应用随机分析 [1]–读书笔记

王梦收

2025 年 7 月 17 日

1 Part1 基础知识

本节将介绍概率论的基本概念和工具，为后续的随机分析内容打下基础。

1.1 基本例子

首先引入一个公平投掷硬币的例子，从两个方面考虑，会得到两个看似矛盾的结果。记

$$X_j = \begin{cases} 1, & \text{如果第 } j \text{ 次试验结果为正面（head）} \\ 0, & \text{如果第 } j \text{ 次试验结果为反面（tail）} \end{cases}$$

一次试验的结果，则 $S_n = X_1 + \cdots + X_n$ 为 n 次试验中正面出现的次数。

结果一：单点概率趋于零

对于 $2n$ 次独立抛硬币实验，S_{2n} 表示正面出现的次数。则恰好有 n 次正面的概率为：

$$\text{Prob}(S_{2n} = n) \to 0, \quad (n \to \infty)$$

结果二：大数定律收敛

对于任意 $\epsilon > 0$，有

$$\text{Prob}\left(\left|\frac{S_{2n}}{2n} - \frac{1}{2}\right| > \epsilon\right) \to 0, \quad (n \to \infty)$$

这说明虽然单点概率趋于零，但频率依然收敛到 $1/2$。

两个结果的证明主要是用到了 *Stirling* 公式 $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$

连续型的例子 考虑圆球面上的单位向量$\tau \in \mathbb{S}^2 \subset \mathbb{R}^3$, 定义$\rho(\mathbf{n}), \mathbf{n} \in \mathbb{S}^2$为取向分布密度，也即

$$\text{Prob}(\tau \in A) = \int_A \rho(\mathbf{n}) dS,$$

如果τ没有优先取向，则$\rho(\mathbf{n}) = \frac{1}{4\pi}$, 称$\tau$为各向同性；如果$\tau$优先取向$\mathbf{n_0}$，则$\rho(\mathbf{n})$在某方向$\mathbf{n_0}$达到峰值。

1.2 概率空间

为了严格表述概率论，这里给出Kolmogorov引入的概论空间的相关定义，记作三元组$(\Omega, \mathcal{F}, \mathbb{P})$。

定义 1.1. *(样本空间)*样本空间Ω是所有可能结果的集合，其元素$\omega \in \Omega$被称为样本点。

定义 1.2. *($\sigma-$代数)* 一个 $\sigma-$代数 *($\sigma-$域)* $\mathcal{F}$ 是 Ω的子集的集合，满足一下条件：

- $\Omega \in \mathcal{F}$ 。
- 如果 $A \in \mathcal{F}$,那么 $A^c \in \mathcal{F}$. $(A^c = \Omega\ A)$
- 如果 $A_1, A_2, \cdots \in \mathcal{F}$, 那么$\cup_{n=1}^{\infty} A_n \in \mathcal{F}$。

每一个集合 $A \in \mathcal{F}$ 被称为事件。Ω的子集的集合$\mathcal{B}$,定义$\sigma(\mathcal{B})$为包含$\mathcal{B}$的最小$\sigma-$`algebra`，也称由$\mathcal{B}$生成的最小$\sigma-$`algebra`。满足上面条件的二元组$(\Omega, \mathcal{F})$称为可测空间。

定义 1.3. *(概率空间)* 定义集函数 $\mathbb{P} : \mathcal{F} \to [0,1]$定义在$\mathcal{F}$上满足，

- $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$。
- 称$A_1, A_2, \cdots \in \mathcal{F}$是*pairwise disjoint*，意味着对于$i \neq j, A_i \cap A_j = \emptyset$，那么有可数可加性（$\sigma-$可加性），

$$\mathbb{P}\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Boole's 不等式 $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$。

1.3 条件概率

定义 1.4. 对于$A, B \in \mathcal{F}, \mathbb{P}(B) \neq 0$,定义给定$B$,$A$的条件概率为

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

定理 1.1. *(Bayes's 定理)* 如果$A_1, A_2, \cdots$是*disjoint*集合使得$\cup_{j=1}^{\infty} A_j = \Omega$,那么有

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{n=1}^{\infty} \mathbb{P}(A_n)\mathbb{P}(B|A_n)}$$

