# 概率论

孙天阳

中国科学技术大学数学科学学院

tysun@mail.ustc.edu.cn

2025 年 12 月 7 日

# 目录

# Chapter 1

# 随机变量

## 1 随机变量

**定义 1.1.** 设 $(\Omega, \mathcal{F}, P)$ 是一个概率空间, 称一个可测函数 $X\colon \Omega \to \mathbb{R}$ 为一个随机变量.

考虑随机变量有如下好处. 一方面, 相较于弄清楚 $\Omega$ 中有哪些可测集, 每个可测集的测度是多少, 随机变量可以帮助我们聚焦在自己关心的事情上. 比如考虑一个只取值 0 和 1 的随机变量 $X$, 其实也就是说我们只关心 $X=0$ 与 $X=1$ 发生的概率. 我们不关心 $\mathcal{F}$ 中所有的元素, 只关心

$$\sigma(X) = \{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathcal{B}(\mathbb{R})\}$$

这是 $\mathcal{F}$ 的一个子 $\sigma$-代数, 只保留了 $X$ 能区分的那些信息. 另一方面, 随机变量将实验结果数值化, 使我们能够用数学工具来分析随机现象, 甚至可以忘掉背后的概率空间 $\Omega$ 具体的样子. 为了用严格的数学语言来说明, 我们需要引入推出测度的概念:

**定义 1.2.** 给定可测映射 $f\colon (E_1, \Sigma_1) \to (E_2, \Sigma_2)$ 和测度 $\mu : \Sigma_1 \to [0, +\infty]$. 定义推出测度 $f_*\mu$ 为

$$f_*\mu(B) = \mu(f^{-1}(B)), \quad \forall B \in \Sigma_2.$$

把这套语言用在概率空间和随机变量上, 也就是说我们可以通过随机变量 $X$ 将概率空间上的测度 $P$ 推出到 $\mathbb{R}$ 上. 事实上, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*P)$ 在某种意义下同构于 $(\Omega, \sigma(X), P)$. 这也就是概率论中我们直接提及服从某个分布的随机变量 $X$ 而可以不关心背后的 $\Omega$ 是谁的逻辑基础. 相比于抽象的空间 $\Omega$ 上的测度 $P$, 实轴 $\mathbb{R}$ 上的测度 $X_*P$ 更方便我们使用微积分的工具. 由 Lebesgue 分解定理,

$$X_*P = \alpha\mu_{\text{discrete}} + \beta\mu_{\text{ac}} + \gamma\mu_{\text{sc}}$$

其中 $\alpha+\beta+\gamma=1$. 当 $\beta=0, \gamma=0$ 时, 是离散型随机变量. 当 $\alpha=0, \gamma=0$ 时, 是连续型随机变量.

关于推出测度, 我们有如下的积分换元公式

**命题 1.1.** 设 $g\colon E_2 \to \mathbb{R}$ 是一个可测函数, 则 $g$ 关于 $f_*\mu$ 可积当且仅当 $g \circ f$ 关于 $\mu$ 可积, 且

$$\int g f_*\mu = \int (g \circ f)\mu$$

取 $f$ 为随机变量 $X$, 取 $g$ 为恒等映射 $x \mapsto x$, 我们便得到 $\int_{\mathbb{R}} x X_*P = \int_\Omega XP$, 这是数学期望. 类似的, 佚名统计学家公式 $\mathbb{E}[g(X)] = \int_{\mathbb{R}} x g_* X_* P = \int_{\mathbb{R}} g(x) X_* P$ 也来自于这个命题. 在这里我们并不给出这个命题的证明, 它可以通过标准的四步法来完成.

## 2 积测度

我们经常提到独立同分布的随机变量, 还会把它们加起来, 但你有没有想过, 它们背后的测度空间是什么? 稍微想想就能知道, 独立同分布的随机变量, 比如独立抛两次硬币的结果, 绝对不可能简单地定义在同一个测度空间 $\{正, 反\}$ 上, 那又是怎么把这两个随机变量加起来的呢? 这就不得不求助于构造乘积空间和乘积测度. 假设有随机变量

$$X\colon \Omega_1 \longrightarrow \mathbb{R}, \quad Y\colon \Omega_2 \longrightarrow \mathbb{R},$$

定义 $\Omega = \Omega_1 \times \Omega_2$, 在乘积空间 $\Omega$ 上可以定义两个新的随机变量 $\tilde{X}$ 和 $\tilde{Y}$ 为

$$\tilde{X}(\omega_1, \omega_2) = X(\omega_1), \quad \tilde{Y}(\omega_1, \omega_2) = Y(\omega_2).$$

现在虽然构造了空间和函数, 但我们还没有定义 $\Omega$ 上的概率测度 $P$. 事实上, $\Omega$ 上不止有一种 $P$ 可以满足 $\pi_{1*}P = P_1$ 且 $\pi_{2*}P = P_2$. 比如我们以抛硬币为例子,

$$\Omega_1 = \Omega_2 = \{0, 1\}, \quad P_1(0) = P_1(1) = P_2(0) = P_2(1) = 0.5, \quad \Omega = \{(0,0), (0,1), (1,0), (1,1)\}$$

$$P_A(0,0) = P_A(0,1) = P_A(1,0) = P_A(1,1) = 0.25, \quad P_B(0,0) = P_B(1,1) = 0.5, P_B(0,1) = P_B(1,0) = 0$$

所以在乘积空间上选择哪个测度, 依赖于我们对问题的认知, 依赖于我们对 $X$ 和 $Y$ 的关系或者说 $\Omega_1$ 和 $\Omega_2$ 的关系的认知, 或者说假设. 我们假设 $X$ 和 $Y$ 独立, 其实相当于在 $\Omega$ 上选择测度 $P_A$.

