathscr{G}$。要证明我们的结论，根据单调类定理，我们只需要证明 $\mathscr{G}$ 是单调类。

我们假设 $A_1\subseteq A_2\subseteq\cdots\in\mathscr{G}$，那由测度的连续性

$$\mathbb{P}\left(\bigcup_{n\geq1}A_n\right)=\lim_{N\to\infty}\mathbb{P}\left(\bigcup_{n=1}^N A_n\right)=\lim_{N\to\infty}\mathbb{Q}\left(\bigcup_{n=1}^N A_n\right)=\mathbb{Q}\left(\bigcup_{n\geq1}A_n\right).$$

换句话说，$\bigcup_{n\geq1}A_n\in\mathscr{G}$。对于 $A_1\supseteq A_2\supseteq\cdots\in\mathscr{G}$ 的情形我们可以同样证明。

7.3.2 单调类定理的证明

我们这一小节来证明单调类定理。我们实际上只需要证明 $\mathscr{M}(\mathscr{F}_0)=\sigma(\mathscr{F}_0)$ 即可（因为根据定义，$\mathscr{M}(\mathscr{F}_0)\subseteq\mathscr{M}$）。容易验证，如果一个集合同时是单调类与代数，那么它就是 σ-代数，因此，我们只需要验证 $\mathscr{M}(\mathscr{F}_0)$ 是代数即可。

由于 $\mathscr{M}(\mathscr{F}_0)$ 包含了 $\mathscr{F}_0$，因此，验证其为代数，只需要验证

1. 其对求补封闭，即 $A \in \mathscr{M}(\mathscr{F}_0) \implies A^c \in \mathscr{M}(\mathscr{F}_0)$；
2. 其对求（有限）交封闭，即 $A, B \in \mathscr{M}(\mathscr{F}_0) \implies A \cap B \in \mathscr{M}(\mathscr{F}_0)$。

我们验证的策略，正如同我们之前应用单调类定理那般，便是把所有满足条件的集合拿出来放一起，再证明它们组成了一个单调类即可。

1. 我们先来验证 $\mathscr{M}(\mathscr{F}_0)$ 对求补封闭。我们设 $\mathscr{G}_1 := \{A \in \mathscr{M}(\mathscr{F}_0) : A^c \in \mathscr{M}(\mathscr{F}_0)\}$. 那么对于 $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{G}_1$，我们有
$$\left(\bigcup_{n\geq 1} A_n\right)^c = \bigcap_{n\geq 1} A_n^c.$$
由于 $A_n^c \in \mathscr{M}(\mathscr{F}_0)$，所以 $\bigcap_{n\geq 1} A_n^c \in \mathscr{M}(\mathscr{F}_0)$，这说明 $\bigcup_{n\geq 1} A_n \in \mathscr{G}_1$。我们对于递降的集合 $A_1 \supseteq A_2 \supseteq \cdots \in \mathscr{G}_1$ 也可以同样说明 $\bigcap_{n\geq 1} A_n \in \mathscr{G}_1$。因此 $\mathscr{G}_1$ 是单调类，而根据定义 $\mathscr{G}_1 \subseteq \mathscr{M}(\mathscr{F}_0)$ 且 $\mathscr{F}_0 \subseteq \mathscr{G}_1$，所以 $\mathscr{G}_1 = \mathscr{M}(\mathscr{F}_0)$。即 $\mathscr{M}(\mathscr{F}_0)$ 对补封闭。
2. 我们接着验证 $\mathscr{M}(\mathscr{F}_0)$ 对求有限交封闭。设 $\mathscr{G}_2 := \{A \in \mathscr{M}(\mathscr{F}_0) : \forall B \in \mathscr{M}(\mathscr{F}_0), A \cap B \in \mathscr{M}(\mathscr{F}_0)\}$. 同样，我们想说 $\mathscr{G}_2 = \mathscr{M}(\mathscr{F}_0)$。这需要验证两件事：首先 $\mathscr{G}_2$ 是单调类，并且 $\mathscr{F}_0 \subseteq \mathscr{G}_2$。
   - 先来验证 $\mathscr{G}_2$ 是单调类。设 $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{G}_2$，对于任何的 $B \in \mathscr{M}(\mathscr{F}_0)$，有
   $$\left(\bigcup_{n\geq 1} A_n\right) \cap B = \bigcup_{n\geq 1} (A_n \cap B).$$
   由于 $A_n \cap B \in \mathscr{M}(\mathscr{F}_0)$，由单调类的性质，$\bigcup_{n\geq 1}(A_n \cap B) \in \mathscr{M}(\mathscr{F}_0)$，也就是说 $\bigcup_{n\geq 1} A_n \in \mathscr{G}_2$。对于递减的 A_n，我们也可以同样验证 $\mathscr{G}_2$ 对于它们的极限封闭。因此 $\mathscr{G}_2$ 是单调类。

   - 接着我们要说明 $\mathscr{F}_0 \subseteq \mathscr{G}_2$。我们要做一个类似于“跷跷板”的操作。定义 $\mathscr{G}_3 := \{A \in \mathscr{M}(\mathscr{F}_0) : \forall B \in \mathscr{F}_0, A \cap B \in \mathscr{M}(\mathscr{F}_0)\}$. 注意到，和 $\mathscr{G}_2$ 的不同在于后者只考虑 $B \in \mathscr{F}_0$。这样可以直接得到 $\mathscr{F}_0 \subseteq \mathscr{G}_3$。我们可以用对于 $\mathscr{G}_2$ 同样的的方法说明 $\mathscr{G}_3$ 是单调类，因此，$\mathscr{G}_3 = \mathscr{M}(\mathscr{F}_0)$。有了这个结论，我们就可以说，对于任意 $B \in \mathscr{F}_0$，其满足对于任何 $A \in \mathscr{M}(\mathscr{F}_0) = \mathscr{G}_3$，$A \cap B \in \mathscr{M}(\mathscr{F}_0)$，也就是说 $B \in \mathscr{G}_2$。

7.4 λ 在 $\mathscr{B}_0((0,1])$ 上 σ-可加性的验证

我们最后来验证定义在 $\mathscr{B}_0((0,1])$ 上的 λ 的 σ-可加性。这是应用 Carathéodory 扩张定理把 λ 从 $\mathscr{B}_0((0,1])$ 扩张到 $\mathscr{B}((0,1])$ 的必要条件。

我们先证明一个重要的引理。

引理 7.2

如果 $I = \bigsqcup_{k\geq 1} I_k$，其中 $I_k = (a_k, b_k]$，则 $\lambda(I) = \sum_{k\geq 1} \lambda(I_k)$。 ♡

这个引理的强大之处在于它从 λ 定义中的有限可加性突破到了可数可加性，实现了从有限到无限的突破。而这一点成立的关键原因是实数集上有界闭区间的紧性。我们接下来的讨论里假设 $I = (a, b]$。

证明

1. 首先简单说明一下如果 $\bigsqcup_{k\geq 1} I_k \subseteq I$，那么 $\sum_{k\geq 1} \lambda(I_k) \leq \lambda(I)$。我们只需要对任意 $N > 0$，证明 $\sum_{k=1}^N \lambda(I_k) \leq \lambda(I)$，再取极限即可。而对于有限的 N，我们可以使用归纳法证明。这个比较简单，留作练习。
2. 我们接着来说明如果 $I \subseteq \bigcup_{k\geq 1} I_k$，那么 $\lambda(I) \leq \sum_{k\geq 1} \lambda(I_k)$。注意到我们这儿不要求 I_k 是不相交的，结论也正确。同样的，有限和的情况是可以用归纳法容易证明的，也就是说，如果对于 N，有 $I \subseteq \bigcup_{k=1}^N I_k$，那么 $\lambda(I) \leq \sum_{k=1}^N \lambda(I_k)$。现在假设 $N = \infty$。根据我们的条件，对于任何 $\varepsilon \in (0, b-a)$，我们均有 $[a+\varepsilon, b] \subseteq \bigcup_{k\geq 1}(a_k, b_k + \varepsilon 2^{-k})$。Heine-Borel 定理说明闭区间是紧集，也就是闭区间的每一个开覆盖一定存在一个有

限的子覆盖。所以，存在一个 $N \in \mathbb{N}$，满足 $[a+\varepsilon, b] \subseteq \bigcup_{k=1}^{N}(a_k, b_k+\varepsilon 2^{-k}]$。根据有限情况的结论，我们有

$$b-(a+\varepsilon) \leq \sum_{k=1}^{N}(b_k+\varepsilon 2^{-k}-a_k) \leq \sum_{k\geq 1}(b_k-a_k)+\varepsilon.$$

由于 ε 可以取任意小，故得证。

有了这个引理，λ 的 σ-可加性便容易验证了。我们假设不相交的 $A_1, A_2, \cdots \in \mathscr{B}_0((0,1])$，并且 $A := \bigsqcup_{k\geq 1} A_k \in \mathscr{B}_0((0,1])$。由于它们都是 $\mathscr{B}_0((0,1])$ 里的元素，因此可以假设 $A = \bigsqcup_{i=1}^{n} I_i$，$A_k = \bigsqcup_{j=1}^{m_k} J_{kj}$，并且 $I_i = \bigsqcup_{k\geq 1, j\in[m_k]}(I_i \cap J_{kj})$。那么，根据刚才的引理，我们有

$$\lambda(A) = \sum_{i=1}^{n}\lambda(I_i) = \sum_{i=1}^{n}\sum_{k\geq 1}\sum_{j=1}^{m_k}\lambda(I_i\cap J_{kj}) = \sum_{k\geq 1}\sum_{j=1}^{m_k}\lambda(J_{kj}) = \sum_{k\geq 1}\lambda(A_k).$$

7.5 参考书籍

本课程并不会对测度论进行全面的介绍，只会引入我们未来将用到的内容和概念。如果想了解更详细的内容，可以参考如下两本非常优秀的教材。

- Probability and Measure, by Patrick Billingsley.
- An Introduction to Measure Theory, by Terence Tao.

第 8 章 一般概率空间上的随机变量

在有了上节课的基础之后，我们终于可以来定义一般概率空间上的随机变量，以及研究它的一些性质。我们会发现，在一般的概率空间上，许多概念与离散的概率空间会有一些不一样，主要的原因在于一般的概率空间的结构内容更丰富了，以致于不少我们直观上认为会正确的东西不再一定正确。但在看到这些性质的时候，请务必回忆一下在离散场合对应的是什么，如果有所不同，想想为什么。

8.1 随机变量与可测函数

给定一个样本集 Ω 以及定义在上面的 σ-代数 $\mathscr{F} \subseteq 2^{\Omega}$。我们称 $(\Omega, \mathscr{F})$ 为一个可测空间。我们假设上面有一个概率测度 $\mathbb{P}$，因此，$(\Omega, \mathscr{F}, \mathbb{P})$ 是一个概率空间。给定一个（实值）函数 $f : \Omega \to \mathbb{R}$，我们说 f 是一个**可测函数**，当且仅当对于任何 Borel 集 $A \in \mathscr{B}$，$f^{-1}(A) := \{\omega \in \Omega : f(\omega) \in A\} \in \mathscr{F}$. 所谓随机变量，实际上就是定义在 Ω 上的可测函数。不过惯例上，我们会用大写的字母 $X, Y, \ldots$ 来表示随机变量。

回忆在离散概率空间的时候，我们把任何 $\Omega \to \mathbb{R}$ 上的函数都称为随机变量，这是因为在那个时候，我们总是把 $\mathscr{F}$ 取成全集 2^{Ω}，因此任何函数都是可测的。可测的要求是非常自然的。因为我们在研究概率的时候，我们需要对于 Borel 集 A，讨论 $\mathbb{P}[X \in A]$ 的概率。由于 $\mathbb{P}$ 是定义在 $\mathscr{F}$ 上的函数，我们必须要要求 $[X \in A] \in \mathscr{F}$（我们有时候用 $X^{-1}(A)$ 来表示 $[X \in A]$）。

8.1.1 验证随机变量

我们对于可测性的要求使得任意给一个函数 $X : \Omega \to \mathbb{R}$，其是否是随机变量是需要验证的。实际上，我们并不需要对于所有的 Borel 集 $A \in \mathscr{B}$，来验证 $[X \in A] \in \mathscr{F}$。我们只需考虑那些形如 $(-\infty, r]$，其中 r 是有理数的集合即可。

命题 8.1

X 是随机变量当且仅当对于每一个有理数 $r \in \mathbb{Q}$，$[X \leq r] \in \mathscr{F}$。 ♠

证明

“仅当”是显然的。我们来验证“当”，也就是说，如果对于每一个 $r \in \mathbb{Q}$，$[X \leq r] \in \mathscr{F}$，那么对于任何 $A \in \mathscr{B}$，$[X \in A] \in \mathscr{F}$。我们定义 $\mathscr{G} = \{A \subseteq \mathbb{R} : [X \in A] \in \mathscr{F}\}$

设 $\mathscr{G}_0 = \{(-\infty, r] : r \in \mathbb{Q}\}$。我们知道 $\mathscr{G}_0 \subseteq \mathscr{G}$。我们先验证 $\mathscr{G}$ 是 σ-代数，于是就有 $\sigma(\mathscr{G}_0) \subseteq \mathscr{G}$。然后我们验证 $\sigma(\mathscr{G}_0) = \mathscr{B}$ 就可以了。

我们直接根据定义来验证 $\mathscr{G}$ 是 σ-代数。首先 $\varnothing \in \mathscr{G}$。如果 $A \in \mathscr{G}$，那么意味着 $[X \in A] \in \mathscr{F}$。因此 $[X \in A^c] = [X \in A]^c \in \mathscr{F}$。所以 $A^c \in \mathscr{G}$。类似的，如果 $A_1, A_2, \cdots \in \mathscr{G}$，那么 $\bigcup_{n\geq 1}[X \in A_n] = [X \in \bigcup_{n\geq 1} A_n] \in \mathscr{F}$。这说明 $\bigcup_{n\geq 1} A_n \in \mathscr{G}$。因此 $\mathscr{G}$ 是 σ-代数。

我们要证明 $\sigma(\mathscr{G}_0) = \mathscr{B}$，稍微思索一下即可发现，我们只需要证明使用求补、求可数交的操作，能够从 $\mathscr{G}_0$ 得到 $\mathscr{B}_0$ 即可。那么，现在给定 $(a, b] \in \mathscr{B}_0$，其中 $a \leq b \in \mathbb{R}$，我们显然有

$$(a, b] = (-\infty, b] \setminus (-\infty, a].$$

而对于任何一个实数 $a \in \mathbb{R}$，我们总可以找到一列递减的有理数 $r_1, r_2, \ldots$，满足 $\lim_{n\to\infty} r_n = a$。于是，

$$(-\infty, a] = \bigcap_{n\geq 1} (-\infty, r_n].$$

命题 8.2

设 $X, Y : \Omega \to \mathbb{R}$ 是随机变量。

1. 对于任意实数 a，aX 是随机变量；
2. $X + Y$ 与 XY 是随机变量；
3. 定义 $Z : \omega \in \Omega \mapsto \begin{cases} Y(\omega)/X(\omega) & \text{if } X(\omega) \neq 0, \\ 0 & \text{if } X(\omega) = 0, \end{cases}$ 那么 Z 是随机变量；
4. 设 $f : \mathbb{R} \to \mathbb{R}$ 为 $\mathscr{B}$ 上的一个可测函数（又称 Borel 函数），那么 $f(X)$ 是随机变量。

我们接下来验证一下 $X + Y$ 是随机变量。剩余的一些，我们留成练习。

我们只需要对于任意 $a \in \mathbb{R}$，说明 $[X + Y > a] \in \mathscr{F}$ 即可（why?）。实际上

$$[X + Y > a] = \bigcup_{r\in\mathbb{Q}} ([X > r] \cap [Y > a - r]).$$

8.1.2 构造概率空间：从有限次试验到无穷次试验

在概率论中，一个完整的概率空间由三元组 $(\Omega, \mathcal{F}, \mathbb{P})$ 构成，其中 Ω 是样本空间，$\mathcal{F}$ 是事件域（一个 σ-代数），而 $\mathbb{P}$ 是定义在 $\mathcal{F}$ 上的概率测度。本小节将通过“独立地扔无穷多个均匀硬币”这一经典模型，系统性地展示如何一步一步地构造出这个三元组，特别是如何引入并定义概率测度 $\mathbb{P}$。

我们考虑一个无限序列的随机试验：独立地抛掷一枚均匀硬币无穷多次。每一次抛掷的结果是正面（记为 1）或反面（记为 0）。因此，一个完整的试验结果可以表示为一个无穷长的二进制序列：

$$\omega = (\omega_1, \omega_2, \omega_3, \ldots), \quad \text{其中} \omega_i \in \{0, 1\}.$$

于是，我们的样本空间 Ω 定义为所有这样的无穷序列的集合：

$$\Omega = \{0, 1\}^{\mathbb{N}}.$$

对于任意一个有限长度的二进制串 $s = (s_1, s_2, \ldots, s_n) \in \{0, 1\}^n$，我们定义一个“基本事件” C_s，它表示“前 n 次抛掷的结果恰好是 s”：

$$C_s := \{\omega \in \Omega \mid \omega_1 = s_1, \omega_2 = s_2, \ldots, \omega_n = s_n\}.$$

这些基本事件 C_s 是我们构建整个概率空间的基石。

为了确保我们的构造是良定义的，首先需要证明，对于任意固定的 n，所有长度为 n 的基本事件 $\{C_s\}_{s\in\{0,1\}^n}$ 构成了样本空间 Ω 的一个划分。

证明 对于任意一个无穷序列 $\omega \in \Omega$，其前 n 位 $\omega_{[1:n]} = (\omega_1, \ldots, \omega_n)$ 必然属于 $\{0, 1\}^n$ 中的某一个元素 s。因此，ω 必然属于 C_s。这说明 $\bigcup_{s\in\{0,1\}^n} C_s = \Omega$。

另一方面，如果 $s_1 \neq s_2$，那么 C_{s_1} 和 C_{s_2} 所要求的前 n 位不同，因此它们不可能有共同的元素，即 $C_{s_1} \cap C_{s_2} = \varnothing$。

综上所述，$\{C_s\}_{s\in\{0,1\}^n}$ 是 Ω 的一个划分。

现在，我们考虑只关心前 n 次抛掷结果的所有可能事件。这些事件构成了一个 σ-代数 $\mathcal{F}_n$，它是包含所有基本事件 $\{C_s\}_{s\in\{0,1\}^n}$ 的最小 σ-代数。我们可以构造一个映射 $f:\mathcal{F}_n\to 2^{\{0,1\}^n}$，它将每个事件 $A\in\mathcal{F}_n$ 映射为其在 $\{0,1\}^n$ 上的“投影”：

$$f(A):=\bigcup_{\omega\in A}\{(\omega_1,\omega_2,\dots,\omega_n)\}.$$

这个映射是一个双射。因为 $\mathcal{F}_n$ 中的任意元素都可以表示为若干个基本事件 C_s 的并集，而 $f(C_s)$ 就是单点集 $\{s\}$。由于 $\{0,1\}^n$ 有 2^n 个元素，所以 $\mathcal{F}_n$ 也有 2^n 个元素，且 f 是一个一一对应。

基于此，我们可以在 $(\{0,1\}^n,\mathcal{F}_n)$ 上定义一个概率测度 $\mathbb{P}_n$，使得每个基本事件 C_s 的概率为 $\frac{1}{2^n}$。由于 $\mathcal{F}_n$ 中的任意事件都是若干个互斥的基本事件的并，其概率自然就是这些基本事件概率之和。这就是独立地抛 n 次均匀硬币所对应的概率空间。

接下来，我们考虑随着试验次数增加，事件域是如何变化的。令 $\mathcal{F}_i$ 表示只关心前 i 次试验结果的事件域。显然，如果我们知道了前 $i+1$ 次的结果，我们当然也知道了前 i 次的结果。因此，$\mathcal{F}_i$ 中的任何一个事件都可以用 $\mathcal{F}_{i+1}$ 中的事件来描述，即 $\mathcal{F}_i\subseteq\mathcal{F}_{i+1}$。这样的一列嵌套的 σ-代数 $\{\mathcal{F}_n\}_{n=1}^{\infty}$ 被称为一个滤链（filtration）。

我们定义 $\mathcal{F}_\infty:=\bigcup_{n\geq 1}\mathcal{F}_n$。这是一个代数（algebra），但还不是 σ-代数，因为它对可数并运算不封闭。

证明 验证 $\mathcal{F}_\infty$ 是一个代数：

1. 空集 $\varnothing$ 属于 $\mathcal{F}_1$，因此 $\varnothing\in\mathcal{F}_\infty$。
2. 对于任意 $A\in\mathcal{F}_\infty$，存在某个 n 使得 $A\in\mathcal{F}_n$，则其补集 A^c 也属于 $\mathcal{F}_n$，故 $A^c\in\mathcal{F}_\infty$。
3. 对于任意 $A,B\in\mathcal{F}_\infty$，存在 n,m 使得 $A\in\mathcal{F}_n,B\in\mathcal{F}_m$。不妨设 $n\leq m$，则 $A,B\in\mathcal{F}_m$，所以它们的并集 $A\cup B\in\mathcal{F}_m\subseteq\mathcal{F}_\infty$。综上，$\mathcal{F}_\infty$ 是一个代数。

为了得到一个真正的 σ-代数，我们需要取 $\mathcal{F}_\infty$ 的闭包。令 $\mathscr{B}(\Omega)=\sigma(\mathcal{F}_\infty)$ 为包含 $\mathcal{F}_\infty$ 的最小 σ-代数。这是我们在无穷样本空间上所能定义的“最大”的事件域。

值得注意的是，对于任意一个单点集 $\{\omega\}$，它不属于 $\mathcal{F}_\infty$，因为没有任何有限次试验能唯一确定一个无穷序列。但它属于 $\mathscr{B}(\Omega)$，因为我们可以将其表示为一列递减的基本事件的交：

$$\{\omega\}=\bigcap_{n=1}^{\infty}C_{(\omega_1,\dots,\omega_n)}.$$

由于每个 $C_{(\omega_1,\dots,\omega_n)}\in\mathcal{F}_n\subseteq\mathcal{F}_\infty\subseteq\mathscr{B}(\Omega)$，并且 $\mathscr{B}(\Omega)$ 对可数交运算封闭，所以 $\{\omega\}\in\mathscr{B}(\Omega)$。

现在，我们定义一个函数 $\mu:\mathcal{F}_\infty\to[0,1]$，对于任意 $A\in\mathcal{F}_n$，令 $\mu(A)=\frac{k}{2^n}$，其中 k 是 A 可以分解成的互斥基本事件 C_s 的个数。

关键问题是：这个定义是否良好？也就是说，同一个事件 A 如果可以用不同的 n 和 k 来表示，其比值 $\frac{k}{2^n}$ 是否恒定？

证明 假设 $A\in\mathcal{F}_n$，且 $A=\bigcup_{j=1}^{k}C_{s_j}$，其中 $s_j\in\{0,1\}^n$。令 n_0 为满足 $A\in\mathcal{F}_{n_0}$ 的最小整数，$k_{n_0}=|f_{n_0}(A)|$。

我们构造一个新的集合 $S'=\{(s_1,\dots,s_{n_0},0),(s_1,\dots,s_{n_0},1)\mid s=(s_1,\dots,s_{n_0})\in f_{n_0}(A)\}$。显然，$S'$ 中的元素都属于 $\{0,1\}^{n_0+1}$，且 $A=\bigcup_{s'\in S'}C_{s'}$。此时，$|S'|=2k_{n_0}$。

因此，如果我们将 A 视为 $\mathcal{F}_{n_0+1}$ 中的事件，其对应的 k 值为 $2k_{n_0}$，分母为 2^{n_0+1}，比值仍为 $\frac{2k_{n_0}}{2^{n_0+1}}=\frac{k_{n_0}}{2^{n_0}}$。

通过归纳法，我们可以证明，对于任意 $n>n_0$，如果 $A=\bigcup_{j=1}^{k_n}C_{s_j}$，其中 $s_j\in\{0,1\}^n$，则总有 $\frac{k_n}{2^n}=\frac{k_{n_0}}{2^{n_0}}$。这意味着比值 $\frac{k}{2^n}$ 只依赖于事件 A 本身，而与我们选择的 n 和 k 无关。因此，μ 在 $\mathcal{F}_\infty$ 上是良定义的。

最后一步，也是最关键的一步，是将 μ 从代数 $\mathcal{F}_\infty$ 扩张到 σ-代数 $\mathscr{B}(\Omega)$ 上。这需要用到著名的 Carathéodory 扩张定理。

该定理指出，如果一个函数 μ 在一个代数 $\mathcal{A}$ 上是 σ-可加的（即对于任意一列互不相交的集合 $A_1,A_2,\dots\in\mathcal{A}$，若它们的并集也在 $\mathcal{A}$ 中，则 $\mu(\bigcup_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}\mu(A_i)$），那么 μ 可以唯一地扩张为定义在 $\sigma(\mathcal{A})$ 上的概率测度 $\mathbb{P}$。

因此，我们只需验证 μ 在 $\mathcal{F}_\infty$ 上的 σ-可加性。

证明 考虑一列互不相交的集合 $A_1,A_2,\dots\in\mathcal{F}_\infty$，且它们的并集 $A=\bigcup_{i=1}^{\infty}A_i$ 也属于 $\mathcal{F}_\infty$。

根据第 6 问的结论，存在某个 n_0 和 k_0，以及一些基本事件 $C_{s_j^{(0)}}$，使得 $A=\bigcup_{j=1}^{k_0}C_{s_j^{(0)}}$，且 $\mu(A)=$

$\sum_{j=1}^{k_0} \mu(C_{s_j^{(0)}})$。

同时，对于每个 A_i，它也可以被分解为若干个基本事件 $C_{s_j^{(i)}}$ 的并。由于 A_i 之间互不相交，这些基本事件 $C_{s_j^{(i)}}$ 之间也必然互不相交。

因此，$\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$ 成立。这证明了 μ 的 σ-可加性。

至此，我们成功地应用 Carathéodory 扩张定理，将 μ 扩张为定义在 $(\Omega, \mathscr{B}(\Omega))$ 上的概率测度 $\mathbb{P}$。我们最终构造出了独立地抛无穷多个均匀硬币所对应的完备概率空间 $(\Omega, \mathscr{B}(\Omega), \mathbb{P})$。

同时说明一个有趣的事实：在上文中，我们使用了记号 $\mathscr{B}$ 来表示 $[0,1]$ 上所有的 Borel 集。这是因为，在构造过程中，我们实际上建立了一个从样本空间 Ω 到区间 $[0,1]$ 的映射 $f : \Omega \setminus S \to (0,1]$，其中 S 是所有尾部为无穷多个 0 的序列的集合。

这个映射 f 将每一个无穷二进制序列 ω 映射为其对应的二进制小数的值，即 $f(\omega) = \sum_{j=1}^{\infty} \omega_j \cdot 2^{-j}$。这个映射几乎是双射的（除了那些尾部为无穷多个 0 或无穷多个 1 的序列，它们会映射到同一个实数）。

通过这个映射，我们将概率空间 $(\Omega, \mathscr{B}(\Omega), \mathbb{P})$ 与区间 $[0,1]$ 上的 Borel 测度空间建立了联系。由于 S 是一个零测集（其测度为 0），忽略它不会影响任何概率计算。因此，我们可以说，独立抛无穷硬币的过程等价于在 $[0,1]$ 上均匀采样一个实数。

8.2 随机变量的分布函数（Distribution Function）

一个随机变量 X 唯一决定了一个 $\mathbb{R} \to [0,1]$ 的函数 F_X:

$$F_X(a) := \mathbb{P}[X \le a].$$

我们把 F_X 称为 X 的分布函数（Distribution Function），或者累积分布函数（Cumulative Distribution Function, CDF）。

图 8.1: Figure 1

设 X 是投掷一个六面骰子得到的的点数，那么它对应的分布函数的图像如 Figure 1 所示。

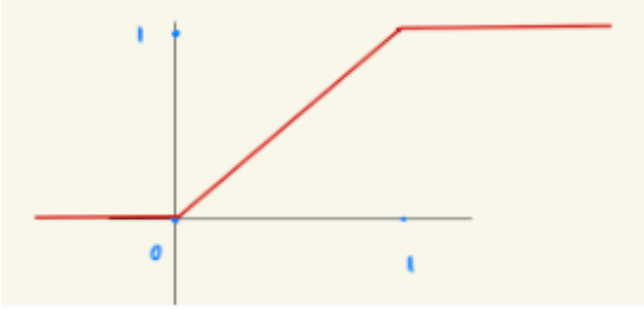


图 8.2: Figure 2

设 X 是从 $[0,1]$ 上均匀取的一个数，那么它对应的分布函数的图像如 Figure 2 所示。

8.2.1 分布函数的基本性质

分布函数有一些基本性质，我们罗列一些。

命题 8.3 (分布函数的基本性质)

设 X 是一个随机变量并且 F 是它的分布函数，那么对于任何的 $x, y \in \mathbb{R}$，以下结论成立。

1. $0 \leq F(x) \leq 1$；
2. $x \leq y \implies F(x) \leq F(y)$；
3. $\lim_{x \to -\infty} F(x) = 0$ 并且 $\lim_{x \to \infty} F(x) = 1$；
4. $\lim_{y \downarrow x} F(y) = F(x)$；
5. $F(x-) := \lim_{y \uparrow x} F(y)$ 存在；
6. F 具有至多可数个间断点。

♠

这些性质大部分使用定义可以直接验证。其中 6 是我们数学分析课中证明过的单调函数的间断点至多可数个的性质。而 5 成立的原因是有上界的单调非降序列极限一定存在。我们来验证一下 4，它告诉我们每一个分布函数都是右连续的。同时满足 4 和 5 的函数被称作 càdlàg 的。

Remark

càdlàg 是个法文词，大概意思是数学中定义在实数集或其子集上的函数，具有处处右连续且左极限存在的特性；

我们定义一列递减的数 $\{x + \frac{1}{n}\}_{n \geq 1}$。那么

$$\lim_{y \downarrow x} F(y) = \lim_{n \to \infty} F\left(x + \frac{1}{n}\right) = \lim_{n \to \infty} \mathbb{P}\left[X \leq x + \frac{1}{n}\right] = \mathbb{P}[X \leq x] = F(x).$$

接下来这些随机变量和分布函数的关系也是比较容易验证的，我列出来，证明留作练习。

命题 8.4

设 X 是一个随机变量并且 F 是它的分布函数。那么如下结论成立。

1. $\mathbb{P}[X < x] = F(x-)$；
2. $\mathbb{P}[X = x] = F(x) - F(x-)$；
3. 如果 $a < b$，则 $\mathbb{P}[a < X \leq b] = F(b) - F(a)$；
4. $\mathbb{P}[X > x] = 1 - F(x)$。

♠

我们之前定义过随机变量 X 的分布 μ_X。它是 $(\mathbb{R}, \mathscr{B})$ 上的一个概率测度，满足对于任何 $A \in \mathscr{B}$，$\mu_X(A) = \mathbb{P}[X = A]$。可以看到，它可以由分布函数 F_X 直接给出：对于任何 $(a, b] \in \mathscr{B}_0$，我们有 $\mu_X((a, b]) = F(b) - F(a)$。然后使用扩张定理把这个测度唯一的扩张到 $\mathscr{B}$ 上即可。

8.2.2 分布函数和随机变量的等价性

我们上面说了每一个随机变量都可以定义出一个分布函数，并且这个分布函数满足若干性质。我们现在想说，如果有一个 $\mathbb{R} \to \mathbb{R}$ 的函数 F，它满足我们上一节第一个命题中前四条性质，那么也能够构造出一个随机变量，使得它的分布函数正好是 F。我们现在给出这个构造。

基本的想法是先找到函数 F 的逆 F^{-1}。但由于 F 可能有间断点，我们没有办法找到完美的逆，因此定义函数 $G : (0, 1) \to \mathbb{R}$ 满足

$$G(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}.$$

注意到，在 F 是连续函数的情况下，$G = F^{-1}$。对于我们这儿的 càdlàg 的 F，容易验证，如下命题依然成立。

命题 8.5

对于任意 $u \in (0,1)$ 和 $x \in \mathbb{R}$，$G(u) \le x \iff u \le F(x)$。

我们现在构造一个以 F 为分布的随机变量。设概率空间为 $((0,1),\mathscr{B}((0,1)),\mathbb{P})$，其中 $\mathbb{P}$ 为 $(0,1)$ 上的均匀分布。设 $U : x \in (0,1) \mapsto x$ 为恒等函数，我们构造的随机变量为

$$G(U) : x \in (0,1) \mapsto G(U(x)).$$

根据我们说明的 G 的性质，

$$\mathbb{P}[G(U) \le x] = \mathbb{P}[U \le F(x)].$$

但由于 U 是 $(0,1)$ 上的均匀分布，所以 $\mathbb{P}[U \le F(x)] = F(x)$。

这个构造还告诉我们一件事情。假设在知道分布函数 F 的情况下，如果从 F 定义的分布中进行采样呢？我们可以先均匀的从 $(0,1)$ 中选一个 u，再输出 $G(u)$。

8.3 随机变量的独立性

我们之前对于离散的随机变量定义了独立的概念，即 X,Y 独立，当且仅当对于任何 x,y，$\mathbb{P}[X = x \wedge Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$。这个定义对于一般的随机变量是不正确的。我们修正如下。

定义 8.1

对于定义在概率空间 $(\Omega,\mathscr{F},\mathbb{P})$ 上的两个随机变量 X,Y，我们说它们是独立的，记作 $X \perp Y$，当且仅当对于任何 $A,B \in \mathscr{B}$，

$$\mathbb{P}[X \in A \wedge Y \in B] = \mathbb{P}[X \in A] \cdot \mathbb{P}[Y \in B].$$

同样，我们接下来说明，如果要验证两个随机变量是不是独立，我们只需要取 A,B 是形如 $(-\infty,r], r \in \mathbb{Q}$ 这样的集合就够了。我们固定一个 $r \in \mathbb{Q}$，然后设 $\mathscr{G} = \{A \subseteq \mathbb{R} : \mathbb{P}[X \le r \wedge Y \in A] = \mathbb{P}[X \le r] \cdot \mathbb{P}[Y \in A]\}$。我们现在来证明 $\mathscr{G}$ 包含 $\mathscr{B}$。设 $\mathscr{B}'$ 为所有可以写成形如 $(a,b], a \le b \in \mathbb{Q}$ 的区间的有限并的集合的集合。显然 $\mathscr{B}'$ 是一个代数（和 $\mathscr{B}_0$ 的区别是这儿我们要求区间的端点是有理数）。我们前面也验证过 $\sigma(\mathscr{B}') = \mathscr{B}$。

我们使用定义，可以比较容易的验证 $\mathscr{B}' \subseteq \mathscr{G}$，这儿不再细说。为了使用单调类定理，我们只需要验证 $\mathscr{G}$ 是单调类即可。考虑 $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{G}$，那么

$$\begin{aligned}
\mathbb{P}\left[X \le r \wedge Y \in \bigcup_{n\ge 1} A_n\right] &= \mathbb{P}\left[\lim_{n\to\infty}\left(X \le r \wedge Y \in \bigcup_{i=1}^{n} A_i\right)\right] \\
&= \lim_{n\to\infty} \mathbb{P}\left[X \le r \wedge Y \in \bigcup_{i=1}^{n} A_i\right] \\
&= \lim_{n\to\infty} \mathbb{P}\left[X \le r\right] \cdot \mathbb{P}\left[Y \in \bigcup_{i=1}^{n} A_i\right] \\
&= \mathbb{P}\left[X \le r\right] \cdot \mathbb{P}\left[Y \in \bigcup_{n\ge 1} A_n\right].
\end{aligned}$$

我们再使用类似的方法，对每一个 $B \in \mathscr{B}$，证明集合 $\mathscr{G}' = \{A \subseteq \Omega : \mathbb{P}[X \in A \wedge Y \in B] = \mathbb{P}[X \in A] \cdot \mathbb{P}[Y \in B]\}$ 是个 σ-代数，便完成了整个证明。

第 9 章 一般随机变量的期望，勒贝格积分

我们之前定义了在离散概率空间上随机变量的期望，现在我们终于有了足够的工具来定义一般的随机变量的期望了。

9.1 一般测度空间上离散随机变量的期望

首先，我们来说明一下，我们之前对于离散概率空间上随机变量的期望的定义，可以直接的推广成一般的概率空间上的 **离散随机变量** 的期望，即使现在的概率空间不一定离散了。我们假设概率空间是 $(\Omega, \mathscr{F}, \mathbb{P})$，随机变量 $X : \Omega \to \mathbb{R}$ 的取值 $\mathsf{Im}(X)$ 是可数集。对于每一个 Ω 的分划 $\Omega = \bigsqcup_{i=1}^{\infty} \Lambda_i$，如果 X 在 Λ_i 上的取值是常数 z_i，并且级数 $\sum_{i=1}^{\infty} |z_i| \mathbb{P}[\Lambda_i] < \infty$，则称 X 是可积的，并定义其期望为

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} z_i \cdot \mathbb{P}[\Lambda_i].$$

我们可以同样的证明，只要 X 在每个 Λ_i 上是常数，定义出来的 $\mathbf{E}[X]$ 与实际选取的 Λ_i 无关。所以，我们可以对于每一个 $x \in \mathsf{Im}(X)$，选取 $\Lambda_i = [X = x]$，那么

$$\mathbf{E}[X] = \sum_{x \in \mathsf{Im}(X)} x \cdot \mathbb{P}[X = x].$$

对于离散随机变量，我们之前证明过的大部分性质，比如对于可积随机变量的期望线性性，独立随机变量的乘积的期望等于期望的乘积等，其证明均可以同样照搬至现在的场合。

9.2 从离散到一般可测函数

现在我们考虑一个一般的随机变量 X，也就是说，是从 Ω 到 $\mathbb{R}$ 的一个可测函数。现在我们没有办法再像离散的场合那样找到一个可数的分划，使得 X 限制在每个分划里是常数了。那我们定义期望的做法便是，正如同数学分析中常见的那样，使用离散的随机变量去逼近它。

对于每一个整数 $n \geq 0$，我们用 2^{-n} 的尺度来离散化实数轴。我们定义一系列随机变量，$\{\overline{X}_n\}_{n\geq 0}, \{\underline{X}_n\}_{n\geq 0}$，分别来表示在 2^{-n} 这个尺度下对于 X 的上逼近和下逼近。具体来说，对于每一个 $n \geq 0$，以及每一个 $\omega \in \Omega$，我们总可以找到一个整数 k，使得 $X(\omega) \in (k \cdot 2^{-n}, (k+1) \cdot 2^{-n}]$。于是，我们定义 $\overline{X}_n(\omega) := (k+1) \cdot 2^{-n}$，$\underline{X}_n(\omega) := k \cdot 2^{-n}$。换句话说，$\overline{X}_n(\omega)$ 和 $\underline{X}_n(\omega)$ 都是 $X(\omega)$ 精确到（二进制）小数点后 n 位的近似，所不同的是对于 n 位后面的部分，一个是向上取整，一个是向下取整。根据这个定义，我们可以马上得到一些性质：

命题 9.1

1. 所有的 $\overline{X}_n, \underline{X}_n$ 均为离散随机变量；
2. $\forall \omega \in \Omega, \underline{X}_n(\omega) \le X(\omega) \le \overline{X}_n(\omega)$；
3. $\overline{X}_n - \underline{X}_n = 2^{-n}$；
4. $\forall n \ge 0, \underline{X}_n \le \underline{X}_{n+1} \le \overline{X}_{n+1} \le \overline{X}_n$；
5. $\forall \omega \in \Omega, \lim_{n\to\infty} \underline{X}_n(\omega) = \lim_{n\to\infty} \overline{X}_n(\omega) = X(\omega)$。

♠

这些性质大部分是不言自明的，请大家自行验证。上面第 3 条性质告诉我们，在任意样本点 ω，$\underline{X}_n(\omega)$ 和 $\overline{X}_n(\omega)$ 相差最多 $2^{-n} \le 1$。因此，所有这些随机变量，是同时可积或者同时不可积的。并且，在它们可积的时候，我们有性质

$$-\infty < \mathbf{E}[\underline{X}_0] \le \mathbf{E}[\underline{X}_1] \cdots \le \mathbf{E}[\overline{X}_1] \le \mathbf{E}[\overline{X}_0] < \infty.$$

所以，我们可以引入如下 $\mathbf{E}[X]$ 的定义。

定义 9.1

我们说随机变量 X 是可积的，当且仅当 X_0 是可积的。如果 X 是可积的，定义其期望为

$$\mathbf{E}[X] := \lim_{n\to\infty} \mathbf{E}[\underline{X}_n] = \lim_{n\to\infty} \mathbf{E}[\overline{X}_n].$$

♣

我们上面的性质说明了，$\lim_{n\to\infty} \mathbf{E}[\underline{X}_n]$ 和 $\lim_{n\to\infty} \mathbf{E}[\overline{X}_n]$ 这两个极限一定是存在并且相等的。因此，这个定义是良定义。

我们有时候也把 $\mathbf{E}[X]$ 记成 $\int_\Omega X d\mathbb{P}$，或者 $\int_\Omega X(\omega)\mathbb{P}(d\omega)$，它被称为在 Ω 上可测函数 X 关于测度 $\mathbb{P}$ 的勒贝格积分（Lebesgue Integral）。

所以，期望就是积分，正如同随机变量就是可测函数一样，这也是为什么我们把存在有限期望称为“可积”的原因。于是，我们便能用分析的工具来研究概率论了，这也是 Kolmogorov 概率公理体系的高明之处。

事实上，我们对于积分（期望）的定义不限于概率测度，可以完全的推广到任何“有限”测度。对于无穷的测度，比如 $\mathbb{R}$ 上的勒贝格测度，也可以用类似的方法定义积分（一个值得注意的点是我们只能使用 $\underline{X}_n$ 从下方来逼近一个可积函数）。感兴趣的同学可以参考 wiki，或者任何一本讲测度或者实分析的教材。我们未来证明的关于期望的大部分性质，都可以无缝推广到无穷测度（事实上，我们需要测度是 σ-有限的）的勒贝格积分上去。对于只在有限测度上成立的性质，我（如果记得的话）会特别指出。

值得注意的是，在未来，如果 $\mathbb{P}$ 是勒贝格测度（定义在某个 $\mathscr{B}(\Omega)$ 上），我们常常用来 dx 表示 $\mathbb{P}(\mathrm{dx})$。

9.2.1 关于无穷的处理

有的时候，我们允许随机变量取无穷值，也就是说，是把 X 当成从 Ω 到 $\mathbb{R} \cup \{\pm\infty\}$ 的映射。这个时候，我们也允许期望取无穷值。在这个场合，我们按照如下方式扩展期望的定义，也把这个定义当成最一般的期望定义。

我们首先考虑非负的随机变量 X，即（包括）$\forall \omega \in \Omega, X(\omega) \ge 0$（包括 $X(\omega) = \infty$）。如果 $\mathbb{P}[X = \infty] > 0$，我们就定义 $\mathbf{E}[X] := \infty$，否则，我们定义一个新的随机变量 $\widehat{X}$，用来把 X 取值为 ∞ 的那些位置的值置为零：

$$\forall \omega \in \Omega, \quad \widehat{X}(\omega) := \begin{cases} 0, & \text{if } X(\omega) = \infty; \\ X(\omega), & \text{otherwise.} \end{cases}$$

由于 $\widehat{X}$ 也是非负随机变量，如果其可积，我们定义 $\mathbf{E}[X] := \mathbf{E}[\widehat{X}]$，如果其不可积（那么定义其期望的级数一定发散），我们定义 $\mathbf{E}[X] := \infty$。

我们现在引入一个方便的记号：我们用二元算符 $\wedge, \vee$ 分别来表示 min 和 max。即 $a \wedge b := \min\{a, b\}$，$a \vee b := \max\{a, b\}$。注意到，这个在组合数学里在 Lattice 上定义的类似算符的意义是一致的。

对于一般的不一定非负的随机变量 X，我们用 X^+ 和 X^- 分别表示其非负的部分和非正的部分，即对于任何 $\omega \in \Omega$，

$$X^+(\omega) = X(\omega) \vee 0, \quad X^-(\omega) = -X(\omega) \vee 0.$$

那么 $X^+, X^- \geq 0$ 并且 $X = X^+ - X^-$。如果 $\mathbf{E}[X^+] = \mathbf{E}[X^-] = \infty$，此时我们称 $\mathbf{E}[X]$ 无定义，否则，我们定义

$$\mathbf{E}[X] := \mathbf{E}[X^+] - \mathbf{E}[X^-].$$

这样，我们便完成了最一般的期望的定义。注意到，到现在为止，随机变量“可积”表示它的期望是有限的，而“不可积”有可能期望不存在，也有可能期望是无穷。对于一个非负的随机变量，它的期望一定存在，在它不可积的时候，期望是无穷。

9.3 期望（积分）的基本性质

我们说某个概率空间上的事件“几乎必然（almost surely）”发生，记作 a.s.，如果该事件发生的概率为 1。

这一节，我们列举期望的一些基本性质，同样，它们大部分的正确性是不言自明的，也可以使用定义直接验证。我们给出一些证明，并把剩下的留作练习。

1. 如果 $X = Y$ a.s., 那么 X 可积当且仅当 Y 可积。如果它们都可积的话，那么 $\mathbf{E}[X] = \mathbf{E}[Y]$。

   证明 显然，$X = Y$ a.s. 可以推出对于任何 $n \geq 0, \overline{X}_n = \overline{Y}_n$ a.s.。而我们有

   $$X \text{ 可积} \iff \overline{X}_0 \text{ 可积} \iff \overline{Y}_0 \text{ 可积} \iff Y \text{ 可积}.$$

   因此，在它们都可积的时候，有 $\mathbf{E}[X] = \lim \mathbf{E}[\overline{X}_n] = \lim \mathbf{E}[\overline{Y}_n] = \mathbf{E}[Y]$。
2. 如果 $|X| \leq |Y|$ a.s. 并且 Y 可积，那么 X 可积。特别的，X 可积当且仅当 $|X|$ 可积。
3. 如果 X 可积，那么对于任何 $a \in \mathbb{R}, aX$ 可积，并且 $\mathbf{E}[aX] = a\mathbf{E}[X]$。
4. 如果 X 和 Y 均可积，那么 $X + Y$ 也可积，并且 $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$。

证明 这一条就是所谓的期望的线性性，我们在之前的离散场合已经证明了，现在对于一般的随机变量我们马上对其再进行验证。注意，这个条件里面的 X 和 Y 均可积是非常重要的，否则该性质不一定成立。比如我们可以构造 $\mathbf{E}[X] = \infty, \mathbf{E}[Y] = -\infty$，但 $\mathbf{E}[X + Y]$ 是任何数。

首先验证 $X + Y$ 的可积性。我们令 $Z = X + Y$。只需要验证 $\overline{Z}_0$ 是可积的即可。显然有

$$|\overline{Z}_0| \leq |Z| + 1 \leq |X| + |Y| + 1 \leq |\overline{X}_0| + |\overline{Y}_0| + 3.$$

由于 X, Y 均是可积的，所以 $\overline{X}_0, \overline{Y}_0$ 均是可积的。因此 $\overline{Z}_0$ 也是可积的。

接着我们验证关于期望的等式。由于 $\mathbf{E}[Z] = \lim_{n\to\infty} \overline{Z}_n$, $\mathbf{E}[X] + \mathbf{E}[Y] = \lim_{n\to\infty}(\overline{X}_n + \overline{Y}_n)$, 对于每一个 $n \geq 0$，我们考察 $\overline{Z}_n - (\overline{X}_n + \overline{Y}_n)$。根据三角不等式，我们有

$$|\overline{Z}_n - (\overline{X}_n + \overline{Y}_n)| \leq |\overline{Z}_n - Z| + |Z - (X + Y)| + |X - \overline{X}_n| + |Y - \overline{Y}_n| \leq 3 \cdot 2^{-n}.$$