1.4 离散分布和连续分布

对于离散分布考虑的是样本空间是有限或可数的情况，即$\Omega = \{\omega_1, \omega_2, \cdots\}$, 定义

$$X(w_j) = x_j, \quad p_j = \mathbb{P}(X = w_j), j = 1, 2, \cdots.$$

对于给定随机变量X的函数f，定义其期望为

$$\mathbb{E}f(X) = \sum_j f(x_j)p_j, \quad m_p = \sum_j x_j^p p_j,$$

且$p = 1$时为均值$\mathtt{mean}(X)$，方差定义为$\mathtt{Var}(X) = m_2 - m_1^2$。

考虑Ω不可数时，我们在下面引入随机变量。首先定义$\mathbb{R}$上包含所有开集的最小$\sigma-$代数为$\mathcal{R}$,也称为$\mathbb{R}$上的Borel $\sigma-$代数。

定义 1.5. *(随机变量)* 随机变量X是$\mathcal{F}$上可测的实值函数 $X : \mathcal{F} \to \mathbb{R}$,也即对任意的$B \in \mathcal{R}, X^{-1}(B) \in \mathcal{F}$。

定义 1.6. *(分布)* 一个随机变量的分布是定义在$\mathbb{R}$上的概率测度μ，满足对任意的集合$B \in \mathcal{R}$有：

$$\mu(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)).$$

定义 1.7. *(分布函数)* 当$B = (-\infty, x]$时,定义$F(x) = \mathbb{P}(x \leq B)$。

定义 1.8. *(概率密度函数)* 如果存在可积函数$\rho(x)$,使得对于任意的$B \in \mathcal{R}$,有

$$\mu(B) = \int_B \rho(x)dx,$$

则称ρ为X的概率密度函数。

对于概率密度函数 $\rho(x)$，如果测度 $\mu(dx)$ 关于 Lebesgue 测度 $m(dx)$ 是绝对连续的（即 $m(B) = 0$ 时 $\mu(B) = 0$），则 $\rho(x) = \frac{d\mu}{dm}$ 被称为 $\mu(dx)$ 关于 $m(dx)$ 的 Radon-Nikodym 导数。

定义 1.9. *(期望)* 如果下面积分是良好定义的，对于随机变量 X的期望为

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}} x\rho(x)dx.$$

(方差) 定义方差为

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

(协方差) 对于两个随机变量 X 和 Y，定义协方差为

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

(协方差矩阵) 对于随机向量 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$，定义协方差矩阵为

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}(X - \mathbb{E}\mathbf{X})(X - \mathbb{E}\mathbf{X})^T.$$

定义 1.10. *(p-阶矩空间)* 设 $1 \le p < \infty$，空间$L^p(\Omega)$*(*或L^p_ω空间*)*是所有满足 $\mathbb{E}[|X|^p] < \infty$ 的随机变量 X 的集合，即

$$L^p(\Omega) = \{X : \mathbb{E}[|X|^p] < \infty\}.$$

对 $X \in L^p(\Omega)$，让

$$\|X\|_p = \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

定理 1.2. *(基本不等式)* 对于 $L^p(\Omega)$ 空间中的随机变量，满足以下不等式:

- *Minkowski* 不等式:

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p, \quad \forall X, Y \in L^p(\Omega).$$

- *Hölder* 不等式:

$$\mathbb{E}|(X, Y)| \le \|X\|_p \|Y\|_q, \quad \forall X \in L^p(\Omega), Y \in L^q(\Omega), \frac{1}{p} + \frac{1}{q} = 1, p > 1,$$

  其中 $\mathbb{E}|(X, Y)|$是$\mathbb{R}$上的标准内积。

- *Schwarz* 不等式:

$$\mathbb{E}|(X, Y)| \le \|X\|_2 \|Y\|_2, \quad \forall X, Y \in L^2(\Omega).$$

可以证明，$L^p(\Omega)$ 空间是一个 Banach 空间，即它是一个完备的范数空间，并且$L^2(\Omega)$ 是一个 Hilbert 空间,有内积

$$(X, Y)_{L^2_\omega} = \mathbb{E}[(X, Y)], \quad \forall X, Y \in L^2(\Omega).$$

引理 1.1. *(Chebyshev* 不等式*)* 对于任意随机变量 X 使得$\mathbb{E}|X|^p < \infty, p > 0$ 和任意 $\lambda > 0$，有

$$\mathbb{P}(|X| \ge \lambda) \le \frac{1}{\lambda^p}\mathbb{E}[|X|^p].$$

上面直接得到对非负递增函数 f 的切比雪夫不等式：

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{\mathbb{E}[|f(X)|]}{f(\lambda)}.$$