**定义 2.1.** *设有两个测度空间 $(\Omega_1, \mathcal{F}_1, \mu_1), (\Omega_2, \mathcal{F}_2, \mu_2)$, 定义其乘积测度空间为*

$$\Omega = \Omega_1 \times \Omega_2, \quad \mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\{A \times B\colon A \in \mathcal{F}_1, B \in \mathcal{F}_2\}), \quad \mu(A \times B) = \mu_1(A) \cdot \mu_2(B)$$

*由 Carathéodory 延拓定理, $\mu$ 可以唯一延拓到整个 $\mathcal{F}$.*

# 3 分布函数

**定义 3.1.** 设 $X$ 是随机变量, 称函数 $F(x) = \int_{-\infty}^{x} X_* P$ 为 $X$ 的分布函数.

**命题 3.1.** 设 $F(x)$ 是分布函数，则

(1) $F(x)$ 单调递增.

(2) $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to+\infty} F(x) = 1$.

(3) $F(x)$ 右连续.

证明. □

**注记.**

(1) 可以证明，满足上述三条的函数必为某概率空间上某随机变量的分布函数

(2) $F(x)$ 单调递增的充分条件是 $F$ 几乎处处可导且 $F' > 0$.

# 4 随机向量

**定义 4.1.**

(1) 设 $X_1,\cdots,X_n$ 为 $(\Omega,\mathcal{F},P)$ 上的随机变量，称 $\vec{X}=(X_1,\cdots,X_n)$ 为 $n$ 维随机向量，称 $n$ 元函数 $F(x_1,\cdots,x_n)=P(X_1\leqslant x_1,\cdots,X_n\leqslant x_n)$ 为 $\vec{X}$ 的联合分布函数.

(2) 若 $\vec{X}$ 只取 $\mathbb{R}^n$ 中的可列个点，则称 $\vec{X}$ 是离散型随机向量，称 $n$ 元函数 $f(x_1,\cdots,x_n)=P(X_1=x_1,\cdots,X_n=x_n)$ 为联合分布列.

(3) 若存在非负可积函数 $f(x_1,\cdots,x_n)$ 使得

$$F(x_1,\cdots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}f(u_1,\cdots,u_n)\mathrm{d}u_1\cdots\mathrm{d}u_n,$$

则称 $\vec{X}$ 为连续型随机向量，称 $f$ 为联合密度函数.

**注记.** 任意两个随机变量都有联合分布函数，但即使是两个连续型随机变量，组合在一起也不一定是连续型随机向量.

# 5 独立性

**例 5.1.** 二项分布

几何分布称随机变量 $X$ 服从以 $p\in(0,1)$ 为参数的几何分布，是指 $X$ 的分布列满足

$$f(k)=P(X=k)=q^{k-1}p,\quad k=1,2,\cdots$$

称随机变量 $X$ 服从以 $\lambda>0$ 为参数的 *Poisson* 分布，是指 $X$ 的分布列满足

$$f(k)=P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\quad k=0,1,2,\cdots$$

记作 $X\sim P(\lambda)$. 若 $X\sim P(\lambda)$，容易算得 $E[X]=\lambda$.

**定义 5.1.** 称离散型随机变量 $X$ 和 $Y$ 独立，若 $\forall x,y\in\mathbb{R}$，有

$$P(X=x,Y=y)=P(X=x)P(Y=y).$$

称 $X_1,\cdots,X_n$ 相互独立，若任意 $x_i\in\mathbb{R}$，

$$P(X_1=x_1,\cdots,X_n=x_n)=P(X_1=x_1)\cdots P(X_n=x_n).$$

更一般地，称可列个随机变量 $\{X_n\}$ 相互独立，是指其中任意有限多个相互独立.

# 6 数学期望

**定义 6.1.** 设 $X$ 是随机变量, 则它的数学期望定义为 $\int xX_*P$, 记作 $\mathbb{E}[X]$.

**命题 6.1.** 设 $X$ 是随机变量, $g\colon \mathbb{R} \to \mathbb{R}$ 可测, 则 $\mathbb{E}[g(X)] = \int g(x)X_*P$.

**定理 6.1.** 期望算子 $E$ 满足

(1) 非负性. 当 $X \geqslant 0$ 时，$E[X] \geqslant 0$.

(2) 归一性. $E[1] = 1$.

(3) 线性性. $E[aX + bY] = aE[X] + bE[Y]$.

**命题 6.2.** 设 $X$ 与 $Y$ 期望均存在. 若 $X$ 与 $Y$ 独立，则

$$E[XY] = E[X]E[Y].$$

# 7 协方差

# 8 依赖性

**定义 8.1.** 联合分布与联合分布列

**引理 8.1.** 离散随机变量 $X$ 和 $Y$ 是独立的当且仅当

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \forall x, y \in \mathbb{R}.$$

更一般地，$X$ 和 $Y$ 是独立的当且仅当 $f_{X,Y}$ 能够被分解为分别只依赖于 $x$ 与 $y$ 的函数 $g(x)$ 与 $h(y)$ 的乘积.

# 9 条件期望

**定义 9.1.** 设 $(\Omega, \mathcal{F}, P)$ 是一个概率空间, 随机变量 $X \in L^1$, 设 $\mathcal{G} \subset \mathcal{F}$ 是一个子 $\sigma$-代数. 定义 $X$ 关于 $\mathcal{G}$ 的条件期望为一个 $\mathcal{G}$ 可测的随机变量, 记作 $\mathbb{E}[X|\mathcal{G}]$, 满足

$$\int_G \mathbb{E}[X|\mathcal{G}]dP = \int_G XdP, \quad \forall G \in \mathcal{G}.$$

不平凡的地方在于 $X$ 不一定是 $\mathcal{G}$ 可测的, 否则取 $X$ 即可. 这也启发我们可以考虑一下 $\sigma(X)$ 与 $\mathcal{G}$ 的大小关系

**定义 9.2.** 设 $(X,Y)$ 是离散型随机向量，若 $P(X=x)>0$，则称

$$f_{Y|X}(y|x)=P(Y=y\big|X=x)$$

是给定 $X=x$ 下 $Y$ 的条件分布列，

$$F_{Y|X}(y|x)=P(Y\leqslant y\big|X=x)$$

是给定 $X=x$ 下 $Y$ 的条件分布，

$$\psi(x)=E[Y\big|X=x]=\sum_y yf_{Y|X}(y\big|x)$$

是给定 $X=x$ 下 $Y$ 的条件期望，

# 10 随机游走

## 10.1 $n$ 步游走最远距离

设 $S_0 = 0$，记 $M_n = \max\{S_0, S_1, \cdots, S_n\}$.