这也意味着，$\mathbf{E}[\overline{Z}_n]$ 和 $\mathbf{E}[\overline{X}_n + \overline{Y}_n]$ 有着相同的极限。

5. 如果 X 可积，那么 $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$。
6. 如果 X 和 Y 都可积，并且 $X \leq Y$ a.s.，那么 $\mathbf{E}[X] \leq \mathbf{E}[Y]$。
7. 如果 X 和 Y 独立，并且都可积，那么 XY 也可积，并且 $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$。

证明 这个性质我们也证明过其离散的版本。对于一般的情况的证明，我们要用到离散时候的结论。我们首先验证 XY 可积。对于任意 $n \geq 0$，根据三角不等式，我们有

$$|XY| \leq |\overline{X}_n\overline{Y}_n| + |XY - \overline{X}_n\overline{Y}_n|.$$

由于 $\overline{X}_n, \overline{Y}_n$ 均是离散随机变量，并且是可积的，我们在离散的时候已经证明过了 $\overline{X}_n\overline{Y}_n$ 是可积的。所以我们只需要验证 $|XY - \overline{X}_n\overline{Y}_n|$ 是可积的即可。To this end，我们有

$$\begin{aligned}\left|XY-\overline{X}_n\overline{Y}_n\right| &= \left|XY-\overline{X}_nY+\overline{X}_nY-\overline{X}_n\overline{Y}_n\right| \\ &\le |Y|\left|X-\overline{X}_n\right|+\left|\overline{X}_n\right|\left|Y-\overline{Y}_n\right| \\ &\le 2^{-n}\left(|Y|+\left|\overline{X}_n\right|\right).\end{aligned}$$

由于 $|Y|$ 和 $\overline{X}_n$ 均是可积的随机变量，使用上面的性质 2 和 4，我们知道 $|XY-\overline{X}_n\overline{Y}_n|$ 也是可积的。因此 XY 是可积的。

接下来验证 $\mathbf{E}[XY]=\mathbf{E}[X]\mathbf{E}[Y]$。我们使用刚才验证的对应变量的可积性，期望的线性性以及离散时候独立随机变量乘法和期望的可交换性，可以得到

$$\begin{aligned}\mathbf{E}[XY] &= \mathbf{E}\left[\overline{X}_n\overline{Y}_n+(XY-\overline{X}_n\overline{Y}_n)\right] \\ &= \mathbf{E}\left[\overline{X}_n\overline{Y}_n\right]+\mathbf{E}\left[XY-\overline{X}_n\overline{Y}_n\right] \\ &= \mathbf{E}\left[\overline{X}_n\right]\mathbf{E}\left[\overline{Y}_n\right]+\mathbf{E}\left[XY-\overline{X}_n\overline{Y}_n\right].\end{aligned}$$

所以，由性质 5 以及我们刚才得到的估计，

$$\begin{aligned}\left|\mathbf{E}[XY]-\mathbf{E}[X]\mathbf{E}[Y]\right| &= \left|\lim_{n\to\infty}\left(\mathbf{E}[XY]-\mathbf{E}\left[\overline{X}_n\right]\mathbf{E}\left[\overline{Y}_n\right]\right)\right| \\ &= \lim_{n\to\infty}\left|\mathbf{E}\left[XY-\overline{X}_n\overline{Y}_n\right]\right| \\ &\le \lim_{n\to\infty}\mathbf{E}\left[\left|XY-\overline{X}_n\overline{Y}_n\right|\right] \\ &\le \lim_{n\to\infty}2^{-n}\left(\mathbf{E}[|Y|]+\mathbf{E}\left[\left|\overline{X}_n\right|\right]\right) \\ &= 0.\end{aligned}$$

以上最后一个等号是由于 $|Y|$ 和 $\overline{X}_n$ 均是可积的。

9.3.1 Remark: 期望、求和和求极限的交换性质

我们这里引入作业题的例子，以说明其不可直接交换性：

(1) 给定概率空间 $(\Omega,\mathcal{F},P)$ 和一列随机变量 $X_1,X_2,\cdots:\Omega\to\mathbb{R}$。如果每个 X_n 都可积 ($\mathbb{E}[|X_n|]<\infty$)，那么

$$\mathbb{E}\left[\sum_{n=1}^{\infty}X_n\right]=\sum_{n=1}^{\infty}\mathbb{E}[X_n].$$

该结论错误。
给定概率空间 $([0,1],\mathcal{B}([0,1]),\mathbb{P})$，其中 $\mathbb{P}$ 是勒贝格测度。对于任意 $n\in\mathbb{N}$，定义 $Y_n=n\cdot\mathbb{I}_{[0,1/n]}$。对于任意 $n\in\mathbb{N}^+$ 定义 $X_n=Y_n-Y_{n-1}$。
显然我们有 $\mathbb{E}[|X_n|]\le 2<\infty$。注意到 $\sum_{n=1}^{\infty}\mathbb{E}[X_n]=\lim_{n\to\infty}\mathbb{E}[Y_n]=1$ 并且 $\mathbb{E}[\sum_{n=1}^{\infty}X_n]=\mathbb{E}[\lim_{n\to\infty}Y_n]=0$。因此，$\mathbb{E}[\sum_{n=1}^{\infty}X_n]\ne\sum_{n=1}^{\infty}\mathbb{E}[X_n]$。

(2) 假设随机变量 $X,X_1,X_2,\ldots$ 满足对于任意 n，$X_n\le X_{n+1}$ 并且 $\lim_{n\to\infty}X_n=X$ a.s. 那么

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

该结论错误。
给定概率空间 $((0,1),\mathcal{B},\mathbb{P})$，其中 $\mathbb{P}$ 是勒贝格测度。定义随机变量 X_n 使得对于任何 $\omega\in(0,1)$，$X_n(\omega)=-\frac{1}{n\omega^2}$。
那么我们有 $X_n\le X_{n+1}$，$\lim_{n\to\infty}X_n=X$ a.s. 并且 $X=0$。
注意到 $\mathbb{E}[X]=0$，而对于任何 n，$\mathbb{E}[X_n]=-\infty$。因此 $\lim_{n\to\infty}\mathbb{E}[X_n]\ne\mathbb{E}[X]$。

(3) 假设随机变量 $X,X_1,X_2,\ldots$ 满足对于任意 n，$X_n\le X_{n+1}$ 并且 $\lim_{n\to\infty}X_n=X$ a.s. 如果对于任意 n，X_n

都可积，那么

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

该结论正确。证明如下：
定义随机变量 $Y_1,Y_2,\ldots$，使得 $Y_n=X_n-X_1$。对于任何 n，Y_n 是非负可积的，并且 $\lim_{n\to\infty}Y_n=X-X_1$。对 $Y_1,Y_2,\ldots$ 使用单调收敛定理，我们得到，

$$\lim_{n\to\infty}\mathbb{E}[Y_n]=\mathbb{E}[X-X_1].$$

因为每个 X_n 都是可积的，所以

$$\lim_{n\to\infty}\mathbb{E}[Y_n]=\lim_{n\to\infty}(\mathbb{E}[X_n]-\mathbb{E}[X_1])=\lim_{n\to\infty}\mathbb{E}[X_n]-\mathbb{E}[X_1].$$

接下来我们只需要证明 $\mathbb{E}[X-X_1]=\mathbb{E}[X]-\mathbb{E}[X_1]$。注意到 $X_1\le X$ a.s.。因为 X_1 可积，我们有 $\mathbb{E}[X^-]\le\mathbb{E}[X_1^-]<\infty$。因此要么 X 可积，要么 $\mathbb{E}[X]=\infty$。当 X 可积时，由期望线性性可知 $\mathbb{E}[X-X_1]=\mathbb{E}[X]-\mathbb{E}[X_1]$。否则我们有 $\mathbb{E}[X-X_1]=\mathbb{E}[X]-\mathbb{E}[X_1]=\infty$。
因此我们可以证明 $\lim_{n\to\infty}\mathbb{E}[X_n]=\mathbb{E}[X]$。

(4) 对于一列非负随机变量 $X_1,X_2,\ldots$，我们有

$$\mathbb{E}\left[\limsup_{n\to\infty}X_n\right]\ge\limsup_{n\to\infty}\mathbb{E}[X_n].$$

该结论错误。
给定概率空间 $([0,1],\mathcal{B}([0,1]),\mathbb{P})$，其中 $\mathbb{P}$ 是勒贝格测度。对于任意 $n\in\mathbb{N}$，定义 $X_n=n\cdot\mathbb{I}_{[0,1/n]}$。显然我们有对于任何 n，$\mathbb{E}[X_n]=1$，并且 $\mathbb{E}[\lim_{n\to\infty}X_n]=0$。因此，$\mathbb{E}[\limsup_{n\to\infty}X_n]<\limsup_{n\to\infty}\mathbb{E}[X_n]$。

第 10 章 MCT，Fatou Lemma 与 DCT

我们今天来学习关于期望，或者更一般的勒贝格积分的几个关于求极限与求积分进行交换的结论。这几个结论在我们未来的学习中会扮演非常重要的角色。

10.1 逐点收敛与几乎处处收敛

我们在数学分析中会遇到函数列收敛的概念，也就是一列函数 $\{f_n\}_{n\geq 1}$ 收敛到一个函数 f。我们最常见的收敛形式是逐点收敛，也就是说对于定义域里面的每一个点 x，数列 $f_1(x), f_2(x), \dots$ 收敛到 $f(x)$。

同样，我们在概率论中会讨论一列随机变量 $\{X_n\}_{n\geq 1}$ 收敛到 X。如果所有的这些随机变量均生活在同一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 中，那么，逐点收敛可以自然的定义为：

$$\forall \omega \in \Omega, \quad \lim_{n\to\infty} X_n(\omega) = X(\omega).$$

事实上，由于概率测度的存在，我们对于收敛会有很多种不同的定义。在未来，我们会专门讨论各种收敛的概念之间的联系。

今天，我们先引入所谓的“几乎必然（almost surely）”收敛，也叫做“以概率 1 收敛”或者“几乎处处收敛（almost everywhere，简称 a.e.）”。它的定义是

$$\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1.$$

换句话说，那些使得 $\lim_{n\to\infty} X_n \neq X$ 的样本集的测度是 0。在概率论里，这是一个很强的收敛准则，但它比逐点收敛（$\forall \omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)$）要弱。在概率论的很多结论中，一个测度为零的集合往往不能掀起什么风浪，所以我们经常可以把逐点收敛的条件弱化成几乎处处收敛。

10.2 期望与极限的交换

假设 X_n 逐点收敛到 X。我们想问，$\mathbf{E}[X_n]$ 会不会收敛到 $\mathbf{E}[X]$ 呢？换句话说，就是期望和极限是否能够交换？

$$\lim_{n\to\infty} \mathbf{E}\left[X_n\right] \stackrel{?}{=} \mathbf{E}\left[\lim_{n\to\infty} X_n\right].$$

首先，我们必须明白，期望和极限不一定总是能够交换。

我们假设概率空间是 $(0,1]$ 上的均匀分布。设 $X_n = n \cdot \mathbb{I}_{(0,1/n)}$。那么，容易验证 X_n 逐点收敛到 0。另一方面，对于每一个 $n \geq 1$，我们有 $\mathbf{E}[X_n] = 1$。所以，

$$\lim_{n\to\infty} \mathbf{E}[X_n] = 1 \neq 0 = \mathbf{E}\left[\lim_{n\to\infty} X_n\right].$$

另一个更实际的例子是下图中的赌博游戏。

考虑被称之为 martingale 的赌博策略来参加比大小的游戏。第一轮我押 1 元，如果赢了就跑路，输了就进入第二轮。第二轮我押 2 元，赢了跑路，输了进入第三轮。第三轮押 4 元，赢了跑路，输了继续，以此类推。总之策略就是赢了就结束游戏，输了就再赌一次，但是赌注翻倍。

我用 X_i 表示第 i 论我赢的钱。假设这个比大小是公平游戏，所以对于任意 $i \geq 1$，$\mathbf{E}[X_i] = 0$。也就是说 $\sum_{i\geq 1} \mathbf{E}[X_i] = 0$。

然而，这个游戏我会以 1 的概率赢钱离场。一旦我赢了，假设是在第 m 轮赢的，我的总收入是 $-\sum_{i=1}^{m-1} 2^{i-1} + 2^{m-1} = 1$，也就是说，我一定会赢 1 元钱。所以 $\mathbf{E}\left[\sum_{i\geq 1} X_i\right] = 1$。

图 10.1: Figure 1

我们接下来会介绍期望与极限交换的三个重要结论，即单调收敛定理（简称 MCT）、Fatou 引理（简称 Fatou）和控制收敛定理（简称 DCT）。这三个结论是可以互推的，我们采取 MCT $\Longrightarrow$ Fatou $\Longrightarrow$ DCT 的顺序介绍。因此，大家可以看到，在我们的处理中，真正打开定义的黑盒进行实质性说明的是第一个对于 MCT 的证明。

10.3 单调收敛定理（Monotone Convergence Theorem）

定理 10.1 (单调收敛定理)

设 $\{X_n\}_{n\geq 1}$ 是一组非负随机变量，满足 $0 \leq X_1 \leq X_2 \leq \cdots$，并且 $\lim_{n\to\infty} X_n = X$ a.s.，那么

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

♡

值得注意的是，这个定理里我们没有要求随机变量可积，因此，即使在 $\mathbf{E}[X] = \infty$ 的时候也是成立的。

为了证明这个定理，我们可以直观上来考察一下这个结论为什么对。X_n 是单调递增的，它越来越接近 X，我们想说明 $\mathbf{E}[X_n]$ 也会越来越接近 $\mathbf{E}[X]$。直观上来说，对于每一个 ε，我们考察 X_n 和 X 差距比较大（比如大于 ε）的那些样本点的集合。随着 n 越来越大，这些样本点的集合的测度会越来越小。我们希望这些样本点对于期望的差距的贡献也是越来越小的。如果我们知道函数的取值本身是有界的，那么这个问题就会很显然。这促使我们先来证明下面这个引理，它实际上是 MCT 的一个特殊情况（$X_n = X \wedge n$），来把一般 MCT 的证明转变为有界函数的场合。

引理 10.1

设 $X \geq 0$。那么 $\lim_{n\to\infty} \mathbf{E}[X \wedge n] = \mathbf{E}[X]$。

♡

证明

如果 $\mathbb{P}[X = \infty] > 0$，那么对于每一个 $\omega \in [X = \infty]$，我们都有 $(X \wedge n)(\omega) = n$。因此，我们有 $\mathbf{E}[X \wedge n] \geq n \cdot \mathbb{P}[X = \infty]$。这说明 $\lim_{n\to\infty} \mathbf{E}[X \wedge n] = \infty$。

因此，我们可以假设 $\mathbb{P}[X < \infty] = 1$。实际上，我们可以进一步的假设 X 是逐点有限的，因为测度为零的部分不会改变期望。由于对于任意 n，我们均有 $X \wedge n \leq X$，所以 $\lim_{n\to\infty} \mathbf{E}[X \wedge n] \leq \mathbf{E}[X]$。所以我们只需证明 $\mathbf{E}[X] \leq \lim_{n\to\infty} \mathbf{E}[X \wedge n]$。事实上，我们有对于任何 n 和 k，

$$\mathbf{E}[X \wedge n] \geq \mathbf{E}[\underline{X}_k \wedge n] \geq \mathbf{E}\left[\underline{X}_k \cdot \mathbb{I}_{\underline{X}_k \leq n}\right] = \sum_{j=0}^{n\cdot 2^k} j \cdot 2^{-k} \cdot \mathbb{P}\left[\underline{X}_k = j \cdot 2^{-k}\right].$$

我们让上式最左边和最右边的 n 同时趋向无穷大，便能得到

$$\lim_{n\to\infty}\mathbf{E}[X\wedge n]\geq\sum_{j=0}^{\infty}j\cdot 2^{-k}\cdot\mathbb{P}\left[\underline{X}_k=j\cdot 2^{-k}\right]=\mathbf{E}[\underline{X}_k].$$

我们再令 $k\to\infty$，便得证。

有了这个引理，我们来证明 MCT。

证明

首先我们简单说明一下我们可以不管题设里的 a.e.，而假设所有性质都是逐点成立的。我们把那些让某个性质不成立的样本点拿出来，记作 Λ。由于 $\mathbb{P}(\Lambda)=0$，我们可以把一个随机变量 Y 都换成 $Y\cdot\mathbb{I}_{\overline{\Lambda}}$。这样所有的性质都是逐点成立了，并且，$\mathbf{E}[Y]=\mathbf{E}[Y\cdot\mathbb{I}_{\overline{\Lambda}}]$.

同样，因为我们知道 $X_n\leq X$，所以 $\lim_{n\to\infty}\mathbf{E}[X_n]\leq\mathbf{E}[X]$。我们只需要证明 $\mathbf{E}[X]\leq\lim_{n\to\infty}\mathbf{E}[X_n]$。根据刚才的引理，我们只需要证明 $\lim_{N\to\infty}\mathbf{E}[X\wedge N]\leq\lim_{n\to\infty}\mathbf{E}[X_n]$。我们将证明，对于任意 $N\geq 0$，$\mathbf{E}[X\wedge N]\leq\lim_{n\to\infty}\mathbf{E}[X_n]$。由于显然 $\mathbf{E}[X_n]\geq\mathbf{E}[X_n\wedge N]$，我们只需要证明对于任意 $N\in\mathbb{N}$，

$$\mathbf{E}[X\wedge N]\leq\lim_{n\to\infty}\mathbf{E}[X_n\wedge N].$$

上面的讨论说明，我们可以不失一般性的假设我们关心的随机变量是有界，即只需证明有上界 N 的随机变量 $0\leq X_1\leq X_2\leq\cdots\leq N$，满足 $\lim_{n\to\infty}\mathbf{E}[X_n]=\mathbf{E}[X]$，就可以证明原问题。

对于有界的随机变量，这个问题变得容易很多。我们将说明，对于任意 $\varepsilon>0$，在 n 足够大的时候，X_n 和 X 差距大于 ε 的那些样本集的测度将任意小，而随机变量在这些样本集上的取值又有上界，因此，他们对于期望的贡献也任意小。

所以我们将说明，对于任意 $\varepsilon>0$，$\lim_{n\to\infty}\mathbf{E}[X_n]\geq\mathbf{E}[X]-\varepsilon$，这等价于 $\lim_{n\to\infty}\mathbf{E}[X_n]\geq\mathbf{E}[X]$。这个技巧叫做“an epsilon of room”。

对于每一个 $n\in\mathbb{N}$ 和 $\varepsilon>0$，我们定义 $A_{n,\varepsilon}=\{\omega:X_n(\omega)<X(\omega)-\varepsilon\}$，也就是 X_n 和 X 差距大于 ε 的那些集合。由于 X_n 关于 n 是非降的，我们有 $A_{n,\varepsilon}$ 是非增的，并且 $\bigcap_{n\geq 1}A_{n,\varepsilon}=\varnothing$。所以

$$\lim_{n\to\infty}\mathbf{E}\left[X-X_n\right]\leq\lim_{n\to\infty}\left(\varepsilon\cdot 1+N\mathbb{P}\left[\bigcap_{i=1}^{n}A_{i,\varepsilon}\right]\right)=\varepsilon.$$

这足够说明我们想证明的结论了（why ?）。

推论 10.1 (MCT 的推论 1)

我们可以考虑非增的随机变量：假设 $X_1\geq X_2\geq\cdots\geq 0$，并且 $\lim_{n\to\infty}X_n=X$ a.e.。如果 X_1 是可积的，那么

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

♡

这个结论的证明也很简单，我们只要令 $Y_n=X_1-X_n$，由于涉及的每一个随机变量的都是可积的，再使用 MCT 即可。条件里的可积性是必要的，不然的话，考虑定义在 $(0,1]$ 均匀测度上的随机变量 $X_n(x)=\frac{1}{nx}$。

另外一个推论是期望的线性性在涉及无穷项的时候，如果每一项都是非负的，那依然成立。也就是说：

推论 10.2 (MCT 的推论 2)

如果 $Y_1,Y_2,\cdots\geq 0$，那么

$$\mathbf{E}\left[\sum_{i=1}^{\infty}Y_i\right]=\sum_{i=1}^{\infty}\mathbf{E}[Y_i].$$

♡

为了说明这个，我们令 $X_n=\sum_{i=1}^{n}Y_i$，并使用 MCT，可以得到

$$\mathbf{E}\left[\sum_{i=1}^{\infty}Y_i\right]=\mathbf{E}\left[\lim_{n\to\infty}X_n\right]\overset{\text{(MCT)}}{=}\lim_{n\to\infty}\mathbf{E}\left[X_n\right]\overset{(\heartsuit)}{=}\lim_{n\to\infty}\sum_{i=1}^{n}\mathbf{E}\left[Y_i\right]=\sum_{i=1}^{\infty}\mathbf{E}\left[Y_i\right].$$

注意到，我们在 (♡) 用到了有限和时候的期望线性性。这个在每一个 Y_i 是可积的时候我们已经证明过了。由于我们这儿 Y_i 都是非负的，即使其中某一个不可积，期望的线性性也成立，因为两边都等于无穷大。

10.4 Fatou 引理（Fatou's Lemma）

我们一开始说的期望和极限交换的反例说明，对于非负的随机变量，取其极限可能让期望变小。下面这个结论确认这个事实。对于非负随机变量，如果直观的把期望想象成围成的面积的话（我们将在下次课 justify 这件事!），极限过程只可能破坏这些面积，而不可能增加面积。

定理 10.2 (Fatou 引理)

对于一列非负随机变量 X_n，我们有

$$\mathbf{E}\left[\liminf_{n\to\infty} X_n\right] \le \liminf_{n\to\infty} \mathbf{E}[X_n].$$

♡

我们注意到，在这儿，我们没有任何收敛性的要求。因此，我们只能谈论 $\liminf$，而不是 $\lim$。

根据定义，对于一列数 x_n，

$$\liminf_{n\to\infty} x_n := \lim_{n\to\infty} \inf_{j\ge n} x_j.$$

因此，如果我们定义 $y_n := \inf_{j\ge n} x_j$，则 $y_n \uparrow \liminf_{n\to\infty} x_n$。于是乎，对于随机变量列 X_n，我们也有

$$\inf_{j\ge n} X_j \uparrow \liminf_{n\to\infty} X_n.$$

由于我们关心的随机变量都是非负的，我们便可以使用 MCT 得到

$$\lim_{n\to\infty} \mathbf{E}\left[\inf_{j\ge n} X_j\right] = \mathbf{E}\left[\liminf_{n\to\infty} X_n\right].$$

另一方面，我们有 $X_n \ge \inf_{j\ge n} X_j$；也就是说

$$\mathbf{E}[X_n] \ge \mathbf{E}\left[\inf_{j\ge n} X_j\right].$$

两边取 $\liminf$，可以得到

$$\liminf_{n\to\infty} \mathbf{E}[X_n] \ge \liminf_{n\to\infty} \mathbf{E}\left[\inf_{j\ge n} X_j\right] \overset{(\spadesuit)}{=} \lim_{n\to\infty} \mathbf{E}\left[\inf_{j\ge n} X_j\right] = \mathbf{E}\left[\liminf_{n\to\infty} X_n\right],$$

其中 (♠) 是由于 $\mathbf{E}[\inf_{j\ge n} X_n]$ 是关于 n 单调的。

注意到，如果把 Fatou 引理里面的 inf 换成 sup 是不对的。考虑一个从 $(0,1)$ 中均匀选出来的实数的二进制表示。我们用 X_n 表示它的第 n 位数字。那么容易验证 $\mathbf{E}[X_n] = 0.5$，但是 $\limsup X_n = 1, \liminf X_n = 0$。

10.5 控制收敛定理（Dominated Convergence Theorem）

我们接下来证明控制收敛定理（DCT），这也是实际上我们最常用的一个结论。

定理 10.3 (控制收敛定理)

设 X_n 为一列随机变量，满足 $\lim_{n\to\infty} X_n = X$ a.e.。如果存在一个随机变量 Y，满足

1. 对所有 $n \in \mathbb{N}, |X_n| \le Y$；
2. Y 是可积的。

那么 $\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X]$。

♡

需要注意的是，我们这儿没有要求 X_n 是非负的。

可以回忆我们一开始提到的反例 $X_n = n \cdot \mathbb{I}_{(0,1/n)}$，我们可以直观的解释它是反例的原因：在取极限的过程中，所围成的面积从上方溜掉了。DCT 便说明，如果能够给所有的 X_n 找一个统一的上界 Y，把它们都给罩住不让跑出去，这种情况便不会发生。

我们来证明 DCT。

证明

考虑非负的随机变量 $Y - X_n$，根据 Fatou 引理，我们有

$$\liminf_{n\to\infty} \mathbf{E}[Y - X_n] \geq \mathbf{E}\left[\liminf_{n\to\infty}(Y - X_n)\right] = \mathbf{E}\left[Y - \limsup_{n\to\infty} X_n\right] = \mathbf{E}[Y - X] = \mathbf{E}[Y] - \mathbf{E}[X].$$

因此，

$$\mathbf{E}[Y] - \mathbf{E}[X] \leq \liminf_{n\to\infty} \mathbf{E}[Y - X_n] = \mathbf{E}[Y] - \limsup_{n\to\infty} \mathbf{E}[X_n].$$

由于 $\mathbf{E}[Y] < \infty$，这等价于

$$\limsup_{n\to\infty} \mathbf{E}[X_n] \leq \mathbf{E}[X].$$

同样的，我们考察随机变量 $Y + X_n$. 根据 Fatou 引理，我们有

$$\liminf_{n\to\infty} \mathbf{E}[Y + X_n] \geq \mathbf{E}\left[\liminf_{n\to\infty}(Y + X_n)\right] = \mathbf{E}[Y + X] = \mathbf{E}[Y] + \mathbf{E}[X].$$

因此，

$$\mathbf{E}[Y] + \mathbf{E}[X] \leq \liminf_{n\to\infty} \mathbf{E}[Y + X_n] = \mathbf{E}[Y] + \liminf_{n\to\infty} \mathbf{E}[X_n].$$

和上面得到的式子放在一起，我们便知道了

$$\limsup_{n\to\infty} \mathbf{E}[X_n] \leq \mathbf{E}[X] \leq \liminf_{n\to\infty} \mathbf{E}[X_n].$$

所以 $\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X]$ 得证。

控制收敛定理的一个显然的推论，有时候又叫 **有界收敛定理**，便是当所有的 X_n 都有一个一致的上界 $|X_n| \leq M$ 时，$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[\lim_{n\to\infty} X_n]$ 成立。这只需要在我们的 DCT 里取 $Y = M$ 就可以了。注意到，有界收敛定理对于无穷测度不成立，原因是有界函数在无穷测度下并不一定是可积的。

第 11 章 积分的换元，期望与分布，概率密度函数

如果概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 是离散的，我们知道，在上面定义的随机变量 X 之期望可以写成下面两种形式，分别对应了对于 X 的不同分划方式：

1. $\mathbf{E}[X] = \sum_{\omega\in\Omega} X(\omega)\cdot\mathbb{P}[\{\omega\}]$
2. $\mathbf{E}[X] = \sum_{x\in \mathsf{Im}(X)} x\cdot\mathbb{P}[X=x]$

直观上，这两种期望的表示方法可以分别看成对 X 的定义域和值域进行加权求和。其中第二种方式通过枚举随机变量取值的方法在有些问题中会特别方便我们计算。我们今天想说明，对于一般的随机变量，我们也能这么做，唯一需要做的是把求和换成合适测度的（勒贝格）积分。

11.1 积分的换元，期望与分布

我们首先引入一个记号。对于测度空间 $(\Omega, \mathscr{F}, \mathbb{P})$，随机变量 X，以及 $A\in\mathscr{F}$，我们定义在 A 上的 X 积分

$$\int_A X d\mathbb{P} := \int_\Omega X\cdot\mathbb{I}_{X\in A} d\mathbb{P}.$$

我们假设 X 是 $(\Omega, \mathscr{F}, \mathbb{P})$ 上定义的一个一般的随机变量。根据我们对于期望的定义，我们可以记作

$$\mathbf{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega).$$

这便是离散空间上随机变量上面第一种写法的推广。我们将要证明，这个期望等于

$$\int_{\mathbb{R}} x\mu(\mathrm{dx}),$$

其中 μ 是我们介绍过的 X 诱导出来的分布，满足对于任何 $A\in\mathscr{B}(\mathbb{R})$，$\mu(A) = \mathbb{P}[X\in A]$。注意到这是一个在测度空间 $(\mathbb{R}, \mathscr{B}, \mu)$ 上定义的勒贝格积分。因此，如果我们直观的把 dx 想象成 $\mathbb{R}$ 中的一小段区间的话，$\mu(\mathrm{dx}) = \mathbb{P}[X\in \mathrm{dx}]$，这在形式上便是离散随机变量的期望第二种写法的推广。

事实上，我们将证明一个更强的定理。我们称定义在 $(\mathbb{R}, \mathscr{B})$ 上的可测函数 $f:\mathbb{R}\to\mathbb{R}$ 为 Borel 函数。那么，对于一个 Borel 函数 f 和随机变量 X，容易验证复合函数 $f(X):\omega\in\Omega\mapsto f(X(\omega))\in\mathbb{R}$ 是可测的，因此也是一个随机变量。我们有下面这个一般版本的 LOTUS(Law of the unconscious statistician)：

定理 11.1

设 X 是 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的随机变量，g 是一个非负 Borel 函数。那么

$$\mathbf{E}[g(X)] = \int_{\mathbb{R}} g(x)\mu(\mathrm{dx}).$$

♡

注意到，因为按照我们上一节课提及之处理无穷随机变量期望的方法，我们只需要考虑非负的 g 就够了。

证明 对于每一个 $n \geq 0$，我们考虑 g 的上近似 $\overline{g}_n(x)$。对于每一个 $k \in \mathbb{N}$，定义 $\Lambda_k = \{x \in \mathbb{R} : \overline{g}_n(x) = k \cdot 2^{-n}\}$。设 $A_k = [X \in \Lambda_k]$。那么，根据 μ 的定义，我们有 $\mathbb{P}(A_k) = \mu(\Lambda_k)$。于是，

$$\mathbf{E}[\overline{g}_n(X)] = \sum_{k=0}^{\infty} k \cdot 2^{-n} \mathbb{P}[A_k] = \sum_{k=0}^{\infty} k \cdot 2^{-n} \mu(\Lambda_k) = \mathbf{E}_{\mu}[\overline{g}_n].$$

令 $n \to \infty$ 即得证。

11.2 随机变量的分类

我们知道每一个随机变量 X 均有一个分布函数 $F_X : \mathbb{R} \to [0, 1]$，满足

$$\forall t, F(t) = \mathbb{P}[X \leq t].$$

我们前面说过，如果一个随机变量的所有可能取值是一个可数集，则它被称为 **离散随机变量**。我们将称一个随机变量为 **连续随机变量**，如果其分布函数是一个连续函数。当然了，一个随机变量可以既不连续也不离散。

我们说一个随机变量的分布函数 F 是绝对连续（absolutely continuous）的，如果存在一个定义在 $\mathbb{R}$ 上的函数 f，满足

$$\forall t, \quad F(t) = \int_{(-\infty, t]} f(x) \mathrm{dx}.$$

这里的积分是勒贝格积分，dx 是勒贝格测度。我们称 f 为 X 的概率密度函数（probability density function, pdf），有时候又简称为密度函数。显然，我们要求 f 非负并且 $\int_{\mathbb{R}} f(x) \mathrm{dx} = 1$。比如说，$X$ 在 $(0, 1]$ 上的均匀分布，它的密度函数为

$$f(x) = \begin{cases} 0, & x < 0; \\ 1, & 0 \leq x \leq 1; \\ 0, & x > 1. \end{cases}$$

一个绝对连续的 F 一定是连续的，但不一定是可导的（比如说 $(0, 1]$ 上的均匀分布，它有 0 和 1 两个不可导的点）。可以证明，它一定是几乎处处可导的。这个帖子里给出了一些单调的，连续的，但是不绝对连续的例子。如果 F 在 $x = t$ 可导，那么 $f(t) = F'(t)$。

绝对连续的概念在概率论里面很重要的一个原因是，我们关心的很多连续分布，是直接通过其概率密度函数定义的。比如说，我们未来会仔细研究的标准正态分布，正是满足 $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ 的分布。我们也可以通过概率密度函数计算随机变量的很多数字特征，如期望、方差等。

当然，别忘了，并不是每一个连续分布均有概率密度函数的。

11.3 概率密度的积分

我们接下来要介绍一个对于计算期望非常重要的结论。再次回忆对于离散随机变量我们有

$$\mathbf{E}[X] = \sum_{x \in \mathsf{Im}(X)} x \cdot \mathbb{P}[X = x].$$

我们之前讲离散随机变量的时候定义过所谓概率质量函数 pmf，即 $p_X(x) = \mathbb{P}[X = x]$。所以上式也可以写成

$$\mathbf{E}[X] = \sum_{x \in \mathsf{Im}(X)} x \cdot p_X(x).$$

这个式子的 LOTUS 版本是对于任何 Borel 函数 g：

$$\mathbf{E}[g(X)] = \sum_{x \in \mathsf{Im}(X)} g(x) \cdot p_X(x).$$

这个式子是我们计算离散随机变量以及关于其函数的期望最常用的公式。根据我们上面的定义可以看出，概率密度函数其实就是“连续版本”的概率质量函数，因此，我们也期望有类似的式子来进行计算。我们现在就给出这个公式。

定理 11.2

对于有概率密度 $f(x)$ 的随机变量 X 以及非负 Borel 函数 g，我们有

$$\mathbf{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)\mathrm{dx}.$$

♡

同样，这里的积分是勒贝格积分，dx 是勒贝格测度。

证明

由于我们知道 $\mathbf{E}[g(X)] = \int_{\mathbb{R}} g(x)\mu(\mathrm{dx})$，我们只需要证明

$$\int_{\mathbb{R}} g(x)\mu(\mathrm{dx}) = \int_{\mathbb{R}} g(x)f(x)\mathrm{dx}$$

即可。

我们分两步来证明上述使用概率密度函数的积分公式。

1. 第一步 我们首先定义一个 $\mathbb{R}$ 上的集合函数 ν：

$$\forall A \in \mathscr{B}, \quad \nu(A) := \int_A f(x)\mathrm{dx}.$$

我们来验证，对于每一个 $A \in \mathscr{B}$，$\mu(A) = \nu(A)$。To this end，定义

$$\mathscr{G} := \{A \subseteq \Omega : \nu(A) = \mu(A)\}.$$

我们将用单调类定理证明 $\mathscr{G}$ 包含了 $\mathscr{B}$。首先容易验证 $\mathscr{G}$ 包含了所有的形如 $(a,b]$ 的区间。这是由于

$$\nu((a,b]) = \int_{(a,b]} f(x)\mathrm{dx} = F(b) - F(a) = \mu((a,b]).$$

根据勒贝格积分的定义，容易验证，所有有限个左开右闭的区间的并也是属于 $\mathscr{G}$。因此，代数 $\mathscr{B}_0 \subseteq \mathscr{G}$。我们现在验证对于单调上升的 $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{G}$，$\bigcup_{n\geq 1} A_n \in \mathscr{G}$。由于 μ 是一个测度，所以我们有

$$\mu\left(\bigcup_{i\geq 1} A_i\right) = \mu\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \mu\left(\bigcup_{i=1}^{n} A_i\right).$$

由于对于每一个 n，$\bigcup_{i=1}^{n} A_i = A_n \in \mathscr{G}$，我们有

$$\mu\left(\bigcup_{i\geq 1} A_n\right) = \lim_{n\to\infty} \nu\left(\bigcup_{i=1}^{n} A_i\right) \overset{(\clubsuit)}{=} \nu\left(\bigcup_{i\geq 1} A_i\right).$$

其中 $(\clubsuit)$ 是使用了 MCT（$\nu\left(\bigcup_{i=1}^{n} A_i\right) = \int_{\mathbb{R}} \mathbb{I}_{\bigcup_{i=1}^{n} A_i} f(x)\mathrm{dx}$，而 $\mathbb{I}_{\bigcup_{i=1}^{n} A_i} f(X)$ 是单调递增的随机变量）。对于单调递减的事件列，其封闭性亦可类似证明。于是，由单调类定理，$\mathscr{B} \subseteq \mathscr{G}$。

2. 第二步 接着，我们使用上一步的结论证明对于任何 $A \in \mathscr{B}$，有

$$\int_A g(x)\mu(\mathrm{dx}) = \int_A g(x)f(x)\mathrm{dx}.$$

对于每一个 n，我们来考察 g 的下近似 $\underline{g}_n$ 的期望。对于每一个 $k \in \mathbb{N}$，我们定义 $A_k := \{x \in \mathbb{R} : \underline{g}_n(x) = k \cdot 2^{-n}\}$。于是，$\underline{g}_n$ 可以写成 $\underline{g}_n(x) = \sum_{k\in\mathbb{N}} k \cdot 2^{-n}\mathbb{I}_{x\in A_k}$。那么

$$\int_A \underline{g}_n(x)\mu(\mathrm{dx}) = \int_A \sum_{k\in\mathbb{N}} k \cdot 2^{-n}\mathbb{I}_{x\in A_k}\mu(\mathrm{dx}) \overset{(\diamondsuit)}{=} \sum_{k\in\mathbb{N}} k \cdot 2^{-n} \int_{\mathbb{R}} \mathbb{I}_{x\in A\cap A_k}\mu(\mathrm{dx}).$$

这里 $(\diamondsuit)$ 这一步积分和求和可以交换的原因是每一项都是非负的，因此可以应用 MCT。于是，上式可以继续写成

$$\sum_{k\in\mathbb{N}} k \cdot 2^{-n}\mu(A \cap A_k) = \sum_{k\in\mathbb{N}} k \cdot 2^{-n}\nu(A \cap A_k) = \sum_{k\in\mathbb{N}} k \cdot 2^{-n} \int_{A\cap A_k} f(x)\mathrm{dx} = \int_A \underline{g}_n(x)f(x)\mathrm{dx}.$$

所以我们就证明了

$$\int_A \underline{g}_n(x)\mu(\mathrm{dx}) = \int_A \underline{g}_n(x)f(x)\mathrm{dx}.$$

对两边取极限，并使用 MCT 把极限和积分进行交换，我们便证明了

$$\int_A g(x)\mu(\mathrm{dx}) = \int_A g(x)f(x)\mathrm{dx}.$$

最后再提一句，设 F 是 X 的分布函数。我们有的时候会使用记号

$$\int_A g(x)dF(x) := \int_A g(x)\mu(\mathrm{dx}).$$

因此，我们有在勒贝格积分的意义下，如果 $F(x)$ 绝对连续，那么

$$\int_A g(x)dF(x) = \int_A g(x)f(x)\mathrm{dx}.$$

11.3.1 勒贝格积分与黎曼积分

在一个区间 $[a,b]$ 上的一个 Borel 函数 f，如果他是有界的，那么它是勒贝格可积的函数。如果它正好也是黎曼可积的，那么两个积分一定相等。在这儿，我简述一下证明。由于 f 是黎曼可积，设其积分是 S，那么对于任何 $\varepsilon > 0$，都存在一个 $\delta > 0$，满足对于 $[a,b]$ 的一个分划 $\{I_i\}_{i\in[n]}$，如果 $\lambda(I_i) \le \delta$，则用它定义的函数的黎曼和满足

$$\left| S - \sum_{i=1}^{n} f(x_i)\lambda(I_i) \right| \le \varepsilon,$$

其中 $x_i \in I_i$ 并且 λ 是勒贝格测度（即区间的长度）。我们定义一个函数 g，满足 $g(x) = \sum_{i=1}^{n} \sup_{x\in I_i} f(x) \cdot \mathbb{I}_{x\in I_i}$。那么显然 $f \le g$，并且由我们关于黎曼和的假设

$$\left| S - \sum_{i=1}^{n} \sup_{x_i\in I_i} f(x_i)\lambda(I_i) \right| \le \varepsilon.$$

而 $\sum_{i=1}^{n} \sup_{x_i\in I_i} f(x_i)\cdot\lambda(I_i)$ 正好是 g 的勒贝格积分 $\int_{[a,b]} g\mathrm{dx}$。因此，我们由勒贝格积分的单调性，$\int_{[a,b]} f\mathrm{dx} \le \int_{[a,b]} g\mathrm{dx} \le S + \varepsilon$。我们同样可以类似证明 $\int_{[a,b]} f\mathrm{dx} \ge S - \varepsilon$。由于这个不等式对于任何 ε 都成立，所以勒贝格积分 $\int_{[a,b]} f\mathrm{dx} = S$。

因此，我们在概率密度存在的情况下，可以通过数学分析课熟悉的对于黎曼积分的计算技巧，来处理随机变量的期望问题。我们将在未来看到更多实际计算的例子。

第 12 章 矩生成函数以及期望的一些基本性质

我们在前面几次课定义了很多数学对象，发展了很多数学工具。今天，我们开始使用他们来做一些具体的计算。

12.1 矩生成函数（Moment-Generating Function）

我们前面说到过，当我们研究一个分布的时候，我们经常需要计算它的各阶矩。设 X 是满足这个分布的一个随机变量，也就是说，我们想对于 $k = 1, 2, \ldots$，计算 $\mathbf{E}[X^k]$。这些量被称为随机变量的数字特征，它们反映了随机变量的各种性质。实际上，我们有一个统一的方法来计算所有阶矩，即计算它的矩生成函数

$$M_X(\theta) := \mathbf{E}[e^{\theta X}], \quad \theta \in \mathbb{R}.$$

当然了，在 $\theta \neq 0$ 的时候，随机变量 $e^{\theta X}$ 不一定可积。但是，如果它在 0 的某个非空邻域可积的话，那我们就可以通过 $M_X(\theta)$ 算出每一阶 $\mathbf{E}[X^k]$。

定理 12.1

假设存在某个 $b > 0$，使得 $M_X(\theta)$ 在 $-b \leq \theta \leq b$ 的时候存在，那么

$$\mathbf{E}[X] = \left.\frac{\mathrm{d}M_X(\theta)}{\mathrm{d}\theta}\right|_{\theta=0};$$

更一般的，对于任何 $k \geq 1$，

$$\mathbf{E}[X^k] = \left.\frac{\mathrm{d}^k M_X(\theta)}{\mathrm{d}\theta^k}\right|_{\theta=0}.$$

♡

证明 实际上，由于

$$\frac{\mathrm{d}^k e^{\theta X}}{\mathrm{d}\theta^k} = X^k e^{\theta X},$$

如果我们能够说明

$$\frac{\mathrm{d}^k \mathbf{E}[e^{\theta X}]}{\mathrm{d}\theta^k} = \mathbf{E}\left[\frac{\mathrm{d}^k e^{\theta X}}{\mathrm{d}\theta^k}\right],$$

也就是说求导和期望能够（对 θ 在 0 附近的时候）交换，那么我们的定理自然成立。

我们只对 $k = 1$ 来进行证明，对于 $k > 1$，可以在归纳法的基础上类似的说明。设 $\varepsilon < b$，θ 和 h 满足

$\theta, \theta+h \in [-b+\varepsilon, b-\varepsilon]$。根据定义，根据在给定范围内矩生成函数的可积性，我们有

$$M'_X(\theta) = \lim_{h\to 0} \frac{1}{h} \cdot (M_X(\theta+h) - M_X(\theta)) = \lim_{h\to 0} \mathbf{E}\left[\frac{1}{h}(e^{(\theta+h)X} - e^{\theta X})\right].$$

因此，我们需要说明，这儿的极限和期望可以交换。我们用 DCT 来说明这件事情，因此，我们只需要找到一个可积的随机变量 Y，对于给定范围内的 θ 和 h，我们说明 $Y \geq \left|\frac{1}{h}(e^{(\theta+h)X} - e^{\theta X})\right|$ 就行了。

根据中值定理，我们可以找到一个 $\theta' \in [\theta, \theta+h]$，满足

$$\frac{1}{h}(e^{(\theta+h)X} - e^{\theta X}) = Xe^{\theta' X}.$$

注意到，由于 $|\theta'| \leq b-\varepsilon$，我们有

$$|Xe^{\theta' X}| \leq |X| e^{|\theta'||X|} \leq e^{b|X|} \cdot |X| e^{-\varepsilon|X|}.$$

由于一定存在一个常数 C，使得 $|X| \cdot e^{-\varepsilon|X|} \leq C$，所以

$$|Xe^{\theta' X}| \leq C \cdot e^{b|X|} \leq C \cdot (e^{bX} + e^{-bX}) =: Y.$$

根据我们 b 的定义可以知道 Y 是可积的。这便证明了定理 $k=1$ 的情况。

12.1.1 一些常见分布的矩生成函数

例题 12.1. 伯努利分布 $\mathsf{Ber}(p)$

对于概率质量函数为 $p_X(0) = 1-p$，$p_X(1) = p$ 的伯努利分布，对于 $X \sim \mathsf{Ber}(p)$，由 LOTUS，我们知道

$$M_X(\theta) = \mathbf{E}[e^{\theta X}] = p \cdot e^{\theta} + (1-p).$$

对其求导，我们得到 $M'_X(\theta) = M''_X(\theta) = p \cdot e^{\theta}$。因此，$\mathbf{E}[X] = \mathbf{E}[X^2] = p$。

例题 12.2. 二项式分布 $\mathsf{Bin}(n,p)$

我们当然可以直接通过概率质量函数来计算二项式分布的矩生成函数。但如果注意到二项式分布的组合意义，对于 $X \sim \mathsf{Bin}(n,p)$，它可以写成 n 个分布为 $\mathsf{Ber}(p)$ 的独立随机变量之和，即 $X = X_1 + X_2 + \cdots + X_n$，其中每个 $X_i \sim \mathsf{Ber}(p)$，并且它们是相互独立的。于是，我们有

$$M_X(\theta) = \mathbf{E}[e^{\theta X}] = \mathbf{E}\left[e^{\theta(\sum_{i=1}^n X_i)}\right] = \prod_{i=1}^n \mathbf{E}[e^{\theta X_i}] = (p \cdot e^{\theta} + (1-p))^n.$$

上面的计算也告诉我们，若干个独立的随机变量之和的矩生成函数，是各个随机变量的矩生成函数的乘积。

例题 12.3. 超几何分布 $H(k,m,n)$

假设袋子里共有 n 个球，其中有 m 个黑球和 $n-m$ 个白球，现在从袋子中无放回地随机取出 k 个球。令随机变量 X 表示取出的黑球个数，我们称 X 服从参数为 (k,m,n) 的超几何分布。为方便计算，我们假设 $n > 2m > 2$ 以及 $n-m \geq k \geq m$。

其概率质量函数为

$$p_X(x) = \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}, \quad x = 0, 1, \ldots, \min(k,m).$$

由 LOTUS，X 的矩生成函数为

$$M_X(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \sum_{x=0}^{m} e^{\theta x} \cdot \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}.$$

对 $M_X(\theta)$ 求导，并在 $\theta = 0$ 处取值，可得

$$\mathbb{E}[X] = \left.\frac{dM_X}{d\theta}\right|_{\theta=0} = \frac{mk}{n}, \qquad \mathbb{E}[X^2] = \left.\frac{d^2M_X}{d\theta^2}\right|_{\theta=0} = \frac{m(m-1)k(k-1)}{n(n-1)} + \frac{mk}{n}.$$