引理 1.2. *(Jensen 不等式)* 让随机变量 X有$\mathbb{E}[|X|] < \infty$，如果 $\phi : \mathbb{R} \to \mathbb{R}$ 是一个凸函数使得$\mathbb{E}|\phi(X)| < \infty$，则有

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

对于具体的连续随机变量的例子这里省略。特别提醒对于高维的高斯分布通过协方差矩阵定义时，协方差矩阵可能不可逆，因此之后可以通过特征函数方法定义。以及关于平衡态统计物理的Gibbs分布。

1.5 独立性

两个事件 $A, B \in \mathcal{F}$ 是独立的，

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

定义 1.11. *(独立性)* 设 X 和 Y 是两个随机变量，如果对任意的两个*Borel*集合 A 和 B， $X^{-1}(A)$和$Y^{-1}(B)$是独立的，即

$$\mathbb{P}(X^{-1}(A) \cap Y^{-1}(B)) = \mathbb{P}(X^{-1}(A))\mathbb{P}(Y^{-1}(B)).$$

在独立性的条件下，随机变量的联合分布，概率密度函数，期望等都有简单的形式。

1.6 条件期望

令$\mathcal{G}$是$\mathcal{F}$的子$\sigma-$代数, X是满足$\mathbb{E}[|X|] < \infty$的随机变量。

定义 1.12. *(条件期望)* 对于给定$\mathcal{G}$，X的条件期望Z满足以下条件

- Z是$\mathcal{G}$-可测的，即Z是$\mathcal{G}$上的随机变量。
- 对于任意的$A \in \mathcal{G}$，有
$$\int_A Z(\omega)\mathbb{P}(d\omega) = \int_A X(\omega)\mathbb{P}(d\omega).$$

对于上面Z的存在性和唯一性，Radon-Nikodym 定理可以保证。具体而言，定义$\mathcal{G}$的测度μ,有$\mu(A) = \int_A X(\omega)\mathbb{P}(d\omega)$, 则$\mu$关于$\mathbb{P}$限制在$\mathcal{G}$上的测度$\mathbb{P}|_{\mathcal{G}}$是绝对连续的，因此$Z$存在且唯一。

下面给出一些条件期望的性质：

定理 1.3. 假设X, Y是满足$\mathbb{E}[|X|] < \infty, \mathbb{E}[|Y|] < \infty$的随机变量。让$a, b$是常数，则有

- $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$.
- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
- 如果X是$\mathcal{G}$-可测的，则有$\mathbb{E}[X|\mathcal{G}] = X$。
- 如果X和Y是独立的，则有$\mathbb{E}[X|Y] = \mathbb{E}[X]$。
- 如果Y是$\mathcal{G}$-可测的，则有$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$。
- 如果$\mathcal{H}$是$\mathcal{G}$的子$\sigma-$代数，则有$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$.

引理 1.3. *(条件Jensen 不等式)* 让随机变量 X有$\mathbb{E}[|X|] < \infty$，如果 $\phi : \mathbb{R} \to \mathbb{R}$ 是一个凸函数使得$\mathbb{E}|\phi(X)| < \infty$，则有

$$\mathbb{E}[\phi(X)|\mathcal{G}] \geq \phi(\mathbb{E}[X|\mathcal{G}]).$$

上面是对于$\sigma-$代数取条件，对于随机变量取条件，有一下定义。给定两个随机变量 X 和 Y，则X 关于 Y 的条件期望定义为X关于Y生成的最小$\sigma-$代数$\mathcal{G} = \sigma(Y)$,

$$\mathcal{G} = \{Y^{-1}(B) : B \in \mathcal{R}\},$$

的条件期望. 并且通过此定义，该条件期望与经典的条件概率结果一致。

条件期望是$\mathcal{G}$中所有可测函数对$L^2(\Omega)$中可测函数的最优近似。即对于任意的$X \in L^2(\Omega)$，有

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

1.7 收敛性

下面考虑四种收敛性：几乎处处收敛、依概率收敛、依分布收敛和L^p收敛。

定义 1.13. *(几乎处处收敛)* 随机变量序列 X_n 几乎处处收敛到 X，如果

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

我们写符号$X_n \xrightarrow{a.s.} X$。

定义 1.14. *(依概率收敛)* 随机变量序列 X_n 依概率收敛到 X，如果对于任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} \mathbb{P}(\omega : |X_n(\omega) - X(\omega)| \geq \epsilon) = 0.$$

我们写符号$X_n \xrightarrow{p} X$。

定义 1.15. *(依分布收敛)* 随机变量序列 X_n 以分布收敛到 X，如果对于任意的$f \in C_b(\mathbb{R})$（即连续且有界的函数），有