- 不是指绝对值的最大，而是指通常序关系 $\cdots < -1 < 0 < 1 < \cdots$ 下的最大.
- 但绝对值最大也是值得关心的问题.

直接考虑 $M_n = r$ 的概率并不容易，往往是先考虑 $M_n \geqslant r$ 的概率，还要用全概率公式讨论 $S_n$ 的位置. 确定下 $S_n$ 的位置的好处在于左右步数的分配被确定，从而概率的计算问题转化为轨道的数数问题.

剩下的计算是自然的：

$$P(M_n \geqslant r, S_n = b) = \begin{cases} P(S_n = b), & b \geqslant r \\ \left(\frac{q}{p}\right)^{r-b} P(S_n = 2r - b) & b < r \end{cases}$$

- $M_n \geqslant r$ 被转化为经过直线 $y = r$ 至少一次.
- 当 $S_n = b$ 落在直线 $y = r$ 下方时要想数过程中经过 $y = r$ 至少一次的轨道当然是把 $S_n = b$ 对称过去. 同时虽然轨道数是相同的，但因为 $2r - b$ 相较 $b$ 靠右 $2r - 2b$ 个格，粒子就得向右多跳 $r - b$ 步，向左少跳 $r - b$ 步，从而会多一个 $\left(\frac{q}{b}\right)^{r-b}$ 的系数.
- 上式对于 $r \geqslant 0$ 都是对的.

对称随机游走的好处是，即使左右步数的分配不同，即终点的位置不同，不同轨道的权重仍然是相同的，反映在上面的式子中，就是系数 $\left(\frac{q}{p}\right)^{r-b}$ 没掉了，从而我们可以直接相加，有

$$P(M_n \geqslant r) = P(S_n \geqslant r) + P(S_n \geqslant r + 1).$$

# Chapter 2

# 生成函数及其应用

## 1 生成函数

### 1.1 数列的生成函数

### 1.2 非负整值随机变量的生成函数

### 1.3 典型分布的特征函数

### 1.4 与数字特征的关系

# Chapter 3

# 正态分布

# 1 例子

## 1.1 正态分布

正态分布的概率密度函数我是记不住的，哪怕能背下来时间久了也会忘，但是我只需要知道

-

**定义 1.1.** 设 $X_1, \cdots, X_n$ 为概率空间 $(\Omega, \mathcal{F}, P)$ 上的随机变量，若对任意 $x_1, \cdots, x_n \in \mathbb{R}$ 成立

$$P(X_1 \leqslant x_1, \cdots, X_n \leqslant x_n) = P(X_1 \leqslant x_1) \cdots P(X_n \leqslant x_n),$$

则称 $X_1, \cdots, X_n$ 是相互独立的.

## 1.2 $\Gamma$ 分布

- $\Gamma(t) = \int_0^{+\infty} x^{t-1} \mathrm{e}^{-x} \mathrm{d}x$
- $\int_0^{+\infty} \frac{1}{\Gamma(t)} x^{t-1} \mathrm{e}^{-x} \mathrm{d}x = 1$
- $\int_0^{+\infty} \frac{1}{\Gamma(t)} (\lambda x)^{t-1} \mathrm{e}^{-\lambda x} \mathrm{d}(\lambda x) = \int_0^{+\infty} \frac{1}{\Gamma(t)} \lambda^t x^{t-1} \mathrm{e}^{-\lambda x} \mathrm{d}x = 1$

## 2 一般理论

**定义 2.1.** 设 $X$ 与 $Y$ 是独立的连续型随机变量，称 $Z := X + Y$ 为 $X$ 与 $Y$ 的卷积.

**命题 2.1.** 设 $(X, Y)$ 有密度 $f(x, y)$，则 $Z = X + Y$ 有密度

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x)\mathrm{d}x = \int_{-\infty}^{+\infty} f(z - y, y)\mathrm{d}y.$$

证明.

$$\begin{aligned} P(Z \leqslant z) &= P(X + Y \leqslant z) \\ &= \iint_{x+y\leqslant z} f(x, y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} f(x, y)\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z} f(x, y - x)\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{z} \int_{-\infty}^{+\infty} f(x, y - x)\mathrm{d}y\mathrm{d}x \\ f_Z(z) &= \int_{-\infty}^{+\infty} f(x, z - x)\mathrm{d}x \end{aligned}$$

□

# 3 条件期望

# 4 多元正态分布

**定义 4.1.** 称 $\vec{X}=(X_1,\cdots,X_n)$ 服从 $n$ 元正态分布，如果 $\vec{X}$ 有概率密度函数

$$f(\vec{x})=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})^T\right\},$$

其中 $\vec{\mu}\in\mathbb{R}^n$，$\Sigma$ 为正定矩阵. 记 $\vec{X}\sim N(\vec{\mu},\Sigma)$.