因此，

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = \frac{m(m-1)k(k-1)}{n(n-1)} + \frac{mk}{n} - \left(\frac{mk}{n}\right)^2.$$

例题 12.4. 几何分布 $\mathsf{Geom}(p)$

对于 $X \sim \mathsf{Geom}(p)$，我们知道其概率质量函数为 $\forall k \geq 1, p_X(k) = (1-p)^{k-1} \cdot p$。因此，矩生成函数为

$$M_X(\theta) = \mathbf{E}[e^{\theta X}] = \sum_{k\geq 1}(1-p)^{k-1}p \cdot e^{\theta k} = \frac{p}{1-p}\sum_{k\geq 1}((1-p)e^{\theta})^k.$$

这是一个等比数列求和，因此，在 $(1-p)e^{\theta} < 1$，或者等价的 $\theta \leq \log(\frac{1}{1-p})$ 的时候，我们有

$$M_X(\theta) = \frac{pe^{\theta}}{1-(1-p)e^{\theta}}.$$

例题 12.5. 指数分布 $\mathsf{Exp}(\lambda)$

指数分布是一个连续分布，其概率密度函数为:

$$\forall x \geq 0,\ p_X(x) = \lambda e^{-\lambda x}; \quad \forall x < 0,\ p_X(x) = 0.$$

因此，由连续版本的 LOTUS，我们有

$$M_X(\theta) = \mathbf{E}[e^{\theta X}] = \int_0^{\infty} e^{\theta x} \cdot \lambda e^{-\lambda x} dx.$$

我们多数时候关心 θ 在 0 附近时候的取值，因此我们假设 $\theta < \lambda$，那么上面积分算出来便是

$$M_X(\theta) = \frac{\lambda}{\lambda - \theta}.$$

例题 12.6. Gamma 分布 $\mathsf{Gamma}(\alpha, \lambda)$

假设我们要考察一台机器的可靠性，考虑随机变量 $X_1, X_2, \ldots, X_n$，其中 X_i 表示这台机器第 $i-1$ 次故障到第 i 次故障之间的间隔时间。假设 $X_1, X_2, \ldots, X_n$ 互相独立，且均服从参数为 λ 的指数分布，即其概率密度函数为 $p_X(x) = \mathsf{Exp}(\lambda)$

令 $T_n = \sum_{i=1}^{n} X_i$ 表示从机器开始运行到发生第 n 次故障之间的总时间（假设机器故障后的检修时间为 0）。令 p_{T_n} 表示 T_n 的概率密度函数，可证明对于任意的 $t \geq 0$，

$$p_{T_n}(t) = \frac{t^{n-1} \cdot e^{-\lambda t} \lambda^n}{(n-1)!}.$$

而 T_n 的分布是 Gamma 分布的一种特殊情况。对于一般的参数为 (α, λ) 的 Gamma 分布（$\alpha, \lambda \in \mathbb{R}_+$），其概率密度函数为

$$\forall t > 0, \quad f(t) = \frac{t^{\alpha-1} \cdot e^{-\lambda t} \lambda^{\alpha}}{\Gamma(\alpha)},$$

其中 $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \mathrm{d}x$ 是 Gamma 函数，对于任意的 $\alpha > 0$，有 $\Gamma(\alpha) < \infty$。这里 α 被称为形状参数，λ 被称为尺度参数。我们把这个分布简记为 $\mathsf{Gamma}(\alpha, \lambda)$。

类似上文过程，我们可以得出，当 $\theta < \lambda$ 时，$M_X(\theta)$ 为

$$M_X(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^{\alpha}.$$

随后对 $M_X(\theta)$ 求导，并在 $\theta = 0$ 处取值，可得

$$\mathbb{E}[X] = \left.\frac{dM_X}{d\theta}\right|_{\theta=0} = \frac{\alpha}{\lambda}, \qquad \mathbb{E}[X^2] = \left.\frac{d^2 M_X}{d\theta^2}\right|_{\theta=0} = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

因此，

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}.$$

例题 12.7. Beta 分布 $\mathsf{Beta}(\alpha, \beta)$

在扔硬币之前，我们对参数 q 一无所知，因此可以将其视为一个在 $(0,1)$ 上均匀分布的随机变量。考虑样本空间 $\Omega = (0,1) \times ([T] \cup \{0\})$，即对于任意的 $\omega_1 \in (0,1), t \in [T] \cup \{0\}$，$\omega = (\omega_1, t) \in \Omega$。令 $\mathcal{F} = \sigma\left(\mathcal{B}(0,1) \times 2^{[T]\cup\{0\}}\right)$ 表示包含 $\mathcal{B}(0,1) \times 2^{[T]\cup\{0\}}$ 的最小 σ-代数，考虑概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$，$\forall A \in \mathcal{B}(0,1), t \in [T] \cup \{0\}$，测度定义为

$$\mathbb{P}(A \times \{t\}) = \int_{q \in A} \binom{T}{t} q^t (1-q)^{T-t} \mathrm{d}q.$$

我们定义 $(\Omega, \mathcal{F}, \mathbb{P})$ 上的随机变量 Q 为 $\forall \omega = (\omega_1, t) \in \Omega$, $Q(\omega) = \omega_1$；以及定义随机变量 Y 为 $\forall \omega = (\omega_1, t) \in \Omega, Y(\omega) = t$。

对于 $t \in [T] \cup \{0\}$，容易计算得：

$$\begin{aligned}
\mathbb{P}(Y = t) &= \binom{T}{t} \int_0^1 q^t (1-q)^{T-t} \mathrm{d}q \\
&= \binom{T}{t} \left(\frac{1}{t+1} q^{t+1} (1-q)^{T-t} \Big|_0^1 + \int_0^1 \frac{T-t}{t+1} q^{t+1} (1-q)^{T-t-1} \mathrm{d}q \right) \\
&= 0 + \binom{T}{t} \int_0^1 \frac{T-t}{t+1} q^{t+1} (1-q)^{T-t-1} \mathrm{d}q \\
&= \mathbb{P}(Y = t+1).
\end{aligned}$$

由此，由二项式定理，Q 的边缘概率密度函数为

$$p_Q(q)=\sum_{t=0}^{T}\binom{T}{t}q^t(1-q)^{T-t}=1.$$

条件概率密度函数 $p_{Q|Y}(q|t)$ 为

$$\begin{aligned}p_{Q|Y}(q|t)&=\lim_{h\to 0}\frac{1}{h}\cdot\frac{\mathbb{P}(Q\in[q,q+h],Y=t)}{\mathbb{P}(Y=t)}\\&=(T+1)\lim_{h\to 0}\frac{1}{h}\int_q^{q+h}\binom{T}{t}q'^t(1-q')^{T-t}\mathrm{d}q'\\&=(T+1)\binom{T}{t}q^t(1-q)^{T-t}.\end{aligned}$$

直观上看，上问中的条件分布表示一共出现了 t 个正面的情况下参数 q 的后验分布，这实际上是 Beta 分布的一种特殊情况。假设随机变量 X 服从参数为 α 和 β 的 Beta 分布 $(\alpha,\beta>0)$，记为 $X\sim\mathsf{Beta}(\alpha,\beta)$，其概率密度函数为

$$\forall x\in(0,1),\quad p_X(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}.$$

其中 $\Gamma(\cdot)$ 是前面定义过的 Gamma 函数。
可以证明：X 的矩生成函数为

$$M_X(\theta)=1+\sum_{n=1}^{\infty}\frac{\theta^n}{n!}\cdot\left(\prod_{k=0}^{n-1}\frac{\alpha+k}{\alpha+\beta+k}\right).$$

由定义，$M_X(\theta)=\mathbb{E}\left[e^{\theta X}\right]$，且其 n 阶导数在 $\theta=0$ 处满足 $\frac{d^n}{d\theta^n}M_X(\theta)\big|_{\theta=0}=\mathbb{E}[X^n]$。$X$ 的 n 阶矩为

$$\begin{aligned}\mathbb{E}[X^n]&=\int_0^1\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{n+\alpha-1}(1-x)^{\beta-1}\mathrm{d}x\\&=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(n+\alpha)\Gamma(\beta)}{\Gamma(n+\alpha+\beta)}\\&=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\cdot\frac{\Gamma(n+\alpha)}{\Gamma(n+\alpha+\beta)}\\&=\prod_{k=0}^{n-1}\frac{\alpha+k}{\alpha+\beta+k}\\&=\frac{\alpha(\alpha+1)\cdots(\alpha+n-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+n-1)}.\end{aligned}$$

因此，

$$M_X(\theta)=\sum_{n=0}^{\infty}\frac{\theta^n}{n!}\mathbb{E}[X^n]=1+\sum_{n=1}^{\infty}\frac{\theta^n}{n!}\cdot\left(\prod_{k=0}^{n-1}\frac{\alpha+k}{\alpha+\beta+k}\right).$$

特别地，$\mathbb{E}[X] = M'_X(0)$ 且 $\mathbb{E}[X^2] = M''_X(0)$。我们计算：

$$\begin{aligned}
M'_X(\theta) &= \sum_{n=1}^{\infty} \frac{n\theta^{n-1}}{n!}\mathbb{E}[X^n] = \sum_{n=1}^{\infty} \frac{\theta^{n-1}}{(n-1)!}\mathbb{E}[X^n], \\
\Rightarrow M'_X(0) &= \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \\
M''_X(\theta) &= \sum_{n=2}^{\infty} \frac{n(n-1)\theta^{n-2}}{n!}\mathbb{E}[X^n] = \sum_{n=2}^{\infty} \frac{\theta^{n-2}}{(n-2)!}\mathbb{E}[X^n], \\
\Rightarrow M''_X(0) &= \mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.
\end{aligned}$$

因此，方差为

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.
\end{aligned}$$

例题 12.8. 标准高斯分布 $\mathcal{N}(0,1)$

标准高斯分布的概率密度函数是 $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$。我们数学分析课学习过计算高斯积分

$$I = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx$$

的技巧。即考虑

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy.$$

再使用极坐标变换 $x = r\cos\theta, y = r\sin\theta$：

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r dr\mathrm{d}\theta = 2\pi.$$

因此，我们可以计算其矩生成函数为

$$M_X(\theta) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\theta x}\cdot e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\theta)^2} dx \cdot e^{\frac{\theta^2}{2}} = e^{\frac{\theta^2}{2}}.$$

12.2 期望的一些常用结论

命题 12.1 (全期望公式（Law of Total Expectation）)

我们在之前定义过全概率公式，即假设事件集 $\{A_i\}_{i\geq 1}$ 是对于样本空间 Ω 的一个分划，那么对于任何事件 B，

$$\mathbb{P}[B]=\sum_{i\geq 1}\mathbb{P}[B\cap A_i]=\sum_{i\geq 1:\mathbb{P}[A_i]>0}\mathbb{P}[B\mid A_i]\cdot\mathbb{P}[A_i].$$

我们现在给出一个类似的 全期望公式。对于一个事件 A，如果 $\mathbb{P}(A)\geq 0$，我们定义随机变量的条件期望

$$\mathbf{E}[X\mid A]:=\frac{\int_A X d\mathbb{P}}{\mathbb{P}(A)}.$$

它可以直观的理解成，在 A 事件发生的情况下，随机变量 X 的平均值。因此，假设 $\{A_i\}_{i\geq 1}$ 是对于样本空间 Ω 的一个分划，我们显然有

$$\mathbf{E}[X]=\sum_{i\geq 1}\mathbf{E}[X\mid A_i]\cdot\mathbb{P}[A_i].$$

我们在之前对于离散随机变量给出了马尔科夫不等式：对于非负随机变量 X 以及任何 $a>0$，由 $\mathbb{P}[X\geq a]\leq\frac{\mathbf{E}[X]}{a}$。我们现在定义了一般随机变量的期望，同样的结论成立。我们可以照搬离散场合的证明，我们这儿等价的用条件期望的语言说一下：

定理 12.2 (马尔科夫不等式（Markov Inequality）)

$$\mathbf{E}[X]=\mathbf{E}[X\mid[X\geq a]]\mathbb{P}[X\geq a]+\mathbf{E}[X\mid[X<a]]\mathbb{P}[X<a]\geq a\cdot\mathbb{P}[X\geq a].$$

于是

$$\mathbb{P}[X\geq a]\leq\frac{\mathbf{E}[X]}{a}.$$

我们在数学分析课中学习过柯西-施瓦茨不等式，它的一般形式是说，两个向量的内积小于等于各自长度的乘积。事实上，两个随机变量的乘积的期望也可以看成是某种内积，所以我们有期望形式的柯西施瓦茨不等式：

定理 12.3 (柯西-施瓦茨不等式（Cauchy-Schwarz Inequality）)

对于平方可积的随机变量 X 和 Y，

$$\mathbf{E}[XY]\leq\sqrt{\mathbf{E}[X^2]}\cdot\sqrt{\mathbf{E}[Y^2]}.$$

证明 这儿实际上我们用到了一个性质，如果 X 是平方可积 ($\mathbf{E}[X^2]<\infty$)，那么 $\mathbf{E}[X]<\infty$。大家可以想想为什么。我们给不等式一个很巧妙的证明。对于任意 λ，我们都有

$$0\leq\mathbf{E}[(X-\lambda Y)^2]=\mathbf{E}[X^2]-2\lambda\mathbf{E}[XY]+\lambda^2\mathbf{E}[Y^2].$$

因此

$$\mathbf{E}[XY]\leq\frac{1}{2\lambda}\cdot\mathbf{E}[X^2]+\frac{\lambda}{2}\cdot\mathbf{E}[Y^2].$$

如果 X 和 Y 中有一个几乎处处是 0，那么柯西-施瓦茨不等式显然成立。否则，我们可以在上式中令 $\lambda=\frac{\sqrt{\mathbf{E}[X^2]}}{\sqrt{\mathbf{E}[Y^2]}}$ 即可。

我们在数学分析课或者优化课程中学习过琴生不等式，但是在那边遇到的也许是有限求和的版本。我们现在介绍一个期望的版本。由于我们对于期望的定义非常的一般化，这推广了我们遇到过有限版本。

定理 12.4 (琴生不等式（Jensen's Inequality）)

假设 $f:(a,b)\to\mathbb{R}$ 是一个凸函数，随机变量 X 的取值是在 (a,b)，并且 $f(X)$ 可积，那么 $\mathbf{E}[f(X)]\geq f(\mathbf{E}[X])$。♡

证明 我们回忆优化课上对于凸函数的各种定义和刻画，其中之一便是对任何一个点 $x\in\mathbb{R}$ 上，均可以找到过 $(x,f(x))$ 的一条直线，称之为支撑线，使得函数 f 的图像全部落在这个直线的上方。我们便利用这一个性质来证明琴生不等式。

我们设 $x_0=\mathbf{E}[X]$。考虑过 x_0 的一条支撑线 $y(x)=f(x_0)+k(x-x_0)$。我们知道 $f(x)$ 落在 $y(x)$ 的上方，因此

$$f(X)\geq y(X)=f(x_0)+k(X-x_0).$$

对两边取期望即得证。

第 13 章 乘积概率空间，富比尼-托内利定理

在我们上周的作业里，我们让大家举反例说明如下命题是 **不对** 的：如果每一个 X_i 均可积，那么

$$\mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathbb{E}[X_i].$$

我们知道，无论是期望还是求和，都可以看成在某个测度下的勒贝格积分。因此，我们今天想来回答，在什么条件下，积分是可以交换的，即

$$\int_A \left(\int_B f dP\right) dQ = \int_B \left(\int_A f dQ\right) dP$$

成立。我们首先来定义乘积测度空间。

13.1 乘积概率空间

假设我们有两个概率空间 $(X, \mathscr{X}, \mu)$ 和 $(Y, \mathscr{Y}, \nu)$。我们可以想象：

- 第一个概率空间是掷一个骰子得到 $\omega_1 \in X$，第二个概率空间是掷另一个骰子，得到 $\omega_2 \in Y$；或者
- 第一个概率空间是在 $[0,1]$ 上均匀取一个数 a，第二个概率空间是在 $[0,2]$ 上均匀取一个数 b。

我们现在可以考虑一个概率空间，称为乘积概率空间，它的样本集是 $X \times Y$，用来表示独立的做两次实验。在我们上面的两个例子里，分别对应了：

- 同时投两枚骰子，得到 (ω_1, ω_2)；
- 在 $[0,1] \times [0,2]$ 上均匀取一个点 (a,b)。

所以一个自然的问题是，这个乘积概率空间里，事件集和测度应该是什么。离散概率空间比较平凡，我们考虑上面第二个例子。这个时候 $\mathscr{X} = \mathscr{B}([0,1])$，$\mathscr{Y} = \mathscr{B}([0,2])$。集合 $\mathscr{X} \times \mathscr{Y} = \{A \times B : A \in \mathscr{X}, B \in \mathscr{Y}\}$ 是 $[0,1] \times [0,2]$ 上所有“矩形”的集合。它显然不是一个 σ-代数，因此，我们可以取乘积概率空间的事件集为 $\sigma(\mathscr{X} \times \mathscr{Y})$，并把它记做 $\mathscr{X} \otimes \mathscr{Y}$。

更一般的，对于两个测度空间 $(X, \mathscr{X}, \mu)$ 和 $(Y, \mathscr{Y}, \nu)$，我们定义它的乘积空间为 $(X \times Y, \mathscr{X} \otimes \mathscr{Y})$，其中 $\mathscr{X} \otimes \mathscr{Y} := \sigma(\mathscr{X} \times \mathscr{Y})$。我们接下来定义 $\mathscr{X} \otimes \mathscr{Y}$ 上的测度。看起来有两种比较自然的定义方式。

给定一个集合 $E \in \mathscr{X} \otimes \mathscr{Y}$，我们定义两个函数 $E_1 : x \in X \mapsto E_1(x) \subseteq Y$ 和 $E_2 : y \in Y \mapsto E_2(y) \subseteq X$，分别表示集合 E 在 $X = x$ 和 $Y = y$ 上的投影，如下图所示。

$$E_1(x) := \{y \in Y : (x,y) \in E\}, \quad E_2(y) := \{x \in X : (x,y) \in E\}.$$

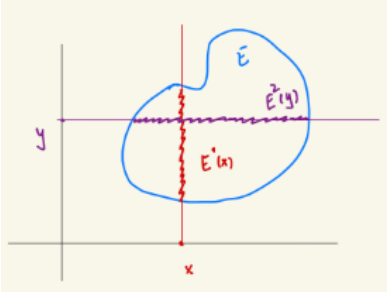


图 13.1: Figure1

我们需要说明，如果 E 是 $\mathscr{X}\otimes\mathscr{Y}$ 可测的，那么对于任意 x 和 y，$E_1(x)$ 与 $E_2(y)$ 分别是 $\mathscr{Y}$ 与 $\mathscr{X}$ 可测的。

命题 13.1

如果 E 是 $\mathscr{X}\otimes\mathscr{Y}$ 可测的，那么对于任意 x 和 y，$E_1(x)$ 与 $E_2(y)$ 分别是 $\mathscr{Y}$ 与 $\mathscr{X}$ 可测的。 ♠

我们把命题的证明放到本次讲义的最后，先来继续我们的讨论。

于是，我们可以定义

$$\pi_1(E):=\int_X \nu(E_1(x))\mu(dx);\quad \pi_2(E):=\int_Y \mu(E_2(y))\nu(dy).$$

换句话说，$\pi_1(E)$ 和 $\pi_2(E)$ 分别对应着在计算 E 的面积的时候“横着积”和“竖着积”。设

$$\mathscr{G}=\{E\in\mathscr{X}\otimes\mathscr{Y}:\pi_1(E)=\pi_2(E)\}$$

为所有 π_1 和 π_2 相等的可测集的集合。对于矩形 $E=A\times B\in\mathscr{X}\times\mathscr{Y}$，我们显然有

$$\pi_1(E)=\pi_2(E)=\mu(A)\times\nu(B),$$

即 $\mathscr{X}\times\mathscr{Y}\subseteq\mathscr{G}$。如果 μ 和 ν 均是 σ-有限的，我们可以使用单调类定理说明 $\mathscr{X}\otimes\mathscr{Y}\subseteq\mathscr{G}$（留作练习）。所以，我们可以定义测度

$$\forall E\in\mathscr{X}\otimes\mathscr{Y},\quad (\mu\otimes\nu)(E):=\pi_1(E)=\pi_2(E).$$

于是，$(X\times Y,\mathscr{X}\otimes\mathscr{Y},\mu\otimes\nu)$ 便是我们构造出来的乘积空间。

我们考虑一个特殊的例子。设 $X=Y=\mathbb{R}$，$\mathscr{X}=\mathscr{Y}=\mathscr{B}(\mathbb{R})$，$\mu=\nu=\lambda$。那么 $(X,\mathscr{X},\mu)$ 和 $(Y,\mathscr{Y},\nu)$ 均为一维的 $\mathbb{R}$ 上的勒贝格测度空间。按照我们的定义，它的乘积空间是 $(\mathbb{R}^2,\mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R}),\lambda\otimes\lambda)$。同样根据定义，我们知道 $\mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R})$ 就是 $\mathscr{B}(\mathbb{R}^2)$，并且根据 Carathéodory 定理，$\lambda\otimes\lambda$ 就是 $\mathbb{R}^2$ 上的勒贝格测度。

我们的定义可以推广到任意有限个概率空间的乘积。对于无穷多个概率空间的乘积，情况比较复杂，超出了这门课的范畴，可以参见 Kolmogorov extension theorem。

13.2 富比尼-托内利定理（Fubini-Tonelli's Theorem）

富比尼-托内利定理说的是，如果一个定义在乘积空间 $(X\times Y,\mathscr{X}\otimes\mathscr{Y},\mu\otimes\nu)$ 上的可测函数 f 非负，或者可积，那么有等式

$$\int_{X\times Y} f(x,y)\mu\otimes\nu(dxdy)=\int_X\left(\int_Y f(x,y)\nu(dy)\right)\mu(dx)=\int_Y\left(\int_X f(x,y)\mu(dx)\right)\nu(dy)$$

成立。

也就是说，积分的顺序可以任意交换。当然，这个等式成立需要先保证对于任意 x，函数 $f(x,\cdot):y\in Y\mapsto f(x,y)\in\mathbb{R}$ 是可测的（对函数 $f(\cdot,y)$ 同理）。这是上一节关于 E_1 和 E_2 函数可测的更一般版本。我们将在最后一节证明。

值得强调的是，关于 f 的可积性要求是 $|f|$ 在 $\mu\otimes\nu$ 这个乘积测度上是可积的，如果仅仅是对于任何 $y\in Y$，$f(\cdot,y)$ 在 μ 上可积，或者对于任何 $x\in X$，$f(x,\cdot)$ 在 ν 上可积，是不够的。我们本次课一开始提到的那个反例就

说明了这一点。

我们简单叙述一下定理的证明，类似的套路大家现在应该很熟悉了，具体的细节请大家自己完成。我们只讨论 $f \geq 0$ 的情况，可积的情形可以转化为分别考虑 f^+ 与 f^- 两个非负函数。

1. 首先，如果 $f = \mathbb{I}_E$，其中 E 是某个可测集。那么此时，我们想要证明的等式即我们上一节验证过的“横着积”等于“竖着积”的测度等式。
2. 然后，我们考虑 f 的下近似 $\underline{f}_n$。对于每一个 n，它都可以写成 $\sum_i c_i \cdot \mathbb{I}_{E_i}$ 的形式，其中 E_i 是可测集。于是，使用期望的线性性，我们可以证明此时的等式。
3. 使用 MCT，我们可以对 $\underline{f}_n$ 的积分求极限，得到对 f 的积分。

在具体使用 Fubini-Tonelli 的时候，我们往往是这样做的：

- 如果 f 非负，那么可以随意的交换积分顺序进行计算。
- 如果 f 可正可负，我们先对 $|f|$ 计算某一个累次积分，比如 $\int_X (\int_Y |f| d\nu) d\mu$。如果其 $< \infty$，那么就说明它满足我们定理的条件。

13.3 $f(x,\cdot), f(\cdot,y)$ 可测的证明

我们接下来证明在乘积测度空间 $(X \times Y, \mathscr{X} \otimes \mathscr{Y}, \mu \otimes \nu)$ 上的可测函数 $f : (x,y) \in X \times Y \mapsto \mathbb{R}$ 的限制 $f(x,\cdot) : y \in Y \mapsto f(x,y)$ 和 $f(\cdot,y) : x \in X \mapsto f(x,y)$ 分别是在 $\mathscr{Y}$ 和 $\mathscr{X}$ 上可测的。

我们证明 $f(x,\cdot)$ 的情况。考虑一个固定的 $x \in X$ 以及一个函数

$$T_x : y \in Y \mapsto (x,y) \in X \times Y.$$

显然，$f(x,\cdot) = f \circ T_x$。我们只需要验证 T_x 是从 $(X, \mathscr{X})$ 到 $(X \times Y, \mathscr{X} \otimes \mathscr{Y})$ 的可测函数即可。因为容易验证，两个可测函数的复合函数还是一个可测函数。

为了验证 T_x 是一个可测函数，我们只需要验证，对于每一个 $\mathscr{X} \otimes \mathscr{Y}$ 里的矩形 $A \times B$，$T_x^{-1}(A \times B) \in \mathscr{Y}$ 即可（我们以前证明过这件事情，还记得吗）。但

$$T_x^{-1}(A \times B) = \begin{cases} B, & \text{if } x \in A; \\ \varnothing, & \text{if } x \notin A. \end{cases}$$

因此 T_x 是可测的。

第 14 章 联合分布，联合密度函数，条件密度函数

我们之前介绍了一个随机变量的分布函数、分布、概率质量/密度函数等，今天，我们开始介绍定义在同一个概率空间上的多个随机变量的“联合”分布。

14.1 联合分布（Joint Distribution）

我们还是固定一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$。对于定义在上面的两个随机变量 $X, Y : \Omega \to \mathbb{R}$，我们定义它们的联合分布函数 $F : \mathbb{R}^2 \to [0,1]$ 为

$$\forall x, y \in \mathbb{R}, \quad F(x,y) := \mathbb{P}[X \le x, Y \le y].$$

这个定义可以直接推广成任意有限个随机变量的联合分布

$$\forall x_1, \dots, x_n \in \mathbb{R}, \quad F(x_1, \dots, x_n) := \mathbb{P}[X_1 \le x_1, \dots, X_n \le x_n].$$

对于一般的 n，联合分布的大部分性质和 $n = 2$ 时并没有本质区别，因此，我们接下来的讨论均以 $n = 2$ 为例。除非额外说明，我们所述的性质都可以被推广到一般的有限 n。

命题 14.1 (多元分布函数的性质)

1. $x \mapsto F(x,y)$ 以及 $y \mapsto F(x,y)$ 均是左极限存在，右连续的非降函数；
2. $F(x,\infty) := \lim_{y\to\infty} F(x,y)$，$F(\infty,y) := \lim_{x\to\infty} F(x,y)$ 均存在；
3. 对于每一个 x, y，$\lim_{x\to-\infty} F(x,y) = \lim_{y\to-\infty} F(x,y) = 0$，$\lim_{x,y\to\infty} F(x,y) = 1$；
4. $\mathbb{P}[X = x, Y = y] = F(x,y) - F(x-,y) - F(x,y-) + F(x-,y-)$，其中 $F(x-,y) := \lim_{u\uparrow x} F(u,y)$（其余类似）。

当我们谈论 X 和 Y 的联合分布函数的时候，有时候会把 X, Y 作为下标，记作 F_{XY}。我们定义 **边缘分布函数** (marginal distribution function)

$$F_X(x) := F_{XY}(x,\infty), \quad F_Y(y) := F_{XY}(\infty,y).$$

显然，$F_X(x)$ 和 $F_Y(y)$ 就是 X 和 Y 对应的分布函数。我们这儿称之为“边缘”的原因是强调它们分别是一个联合分布的一部分。

14.1.1 概率质量函数与概率密度函数

对于离散随机变量 X 和 Y，我们有 **联合质量函数**

$$p_{XY}(x,y) := \mathbb{P}[X=x, Y=y].$$

显然，随机变量 X 和 Y 的概率质量函数 p_X 和 p_Y 分别满足

$$p_X(x) = \sum_{y\in \mathsf{Im}(Y)} p(x,y), \quad p_Y(y) = \sum_{x\in \mathsf{Im}(X)} p(x,y).$$

我们有时候把它们称为 X 或者 Y 对应的 **边缘概率质量函数**（marginal probability mass function）。

类似的，假设 X 和 Y 是连续随机变量，如果存在一个非负的函数 $f(x,y)$，满足

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv,$$

则称 $f(x,y)$ 是 X 和 Y 的 **联合概率密度**。我们可以使用单调类定理说明，对于任何的 $A\in\mathscr{B}(\mathbb{R}^2)$，

$$\mathbb{P}[(X,Y)\in A] = \int_A f(x,y)dx\otimes dy.$$

由微积分基本定理，如果 f 在 (x,y) 连续，那么

$$f(x,y) = \frac{\partial^2}{\partial x\partial y}F(x,y).$$

我们同样可以定义 **边缘密度函数** f_X 和 f_Y 为

$$f_X(x) := \int_{-\infty}^{\infty} f(x,y)dy, \quad f_Y(y) := \int_{-\infty}^{\infty} f(x,y)dx.$$

容易验证，它们实际上分别是 F_X 和 F_+Y 的密度函数。

值得注意的是，如果 (X,Y) 具有联合密度函数，那么它们就有边缘密度函数，但反过来不一定成立。比如 X 是 $(0,1)$ 上均匀取的一个数，$Y=X$，容易验证，(X,Y) 不存在联合密度函数（why?）。

如果 X 和 Y 有连续的联合密度函数 $f_{XY}(x,y)$，那么 X 和 Y 独立当且仅当 $f_{XY}(x,y)=f_X(x)f_Y(y)$。为了说明这一点，我们只需要注意到

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x\partial y}F_{XY}(x,y) = \frac{\partial^2}{\partial x\partial y}(F_X(x)F_Y(y)) = f_X(x)f_Y(y)$$

即可。

14.2 条件分布与条件密度（Conditional Distribution）

我们接下来讨论条件概率。我们在之前介绍概率空间的时候已经定义过条件概率了。给定两个事件 $A,B\in\mathscr{F}$，如果 $\mathbb{P}[B]\neq 0$，那么我们定义条件概率

$$\mathbb{P}[A\mid B] = \frac{\mathbb{P}[A\cap B]}{\mathbb{P}[B]}.$$

这个定义可以自然的给出离散的随机变量的条件期望的定义。假设 X 是一个离散的随机变量，那么，对于任何可测集 A 和 x，如果 $\mathbb{P}[X=x]>0$，那么，我们可以无缝使用上面的定义得到

$$\mathbb{P}[Y\in A\mid X=x] = \frac{\mathbb{P}[Y\in A\wedge X=x]}{\mathbb{P}[X=x]}.$$

如果 $\mathbb{P}[X=x]=0$，这个时候 $\mathbb{P}[Y\in A\mid X=x]$ 是无定义的。我们可以同时自然的定义出条件分布函数

$$F_{Y|X}(y\mid x) := \mathbb{P}[Y\le y\mid X=x];$$

以及得到对应的条件质量函数

$$p_{Y|X}(y\mid x) = \begin{cases} \dfrac{p_{YX}(y,x)}{p_X(x)}, & \text{if } p_X(x)>0; \\ 0, & \text{otherwise.}\end{cases}$$

我们可以同时给出条件期望的定义。如果 Y 可积并且 $\mathbb{P}[X=x]>0$，那么定义

$$\mathbb{E}[Y \mid X=x] := \frac{\mathbb{E}[Y \cdot \mathbb{I}_{X=x}]}{\mathbb{P}[X=x]}.$$

上面这些定义都是非常自然，而且我们之前在作业里也多次显式或者隐式的使用过了。但是，当 X 不是离散随机变量的时候，这样的定义就会出现一些问题。比如说，假设 X 和 Y 是独立的从 $[0,1]$ 中均匀得到的两个数，那么直观上，我们应该有 $\mathbb{P}[Y \le \frac{1}{2} \mid X=\frac{1}{3}] = \frac{1}{2}$。但由于 $\mathbb{P}[X=\frac{1}{3}]=0$，我们上述给出的条件概率定义是一个形如 $\frac{0}{0}$ 的没有意义的数。因此，我们需要对条件概率有新的定义。实际上，在概率论里面，条件概率是条件期望的特殊情况，而最一般的条件期望的定义，我们现在还没有准备好。

大约在这门课的最后，我们会给出定义。今天，我们先讨论一个特殊情况，即在 X 和 Y 有连续的联合密度函数 f_{XY} 的时候，定义条件期望与条件概率。

我们刚才说了，由于 $\mathbb{P}[X=x]=0$，我们从近似的角度来考虑这个问题。根据微积分基本定理，对于一个很小的 $h>0$，我们有

$$\begin{aligned}\mathbb{P}\left[Y \le y \mid X \in [x, x+h]\right] &= \frac{\int_{-\infty}^{y}\int_{x}^{x+h} f_{XY}(u,v)\,du\,dv}{\int_{x}^{x+h} f_X(u)\,du} \\ &= \frac{\int_{-\infty}^{y} h\cdot f_{XY}(x,v)+o(h)\,dv}{(h+o(h))f_X(x)} \\ &= \frac{\int_{-\infty}^{y} f_{XY}(x,v)\,dv + h^{-1}\int_{-\infty}^{y} o(h)\,dv}{f_X(x)+o(1)}.\end{aligned}$$

如果我们假设 f_{XY} 有一定的正则性使得 $\lim_{h\to 0} h^{-1}\int_{-\infty}^{y} o(h)dv = \int_{-\infty}^{y}\lim_{h\to 0} h^{-1}o(h) = 0$，则我们可以对于可测的 A，定义

$$\mathbb{P}[Y\in A \mid X=x] := \lim_{h\to 0}\mathbb{P}[Y\in A \mid X\in[x,x+h]].$$

更一般的（f_{XY} 不一定连续），我们可以自然的定义条件分布函数

$$F_{Y|X}(y \mid x) := \begin{cases}\int_{-\infty}^{y}\frac{f_{XY}(x,v)}{f_X(x)}dv, & \text{if } f_X(x)>0,\\ 0, & \text{if } f_X(x)=0.\end{cases}$$

其对应的条件密度函数为

$$f_{Y|X}(y \mid x) = \begin{cases}\frac{f_{XY}(x,y)}{f_X(x)}, & \text{if } f_X(x)>0,\\ 0, & \text{if } f_X(x)=0.\end{cases}$$

我们也定义条件期望

$$\mathbb{E}[Y \mid X=x] := \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)dy.$$

条件期望是一个非常重要的概念，我们在未来会专门讨论条件期望的性质并给出对应的应用，今天，我们暂时了解这个定义即可。

我们接着验证一下，所谓全概率公式，对于具有连续联合密度的随机变量也成立。

命题 14.2 (全概率公式)

$$\mathbb{P}[Y\in A] = \int_{-\infty}^{\infty}\int_{A} f_{Y|X}(y \mid x) f_X(x)dydx.$$

我们仅需要把定义代进去，并使用 Fubini-Tonelli 交换积分顺序即可证明。注意到

$$\begin{aligned}\int_{-\infty}^{\infty}\int_A f_{Y|X}(y|x)f_X(x)\,dy\,dx &= \int_{-\infty}^{\infty}\int_A \frac{f_{XY}(x,y)}{f_X(x)}\cdot f_X(x)\,dy\,dx\\ &= \int_{-\infty}^{\infty}\int_A f_{XY}(x,y)\,dy\,dx\\ &= \int_A f_Y(y)\,dy\\ &= \mathbb{P}\left[Y\in A\right].\end{aligned}$$

使用类似的证明，我们可以更一般的得到，对于 $A,B\in\mathscr{F}$，

$$\mathbb{P}[Y\in A\wedge X\in B]=\int_B\int_A f_{Y|X}(y\mid x)f_X(x)dydx.$$

14.2.1 积分的换元

我们现在考虑一个在计算中经常会遇到的问题，假设我们知道随机变量 X 和 Y 的联合密度函数 f_{XY}，那么对于新的随机变量 $(U,V)=g(X,Y)=(g_1(X,Y),g_2(X,Y))$，它们的联合密度函数 f_{UV} 是什么？这儿 $g_1,g_2:\mathbb{R}^2\to\mathbb{R}$ 是两个可测函数，并且我们假设它们是可微的。

对于一个可积的测试函数 $\phi:\mathbb{R}^2\to\mathbb{R}$，我们考虑用两种方法来计算 $\mathbb{E}[\phi(U,V)]$。首先是通过 U,V 的联合密度函数 f_{UV}：

$$\mathbb{E}[\phi(U,V)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(u,v)f_{UV}(u,v)dudv.$$

接着是通过 X,Y 的联合密度函数 f_{XY}：

$$\mathbb{E}[\phi(U,V)]=\mathbb{E}[\phi(g(X,Y))]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(g(x,y))f_{XY}(x,y)dxdy.$$

我们再把上面第一个式子使用换元公式得到

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(u,v)f_{UV}(u,v)dudv=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(g(x,y))f_{UV}(g(x,y))|\det J_g(x,y)|dxdy,$$

其中 $J_g(x,y)$ 是 g 在 (x,y) 处的雅可比矩阵

$$J_g(x,y)=\begin{pmatrix}\frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y}\\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y}\end{pmatrix}.$$

所以，我们可以得到如下命题：

命题 14.3

$$f_{XY}(x,y)=f_{UV}(g(x,y))|\det J_g(x,y)|.$$

♠

例题 14.1. 极坐标的例子

我们考虑下面的例子，假设 X 和 Y 是两个独立的标准正态分布随机变量，那么它们的联合密度函数为 $f_{XY}(x,y)=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}$。我们可以把 (X,Y) 看成 $\mathbb{R}^2$ 上的随机的点。我们考虑这些点的极坐标 (R,Θ)，其中 $R=\sqrt{X^2+Y^2}$，$\Theta=\arctan\frac{Y}{X}$。我们想知道 (R,Θ) 的联合密度函数是什么。

我们首先知道，$X=R\cos\Theta$，$Y=R\sin\Theta$。这个变换的雅可比矩阵的行列式是 r。因此，根据命题 14.3，我们有

$$f_{R\Theta}(r,\theta)=f_{XY}(r\cos\theta,r\sin\theta)\cdot r=\frac{r}{2\pi}e^{-\frac{r^2}{2}}.$$

大家会发现这个式子是与 θ 无关的，这说明关于 θ 的边缘分布是均匀分布。这件事情的一个推论是，如果我们希望从二维的单位圆上均匀的取出一个点来，我们只需独立的取两个标准高斯变量 (X,Y)，然后把

它归一化成长度为 1 的向量 $\left(\frac{X}{\sqrt{X^2+Y^2}}, \frac{Y}{\sqrt{X^2+Y^2}}\right)$ 即可。这件事情对于高维也是成立的，对于算法设计很有意义。

例题 14.2. 随机变量的和

假设知道 X 和 Y 的联合概率密度 f_{XY}，我们来考虑两个随机变量的和 $Z = X + Y$ 的概率密度。我们首先引入一个辅助变量 $W = Y$，于是对于 $(Z, W), (X, Y) = g(Z, W)$，我们有

$$|\det J_g(z, w)| = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

所以根据命题 14.3，我们有

$$f_{ZW}(z, w) = f_{XY}(z - w, w).$$

我们可以计算出 Z 的边缘密度函数为

$$f_Z(z) = \int_{-\infty}^{\infty} f_{ZW}(z, w)dw = \int_{-\infty}^{\infty} f_{XY}(z - w, w)dw = \int_{-\infty}^{\infty} f_Y(w) \cdot f_{X|Y}(z - w \mid w)dw.$$

特别的，如果 X 和 Y 独立，那么

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - w)f_Y(w)dw.$$

第 15 章 指数分布与泊松过程

我们接下来的几次课都会使用之前介绍的工具来解决一些实际的问题。我们通过研究一些具体的分布，理解它们直观上的含义，证明一些有趣的性质，解决一些初看起来并不是那么容易处理的问题。便让我们从泊松分布开始。

15.1 泊松分布

考虑一个场景：某餐厅过去五天的顾客数量分别是 100、120、80、75 和 110。为了确保明天的食材准备足够，我们需要根据前几天的顾客数量预测明天的顾客数量。一个常见的方法是计算顾客数量的平均值（在本例中为 97）。然而，仅仅使用期望的信息往往是不够的，有可能造成大量天数出现食材短缺的情况。因此，我们尝试建模每天顾客到来人数的分布。为了得到这个分布，我们需要做一些假设。一个最常见的假设是假设顾客的到来是均匀且独立的。我们可以假设一天被分为 n 个等长的时间段，每个时间段足够短，以至于在该时间段内最多只能有一位顾客进入餐厅，并且在每个时间段以 p 的概率独立有一位顾客进入餐厅。

形式化地表示，对于 $i \in [n]$，我们令 $X_i = \mathbb{I}_{\text{第}i\text{ 个时间段有顾客进入}}$，则 $X_i \sim \texttt{Ber}(p)$，并且 X_i 相互独立。设

$$Z_n = \sum_{i=1}^{n} X_i, \quad \lambda = \mathbf{E}[Z_n] = p \cdot n.$$

现在我们计算顾客数量 Z_n 的分布。对于任何常数 $k \in \mathbb{N}$，有：

$$\begin{aligned}\mathbb{P}(Z_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

由于我们假设 k 和 λ 是常数，当 $n \to \infty$，上述表达式收敛于：

$$\mathbb{P}(Z_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

我们便把概率质量函数满足任意 $k \in \mathbb{N}$，$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ 的分布称为均值为 λ 的泊松分布，并把这个分布记作 $\texttt{Pois}(\lambda)$。

设 $X \sim \texttt{Pois}(\lambda)$。由于我们在上面是通过取极限的方式定义了泊松分布，因此需要验证它确实是一个分布，并且其均值确实为 λ：

- 验证分布性质

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=1.$$

因此它确实是一个分布。

- 验证均值

$$\mathbf{E}[X]=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\lambda.$$

两个独立的满足泊松分布的随机变量之和依旧满足泊松分布。

命题 15.1

假设 $X_1\sim\texttt{Pois}(\lambda_1)$ 并且 $X_2\sim\texttt{Pois}(\lambda_2)$ 是两个独立的随机变量。那么：

$$X_1+X_2\sim\texttt{Pois}(\lambda_1+\lambda_2).$$

证明 对于任意的 $n\geq 0$，

$$\begin{aligned}\mathbb{P}(X_1+X_2=n)&=\sum_{m=0}^{n}\mathbb{P}(X_1=m)\cdot\mathbb{P}(X_2=n-m)\\&=\sum_{m=0}^{n}\frac{\lambda_1^m}{m!}e^{-\lambda_1}\cdot\frac{\lambda_2^{n-m}}{(n-m)!}e^{-\lambda_2}\\&=e^{-(\lambda_1+\lambda_2)}\cdot\sum_{m=0}^{n}\binom{n}{m}\frac{\lambda_1^m\lambda_2^{n-m}}{n!}\\&=e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!}.\end{aligned}$$

这个结论可以推广到任意 n 个满足泊松分布的随机变量：

如果 $X_1,X_2,\ldots,X_n$ 是 n 个相互独立的随机变量，且 $X_i\sim\texttt{Pois}(\lambda_i)$，则：

$$\sum_{i=1}^{n}X_i\sim\texttt{Pois}\left(\sum_{i=1}^{n}\lambda_i\right).$$

15.2 泊松过程（Poisson Process）

15.2.1 泊松过程的定义

我们刚才说了，均值为 λ 的泊松分布可以用来描述单位时间内平均顾客数为 λ 人的时候，顾客人数的分布。如果我们统计一段时间内的顾客数量（例如从时间 t_1 到时间 t_2），假设 t_2-t_1 是整数，并且顾客的到来依然是均匀的话，那么按照我们上述的多个泊松分布变量之和的结论，这段时间内来的顾客人数应该符合 $\texttt{Pois}((t_2-t_1)\lambda)$ 分布。我们可以使用泊松过程来描述一段时间内到来的顾客人数。

我们说一族随机变量 $\{N(s)\}_{s\geq 0}$ 为速率为 λ 的泊松过程，当且仅当其满足以下条件：

1. $N(0)=0$；
2. 对于任意 $t,s\geq 0$，有：

$$N(t+s)-N(s)\sim\texttt{Pois}(\lambda t);$$

3. 对于任意 $t_0\leq t_1\leq\cdots\leq t_n$，随机变量：

$$N(t_1)-N(t_0),\ N(t_2)-N(t_1),\ \ldots,\ N(t_n)-N(t_{n-1})$$

相互独立。

实际上，满足条件的这样一族随机变量的概率空间长什么样子，这些随机变量作为概率空间上的可测函数为什么存在，如何构造，并不是一件容易的事情，是随机过程课需要讨论的内容。在我们这个课上，我们不妨假

设泊松过程是存在的，然后研究它的性质。

我们可以从另外一个角度来描述一个泊松过程，也就是考虑相邻两个顾客到达的间隔时间。事实上，对于一个速率为 λ 的泊松过程，两名顾客之间的间隔时间满足速率为 λ 的指数分布 $\texttt{Exp}(\lambda)$。为了说明这一点，我们先来考虑指数分布的一些性质。

15.2.2 指数分布

回忆到我们曾经说过，速率为 $\lambda > 0$ 的指数分布的概率密度函数为：

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

它的累积分布函数为：

$$F(t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}.$$

指数分布可以用来建模顾客到来的时间。比如说，速率为 λ 的指数分布可以用来表示，在一个速率为 λ 的泊松过程里，从 0 时刻开始，第一个顾客在 t 时刻还未到来的概率是 $1 - F(t) = e^{-\lambda t}$。我们今天主要的目标之一便是严格的证明这件事情。

我们之前计算过指数分布的概率矩生成函数为 $\forall \theta < \lambda, M_X(\theta) = \frac{\lambda}{\lambda - \theta}$。因此可以容易的计算出，对于 $X \sim \texttt{Exp}(\lambda)$，我们有

$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \mathbf{Var}[X] = \frac{1}{\lambda^2}.$$

指数分布的一个重要性质就是所谓的“无记忆性”，直观上来说，就是“假设已经等了第一个顾客 s 时间之后，他还没来，那此时还需等待的他到来时间的分布与一开始他到来时间的分布相同”。

命题 15.2 (指数分布的无记忆性)

设 $X \sim \text{Exp}(\lambda)$。那么对于任意 $t, s > 0$，有

$$\mathbb{P}(X > t + s | X > s) = \mathbb{P}(X > t).$$

这个性质用定义可以简单证明，但是它却意义深远…

证明 我们有

$$\begin{aligned} \mathbb{P}(X > t + s | X > s) &= \frac{\mathbb{P}(X > t + s,\ X > s)}{\mathbb{P}(X > s)} \\ &= \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= \mathbb{P}(X > t). \end{aligned}$$

指数分布另外一个有趣的性质是所谓“指数竞赛”。假设有两家商店，第一家商店第一位顾客到来时间服从 $\texttt{Exp}(\lambda_1)$ 分布，第二家商店第一位顾客到来时间服从 $\texttt{Exp}(\lambda_2)$ 分布。那么，这两家店同时开门后，第一位顾客的到来时间符合什么分布？答案是 $\texttt{Exp}(\lambda_1 + \lambda_2)$。

命题 15.3 (指数竞赛)

设 $X_1 \sim \texttt{Exp}(\lambda_1)$，$X_2 \sim \texttt{Exp}(\lambda_2)$ 为两个独立的随机变量。那么 $Y := X_1 \wedge X_2 \sim \texttt{Exp}(\lambda_1 + \lambda_2)$。