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

我们写符号$X_n \xrightarrow{d} X$或者$\mu_n \xrightarrow{d} \mu$或者 $\mu_n \Rightarrow \mu$。

定义 1.16. *(L^p收敛)* 随机变量序列 X_n 在 L^p 意义下收敛到 X，如果

$$\lim_{n\to\infty} \mathbb{E}\|X_n - X\|^p = 0.$$

当$p = 1$时，L^1 收敛也称为均值收敛；当$p = 2$时，L^2 收敛也称为均方收敛。

对于上面四种收敛性，有以下关系：

定理 1.4. *(收敛性关系)*

- 几乎处处收敛 $\Rightarrow$ 依概率收敛 $\Rightarrow$ 存在一个子序列几乎处处收敛。
- 如果$p < q$, 则 L^p 收敛 $\Rightarrow L^q$ 收敛。
- L^p 收敛 $\Rightarrow$ 依概率收敛 $\Rightarrow$ 依分布收敛。

1.8 特征函数

随机变量X的特征函数定义为

$$f(\xi) = \mathbb{E}[e^{i\xi X}] = \int_{\mathbb{R}} e^{i\xi X}\mu(dx).$$

命题 1.1. *(特征函数的性质)*

- $\forall \xi \in \mathbb{R}, |f(\xi)| \leq 1, f(\xi) = \overline{f(-\xi)}, f(0) = 1;$
- f 在$\mathbb{R}$ 上一致连续。

定理 1.5. *(Levy 连续性定理)* 令$\{\mu_n\}$是一个概率测度序列，且μ_n的特征函数为$f_n(\xi)$。假设

- f_n 在$\mathbb{R}$上处处收敛到极限函数f。
- f 在$\xi = 0$处连续。

那么存在一个概率分布μ，使得$\mu_n \xrightarrow{d} \mu$。而且$f$是$\mu$的特征函数.

反过来，如果$\mu_n \xrightarrow{d} \mu$,其中$\mu$是某个概率分布，那么$f_n$在每个有限区间上一致收敛到$f$，其中$f$是$\mu$的特征函数。

类似Fourier变换，我们也能定义特征函数的逆变换

$$\rho(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x} f(\xi) d\xi.$$

我们感兴趣的是对于f满足什么条件时，ρ是一个概率密度函数。

定义 1.17. 一个函数f是半正定的，如果对于任意的有限集合 $\{\xi_1, \xi_2, \ldots, \xi_n\}, n \in \mathbb{N}$, 矩阵$(f(\xi_i - \xi_j))_{i,j=1}^n$ 是半正定的，即对于任意的复数 $c_1, c_2, \ldots, c_n$，有

$$\sum_{i,j=1}^{n} c_i \overline{c_j} f(\xi_i - \xi_j) \geq 0.$$

定理 1.6. *(Bochner 定理)* 设 f 是一个概率测度的特征函数当且仅当 f 是一个半正定函数且在*0*处连续有值$f(0) = 1$。

1.9 生成函数和矩母函数

对于离散随机变量，它的概率生成函数定义为

$$G(z) = \sum_{k=0}^{\infty} P(X = x_k) x^k,$$

那么立即有

$$P(X = x_k) = \frac{1}{k!} \left.\frac{d^k G(z)}{dz^k}\right|_{z=0}.$$

定义 1.18. *(卷积)* 定义序列$\{a_k\}$和$\{b_k\}$的卷积为$\{c_k\} = \{a_k\} * \{b_k\}$,

$$c_k = \sum_{j=0}^{k} a_j b_{k-j}.$$

对于生成函数$\{a_k\},\{b_k\}$和$\{c_k\}$的生成函数

$$A(x)=\sum_{k=0}^{\infty}a_kx^k,B(x)=\sum_{k=0}^{\infty}b_kx^k,C(x)=\sum_{k=0}^{\infty}c_kx^k,$$

则有

$$C(x)=A(x)B(x).$$

定理 1.7. 令X和Y是两个独立的离散随机变量对应的分布分别为

$$P(X=j)=a_j,\quad P(Y=k)=b_k,$$

并且$A(x)$和$B(x)$是相关的生成函数,那么有$X+Y$的生成函数为$C(x)=A(x)B(x)$.