验证 $f(\vec{x})$ 确为概率密度函数：令 $\vec{X}=\vec{\mu}+\vec{Y}O$，其中 $O$ 是使得 $O\Sigma O^T=\Lambda=diag(\lambda_1,\cdots,\lambda_n)$ 的正交矩阵. 由线性代数的知识，这样的 $O$ 是存在的. 则

$$\begin{aligned}\int_{\mathbb{R}^n}f(\vec{x})\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int_{\mathbb{R}^n}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})^T\right\}\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\\&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int_{\mathbb{R}^n}\exp\left\{-\frac{1}{2}\vec{y}\Lambda^{-1}\vec{y}^T\right\}\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^n\\&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\prod_{i=1}^n\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2}\frac{y_i^2}{\lambda_i}\right\}\mathrm{d}y^i\\&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\prod_{i=1}^n\sqrt{2\pi\lambda_i}=1.\end{aligned}$$

**命题 4.1.** 设 $\vec{X}\sim N(\vec{\mu},\Sigma)$，则

(1) $E[\vec{X}]=\vec{\mu}$.

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

证明.

(1)

$$\begin{aligned}E[X_i]&=\int_{\mathbb{R}^n}x_if(\vec{x})\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\\&=\int_{\mathbb{R}^n}(\mu_i+y_jo_{ji})\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{1}{2}\vec{y}\Lambda^{-1}\vec{y}^T\right\}\mathrm{d}y^1\wedge\cdots\wedge\mathrm{d}y^n\\&=\mu_j\end{aligned}$$

(2)

□

**定理 4.1.** 设 $\vec{X}\sim N(\vec{\mu},\Sigma)$，$D$ 是 $n$ 阶非奇异方阵. 令 $\vec{Y}=\vec{X}D$，则 $\vec{Y}\sim N(\vec{\mu}D,D^T\Sigma D)$.

证明.

$$\begin{aligned}f_X(\vec{x})\mathrm{d}\vec{x}&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})\Sigma^{-1}(\vec{x}-\vec{\mu})^T\right\}\mathrm{d}\vec{x}\\&=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{1}{2}(\vec{y}D^{-1}-\vec{\mu})\Sigma^{-1}(\vec{y}D^{-1}-\vec{\mu})^T\right\}\mathrm{d}\vec{y}D^{-1}\\&=\frac{1}{\sqrt{(2\pi)^n|D^T\Sigma D|}}\exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu}D)(D^T\Sigma D)^{-1}(\vec{y}-\vec{\mu}D)^T\right\}\mathrm{d}\vec{y}\end{aligned}$$

$$= f_Y(\vec{y})\mathrm{d}\vec{y}$$

□

**定理 4.2.** 设 $\vec{X} \sim N(\vec{\mu}, \Sigma)$，$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$，则 $\vec{X}^{(i)} \sim N(\vec{\mu}^{(i)}, \Sigma_{ii})$.

**定理 4.3.** 设 $\vec{X} \sim N(\vec{\mu}, \Sigma)$，则 $X_1, \cdots, X_n$ 相互独立 $\iff$ $\Sigma$ 为对角阵.

# 5 高斯混合模型

有许多现象无法用单个正态分布来表示, 我们可以将多个正态分布结合起来, 以表达多个峰值,

$$p(\vec{x}) = \sum_{k=1}^{K} \phi_k N(\vec{x}; \vec{\mu}_k, \Sigma_k)$$

可以看到这里是对概率密度函数进行了凸组和. 注意与随机变量的加法做区分, 随机变量做加法反映到概率密度函数上是做卷积. 在实践中要在一个高斯混合模型中采样时, 先以 $\phi_k$ 为概率决定在哪个正态分布中采样, 再在那个正态分布中采样. 可以考虑一个离散型随机变量 $z$ 满足 $P(z=k)=\phi_k$, 我们称它为隐变量. 举个例子, 考虑某所学校初一学生的身高, 会发现出现了双峰, 这是因为学生里混合了男生和女生, 这个时候隐变量就是性别.

高斯混合模型的极大似然估计比单个正态分布的极大似然估计要复杂. 假设我们有 $N$ 个观测数据点 $\vec{x}_1, \cdots, \vec{x}_N$, 假设模型由两个正态分布混合而成

$$p(\vec{x}) = \phi_1 N(\vec{x}; \vec{\mu}_1, \Sigma_1) + \phi_2 N(\vec{x}; \vec{\mu}_2, \Sigma_2)$$

对数似然是

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ln \left( \underbrace{\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}_{\text{Sum inside Log}} \right)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_1} = \sum_{i=1}^{N} \left( \frac{1}{\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \dots) + \pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \dots)} \right) \cdot \frac{\partial}{\partial \boldsymbol{\mu}_1} \left( \pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \right)$$

对于 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$，其关于 $\boldsymbol{\mu}$ 的导数是：

$$\frac{\partial \mathcal{N}}{\partial \boldsymbol{\mu}} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

代回原式：

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_1} = \sum_{i=1}^{N} \frac{\pi_1 \cdot \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)}{\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} = \mathbf{0}$$

我们将上式中 $\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)$ 以外的部分单独拿出来看，这正是贝叶斯公式的形式，也就是样本 $\mathbf{x}_i$ 属于第一个分量的后验概率（Responsibility），记为 $\gamma_{i1}$：

$$\gamma_{i1} = \frac{\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{\pi_1 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}$$

于是，原本复杂的导数方程可以被简化写为：

$$\sum_{i=1}^{N} \gamma_{i1} \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1) = \mathbf{0}$$

假设 $\boldsymbol{\Sigma}_1$ 可逆，我们可以把 $\boldsymbol{\Sigma}_1^{-1}$ 约掉（或者两边同乘 $\boldsymbol{\Sigma}_1$），得到：

$$\sum_{i=1}^{N} \gamma_{i1}(\mathbf{x}_i - \boldsymbol{\mu}_1) = \mathbf{0}$$

$$\sum_{i=1}^{N} \gamma_{i1} \mathbf{x}_i = \sum_{i=1}^{N} \gamma_{i1} \boldsymbol{\mu}_1$$