证明

$$\mathbb{P}(Y > t) = \mathbb{P}(X_1 > t \wedge X_2 > t) = \mathbb{P}(X_1 > t)\,\mathbb{P}(X_2 > t) = e^{-(\lambda_1 + \lambda_2)t}.$$

这个性质可以自然推广到 n 个独立指数分布，即如果 $\forall i \in [n]$，$X_i \sim \texttt{Exp}(\lambda_i)$，那么 $\min\{X_1, \ldots, X_n\} \sim \texttt{Exp}(\lambda_1 + \cdots + \lambda_n)$。

我们接下来考虑“谁赢了竞赛”的问题，也就是说 $X_1 < X_2$ 的概率。我们用 f_λ 来表示 $\texttt{Exp}(\lambda)$ 的概率密度函数，那么使用全概率公式，我们有

$$\begin{aligned}\mathbb{P}(X_1 < X_2) &= \int_0^\infty f_{\lambda_1}(t) \cdot \mathbb{P}(X_2 > t)\,\mathrm{d}t \\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 t} e^{-\lambda_2 t}\,\mathrm{d}t \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}.\end{aligned}$$

这个结论同样可以自然推广到 n 个指数分布的场合：X_i 是 $X_1, \ldots, X_n$ 中最小的概率为 $\frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$。

15.2.3 泊松过程的指数分布刻画

我们接下来证明一个重要的结论，它说明我们一开始定义的泊松过程可以用指数分布的间隔来描述。

定理 15.1

设 $\tau_1, \tau_2, \ldots, \tau_n$ 为一系列独立的指数分布随机变量，满足 $\tau_i \sim \texttt{Exp}(\lambda)$。设 $T_n = \sum_{i=1}^n \tau_i$。对于 $t \geq 0$，定义 $N(t) := \max\{n \mid T_n \leq t\}$。那么 $\{N(t)\}_{t\geq 0}$ 是一个泊松过程。♡

根据定理描述，我们可以想象 τ_i 就是第 $i-1$ 个顾客和第 i 个顾客到达的时间间隔。T_n 就表示第 n 个顾客到达的时间。那么 $N(t)$ 就表示 t 时刻之前到达了多少顾客。

证明 我们在上次的作业里已经知道，T_n 满足所谓的 Gamma 分布 $\Gamma(n, \lambda)$，其概率密度为

$$g_n(t) = \begin{cases} \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{n-1}}{(n-1)!}, & t \geq 0; \\ 0, & t < 0. \end{cases}$$

于是，我们便可以使用全概率公式来计算 $N(t)$ 的分布。

$$\begin{aligned}\mathbb{P}(N(t) = n) &= \mathbb{P}(T_n \geq t \wedge T_{n+1} > t) \\ &= \int_0^t g_n(s) \cdot \mathbb{P}(\tau_{n+1} > t - s)\,\mathrm{d}s \\ &= \int_0^t \lambda e^{-\lambda s} \cdot \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot e^{-\lambda(t-s)}\,\mathrm{d}s \\ &= \lambda^n e^{-\lambda t} \cdot \frac{t^n}{n!}.\end{aligned}$$

因此，$N(t) \sim \texttt{Pois}(\lambda t)$。我们接着来一一验证 $\{N(t)\}_{t\geq 0}$ 满足泊松过程定义的三个条件。

首先，$N(0) = 0$ 是显然的。根据指数分布的无记忆性，我们知道 $N(s+t) - N(s)$ 的分布和 $N(t) - N(0)$ 一样是 $\texttt{Pois}(\lambda t)$，因此第二个条件也满足。我们同样可以使用指数分布的无记忆性验证第三个独立性条件。

15.2.4 泊松过程的稀疏化（Thinning）

在我们用泊松过程建模顾客到店的例子里，我们可以分别考虑男顾客和女顾客。假设男女比是 1:1，我们可以用下面的方式等价的建模：对于每一个到达的顾客，投掷一个公平的硬币，如果是正面，这位顾客就是男顾客，如果是反面，这位顾客就是女顾客。泊松过程的稀疏化定理就是说，男女顾客人数分别为一个满足 $\texttt{Pois}(\frac{\lambda}{2})$ 的泊松过程，并且这两个过程是**独立**的。

严格的说，给定一个泊松过程 $\{N(t)\}_{t\geq 0}$，我们可以给第 i 位到达的顾客引入一个独立同分布的随机变量 Y_i，用来表示该顾客的种类，从而把泊松过程划分成若干个子过程。我们假设 Y_i 的取值是有限的非负整数，并且设 $p_k = \mathbb{P}(Y_i = k)$ 为其概率质量函数。对于每一个 Y_i 可能的取值 k，我们用 $N_k(t)$ 表示截止时间 t 时具有种类 k 顾客的到达数量。那么 $\{N_k(t)\}_{t\geq 0}$ 被称为泊松过程的一个“稀疏化”。我们有如下一个有用且令人惊讶的命题。

命题 15.4

对于每个 k，$\{N_k(t)\}_{t\geq 0}$ 是一个参数为 λp_k 的泊松过程。此外，所有过程的集合

$$\left\{\{N_k(t)\}_{t\geq 0} : k \in \mathtt{Im}(Y_1)\right\}$$

是相互独立的。

注意到，我们说两个随机过程 $N_1(t)$ 和 $N_2(t)$ 独立，指的是对于任意 $t_0 \leq t_1 \leq \cdots \leq t_n$，随机变量 $\bigcup_{i\in[n]}\{N_1(t_i)-N_1(t_{i-1}), N_2(t_i)-N_2(t_{i-1})\}$ 是相互独立的。

我觉得这个命题的结论是比较让人意外的。举一个栗子，假设进入餐厅的顾客是一个速率为 1 的泊松过程，并且每位顾客是男性或女性的概率分别为 $\frac{1}{2}$。实际上，我们可以假设通过抛硬币来决定到来的顾客是男性还是女性。直观上来看，如果在一个小时内到来的男性数量很大（如 50 人），这可能意味着业务量很大，从而女性顾客的数量也会比较多。然而，这个命题告诉我们，男性顾客和女性顾客的数量是独立的。

证明 我们只需要证明 Y_i 取两种值的情况即可，多种值的情况可以类似证明。假设 $Y_i \in \{0,1\}$。由于泊松分布本身的独立增量性质，为了验证独立性，我们只需要验证对任何 $s,t\geq 0$，$N_0(t+s)-N_0(s)$ 和 $N_1(t+s)-N_1(s)$ 独立即可（why）。我们下面计算 $N_0(t)$ 和 $N_1(t)=n$ 的联合分布。这验证了每一个 $N_i(t)$ 都是泊松过程。由泊松过程的无记忆性，便可以得到我们的独立性结果。

$$\begin{aligned}
\mathbb{P}\left(N_0(t)=m \wedge N_1(t)=n\right) &= \mathbb{P}\left(N_0(t)=m \wedge N(t)=m+n\right)\\
&= \mathbb{P}\left(N(t)=m+n\right)\cdot \mathbb{P}\left(N_0(t)=m \mid N(t)=m+n\right)\\
&= \frac{e^{-\lambda t}(\lambda t)^{m+n}}{(m+n)!}\cdot \binom{m+n}{m} p_0^m p_1^n\\
&= \frac{e^{-\lambda p_0 t}(\lambda p_0 t)^m}{m!}\cdot \frac{e^{-\lambda p_1 t}(\lambda p_1 t)^n}{n!}.
\end{aligned}$$

15.2.5 泊松过程的一个应用

我们来看泊松过程最大似然估计（maximum likelihood estimation）的一个应用。

假设有两位编辑在阅读一本 300 页的书。编辑 A 在书中发现了 100 个错别字，编辑 B 发现了 120 个错别字，其中 80 个是两人都发现的。

假设作者的错别字遵循一个参数为 λ 的泊松过程，每页的错误率未知。两位编辑分别以未知的成功概率 p_A 和 p_B 抓到错别字。我们想知道总共有多少错别字。这可以通过估计 λ、p_A 和 p_B 来解决这个问题。

显然，总共有四种类型的错别字：

1. **类型 1**：未被两位编辑发现的错别字。发生概率为 $p_1=(1-p_A)(1-p_B)$。
2. **类型 2**：仅被编辑 A 发现的错别字。发生概率为 $p_2=p_A(1-p_B)$。
3. **类型 3**：仅被编辑 B 发现的错别字。发生概率为 $p_3=(1-p_A)p_B$。
4. **类型 4**：被两位编辑都发现的错别字。发生概率为 $p_4=p_Ap_B$。

因此，每种类型的错别字发生过程是一个独立的泊松过程，参数为 λp_i。令 N_1,N_2,N_3,N_4 分别表示书中对应类型的错别字数量，则 $N_i \sim \mathtt{Pois}(300\lambda p_i)$。

书中有 20 个类型 2 的错别字，40 个类型 3 的错别字，80 个类型 4 的错别字。我们声称最可能的参数值满足方程组：

$$\begin{cases}
300\lambda p_A(1-p_B)=20,\\
300\lambda(1-p_A)p_B=40,\\
300\lambda p_Ap_B=80.
\end{cases}$$

换句话说，我们在得到一个 $X\sim\mathtt{Pois}(\mu)$ 的样本 x 之后，认为最有可能的 μ 的值就是 x。这是泊松分布的

最大似然估计，我们马上进行验证。假设承认这件事情，我们可以马上解得

$$\lambda = \frac{1}{2}, \quad p_A = \frac{2}{3}, \quad p_B = \frac{4}{5}.$$

我们最后证明这个最大似然估计。假设 $N \sim \texttt{Pois}(\lambda)$，其中 λ 未知。给定 $N = n$，我们的目标是找到：

$$\arg\max_{\lambda} \mathbb{P}(N = n) \text{ provided } N \sim \texttt{Pois}(\lambda).$$

注意到在已知 $N \sim \texttt{Pois}(\lambda)$ 时，$\mathbb{P}(N = n) = \frac{e^{-\lambda}\lambda^n}{n!}$，且 $\log \mathbb{P}(N = n) = -\lambda + n \log \lambda - \log n!$。

因此，最大化的目标等价于：

$$\arg\max_{\lambda} \{-\lambda + n \log \lambda\}.$$

对上式求导并令其为 0，得到 $\lambda = n$。

第 16 章 使用泊松做近似

16.1 非均匀奖券收集（Coupon Collector）问题

我们以前讲过奖券收集问题。

考虑玩一个抽卡手游。现在总共有 n 种不同类型的卡，每一抽可以均匀的得到其中一种。现在想问平均要抽多少次，可以集齐一套，即 n 种卡每种至少一张。

如果每一次抽卡每一种类型的卡出现的概率都是等概率 $\frac{1}{n}$，那么我们期望需要抽 nH_n 次才能收集到所有类型的卡，其中 $H_n = \sum_{k=1}^{n} \frac{1}{k} \xrightarrow{n\to\infty} n\log n + \gamma n$ 是调和级数。

在实际中，手游公司往往会对每一种卡有稀有度的设定，比如，对于每一抽，第 i 种卡被开出来的概率是 p_i （$\sum_{i=1}^{n} p_i = 1$ ）。那么，在这样一种设定下，集齐一套平均要抽多少次呢？稍微想一下就会明白，因为现在每种卡不再有对称性，我们之前基于期望的线性性的简单技巧不再有效了。

令 N_i 表示第一次获得第 i 种卡片需要抽卡的次数，那么 N_i 服从参数为 p_i 的几何分布。令 N 表示收集到所有 n 种类型卡片所需的抽卡次数，那么 $N = \max_{i\in[n]} N_i$。

我们问题便是计算 $\mathbf{E}[N]$。然而，由于 N_i 之间不是独立的，直接计算 $\mathbf{E}[N]$ 并不容易。

16.1.1 泊松抽卡法

我们可以脑补如下一个抽卡的方式，和我们要研究的问题是等价的:

玩家在一个线下的商店柜台购买卡包进行抽卡。每一分钟过来一位顾客，进行一次抽卡。现在问平均第几位顾客（也就是第几分钟）抽完之后，前面所有的顾客抽的卡放在一起能够集齐所有 n 种。

我们现在稍稍修改上面这个场景，假设柜台的顾客并不是严格的每一分钟过来一位，而是按照速率为 1 的泊松过程过来。同样，每一位过来的顾客随机抽一张卡。我们同样考虑当所有顾客抽的卡放在一起集齐全套的时间 T。注意这里 $T \in \mathbb{R}$ 是一个实数。

回忆我们在上一讲讨论过的泊松过程的稀疏化（thinning）。令 $X_i(t)$ 表示在时间区间 $[0,t]$ 内，通过这个泊松过程收集到的类型 i 卡片的数量。那么 $\{X_i(t)\}_{t\geq 0}$ 是一个速率为 p_i 的泊松过程，并且所有这些 $X_i(t), i\in[n]$ 是相互独立的。换句话说，我们看那些抽到了第 i 种卡片的人的队伍，它们的人数是各自独立的泊松过程!

对于 $i\in[n]$，令 $T_i = \min\{t \mid X_i(t) = 1\}$ 表示第一次开出类型 i 的卡片的时间。显然 T_i 的分布是 $\mathtt{Exp}(p_i)$，并且 $T = \max_{i\in[n]} T_i$。

于是，由于独立性，我们可以计算

$$\mathbf{E}[T] = \int_0^\infty \mathbb{P}(T \geq t)\,\mathrm{d}t = \int_0^\infty \left(1 - \prod_{i=1}^{n}(1-e^{-p_i t})\right)\mathrm{d}t.$$

16.1.2 泊松抽卡法和标准抽卡法的比较

上面的计算可以看出，由于泊松抽卡法在稀疏化后有非常神奇的独立性质，我们可以方便的计算收集完全套的时间。直观上来说，由于是速率为 1 的泊松过程，平均一分钟抽一张卡，所以，从平均的意义上来看，$\mathbf{E}[T]$ 很有可能和 $\mathbf{E}[N]$，即固定一分钟抽一张卡的平均集齐时间是很接近的。接下来，我们通过**耦合（coupling）**的方法说明事实上 $\mathbf{E}[T]=\mathbf{E}[N]$。

想象现在有两个柜台，壹号柜台是固定一分钟来一个顾客抽卡，贰号柜台是按照速率为 1 的泊松过程到来顾客抽卡。我们想象，两个柜台的各自第 i 位顾客总是抽到同样的卡片（这种定义联合分布的思想实验便叫做耦合）。显然，假设壹号柜台上第 N 位顾客抽完卡后集齐了一套（N 是随机变量），那么在贰号柜台上，也是第 N 位顾客抽完卡后集齐一套。如果我们用 τ_i 表示贰号柜台第 $i-1$ 位顾客和第 i 位顾客到达的间隔时间，那么 T 和 $\sum_{i=1}^{N}\tau_i$ 有同样的分布。于是，

$$\mathbf{E}[T]=\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right].$$

由于 $\tau_i\sim\mathrm{Exp}(1)$，所以 $\mathbf{E}[\tau_i]=1$。在上面的式子里，如果 $N=n$ 是一个常数，那么就有期望的线性性 $\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\sum_{i=1}^{n}\mathbf{E}[\tau_i]=n$ 成立。但是，我们这里 N 是一个随机变量，在最一般的情况下，期望和求和是不一定可以交换的。但在我们这儿，N 和 $\{\tau_i\}$ 是独立的，Wald's Equation 可以保证这儿交换是成立的。对于最一般的 Wald's equation，我们要学习了鞅相关的知识后才能够比较方便的证明。但在现在这个特殊情况，我们可以直接使用定义证明。

命题 16.1

在我们上述例子里

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\mathbf{E}[N].$$

♠

证明

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\mathbf{E}\left[\sum_{i=1}^{\infty}\tau_i\cdot\mathbb{I}_{[i\le N]}\right]\overset{\text{(Fubini)}}{=}\sum_{i=1}^{\infty}\mathbf{E}\left[\tau_i\cdot\mathbb{I}_{[i\le N]}\right].$$

由于 τ_i 和 $[i\le N]$ 独立，所以 $\mathbf{E}\left[\tau_i\cdot\mathbb{I}_{[i\le N]}\right]=\mathbf{E}[\tau_i]\cdot\mathbb{P}(i\le N)$。于是

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\sum_{i=1}^{\infty}\mathbb{P}(N\ge i)=\mathbf{E}[N].$$

这便说明了，对于非均匀的奖券收集问题，平均集齐一套的时间

$$\mathbf{E}[N]=\mathbf{E}[T]=\int_0^{\infty}\left(1-\prod_{i=1}^{n}(1-e^{-p_it})\right)\mathrm{d}t.$$

上述式子看起来比较吓人，我们接着进行一个合理性检查，即对每一个 $p_i=\frac{1}{n}$ 的时候计算一下这个积分。

于是乎，

$$\begin{aligned}\mathbf{E}[N] &= \int_0^\infty 1-\prod_{i=1}^n\left(1-e^{-\frac{t}{n}}\right)\mathrm{d}t\\ &\overset{(x=e^{-\frac{t}{n}})}{=} -n\int_0^1 1-(1-x)^n\,\mathrm{d}\log x\\ &= -n\int_0^1 \frac{1}{x}-\frac{(1-x)^n}{x}\,\mathrm{d}x\\ &= -n\int_0^1\sum_{k=1}^n\frac{(1-x)^{k-1}}{x}-\frac{(1-x)^k}{x}\,\mathrm{d}x\\ &\overset{\text{(Fubini)}}{=} -n\sum_{k=1}^n\int_0^1(1-x)^{k-1}\,\mathrm{d}x\\ &= n\sum_{k=1}^n\frac{1}{k}=nH_n.\end{aligned}$$

神奇！

16.2 泊松近似

假设将 m 个相同的球随机投放到 n 个箱子中的随机试验。这个实验叫做球与箱子（Balls-into-Bins）模型。对任意 $i\in[n]$，令 X_i 表示第 i 个箱子中的球的数量。如果投放是均匀随机的，那么我们有 $X_i\sim\texttt{Binom}\left(m,\frac{1}{n}\right)$ 且 $\mathbb{E}[X_i]=\frac{m}{n}$。

这个模型可以用来建模很多概率问题，比如我们刚刚讨论过的奖券收集问题，以及生日悖论（birthday paradox）等等。在计算机科学中，它也很自然的被用来建模随机映射的哈希表。为了理解哈希表中的冲突，我们自然会关注 $\max_{i\in[n]}X_i$ 的值，这个被称为最大负载（maxload）。然而，最大负载 $\max_{i\in[n]}X_i$ 并不是一个特别容易计算的量，原因在于 X_i 之间不是相互独立的。然而，我们可以使用泊松分布来近似计算它。我们将要发展一个研究 Balls-into-Bins 问题的很一般化的工具。

定理 16.1

设 $\forall i\in[n],Y_i\sim\texttt{Pois}(\lambda)$ 是一组独立的泊松分布，其中 $\lambda>0$ 为任意固定常数。那么，在 $\sum_{i=1}^n Y_i=m$ 的条件下，$(Y_1,\ldots,Y_n)$ 和 $(X_1,\ldots X_n)$ 具有相同的分布。♡

换句话说，对于任何 $a_1,\ldots,a_n$，

$$\mathbb{P}\left((Y_1,\ldots,Y_n)=(a_1,\ldots,a_n)\,\middle|\,\sum_{i=1}^n Y_i=m\right)=\mathbb{P}\left((X_1,\ldots,X_n)=(a_1,\ldots,a_n)\right).$$

证明 对于任意给定的 $(a_1,\ldots,a_n)\in\mathbb{N}^n$ 满足 $\sum_{i=1}^n a_i=m$，我们有

$$\mathbb{P}\left((X_1,\ldots,X_n)=(a_1,\ldots,a_n)\right)=\frac{1}{n^m}\cdot\frac{m!}{a_1!a_2!\cdots a_n!}.$$

另一方面，

$$\begin{aligned}&\mathbb{P}\left((Y_1,\ldots,Y_n)=(a_1,\ldots,a_n)\,\middle|\,\sum_{i=1}^n Y_i=m\right)\\ &=\frac{\mathbb{P}\left((Y_1,\ldots,Y_n)=(a_1,\ldots,a_n)\right)}{\mathbb{P}\left(\sum_{i=1}^n Y_i=m\right)}\\ &=\frac{\prod_{i=1}^n\mathbb{P}\left(Y_i=a_i\right)}{\mathbb{P}\left(\sum_{i=1}^n Y_i=m\right)}\\ &=\frac{\prod_{i=1}^n e^{-\lambda}\frac{\lambda^{a_i}}{a_i!}}{e^{-\lambda n}\frac{(\lambda n)^m}{m!}}=\frac{1}{n^m}\cdot\frac{m!}{a_1!a_2!\cdots a_n!}.\end{aligned}$$

上面这个等分布的结论说明，当我们想计算 $(X_1,\dots,X_n)$ 的某个性质的时候，我们可以转而计算独立的 $(Y_1,\dots,Y_n)$ 的性质。当然了，我们等分布的结论需要 $\sum_i Y_i = m$ 这个条件，因此处理的方法便是把所有不满足这个条件的贡献全部扔掉。

推论 16.1

设 $f:\mathbb{N}^n\to\mathbb{N}$ 是可测函数，且 $Y_1,Y_2,\dots,Y_n$ 是独立的泊松随机变量，其速率为 $\lambda=\frac{m}{n}$。则有：

$$\mathbf{E}\left[f(X_1,X_2,\dots,X_n)\right]\le e\sqrt{m}\cdot\mathbf{E}\left[f(Y_1,Y_2,\dots,Y_n)\right].$$

我们有时候把这个不等式称作泊松近似公式。

证明 根据全期望公式，我们有：

$$\mathbf{E}\left[f(Y_1,\dots,Y_n)\right]=\sum_{k=0}^{\infty}\mathbf{E}\left[f(Y_1,\dots,Y_n)\,\middle|\,\sum_{i=1}^n Y_i=k\right]\cdot\mathbb{P}\left(\sum_{i=1}^n Y_i=k\right).$$

由于 f 是非负函数，我们扔掉所有 $k\neq m$ 的项，可以得到

$$\mathbf{E}\left[f(Y_1,\dots,Y_n)\right]\ge\mathbf{E}\left[f(Y_1,\dots,Y_n)\,\middle|\,\sum_{i=1}^n Y_i=m\right]\cdot\mathbb{P}\left(\sum_{i=1}^n Y_i=m\right).$$

我们知道，假设 $Y_i\sim\texttt{Pois}(\lambda)$，则 $\sum_{i=1}^n Y_i\sim\texttt{Pois}(\lambda n)$，并且上述等式对于任意 λ 均成立。我们希望 $\mathbb{P}\left(\sum_{i=1}^n Y_i=m\right)$ 尽量的大，根据我们计算过的泊松分布的最大似然原理，我们取 $\lambda=\frac{m}{n}$，于是根据 Stirling 公式（需要对常数进行仔细的讨论），有

$$\mathbb{P}\left(\sum_{i=1}^n Y_i=m\right)=e^{-m}\frac{m^m}{m!}>\frac{1}{e\sqrt{m}}.$$

所以

$$\mathbf{E}\left[f(Y_1,\dots,Y_n)\right]\ge\frac{1}{e\sqrt{m}}\mathbf{E}\left[f(Y_1,\dots,Y_n)\,\middle|\,\sum_{i=1}^n Y_i=m\right]=\frac{1}{e\sqrt{m}}\mathbf{E}\left[f(X_1,\dots,X_n)\right].$$

16.2.1 最大负载

我们现在研究在 $m=n$ 时候的最大负载问题。这个问题若干年前在水源上有人问过。我们将证明，$m=n$ 时，最大负载 $X=\max_{i\in[n]}X_i$ 以 $1-o(1)$ 的概率，满足

$$X=\Theta\left(\frac{\log n}{\log\log n}\right).$$

首先证明上界，即存在常数 $c_1>0$，使得 $\mathbb{P}\left(X\ge c_1\frac{\log n}{\log\log n}\right)=o(1)$。令 $k=\frac{c_1\log n}{\log\log n}$。通过 union-bound，我们有：

$$\mathbb{P}\left(X\ge k\right)=\mathbb{P}\left(\exists i\in[n],X_i\ge k\right)\le\sum_{i=1}^n\mathbb{P}\left(X_i\ge k\right)=n\cdot\mathbb{P}\left(X_1\ge k\right).$$

再次使用 union-bound，可以得到

$$\mathbb{P}\left(X\ge k\right)\le n\cdot\binom{n}{k}n^{-k}\le n\left(\frac{e}{k}\right)^k.$$

注意到

$$k\log k=\frac{c_1\log n}{\log\log n}\left(\log\log n-\log\log\log n+\log c_1\right)$$

取 $c_1=6$，我们有

$$\log n+k-k\log k<-\log n.$$

于是，$\mathbb{P}\left(X\ge k\right)\le n\left(\frac{e}{k}\right)^k<\frac{1}{n}=o(1)$。

我们接着使用泊松近似公式证明下界，即存在常数 c_2，使得：

$$\mathbb{P}\left(X \leq \frac{c_2 \log n}{\log\log n}\right) = o(1).$$

设 $h = \frac{c_2 \log n}{\log\log n}$。我们定义函数 $f(X_1, \ldots, X_n) := \mathbb{I}_{[X \leq h]} = \mathbb{I}_{[\max_i X_i \leq h]}$。于是，根据泊松近似公式

$$\begin{aligned}\mathbb{P}(X \leq h) &= \mathbf{E}[f(X_1, \ldots, X_n)] \\ &\leq e\sqrt{n}\mathbf{E}[f(Y_1, \ldots, Y_n)] \\ &= e\sqrt{n} \cdot \mathbb{P}\left(\max_{i\in[n]} Y_i \leq h\right).\end{aligned}$$

根据 Y_i 的定义，我们有

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[n]} Y_i \leq h\right) &= (\mathbb{P}(Y_1 \leq h))^n = (1 - \mathbb{P}(Y_1 > h))^n \\ &\leq (1 - \mathbb{P}(Y_1 = h+1))^n = \left(1 - \frac{1}{(h+1)!e}\right)^n \leq e^{-\frac{n}{e(h+1)!}}.\end{aligned}$$

注意到

$$\begin{aligned}\log(h+1)! &= \sum_{i=1}^{h+1} \log i < \int_1^{h+2} \log x \,\mathrm{d}x \\ &= (h+2)\log(h+2) - h - 1 \leq (h+2)\log h - h + 3 \\ &= \frac{c_2 \log n + 2\log\log n}{\log\log n}(\log\log n - \log\log\log n + \log c_2) - \frac{c_2 \log n}{\log\log n} + 3 \\ &\leq c_2 \log n - \log\log n - 2.\end{aligned}$$

设 $c_2 = 1$, 我们有 $\log(h+1)! \leq \log n - \log\log n - 2$。因此

$$e(h+1)! \leq \frac{n}{e\log n}.$$

所以

$$\mathbb{P}\left(\max_{i\in[n]} Y_i \leq h\right) \leq e^{-\frac{n}{(h+1)!e}} \leq e^{-e\log n} = n^{-e}.$$

第 17 章 一些经典的概率问题

我们今天使用学过的工具来解决一些经典的概率问题。这一部分的内容主要来自 William Feller 的名著 概率论及其应用。

17.1 等待时间悖论

假设公交车以 Pois(1) 的方式到达车站，周而复始，日夜无休。现在你在 12:00 （或者任何一个固定时间）到达车站。你平均需要等待多久才能等到下一辆公交车？

有两个自然的答案：

- 由于两辆公交车之间的间隔服从 Exp(1)，而指数分布具有无记忆性。因此，平均等待时间应该是 1。
- 由于两辆公交车之间的间隔平均长度为 1，而平均来说会在间隔中点到达。因此，平均等待时间应该是 $\frac{1}{2}$。

哪个答案是正确的？

正确答案是 1。第二个直观错在哪呢？“在间隔中点到达”的观察是正确的，但较大的间隔发生的概率更高。因此，你遇到的间隔的平均长度实际上大于 1。为了严格说明这一点，我们将计算等待时间的累积分布函数，然后确定它的期望值。

首先，定义一些符号。我们用 S_i 表示第 i 次公交车到达的时间，用 τ_i 表示相邻公交车之间的间隔，其中 $\tau_i = S_i - S_{i-1}$ 服从指数分布 Exp(1)。假设我们到达车站的时间远晚于泊松过程的开始时间，用 L_t 表示包含时间 t 的间隔的长度。回忆一下，S_n 服从伽马分布 $\Gamma(n, \lambda)$，其概率密度函数为：

$$g_n(t) = \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{n-1}}{(n-1)!}, \text{ 对} t \geq 0.$$

现在我们计算 $\mathbb{P}(L_t \leq x)$，分为两种情况：

1. 如果 $x \leq t$，这意味着一些公交车必须在时间 t 之前到达。我们将枚举在 t 之前到达的公交车数量，并相应

地计算概率。

$$\begin{aligned}
\mathbb{P}(L_t \le x) &= \sum_{n=1}^{\infty} \int_0^t g_n(y) \cdot \mathbb{P}(t-y < \tau_{n+1} \le x)\,\mathrm{d}y \\
&= \sum_{n=1}^{\infty} \int_{t-x}^t g_n(y) \cdot \mathbb{P}(t-y < \tau_{n+1} \le x)\,\mathrm{d}y \qquad (y > t-x) \\
&= \sum_{n=1}^{\infty} \int_{t-x}^t e^{-y} \frac{y^{n-1}}{(n-1)!} \cdot \left(e^{y-t} - e^{-x}\right)\mathrm{d}y \\
&= \int_{t-x}^t \sum_{n=1}^{\infty} e^{-y} \frac{y^{n-1}}{(n-1)!} \cdot \left(e^{y-t} - e^{-x}\right)\mathrm{d}y \qquad \text{(Fubini-Tonelli)} \\
&= \int_{t-x}^t \left(e^{y-t} - e^{-x}\right)\mathrm{d}y = 1 - e^{-x} - xe^{-x}.
\end{aligned}$$

2. 如果 $x > t$，我们需要单独处理 $\tau_1 > t$ 的情况，其余部分与前一种情况类似。

$$\begin{aligned}
\mathbb{P}(L_t \le x) &= \mathbb{P}(t < \tau_1 \le x) + \sum_{n=1}^{\infty} \int_0^t g_n(y) \cdot \mathbb{P}(t-y < \tau_{n+1} \le x)\,\mathrm{d}y \\
&= e^{-t} - e^{-x} + \int_0^t e^{y-t} - e^{-x}\mathrm{d}y \\
&= e^{-t} - e^{-x} + 1 - e^{-t} - te^{-x} = 1 - (1+t)e^{-x}.
\end{aligned}$$

因此，L_t 的概率密度函数为：

$$f(x) = \begin{cases} xe^{-x} & \text{if } x \le t, \\ (1+t)e^{-x} & \text{if } x > t. \end{cases}$$

尽管 t 会影响 L_t 的期望值，但对于足够大的 t，积分 $\int_t^{\infty}(1+t)xe^{-x}\,\mathrm{d}x$ 趋于 0。因此，当 t 趋于无穷大时，$\mathbf{E}[L_t]$ 收敛到：$\mathbf{E}[\lim_{t\to\infty} L_t] = \int_0^{\infty} x^2 e^{-x}\mathrm{d}x = 2$，这也表明下一辆公交车的平均等待时间为 $\frac{1}{2} \times 2 = 1$。

17.2 线段和圆环上的点

在 $[0,1]$ 线段上均匀的扔 n 个点 $X_1, X_2, \ldots, X_n$。我们把它们从小到大排列后叫做 $X^{(1)} \le X^{(2)} \le \cdots \le X^{(n)}$。这 n 个点把 $[0,1]$ 分成了 $n+1$ 个线段 $[0, X^{(1)}], [X^{(1)}, X^{(2)}], \ldots, [X^{(n-1)}, X^{[n]}], [X^{[n]}, 1]$。我们分别用 $L_1, L_2, \ldots, L_{n+1}$ 来表示这些线段长度。现在的问题是，对于 $i \in [n+1]$，L_i 是同分布的吗？

它们确实是同分布的。我们可以做这样一个思想实验。假设我们在周长为 1 的圆周上均匀的扔 $n+1$ 个点，这 $n+1$ 个点把圆周分成了 $n+1$ 段，根据对称性，这 $n+1$ 段的长度一定是同分布的。我们随便选一个点，把它标记为原点，把圆周从这里断开成一个 $[0,1]$ 线段，从它开始按照顺时针经过的点叫做 $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$。这样得到的 n 个点和我们在 $[0,1]$ 上均匀的扔 n 个点一定是同分布的，因此，他们划分的线段长度也是同分布的。

我们容易计算出 $\mathbb{P}(L_1 > t) = (1-t)^n$。由于所有的 L_i 都同分布，所以对于 $\forall i \in [n+1], \mathbb{P}(L_i > t) = (1-t)^n$。

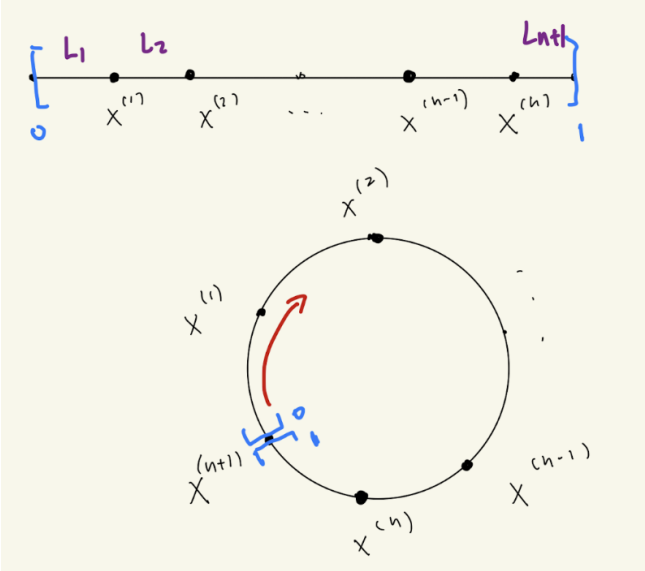
假设我们在圆环上均匀随机选两个点，这两个点把圆环分成了两个弧。很容易验证，每一条弧的平均长度是 $\frac{1}{2}$。我们现在先在圆环上事先画一个标记点，然后再随机选两个点，请问包含了标记点的那个弧平均有多长？现在，你应该有正确的直观了，这个弧长的平均值应该大于 $\frac{1}{2}$！

我们设这段弧长为 L，我们来计算一下它的分布函数。我们可以想象从标记点把圆环割开形成一个 $[0,1]$ 线段，那么不包含标记点的那一段就对应了我们一开始所定义的 L_2。因此

$$\mathbb{P}(L \le t) = \mathbb{P}(L_2 > 1-t) = t^2.$$

所以

$$\mathbf{E}[L] = \int_0^1 \mathbb{P}(L > t)\,\mathrm{d}t = \int_0^1 1 - t^2\,\mathrm{d}t = \frac{2}{3}.$$



17.3 随机分裂

假设有一个初始质量为 1 的物质，其每单位时间会独立的分裂一次。分裂的方式是变成两块，质量分别为原来的 p 和 $1-p$ 倍，其中 p 是 $[0,1]$ 中的一个均匀数字。那么，当分裂了 n 次之后，一块物质质量的分布是怎么样的呢?

我们用 Z_n 表示分裂了 n 次之后一块的质量。那么 $Z_0=1$ 并且 $Z_n=\prod_{i=1}^n X_i$，其中 X_i 为取值为 $[0,1]$ 的独立随机变量。我们可以看到，如果令 $Y_i=-\log X_i$，那么

$$\log Z_n=\sum_{i=1}^n \log X_i=-\sum_{i=1}^n Y_i$$

是 n 个独立的随机变量之和。我们现在来看 Y_i 的分布。

$$\mathbb{P}\left(Y_i\geq t\right)=\mathbb{P}\left(-\log X_i\geq t\right)=\mathbb{P}\left(X_i\leq e^{-t}\right)=e^{-t}.$$

因此，$Y\sim \texttt{Exp}(1)$。于是

$$\mathbb{P}\left(Z_n\leq t\right)=\mathbb{P}\left(\sum_{i=1}^n Y_n\geq -\log t\right)=1-G_n\left(\log\left(\frac{1}{t}\right)\right).$$

这儿 $G_n(x)=1-e^{-x}\left(\sum_{k=0}^{n-1}\frac{x^k}{k!}\right)$ 是速率为 1 的 $\Gamma(n,1)$ 分布的分布函数。

17.4 顺序统计量的分布

我们把 $[0,1]$ 上随机扔的 n 个点按照大小顺序排好序的随机变量 $X^{(1)},\ldots,X^{(n)}$ 称为顺序统计量。我们来研究它的分布。我们用 F_k 表示 $X^{(k)}$ 的分布函数，那么，简单思考之后可以看出来

$$F_k(t)=\mathbb{P}\left(X^{(k)}\leq t\right)=\sum_{j=k}^n\binom{n}{j}t^j(1-t)^{n-j}.$$

我们可以直接求导来得到 $X^{(k)}$ 的概率密度函数 f_k。但这儿我们直接从导数的定义来看可以更容易的得到 f_k 的表达式。

$$\begin{aligned} f_k(t) &= \lim_{h\to 0}\frac{1}{h}\cdot\mathbb{P}\left(X^{(k)}\in[t,t+h]\right)\\ &= \lim_{h\to 0}\frac{1}{h}\cdot n\binom{n-1}{k-1}h\cdot(t+O(h))^{k-1}\cdot(1-t+O(h))^{n-k}\\ &= n\binom{n-1}{k-1}t^{k-1}(1-t)^{n-k}. \end{aligned}$$

我们来考察 $nX^{(k)}$ 的分布，这等价于我们把 $[0,1]$ scale 成 $[0,n]$，并且 $\mathbf{E}\left[nX^{(k)}\right]=k\cdot\frac{n}{n+1}\overset{n\to\infty}{\to}k$.

$$\mathbb{P}\left(nX^{(k)}>t\right)=\mathbb{P}\left(X^{(k)}>\frac{t}{n}\right)=\sum_{j=0}^{k-1}\binom{n}{j}\left(\frac{t}{n}\right)^j\left(1-\frac{t}{n}\right)^{n-j}\overset{n\to\infty}{\to}\sum_{j=0}^{k-1}\frac{t^j}{j!}e^{-t}.$$

注意到，这是 Gamma 分布 $\Gamma(k,1)$ 的分布函数。换句话说，当 n 足够大的时候，$n\cdot X^{(k)}$ 的分布会收敛到 $\Gamma(k,1)$ 分布！

17.5 均匀分布的和与覆盖概率

我们现在考虑另外一个问题，假设 $X_1,\dots,X_n\sim\texttt{Unif}([0,a])$ 是 n 个独立随机变量，那么它们的和 $S_n=\sum_{i=1}^n X_i$ 的分布是什么样的呢？我们用 U_n 和 u_n 分别来表示 S_n 的分布函数和概率密度函数。我们可以使用全概率公式得到

$$\begin{aligned} u_1(x)&=\frac{1}{a},\quad x\in[0,a]\\ u_{n+1}(x)&=\frac{1}{a}\int_0^a u_n(x-y)\,\mathrm{d}y=\frac{1}{a}\left(U_n(x)-U_n(x-a)\right),\forall n\geq 1. \end{aligned}$$

因此，u_n 具有所谓“卷积”形式的递推式。为了方便叙述，对于 $x\in\mathbb{R}$，我们用 x_+ 来表示 $\max\{x,0\}$。那么，

$$U_1(x)=a^{-1}\cdot(x_+-(x-a)_+).$$

我们接着证明

定理 17.1

$$\begin{aligned} U_n(x)&=\frac{1}{a^n n!}\sum_{j=0}^{n}(-1)^j\binom{n}{j}\left((x-j\cdot a)_+\right)^n\\ u_{n+1}(x)&=\frac{1}{a^{n+1}n!}\sum_{j=0}^{n+1}(-1)^j\binom{n+1}{j}\left((x-j\cdot a)_+\right)^n. \end{aligned}$$

♡

既然表达式都告诉你了，我们使用归纳法进行验证即可。边界条件 $n=1$ 是容易验证的。对于更大的 n，使用归纳假设和 Fubini-Tonelli，我们有

$$\begin{aligned} U_n(x)&=\int_0^x u_n(t)\,\mathrm{d}t\\ &=\frac{1}{a^n(n-1)!}\sum_{j=0}^{n}(-1)^j\binom{n}{j}\int_0^x(t-j\cdot a)_+^{n-1}\,\mathrm{d}t\\ &=\frac{1}{a^n n!}\sum_{j=0}^{n}(-1)^j\binom{n}{j}(x-j\cdot a)_+^n. \end{aligned}$$

当 $j\notin\{0,1,\dots,n\}$ 时，按照惯例我们让 $\binom{n}{j}=0$。这样我们在上面求和的时候，可以让 j 的取值范围是所有的

整数。那么，我们有

$$\begin{aligned} u_{n+1}(x) &= a^{-1}\cdot(U_n(x)-U_n(x-a)) \\ &= \frac{1}{a^{n+1}n!}\sum_{j\in\mathbb{Z}}(-1)^j\binom{n}{j}\left((x-j\cdot a)_+^n-(x-(j+1)\cdot a)_+^n\right) \\ &= \frac{1}{a^{n+1}n!}\sum_{j\in\mathbb{Z}}(-1)^j\left(\binom{n}{j}+\binom{n}{j-1}\right)(x-j\cdot a)_+^n \\ &= \frac{1}{a^{n+1}n!}\sum_{j\in\mathbb{Z}}(-1)^j\binom{n+1}{j}(x-j\cdot a)_+^n. \end{aligned}$$

我们使用上面的结论研究一个有趣的问题：往一个圆环上均匀随机扔 n 个长度为 a 的圆弧，有多大概率能够盖住整个圆环？这个问题在水源上也有人问过。

我们用 $\varphi_n(t)$ 表示在长度为 t 的圆环上均匀扔 n 个长度为 a 的圆弧盖住圆环的概率。所谓 n 个圆弧盖住圆环，这等价于在 $[0,t]$ 的区间上扔 $n-1$ 个点，使得相邻两个点（包括两个顶点）的距离不大于 a。简单思考一下，我们可以写出递推式

$$\varphi_n(t)=(n-1)\int_0^a\varphi_{n-1}(t-x)\left(\frac{t-x}{t}\right)^{n-2}\cdot\frac{1}{t}\,\mathrm{d}x.$$

注意到 $\varphi_n(t)$ 的递推式也是“卷积”形式，和 u_n 有一点像。我们可以凑出 φ_n 和 u_n 的关系。实际上，我们可以验证，

$$\varphi_n(t)=a^n(n-1)!u_n(t)\frac{1}{t^{n-1}}.$$

第 18 章 随机变量收敛的模式，Borel-Cantelli 引理

18.1 随机变量的收敛（Convergence of Random Variables）

在概率论的研究里面，我们经常会讨论（定义在同一个概率空间上的）随机变量列 $\{X_n\}_{n\geq 1}$ 收敛到 X。我们之前已经遇到过所谓的“几乎必然”收敛 $X_n \overset{a.s.}{\to} X$。我们今天将介绍其它几种常见的收敛模式。这些收敛模式会出现在概率论研究的不同场合。比如说，在这门课未来学习大数定律的时候我们会接触“依概率”收敛，在未来学习随机过程的时候会遇到“依 L^2 ”收敛，以及我们在计算中常常用到的“依分布”收敛。我们现在先给出各自的定义，并研究互相之间的关系。

如果不特别说明，我们下面均设 $\{X_n\}_{n\geq 1}$ 和 X 为概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的随机变量。

18.1.1 几乎必然收敛 (almost surely convergence)

也叫做“以概率 1 收敛”，即存在一个可测集 $\Omega' \subseteq \Omega$ 满足 $\mathbb{P}(\Omega') = 1$，并且 $\forall \omega \in \Omega'$，$\lim_{n\to\infty} X_n(\omega) = X(\omega)$。我们可以等价的写作

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

我们一般记作 $X_n \overset{a.s.}{\to} X$.

18.1.2 依概率收敛（converge in probability）

指的是对于任何 $\varepsilon > 0$，

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

我们一般记作 $X_n \overset{P}{\to} X$.

18.1.3 依 L^p 收敛（converge in L^p ）

这里 $p \geq 1$ 是一个整数。它的定义是

$$\lim_{n\to\infty} \mathbf{E}\left[|X_n - X|^p\right] = 0.$$

我们一般记作 $X_n \overset{L^p}{\to} X$.

18.1.4 依分布收敛（converge in distribution）

依分布收敛和上述几种收敛模式不一样，它不要求这些随机变量生活在同一个概率空间中。我们可以假设 X_n 的分布函数是 F_n，X 的分布函数是 F。它的定义是，对于每一个 $F(x)$ 连续的点 x，有

$$\lim_{n\to\infty} F_n(x) = F(x).$$

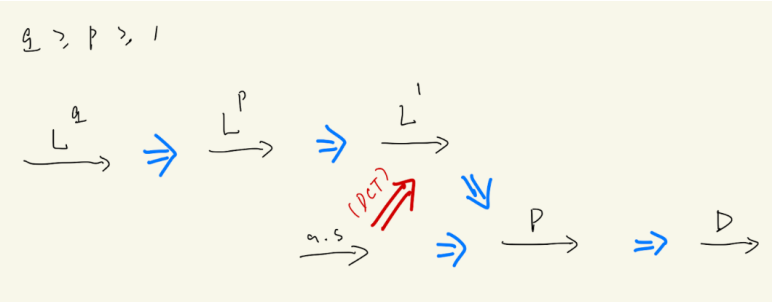
我们一般记作 $X_n \xrightarrow{D} X$.