因为 $\boldsymbol{\mu}_1$ 与求和下标 $i$ 无关，提出来：

$$\boldsymbol{\mu}_1 = \frac{\sum_{i=1}^{N} \gamma_{i1} \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{i1}}$$

既然等式两边都有 $\boldsymbol{\mu}_1$，没法直接解，数学家们就想了一个“偷懒”的办法——迭代（Iterative）：E 步（Expectation）：先随便猜一个 $\boldsymbol{\mu}_1^{(old)}$ 的值，代入到等式右边的 $\gamma_{i1}$ 里。把 $\gamma_{i1}$ 算出来并固定住（当成常数）。M 步（Maximization）：现在 $\gamma_{i1}$ 是常数了，那么等式就变成了 $\boldsymbol{\mu}_1^{(new)}$ = 常数加权平均。这一步就能直接算出新的 $\boldsymbol{\mu}_1$。循环：把算出来的新 $\boldsymbol{\mu}_1$ 再代回 E 步，周而复始，直到收敛。

这里是通过求梯度并令它等于 0 的方式得到了 $\gamma$ 和 $\mu$ 之间的关系，也让我们看到为什么后验概率会出现，为什么我们需要知道后验概率。

对任意 $z$ 有 $p_\theta(x) = \frac{p_\theta(x,z)}{p_\theta(z|x)}$, 而对固定的 $x$ 又有 $\sum_z p_\theta(z|x) = 1$

$$\log p_\theta(x) = \sum_z p_\theta(z|x) \log \frac{p_\theta(x,z)}{p_\theta(z|x)}$$

这是一个非常深刻且令人着迷的视角，它连接了统计学（信息论）与统计力学（热力学）。你给出的公式，在物理学中对应的正是 ** 亥姆霍兹自由能（Helmholtz Free Energy）** 的定义。我们可以通过构建一个“物理词典”来将这两者一一对应，从而揭示 EM 算法背后的物理直觉：最大化似然，等价于最小化系统的自由能。1. 建立映射：统计学 vs 统计力学为了对应物理概念，我们需要定义“能量”和“温度”。在统计物理中，玻尔兹曼分布（Boltzmann Distribution）告诉我们，一个系统处于某个微观状态 $z$ 的概率为：

$$P(z) = \frac{1}{Z} e^{-\frac{E(z)}{k_B T}}$$

为了简化，我们设定：温度 $T = 1$ 玻尔兹曼常数 $k_B = 1$ 现在，我们进行如下映射：统计学 / 机器学习 (GMM) 统计力学 (Thermodynamics) 隐变量 $z$ 微观状态 (Microstate) 观测数据 $x$ 外部环境场 (Fixed Environment) 联合概率的负对数 $-\log p_\theta(x,z)$ 能量 $E(z)$ (Energy function) 边缘似然 $p_\theta(x) = \sum_z p(x,z)$ 配分函数 $Z$ (Partition Function) 负对数似然 $-\log p_\theta(x)$ 自由能 $F$ (Helmholtz Free Energy)2. 公式推导：从似然到自由能让我们把你给出的公式翻译成物理语言。你给出的公式是：

$$\log p_\theta(x) = \sum_z p_\theta(z|x) \log p_\theta(x,z) - \sum_z p_\theta(z|x) \log p_\theta(z|x)$$

我们对等式两边取负号（因为物理中习惯通过最小化能量来达到稳定，而机器学习习惯最大化似然）：

$$-\log p_\theta(x) = \sum_z p_\theta(z|x) \underbrace{(-\log p_\theta(x,z))}_{E(z)} - \underbrace{\left( -\sum_z p_\theta(z|x) \log p_\theta(z|x) \right)}_{\text{Entropy } S}$$

代入物理映射：左边：$-\log p_\theta(x)$ 对应配分函数的负对数，即自由能 $F$。右边第一项：$\sum p(z|x)E(z)$ 是能量的期望值，即内能 $U$ (Internal Energy)。右边第二项：$-\sum p \log p$ 正是香农熵，对应物理中的熵 $S$ (Entropy)。于是，你的公式完美复现了热力学第二定律中最著名的公式之一：

$$F = U - TS$$

(注：前面设定了 $T = 1$)3. 物理直觉：内能与熵的博弈这个公式 $F = U - S$ 揭示了系统演化的两个驱动力，它们是相互竞争的：(1) 内能 $U$ (Energy term) 对应项：$\mathbb{E}_{z|x}[-\log p(x,z)]$ 物理倾向：自然界喜欢低能量状态。ML 含义：模型希望找到最“准”的解释。即找到那个 $z$（比如属于哪个高斯分量），使得 $p(x,z)$ 最大（能量最低）。这推动分布 $p(z|x)$ 集中到最可能的那个 $z$ 上（也就是变成一个 delta 函数）。(2) 熵 $S$ (Entropy term) 对应项：$H[p(z|x)]$ 物理倾向：自然界喜欢高熵（混乱）状态。ML 含义：模型不希望过早下定论，希望保留不确定性。这推动分布 $p(z|x)$ 尽可能平摊（Uniform），不要把鸡蛋都放在一个篮子里。(3) 平衡点 (Equilibrium)EM 算法中的 E 步算出的真实后验概率 $p_\theta(z|x)$，正是这两个力博弈后的平衡态。如果我们只最小化内能 $U$，我们会把所有概率都给那个似然最大的 $z$（过拟合，Hard Assignment）。如果我们只最大化熵 $S$，我们会认为 $x$ 属于任何 $z$ 的概率都一样（完全瞎猜）。自由能最小化（即 $F$）找到了最佳折中：在解释数据的同时，保留合理的混乱度。4. 为什么叫“自由”能？在物理学中，Free Energy 代表了一个系统在恒温下 **“可用于做功的能量”**。在 EM 算法的语境下，我们可以这样理解：我们的目标是提升 $\log p(x)$（降低

$F$）。但是 $\log p(x)$ 很难直接算（因为要积分/求和）。于是我们引入了一个“变分分布”$q(z)$。变分推断中的 ELBO (Evidence Lower Bound) 其实就是负的变分自由能 (Variational Free Energy)。

$$\mathcal{F}_{var}(q) = \mathbb{E}_q[E(z)] - H(q) \geq F_{true}$$

这里的物理含义是：任何非平衡态（$q \neq p_{posterior}$）的自由能，都总是大于真实平衡态的自由能。EM 算法的过程，就是让我们的模型系统从一个非平衡态，自然演化（弛豫）到自由能最低的平衡态的过程：E 步：固定环境参数 $\theta$，让微观状态分布 $q(z)$ 演化，直到它达到热力学平衡（$q(z) = p(z|x)$），此时变分自由能降到最低，等于真实自由能。M 步：改变环境参数 $\theta$（比如改变容器形状），使得在这个新环境下，系统的整体自由能进一步降低。总结你给出的公式不仅是数学恒等式，它在物理上描述了一个处于热力学平衡态的系统：

$$\underbrace{\text{Free Energy}}_{\text{Log-Likelihood}} = \underbrace{\text{Internal Energy}}_{\text{Fit to Data}} - \underbrace{\text{Entropy}}_{\text{Uncertainty}}$$

这就是为什么 EM 算法（以及变分自编码器 VAE、扩散模型 Diffusion Models）如此有效——它们不仅仅是优化技巧，它们顺应了自然界最基本的演化法则：最小化自由能。

哲学是, 复杂的现实世界 $p(\vec{x})$, 其实是无数个简单的可能性 $p(\vec{x}|z)$ 在隐变量空间上以 $p(z)$ 为权重的加权混合. 可以考虑联合分布

$$p(x,z) = p(x|z) \cdot p(z), \quad p(x, z=k) = \pi_k \cdot \mathcal{N}(x; \mu_k, \Sigma_k).$$

# Chapter 4

# 特征函数与极限定理

## 1 再谈期望

- 一方面，随机变量 $X$ 的期望 $E[X]$ 就是可测函数 $X$ 关于测度 $P$ 的积分 $\int_\Omega X\mathrm{d}P$.
- 另一方面，给定随机变量 $X$，设其分布函数为 $F$. $X$ 将测度 $P$ 推出到 $\mathbb{R}$ 上得到测度 $\mu_F$. 称 Borel 可测函数 $g:\mathbb{R}\to\mathbb{R}$ 关于 $\mu_F$ 的积分为 Lebesgue-Stieltjes 积分，记作 $\int_\mathbb{R} g\mathrm{d}F$.

**定理 1.1.**

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

# 2 特征函数

- 特征函数可将求各阶矩的积分运算化成微分运算.
- $X$ 的函数的特征函数比 $X$ 的函数的概率密度函数容易求.

**定义 2.1.** 复随机变量

- 复随机变量与实二维随机向量的本质区别在于复随机变量可以做乘法.
- 复随机变量的独立性的定义
- 复随机变量的期望的定义
- $Z_1$ 与 $Z_2$ 独立 $\Longrightarrow E[Z_1Z_2]=E[Z_1]E[Z_2]$.

**定义 2.2.** $\phi_X(t):=E[\mathrm{e}^{\mathrm{i}tX}]$

**定理 2.1.**

(1) $\phi(0)=1$

(2) $\phi$ 在 $(-\infty,+\infty)$ 上一致连续.

(3) $\phi$ 非负定. $\forall t_1,\cdots,t_n\in\mathbb{R},\ \forall z_1,\cdots,z_n\in\mathbb{C}$，成立

$$\sum_{j,k=1}^{n} z_j\bar{z}_k\phi(t_j-t_k)\geqslant 0.$$

证明.

(2)

$$|\phi(t+h)-\phi(t)|=\left|\int \mathrm{e}^{\mathrm{i}tx}(\mathrm{e}^{\mathrm{i}hx}-1)\mathrm{d}F\right|\leqslant\int|\mathrm{e}^{\mathrm{i}hx-1}|\mathrm{d}F\xrightarrow{DCT}0,\quad h\to 0.$$

□

**定理 2.2.** 当 $X$ 与 $Y$ 独立时，

$$\phi_{X+Y}(t)=\phi_X(t)\phi_Y(t).$$

证明.

$$\phi_{X+Y}(t)=E[\mathrm{e}^{\mathrm{i}t(X+Y)}]=E[\mathrm{e}^{\mathrm{i}tX}\mathrm{e}^{\mathrm{i}tY}]=E[\mathrm{e}^{\mathrm{i}tX}]E[\mathrm{e}^{\mathrm{i}tY}]=\phi_X(t)\phi_Y(t).$$

□

当 $X$ 与 $Y$ 独立时，$X$ 与 $Y$ 的加法对应于概率密度函数的卷积，而卷积的傅里叶变换对应于傅里叶变换的乘法.

5.7.1 给出两个不独立的随机变量 $X,Y$ 满足 $\phi_{X+Y}(t)=\phi_X(t)\phi_Y(t),\forall\, t\in\mathbb{R}$.

证明. □

-

**定理 2.3.** 设 $(X,\Sigma,\mu)$ 是一个测度空间，$G$ 是 $\mathbb{C}$ 中的开集，$f:X\times G\to\mathbb{C}$. 若

(1) 对任意取定的 $x\in X$，$f(x,z)$ 是关于 $z$ 的全纯函数.

(2) 对任意取定的 $z\in G$，$F(z):=\int_X f(x,z)\mathrm{d}\mu$ 存在.

(3) 存在 $X$ 上的非负可积函数 $g$，使得

$$\left|\frac{\partial f}{\partial z}(x,z)\right|\leqslant g(x),\quad \forall x\in X,z\in G.$$

那么 $F$ 全纯且 $F'(z)=\int_X \frac{\partial f}{\partial z}(x,z)\mathrm{d}\mu$.