18.2 收敛之间的关系

给出了这么多收敛的定义，小朋友一定有一堆疑问。为什么要有这么多不同种类的收敛？它们之间有什么关系？各自的直观又是什么？

对于第一个问题，我们在未来的学习中会逐渐体会到。但一个主要原因是，这些收敛有强有弱，我们关心的概率结论，在有的时候往往只能在比较弱的收敛意义下成立，或者在强的意义下成立需要更多的限制条件，或者更复杂的证明。我们现在来回答后两个问题。

我们假设 $q \geq p \geq 1$ 是整数。我们在下图中展示了各个收敛模式的关系。



18.2.1 $\xrightarrow{a.s}$ 与 $\xrightarrow{P}$

首先我们通过一个例子说明，$\xrightarrow{P}$ 不能推出 $\xrightarrow{a.s.}$。这个例子如图所示：设 X_n 和 X 都是定义在 $[0,1]$ 的均匀测度上的随机变量，其中 $X \equiv 0$。对于任意的 $n \geq 1$，我们设 $m = \lceil \log_2(n+1) \rceil - 1$。则我们知道 $n \in [2^m, 2^{m+1}-1]$。我们取 $k = n - (2^m - 1) \in [1, 2^m]$。我们定义

$$X_n = \begin{cases} 1, & \omega \in [(k-1)\cdot 2^{-m}, k\cdot 2^{-m}) \\ 0, & \text{otherwise.} \end{cases}$$

直观上说，我们让 X_n 在图中为红色的线段部份取值为 1，其余的部分为 0。令 $X = 0$。可以看出，红色部分是越来越少的，即 $X_n \neq X$ 的测度越来越小。即

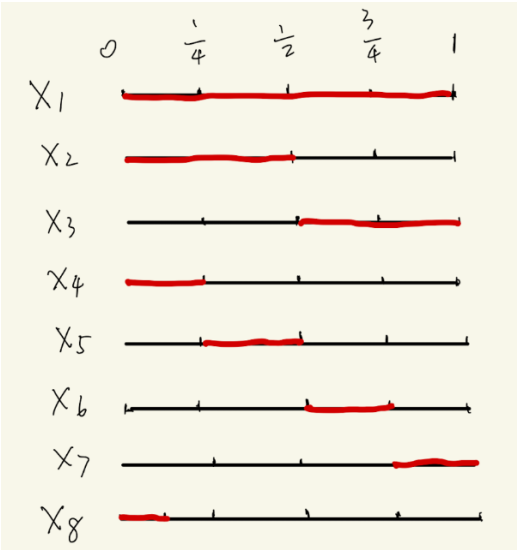
$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

但是，显然我们有 $X_n \xrightarrow{a.s} X$ 不成立。事实上，对于任意的 $\omega \in [0,1]$，我们都有 $X_n(\omega)$ 是不收敛的，因为红色的部分会无穷次的扫过 ω 这个点。

这个反例很好的展示了这两种收敛模式的区别：即 X_n 与 X 不同的地方测度尽管越来越小，但是这个位置是可以移动，这种移动阻止了几乎处处收敛。

我们接着来说明，$\xrightarrow{a.s.}$ 可以推出 $\xrightarrow{P}$。我们来证明，

$$X_n \xrightarrow{a.s} X \iff \forall \varepsilon > 0,\ \lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} |X_k - X| > \varepsilon\right) = 0.$$



由于 $|X_n - X| > \varepsilon \implies \sup_{k\geq n} |X_k - X| > \varepsilon$，所以说明了 $X_n \xrightarrow{P} X$。对于每一个 n，我们定义 $Z_n = \sup_{k\geq n} |X_k - X|$。注意到

$$\begin{aligned}
X_n \xrightarrow{a.s} X &\iff \exists \Omega' \subseteq \Omega \text{ s.t. } \mathbb{P}(\Omega') = 1 \text{ and } \forall \omega \in \Omega',\ X_n(\omega) \to X(\omega) \\
&\iff \exists \Omega' \subseteq \Omega \text{ s.t. } \mathbb{P}(\Omega') = 1 \text{ and } \forall \omega \in \Omega',\ Z_n(\omega) \to 0 \\
&\iff \forall \varepsilon \in \mathbb{Q}_{>0}, \exists \Omega' \subseteq \Omega \text{ s.t. } \mathbb{P}(\Omega') = 1 \text{ and } \forall \omega \in \Omega',\ \lim_{n\to\infty} Z_n(\omega) \leq \varepsilon \\
&\iff \forall \varepsilon \in \mathbb{Q}_{>0}, \mathbb{P}\left(\lim_{n\to\infty} \{\omega : Z_n(\omega) > \varepsilon\}\right) = 0 \\
&\iff \forall \varepsilon > 0,\ \lim_{n\to\infty} \mathbb{P}(Z_n > \varepsilon) = 0.
\end{aligned}$$

18.2.2 $\xrightarrow{L^p}$ 与 $\xrightarrow{P}$

我们接着说明，对于 $p \geq 1$，“依 L^p 收敛”可以推出“依概率收敛”，但是反过来不成立。我们只需要对 $p = 1$ 的情况进行证明，因为，我们接着马上要说明，如果 $p > 1$，那么“依 L^p 收敛”可以推出“依 L^1 收敛”。

这件事情正确的直观也很容易，“依概率收敛”是说的随机变量 X_n 和 X 不一样的位置的测度趋向于 0。而“依 L^p 收敛”要求的是在不一样的地方，这个测度还要乘上“两者所差”的值之后依旧趋向于 0。因此，这个要求更强一些。证明使用马尔可夫不等式即可：对于任何 $\varepsilon > 0$，

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} \frac{\mathbf{E}[|X_n - X|]}{\varepsilon} = 0.$$

反过来不成立的例子是我们很熟悉的：设 X_n, X 均是定义在 $(0,1)$ 的均匀测度上的随机变量。我们令 $X_n = n \cdot \mathbb{I}_{(0,\frac{1}{n})}$，$X = 0$.

18.2.3 $\xrightarrow{L^q}$ 与 $\xrightarrow{L^p}$

和上面相同的直观指出，如果 $q > p$，那么 $X_n \xrightarrow{L^q} X$ 可以推出 $X_n \xrightarrow{L^p} X$。证明如下：

$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|^p] = \lim_{n\to\infty} \mathbf{E}\left[(|X_n - X|^q)^{\frac{p}{q}}\right] \leq \lim_{n\to\infty} (\mathbf{E}[|X_n - X|^q])^{\frac{p}{q}} = 0.$$

其中上式的不等号是利用琴生不等式以及 $f(x) = x^{\frac{p}{q}}$ 是一个 concave 函数的事实。

反过来不成立的例子可以类似前一种情况给出，留作练习。

18.2.4 $\xrightarrow{P}$ 与 $\xrightarrow{D}$

根据定义就可以知道，“依分布收敛” $\xrightarrow{D}$ 是一个很弱的概念，它关心的是分布函数的收敛性，甚至都不要求随机变量们生活在同一个概率空间上。

固定概率空间 $(\Omega, 2^{\Omega}, \mathbb{P})$ 为在 $\Omega = \{0,1\}$ 上的均匀分布。对于任意 n，定义 $X_n(0) = 0, X_n(1) = 1$。定义 $X(0) = 1, X(1) = 0$。那显然 $X_n \xrightarrow{D} X$ 但是 $X_n \not\xrightarrow{P} X$.

我们现在来说明 $X_n \xrightarrow{P} X$ 可以推出 $X_n \xrightarrow{D} X$。我们想把事件 $|X_n - X| > \varepsilon$ 与 X_n 的分布函数联系起来。我们使用下面一个基本的事实：对于任意 $\varepsilon > 0$，任意两个随机变量 X, Y 和实数 a：

$$Y \le a \implies X \le a + \varepsilon \text{ or } |Y - X| > \varepsilon,$$

即如果已知 Y 不大于 a，则要么 X 不大于 $a + \varepsilon$，要么 X 和 Y 的差距比较大。我们使用这个不等式两次，并使用 union-bound，可以得到

$$\mathbb{P}(X_n \le a) \le \mathbb{P}(X \le a + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$
$$\mathbb{P}(X < a - \varepsilon) \le \mathbb{P}(X_n \le a) + \mathbb{P}(|X_n - X| > \varepsilon)$$

这便得到了

$$\mathbb{P}(X \le a - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \le \mathbb{P}(X_n \le a) \le \mathbb{P}(X \le a + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

我们让 n 趋向于无穷大并让 $\varepsilon \to 0$ 便得到了想要的结论。

18.2.5 $\xrightarrow{a.s.}$ 与 $\xrightarrow{L^1}$

这两者一般来说是不可比较的。事实上，在一定条件下，我们有 $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{L^1} X$。如果我们存在一个可积的随机变量 Y，满足对于每一个 n，$|X_n| \le Y$ 并且 $|X| \le Y$。那么容易验证 $|X_n - X| \le 2Y$。显然我们也有 $|X_n - X| \xrightarrow{a.s.} 0$。因此由 DCT

$$\lim_{n\to\infty} \mathbf{E}[|X_n - X|] = \mathbf{E}\left[\lim_{n\to\infty} |X_n - X|\right] = 0.$$

18.3 集合的极限

我们之前定义过集合的极限的概念。如果 $\{A_n\}_{n\ge1}$ 是单调递增的（$\forall n, A_n \subseteq A_{n+1}$），那么

$$\lim_{n\to\infty} A_n := \bigcup_{n\ge1} A_n.$$

类似的，如果 $\{A_n\}_{n\ge1}$ 是单调递减的（$\forall n, A_n \supseteq A_{n+1}$），那么

$$\lim_{n\to\infty} A_n := \bigcap_{n\ge1} A_n.$$

这可以类比于数列的极限。假设我们有一列实数 a_n，如果它是单调的数列，那么它一定存在极限（允许极限是正负无穷大的话）。而如果数列不单调的话，那么极限就不一定存在了。但是，我们可以定义它的上极限和下极限：

$$\limsup_{n\to\infty} a_n := \lim_{n\to\infty} \left(\sup_{k\ge n} a_k\right)$$
$$\liminf_{n\to\infty} a_n := \lim_{n\to\infty} \left(\inf_{k\ge n} a_k\right).$$

上极限和下极限总是存在的，这是因为 $\left(\sup_{k\ge n} a_k\right)_{n\ge1}$ 与 $\left(\inf_{k\ge n} a_k\right)_{n\ge1}$ 分别是单调递减和单调递增的数

列。我们尝试类似的定义集合列的上极限与下极限。设 $(A_n)_{n\geq 1}$ 是一列（不一定单调的）集合。我们定义

$$\limsup_{n\to\infty} A_n := \lim_{n\to\infty}\left(\sup_{k\geq n} A_k\right)$$
$$\liminf_{n\to\infty} A_n := \lim_{n\to\infty}\left(\inf_{k\geq n} A_k\right).$$

当然，我们还没有说 $\sup_{k\geq n} A_k$ 和 $\inf_{k\geq n} A_k$ 是怎么定义的。但是，我们可以很自然的想到，对于一个集族 $\{B_n\}_{n\in I}$，其上确界应该是包含每一个 B_n 的最小的集合，而下确界应该是被每一个 B_n 包含的最大的集合。因此

$$\sup_{n\in I} B_n := \bigcup_{n\in I} B_n$$
$$\inf_{n\in I} B_n := \bigcap_{n\in I} B_n.$$

使用这个定义，以及对于单调集合族极限的定义，我们有：

$$\limsup_{n\to\infty} A_n := \lim_{n\to\infty}\left(\sup_{k\geq n} A_k\right) = \bigcap_{n\geq 1}\bigcup_{k\geq n} A_k$$
$$\liminf_{n\to\infty} A_n := \lim_{n\to\infty}\left(\inf_{k\geq n} A_k\right) = \bigcup_{n\geq 1}\bigcap_{k\geq n} A_k.$$

另外一个比较重要的事情是我们来看看 $\limsup A_n$ 和 $\liminf A_n$ 究竟包含的哪些元素。简单的思考之后（记得思考哦），我们可以发现

$$\limsup A_n = \{x : x \text{ 在无穷多个} A_n \text{ 中出现过}\},$$
$$\liminf A_n = \{x : x \text{ 只在有限个} A_n \text{ 中没出现过}\}.$$

基于这种直观含义，我们有的时候会把“$\limsup A_n$”记作“A_n i.o”，其中“i.o”是“infinitely often”的意思。

我们使用定义以及集合的 De-Morgan 律，可以马上得到

$$\limsup A_n = (\liminf A_n^c)^c.$$

18.4 波莱尔-坎泰利引理（Borel-Cantelli proposition）

我们接着介绍一个很常用的工具。它通常处理的问题是这样的：假设在一个固定的概率空间里，我们有一些坏事件 $\{A_n\}_{n\geq 1}$。我们想知道，有多大的概率，这些坏事件不会总发生。

命题 18.1 (Borel-Cantelli)

如果 $\sum_{n\geq 1}\mathbb{P}(A_n) < \infty$，那么 $\mathbb{P}(A_n \text{ i.o}) = 0$.

我们前面刚说过

$$A_n \text{ i.o} = \limsup_n A_n = \{\omega : \omega \text{ 在无穷多个} A_n \text{ 中出现过}\}.$$

因此，Borel-Cantelli 说的是，如果所有的坏事件（它们可能互相相关）发生的概率之和是一个有限数的话，那么，几乎一定（almost surely）这些坏事件不会不停发生。

Borel-Cantelli 的证明非常简单：

$$\mathbb{P}(A_n \text{ i.o}) = \mathbb{P}\left(\lim_n\left(\sup_{k\geq n} A_k\right)\right) = \lim_n \mathbb{P}\left(\sup_{k\geq n} A_k\right) \leq \lim_n \sum_{k\geq n}\mathbb{P}(A_k).$$

上面式子里第二个等号是因为概率测度的连续性，不等号是使用了 union-bound。根据我们的条件，$\sum_{n\geq 1}\mathbb{P}(A_n) < \infty$，而一个收敛级数的 tail 一定是 0。所以我们有 $\mathbb{P}(A_n \text{ i.o}) = 0$.

Borel-Cantelli 反过来就不一定正确了，也就是说如果 $\sum_{n\geq 1}\mathbb{P}(A_n)=\infty$，不一定有 $\mathbb{P}(A_n \text{ i.o})>0$。但是，如果这些坏事件是相互独立的，那么 $\mathbb{P}(A_n \text{ i.o})=1$。这个结论又被称为第二 Borel-Cantelli 引理。

命题 18.2 (Second Borel-Cantelli)

如果 A_n 相互独立，那么

1. $\mathbb{P}(A_n \text{ i.o.})=0 \iff \sum_{n\geq 1}\mathbb{P}(A_n)<\infty$.
2. $\mathbb{P}(A_n \text{ i.o.})=1 \iff \sum_{n\geq 1}\mathbb{P}(A_n)=\infty$.

♠

这个引理也说明，概率 $\mathbb{P}(A_n \text{ i.o.})$ 只有 0 或者 1 两种取值。这实际上是一种更一般的现象，被称为 0-1 律，我们在未来会介绍。

我们现在证明 Second Borel-Cantelli。事实上，我们只要证明 $\sum_{n\geq 1}\mathbb{P}(A_n)=\infty \implies \mathbb{P}(A_n \text{ i.o.})=1$ 就可以了（why）。于是，我们利用独立的条件和 De-Morgen 律可以得到

$$\mathbb{P}\left(\limsup_n A_n\right)=1-\mathbb{P}\left(\liminf_n A_n^c\right)=1-\lim_n \mathbb{P}\left(\bigcap_{k\geq n}A_k^c\right)=1-\lim_n\prod_{k\geq n}\mathbb{P}(A_k^c).$$

如果我们设 $x_k:=\mathbb{P}(A_k)$，那么

$$\mathbb{P}\left(\limsup_n A_n\right)=1-\lim_n\prod_{k\geq n}(1-x_k)\geq 1-\lim_n e^{-\sum_{k\geq n}x_k}=1.$$

其中最后一个等号是因为 $\sum_n x_n=\infty$ 是一个发散的级数（因此它的 tail 是发散的）。

我们现在来使用 Borel-Cantelli 来证明一个有用的结论。即如果 $X_n \xrightarrow{P} X$，那么存在一个子序列 $n_1,n_2,\dots$，满足 $X_{n_k}\xrightarrow{a.s.}X$.

大家可以先想想，在我们前面说明 $X_n\xrightarrow{P}X \not\xrightarrow{a.s.} X$ 的例子里，这样一个子序列如何挑。

我们只用挑那些红色都在最左边的 X_{n_k} 即可。

以下的证明，是 Borel-Cantelli 引理的一个典型应用。对于每一个 $k\geq 1$，我们选取 n_k 满足 $\mathbb{P}\left(|X_{n_k}-X|\geq\frac{1}{k}\right)\leq\frac{1}{2^k}$。由于 $X_n\xrightarrow{P}X$，这样的 n_k 总是可以挑出来。我们用 A_k 来表示坏事件 “$|X_{n_k}-X|\geq\frac{1}{k}$”。那么根据定义 $\sum_{k\geq 1}\mathbb{P}(A_k)\leq\sum_{k\geq 1}2^{-k}<\infty$。于是使用 Borel-Cantelli，我们可以得到 $\mathbb{P}(A_n \text{ i.o.})=0$.

我们需要仔细解读一下 $\mathbb{P}(A_n \text{ i.o.})=0$ 意味着什么。它说明，存在 $\Omega'\subseteq\Omega$，满足 $\mathbb{P}(\Omega')=1$，对于任何 $\omega\in\Omega'$，只存在有限个 k，使得 $|X_{n_k}(\omega)-X(\omega)|\geq\frac{1}{k}$ 成立。这意味着对于每一个这样的 $\omega\in\Omega'$，$X_{n_k}(\omega)\to X(\omega)$.

我们可以使用这个结论加强我们的老熟人控制收敛定理：

定理 18.1 (控制收敛定理)

设 X_n 为一列随机变量，满足 $\lim_{n\to\infty}X_n=X$ a.e.。如果存在一个随机变量 Y，满足

1. 对所有 $n\in\mathbb{N}$, $|X_n|\leq Y$;
2. Y 是可积的。

那么 $\lim_{n\to\infty}\mathbf{E}[X_n]=\mathbf{E}[X]$。

♡

我们现在说明，我们可以把条件里的 $X_n\xrightarrow{a.s.}X$ 弱化成 $X_n\xrightarrow{P}X$。我们使用反证法。假设 $\lim_n\mathbf{E}[X_n]=\mathbf{E}[X]$ 不成立。那么，一定存在一个子序列 $\{n_k\}_{k\geq 1}$ 满足 $\lim_{k\to\infty}\mathbf{E}[X_{n_k}]=L\neq\mathbf{E}[X]$。根据条件，我们知道 $X_{n_k}\xrightarrow{P}X$。因此，使用我们刚才证明的结论，从 $\{n_k\}_{k\geq 1}$ 中我们能再找到一个子序列 $\{m_j\}_{j\geq 1}\subseteq\{n_k\}_{k\geq 1}$，满足 $X_{m_j}\xrightarrow{a.s.}X$。根据 a.s. 版本的 DCT，我们知道 $\lim_{j\to\infty}\mathbf{E}\left[X_{m_j}\right]=\mathbf{E}[X]\neq L$，这与 $\left\{X_{m_j}\right\}$ 是 $\{X_{n_k}\}$ 的子序列矛盾，因为 $\mathbf{E}\left[X_{m_j}\right]$ 与 $\mathbf{E}[X_{n_k}]$ 理应有一样的极限。

第 19 章 大数定律，矩方法

19.1 大数定律（Law of large numbers）

有一个六面的骰子，我们想知道它扔到每一面的概率是不是相同的。那我们便可以通过反复投掷该骰子，然后统计每一面出现的频率。直观上，当我们投掷的次数足够多之后，每一面出现的频率就应该越来越接近投掷一次骰子时该面出现的概率。大数定律就是严格的来说明这件事情。

有很多种不同形式的大数定律，但是它们都遵循着同样的模式。设 $X_1, X_2, \dots$ 是定义在同一个概率空间上的随机变量。对于 $n \in \mathbb{N}$, 我们定义部分和 $S_n = \sum_{i\in[n]} X_i$。一个大数定律总是如下模式

- 如果 $X_1, X_2, \dots$ 满足 [某条件],
- 那么 $\frac{S_n}{n}$ 以 [某种模式] 收敛到 [某一个数]。

我们可以对 [某条件], [某种模式], [某一个数] 取不同的东西，得到不同的大数定律。但在一般情况下，我们关心的收敛模式主要是“依概率收敛”和“几乎必然收敛”两种。而具有这两种收敛模式的结论的大数定律，我们分别叫做**弱大数定律（Weak Law of Large Numbers, WLLN）**和**强大数定律（Strong Law of Large Numbers, SLLN）**。由于我们前面说明过，“几乎必然收敛”是比“依概率收敛”更强的收敛模式，因此，强大数定律的成立要么需要更强的条件，要么其证明需要更加细致的分析。

我们接下来给出几个大数定律。

1. （弱大数定律）设 $X_1, X_2, \dots$ 是**独立**的随机变量，并且每一个 X_i 均是可积的且有相同的期望，即 $\mathbf{E}[X_i] = \mu$。如果对于每一个 X_i，其**二阶矩有统一上界**，即 $\mathbf{E}\left[X_i^2\right] \le \sigma^2$。那么 $\frac{S_n}{n} \xrightarrow{P} \mu$。
2. （Cantelli 强大数定律）设 $X_1, X_2, \dots$ 是**独立**的随机变量，并且每一个 X_i 均是可积的且有相同的期望，即 $\mathbf{E}[X_i] = \mu$。如果对于每一个 X_i，其**四阶矩有统一上界**，即 $\mathbf{E}\left[X_i^4\right] \le \tau$。那么 $\frac{S_n}{n} \xrightarrow{a.s} \mu$。
3. （辛钦（Khinchin）弱大数定律）设 $X_1, X_2, \dots$ 是**独立同分布**的随机变量，并且每一个 X_i 均是可积的，满足 $\mathbf{E}[X_i] = \mu$。那么 $\frac{S_n}{n} \xrightarrow{P} \mu$。
4. （科尔莫格洛夫（Kolmogorov）强大数定律）设 $X_1, X_2, \dots$ 是**独立同分布**的随机变量，并且每一个 X_i 均是可积的，满足 $\mathbf{E}[X_i] = \mu$。那么 $\frac{S_n}{n} \xrightarrow{a.s.} \mu$。

注意到，前两个大数定律对随机变量的高阶矩有要求，而后面两个大数定律只要求随机变量可积。但是，前两个大数定律允许随机变量具有不同的分布，但后面两个要求随机变量必须是同分布的，否则结论不一定正确。

19.2 矩方法

我们今天先来证明前两个大数定律，即弱大数定律与 Cantelli 强大数定律。它们的共同点是我们分别要求了随机变量的二阶矩和四阶矩有上界。这便让我们可以使用切比雪夫不等式来进行证明。我们先来证明弱大数定律。

由于显然 $\mathbf{E}\left[\frac{S_n}{n}\right]=\mu$，对于任何 $\varepsilon>0$，使用切比雪夫不等式，我们有

$$\mathbb{P}\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right)\leq\frac{\mathbf{Var}\left[S_n\right]}{n^2\varepsilon^2}\leq\frac{\sigma^2}{n\varepsilon^2}\to 0.$$

我们接着来证明 Cantelli 强大数定律。我们可以不妨假设 $\mu=0$（否则可以使用 $X_n-\mu$ 来代替 X_n）。我们如果想说明“几乎必然收敛”，可以回忆我们上节课证明依概率收敛的随机变量列里存在几乎必然收敛的子列的证明。我们希望这里挑出来的子序列就是序列 $\frac{S_n}{n}$ 自己。我们是使用 Borel-Cantelli 来说明几乎必然收敛的，也因此要求概率 $\mathbb{P}\left(\left|\frac{S_n}{n}\right|>\varepsilon\right)$ 不仅仅收敛到零，而且还要收敛的足够快，快到级数 $\sum_{n\geq 1}\mathbb{P}\left(\left|\frac{S_n}{n}\right|>\varepsilon\right)$ 是收敛的。我们前面使用二阶矩的切比雪夫不等式得到的界显然是不够的，这也是为什么在 Cantelli 强大数定律的条件里面出现了更强的“四阶矩有界”的条件。

在这个条件下，使用马尔科夫不等式，我们有

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right|>\varepsilon\right)=\mathbb{P}\left(\left(\frac{S_n}{n}\right)^4>\varepsilon^4\right)\leq\frac{\mathbf{E}\left[S_n^4\right]}{n^4\varepsilon^4}.$$

我们需要来估计 $\mathbf{E}\left[S_n^4\right]$。注意到

$$\mathbf{E}\left[S_n^4\right]=\mathbf{E}\left[\left(\sum_{i\in[n]}X_i\right)^4\right]=\sum_{i,j,k,\ell\in[n]}\mathbf{E}\left[X_iX_jX_kX_\ell\right].$$

由于 X_i 是相互独立的均值为零的随机变量，我们知道如果 $\mathbf{E}\left[X_iX_jX_kX_\ell\right]$ 在出现一次项的时候一定为零（比如 $\mathbf{E}\left[X_1X_4^3\right]=\mathbf{E}\left[X_1\right]\mathbf{E}\left[X_4^3\right]=0$）。所以

$$\sum_{i,j,k,\ell\in[n]}\mathbf{E}\left[X_iX_jX_kX_\ell\right]=3n(n-1)\mathbf{E}\left[X_1^2X_2^2\right]+n\mathbf{E}\left[X_1^4\right].$$

我们知道 $\mathbf{E}\left[X_1^4\right]\leq\tau$。根据 Cauchy-Schwarz 不等式，

$$\mathbf{E}\left[X_1^2X_2^2\right]\leq\sqrt{\mathbf{E}\left[X_1^4\right]\mathbf{E}\left[X_2^4\right]}\leq\tau.$$

所以，我们有 $\mathbf{E}\left[S_n^4\right]\leq(3n^2-2n)\tau$。如果我们选 $\varepsilon=\frac{1}{n^{0.1}}$ 的话，马上可以得到

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right|>\frac{1}{n^{0.1}}\right)\leq\frac{3\tau}{n^{1.6}}.$$

我们用 A_n 表示坏事件 $\left|\frac{S_n}{n}\right|>\frac{1}{n^{0.1}}$，那么由于

$$\sum_{n\geq 1}\mathbb{P}\left(A_n\right)\leq\sum_{n\geq 1}\frac{3\tau}{n^{1.6}}<\infty,$$

根据 Borel-Cantelli，我们有 $\mathbb{P}\left(A_n\text{ i.o.}\right)=0$，即 $\frac{S_n}{n}\overset{a.s.}{\to}0$。

19.3 两个大数定律的应用

19.3.1 Glivenko-Cantelli 定理

我们假设有一个随机源，可以反复、独立的产生分布为 μ 的样本。我们从中进行 n 次采样得到 n 个独立同分布的样本 $X_1,X_2,\ldots,X_n$。设 F 是 μ 的分布函数，我们现在想根据这些样本来估计 F，应该怎么做？一个很自然的做法如下：我们定义**经验分布函数**

$$F_n(x)=\frac{1}{n}\cdot\left|\{i\in[n]\,:\,X_i\leq x\}\right|.$$

对于一个固定的 x，我们定义 $Y_i = \mathbb{I}_{X_i \le x}$。那么 $F_n(x) = \frac{1}{n}\sum_{i\in[n]} Y_i$。使用期望的线性性容易验证，对于任何 x，我们有

$$\mathbf{E}\left[F_n(x)\right] = \frac{1}{n}\sum_{i\in[n]} \mathbf{E}\left[Y_i\right] = \mathbb{P}\left(X_i \le x\right) = F(x).$$

换句话说，$F_n(x)$ 是对 $F(x)$ 的一个无偏估计。

Glivenko-Cantelli 定理说的是 almost surely，函数列 F_n 一致收敛到 F。注意到，我们这里有两个修饰的词，几乎一定（almost surely）和一致（uniformly）。因为 F_n 是随机变量，所以实际上可以看成 $(\omega, x) \times \Omega \in \mathbb{R} \to \mathbb{R}$ 的函数。所以这句话的意思是，存在一个测度为 1 的样本集 $\Omega' \subseteq \Omega$，使得对于任意 $\omega \in \Omega'$，函数 $F_n(\omega, \cdot)$ 一致收敛到 $F(\cdot)$。我们先不妨假设 F 是 $[0,1]$ 上的均匀分布的分布函数，我们接着再说明，对于一般的分布，可以简单的转化成均匀分布的情形。

对于固定的 x，由于 Y_i 是一个伯努利分布的变量，它的每一阶矩都是有限的。Cantelli 大数定律告诉我们 $F_n(x) \xrightarrow{a.s.} F(x)$。我们现在需要进一步说明，这一个 almost surely 可以对于所有的 $x \in [0,1]$ 同时保持，并且，收敛速度是一致的（与 x 无关）。我们固定一个让 $F_n(x) \to F(x)$ 的样本点。对于任何 $\varepsilon > 0$，我们取一个整数 m 满足 $\frac{1}{m} < \varepsilon$。对于 $i = 0, 1, \ldots, m$，我们设 $x_i = \frac{i}{m}$。由于 m 是一个整数，所以我们知道 almost surely, 存在一个 N (可以与 m 有关)，使得当 $n > N$ 的时候，对于任何 $i = 0, 1, \ldots, m$，$|F_n(x_i) - F(x_i)| \le \varepsilon$。

于是，对于任何的 $x \in [0,1]$，我们总是可以找到一个 i，使得 $x \in [x_i, x_{i+1}]$。于是，根据 F_n 和 F 的单调性，对于 $n > N$，有

$$F_n(x) - F(x) \le F_n(x_{i+1}) - F(x_i) \le |F_n(x_{i+1}) - F(x_{i+1})| + F(x_{i+1}) - F(x_i) \le 2\varepsilon;$$

$$F_n(x) - F(x) \ge F_n(x_i) - F(x_{i+1}) = (F_n(x_i) - F(x_i)) + (F(x_i) - F(x_{i+1})) \ge -2\varepsilon.$$

这说明，在这个样本点上，$|F_n(x) - F(x)| \le 2\varepsilon$。由于 N 是一个与 x 无关的数，这个收敛是一致收敛。

我们现在来考虑 F 是一般的分布函数的情况。如果 X 的分布是 F，那么对于任何 t

$$F(t) = \mathbb{P}\left(X \le t\right) = \mathbb{P}\left(F(X) \le F(t)\right)$$

这说明，如果我们定义 $Z = F(X) \in [0,1]$，那么 $\mathbb{P}\left(Z \le F(t)\right) = F(t)$。这意味着 Z 的分布是 $[0,1]$ 上的均匀分布。我们用 U 来表示 $[0,1]$ 上均匀分布的分布函数。

对于每一个 $i \in [n]$，我们设 $Z_i = F(X_i)$。因此，我们可以想想此时到来的是 n 个 $[0,1]$ 上均匀分布的随机变量，这便把问题转化为了我们刚解决的均匀分布的情况。实际上，我们假设 U_n 是使用这些假想的 Z_i 估计出来的 U 的近似，那么就有对于任何 n 和任何 $x \in \mathbb{R}$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \le x} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{Z_i \le F(x)} = U_n(F(x)).$$

于是，对于每一个 n，

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{x\in\mathbb{R}} |U_n(F(x)) - U(F(x))| = \sup_{z\in[0,1]} |U_n(z) - U(z)|.$$

而上述最后一项我们已经证明了 almost surely uniformly 收敛到 0。

19.3.2 Bernstein 多项式 与 Weierstrass 定理

Weierstrass 定理说的是每一个定义在 $[0,1]$ 上的连续函数 $f(x)$，均可以找到一列多项式 $p_n(x)$ 使得

$$\lim_{n\to\infty} \left(\sup_{x\in[0,1]} |f(x) - p_n(x)| \right) = 0,$$

即 p_n 一致收敛到 f。我们现在使用“概率”的想法给出一个简洁的证明。

首先我们构造 p_n 收敛到 f。对于每一个 $x \in [0,1]$，我们考虑随机变量 $Y_n^x \sim \texttt{Bin}(n, x)$。由于 Y_n^x 可以写成 n 个独立的满足 $\texttt{Ber}(x)$ 分布的随机变量之和，根据弱大数定律，我们有 $\frac{Y_n^x}{n} \xrightarrow{P} x$。由于 f 是 $[0,1]$ 上的连续函数，所以 $f\left(\frac{Y_n^x}{n}\right) \xrightarrow{P} f(x)$ (why?)。再根据 DCT，我们有 $\mathbf{E}\left[f\left(\frac{Y_n^x}{n}\right)\right] \to f(x)$。我们便令 $p_n(x) = \mathbf{E}\left[f\left(\frac{Y_n^x}{n}\right)\right]$，可以

看到，根据 LOTUS 以及二项式分布的定义，

$$p_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}$$

是一个关于 x 的多项式。

我们再来验证收敛的一致连续性。由于 f 是连续函数，对于任何 $\varepsilon > 0$，存在 $\delta > 0$ 满足 $|y-x| < \delta \implies |f(y)-f(x)| < \varepsilon$。于是，

$$\begin{aligned}\left|\mathbf{E}\left[f\left(\frac{Y_n^x}{n}\right)\right] - f(x)\right| &= \mathbf{E}\left[\left|f\left(\frac{Y_n^x}{n}\right) - f(x)\right| \,\middle|\, \left|\frac{Y_n^x}{n} - x\right| < \delta\right] \cdot \mathbb{P}\left(\left|\frac{Y_n^x}{n} - x\right| < \delta\right) \\ &+ \quad \mathbf{E}\left[\left|f\left(\frac{Y_n^x}{n}\right) - f(x)\right| \,\middle|\, \left|\frac{Y_n^x}{n} - x\right| \geq \delta\right] \cdot \mathbb{P}\left(\left|\frac{Y_n^x}{n} - x\right| \geq \delta\right) \\ &\leq \varepsilon + 2\|f\|_\infty \cdot \mathbb{P}\left(\left|\frac{Y_n^x}{n} - x\right| \geq \delta\right).\end{aligned}$$

使用切比雪夫不等式，我们可以得到 $\mathbb{P}\left(\left|\frac{Y_n^x}{n} - x\right| \geq \delta\right) \leq \frac{1}{4n\delta^2}$。因此

$$\left|\mathbf{E}\left[f\left(\frac{Y_n^x}{n}\right)\right] - f(x)\right| \leq \varepsilon + \frac{\|f\|_\infty}{2n\delta^2}$$

随着 n 增大趋向于零，并且这个上界是一个与 x 无关的数。

第 20 章 截断法，辛钦弱大数定律

我们今天来首先证明辛钦弱大数定律，需要引入一种称为“截断法”的技巧。我们再使用类似技巧来讨论我们第一次课提到的圣彼得堡悖论问题。

20.1 辛钦大数定律

我们回忆上节课提到过的辛钦弱大数定律：

定理 20.1 (辛钦弱大数定律)

设 $X_1, X_2, \dots$ 是独立同分布的随机变量，并且每一个 X_i 均是可积的，满足 $\mathbf{E}[X_i] = \mu$。那么 $\frac{S_n}{n} = \frac{\sum_{i\in[n]} X_i}{n} \xrightarrow{P} \mu$。 ♡

这个结论和我们之前证明过的弱大数定律主要有两点不同：

- 对随机变量列的要求加强为独立同分布，而在之前的结论里只要求独立；
- 对随机变量不要求二阶矩有界了。

注意到我们之前是使用的矩方法，也就是切比雪夫不等式证明的弱大数定律。在我们现在只知道每一个 $\mathbf{E}[X_i]$ 可积，即一阶矩有界的情况下是不够用的。比如说，我们不妨假设 $X_n > 0$, 那么根据马尔科夫不等式

$$\mathbb{P}\left(\frac{S_n}{n} - \mu > a\right) \le \frac{\mathbf{E}[X_1]}{\mu + a} = \frac{\mu}{\mu + a}.$$

不等式右边与 n 没有关系，因此我们不能够得到依概率收敛的结果。

注意到这里主要的困难在于 X_i 的二阶矩是无界的，因此我们不能使用切比雪夫不等式来加强上式，使得右边出现 $o_n(1)$ 项。我们接下来将介绍一个我很喜欢的证明。我们现在做这样一个操作：选一个数 $M > 0$，把随机变量 X_i 分成其绝对值 $\le M$ 和 $> M$ 的两部分之和。对于绝对值 $\le M$ 的那一部分，我们知道由于其绝对值不超过 M，二阶矩是有界的，所以我们可以用切比雪夫不等式控制。对于绝对值 $> M$ 的那一部分，我们尝试说明它取到这里的值的概率为零（即所谓零阶矩方法）。To this end，我们定义

$$X_i^{\le M} := X_i \cdot \mathbb{I}_{|X_i| \le M}, \quad X_i^{> M} := X_i \cdot \mathbb{I}_{|X_i| > M}.$$

于是 $X_i = X_i^{\le M} + X_i^{> M}$。我们同样定义

$$S_n^{\le M} = \sum_{i\in[n]} X_i^{\le M}, \quad S_n^{> M} := \sum_{i\in[n]} X_i^{> M}.$$

于是 $S_n = S_n^{\leq M} + S_n^{>M}$。对于常数 $\varepsilon > 0$，我们有

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) &= \mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu + \frac{S_n^{>M}}{n}\right| > \varepsilon\right) \\ &= \mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu + \frac{S_n^{>M}}{n}\right| > \varepsilon \wedge S_n^{>M} = 0\right) \\ &\quad + \mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu + \frac{S_n^{>M}}{n}\right| > \varepsilon \wedge S_n^{>M} \neq 0\right) \\ &\leq \mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu\right| > \varepsilon\right) + \mathbb{P}\left(S_n^{>M} \neq 0\right).\end{aligned}$$

按照我们刚才说的计划，我们需要选一个合适的 M，使得在 n 趋向于无穷大的时候，上式两个概率均趋向于 0。显然，上式两个概率让我们需要仔细权衡 M 的选择，因为 M 越小，第一项 $\mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu\right| > \varepsilon\right)$ 越小，而第二项 $\mathbb{P}\left(S_n^{>M} \neq 0\right)$ 越大。我们这儿令 $M = n$.

我们先来控制第一项 $\mathbb{P}\left(\left|\frac{S_n^{\leq n}}{n} - \mu\right| > \varepsilon\right)$。首先有一个问题在于，我们不能够直接对这个式子使用切比雪夫不等式，原因在于虽然 $\mathbf{E}\left[\frac{S_n}{n}\right] = \mu$，但截断后的 $\mu_n := \mathbf{E}\left[\frac{S_n^{\leq n}}{n}\right]$ 就不一定是 μ 了。根据定义

$$\mu_n = \mathbf{E}\left[X_1^{\leq n}\right] = \mathbf{E}\left[X_1 \cdot \mathbb{I}_{|X_i| \leq n}\right].$$

我们注意到，对于任意 n，都有 $\left|X_1^{\leq n}\right| \leq |X_1|$。因此由 DCT，我们知道

$$\lim_{n\to\infty} \mu_n = \mathbf{E}\left[\lim_{n\to\infty} X_1^{\leq n}\right] = \mathbf{E}\left[X_1\right] = \mu.$$

换句话说，在 n 足够大的时候 μ_n 和 μ 可以任意接近。因此，我们只需要证明

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n^{\leq n}}{n} - \mu_n\right| > \frac{\varepsilon}{2}\right) = 0$$

即可。对于每一个 n，我们使用切比雪夫不等式，可以得到

$$\mathbb{P}\left(\left|\frac{S_n^{\leq n}}{n} - \mu_n\right| > \frac{\varepsilon}{2}\right) \leq \frac{\mathbf{Var}\left[X_1^{\leq n}\right]}{n\varepsilon^2} \leq \frac{\mathbf{E}\left[\left(X_1^{\leq n}\right)^2\right]}{n\varepsilon^2} = \varepsilon^{-2}\mathbf{E}\left[\frac{X_1^2 \cdot \mathbb{I}_{|X_1| \leq n}}{n}\right].$$

对于任意 n，我们都有 $\frac{X_1^2 \cdot \mathbb{I}_{|X_1| \leq n}}{n} \leq |X_1|$。并且，$\lim_{n\to\infty} \frac{X_1^2 \cdot \mathbb{I}_{|X_1| \leq n}}{n} = 0$ (逐点哦)。所以再次根据 DCT，

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n^{\leq n}}{n} - \mu_n\right| > \frac{\varepsilon}{2}\right) \leq \varepsilon^{-2}\mathbf{E}\left[\lim_{n\to\infty} \frac{X_1^2 \cdot \mathbb{I}_{|X_1| \leq n}}{n}\right] = 0.$$

接着我们来考虑第二项 $\mathbb{P}\left(S_n^{>n} \neq 0\right)$。由于 $S_n^{>n} \neq 0$ 的话必然对于某个 $i \in [n], X_i^{>n} \neq 0$。于是使用 union-bound，我们有

$$\mathbb{P}\left(S_n^{>n} \neq 0\right) \leq \sum_{i\in[n]} \mathbb{P}\left(X_i^{>n} \neq 0\right) = n \cdot \mathbb{P}\left(X_1^{>n} \neq 0\right) = \mathbf{E}\left[n \cdot \mathbb{I}_{|X_1| > n}\right].$$

注意到对于每一个 n，我们依旧有 $n \cdot \mathbb{I}_{|X_1|>n} \leq |X_1|$。同时 $\lim_{n\to\infty} n \cdot \mathbb{I}_{|X_1|>n} = 0$ （还是逐点哦）。所以第三次使用 DCT，可以得到

$$\lim_{n\to\infty} \mathbb{P}\left(S_n^{>n} \neq 0\right) \leq \mathbf{E}\left[\lim_{n\to\infty} n \cdot \mathbb{I}_{|X_1|>n}\right] = 0.$$

这便完成了证明。

如果大家仔细查看整个证明，特别是三次使用 DCT 时候需要验证的条件，会发现 $M = n$ 的选择似乎让两种情况的证明都“刚刚好”正确。整个证明过程如同在钢丝上跳舞，非常美妙。

20.2 圣彼得堡悖论

我们接着讨论第一次课提到过的圣彼得堡悖论。

假设有一个基于掷硬币赌博游戏。首先庄家扔一个公平硬币，如果结果是正面，则给玩家 2 元钱，游戏结束；如果结果是反面，庄家再扔一次硬币，如果结果是正面，则给玩家 4 元钱，游戏结束，否则按照同样的规则继续扔硬币，每一轮奖金翻倍。换句话说，庄家会生成一个无限长的投掷硬币的结果序列，如果这个序列里第一次正面是第 k 个，则玩家获得 2^k 元的奖金。现在的问题是，你愿意花多少钱去购买一次玩这个游戏的机会？或者说，假设你可以无限次的玩这个游戏，但是每一次需要付门票 a 元，那你认为 a 设置成多少是合理的？

我们现在来这样建模这个问题。我们用 X_i 表示第 i 次游戏玩家能够得到的钱。那么 $S_n = \sum_{i\in[n]} X_i$ 就是玩了 n 次游戏后的总收入。所有的 X_i 都是独立同分布的。我们知道

$$\mathbf{E}[X_i] = \sum_{k=1}^{\infty} 2^k \cdot 2^{-k} = \infty.$$

因此，我们不能用前面的辛钦大数定律来描述 $\frac{S_n}{n}$ 的行为。我们直观上 $\frac{S_n}{n}$ 应该是会发散到无穷大的。现在我们便用截断法来说明这一点。我们同样希望找到合适的 μ （和 ε ）使得 $\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right)$ 随着 n 变大趋向于 0。使用和前面证明同样的记号和方法，我们有

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n^{\le M}}{n} - \mu\right| > \varepsilon\right) + \mathbb{P}\left(S_n^{>M} \neq 0\right).$$

我们这儿为了方便设 $M = 2^m$，其中 m 是一个待定参数。我们同样想对上面的第一项使用切比雪夫不等式，于是计算

$$\mathbf{E}\left[\frac{S_n^{\le 2^m}}{n}\right] = \mathbf{E}\left[X_1^{\le 2^m}\right] = \sum_{k=1}^{m} 2^k \cdot 2^{-k} = m.$$

于是我们可以选择 $\mu = m$，并且使用切比雪夫不等式得到

$$\mathbb{P}\left(\left|\frac{S_n^{\le M}}{n} - m\right| > \varepsilon\right) \le \frac{\mathbf{Var}\left[X_1^{\le 2^m}\right]}{n\varepsilon^2} \le \frac{\mathbf{E}\left[\left(X_1^{\le 2^m}\right)^2\right]}{n\varepsilon^2} \le \frac{2^{m+1}}{n\varepsilon^2},$$

其中最后一个不等式是由于 $\mathbf{E}\left[\left(X_1^{\le 2^m}\right)^2\right] = \sum_{k=1}^{m} 2^{2k} \cdot 2^{-k} \le 2^{m+1}$。

另一方面，我们依旧使用 union-bound，可以得到

$$\mathbb{P}\left(S_n^{>M} \neq 0\right) \le n \cdot \mathbb{P}\left(X_1^{>2^m} \neq 0\right) = n \cdot \sum_{k=m+1}^{\infty} 2^{-k} = n \cdot 2^{-m}.$$

把这两项放在一起，我们得到了

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \le \frac{2^{m+1}}{n\varepsilon^2} + n \cdot 2^{-m}.$$

我们因此又需要权衡截断的阈值 $M = 2^m$。盯一会儿上式之后不难发现，我们可以取

$$\varepsilon = \sqrt{\log_2 n}, \quad m = \log_2 n + \frac{1}{2}\log_2 \log_2 n.$$

就能得到

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \left(\log_2 n + \frac{1}{2}\log_2 \log_2 n\right)\right| > \sqrt{\log_2 n}\right) \le \frac{3}{\sqrt{\log_2 n}},$$

即

$$\mathbb{P}\left(\left|\frac{S_n}{n\log_2 n} - (1 + o(1))\right| > \frac{1}{\sqrt{\log_2 n}}\right) \le \frac{3}{\sqrt{\log_2 n}}.$$

这说明，$\frac{S_n}{n} \xrightarrow{P} (1 + o(1))\log_2 n$。这说明，如果游戏的门票是 n 轮捆绑销售的话，每一轮 $\log_2 n$ 的价格是一个合适的门票定价。

第 21 章 作为信息的 σ-代数，Kolmogorov 0-1 律

21.1 σ-代数与信息

我们今天从另外一视角来看 σ-代数，即看成信息的集合。为了说明这一点，我们回顾一下随机变量的定义。给定一个概率空间 $(\Omega,\mathscr{F},\mathbb{P})$，我们说函数 $X\colon\Omega\to\mathbb{R}$ 是一个随机变量，当且仅当 X 是一个可测函数，也就是说对于任何 $B\in\mathscr{B}(\mathbb{R})$，我们有 $X^{-1}(B)\in\mathscr{F}$。这个时候，我们也称 X 是 $\mathscr{F}$-可测的。同理，对于定义在 Ω 上的任意一个 σ-代数 $\mathscr{G}$ 和一个函数 $Y\colon\Omega\to\mathbb{R}$，我们说 Y 是 $\mathscr{G}$-可测的，当且仅当对于任何 $B\in\mathscr{B}(\mathbb{R})$，$Y^{-1}(B)\in\mathscr{G}$。

反过来，给定一个函数 $X\colon\Omega\to\mathbb{R}$，我们用 $\sigma(X)$ 表示使得 X 可测的最小的 σ-代数，容易验证，$\sigma(X)$ 总是存在的。直观上，对于**离散**的 X 我们可以把 $\sigma(X)$ 理解成 $\left\{X^{-1}(x)\,:\,x\in\mathtt{Im}(X)\right\}$ 所构成的 Ω 的分划所生成的 σ-代数。这个直观帮助我们理解这个概念很重要。实际上，对于一般的 X 我们有下面命题。

命题 21.1

$\sigma(X)=\left\{X^{-1}(B)\,:\,B\in\mathscr{B}(\mathbb{R})\right\}$.