证明. □

# 3 极限定理

**定理 3.1.** 设 $\{X_n\}$ 是独立同分布的随机变量序列，且 $\mu = E[X_1]$ 存在，令 $S_n = \sum_{k=1}^{n} X_k$，则

$$\frac{1}{n} S_n \xrightarrow{D} \mu.$$

证明. 由 Levy-Cramer 连续性定理，只需证明 $\phi_n(t) = E[\mathrm{e}^{\mathrm{i}t\frac{S_n}{n}}]$ 逐点收敛到 $\mathrm{e}^{\mathrm{i}\mu t}$.

记 $\phi(t) = E[\mathrm{e}^{\mathrm{i}tX_1}]$，则 $\phi_n(t) = (\phi(\frac{t}{n}))^n$.

又 $\phi(\frac{t}{n}) = 1 + \frac{\mathrm{i}\mu t}{n} + o(\frac{t}{n})$，所以

$$\lim_{n\to+\infty} \phi_n(t) = \left(1 + \frac{\mathrm{i}\mu t}{n} + o\left(\frac{t}{n}\right)\right)^n = \mathrm{e}^{\mathrm{i}\mu t}.$$

□

# Chapter 5

# 几种收敛

## 1 四种收敛

**定义 1.1.** 设 $X, X_n$ 为 $(\Omega, F, P)$ 上的随机变量，

(1)

(2)

(3)

设 $X, X_n$ 为随机变量，不必定义在同一测度空间，

(4) 对 $F_X(x) = P(X \leqslant x)$ 的连续点处，有 $F_{X_n}(x) \to F_X(x)$. 记作 $X_n \xrightarrow{D} X$.

**命题 1.1.** $X_n \xrightarrow{P} X \Longrightarrow X_n \xrightarrow{D} X$.

证明. □

**定理 1.1.**

(1) 若 $X_n \xrightarrow{D} c$，则 $X_n \xrightarrow{P} c$，其中 $c \in \mathbb{R}$ 是常数.

(2)

(3)

证明.

(1) $P(|X_n - c| > \varepsilon) = P(X_n > c + \varepsilon) + P(X_n < c - \varepsilon) = 1 - P(X_n \leqslant c + \varepsilon) + P(X_n < c - \varepsilon)$

(2)

(3)

□

# 2 结论拾零

## 2.1 Markov 不等式

## 2.2 Borel-Cantelli 引理

**例 2.1.** 设 $\{X_n\}$ 相互独立，服从参数为 1 的指数分布，试证

$$P(\limsup_{n\to+\infty}\frac{X_n}{\log n}=1)=1.$$

证明. $P(\frac{X_n}{\log n}\geqslant 1+\varepsilon)=\frac{1}{n^{\alpha+1}}$ □

**例 2.2.** 设 $\{X_n\}$ 相互独立，服从标准正态分布 $N(0,1)$，试证

$$P(\limsup_{n\to+\infty}\frac{|X_n|}{\sqrt{\log n}}=\sqrt{2})=1.$$

证明.

$$P(\limsup_{n\to+\infty}\frac{|X_n|}{\sqrt{\log n}}\geqslant\sqrt{2}(1+\varepsilon))=0,\quad \varepsilon>0$$

$$P(\limsup_{n\to+\infty}\frac{|X_n|}{\sqrt{\log n}}\geqslant\sqrt{2}(1+\varepsilon))=1,\quad \varepsilon\leqslant 0$$

$$P(\limsup_{n\to+\infty}\frac{|X_n|}{\sqrt{\log n}}=\sqrt{2})=1$$

□

# 3 强大数律

**定理 3.1.** 设 $\{X_k\}$ 独立同分布，$E[X_1^2] < +\infty$，$E[X_1] = \mu$，则

(1) $\frac{1}{n}\sum_{n=1}^{\infty} X_i \xrightarrow{2} \mu$

(2) $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$

太长不看版：样本均值收敛于期望值.

# 4 习题

7.11.7 Show that $X_n \xrightarrow{a.e.} X$ whenever $\sum_n E(|X_n - X|^r) < \infty$.

# Chapter 6

# 深度学习

## 1 KL 散度和极大似然估计

假设真实数据服从的分布为 $p_{\text{data}}$, 而我们希望通过一个参数化分布 $p_\theta$ 来拟合 $p_{\text{data}}$（大家心中要有一些简单的小例子，比如混合高斯模型，$p_\theta$ 一般都是要有隐变量的，$p_\theta(x|z)p(z)$ 其中 $p(z)$ 是 $z$ 的先验分布,$p_\theta(x|z)$ 一般比较简单，可以是高斯分布，或者是一个神经网络）. 衡量两个分布的相似程度的数学概念是 KL 散度

$$D_{\text{KL}}(p_{\text{data}} \parallel p_\theta) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(x)}{p_\theta(x)}\right] = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_{\text{data}}(x)] - \mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)]$$

$$\arg\min_\theta D_{\text{KL}}(p_{\text{data}} \parallel p_\theta) \iff \arg\max_\theta \mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)] \approx \frac{1}{m}\sum_{i=1}^{m} \log p_\theta(x_i)$$

现在假设我们只采样了一个数据点, 也就是要极大化 $\log p_\theta(x_1)$。但直接极大化 $\log p_\theta(x_1)$ 并不容易，为此，我们要引入一些统计物理的视角

# Chapter 7

# 外篇

## 1 信息熵

## 2 Lindeberg 替換术

# Chapter 8

# 我的一些观察

## 1

如果 $f(x)$ 是概率密度函数，那么 $f_\lambda(x) := \lambda f(\lambda x)$ 也是概率密度函数，对于任意的 $\lambda > 0$.

- 期望

$$E[X_\lambda] = \int x f_\lambda(x) \mathrm{d}x = \frac{1}{\lambda} \int \lambda x f(\lambda x) \mathrm{d}\lambda x = \frac{1}{\lambda} E[X]$$

## 2

### 2.1

想证明随机变量 $A$ 是随机变量 $B$，首先你要知道 $B$ 的分布/概率密度函数，然后直接验证 $A$ 的分布/概率密度函数就是 $A$ 的分布/概率密度函数.

**例 2.1.** 设 $X_1, X_2, \cdots, X_n$ 是独立的标准正态分布，证明：

(1) $X_1^2$ 服从 $\chi^2(1)$.

(2)

(3)

证明. □

# 3 独立性

**定义 3.1.** 设 $X_1, \cdots, X_n$ 为 $(\Omega, F, P)$ 上的随机变量，称它们是独立的，如果对任意 $x_1, \cdots, x_n \in \mathbb{R}$ 有

$$P(X_1 \leqslant x_1, \cdots, X_n \leqslant x_n) = P(X_1 \leqslant x_1) \cdots P(X_n \leqslant x_n).$$

**命题 3.1.** $X_1, \cdots, X_n$ 相互独立，如果对任意 $B_1, \cdots, B_n \in B(\mathbb{R})$，有

$$P(X_1 \in B_1, \cdots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

# 4 用到 Markov 不等式的习题

7.11.10 Show that $X_n \xrightarrow{P} 0$ iff

$$\lim_{n\to+\infty} E\left(\frac{|X_n|}{1+|X_n|}\right) = 0$$

证明. 注意到 $g(u) = \frac{u}{1+u}$ 是 $[0,\infty)$ 上的单调递增函数. 因此，对任意 $\varepsilon > 0$，由 Markov 不等式，

$$P(|X_n| > \varepsilon) = P\left(\frac{|X_n|}{1+|X_n|} > \frac{\varepsilon}{1+\varepsilon}\right) \leqslant \frac{1+\varepsilon}{\varepsilon} \cdot E\left(\frac{|X_n|}{1+|X_n|}\right).$$

□

## 5

4.1.3

## 6 最大值与最小值

4.5.4

# Chapter 9

# 习题

## 1 依赖性习题

3.6.8 设 $X$ 和 $Y$ 有联合质量分布函数

$$f(j,k)=\frac{c(j+k)a^{j+k}}{j!k!},\quad j,k\geqslant 0,$$

其中 $a$ 是一个常数，求 $c, P(X=j), P(X+Y=r)$ 和 $E[X]$.

证明.

$$\begin{aligned}
\sum_{j,k\geqslant 0}\frac{(j+k)a^{j+k}}{j!k!} &= \sum_{n=0}^{\infty}\sum_{j=0}^{n}\frac{na^n}{j!(n-j)!}\\
&= \sum_{n=0}^{\infty}\frac{na^n}{n!}\sum_{j=0}^{n}\frac{n!}{j!(n-j)!}\\
&= \sum_{n=0}^{\infty}\frac{n(2a)^n}{n!}\\
&= 2a\sum_{n=0}^{\infty}\frac{(2a)^n}{n!}\\
&= 2a\mathrm{e}^{2a}\\
P(X=j) &= \frac{1}{2a\mathrm{e}^a}\frac{a^j}{j!}(a+j)\\
E[X] &= \sum_{j=1}^{\infty}\frac{1}{2a\mathrm{e}^a}\frac{a^j}{j!}(a+j)j\\
&= \frac{1}{2a\mathrm{e}^a}\left[a^2\sum_{j=1}^{+\infty}\frac{a^{j-1}}{(j-1)!}+a\sum_{j=1}^{+\infty}\frac{a^{j-1}}{(j-1)!}j\right]\\
x\mathrm{e}^x &= \sum_{n=0}^{\infty}\frac{x^{n+1}}{n!}\\
\mathrm{e}^x+x\mathrm{e}^x &= \sum_{n=0}^{\infty}\frac{x^n}{n!}(n+1)\\
E[X] &= a+\frac{1}{2}
\end{aligned}$$

□

## 2 条件期望习题

3.11.4

证明. 按定义计算 $f_{Y|Z}$ 和 $f_{Z|Y}$，帮助你熟悉概念的基础得不能再基础的题. □

3.7.5 设机器的寿命（单位：天）是以 $f$ 为质量函数的随机变量 $T$. 假设机器已经工作了 $t$ 天，问机器的剩余寿命的期望是：

(1) $f(x) = (N+1)^{-1}, x \in \{0, 1, \cdots, N\}$

(2) $f(x) = 2^{-x} x = 1, 2, \cdots$

证明. 本题是对事件求条件期望，所以仅仅是权重重新分配的加权平均. □

3.7.6 设 $X_1, X_2, \cdots$ 独立同分布，期望是 $\mu$，设 $N$ 是取非负整数值的随机变量，与 $X_i$ 独立. 令 $S = X_1 + X_2 + \cdots + X_N$. 证明 $E(S|N) = \mu N$.

证明. 我想这背后应该有一些东西，但目前按照朴素的想法来理解还是很容易的. □

# 3 随机游走习题

3.9.1 设 $T_k$ 是粒子从 $S_0 = k$ 出发，到最终被 0 和 $N$ 处的吸收壁吸收的概率，其中 $0 \leqslant k \leqslant N$. 如 $T_0 = T_N = 0$. 证明

(1) $P(T_k < \infty) = 1$ 对任意 $0 \leqslant k \leqslant N$ 成立.

(2) $E[T_k^m] < \infty$ 对任意 $0 \leqslant k \leqslant N$ 和 $m \geqslant 1$ 成立.

设 $S_n$ 为简单对称随机游走，$S_0 = 0$，记 $\tau_k = \inf\{n \geqslant 1 : S_n = k\}$，