命题的验证很简单，首先根据定义 $\sigma(X)$ 必须包含 $\left\{X^{-1}(B)\,:\,B\in\mathscr{B}(\mathbb{R})\right\}$。其次，我们已经验证过，$\left\{X^{-1}(B)\,:\,B\in\mathscr{B}(\mathbb{R})\right\}$ 本身是一个 σ-代数。

我们可以自然的把定义推广到多个随机变量 $X_1,\ldots,X_n$ 上。我们用 $\sigma(X_1,\ldots,X_n)$ 表示使得 $(X_1,\ldots,X_n)$ 的联合分布可测的最小的 σ-代数。容易验证

$$\sigma(X_1,\ldots,X_n)=\sigma\left(\bigcup_{i\in[n]}\sigma(X_i)\right).$$

同样，如果是无穷多个随机变量 $\{X_\alpha\,:\,\alpha\in I\}$，那么 $\sigma(\{X_\alpha\,:\,\alpha\in I\}):=\sigma\left(\bigcup_{\alpha\in I}\sigma(X_i)\right)$。

我们今天会有很多比较抽象的概念，因此，脑子里一直有下面这个 running example 是比较重要的。我们考虑投掷一个公平的六面骰子的概率空间 $(\Omega,\mathscr{F},\mathbb{P})$，其中 $\Omega=[6]$，$\mathscr{F}=2^\Omega,\forall i\in\Omega,\mathbb{P}\left(\{i\}\right)=\frac{1}{6}$。我们定义四个随机变量

1. $X_1\colon i\in\Omega\mapsto i$，即 X_1 表示掷出来的点数；
2. $X_2\colon i\in\Omega\mapsto\mathbb{I}_{[i\geq 4]}$，即 X_2 表示掷出来的点数是“大”还是“小”；
3. $X_3\colon i\in\Omega\mapsto i\bmod 2$，即 X_3 表示掷出来的点数除 2 之后的余数；
4. $X_4\colon i\in\Omega\mapsto i\bmod 4$，即 X_4 表示掷出来的点数除 4 之后的余数。

我们可以分别计算 $\sigma(X_i)$。由于 X_i 是离散的随机变量，我们只需要给出分划 $\left\{X^{-1}(x)\,:\,x\in\mathtt{Im}(X)\right\}$ 就可以了。回忆到我们之前介绍过，对于一个集族 $\mathscr{A}\subseteq 2^\Omega$，$\sigma(\mathscr{A})$ 为包含 $\mathscr{A}$ 的最小 σ-代数。于是，稍作思索可以

得到

1. $\mathscr{F}_1 = \sigma(X_1) = \sigma(\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\})$;
2. $\mathscr{F}_2 = \sigma(X_2) = \sigma(\{\{1,2,3\},\{4,5,6\}\})$;
3. $\mathscr{F}_3 = \sigma(X_3) = \sigma(\{\{1,3,5\},\{2,4,6\}\})$;
4. $\mathscr{F}_4 = \sigma(X_4) = \sigma(\{\{4\},\{1,5\},\{2,6\},\{3\}\})$。

回忆我们说一个函数 $f\colon \mathbb{R}\to\mathbb{R}$ 是 Borel 的，当且仅当 f 是 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 可测的。下面命题可以说明，为什么我们把 σ-代数称为信息的集合。

命题 21.2

随机变量 Y 是 $\sigma(X)$-可测的当且仅当存在一个 Borel f 使得 $Y = f(X)$。

这个命题想说明这样一件事情：一个随机变量 Y 是另一个随机变量 X 生成的 σ-代数可测，意味着如果知道了 X 的取值，那么 Y 的取值也就知道。换句话说，X 包含了 Y 的所有信息，这等价于 $\sigma(Y)\subseteq\sigma(X)$。也就是说，如果我想知道随机变量 Y 的取值，我并不需要知道随机试验得到了哪个样本点 $\omega\in\Omega$，而只需知道随机试验得到的样本点在 X 上的取值即可。

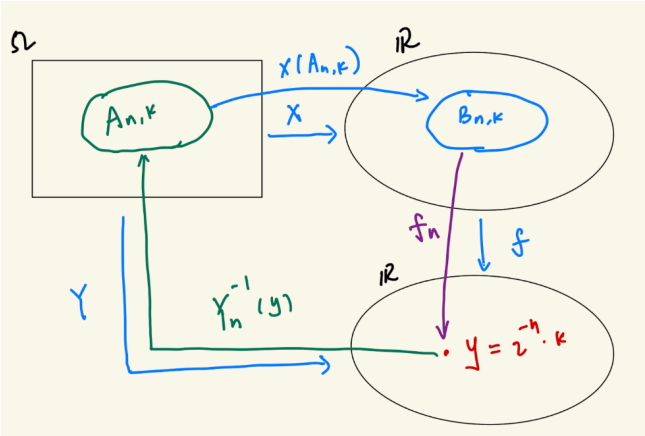
我们用前面的例子来检查一下这个结论，希望大家能够仔细弄清楚。

1. 首先，由于 $\mathscr{F}_1=\mathscr{F}$，因此 X_1, X_2, X_3, X_4 均是 $\mathscr{F}_1$-可测的。这是很显然的，因为 $X_1(i)=i$ 就返回随机实验得到的样本点本身，$\mathscr{F}_1$ 包含了“投一个公平六面骰子”的全部信息。
2. X_3 是 $\mathscr{F}_4$ 可测的。这个从 $\mathscr{F}_3$ 和 $\mathscr{F}_4$ 的定义上可以看出来，但直观上它想说的事情是，“如果我们知道一个数除 4 的余数，那自然也就知道其除 2 的余数”。因此，X_3 可以写成 X_4 的函数（$X_3 = X_4 \bmod 2$）。但是反过来就不对，因为我们知道一个数除 2 的余数，并不能够得到其除 4 的余数，$\sigma(X_3)$ 包含的信息严格少于 $\sigma(X_4)$。
3. $\mathscr{F}_2$ 和 $\mathscr{F}_3$ 是不能够比较的，因此，X_2 和 X_3 互相不能写成对方的函数。因为，知道一个数是否大于等于 4 不能确定其除 2 的余数，反之亦然。

我们接着来证明这个命题。

“当”是比较容易的。如果对于某个 Borel f，$Y=f(X)$，那么，对于任何 $B\in\mathscr{B}(\mathbb{R})$，

$$[Y\in B] = [f(X)\in B] = [X\in f^{-1}(B)]\in\sigma(X).$$



我们接着来说明“仅当”。也就是说，当 Y 是 $\sigma(X)$-可测的时候，我们要构造一个 Borel f 使得 $Y=f(X)$。我们先把 Y 进行离散化，对于任意 $n\in\mathbb{N}$，我们考虑 $\underline{Y}_n$。回忆其定义为

$$\forall\omega\in\Omega,\ \underline{Y}_n(\omega) = 2^{-n}\cdot k \ \text{ if } \ Y(\omega)\in(2^{-n}\cdot k, 2^{-n}\cdot(k+1)].$$

显然 $\underline{Y}_n$ 也是 $\sigma(X)$-可测的。因此，对于任何的 $k\in\mathbb{Z}$，我们考虑集合 $A_{n,k} := \underline{Y}_n^{-1}(2^{-n}\cdot k)\in\sigma(X)$。根据我们前面对于 $\sigma(X)$ 的命题，一定存在一个 $B_{n,k}\in\mathscr{B}(\mathbb{R})$ 使得 $A_{n,k} = X^{-1}(B_{n,k})$。显然，对于固定的 n，所有的 $B_{n,k}$ 是互相不相交的，并且构成了 $\mathbb{R}$ 的一个分划。对于每一个 $x\in B_{n,k}$，我们定义 $f_n(x) = 2^{-n}\cdot k$，

或者等价的 $f_n(x)=\sum_k 2^{-n}\cdot k\cdot \mathbb{I}_{x\in B_{n,k}}$。那么显然 $f_n(X(\omega))=\underline{Y}_n(\omega)$。我们让左右两边的 n 趋于无穷，即 $f(x)=\lim_{n\to\infty} f_n(x)$，即可得到 $f(X(\omega))=Y(\omega)$, as desired.

21.2 σ-代数的独立

我们之前定义过随机变量的独立。我们现在从更一般的角度来重新定义这个概念。给定同一个概率空间的样本集上的两个 σ-代数 $\mathscr{F}$ 和 $\mathscr{G}$，我们说 $\mathscr{F}$ 和 $\mathscr{G}$ 独立，记作 $\mathscr{F}\perp\mathscr{G}$，如果

$$\forall A\in\mathscr{F}, B\in\mathscr{G},\ \mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B).$$

类似的，对于有限个 σ-代数 $\mathscr{F}_1,\mathscr{F}_2,\ldots,\mathscr{F}_n$，我们说它们是独立的当且仅当

$$\forall A_1\in\mathscr{F}_1,\ldots,A_n\in\mathscr{F}_n,\ \mathbb{P}\left(\bigcap_{i\in[n]}A_i\right)=\prod_{i\in[n]}\mathbb{P}(A_i).$$

对于任意一族 σ-代数 $\{\mathscr{F}_\alpha:\alpha\in I\}$，我们说它们是独立的当且仅当它的任何一个有限子集是独立的。

我们现在说明，我们之前的定义的随机变量的独立性是一种特殊情况。

命题 21.3

随机变量 X 和 Y 独立当且仅当 $\sigma(X)$ 和 $\sigma(Y)$ 独立。

我们可以把上述命题自然的推广到一族随机变量 $\{X_\alpha:\alpha\in I\}$ 独立。

我们先证明“当”。对于任何 $A,B\in\mathscr{B}(\mathbb{R})$，我们知道 $[X\in A]\in\sigma(X),[Y\in B]\in\sigma(Y)$，因此 $\mathbb{P}(X\in A\wedge Y\in B)=\mathbb{P}(X\in A)\cdot\mathbb{P}(Y\in B)$。

然后来说明“仅当”。对于任何 $A\in\sigma(X)$ 和 $B\in\sigma(Y)$，我们知道一定存在 $A',B'\in\mathscr{B}(\mathbb{R})$，使得 $A=X^{-1}(A')$，$B=Y^{-1}(B')$。于是

$$\mathbb{P}(A\cap B)=\mathbb{P}(X\in A'\cap Y\in B')=\mathbb{P}(X\in A')\cdot\mathbb{P}(Y\in B')=\mathbb{P}(A)\cdot\mathbb{P}(B).$$

这个命题有一个很有用的推论：如果 $X_1,\ldots,X_n,Y$ 独立，并且 $f:\mathbb{R}^n\to\mathbb{R}$ 是一个 Borel 函数，那么 $f(X_1,\ldots,X_n)$ 和 Y 也独立。

证明是显然的，因为我们前面已经说明了 $\sigma(f(X_1,\ldots,X_n))\subseteq\sigma(X_1,\ldots,X_n)$。

21.3 Kolmogorov 0-1 律

我们之前学过几个结论：

1. （Kolmogorov 强大数定律）设 $X_1,\ldots,X_n,\ldots$ 是独立同分布的随机变量，满足 $\mathbf{E}[X_1]=\mu<\infty$。那么
$$\mathbb{P}\left(\lim_{n\to\infty}\frac{\sum_{i=1}^n X_i}{n}=\mu\right)=1.$$
2. （Second Borell-Cantelli）设 $A_1,\ldots,A_n,\ldots$ 是独立的事件，那么
$$\mathbb{P}(A_n\text{ i.o.})=\begin{cases}1, & \text{if } \sum_{i=1}^\infty\mathbb{P}(A_i)=\infty\\ 0, & \text{if } \sum_{i=1}^\infty\mathbb{P}(A_i)<\infty.\end{cases}$$

这俩结论都有几个共同点：都涉及独立的随机变量或者事件；都是讨论某一个极限事件发生的概率；这个事件发生的概率要么是 0 要么是 1 而不是其它的数。事实上，这个并不是巧合。Kolmogorov 0-1 律说明，对于一大类事件，它发生的概率非零即一。

为了说明这个定律，我们考虑在同一个概率空间 $(\Omega,\mathscr{F},\mathbb{P})$ 下的一列随机变量 $X_1,X_2,\ldots,X_n,\ldots$。对于每一个 $n\geq 1$，我们定义 $\mathscr{F}_n=\sigma(X_1,X_2,\ldots,X_n)$。于是，$\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\ldots$。我们通常把这样一个递增的 σ-代数链称为滤链（filtration），用来表示逐渐增多的信息。我们定义 $\mathscr{F}_\infty=\sigma\left(\bigcup_{i=1}^\infty\mathscr{F}_i\right)$。

比如说，我们考虑一个不停投掷均匀硬币的随机试验（比如我们在作业里定义过的几何分布随机变量的概率空间）。我们用 X_n 表示第 n 枚硬币的结果。直观上，$\mathscr{F}_n$ 包含了前 n 次硬币投掷结果的所有信息。

于是，一个随机变量 X 是 $\mathscr{F}_n$-可测的，当且仅当它的值可以被前 n 枚硬币投掷的结果所决定。比如 $X =$ “是否从一开始连续投出了 5 个正面” 这个随机变量便是当 $k \geq 5$ 时，$\mathscr{F}_k$-可测的，但不是 $\mathscr{F}_1, \mathscr{F}_2, \mathscr{F}_3, \mathscr{F}_4$-可测的。

我们接着定义一系列记号。对于每一个 $n \geq 0$，定义 $\mathscr{F}_n^* := \sigma(X_{n+1}, X_{n+2}, \dots)$。我们定义 $\mathscr{F}_\infty^* := \bigcap_{n\geq 0} \mathscr{F}_n^*$。它被形象的称为尾代数，而 $\mathscr{F}_\infty^*$ 中的事件被称为尾事件。

尾代数的定义看起来有一些抽象，根据定义，它里面的事件满足 “发生与否与任意前面有限个 X_n 无关”。实际上，几乎所有关于 X_n 序列极限的事件都是尾事件，正如我们在大数定律以及 Borel-Cantelli 里面遇到的那样（why?）。

Kolmogorov 0-1 律是下面这个有些惊人的结论。

定理 21.1

设 $X_1, X_2, \dots$ 是一列独立的随机变量。那么其任意尾事件发生的概率要么是 0 要么是 1。 ♡

证明 取一个尾事件 $B \in \mathscr{F}_\infty^*$。我们定义

$$\mathscr{G} = \{A \in \mathscr{F} : \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)\}.$$

我们接下来的目标是说明 B 自己也属于 $\mathscr{G}$，也就是说 $\mathbb{P}(B) = \mathbb{P}(B)^2$。因此 $\mathbb{P}(B) = 0$ or 1。

注意到，对于每一个 $n \geq 0$，我们有 $\mathscr{F}_n$ 和 $\mathscr{F}_n^*$ 是独立的（因为它们分别涉及不相交的独立随机变量）。而 $B \in \mathscr{F}_n^*$，所以 $\mathscr{F}_n \subseteq \mathscr{G}$。于是 $\bigcup_n \mathscr{F}_n \subseteq \mathscr{G}$。我们想说明 $\mathscr{F}_\infty = \sigma(\bigcup_n \mathscr{F}_n) \subseteq \mathscr{G}$。由于 $\bigcup_n \mathscr{F}_n$ 是一个代数而不一定是一个 σ-代数（why？回忆我们作业里投掷无穷硬币的概率空间的例子），根据单调类定理，我们只需要说明 $\mathscr{G}$ 是一个单调类就行了。设 $A_1 \subseteq A_2 \subseteq \dots \in \mathscr{G}$，那么

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \cap B\right) = \lim_{n\to\infty} \mathbb{P}(A_n \cap B) = \lim_{n\to\infty} \mathbb{P}(A_n) \cdot \mathbb{P}(B) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \cdot \mathbb{P}(B).$$

这说明 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{G}$。对于 $A_1 \supseteq A_2 \supseteq \dots$ 的情况也类似可以说明。于是，$\mathscr{G}$ 是单调类。

但显然（Why?），$B \in \mathscr{F}_\infty$。因此，我们有 $B \in \mathscr{G}$，证明结束。

第 22 章 依分布收敛，De Moivre 中心极限定理

22.1 中心极限定理的动机

我们之前讨论了大数定律，其想说的事情是给定一系列随机变量 $X_1, X_2, \ldots$，如果每一个 X_i 的期望均是 μ，那么在某些条件下，其部分和 $S_n = \sum_{i\in[n]} X_i$ 满足 $\frac{S_n}{n}$ 会收敛到 μ。注意到，每一个 $\frac{S_n}{n}$ 都是一个随机变量，而 μ 本身是一个固定数（常值随机变量）。为什么随机变量会收敛到一个固定的数呢？如果我们再假设每一个 X_i 满足 $\mathbf{Var}[X_i] \le \sigma^2$，并且它们是相互独立的，那么我们知道

$$\mathbf{Var}\left[\frac{S_n}{n}\right] = \frac{\sum_{i=1}^n \mathbf{Var}[X_i]}{n^2} \le \frac{\sigma^2}{n} \to 0.$$

也就是说，我们把 S_n 除 n 的操作使得当 n 足够大时，$\frac{S_n}{n}$ 的方差趋向于零了，因此，其也自然的收敛到一个常数。

一个很自然的问题是，对于任意一个关于 n 的递增函数 $f(n)$，随机变量 $\frac{S_n}{f(n)}$ 的收敛情况是怎么样的呢。简单计算就知道，当 $f(n) = \omega(\sqrt{n})$ 的时候，$\mathbf{Var}\left[\frac{S_n}{f(n)}\right]$ 同样收敛到 0；而当 $f(n) = o(\sqrt{n})$ 的时候，其方差会趋向于无穷大。对于这样的 $f(n)$，有一些有趣的性质可以讨论（我们在作业中会遇到），但最有意思的事情发生在 $f(n) = \sqrt{n}$ 的时候，这便是中心极限定理所讨论的问题。

我们先严格化我们的设定。假设 $X_1, X_2, \ldots$ 是定义在同一个概率空间上的独立同分布的随机变量，满足 $\mathbf{E}[X_1] = \mu$，$\mathbf{Var}[X_1] = \sigma^2$ 均是有限的。我们关心 $\frac{S_n}{\sqrt{n}}$ 的极限行为。我们首先证明一个有一些惊人的结论：

命题 22.1

如果 $\mathbf{E}[X_1] = 0$ 并且 $\mathbf{E}\left[X_1^4\right] < \infty$，那么不存在一个随机变量 X 使得 $\frac{S_n}{\sqrt{n}} \xrightarrow{P} X$。 ♠

我们假设这样一个 X 存在。前面的课上证明过 $\frac{S_n}{\sqrt{n}} \xrightarrow{P} X$ 可以推出一定存在一个子序列 $\{n_j\}$ 满足 $\frac{S_{n_j}}{\sqrt{n_j}} \xrightarrow{a.s.} X$。由于我们知道对于每一个 j，$\mathbf{E}\left[\frac{S_{n_j}}{\sqrt{n_j}}\right] = 0$ 并且 $\mathbf{Var}\left[\frac{S_{n_j}}{\sqrt{n_j}}\right] = \sigma^2$。下面的引理可以保证 X 一定也满足 $\mathbf{E}[X] = 0$，$\mathbf{Var}[X] = \sigma^2$。这是一个类似 DCT 和 MCT 的 yet another 保证极限和期望可以交换的充分条件，我们把它的证明放在本次讲义最后（事实上，极限和期望交换的充要条件是所谓的“一致可积（uniform integrability）”，因为课时原因我们不再介绍）。

引理 22.1

设 $X_1, X_2, \ldots$ 是一族随机变量满足 $X_n \xrightarrow{a.s} X$。如果存在 $\varepsilon > 0$ 和常数 M，使得对于每一个 n，$\mathbf{E}\left[|X_n|^{1+\varepsilon}\right] \leq M$，那么

$$\lim_{n\to\infty} \mathbf{E}\left[X_n\right] = \mathbf{E}\left[X\right].$$

♡

但是另一方面，我们知道 X 是一个所谓的“尾变量”，也就是说，对于任何 Borel 集 $B \in \mathscr{B}(\mathbb{R})$，事件 $[X \in B]$ 都是 $\mathscr{F}_\infty^*$ 中的一个尾事件，这儿显然 $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$。根据 Kolmogorov 0-1 律，$[X \in B]$ 发生的概率要么是 0 要么是 1。这也说明 X 一定等于某一个常数。但这与它的方差是 $\sigma^2 > 0$ 矛盾。

上面的讨论说明，我们不能期待对于 $\frac{S_n}{\sqrt{n}}$ 在依概率收敛或者几乎处处收敛的意义上说什么。我们考虑一个更弱的收敛定义，即依分布收敛。回忆我们之前定义过的依分布收敛：设 $X_1, X_2, \ldots$ 分别有分布函数 F_n，并且 X 有分布函数 F。我们说 $X_n \xrightarrow{D} X$ 当且仅当对于 F 的每一个连续的点 x，有 $\lim_{n\to\infty} F_n(x) = F(x)$。最基本的中心极限定理便是如下结论：

定理 22.1

如果独立同分布的随机变量 $X_1, X_2, \ldots$ 满足 $\mathbf{E}\left[X_1\right] = \mu, \mathbf{Var}\left[X_1\right] = \sigma^2$ 均为有限的，那么

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} Y \sim \mathcal{N}(0,1).$$

♡

换句话说，如果我们把 X_i 归一化成期望为 0 方差为 1 的随机变量（即 $X_i' = \frac{X_i - \mu}{\sigma}$），并令 $S_n' = \sum_{i=1}^n X_i'$ 为归一化后的部分和。那么 $\frac{S_n'}{\sqrt{n}}$ 依分布收敛到标准正态分布。

$$\frac{\sum_{i=1}^n X_i'}{\sqrt{n}} \xrightarrow{D} Y \sim \mathcal{N}(0,1).$$

相比大数定律，我认为这是一个非常让人意外的结果，因为它并没有对 X_i 的分布有要求，只规定了它的期望和方差。而中心极限定理告诉我们，不管 X_i 本身的分布是什么，当足够多的独立的 X_i 加在一起的时候，一定呈现出正态分布的样子。这也是正态分布在我们生活中经常出现的原因。 但我上课试图解释为什么教务处总希望同学的成绩是正态分布但发现好像不管怎么 model 正态分布需要的条件都不太成立所以教务处为什么

22.2 棣莫弗-拉普拉斯中心极限定理（De Moivre–Laplace theorem）

我们今天先证明一个简单版本的中心极限定理，即当每一个 X_i 都具有独立同分布的伯努利分布的情形。它被称作棣莫弗-拉普拉斯中心极限定理。我们证明的方法也很暴力，就是直接计算并估计 $\frac{S_n}{\sqrt{n}}$ 的分布函数。在这个证明中使用的一些估计技巧是非常有用并且常见的。

我们严格的陈述一下要证明的结论。设 $X_1, X_2, \ldots$ 是独立的满足 $\texttt{Ber}(p)$ 分布的随机变量，其中 $p \in (0,1)$ 是一个常数，那么

$$\frac{S_n - pn}{\sqrt{(p-p^2)n}} \xrightarrow{D} Y \sim \mathcal{N}(0,1).$$

为了方便，我们只证明 $p = \frac{1}{2}$ 的情况，对于一般的 p 可以完全类似的证明。

回忆我们用 $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ 来表示标准正态分布 $\mathcal{N}(0,1)$ 的概率密度函数。事实上，我们知道 S_n 是满足 $\texttt{Binom}(n, \frac{1}{2})$ 分布的。因此，我们可以直接计算出它的概率质量函数 $p_n(k) := \mathbb{P}\left(S_n = k\right) = \binom{n}{k} 2^{-n}$。我们实际上想证明的是，当 n 足够大的时候，$p_n(k)$ 所决定的离散的点列（图 1），和 $\phi(x)$ 对应的曲线（图 2），在做适当放缩之后是逐渐趋向于一致的（图 3）。

便让我们来严格证明这件事情。我们首先说明，当 k 比较接近其平均值 $\frac{n}{2}$ 时，$p_n(k)$ 的值和进行合适放缩后的 ϕ 的值的差距是非常小的。

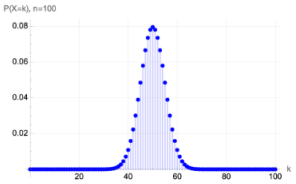


图 1

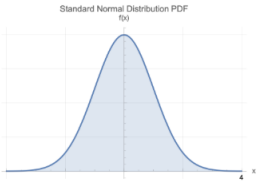


图 2

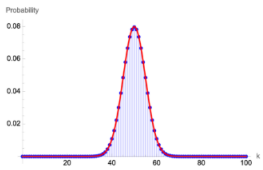


图 3

引理 22.2

设 $C>0$ 是一个常数。那么

$$\max_{|k-n/2|\leq C\cdot\sqrt{n}}\left|\frac{p_n(k)\cdot\sqrt{n/4}}{\phi\left(\frac{k-n/2}{\sqrt{n/4}}\right)}-1\right|=\mathcal{O}\left(n^{-1}\right).$$

♡

我们先解释一下上面引理里是如何缩放 $\phi(x)$ 的。对于 $k\sim\texttt{Binom}(n,\frac{1}{2})$，我们知道其期望是 $n/2$，方差是 $n/4$。因此 $\frac{k-n/2}{\sqrt{n/4}}$ 就是一个期望为 0 方差为 1 的随机变量，所以 $p_n(k)$ 应该是和 $\phi\left(\frac{k-\frac{n}{2}}{\sqrt{n/4}}\right)$ 相比较。但此时，我们需要把它除掉 $\sqrt{n/4}$ 才能保证其（对 k 的）积分为 1。

我们证明引理的主要工具就是斯特林公式：$n!=\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1+\mathcal{O}\left(n^{-1}\right)\right)$。使用这个公式以及 $k\approx\frac{n}{2}$ 的事实，我们可以得到

$$\begin{aligned}p_n(k)&=\binom{n}{k}2^{-n}\\&=\frac{n!}{k!(n-k)!}2^{-n}\\&=\frac{\sqrt{2\pi n}}{\sqrt{2\pi k}\sqrt{2\pi(n-k)}}\frac{n^n}{k^k(n-k)^{n-k}}\cdot 2^{-n}\cdot\left(1+\mathcal{O}\left(n^{-1}\right)\right)\\&=\frac{1}{\sqrt{2\pi n}}\cdot\frac{1}{\sqrt{k/n}\cdot\sqrt{1-k/n}}\cdot\frac{2^{-n}}{(k/n)^k\left(1-k/n\right)^{n-k}}\cdot\left(1+\mathcal{O}\left(n^{-1}\right)\right).\end{aligned}$$

我们的目标是把上式和 $\phi\left(\frac{k-n/2}{\sqrt{n/4}}\right)=\exp\left(-\frac{1}{2n}\cdot(2k-n)^2\right)$ 进行比较。我们的策略如下：我们知道 $k\approx\frac{n}{2}$，所以我们把分母凑出形如 $2k/n\approx 1$ 的项，然后尝试把这些接近于 1 的项用 $1+x\approx e^x$ 换成指数形式，并期待能够变成 $\phi(\cdot)$ 的形式。因此，我们可以继续变形得到

$$p_n(k)=\sqrt{\frac{2}{\pi n}}\cdot\frac{1}{\sqrt{1+\frac{2k-n}{n}}\cdot\sqrt{1-\frac{2k-n}{n}}}\cdot\frac{1}{\left(1+\frac{2k-n}{n}\right)^k\cdot\left(1-\frac{2k-n}{n}\right)^{n-k}}\cdot\left(1+\mathcal{O}\left(n^{-1}\right)\right).$$

注意到，我们目标 $\phi(\cdot)$ 里出现的是形如 $e^{-(2k-n)^2}$ 的项，其指数上关于 $(2k-n)$ 的依赖是二次的。我们仅仅使用 $1\pm x\approx e^{\pm x}$ 这个一阶近似是不够的，所以我们来计算 $1\pm x$ 的二阶近似。

事实上我们注意到由于 $k\approx n/2$，所以 $\left(1+\frac{2k-n}{n}\right)^k\cdot\left(1-\frac{2k-n}{n}\right)^{n-k}\approx\left(1-\left(\frac{2k-n}{n}\right)^2\right)^k\approx e^{-k\left(\frac{2k-n}{n}\right)^2}$。直接估计这儿 $\approx$ 的误差也可以，但我认为下面的技巧更有通用性。

使用 $\log(1+x)=x-\frac{x^2}{2}+\mathcal{O}\left(x^3\right)$，我们有 $1+x=e^{\log(1+x)}=e^{x-\frac{x^2}{2}}\cdot\left(1+\mathcal{O}\left(|x|^3\right)\right)$。注意我们这里要利用 $\left|k-\frac{n}{2}\right|=C\cdot\sqrt{n}$ 仔细估计误差项。可以得到

$$p_n^k=\sqrt{\frac{2}{\pi n}}\cdot e^{\frac{1}{2}\cdot\left(\frac{2k-n}{n}\right)^2}\cdot e^{-\frac{1}{2}\cdot\frac{(2k-n)^2}{n}}\left(1+\mathcal{O}\left(n^{-1}\right)\right)=\sqrt{\frac{2}{\pi n}}\cdot e^{-\frac{1}{2}\cdot\frac{(2k-n)^2}{n}}\cdot\left(1+\mathcal{O}\left(n^{-1}\right)\right).$$

而这正是 $\frac{\phi\left(\frac{k-n/2}{\sqrt{n/4}}\right)}{\sqrt{n/4}}\left(1+\mathcal{O}\left(n^{-1}\right)\right)$。

设 F_n 是 $\frac{S_n-\frac{n}{2}}{\sqrt{n/4}}$ 的分布函数。我们接下来只需要说明（why?）对于任何固定的常数 $a<b\in\mathbb{R}$，我们有 $\lim_{n\to\infty}F_n(b)-F_n(a)=\int_a^b\phi(x)\mathrm{d}x$。

注意到

$$\left|F_n(b)-F_n(a)-\int_a^b \phi(x)\mathrm{d}x\right| \leq \left|F_n(b)-F_n(a)-\sum_{k\in\mathbb{N}:\ a\leq\frac{k-n/2}{\sqrt{n/4}}\leq b}\frac{\phi\left(\frac{k-n/2}{\sqrt{n/4}}\right)}{\sqrt{n/4}}\right| + \left|\sum_{k\in\mathbb{N}:\ a\leq\frac{k-n/2}{\sqrt{n/4}}\leq b}\frac{\phi\left(\frac{k-n/2}{\sqrt{n/4}}\right)}{\sqrt{n/4}}-\int_a^b \phi(x)\mathrm{d}x\right|$$

上式前一项的求和里最多有 $\mathcal{O}(\sqrt{n})$ 项，根据我们前面证明的引理，这一项带来的误差是 $\frac{1}{\sqrt{n}}$ 级别。而上式第二项实际上是给出了积分和它的黎曼和的差，因此在 n 趋向于无穷大的时候趋向于 0。这便证明了我们的结果。

22.2.1 省略的证明

我们现在来证明

引理 22.3

设 $X_1, X_2, \ldots$ 是一族随机变量满足 $X_n \xrightarrow{a.s} X$。如果存在 $\varepsilon > 0$ 和常数 M，使得对于每一个 n，$\mathbf{E}\left[|X_n|^{1+\varepsilon}\right] \leq M$，那么

$$\lim_{n\to\infty}\mathbf{E}\left[X_n\right]=\mathbf{E}\left[X\right].$$

♡

这个证明来自这个 notes。

对于任意 $m>0$ 和随机变量 Y，我们定义一个新的随机变量

$$Y^{[m]}(\omega):=\begin{cases} m & \text{if } Y(\omega)>m \\ -m & \text{if } Y(\omega)<-m \\ Y(\omega) & \text{if } Y(\omega)\in[m,-m]. \end{cases}$$

即 $Y^{[m]}$ 把 Y 大于 m 和小于 $-m$ 的部分分别换成 m 和 $-m$。于是 $\left|Y^{[m]}\right|\leq m$。根据 DCT，我们显然有

$$\lim_{n\to\infty}\mathbf{E}\left[X_n^{[m]}\right]=\mathbf{E}\left[X^{[m]}\right].$$

根据定义，我们又有

$$\left|X_n-X_n^{[m]}\right|\leq\left(\frac{|X_n|}{m}\right)^{\varepsilon}|X_n|=m^{-\varepsilon}|X_n|^{1+\varepsilon}.$$

于是，$\mathbf{E}\left[X_n\right]=\mathbf{E}\left[X_n^{[m]}\right]\pm\mathcal{O}\left(m^{-\varepsilon}M\right)$。

根据 Fatou 引理，我们有

$$\mathbf{E}\left[|X|^{1+\varepsilon}\right]\leq\mathbf{E}\left[\liminf|X_n|^{1+\varepsilon}\right]\leq\liminf\mathbf{E}\left[X_n^{1+\varepsilon}\right]=M.$$

所以我们使用类似的推理可以得到

$$\mathbf{E}\left[X\right]=\mathbf{E}\left[X^{[m]}\right]\pm\mathcal{O}\left(m^{-\varepsilon}M\right).$$

于是

$$\limsup_{n\to\infty}\mathbf{E}\left[X_n\right],\liminf_{n\to\infty}\mathbf{E}\left[X_n\right]=\lim_{n\to\infty}\mathbf{E}\left[X_n^{[m]}\right]\pm\mathcal{O}\left(m^{-\varepsilon}M\right)=\mathbf{E}\left[X\right]\pm\mathcal{O}\left(m^{-\varepsilon}M\right).$$

令 $m\to\infty$ 便得证。

第 23 章 高维概率，协方差矩阵，高斯分布

23.1 高维概率

我们之前讨论过定义在同一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的两个随机变量 X 和 Y 的联合分布。除了可以定义各自的期望 $\mathbf{E}[X], \mathbf{E}[Y]$ 方差 $\mathbf{Var}[X], \mathbf{Var}[Y]$ 之外，我们还能够定义协方差（covariance）的概念

$$\mathbf{Cov}[X,Y] := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

从定义式就可以看出，协方差是方差概念的推广，因为 $\mathbf{Cov}[X,X] = \mathbf{Var}[X]$。它可以用来衡量随机变量之间的“相关度”。从定义式 $\mathbf{Cov}[X,Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$ 可以看出，如果 X 和 Y 独立，那么 $\mathbf{Cov}[X,Y] = 0$。但是反过来一般不成立:

例题 23.1.

设 $X \sim \mathtt{Ber}\left(\frac{1}{2}\right)$; $Y = \begin{cases} \frac{1}{2} & \text{if } X = 0 \\ \sim \mathtt{Ber}\left(\frac{1}{2}\right) & \text{if } X = 1 \end{cases}$。那么

$$\mathbf{E}[XY] = \mathbf{E}[XY \mid X = 0]\,\mathbb{P}(X = 0) + \mathbf{E}[XY \mid X = 1]\,\mathbb{P}(X = 1) = \frac{1}{4};$$

并且 $\mathbf{E}[X]\mathbf{E}[Y] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$。所以 $\mathbf{Cov}[X,Y] = 0$。但是显然 X 和 Y 不独立，因为

$$\mathbb{P}\left(X = 0 \wedge Y = \frac{1}{2}\right) = \frac{1}{2}, \quad \mathbb{P}(X = 0)\,\mathbb{P}\left(Y = \frac{1}{2}\right) = \frac{1}{4}.$$

我们同样可以把相关概念推广到 n 个随机变量。我们会用一个向量 $\mathbf{X} = (X_1, \ldots, X_n)^\top$ 来表示 n 个随机变量 $X_1, \ldots, X_n$ 所组成的向量，并把它看成 $\omega \in \Omega \mapsto \mathbf{X}(\omega) \in \mathbb{R}^n$ 中的函数。我们有时候也直接称 $\mathbf{X}$ 为随机向量。我们定义它的期望为

$$\mathbf{E}[\mathbf{X}] = (\mathbf{E}[X_1], \mathbf{E}[X_2], \ldots, \mathbf{E}[X_n])^\top \in \mathbb{R}^n.$$

高维随机变量在计算机科学和数据科学中都是核心的研究对象，我们这门课不会涉及太多相关的内容，感兴趣的同学可以参考两本著名的教科书 High-Dimensional Probability 和 Probability in High Dimension。其中前者从统计与计算机科学的角度研究高维概率，而后者从概率论的角度研究类似的对象。

对于一个 n-维随机变量，我们可以定义它的协方差矩阵（covariance matrix）为

$$\mathbf{Cov}\left[\mathbf{X}\right] := \mathbf{E}\left[(\mathbf{X}-\mathbf{E}\left[\mathbf{X}\right])(\mathbf{X}-\mathbf{E}\left[\mathbf{X}\right])^{\top}\right].$$

换句话说，$\mathbf{Cov}\left[\mathbf{X}\right]$ 是一个 $n\times n$ 的矩阵，其第 i 行 j 列的元素是 $\mathbf{Cov}\left[X_i, X_j\right]$。协方差矩阵有一些基本的性质：

1. $\mathbf{Cov}\left[\mathbf{X}\right]$ 总是半正定的，这是因为对于任何 $\mathbf{x}\in\mathbb{R}^n$，
$$\mathbf{x}^{\top}\mathbf{Cov}\left[\mathbf{X}\right]\mathbf{x} = \mathbf{x}^{\top}\mathbf{E}\left[(\mathbf{X}-\mathbf{E}\left[\mathbf{X}\right])(\mathbf{X}-\mathbf{E}\left[\mathbf{X}\right])^{\top}\right]\mathbf{x} = \mathbf{E}\left[\langle\mathbf{x}, \mathbf{X}-\mathbf{E}\left[\mathbf{X}\right]\rangle^2\right] \geq 0.$$
2. 如果 $X_1,\ldots,X_n$ 两两独立，那么 $\mathbf{Cov}\left[\mathbf{X}\right]$ 是对角阵，并且其对角线第 i 位的元素是 $\mathbf{Var}\left[X_i\right]$。

注意到，我们以前介绍过的联合分布函数和联合密度函数均可以无缝推广到 n-维随机变量上。

23.2 高斯分布

我们开始介绍也许是最重要的高维分布，高维的高斯分布。对于任意 $i\in[n]$，我们定义 ξ_i 为一个独立的 $\mathcal{N}(0,1)$ 随机变量，$\xi=(\xi_1,\ldots,\xi_n)$。我们把它的分布记作 $\mathcal{N}(0,\mathtt{Id}_n)$，其中 $\mathtt{Id}_n$ 是 n-维的单位矩阵，它是 ξ 的协方差矩阵。

我们在 ξ 的基础上定义一般的高维高斯向量。我们说一个向量 $\mathbf{X}$ 是高维高斯（*multi-dimensional Gaussian random variable*），如果他可以写成 $\mathbf{X}=A\xi+\mu$ 的形式，其中 $A\in\mathbb{R}^{n\times n}$，$\mu\in\mathbb{R}^n$。

我们可以计算一下 $\mathbf{X}$ 的期望和协方差。首先

$$\mathbf{E}\left[\mathbf{X}\right] = \mathbf{E}\left[A\xi+\mu\right] = A\mathbf{E}\left[\xi\right]+\mu = \mu.$$

其次

$$\mathbf{Cov}\left[\mathbf{X}\right] = \mathbf{E}\left[(\mathbf{X}-\mathbf{E}\left[X\right])(\mathbf{X}-\mathbf{E}\left[\mathbf{X}\right])^{\top}\right] = \mathbf{E}\left[(A\xi)(A\xi)^{\top}\right] = A\mathbf{E}\left[\xi\xi^{\top}\right]A^{\top} = AA^{\top}.$$

如果我们定义 $\Sigma=AA^{\top}$，那么 $\mathbf{X}$ 就是一个期望为 μ，协方差矩阵为 Σ 的随机向量。我们把它的分布记为 $\mathcal{N}(\mu,\Sigma)$。可以看到，一个高维的高斯向量，其分布由期望和协方差矩阵唯一确定（这对于一般的随机向量显然是不对的）。

我们接下来推导 $\mathbf{X}\sim\mathcal{N}(\mu,\Sigma)$ 的联合密度函数。我们知道，对于一个 $\mathcal{N}(0,1)$ 的标准高斯随机变量，其概率密度函数为 $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$。于是，对于 $\xi\sim\mathcal{N}(0,\mathtt{Id}_n)$，由于其各个维度均是独立的 $\mathcal{N}(0,1)$ 随机变量，其概率密度函数为 $\phi_{\xi}(x_1,\ldots,x_n)=(2\pi)^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)$。对于一般的 $\mathbf{X}\sim\mathcal{N}(\mu,\Sigma)$，我们先贷款下面这个结论（将在下次课证明）：

引理 23.1

设 F 和 G 是两个分布函数。如果对于任何一个定义在紧集上的光滑函数 h，都有 $\int_{\Omega}h\,dF=\int_{\Omega}h\,dG$，那么在那些 $F(x)$ 连续的点 x 上，$F(x)=G(x)$。 ♡

在上述引理的加持下，我们计算对于一个定义在紧集上的光滑函数 h 的期望 $\mathbf{E}\left[h(\mathbf{X})\right]$。设 $\mathbf{X}$ 的密度函数是 $\phi_{\mathbf{X}}$，那么由 LOTUS：

$$\mathbf{E}\left[h(\mathbf{X})\right] = \int_{\mathbb{R}^n} h(\mathbf{x})\phi_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

我们做换元 $\mathbf{x}=A\mathbf{y}+\mu$，并且注意到这个线性变换的 Jacobian 就是 A，于是

$$\mathbf{E}\left[h(\mathbf{X})\right] = \int_{\mathbb{R}^n} h\left(A\mathbf{y}+\mu\right)\phi_{\mathbf{X}}(A\mathbf{y}+\mu)\left|\det A\right|\,\mathrm{d}\mathbf{y}$$

从另外一方面来说，如果我们按照 $\mathbf{y}\sim\mathcal{N}\left(0,\mathtt{Id}_n\right)$ 来积分函数 $h(A\mathbf{y}+\mu)$，可以同样得到 $\mathbf{E}\left[h(\mathbf{X})\right]$：

$$\mathbf{E}\left[h(\mathbf{X})\right] = \int_{\mathbb{R}^n} h\left(A\mathbf{y}+\mu\right)\phi_{\xi}(\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

比较上面的系数，我们可以得到 $\phi_{\mathbf{X}}(A\mathbf{y}+\mu)\,|\det A| = \phi_{\xi}(\mathbf{y})$，或者等价的

$$\phi_{\mathbf{X}}(\mathbf{x}) = |\det A|^{-1}\,\phi_{\xi}\left(A^{-1}(\mathbf{x}-\mu)\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\,(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right).$$

同样，我们可以看出来，如果 Σ 是对角阵，那么 $X_1,\ldots,X_n$ 是相互独立的。这是高斯向量特有的性质。

23.2.1 高斯分布的和

我们现在证明，对于两个**独立**的高斯分布，$X_1\sim\mathcal{N}(\mu_1,\sigma_1^2)$，$X_2\sim\mathcal{N}(\mu_2,\sigma_2^2)$，它们的和 $X_1+X_2\sim\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$。设 $Z=X_1+X_2$，我们直接计算 Z 的概率密度函数 f_Z。我们之前计算过两个随机变量的和的概率密度公式：

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_X(x)f_Y(y-x)\,\mathrm{d}x \\
&= \frac{1}{2\pi\cdot\sigma_1\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(z-x-\mu_2)^2}{2\sigma_2^2}}\,\mathrm{d}x \\
&= \frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}\int_{-\infty}^{\infty} e^{-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(x-\frac{\sigma_2^2\mu_1+\sigma_1^2(z-\mu_2)}{\sigma_1^2+\sigma_2^2}\right)^2} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left(-\frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}\right).
\end{aligned}$$

这便是 $\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ 的概率密度函数。上述结论可以推广到任意个 n 个**相互独立**的高斯分布之和，即如果 $X_1,\ldots,X_n$ 是相互独立且 $X_i\sim\mathcal{N}(\mu_i,\sigma_i^2)$，那么

$$\sum_{i\in[n]}X_i\sim\mathcal{N}\left(\sum_{i\in[n]}\mu_i,\sum_{i\in[n]}\sigma_i^2\right).$$

注意到，上面要求的相互独立是必须的。如果两个高斯随机变量变量 X_1，X_2 不独立，那么它们的和不一定是高斯随机变量（你能想到反例吗？）。事实上，如果我们要求 X_1 和 X_2 的任意线性组合都是一个高斯变量，这等价于要求 (X_1,X_2) 是一个二维高斯向量。我们会在未来证明这给出了高维高斯向量的另一个等价定义。

23.3 最大割的近似算法

我们小讲一个高维向量在算法设计里面应用。我们之前研究过在一个图上求“最小割”的问题。我们今天来研究它的姊妹问题 - “最大割”。给定一个无向图 $G=(V,E)$，我们想把顶点集 V 分成两部分 S 和 $V\backslash S$，满足 S 和 $V\backslash S$ 之间的边尽量多。在这里 $(S,V\backslash S)$ 就被称为图上的一个割。它的大小定义为 $|\{\{u,v\}\in E:u\in S,v\in V\setminus S\}|$。

和最小割问题不一样，最大割问题是 NP-hard 的，也就是说，如果 $\mathbf{NP}\neq\mathbf{P}$，那么最大割问题不存在多项式时间的算法能找到最优解。因此，我们期待在多项式时间内找到近似解。严格来说，我们说一个算法是 α-近似（$\alpha\in[0,1]$）的，当且仅当给定一个图作为输入之后，如果其最大割的值本身是 $\texttt{OPT}$，算法可以输出一个割，其大小至少为 $\alpha\cdot\texttt{OPT}$。如果 $\alpha=1$，那这就是一个最优算法。我们希望算法能够保证的 α 越大越好。

这是一个经典的组合优化问题，我们将介绍使用半正定规划得到的一个近似算法，这个算法大家猜想是最优的（in terms of α）。

23.3.1 半正定规划（Positive Semi-Definite Programming, SDP）

我们先简单介绍半正定规划，它是线性规划的推广。我们在算法或者优化课中学过线性规划：

$$\begin{aligned}
&\text{s.t. } x+y\le 2,\\
\max\ 2x-3y\quad &3x-y\le 1,\\
&x,y\ge 0.
\end{aligned}$$

我们可以将其改写为等价的矩阵形式：

$$\begin{aligned} &\text{s.t.} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \leq 2, \\ \max \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix} \bullet \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \quad &\begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \bullet \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \leq 1, \\ &\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix} \succeq 0. \end{aligned}$$

在这里，对于两个 $n \times n$ 矩阵 $A = (a_{i,j})_{1 \leq i,j \leq n}$ 和 $B = (b_{i,j})_{1 \leq i,j \leq n}$，它们 Frobenius 积定义为：

$$A \bullet B = \sum_{1 \leq i,j \leq n} a_{i,j} \cdot b_{i,j}.$$

在上述例子中，每个矩阵都为对角矩阵，并附加了一个正半定约束 $X \succeq 0$。

23.3.1.1 SDP 的一般形式

如上所示，标准形式的线性规划可以写成矩阵形式，其中线性规划的变量 $\{x_i\}_{i \in n}$ 被收集在一个对角矩阵 $X = \text{diag}(x_1, \ldots, x_n)$ 中。正半定规划将对角矩阵 X 推广到任意对称矩阵：

$$\max C \bullet X \text{s.t. } A_k \bullet X \leq b_k, \forall k \in [m], X \succeq 0.$$

其中，$C = (c(i,j))_{1 \leq i,j \leq n}$，$X = (x(i,j))_{1 \leq i,j \leq n}$，$A_k = (a_k(i,j))_{1 \leq i,j \leq n}$ 是 $n \times n$ 的矩阵，$k \in [m]$。

我们知道一个半正定矩阵 X 可以写成 $X = U^\top U$，其中 $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ 满足 $\mathbf{u}_i \in \mathbb{R}^n$ 是一个向量。于是 $x(i,j) = \mathbf{u}_i^\top \mathbf{u}_j$。我们可以使用 $\mathbf{u}_i$ 们重写上面的半正定规划，从而得到

$$\max \sum_{1 \leq i,j \leq n} c(i,j) \cdot \mathbf{u}_i^\top \mathbf{u}_j, \text{s.t.} \sum_{1 \leq i,j \leq n} a_k(i,j) \cdot \mathbf{u}_i^\top \mathbf{u}_j \leq b_k, \forall k \in [m], \mathbf{u}_i \in \mathbb{R}^n, i \in [n].$$

这被称为向量规划，与半正定规划是等价的。与线性规划类似，只要提供一个高效的 seperation oracle，就可以使用椭球法（ellipsoid method）在多项式时间内（近似）求解。一个 seperation oracle 指的是如下一个算法：给定一个点 $\mathbf{x} \in \mathbb{R}^n$：

1. 如果 $\mathbf{x}$ 是可行解，算法输出 YES；
2. 如果 $\mathbf{x}$ 不是可行解，算法输出一个它所不满足的约束。

我们并不想详细介绍 SDP，我们只是把上述结果当成黑盒使用。

23.3.2 Goemans-Williamson 的舍入算法

实际上，我们可以把最大割问题自然的建模成一个二次规划（quadratic programming）：

$$\max \frac{1}{2} \sum_{e=\{u,v\} \in E} (1 - x_u x_v) \text{s.t. } x_u \in \{-1, 1\}, \forall u \in V.$$

对于这样一个规划，由于其与最大割问题是等价的，我们知道它也是 NP-hard 的。近似算法设计的一个常用技巧是，把这个规划放松（relax）成一个取值范围更大，也因此更容易解的优化问题。然后从这个容易解的优化问题的最优解得到原规划的最优解。后面一步通常被称为舍入（rounding）。在我们这个问题里，如果将每个 x_u 视为一个一维向量 $x_u \in \mathbb{R}^1$，我们可以将其放松为一个 n 维向量 $\mathbf{w}_u \in \mathbb{R}^n$。于是，我们可以得到如下向量规划：

$$\max \frac{1}{2} \sum_{e=\{u,v\} \in E} \left(1 - \mathbf{w}_u^\top \mathbf{w}_v\right) \text{s.t. } \mathbf{w}_u \in \mathbb{R}^n, \forall u \in V; \quad \mathbf{w}_u^\top \mathbf{w}_u = 1, \forall u \in V$$

使用之前提到的求解半正定规划的黑盒，我们可以高效的求解上述规划，并得到向量规划的最优解 $\{\mathbf{w}_u^*\}_{u\in V}$。注意到，每一个 $\mathbf{w}_u^*$ 均是 $\mathbb{R}^n$ 中的一个向量。直观上，这些向量包含了一个好的割的信息：如果 $\mathbf{w}_u^\top \mathbf{w}_v$ 比较小，说明 u 和 v 更应该被割开。那我们如何从它舍入成一个割，并严格的证明其近似比呢?

Goemans-Williamson 舍入算法通过随机采样一个穿过原点的超平面，将向量分为两部分：超平面一侧的向量，对应于 S，和另一侧的向量，对应于 $V\setminus S$。于是，角度较大的向量对更可能被分开，这也与我们前面说的直观是一致的。

23.3.2.1 如何随机采样超平面？

这并不是一个简单的问题，因为我们首先要定义随机超平面的概率空间，这稍微有一点麻烦。我们便如同大部分计算机科学中做的一样，稍微不那么严格一点，并把正确性的验证付诸于直观（但我并不支持这样做，大家可以参看前文提到的 High-Dimensional Probability 教材来寻找严格的处理）。我们做的具体方法是从 $n-1$ 维单位球（$S^{n-1}=\{x\in\mathbb{R}^n : \|x\|=1\}$）中均匀地采样一个点作为超平面的法向量。$S_{n-1}$ 上的均匀测度是一个 Haar 测度，它是存在的，我们暂时接受这个设定。我们现在来说明如何从这个测度中进行采样。事实上，我们只需要取一个 $\mathbf{x}\sim\mathcal{N}(0,\mathtt{Id}_n)$，然后输出 $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ 即可。这是因为我们知道

$$\phi_\xi(\mathbf{x})\propto\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right),$$

仅仅依赖于 $\mathbf{x}$ 的模长 $\|\mathbf{x}\|$。因此归一化后，$\frac{\mathbf{x}}{\|\mathbf{x}\|}$ 在 S^{n-1} 上均匀分布。

例题 23.2. Goemans-Williamson 舍入算法

1. 计算 $\{\mathbf{w}_u^*\}_{u\in V}$。
2. 随机选择一个向量 $\mathbf{x}=(x_1,\dots,x_n)\in S^{n-1}$。
3. 定义集合 $S=\left\{u\in V : \mathbf{x}^\top\mathbf{w}_u^*\geq 0\right\}$。

定理 23.1

Goemans-Williamson 舍入算法是最大割问题的一个随机 α^*-近似算法，其中 $\alpha^*>0.878$。♡

记 w_u^* 和 w_v^* 之间的夹角为 $\theta_{u,v}$，即 $\theta_{u,v}=\arccos(\langle w_u^*,w_v^*\rangle)$。由于分隔超平面是均匀选取的，对于任意边 $e=\{u,v\}\in E$，顶点 u 和 v 被分隔（位于超平面两侧）的概率为：

$$\mathbb{P}\left(u \text{ and } v \text{ are seperated}\right)=\frac{\theta_{u,v}}{\pi}.$$

我们用随机变量 X 表示割的大小，则有：

$$\mathbf{E}\left[X\right]=\sum_{\{u,v\}\in E}\mathbb{P}\left(u \text{ and } v \text{ are seperated}\right)=\sum_{\{u,v\}\in E}\frac{\arccos(\mathbf{w}_u^*\cdot\mathbf{w}_v^*)}{\pi}$$

设 $\alpha^*=\min_{-1\leq x\leq 1}\frac{2\arccos x}{\pi(1-x)}>0.878$。则有

$$\mathbf{E}\left[X\right]\geq\alpha^*\cdot\mathtt{OPT\text{-}VP}\geq\alpha^*\cdot\mathtt{OPT}.$$

其中 OPT-VP 是向量规划的最优解，而 OPT 是最大割的最优解。

Goemans-Williamson 舍入算法看似是一个对于这个问题很特殊的算法，也看似有很多改进的空间。同样 0.878 也看似是一个无厘头的数。但在某一些计算复杂性假设下，这个近似比是最优的。

第 24 章 依分布收敛，Lindeberg 证明

我们今天给出中心极限定理的的第一个证明。在开始之前，我们先要介绍依分布收敛的一个等价刻画。回忆一下，我们说一族向量 $X_1, X_2, \dots$ 依分布收敛到 X，记作 $X_n \xrightarrow{D} X$，当且仅当在 $F(x)$ 连续的那些点 x 上，

$$\lim_{n\to\infty} F_n(x) = F(x),$$

其中 $F(x)$ 是 X 的分布函数，F_n 是 X_n 的分布函数。

命题 24.1

以下二者等价

1. 对于每一个只在有界闭集上取非零值的连续函数 h，有 $\lim_{n\to\infty} \mathbf{E}\left[h(X_n)\right] = \mathbf{E}\left[h(X)\right]$。
2. $X_n \xrightarrow{D} X$。

(2) $\Longrightarrow$ (1) 是比较显然的。假设 h 的只在 $[a,b]$ 上取非零值。根据定义，我们对 f 的勒贝格积分可以写成黎曼和

$$\mathbf{E}\left[h(X_n)\right] = \int_{\mathbb{R}} h(x)\,\mathrm{d}F_n(x) = \lim_{m\to\infty} \sum_{k=1}^{m} h(x_{k-1}^{[m]}) \cdot \left(F_n(x_k^{[m]}) - F_n(x_{k-1}^{[m]})\right).$$

这儿对于每一个 $m \in \mathbb{N}$，$x_k^{[m]} = a + (b-a)\cdot\frac{k}{m}$ 是对 $[a,b]$ 间距为 $\frac{b-a}{m}$ 的划分点。我们总是可以假设 $x_k^{[m]}$ 是那些 $F(x)$ 的连续点（why?），于是，当 $n\to\infty$ 的时候，每一个 $F_n(x_k^{[m]})$ 均会收敛到 $F(x_k^{[m]})$。

要说明 (1) $\Longrightarrow$ (2)，我们首先构造一族“测试函数”$\{h_a\}_{a\in\mathbb{R}}$ 满足如果 $F(x)$ 在 $x=a$ 连续，那么 $\lim_{n\to\infty}\mathbf{E}\left[h_a(X_n)\right] = \mathbf{E}\left[h_a(X)\right] \Longrightarrow \lim_{n\to\infty} F_n(a) = F(a)$。我们接着说明可以魔改 $\{h_a\}$ 们得到一族只在有界闭集上取非零值的连续函数作为测试函数们。

显然，我们只需要让 $h_a(x) = \mathbb{I}[x\le a]$ 即可。这样的 $h_a(x)$ 明显不连续，但由于 $x=a$ 是 $F(x)$ 的连续点，我们对于每一个足够小的 $\varepsilon>0$，我们可以定义 $h_{a,\varepsilon}(x)$ 为在 $(\infty, a-\varepsilon)$ 上值为 1，在 $[a,\infty)$ 上值为 0，并在 $[a-\varepsilon, a)$ 上连续的函数即可。当然，这样的 $h_{a,\varepsilon}$ 们的定义域还不是有界闭集，但是对于给定的 X_n 以及任何 $\delta>0$，我们显然可以把它限制到一个有界闭集上得到一个函数 $h'_{a,\varepsilon}$，使得 $\left|\mathbf{E}\left[h_{a,\varepsilon}(X_n)\right] - \mathbf{E}\left[h'_{a,\varepsilon}(X_n)\right]\right| < \delta$（why?）。

事实上，我们可以把上述命题中 (1) 的“连续函数”加强成“光滑函数”（任意阶导数存在），这只需要在我们刚才的构造中，用“光滑”的方式定义 $h_{a,\varepsilon}$ 在 $[a-\varepsilon, a)$ 上的取值即可。

24.1 Lindeberg 对于中心极限定理的证明

我们现在给出中心极限定理的第一个完整证明。我们把定理复述如下：

定理 24.1 (中心极限定理)

如果独立同分布的随机变量 $X_1, X_2, \dots$ 满足 $\mathbf{E}[X_1] = \mu, \mathbf{Var}[X_1] = \sigma^2$ 均为有限的，那么

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} Y \sim \mathcal{N}(0,1).$$

♡

我们不失一般性的假设 $\mu = 0$ 并且 $\sigma = 1$。根据上一节对于依分布收敛的刻画，我们只需要验证对于任意定义在有界闭集上的光滑函数 h 都有

$$\mathbf{E}\left[h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)\right] \overset{n\to\infty}{\to} \mathbf{E}[h(\xi)], \quad \xi \sim \mathcal{N}(0,1).$$

Lindeberg 的方法说的是，假设中心极限定理对一组特殊的随机变量 $Y_1, Y_2, \dots$ 正确，那么我们只需要证明

$$\mathbf{E}\left[h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right)\right] \overset{n\to\infty}{\to} 0$$

即可。显然这样的 Y_i 是存在的，比如我们让每一个 Y_i 是独立的 $\mathcal{N}(0,1)$...

我们首先假设对于每一个 i，$\mathbf{E}\left[|X_i|^3\right] < \infty$。之后我们会说明如何去掉这个条件。

我们要对每一个 $n \geq 1$，直接比较 $h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)$ 和 $h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right)$。我们可以把二者之差写成 telescopically 和：

$$\begin{aligned}
&h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right) - h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right) \\
&= \sum_{k=1}^n h\left(\frac{Y_1 + \cdots + Y_{k-1} + X_k + \cdots + X_n}{\sqrt{n}}\right) - h\left(\frac{Y_1 + \cdots + Y_k + X_{k+1} + \cdots + X_n}{\sqrt{n}}\right).
\end{aligned}$$

因此，对于每一个 $k \in [n]$，我们只需要计算 $h\left(\frac{Y_1+\cdots+Y_{k-1}+X_k+\cdots+X_n}{\sqrt{n}}\right) - h\left(\frac{Y_1+\cdots+Y_k+X_{k+1}+\cdots+X_n}{\sqrt{n}}\right)$，这两个函数的输入求和式只在第 k 项不一样。我们可以不失一般性的假设 $k = n$，并把那些相同的求和项记作 Z，然后计算

$$\mathbf{E}\left[h\left(Z + \frac{X_n}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(Z + \frac{Y_n}{\sqrt{n}}\right)\right].$$

值得注意的是，由于 telescopic 和有 n 项，我们需要把上式估计得到的误差乘上 n 才是 $\mathbf{E}\left[h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right)\right]$ 的误差。

使用泰勒级数，我们可以得到

$$\mathbf{E}\left[h\left(Z + \frac{X_n}{\sqrt{n}}\right)\right] = \mathbf{E}[h(Z)] + \frac{1}{\sqrt{n}}\mathbf{E}[h'(Z)X_n] + \frac{1}{2n}\mathbf{E}\left[h''(Z)X_n^2\right] + \mathcal{O}\left(\frac{1}{n^{3/2}}\mathbf{E}\left[|X_n|^3\right]\right).$$

我们注意到 $\mathbf{E}\left[|X_n|^3\right] < \infty$，并且 $\mathbf{E}\left[h\left(Z + \frac{Y_n}{\sqrt{n}}\right)\right]$ 做同样泰勒展开后前三项是一样的。因此，

$$\mathbf{E}\left[h\left(Z + \frac{X_n}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(Z + \frac{Y_n}{\sqrt{n}}\right)\right] = \mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

这说明

$$\mathbf{E}\left[h\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right)\right] = \mathcal{O}\left(\frac{1}{n^{1/2}}\right).$$

24.1.1 使用截断法去除三阶矩要求

在上面的分析中，我们看到由于需要把最终的误差控制到 $o(1)$，需要把泰勒级数计算到第三项。而也因此需要 $\mathbf{E}\left[|X_i|^3\right] < \infty$ 的条件。如何去掉这个条件呢？我们以前证明大数定律的时候使用过的截断技巧特别擅长处理这种需要随机变量的矩的上界的问题。我们这里再次施展一发。

固定任意 $\varepsilon > 0$。对于每一个 $i \geq 1$，我们设 $X_i = X_i^{\leq} + X_i^{>}$，其中

$$\begin{aligned}
X_i^{\leq} &:= X_i \cdot \mathbb{I}[|X_i| \leq \varepsilon\sqrt{n}] - \mu_n; \\
X_i^{>} &:= X_i \cdot \mathbb{I}[|X_i| > \varepsilon\sqrt{n}] + \mu_n,
\end{aligned}$$

其中 $\mu_n := \mathbf{E}\left[X_i \cdot \mathbb{I}[|X_i| \leq \varepsilon\sqrt{n}]\right]$。换句话说，我们用阈值 $\varepsilon\sqrt{n}$ 截断 X_i。但为了让 $X_i^{\leq}$ 的期望为零（以便无缝使用我们刚证明的三阶矩有上界时候的结论），我们把它平移一个期望 μ_n。使用 DCT，我们可以得到 $\mu_n = \mathbf{E}\left[X_1 \cdot \mathbb{I}[|X_1| \leq \varepsilon\sqrt{n}]\right] \overset{n\to\infty}{\to} 0$。和以往使用截断法的原因类似，我们这样操作的目标是保证 $X_n^{\leq}$ 的三阶矩是有界的，而 $X_n^{>0}$ 以高概率为零。

我们首先注意到对于任意 $i \geq 1$，$\mathbf{E}\left[\left|X_i^{\leq}\right|^3\right] \leq \mathbf{E}\left[\varepsilon\sqrt{n}\left|X_i\right|^2\right] = \varepsilon\sqrt{n}$。于是我们重复上面的计算，可以得到

$$\mathbf{E}\left[h\left(\frac{\sum_{i=1}^n X_i^{\leq}}{\sqrt{n}}\right)\right] - \mathbf{E}\left[h\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}\right)\right] = \mathcal{O}\left(\frac{1}{n^{1/2}}\mathbf{E}\left[\left|X_1^{\leq}\right|^3\right]\right) = \mathcal{O}(\varepsilon).$$

剩余的项是

$$\mathbf{E}\left[\left|\frac{\sum_{i=1}^n X_i^{>}}{\sqrt{n}}\right|\right] \overset{\text{(Cauchy-Schwarz)}}{\leq} \sqrt{\mathbf{E}\left[\left(\frac{\sum_{i=1}^n X_i^{>}}{\sqrt{n}}\right)^2\right]} \to \sqrt{\mathbf{Var}\left[X_1^{>}\right]}.$$

根据 DCT，我们又知道

$$\mathbf{Var}\left[X_1^{>}\right] = \mathbf{E}\left[X_1^2\mathbb{I}[|X_1| > \varepsilon\sqrt{n}]\right] - \mathbf{E}\left[X_1^{>}\right]^2 \overset{n\to\infty}{\to} 0.$$

第 25 章 随机变量的特征函数

25.1 特征函数的定义以及基本性质

我们之前介绍过随机变量 X 的矩生成函数 $M_X(t) = \mathbf{E}\left[e^{tX}\right]$。它是研究随机变量的矩的有力工具。我们证明过的一个重要的性质是如果 $M_X(t)$ 在 $t = 0$ 附近的一个邻域内存在的话，那么 X 的任意一阶矩都存在，并且对于 $k \in \mathbb{N}$，我们有

$$\mathbf{E}\left[X^k\right] = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(0).$$

这个结论同时告诉我们，如果存在某个 $m \in \mathbb{N}$，使得 X^m 是不可积的，那么 $M_X(t)$ 就不存在。但往往我们还是会对 $k < m$ 时候 X 的 k 阶矩感兴趣，矩生成函数就无能为力了。这个时候就要小修改一下矩生成函数的定义。我们定义 X 的特征函数

$$\varphi_X(t) := \mathbf{E}\left[e^{itX}\right],$$

这儿 $i = \sqrt{-1}$ 是虚数单位。注意到特征函数和矩生成函数唯一的不同是把 t 换成了 it。在 X 有概率密度 $f(x)$ 的时候，它们分别对应了对 $f(x)$ 的傅里叶变换和拉普拉斯变换。我们马上可以看到，$\varphi_X(t)$ 总是存在的。

我们先来研究一下特征函数的一些基本性质。

1. 根据欧拉公式

   $$\varphi_X(t) = \mathbf{E}\left[\cos tX\right] + i\mathbf{E}\left[\sin tX\right].$$

   由于 sin 和 cos 都是有界函数，因此可以看出 $\varphi_X(t)$ 对于任意 t 都是存在的。这也是和 $M_X(t)$ 的一个本质区别。

2. 如果 X 存在密度函数 $f_X(x)$，那么

   $$\varphi_X(t) = \int_{-\infty}^{\infty} f_X(x) \cdot e^{itx}\,\mathrm{d}x$$

   是 $f_X(x)$ 的傅立叶变换。根据傅里叶逆变换定理，

   $$f_X(x) = \lim_{T\to\infty} \int_{-T}^{T} \varphi_X(t) \cdot e^{-itx}\,\mathrm{d}t.$$

   注意到对于这个不定积分我们使用了柯西主值，因为 $\varphi_X(t)$ 不一定可积。这个结论告诉我们，密度函数唯一确定了特征函数，而特征函数也唯一确定了密度函数（也就是 X 的分布）。

3. 上述结论可以被推广到一般的随机变量（当 X 不存在密度函数的情形），被称作 Lévy 逆定理。设 $\bar{F}(x) =$

$\frac{1}{2}(F(x)+F(x-))$，其中 $F(x-):=\lim_{z\uparrow x}F(z)$。那么

$$\forall a<b,\quad \bar{F}(b)-\bar{F}(a)=\lim_{T\to\infty}\int_{-T}^{T}\frac{i}{2\pi t}\left(e^{-ibt}-e^{-iat}\right)\cdot\varphi_X(t)\,\mathrm{d}t.$$

我们不会证明这个结论（证明可以查看上面链接），但强调一下 Lévy 逆定理说明了随机变量的分布和其特征函数是相互唯一对应的（当然了，如果两个分布只在一个零测集上不一样，我们也认为它们相同）。

4. 根据定义容易验证，如果两个随机变量 X 和 Y 独立，那么 $\varphi_{X+Y}(t)=\varphi_X(t)\cdot\varphi_Y(t)$。类似的结论可以推广到 n 个相互独立的随机变量 $X_1,\dots,X_n$：$\varphi_{\sum_{i=1}^n X_i}(t)=\prod_{i=1}^n\varphi_{X_i}(t)$。

25.1.1 联合分布的特征函数

对于定义在同一个概率空间上的两个随机变量 X 和 Y，我们可以定义它们的联合特征函数

$$\varphi_{(X,Y)}(s,t)=\mathbf{E}\left[e^{i(sX+tY)}\right].$$

这个定义可以推广到任意 n 个随机变量 $\mathbf{X}=(X_1,\dots,X_n)$。它们的联合特征函数是

$$\varphi_{\mathbf{X}}:\mathbf{t}\in\mathbb{R}^n\mapsto\mathbf{E}\left[e^{i\mathbf{t}^\top\mathbf{X}}\right].$$

Lévy 逆定理可以被推广到 $\mathbf{X}$ 这样的随机向量的场合。类似的，联合特征函数也唯一（up to 零测集）确定了联合分布。

25.2 特殊分布的特征函数计算

我们来给那几位老伙计算算特征函数。

例题 25.1. $X\sim\texttt{Ber}(p)$

显然有 $\varphi_X(t)=\mathbf{E}\left[e^{itX}\right]=1-p+pe^{it}$.

例题 25.2. $X\sim\texttt{Binom}(n,p)$

由于 X 可以写成 n 个分布为 $\texttt{Ber}(p)$ 的独立随机变量之和，根据我们前面提到的性质，$\varphi_X(t)=\left(1-p+pe^{it}\right)^n$。

例题 25.3. $X\sim\texttt{Exp}(\lambda)$

由于指数分布的概率密度是 $\forall x\geq 0,\ f_X(x)=\lambda e^{-\lambda x}$，所以

$$\varphi_X(t)=\mathbf{E}\left[e^{itX}\right]=\int_0^\infty\lambda e^{-\lambda x}e^{itx}\,dx=\frac{\lambda}{\lambda-it}.$$

例题 25.4. $X\sim\texttt{Pois}(\lambda)$

泊松分布的概率质量函数是 $\forall n\in\mathbb{N},\ p_X(n)=\frac{\lambda^n}{n!}e^{-\lambda}$，所以

$$\varphi_X(t)=\mathbf{E}\left[e^{itX}\right]=\sum_{n=0}^\infty\frac{\lambda^n}{n!}e^{-\lambda}e^{itn}=e^{-\lambda}\sum_{n=0}^\infty\frac{\left(\lambda e^{it}\right)^n}{n!}=e^{-\lambda}e^{\lambda e^{it}}.$$

例题 25.5. $X \sim \mathcal{N}(0,1)$

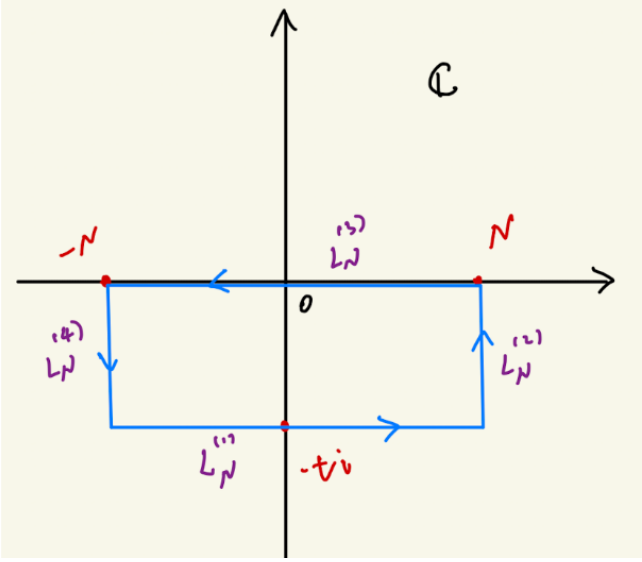
标准高斯分布的概率密度是 $\phi_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$，所以

$$\varphi_X(t) = \mathbf{E}\left[e^{itX}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - \frac{x^2}{2}}\, dx.$$

我们把积分里面的指数部分进行配方，然后做 $z = x - it$ 的换元，可以得到

$$\varphi_X(t) = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{\Im z = -t} e^{-\frac{1}{2}z^2}\, \mathrm{d}z.$$

这是一个在复平面上的积分，我们的积分范围是 $\Im z = -t$ 的直线。对于给定的 $N > 0$，我们计算如下图所示的围道积分。



我们把蓝色曲线记作 L_N，那么根据柯西定理

$$\int_{L_N} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = \int_{L_N^{(1)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z + \int_{L_N^{(2)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z + \int_{L_N^{(3)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z + \int_{L_N^{(4)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = 0.$$

注意到 $\lim_{N\to\infty} \int_{L_N^{(2)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = \lim_{N\to\infty} \int_{L_N^{(4)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = 0$，所以

$$\lim_{N\to\infty} \int_{L_N^{(1)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = -\lim_{N\to\infty} \int_{L_N^{(3)}} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = \sqrt{2\pi}.$$

也就是说 $\int_{\Im z = -t} e^{-\frac{1}{2}z^2}\, \mathrm{d}z = \sqrt{2\pi}$. 代回 $\varphi_X(t)$ 的表达式我们便能得到

$$\varphi_X(t) = e^{-\frac{t^2}{2}}.$$

Wow，居然和 $\phi_X(x)$ 的形式一致。高斯的傅里叶变换还是高斯。

例题 25.6. $X \sim \mathcal{N}(\mu, \sigma^2)$

如果 $\xi \sim \mathcal{N}(0,1)$，那么 $X := \sigma\xi + \mu \sim \mathcal{N}(\mu, \sigma^2)$。根据定义，我们显然有

$$\varphi_X(t) = \mathbf{E}\left[e^{it(\sigma\xi+\mu)}\right] = e^{it\mu}\mathbf{E}\left[e^{it\sigma\xi}\right] = e^{it\mu}\varphi_\xi(\sigma t) = e^{-\frac{\sigma^2 t^2}{2} + it\mu}.$$

例题 25.7. $X \sim \mathcal{N}(\mu, \Sigma)$

注意到这儿 $X = (X_1, \dots, X_n)$ 是一个 n 维随机向量。我们计算它的联合特征函数。对于一个 $\mathbf{t} = (t_1, \dots, t_n)^\top$，我们有

$$\varphi_X(\mathbf{t}) = \mathbf{E}\left[e^{i\mathbf{t}^\top X}\right].$$

我们注意到 $Y := \mathbf{t}^\top X$ 是一个高斯随机变量，我们只要计算出 Y 的期望和方差，就能得到 Y 的特征函数 $\varphi_Y(t)$。而 $\varphi_X(\mathbf{t}) = \varphi_Y(1)$。

我们显然有 $\mathbf{E}[Y] = \mathbf{t}^\top \mathbf{E}[X] = \mathbf{t}^\top \mu$。我们知道 $X = A\xi + \mu$，其中 $\xi \sim \mathcal{N}(0, \mathtt{Id}_n)$，矩阵 A 满足 $AA^\top = \Sigma$。于是

$$\mathbf{Var}[Y] = \mathbf{E}\left[\left(\mathbf{t}^\top A\xi\right)^2\right] = \mathbf{t}^\top A\mathbf{E}\left[\xi\xi^\top\right]A^\top\mathbf{t} = \mathbf{t}^\top\Sigma\mathbf{t}.$$

这便得到

$$\varphi_X(\mathbf{t}) = \varphi_Y(1) = e^{-\frac{1}{2}\mathbf{t}^\top\Sigma\mathbf{t} + i\mathbf{t}^\top\mu}.$$

25.2.1 多元高斯分布的刻画

我们之前说一个 n-维随机变量 $X = (X_1, \dots, X_n)^\top$ 是多元高斯（或称高维高斯，联合高斯）的，当且仅当存在 $n \times n$ 维矩阵 A 和 n 维向量 μ 使得 $X = A\xi + \mu$，其中 ξ 是一个每一维是独立 $\mathcal{N}(0,1)$ 随机变量的 n 维向量。并且我们知道 $X \sim \mathcal{N}(\mu, \Sigma)$，其中 $\Sigma = AA^\top$。我们现在给它一个新的刻画：

定理 25.1

$X = (X_1, \dots, X_n)^\top$ 是一个高维高斯向量当且仅当 $X_1, \dots, X_n$ 的任意线性组合是一个一维高斯随机变量。♡

定理的“仅当”方向是显然的，即如果 X 是高维高斯，那么 $X_1, \dots, X_n$ 的任意线性组合也是高斯。这是由于每一个 X_i 都可以写成 $\xi_1, \dots, \xi_n$ 的线性组合，于是任意 $X_1, \dots, X_n$ 的线性组合也可以写成 $\xi_1, \dots, \xi_n$ 的线性组合。而我们知道，独立高斯的线性组合依旧是高斯的。

我们现在来证明“当”。设 $\mu = \mathbf{E}[X]$，$\Sigma = \mathbf{Cov}[X]$。根据联合分布的 Lévy 逆定理，我们只要说明 X 的（联合）特征函数是高维高斯的就行。也就是对于任何 $\mathbf{t} \in \mathbb{R}^n$，计算 $\varphi_X(\mathbf{t}) = \mathbf{E}\left[e^{i\mathbf{t}^\top X}\right]$。根据条件，我们知道 $Y := \mathbf{t}^\top X$ 作为 $X_1, \dots, X_n$ 的一个线性组合是满足高斯分布的。因此我们有 $\varphi_X(\mathbf{t}) = \varphi_Y(1)$。同样我们只需要计算 Y 的期望和方差就能得到 φ_Y。

$$\mathbf{E}[Y] = \mathbf{t}^\top\mathbf{E}[X] = \mathbf{t}^\top\mu,$$
$$\mathbf{Var}[Y] = \mathbf{E}\left[\left(\mathbf{t}^\top X - \mathbf{t}^\top\mu\right)^2\right] = \mathbf{t}^\top\mathbf{E}\left[(X-\mu)(X-\mu)^\top\right]\mathbf{t} = \mathbf{t}^\top\Sigma\mathbf{t}.$$

这便说明 $\varphi_X(\mathbf{t}) = \varphi_Y(1) = e^{-\frac{1}{2}\mathbf{t}^\top\Sigma\mathbf{t} + i\mathbf{t}^\top\mu}$。也就是说 $X \sim \mathcal{N}(\mu, \Sigma)$。

25.3 特征函数与随机变量的矩

我们在一开始便提到过，我们对矩生成函数求导可以得到随机变量的矩。但这一操作的可行性需要随机变量的任意一阶矩均存在。如果 X 对于 m 阶矩不存在，但对于 $k < m$ 阶矩存在（回忆我们以前用 Jensen 不等式证明过对于 $p > 1$，X^p 可积可以蕴含 X^{p-1} 可积），使用矩生成函数便不行了。下面这个结论，除了告诉我们可以使用特征函数来计算 up to $m-1$ 阶矩之外，对于研究特征函数本身的性质有着重要的意义。

定理 25.2

如果随机变量 X 满足 $\mathbf{E}[|X|^n] < \infty$，那么

$$\varphi_X(t) = \sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbf{E}[X^k] + o(|t|^n).$$

特别的，对于 $k=1,2,\dots,n$，$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\varphi_X(0)=i^k\mathbf{E}\left[X^k\right]$.

函数 e^{itX} 的泰勒级数的前 n 项是 $\sum_{k=0}^{n}\frac{1}{k!}X^k(it)^k$。因此，我们为了证明这个定理，需要讨论的事情主要是控制级数的余项，并据此说明可以交换求和和期望。

我们使用带有积分余项的泰勒级数：

$$f(z)=\sum_{k=0}^{n}\frac{f^{(k)}(0)}{k!}\cdot z^k+\frac{1}{n!}\int_0^z(z-t)^nf^{(n+1)}(t)\,\mathrm{d}t.$$

我们用 R_n 表示把 e^z 展开到 n 次之后的余项，于是

$$e^{ix}=\sum_{k=0}^{n}\frac{(ix)^k}{k!}+R_{n+1},\quad R_{n+1}=\frac{i^{n+1}}{n!}\int_0^x(x-t)^ne^{it}\,\mathrm{d}t.$$

从这个表达式看起来，$|R_{n+1}|\le\frac{|x|^{n+1}}{(n+1)!}$。这在 $x\to0$ 的时候是一个很好的上界，但是我们的条件是 $\mathbf{E}\left[|X|^n\right]<\infty$，在 $x\to\infty$ 的时候 $|x|^{n+1}$ 太大了。我们需要找一个 $x\to\infty$ 时更好的上界。注意到

$$R_{n+1}=R_n-\frac{(ix)^n}{n!}=\frac{i^n}{n!}\left(n\int_0^x(x-t)^{n-1}e^{it}\,\mathrm{d}t-x^n\right).$$

于是 $|R_{n+1}|\le\frac{2|x|^n}{n!}$。我们便得到了

$$|R_{n+1}|\le\frac{|x|^{n+1}}{(n+1)!}\wedge\frac{2\,|x|^n}{n!}.$$

根据条件 $\mathbf{E}\left[|X|^n\right]<\infty$，我们知道

$$e^{itX}=\sum_{k=0}^{n}\frac{(itX)^k}{k!}+R'_{n+1}(X),\quad\left|R'_{n+1}(X)\right|\le\frac{2\,|t|^n\,|X|^n}{n!}$$

满足求和的每一项都是可积的。因此，根据期望的线性性，我们有

$$\mathbf{E}\left[e^{itX}\right]=\sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbf{E}\left[X^k\right]+\mathbf{E}\left[R'_{n+1}(X)\right].$$

我们接着说明 $\mathbf{E}\left[\left|R'_{n+1}(X)\right|\right]=o(|t|^n)$ as $t\to0$。这等价于 $\lim_{t\to0}t^{-n}\mathbf{E}\left[\left|R'_{n+1}(X)\right|\right]=0$。根据我们上面的对于余项的上界可以知道 $t^{-n}\left|R'_{n+1}(X)\right|\le\frac{2|X|^n}{n!}$. 因此使用 DCT，

$$\lim_{t\to0}t^{-n}\mathbf{E}\left[\left|R'_{n+1}(X)\right|\right]=\mathbf{E}\left[\lim_{t\to0}t^{-n}\left|R'_{n+1}(X)\right|\right]\le\mathbf{E}\left[\lim_{t\to0}\frac{t\,|X|^{n+1}}{(n+1)!}\right]=0.$$

注意到我们在上面的分析中，同时用到了 R_{n+1} 的两个上界，分别对应于 x 很大和 x 很小的时候。

25.4 Lévy 连续性定理及应用

特征函数的另一个重要性质说的是它与依分布收敛的关系。我们不加证明的给出结论

定理 25.3 (Lévy 连续性定理)

给定一族（不一定定义在同一概率空间上的）随机变量 $X_1,X_2,\dots$。对于每一个 $n\ge1$，我们用 φ_n 表示 X_n 的特征函数。如果 $\varphi_n(t)$ 逐点收敛到一个函数 $\varphi(t)$。那么下面两件事情等价。

- $\varphi(t)$ 在 $t=0$ 连续；
- 存在一个随机变量 X，它的特征函数是 φ，并且 $X_n\xrightarrow{D}X$。

另一方面，我们可以很容易验证，如果 $X_n\xrightarrow{D}X$，那么 $\varphi_n(t)$ 逐点收敛到 $\varphi(t)$（类似我们之前用测试函数说明依分布收敛的证明，这儿对应的测试函数是 $h_t(x)=e^{itx}$）。所以，我们知道，在 $\varphi(t)$ 在 $t=0$ 连续的条件下，依分布收敛和特征函数的逐点收敛是等价的。

我们简单说明一下，$\varphi(t)$ 在 $t=0$ 这一点连续的条件是必要的，否则 $\varphi(t)$ 有可能并不是任何函数的特征函

数。比如说，我们让 $X_n \sim \mathcal{N}(0,n)$，那么 $\varphi_n(t) = e^{-\frac{nt^2}{2}}$。它的极限是 $\varphi(t) = \mathbb{I}[t=0]$，在 $t=0$ 不连续，也不是任何随机变量的特征函数。

Lévy 连续定理是我们研究依分布收敛的重要工具。接下来我们看几个应用。

25.4.1 泊松分布作为二项式分布的极限

我们之前介绍泊松分布 $\texttt{Pois}(\lambda)$ 的时候是把它看成二项式分布 $\texttt{Binom}(n,p)$ 在 $np=\lambda$ 时候的极限。这个事实可以用特征函数一句话说明。注意到 $X \sim \texttt{Binom}(n,p)$ 的特征函数是

$$\left(1-p+pe^{it}\right)^n = \left(1-\frac{np-npe^{it}}{n}\right)^n = \left(1-\frac{\lambda-\lambda e^{it}}{n}\right)^n \overset{n\to\infty}{\to} e^{-\lambda}e^{\lambda e^{it}}.$$

而这正是 $\texttt{Pois}(\lambda)$ 的特征函数。

25.4.2 中心极限定理的特征函数证明

我们现在用特征函数来证明中心极限定理。

定理 25.4 (中心极限定理)

如果独立同分布的随机变量 $X_1, X_2, \ldots$ 满足 $\mathbf{E}[X_1] = \mu, \mathbf{Var}[X_1] = \sigma^2$ 均为有限的，那么

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} Y \sim \mathcal{N}(0,1).$$

♡

我们不失一般性的假设 $\mu=0, \sigma^2=1$。我们已经证明了

$$\varphi_{X_1}(t) = \mathbf{E}\left[e^{itX_1}\right] = \mathbf{E}\left[1+itX_1-\frac{1}{2}t^2X_1^2+\varepsilon(t^2)\right] = 1-\frac{1}{2}t^2+\varepsilon(t^2).$$

这儿 $\varepsilon(t)$ 是一个满足 $\lim_{t\to 0}\frac{\varepsilon(t)}{t}=0$ 的函数。于是根据独立性

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right)^n = \left(1-\frac{t^2}{2n}+\varepsilon\left(\frac{t^2}{n}\right)\right)^n \overset{n\to\infty}{\to} e^{-\frac{1}{2}t^2}.$$

由 Lévy 连续性定理 $\frac{S_n}{\sqrt{n}} \xrightarrow{D} \xi \sim \mathcal{N}(0,1)$。

可以看到，特征函数是处理独立随机变量和的有力工具。我们将在作业里尝试使用它去掉 CLT 里对于同分布的要求。

第 26 章 条件期望

我们今天继续来引入概率论里面的一个核心概念：条件期望。这是我们在未来学习随机过程的时候必不可少的语言。在今天的讨论里，我们还是固定一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$。

26.1 条件期望的定义

给定事件 $A, B \in \mathscr{F}$，在 $\mathbb{P}(B) > 0$ 的时候，我们定义过条件概率 $\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$。我们也定义过给定事件 B 之后随机变量 X 的条件概率：$\mathbf{E}[X \mid B] := \frac{\mathbf{E}[X \cdot \mathbb{I}_B]}{\mathbb{P}(B)}$。对于两个随机变量 X 和 Y，我们今天的目标是定义记号 $\mathbf{E}[Y \mid X]$。在我们今天所有的讨论中，均假设 X 和 Y 是可积的。

26.1.1 X 是离散随机变量的场合

我们首先假设 X 是离散的随机变量，即 X 的取值 $\mathtt{Im}(X) = \{x_1, x_2, \dots\}$。对于每一个 x_i，我们知道 $[X = x_i]$ 是一个概率非零的事件，因此按照我们上面的定义

$$\mathbf{E}[Y \mid X = x_i] = \frac{\mathbf{E}[Y \cdot \mathbb{I}_{X=x_i}]}{\mathbb{P}(X = x_i)}.$$

显然这是一个关于 x_i 的函数，换句话说，我们可以找到一个 Borel 函数 $f : \mathbb{R} \to \mathbb{R}$ 满足

$$f : x_i \mapsto \mathbf{E}[Y \mid X = x_i].$$

于是，我们定义 $\mathbf{E}[Y \mid X] := f(X)$。换句话说，$\mathbf{E}[Y \mid X]$ 是一个随机变量，满足

$$\mathbf{E}[Y \mid X] : \omega \in \Omega \mapsto f(X(\omega)).$$

我们应该这样看待这个定义：X 定义了样本空间的一个分划 $\Omega = \sqcup_{n \geq 1} \Lambda_n$，其中 $\Lambda_n = X^{-1}(x_n)$。对于每一个 $\omega \in \Omega$，如果其属于 Λ_k，则 $\mathbf{E}[Y \mid X](\omega)$ 的值为 Y 在 Λ_k 上的条件期望，即 $\mathbf{E}[Y \mid \Lambda_k]$。这是理解条件期望以及它的相关性质的最重要的直观。

回忆一下我们以前在讲 σ-代数时候的一个 running example，它对于理解今天的概念也非常的重要：考虑投掷一个公平的六面骰子的概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$，其中 $\Omega = [6]$，$\mathscr{F} = 2^{\Omega}, \forall i \in \Omega, \mathbb{P}(\{i\}) = \frac{1}{6}$。我们定义四个随机变量

1. $X_1 : i \in \Omega \mapsto i$，即 X_1 表示掷出来的点数；
2. $X_2 : i \in \Omega \mapsto \mathbb{I}_{[i \geq 4]}$，即 X_2 表示掷出来的点数是“大”还是“小”；
3. $X_3 : i \in \Omega \mapsto i \bmod 2$，即 X_3 表示掷出来的点数除 2 之后的余数；

4. $X_4 : i \in \Omega \mapsto i \bmod 4$，即 X_4 表示掷出来的点数除 4 之后的余数。

那么我们有

1. $\mathbf{E}\left[X_1 \mid X_1\right](i) = X_1(i)$.
2. $\mathbf{E}\left[X_1 \mid X_2\right](i) = \begin{cases} 5 & \text{if } i \geq 4; \\ 2 & \text{if } i < 4. \end{cases}$
3. $\mathbf{E}\left[X_3 \mid X_4\right](i) = X_3(i)$.
4. $\mathbf{E}\left[X_4 \mid X_2\right](i) = \begin{cases} \frac{2+0+2}{3} & \text{if } i \text{ is even}; \\ \frac{1+3+1}{3} & \text{if } i \text{ is odd}. \end{cases}$

此外，我们还可以注意到一个事实，就是 $\mathbf{E}[Y \mid X]$ 的定义实际上只与 X 所定义出来的分划 $\Lambda_1, \Lambda_2, \ldots$ 有关，而与 $x_1, x_2, \ldots$ 的具体取值无关。换句话说，$\mathbf{E}[Y \mid X]$ 实际上只与 X 生成的 σ-代数 $\sigma(X)$ 有关。

26.1.2 X 与 Y 有联合密度函数的场合

设 $f_{XY}(x, y)$ 是 X 与 Y 的联合密度函数。在边缘密度函数 $f_X(x) \neq 0$ 的时候，我们之前定义过条件期望

$$\mathbf{E}[Y \mid X = x] = \frac{\int_{\mathbb{R}} y f_{XY}(x, y) \mathrm{d}y}{f_X(x)}.$$

可以看出，这也是一个关于 x 的函数。我们可以找到一个 Borel 函数 $f : \mathbb{R} \to \mathbb{R}$ 满足

$$f : x \mapsto \mathbf{E}[Y \mid X = x].$$

于是我们可以类似离散场合定义 $\mathbf{E}[Y \mid X] := f(X)$。

26.1.3 一般随机变量的场合

对于一般的随机变量，合理的定义出 $\mathbf{E}[Y \mid X]$ 不是一件简单的事情。事实上，我们先抽象出前面两种特殊场合定义的条件期望满足的两个重要性质。

- $\mathbf{E}[Y \mid X]$ 是 $\sigma(X)$-可测的。
- 对于任何 $A \in \sigma(X)$，$\int_A Y \,\mathrm{d}\mathbb{P} = \int_A \mathbf{E}[Y \mid X] \,\mathrm{d}\mathbb{P}$。

因为我们知道存在一个 Borel f 满足 $\mathbf{E}[Y \mid X] = f(X)$。所以 $\mathbf{E}[Y \mid X]$ 是 $\sigma(X)$-可测是显然的。我们现在分别对于离散和具有联合分布函数的两种场合验证第二点。

1. 离散随机变量：我们知道 $\Omega = \sqcup_{n \geq 1} \Lambda_n$。我们知道对于每一个 $A \in \sigma(X)$，都可以写成若干 Λ_k 的并，因此根据积分的可加性，我们只需要对 $A = \Lambda_k$ 证明即可。这个时候，我们知道根据定义，对于每一个 $\omega \in \Lambda_k$，$\mathbf{E}[Y \mid X](\omega)$ 的取值均为 $\mathbf{E}[Y \mid X = x_k]$，所以
$$\int_{\Lambda_k} \mathbf{E}[Y \mid X] \,\mathrm{d}\mathbb{P} = \mathbf{E}[Y \mid X = x_k] \cdot \mathbb{P}(\Lambda_k) = \mathbf{E}\left[Y \cdot \mathbb{I}_{[X = x_k]}\right] = \int_{\Lambda_k} Y \,\mathrm{d}\mathbb{P}.$$
2. 具有联合分布的随机变量：对于 $A \in \sigma(X)$，我们知道，存在 $B \in \mathcal{B}$ 使得 $A = X^{-1}(B)$。我们首先有
$$\int_A Y \,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} \int_{\mathbb{R}} y \cdot f_{XY}(x, y) \cdot \mathbb{I}_{[x \in B]} \,\mathrm{d}x \otimes \mathrm{d}y.$$

另一方面

$$\begin{aligned} \int_A \mathbf{E}[Y \mid X] \,\mathrm{d}\mathbb{P} &= \int_B f_X(x) \cdot \mathbf{E}[Y \mid X = x] \,\mathrm{d}x \\ &= \int_B f_X(x) \cdot \left(\int_{\mathbb{R}} y \cdot f_{Y|X}(y|x) \,\mathrm{d}y \right) \mathrm{d}x \\ &\overset{(Fubini)}{=} \int_{\mathbb{R}} y \cdot \left(\int_B f_{XY}(x, y) \,\mathrm{d}x \right) \mathrm{d}y \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} y \cdot f_{XY}(x, y) \cdot \mathbb{I}_{[x \in B]} \,\mathrm{d}x \,\mathrm{d}y. \end{aligned}$$

我们把上面两个性质当做条件期望的定义。我们更一般的给出一个随机变量 Y 在一个 σ-代数的条件下的条

件期望定义。

定义 26.1 (条件期望)

设 $(\Omega, \mathscr{F}, \mathbb{P})$ 是个概率空间，X 是一个定义在其上的可积的随机变量，$\mathscr{G} \subseteq \mathscr{F}$ 是一个子 σ-代数。我们说一个随机变量 $Z\colon \Omega \to \mathbb{R}$ 是给定 $\mathscr{G}$ 后 X 的条件期望，并记作 $Z = \mathbf{E}[X \mid \mathscr{G}]$，当且仅当其满足

1. Z 是 $\mathscr{G}$-可测的；
2. 对于每一个 $A \in \mathscr{G}$，$\int_A Z \, d\mathbb{P} = \int_A X \, d\mathbb{P}$。

在这个定义的基础上，对于随机变量 X，我们定义 $\mathbf{E}[Y \mid X] := \mathbf{E}[Y \mid \sigma(X)]$。

我们有的时候也使用“条件概率” $\mathbb{P}(A \mid \mathscr{G})$ 的记号，它被定义为 $\mathbf{E}[\mathbb{I}_A \mid \mathscr{G}]$。

我们对于条件期望的定义比较抽象，它和我们之前遇到过的大多数数学对象都不一样，是通过“描述性质”的方法来定义的。所以我们必须说明其合理性。首先是“唯一性”

命题 26.1

如果 Z 和 Z' 都是满足上面两个条件的随机变量，那么 $Z = Z'$ a.e.

根据定义的第二条，我们知道 Z 和 Z' 都是可积的。设 $A = \{\omega \in \Omega : Z(\omega) > Z'(\omega)\}$。于是

$$\int_A Z - Z' \, \mathrm{d}\mathbb{P} = \int_A Z \, \mathrm{d}\mathbb{P} - \int_A Z' \, \mathrm{d}\mathbb{P}' = \int_A X \, \mathrm{d}\mathbb{P} - \int_A X \, \mathrm{d}\mathbb{P} = 0.$$

即 $\mathbb{P}(Z > Z') = 0$。同理 $\mathbb{P}(Z < Z') = 0$。因此 $\mathbb{P}(Z = Z') = 1$。

最后，我们要说明这样一个 Z 总是存在的。它是测度论里面的 Radon-Nikodym 定理的推论，它的证明超出了这门课的范畴，我们简单讨论一下。

26.1.4 条件期望的存在性与 Radon-Nikodym 定理

我们之前说过一个分布是“绝对连续 (absolutely continuous)”的，当且仅当其分布函数 $F(x)$ 存在概率密度 f 满足

$$\forall t,\ F(t) = \int_{-\infty}^{t} f(x) \, \mathrm{d}x.$$

事实上，我们更应该把上述定义看成绝对连续的性质。我们引入如下更加一般的“绝对连续”的定义。

定义 26.2 (绝对连续)

设 μ 和 ν 为可测空间 $(\Omega, \mathscr{F})$ 上的两个测度。我们说 ν 相对于 μ 是绝对连续的，当且仅当对于任何 $A \in \mathscr{F}$，$\mu(A) = 0 \implies \nu(A) = 0$，记作 $\nu \ll \mu$。

Radon-Nikodym 定理则说，如果 $\nu \ll \mu$ 并且 μ 和 ν 均是 σ-有限的，那么 ν 具有“相对于” μ 的密度（又被称为 Radon-Nikodym 导数）f，满足：

- $f\colon \Omega \to \mathbb{R}_{\geq 0}$ 是一个 $\mathscr{F}$-可测的函数；
- 对于任何 $A \in \mathscr{F}$，$\nu(A) = \int_A f \, \mathrm{d}\mu$。

并且，在 up to μ 的零测集的意义下 f 是唯一的。我们一般把 f 记作 $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$。可以看到，我们之前以前定义的概率密度函数就是 μ 为勒贝格测度的特例（这个时候我们把 $\mu(\mathrm{d}x)$ 习惯性记作 $\mathrm{d}x$）。

回到我们的条件期望。在我们的概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上，对于一个给定的随机变量 X 以及一个子 σ-代数 $\mathscr{G}$，我们可以定义两个 $\mathscr{G}$ 上的测度：对于任何 $A \in \mathscr{G}$，

- $\mu(A) := \mathbb{P}(A)$，
- $\nu(A) := \int_A X \, \mathrm{d}\mathbb{P}$。

我们定义 $\mathbf{E}[X \mid \mathscr{G}] := \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$。

26.2 条件期望的性质

我们现在讨论条件期望的性质。可以很容易验证，我们定义的期望，本身也是条件期望的一个特殊情况，即 $\mathscr{G}=\{\varnothing,\Omega\}$ 是最简单 σ-代数。

1. $\mathbf{E}[X]=\mathbf{E}[X\mid\{\varnothing,\Omega\}]$。

命题 26.2

如果 X 是 $\mathscr{G}$-可测的，那么 $\mathbf{E}[X\mid\mathscr{G}]=X$ a.e.

证明 根据条件期望的定义，这是显然的（X 是 $\mathscr{G}$-可测且对任意 $A\in\mathscr{G}$，$\int_A X\,\mathrm{d}\mathbb{P}=\int_A X\,\mathrm{d}\mathbb{P}$）。

条件期望的一个核心的性质是所谓的"tower rule"。假设 $\mathscr{G}_1,\mathscr{G}_2\subseteq\mathscr{F}$，并且满足 $\mathscr{G}_1\subseteq\mathscr{G}_2$。换句话说，$\mathscr{G}_1$ 是比 $\mathscr{G}_2$ 更粗的 σ-代数。那么

3. $\mathbf{E}[\mathbf{E}[X\mid\mathscr{G}_1]\mid\mathscr{G}_2]=\mathbf{E}[\mathbf{E}[X\mid\mathscr{G}_2]\mid\mathscr{G}_1]=\mathbf{E}[X\mid\mathscr{G}_1]$。

也就是说，当条件期望复合出现的时候，最终剩下的总是更“粗”的 σ-代数。

证明 $\mathbf{E}[X\mid\mathscr{G}_1]$ 是 $\mathscr{G}_1$-可测的，因此也是 $\mathscr{G}_2$-可测的。于是根据性质 (2)，$\mathbf{E}[\mathbf{E}[X\mid\mathscr{G}_1]\mathscr{G}_2]=\mathbf{E}[X\mid\mathscr{G}_1]$。另一方面，对于任意一个 $A\in\mathscr{G}_1$，我们知道其也 $\in\mathscr{G}_2$。

$$\int_A\mathbf{E}[X\mid\mathscr{G}_2]\,\mathrm{d}\mathbb{P}=\int_A X\,\mathrm{d}\mathbb{P}=\int_A\mathbf{E}[X\mid\mathscr{G}_1]\,\mathrm{d}\mathbb{P}.$$

又 $\mathbf{E}[X\mid\mathscr{G}_1]$ 是 $\mathscr{G}_1$-可测的，所以

$$\mathbf{E}[\mathbf{E}[X\mid\mathscr{G}_2]\mid\mathscr{G}_1]=\mathbf{E}[X\mid\mathscr{G}_1].$$

4. $\mathbf{E}[\mathbf{E}[X\mid\mathscr{G}]]=\mathbf{E}[X]$

这个性质是 (1) 和 (3) 的简单推论，但是在很多概率的计算中非常有用。我们通常使用的方式是 $\mathbf{E}[X]=\mathbf{E}[\mathbf{E}[X\mid Y]]$，它可以解读成“为了计算 X 的平均值，我们先按照 Y 分类，对 Y 的每种情况计算对应 X 的平均值，再对 Y 的取值求平均”。比如说，我们让 $(\Omega,\mathscr{F},\mathbb{P})$ 为班上所有同学的均匀分布。X 为同学的身高，Y 为同学的性别，那么 $\mathbf{E}[X\mid Y]$ 就表示随机抽一个同学，和该同学同性别的同学的平均身高。而 $\mathbf{E}[\mathbf{E}[X\mid Y]]=\mathbf{E}[X]$ 的直观含义就是，先统计男生平均身高和女生平均身高，然后再按照男生女生人数的比例对这两个数平均，就得到了全班同学的身高。（很遗憾，上课的时候 Y 是一个常数）

5. 如果 X 是 $\mathscr{G}$-可测的，并且 XY 是可积的，那么 $\mathbf{E}[XY\mid\mathscr{G}]=X\mathbf{E}[Y\mid\mathscr{G}]$ a.e.

这个性质和（2）一样告诉我们，在做计算的时候，如果 X 是 $\mathscr{G}$-可测的，说明它在“已知 $\mathscr{G}$”的信息下，它没有什么随机性，因此可以当成一个常数一样从期望里拿出来。

对于离散的随机变量 X，可以使用定义简单的证明。对于一般的 X，我们可以通过对 $\underline{X}_k$ 的情况取极限得到。这个证明留做练习。

大量关于期望的性质都可以推广到条件期望，我们罗列如下。他们均可以通过定义简单证明。

6. $\mathbf{E}[aX+bY\mid\mathscr{G}]=a\mathbf{E}[X\mid\mathscr{G}]+b\mathbf{E}[Y\mid\mathscr{G}]$ a.e.
7. 如果 $X\geq 0$ a.e., 那么 $\mathbf{E}[X\mid\mathscr{G}]\geq 0$ a.e.
8. $|\mathbf{E}[X\mid\mathscr{G}]|\leq\mathbf{E}[|X|\mid\mathscr{G}]$ a.e.
9. 如果 X 和 $\mathscr{G}$ 独立，那么 $\mathbf{E}[X\mid\mathscr{G}]=\mathbf{E}[X]$。
10. 如果 X_n 和 X 均可积，并且 $X_n\uparrow X$，那么 $\mathbf{E}[X_n\mid\mathscr{G}]\to\mathbf{E}[X\mid\mathscr{G}]$ a.e.
11. 琴生不等式：如果函数 ϕ 在定义域内是 convex 的，并且 $\phi(X)$ 是可积的。那么 $\phi(\mathbf{E}[X\mid\mathscr{G}])\leq\mathbf{E}[\phi(X)\mid\mathscr{G}]$。

我们将在之后几次课大量使用这些性质进行计算。但在计算的时候，需要非常小心。试看下面一例：

例题 26.1.

假设我独立的投掷两个 6 面骰子，X 表示第一个的点数，Y 表示第二个的点数。那么 $\mathbf{E}[\mathbf{E}[X+Y\mid X]\mid Y]$ 是多少？

根据定义，我们知道 $\mathbf{E}[X+Y \mid X] = X + \mathbf{E}[Y \mid X] = X + 3.5$ 是一个 $\sigma(X)$ 可测的随机变量。于是 $\mathbf{E}[X+3.5 \mid Y] = 3.5 + \mathbf{E}[X] = 7$。

即使非常 senior 的学者也容易在这个问题上犯错。

第 27 章 离散鞅简介

最后两次课，我们简单介绍一下离散鞅论（martingale）。这是现代概率论的核心工具，有着丰富的内容和非凡的应用。因为课时原因，我们仅仅能够简单一瞥其芳容。

鞅的概念首先来自公平赌博游戏。假设你在玩一个猜大小的游戏。游戏的每一轮开始前，你可以“押大”或者“押小”。你可以押任意多的赌注，但无论如何，在每一轮中期望收益为零。因此，你的总财产的“期望”是保持不变。用数学的语言来描述这个性质，我们用 X_t 表示第 t 轮的收益，用 Z_t 表示第 t 轮结束之后的总财产。那么对于任何 $T>0$，$Z_T = Z_0 + \sum_{t=1}^{T} X_t$。这个游戏是一个公平游戏指的是

$$\forall t \geq 0,\ \mathbf{E}\left[X_{t+1} \mid X_1, \ldots, X_t\right] = 0,$$

或者等价的

$$\forall t \geq 0,\ \mathbf{E}\left[Z_{t+1} \mid X_1, \ldots, X_t\right] = Z_t.$$

我们把这样一个性质抽象出来，便是一般的鞅的定义。注意到我们并不要求 X_{t+1} 与 $X_1, \ldots, X_t$ 独立，也就是你的赌注可以与之前的胜负有关，as long as 我们玩的是一个公平游戏就行。

- 设 $\{X_n\}_{n\geq 0}$ 和 $\{Z_n\}_{n\geq 0}$ 是两个随机变量序列。如果每一个 Z_n 都是可积的，并且它们满足

$$\forall n \geq 0,\ \mathbf{E}\left[Z_{n+1} \mid X_0, X_1, \ldots, X_n\right] = Z_n,$$

  我们则称 $\{Z_n\}_{n\geq 0}$ 是相对于 $\{X_n\}_{n\geq 0}$ 的一个鞅。
- 有的时候，我们也会直接称一个序列 $\{Z_n\}_{n\geq 0}$ 是鞅。这个的意思是 $\{Z_n\}$ 相对于自己是鞅，即

$$\forall n \geq 0,\ \mathbf{E}\left[Z_{n+1} \mid Z_0, Z_1, \ldots, Z_n\right] = Z_n.$$

为了简化记号，接下来我们用 $\overline{X}_{i,j}$ 来表示随机变量 $(X_i, X_{i+1}, \ldots, X_j)$。

我们注意到，条件期望 $\mathbf{E}\left[Z_{n+1} \mid \overline{X}_{0,n}\right]$ 实际上的定义是 $\mathbf{E}\left[Z_{n+1} \mid \sigma(\overline{X}_{0,n})\right]$。这便让我们可以用 σ-代数的语言更一般的定义鞅。

为此我们先引入滤链的概念。考虑概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$。设 $\{\mathscr{F}_n\}_{n\geq 0}$ 是一列 σ-代数。如果满足：

$$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}_n \subseteq \mathscr{F}_{n+1} \subseteq \cdots \subseteq \mathscr{F},$$

则称 $\{\mathscr{F}_n\}_{n\geq 0}$ 为一个滤链（filtration）。直观上，它编码了逐渐增多的信息。

定义 27.1 (鞅)

设 $\{Z_n\}_{n\geq 0}$ 是一列可积的随机变量，并且对于每个 $n \geq 0$，Z_n 是 $\mathscr{F}_n$-可测的。如果

$$\forall n \geq 0,\ \mathbf{E}\left[Z_{n+1} \mid \mathscr{F}_n\right] = Z_n$$

成立，则称 $\{Z_n\}_{n\geq 0}$ 是相对于 $\{\mathscr{F}_n\}_{n\geq 0}$ 的鞅。

这个定义直观上就是在说，基于前 n 轮的信息，Z_{n+1} 的平均值是保持不变的（$=Z_n$）。

在上述定义中，如果把要求改为 $\forall n\geq 0,\ \mathbf{E}\left[Z_{n+1}\mid \mathscr{F}_n\right]\leq Z_n$，则称 $\{Z_n\}_{n\geq 0}$ 是一个上鞅 (super-martingale)。类似地，如果 $\forall n\geq 0, \mathbf{E}\left[Z_{n+1}\mid \mathscr{F}_n\right]\geq Z_n$，则称其为下鞅（sub-martingale）。

我们接下来看一些鞅的例子。

例题 27.1. 一维随机游走

考虑一个在整数集合 $\mathbb{Z}$ 上的随机游走，起点为 0。在每一步中，向左和向右移动的概率均为 $\frac{1}{2}$。设第 n 步的移动由一个取值为 $\{-1,+1\}$ 的均匀随机变量 X_n 表示。对于任何 $n\geq 1$，我们令

$$S_n=S_{n-1}+X_n.$$

容易可以验证 $\{S_n\}_{n\geq 0}$ 是相对于 $\{X_n\}_{n\geq 1}$（或 $\{S_n\}_{n\geq 0}$）的鞅：

$$\forall n\geq 0,\ \mathbf{E}\left[S_{n+1}\mid \overline{X}_{0,n}\right]=\mathbf{E}\left[S_n+X_{n+1}\mid \overline{X}_{0,n}\right]=S_n+\mathbf{E}\left[X_{n+1}\mid \overline{X}_{0,n}\right]=S_n.$$

更一般地，如果 $\mathbf{E}\left[X_{n+1}\mid \overline{X}_{0,n}\right]=\mu$，我们定义：

$$\forall n\geq 1, Y_n=X_n-\mu,\ S'_n=S'_{n-1}+Y_n=S_n-n\cdot\mu.$$

则 $\{S'_n\}_{n\geq 0}$ 是相对于 $\{Y_n\}_{n\geq 1}$ 的鞅。

例题 27.2. 均值为 1 的乘积

考虑一个随机变量序列 $\{X_n\}_{n\geq 1}$，其中对于所有 $n\geq 1$，有 $\mathbf{E}\left[X_n\mid \overline{X}_{1,n-1}\right]=1$。令：

$$P_n=\prod_{k=1}^{n}X_k.$$

我们可以验证 $\{P_n\}_{n\geq 0}$ 是相对于 $\{X_n\}_{n\geq 1}$ 的鞅：

$$\mathbf{E}\left[P_{n+1}\mid \overline{X}_{1,n}\right]=\mathbf{E}\left[P_n\cdot X_{n+1}\mid \overline{X}_{1,n}\right]=P_n\cdot\mathbf{E}\left[X_{n+1}\mid \overline{X}_{1,n}\right]=P_n.$$

例题 27.3. Galton-Watson 过程

Galton-Watson 过程是一个用来建模和研究某一个贵族姓氏灭绝概率的模型。在这个模型里，用 G_t 表示第 t 代的个体数量（为了方便，只考虑男性）。假设所有个体彼此独立的繁殖，并且后代数量同分布。我们用 $X_{t,k}$ 表示第 t 代第 k 个个体的（男性）后代个数。设 $\mu=\mathbf{E}\left[X_{t,k}\right]$。于是我们有

$$G_{t+1}=\sum_{k=1}^{G_t}X_{t,k}.$$

所以

$$\mathbf{E}\left[G_{t+1} \mid G_t\right] = \mathbf{E}\left[\sum_{k=1}^{G_t} X_{t,k} \,\middle|\, G_t\right] = G_t \cdot \mathbf{E}\left[X_{t,1}\right] = \mu \cdot G_t.$$

定义 $M_t := \mu^{-t} G_t$，则：

$$\mathbf{E}\left[M_{t+1} \mid G_t\right] = \mu^{-(t+1)} \mathbf{E}\left[G_{t+1} \mid G_t\right] = \mu^{-t} G_t = M_t.$$

因此，$\{M_t\}_{t\geq 1}$ 是相对于 $\{G_t\}_{t\geq 1}$ 的鞅。

例题 27.4. 波利亚的罐子（Pólya's Urn）

假设一个罐子中有若干白球和黑球，除了颜色外，所有球完全相同。考虑以下随机过程：每轮随机取一个球，记录其颜色，然后将该球放回，并添加一个相同颜色的球到罐中。假设初始时罐中只有一个白球和一个黑球，为了方便，我们让轮次从 2 开始计数，这样第 n 轮后罐中总共正好有 n 个球。令 X_n 表示第 n 轮后黑球的数量，$Z_n = \frac{X_n}{n}$ 表示第 n 轮后黑球所占的比例。显然有 $Z_2 = \frac{1}{2}$。对于所有 $n \geq 2$，

$$\mathbf{E}\left[Z_{n+1} \mid \overline{X}_{2,n}\right] = \frac{1}{n+1}\mathbf{E}\left[X_{n+1} \mid \overline{X}_{2,n}\right] = \frac{1}{n+1}\left(Z_n(X_n+1) + (1-Z_n)X_n\right) = Z_n.$$

因此，$\{Z_n\}_{n\geq 2}$ 是相对于 $\{X_n\}_{n\geq 2}$ 的鞅。

例题 27.5. Doob 鞅

有一类一般的构造鞅的方法叫做 Doob 鞅。它在随机过程的研究中有非常重要的应用，我们这门课并不会仔细讨论，但我觉得大家在未来的某一天会遇到。

给定一个 Borel 函数 $f\colon \mathbb{R}^n \to \mathbb{R}$，以及 n 个随机变量 $X_1, \ldots, X_n$ （$n \in \mathbb{N} \cup \{\infty\}$，但我们接下来为了方便就假设 n 是自然数）。对于 $k = 0, 1 \ldots, n$，定义 $Z_k = \mathbf{E}\left[f \mid \overline{X}_{1,k}\right]$。对于 $k > n$，定义 $Z_k = Z_n$。那么 $\{Z_k\}_{k\geq 0}$ 是相对于 $\{X_k\}_{k\geq 1}$ 的鞅，被称为 Doob 鞅。证明很简单，我们只需要对 $k = 0, 1, 2, \ldots, n-1$ 验证定义即可。使用条件期望的 Tower Rule：$\forall k = 0, 1, \ldots, n-1$，

$$\begin{aligned}
\mathbf{E}\left[Z_{k+1} \mid \overline{X}_{1,k}\right] &= \mathbf{E}\left[\mathbf{E}\left[f(X_1, \ldots, X_n) \mid \overline{X}_{1,k+1}\right] \mid \overline{X}_{1,k}\right] \\
&= \mathbf{E}\left[f(X_1, \ldots, X_n) \mid \overline{X}_{1,k}\right] \\
&= Z_k.
\end{aligned}$$

这个鞅的定义非常一般且抽象。我们来看一个具体例子。假设 $X_1, \ldots, X_n$ 分别对应了一个图片的每一个像素点的颜色。$(X_1, \ldots, X_n) \sim \mu$ 是一个随机的图片。而 $f(x_1, \ldots, x_n) = \mathbb{I}_{[(x_1, \ldots, x_n)]}$。那么 $Z_0 = \mathbf{E}[f] = \mathbb{P}(\,)$ 指的是我们完全没有看到这张图的时候该图片是一只猫的概率。而对于 $1 \leq k \leq n$，$Z_k = \mathbb{P}(\,|\,(X_1, \ldots, X_k))$ 指的是我们看到了前 k 个像素之后，我们对于这个图片是不是猫的猜测概率。特别的，当 $k = n$ 的时候，我们已经不需要猜测，因为 $Z_n = f(X_1, \ldots, X_n)$ 是 $\sigma(X_1, \ldots, X_n)$-可测的，它的值要么是 1 要么是 0，取决于给定的 $(X_1, \ldots, X_n)$ 是不是一只猫。

第 28 章 可选停时定理

回忆我们上节课介绍过的离散鞅的概念。固定概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 以及上面的一个滤链 $\{\mathscr{F}_t\}_{t\geq 0}$。我们说一族可积的随机变量 $\{X_t\}_{t\geq 0}$ 是相对于 $\{\mathscr{F}_t\}_{t\geq 0}$ 的鞅当且仅当

$$\forall t \geq 0,\ \mathbf{E}\left[X_{t+1} \mid \mathscr{F}_t\right] = X_t.$$

对于每一个 $t \geq 0$，我们对上式两边同时取期望并使用 tower rule，可以得到 $\mathbf{E}\left[X_{t+1}\right] = \mathbf{E}\left[X_t\right]$。换句话说，对于每一个固定的 t，我们都有 $\mathbf{E}\left[X_t\right] = \mathbf{E}\left[X_0\right]$ 成立。但我们现在假设 τ 是一个随机的时间，是一个随机变量，$\mathbf{E}\left[X_\tau\right] = \mathbf{E}\left[X_0\right]$ 依然成立吗？答案是不一定，我们可以看下面这个例子：

考虑一个公平游戏中的如下赌博策略。在第一轮中，赌徒下注 1 元。然后，他简单地将赌注翻倍，直到他赢得游戏（有趣的是，这个赌博策略也被称为 martingale）。如果我们用 Z_t 表示他 t 轮之后赢得的金额，在上节课我们已经知道 $\{Z_t\}_{t\geq 0}$ 是一个鞅。令 τ 表示他第一次赢得游戏的时间。首先我们知道 $\mathbb{P}\left(\tau < \infty\right) = 1$。注意到

1. 如果 $\tau = 1$，他赢得 1 元。
2. 如果 $\tau = 2$，他赢得 $-1 + 2 = 1$ 元。
3. 如果 $\tau = 3$，他赢得 $-1 - 2 + 4 = 1$ 元。
4. ...

他在赢得游戏的时候一定是赢得 1 元钱。也就是说 $\mathbf{E}\left[Z_\tau\right] = 1 \neq \mathbf{E}\left[Z_0\right] = 0$。我们马上会看到，理解在什么时候 $\mathbf{E}\left[Z_\tau\right] = \mathbf{E}\left[Z_0\right]$ 是一件很重要的事情。

28.1 可选停时定理（Optional Stopping Theorem）

我们首先要定义停时（stopping time）的概念。

定义 28.1 (停时)

设 $\{\mathscr{F}_t\}_{t\geq 0}$ 是概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的一个滤链。我们说取值为自然数的随机变量 $\tau\colon \Omega \to \mathbb{N}$ 为一个停时，当且仅当对于任何 $t \in \mathbb{N}$，事件 $[\tau \leq t]$ 是 $\mathscr{F}_t$-可测的。

我们有的时候也会说 τ 是关于 $X_0, X_1, \ldots$ 的停时，这个时候的意思是在上面的定义里取 $\mathscr{F}_i = \sigma\left(X_0, \ldots, X_i\right)$。

直观上说，如果想象在玩一个一轮又一轮的游戏，$\mathscr{F}_t$ 表示前 t 轮所有的信息。那么 τ 是一个停时的意思是在每一轮游戏结束后，玩家应该就有足够的信息判断是否应该在此时结束游戏了。举个例子，我们平时晚上打游戏的时候经常说的“赢一把就睡”就是一个停时，但比如“今晚赢的最多的一局后结束”就不是一个停时，因为在当前你不知道这一局是不是赢得最多的一局。

假设 $\{X_t\}_{t\geq 0}$ 是相对于 $\{\mathscr{F}_t\}_{t\geq 0}$ 的鞅，τ 是一个停时。可选停时定理给出了 $\mathbf{E}[X_\tau]=\mathbf{E}[X_0]$ 成立的几个充分条件。我们将在今天最后证明这个定理。从这个定理的证明中也容易看出来，$\mathbf{E}[X_\tau]=\mathbf{E}[X_0]$ 成立的本质原因是什么样的。

定理 28.1 (可选停时定理（Optional Stopping Theorem, OST）)

如果以下任意一个条件成立，则 $\mathbf{E}[X_\tau]=\mathbf{E}[X_0]$：

OST1 τ almost surely 有界，也就是说，存在某个 $n\in\mathbb{N}$ 使得 $\mathbb{P}(\tau\leq n)=1$；

OST2 $\mathbb{P}(\tau<\infty)=1$，且存在有限的 M，使得 $|X_t|\leq M$ 对于所有 $t\leq\tau$ 成立；

OST3 $\mathbf{E}[\tau]<\infty$，且存在一个常数 c，使得 $\mathbf{E}[|X_{t+1}-X_t|\mid\mathscr{F}_t]\leq c$ almost surely 对于所有 $t\leq\tau$ 成立。 ♡

我们之后会用 OST 来简称可选停时定理，会用 [OST1], [OST2], [OST3] 来分别指代我们使用 OST 的时候验证的充分条件。值得注意的是，[OST2] 里面的要求 $\mathbb{P}(\tau<\infty)=1$ 是比 [OST3] 里面的 $\mathbf{E}[\tau]<\infty$ 要更弱的条件，但作为代价，$|X_t|$ 有界是比 $\mathbf{E}[|X_{t+1}-X_t|\mid\mathscr{F}_t]$ 有界更强的条件。在具体应用中，我们要根据需要选择合适的条件。

首先来看一个耳熟能详的例子。假设有一个村庄，那儿的人重男轻女。那么在以下三种情况下，长期来看该村庄的性别比例是多少?

1. 每个家庭持续生育，直到他们生了一个男孩。
2. 每个家庭持续生育，直到男孩的数量多于女孩。
3. 每个家庭持续生育，直到男孩的数量多于女孩或孩子总数达到 10。

我们可以将问题建模为一个随机游走。假设有一个人在一维整数轴上随机游走。令 $\{X_t\}_{t\geq 0}$ 表示每个时间点上的位置，其中 X_t 表示一个家庭在前 t 个孩子中**男孩数量减去女孩数量**。初始时 $X_0=0$，在时间 $t=0$ 时，该人向 $c_t\in\{-1,1\}$ 方向随机迈出一步，到达 X_{t+1}，即 $X_{t+1}=X_t+c_t$。很容易验证 $\{X_t\}_{t\geq 0}$ 是一个鞅。

上述三种情况对应于停止时间 τ 的三种不同定义。我们可以简单的认为 $\mathbf{E}[X_\tau]=\mathbf{E}[X_0]$ 表明性别比例是平衡的。我们分别检查这些情况是否满足 OST。

1. τ 是 $c_t=1$ 的第一次出现的时间: 在这种情况下，由于 $\tau\sim\texttt{Geom}\left(\frac{1}{2}\right)$，所以 $\mathbf{E}[\tau]=\frac{1}{2}<\infty$。并且 $|X_{t+1}-X_t|\leq 1$ 对于所有 $t<\tau$ 成立。因此，根据 [OST3]，我们有 $\mathbf{E}[X_\tau]=\mathbf{E}[X_0]=0$。换句话说，如果人在 $c_t=1$ 第一次出现时停止，那么他停止的位置的期望是 0。
2. τ 是 $X_t=1$ 的第一次出现的时间: 在这种情况下，显然 $\mathbf{E}[X_\tau]=1\neq\mathbf{E}[X_0]$。这个过程被称为“一维随机游走的单吸收屏障问题”。容易证明（使用递推），$\mathbb{P}(\tau<\infty)=1$ 但是 $\mathbf{E}[\tau]=\infty$，没有任何一个 OST 条件被满足。在随机过程里，这个性质被称为“零返性”
3. τ 是 $t=10$ 和 $X_t=1$ 中最早发生的那个 t: 在这种情况下，τ 至多为 10，满足 [OST1]。因此我们有 $\mathbf{E}[X_\tau]=\mathbf{E}[X_0]=0$。

28.2 可选停时定理的应用

我们接下来介绍可选停时定理的一些经典应用，可以从中瞥见其威力。

28.2.1 Doob 的鞅不等式

通过可选停时定理，我们可以获得随机变量序列中最大值的集中性质。

引理 28.1

令 $\{X_t\}_{t\geq 0}$ 是一个关于自身的鞅，且对于每个 t 都有 $X_t\geq 0$。那么对于任何 $n\in\mathbb{N}$ 和 $\alpha>0$，有：

$$\mathbb{P}\left(\max_{0\leq t\leq n}X_t\geq\alpha\right)\leq\frac{\mathbf{E}[X_0]}{\alpha}.$$

♡

注：如果我们把引理里面的鞅改成下鞅，则不等式右边的 $\mathbf{E}[X_0]$ 可以换成 $\mathbf{E}[X_n \vee 0]$。这个的证明我们留作练习。

证明 定义停时 τ 为第一次出现大于 α 的 X_t 所对应的 t。如果不存在这样的 t，则令 $\tau = n$。即：

$$\tau := \min\left\{n, \min_{0\le t\le n}\{t \mid X_t \ge \alpha\}\right\}.$$

根据 τ 的定义，有：

$$\mathbb{P}\left(\max_{0\le t\le n} X_t \ge \alpha\right) = \mathbb{P}(X_\tau \ge \alpha).$$

由于 τ 有界（$\tau \le n$），[OST1] 成立，于是有 $\mathbf{E}[X_\tau] = \mathbf{E}[X_0]$。使用马尔科夫不等式便得到

$$\mathbb{P}(X_\tau \ge \alpha) \le \frac{\mathbf{E}[X_\tau]}{\alpha} = \frac{\mathbf{E}[X_0]}{\alpha}.$$

28.2.2 具有两侧吸收壁的一维随机游走

设 $a, b > 0$ 为两个正整数。一个人从位置 0 开始随机游走，每次等概率往左或者往右移动距离 1。当他到达 $-a$ 或 b 时停止。令 τ 表示他第一次到达 $-a$ 或 b 的时间，即 $\tau = \min\{t \ge 0 : X_t = -a \vee X_t = b\}$。显然 τ 是一个停时。该模型被称为**一维随机游走的双吸收屏障问题**。我们希望计算 $\mathbf{E}[\tau]$，即平均停止时间。

我们首先来计算 $\mathbb{P}(X_\tau = -a)$，即该人停在位置 $-a$ 的概率。令 $P_a := \mathbb{P}(X_\tau = -a)$。我们希望通过 OST 证明 $\mathbf{E}[X_\tau] = \mathbf{E}[X_0]$。为此，我们验证 OST 的几个条件。

1. 首先 [OST1] 在 a, b 不同时是 1 的时候显然不成立。
2. 我们可以验证更强的条件 $\mathbf{E}[\tau] < \infty$ 是成立的。我们使用一个基于耦合的证明。想象我每 $(a+b)$ 步投掷一枚 $\texttt{Ber}\left(2^{-(a+b)}\right)$ 的硬币。如果硬币是正面，就想象我在接下来的 $a+b$ 步里一直往右走。如果这件事情发生，那我在未来 $a+b$ 步内一定会停止。因此，我停止的时间 τ 可以被一个 $\texttt{Geom}\left(2^{-(a+b)}\right)$ 的随机变量乘上 $(a+b)$ 给控制住。因此 $\mathbf{E}[\tau] \le (a+b)\cdot 2^{a+b}$。此外，显然对于每个 $0 \le t \le \tau$，都有 $\mathbf{E}[|X_{t+1} - X_t| \mid \mathscr{F}_t] \le 1$，所以根据 [OST3]，$\mathbf{E}[X_\tau] = \mathbf{E}[X_0] = 0$。

另一方面，我们有 $\mathbf{E}[X_\tau] = P_a \cdot (-a) + (1 - P_a)\cdot b$。结合 $\mathbf{E}[X_\tau] = \mathbf{E}[X_0] = 0$，我们有 $P_a = \frac{b}{a+b}$。

我们接下来计算 $\mathbf{E}[\tau]$。对于每个 $t \ge 0$，我们定义一个新的随机变量 $Y_t := X_t^2 - t$。我们可以通过定义验证 $\{Y_t\}_{t\ge 0}$ 是一个鞅：

$$\forall t \ge 0,\ \mathbf{E}[Y_{t+1} \mid \mathscr{F}_t] = \mathbf{E}\left[X_{t+1}^2 - (t+1) \,\middle|\, \mathscr{F}_t\right].$$

根据 $X_{t+1} = X_t + c_t$，$c_t \in \{-1, 1\}$，有 $X_{t+1}^2 = (X_t + c_t)^2 = X_t^2 + 2X_t c_t + c_t^2$。代入后得到

$$\mathbf{E}[Y_{t+1} \mid \mathscr{F}_t] = \mathbf{E}\left[X_t^2 + 2X_t c_t + c_t^2 - (t+1) \,\middle|\, \mathscr{F}_t\right].$$

由于 X_t 是 $\mathscr{F}_t$-可测的，且 $\mathbf{E}[c_t \mid \mathscr{F}_t] = 0$，$\mathbf{E}\left[c_t^2 \,\middle|\, \mathscr{F}_t\right] = 1$。因此

$$\mathbf{E}[Y_{t+1} \mid \mathscr{F}_t] = X_t^2 + 0 + 1 - (t+1) = X_t^2 - t = Y_t.$$

注意 $X_t \in [-a, b]$，因此对于所有 $t \le \tau$，$|Y_{t+1} - Y_t|$ 是有界的，满足 [OST3]。我们使用 OST 得到

$$\mathbf{E}[Y_\tau] = \mathbf{E}[Y_0] = 0.$$

另一方面，由 Y_t 的定义，有 $\mathbf{E}[Y_\tau] = \mathbf{E}\left[X_\tau^2\right] - \mathbf{E}[\tau]$。因此 $\mathbf{E}[\tau] = \mathbf{E}\left[X_\tau^2\right]$。我们知道：

$$\mathbf{E}\left[X_\tau^2\right] = P_a \cdot (-a)^2 + (1 - P_a)\cdot b^2.$$

代入 $P_a = \frac{b}{a+b}$，得到：

$$\mathbf{E}[\tau] = \mathbf{E}\left[X_\tau^2\right] = \frac{b}{a+b}\cdot a^2 + \frac{a}{a+b}\cdot b^2 = ab.$$

28.2.3 模式的期望出现时间

假设有一个由 {H, T} 组成的字符串 P（称为模式串），长度为 ℓ（H 表示“正面”，T 表示“反面”）。我们连续抛掷硬币，直到最后的 ℓ 次结果形成一个与 P 完全相同的字符串。我们希望计算需要抛掷硬币的期望次数。

首先，如果我们抛掷硬币 N 次，并观察结果字符串 S。根据期望的线性性，无论 P 是什么，它在字符串 S 中的期望出现次数为

$$\mathbf{E}\left[\#\text{ of occurrences of } P \text{ in } S\right] = \sum_{i=1}^{n-\ell+1} \mathbf{E}\left[\mathbb{I}_{S_{i,i+1,\ldots,i+\ell-1}=P}\right] = (n-\ell+1)\cdot 2^{-\ell}.$$

也就是说，不管 P 是什么，它期望出现的次数总是一定的。但如果我们考虑第一次出现 P 的平均时间，就会有所不同。比如说，我们考虑以下两种模式串 HT 和 HH:

- 假设第一次抛掷结果是 H。如果第二次抛掷失败：
- 如果目标模式是 HT，那么尽管失败了，我们仍然得到了一个 H，模式串的第一位被匹配上了。
- 如果目标模式是 HH，第二次抛掷结果是 T，那么我们什么都没有得到，前两次抛掷的结果完全浪费。

因此，直观上我们应该相信，模式串 HT 的第一次出现的期望时间比 HH 更小。我们接下来严格的说明这件事情。

令模式 $P = p_1 p_2 \ldots p_\ell$。对于每个 $n \geq 0$，假设在第 $n+1$ 次抛掷之前，有一个新的赌徒 G_{n+1} 带着 1 单位的资金下注，赌接下来的 ℓ 次结果（即第 $n+1$ 到 $n+\ell$ 次结果）完全和 P 相同。换句话说，在第 $n+k$ 次抛掷时，赌徒 G_{n+1} 会采用全押策略下注第 $n+k$ 次的结果是 p_k：

- 如果第 $n+k$ 次结果是 p_k，赌徒的资金会翻倍；
- 否则，赌徒会输掉所有资金。

假设 P = HTHTH，抛掷结果为 HTHHTHTH。下表展示了每位赌徒在每次抛掷后的总资金：

Gambler	H	T	H	H	T	H	T	H	Money	
1	H	T	H	T					0	1→2→4→8→0
2		H							0	1→0
3			H	T					0	1→2→0
4				H	T	H	T	H	32	1→2→4→8→16→32
5					H				0	1→0
2						H	T	H	8	1→2→4→8
5							H		0	1→0
5								H	2	1→2

令 X_t 表示第 t 次抛掷的结果，$M_i(t)$ 表示赌徒 G_i 在第 t 次抛掷后的资金，定义：

$$Z_t := \sum_{i=1}^{t} \left(M_i(t) - 1\right),$$

即在第 t 次抛掷后，所有赌徒的总收入。可以验证对于每一个 i，$\{M_i(t)\}_{t\geq 0}$ 是一个关于抛掷结果 $\{X_t\}_{t\geq 1}$ 的鞅：

$$\mathbf{E}\left[M_i(t+1) \mid \overline{X}_{1,t}\right] = \frac{1}{2}\cdot 2M_i(t) + \frac{1}{2}\cdot 0 = M_i(t).$$

根据期望的线性性，我们也可以得到 $\{Z_t\}_{t\geq 0}$ 是一个鞅。令 τ 表示第一次某个赌徒赢得比赛的时间，即模式串 P 首次被掷出的时间。我们容易验证 [OST3] 是满足的（why?）。于是，$\mathbf{E}[Z_\tau] = \mathbf{E}[Z_0] = 0$。因此，我们得到

$$\mathbf{E}\left[\sum_{i=1}^{\tau} M_i(\tau) - \tau\right] = 0,$$

期望里的两项都是可积的（why?），于是根据期望的线性性，我们有

$$\mathbf{E}[\tau] = \mathbf{E}\left[\sum_{i=1}^{\tau} M_i(\tau)\right].$$

注意到对于 $i \le \tau - \ell$，$M_i(\tau) = 0$，而对于 $i > \tau - \ell$，有

$$M_i(\tau) = 2^{\tau - i + 1} \chi_{\tau - i + 1},$$

其中 $\chi_j = \mathbb{I}_{[p_1 p_2 \dots p_j = p_{\ell - j + 1} \dots p_\ell]}$。因此 $\mathbf{E}[\tau] = \sum_{i=1}^{\ell} 2^i \chi_i$。

我们对 HH 和 HT 来做一个 sanity check。

- 如果 $P = \texttt{HH}$，那么 $\mathbf{E}[\tau] = 2 + 4 = 6$。
- 如果 $P = \texttt{HT}$，那么 $\mathbf{E}[\tau] = 4$。

这验证了我们之前的直观：HH 的第一次出现的期望时间比模式 HT 更大。

28.2.4 Wald 等式

在实际中，我们经常需要分析如下过程的（期望）运行时间，其中 cond 和 compute() 都是随机的：

```
while cond do
    compute();
end while
```

假设第 i 次调用 compute() 的耗时为 X_i，算法在经过 T 次迭代后终止（T 有可能是随机的）。那么总运行时间为 $N := \sum_{i=1}^{T} X_i$。我们现在想来计算 $\mathbf{E}[N]$。

定理 28.2 (Wald 等式)

如果以下条件成立

1. $X_1, X_2, \dots$ 是非负的、独立的可积随机变量，并且期望均为 μ；
2. T 是 $X_1, X_2, \dots$ 的停时；
3. $\mathbf{E}[T] < \infty$。

那么就有 $\mathbf{E}\left[\sum_{i=1}^{T} X_i\right] = \mathbf{E}[T] \cdot \mu$ 成立。

♡

证明 对于 $i \ge 1$，定义随机变量

$$Z_i := \sum_{j=1}^{i} (X_j - \mu).$$

显然，序列 $\{Z_i\}_{i \ge 1}$ 是关于 $X_1, X_2, \dots$ 的鞅，并且 $\mathbf{E}[Z_1] = 0$。同时，我们有以下等式：

$$\mathbf{E}[|Z_{i+1} - Z_i| \mid \mathscr{F}_i] = \mathbf{E}[|X_{i+1} - \mu| \mid \mathscr{F}_i] \le \mathbf{E}[X_{i+1} + \mu \mid \mathscr{F}_i] \le 2\mu.$$

我们知道 $\mathbf{E}[T] < \infty$，因此满足 [OST3]，所以 $\mathbf{E}[Z_T] = \mathbf{E}[Z_1] = 0$。于是

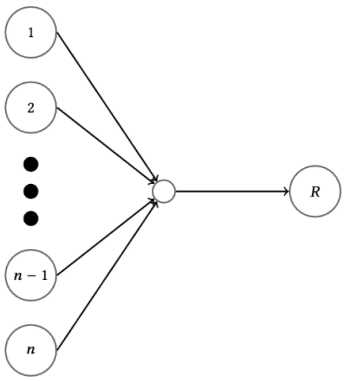
$$\mathbf{E}[Z_T] = \mathbf{E}\left[\sum_{j=1}^{T} (X_j - \mu)\right] = \mathbf{E}\left[\sum_{j=1}^{T} X_j\right] - \mathbf{E}[T] \cdot \mu = 0.$$

由此可得 $\mathbf{E}\left[\sum_{j=1}^{T} X_j\right] = \mathbf{E}[T] \cdot \mu$。

例题 28.1. Wald 等式的应用：一个路由问题

我们来看一个 Wald 等式的应用。如下图所示，假设有 n 个发送者和一个接收者。在每一轮中，每个发送者以概率 $\frac{1}{n}$ 向接收者发送一个数据包。由于所有发送者共享同一个信道，如果有多个数据包同时被发送，则所有数据包都会失败。

我们的问题是，每个发送者至少成功发送一个数据包所需的平均轮次是多少？



令 X_i 为接收者在成功接收到第 $i-1$ 个数据包后，再次成功接收到一个数据包所需的轮数。令 T 表示接收者第一次成功收到每个发送者至少一个数据包所需的数据包总数。那么我们感兴趣的总时间为：

$$N := \sum_{i=1}^{T} X_i.$$

显然，$X_1, X_2, \ldots$ 是独立同分布的，且 $\mathbf{E}[T]$ 有界。因此，根据 Wald 等式有 $\mathbf{E}[N] = \mathbf{E}[T] \cdot E[X_1]$。注意到根据定义，$T$ 是奖券收集问题中收集全一套所需要的开包总数，因此 $\mathbf{E}[T] = nH_n = \Theta(n \log n)$。另一方面，$X_1$ 服从几何分布 $\texttt{Geom}(p)$，其中

$$p = n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \approx e^{-1}.$$

所以 $\mathbf{E}[X_1] = \frac{1}{p} = e$。我们最终得到 $\mathbf{E}[N] = \mathbf{E}[T] \cdot \mathbf{E}[X_1] \approx e \cdot nH_n$。

28.3 可选停时定理的证明

我们重述一下 OST 并证明之。

定理 28.3 (可选停时定理（Optional Stopping Theorem, OST）)

如果以下任意一个条件成立，则 $\mathbf{E}[X_\tau] = \mathbf{E}[X_0]$：

OST1 τ almost surely 有界，也就是说，存在某个 $n \in \mathbb{N}$ 使得 $\mathbb{P}(\tau \le n) = 1$；

OST2 $\mathbb{P}(\tau < \infty) = 1$，且存在有限的 M，使得 $|X_t| \le M$ 对于所有 $t \le \tau$ 成立；

♡

证明 首先我们可以不失一般性的把 almost surely 去掉，因为在零测集上的取值并不影响期望。然后我们注意到，对于每一个 $n \in \mathbb{N}$，鞅的性质保证了 $\mathbf{E}[X_n] = \mathbf{E}[X_0]$。接下来，我们证明对于任意 $n \in \mathbb{N}$，有 $\mathbf{E}[X_{\tau \wedge n}] = \mathbf{E}[X_0]$ 成立。为了说明这个，我们定义

$$\forall n \ge 0,\ Z_n := X_{n \wedge \tau} = X_0 + \sum_{i=0}^{n-1} (X_{i+1} - X_i)\, \mathbb{I}_{[i<\tau]}.$$

我们接着验证 $\{Z_n\}_{n \ge 0}$ 是一个鞅。根据定义

$$\mathbf{E}[Z_{n+1} \mid \mathscr{F}_n] = \mathbf{E}\left[Z_n + (X_{n+1} - X_n)\, \mathbb{I}_{[n<\tau]} \mid \mathscr{F}_n\right].$$

由于 $[\tau > n]$ 是 $\mathscr{F}_n$-可测的，且 $\mathbf{E}[X_{n+1} \mid \mathscr{F}_n] = X_n$，我们有

$$\mathbf{E}[Z_{n+1} \mid \mathscr{F}_n] = Z_n + \mathbb{I}_{[n<\tau]} \cdot (\mathbf{E}[X_{n+1} \mid \mathscr{F}_n] - X_n) = Z_n.$$

因此，$\{Z_n\}_{n \ge 0}$ 是一个鞅。所以

$$\mathbf{E}[X_{\tau \wedge n}] = \mathbf{E}[Z_n] = \mathbf{E}[Z_0] = \mathbf{E}[X_0].$$

上述讨论促使我们把 X_τ 分解成以下两部分：

$$X_\tau = X_{\tau\wedge n} + \mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n).$$

对两边取期望并令 $n \to \infty$，得到：

$$\mathbf{E}[X_\tau] = \mathbf{E}[X_0] + \lim_{n\to\infty} \mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n)\right].$$

因此，我们只需验证 [OST1], [OST2], [OST3] 都能保证保证 $\lim_{n\to\infty} \mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n)\right] = 0$。

例题 28.2. OST1

如果 τ 几乎必然有界，则存在某个 n 使得 $\mathbb{P}(\tau \le n) = 1$。因此，对于足够大的 n，有：

$$\mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n)\right] = 0.$$

例题 28.3. OST2

在这种情况下：

$$\mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n)\right] \le \mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (|X_\tau| + |X_n|)\right].$$

由于 $|X_t| \le M$ 对于所有 $t < \tau$ 成立，因此：

$$\mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (|X_\tau| + |X_n|)\right] \le 2M \cdot \mathbb{P}(\tau > n).$$

而当 $n \to \infty$ 时，$\mathbb{P}(\tau > n) \to 0$。

例题 28.4. OST3

在这种情况下，我们要利用 X_t 之间的增量有界。于是

$$\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n) = \sum_{t=n}^{\tau-1} (X_{t+1} - X_t) \le \sum_{t=n}^{\tau-1} |X_{t+1} - X_t| = \sum_{t=n}^{\infty} |X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]}.$$

对两边取期望，并使用 MCT，可以得到

$$\mathbf{E}\left[\mathbb{I}_{[n<\tau]} \cdot (X_\tau - X_n)\right] \le \mathbf{E}\left[\sum_{t=n}^{\infty} |X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]}\right] = \sum_{t=n}^{\infty} \mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]}\right].$$

使用 tower rule，并注意到 $\mathbb{I}_{[\tau>t]}$ 是 $\mathscr{F}_t$-可测的，我们知道对于任何 $t \ge n$ 有

$$\mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]}\right] = \mathbf{E}\left[\mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]} \mid \mathscr{F}_t\right]\right] = \mathbf{E}\left[\mathbf{E}\left[|X_{t+1} - X_t| \mid \mathscr{F}_t\right] \cdot \mathbb{I}_{[\tau>t]}\right].$$

根据 [OST3]，对于 $t \le \tau$，$\mathbf{E}\left[|X_{t+1} - X_t| \mid \mathscr{F}_t\right] \le c$，我们有

$$\sum_{t=n}^{\infty} \mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{I}_{[\tau>t]}\right] \le \sum_{t=n}^{\infty} c \cdot \mathbb{P}(\tau > t).$$

我们又知道 $\mathbf{E}[\tau] = \sum_{t=0}^{\infty} \mathbb{P}(\tau > t) < \infty$。因此上式在 $n \to \infty$ 的时候是一个有限级数的 tail，所以收敛到零。