下面定义随机变量X的矩母函数

$$M(t)=\mathbb{E}[e^{tX}]=\begin{cases}\sum_{k=0}^{\infty}P(X=k)e^{tk}, & \text{如果 } X \text{ 是离散型随机变量}\\ \int_{\mathbb{R}}\rho(x)e^{tx}dx, & \text{如果 } X \text{ 是连续型随机变量,}\end{cases}$$

假设e^{tX}是可积的，且显然有 $M(0)=\mathbb{E}[1]=1$.

一旦$M(t)$存在，则$M(t)$在它的定义域内有$M(t)\in C^{\infty}$，且有

$$M^{(n)}(t)=\mathbb{E}(X^ne^{tX}),\quad \text{并且}\quad \mu_n\coloneqq\mathbb{E}X^n=M^{(n)}(0),n\in\mathbb{N}.$$

因此

$$M(t)=\sum_{n=0}^{\infty}\frac{\mu_n}{n!}t^n,$$

这也$M(t)$被称为矩母函数的原因。

定理 1.8. 定义$M_X(t),M_Y(t),M_{X+Y}(t)$分别为随机变量$X,Y,X+Y$的矩母函数，如果X,Y是独立的，则有

$$M_{X+Y}(t)=M_X(t)M_Y(t).$$

定义累积量生成函数$\Lambda(t)$为

$$\Lambda(t)=\log M(t)=\log\mathbb{E}e^{tX}=\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!},$$

根据定义，我们有$\kappa_0=0$并且通过此定义

$$\kappa_n=\frac{d^n}{dt^n}\Lambda(0),n\in\mathbb{N}.$$

对于随机向量的矩母函数和累积量生成函数，定义为

$$M(\mathbf{t})=\mathbb{E}e^{\mathbf{t}\cdot\mathbf{X}},\Lambda(\mathbf{t})=\log M(\mathbf{t}),\mathbf{t}\in\mathbb{R}^d,$$

这些记号在统计力学中相当有用。

1.10 Borel-Cantelli 引理

对于给定事件序列 $\{A_n\}_{n=1}^{\infty}$，我们有兴趣的是事件A_n发生无限次，也即

$$\{A_n \quad \text{i.o.}\} = \{\omega : \omega \in A_n \quad \text{i.o.}\} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k.$$

引理 1.4. • 如果$\sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty$,那么$\mathbb{P}(\{A_n \quad i.o.\}) = 0$.

• 如果$\sum_{n=1}^{\infty}\mathbb{P}(A_n) = \infty$,且$A_n$是独立的，那么$\mathbb{P}(\{A_n \quad i.o.\}) = 1$.

引理 1.5. 令$\{X_n\}_{n\in\mathbb{N}}$是同分布的随机变量（不用独立性），使得$\mathbb{E}[|X_n|] < \infty$，那么

$$\lim_{n\to\infty}\frac{X_n}{n} = 0 \quad a.s..$$

2 Part2 极限定理

令$\{X_j\}_{j=1}^{\infty}$是独立同分布（i.i.d.）的随机变量，且令$\eta = \mathbb{E}X_1$和$S_n = \sum_{j=1}^{n} X_j$。我们感兴趣下面三个类型的极限定理：

- 大数法则（Law of Large Numbers, LLN）：当$n \to \infty$时，样本均值$\frac{S_n}{n}$收敛到真实的平均η。
- 中心极限定理（Central Limit Theorem, CLT）：样本均值与真实均值的误差经过适当缩放后收敛到标准正态分布。
- 大偏差原理（Large Deviation Principle, LDP）：此估计了样本均值趋近于真实均值之外的某个值的概率。

在此章节的最后也将讨论一系列随机变量的最大值或最小值的分布。

2.1 大数法则

定理 2.1. *(弱大数法则)*让$\{X_j\}_{j=1}^{\infty}$是一系列*i.i.d.*的随机变量，当$\mathbb{E}|X_j| < \infty$时，那么

$$\frac{S_n}{n} \xrightarrow{p} \eta.$$

定理 2.2. *(强大数法则)*让$\{X_j\}_{j=1}^{\infty}$是一系列*i.i.d.*的随机变量，当$\mathbb{E}|X_j| < \infty$时，那么

$$\frac{S_n}{n} \xrightarrow{a.s.} \eta.$$

特别注意当$\mathbb{E}|X_j| < \infty$条件不成立时，大数法则可能不成立(Cauchy-Lorentz distribution)。

2.2 中心极限定理

定理 2.3. *(Lindeberg-Lévy 中心极限定理)* 让$\{X_j\}_{j=1}^{\infty}$是一系列*i.i.d.*的随机变量，假设$\mathbb{E}X_j^2 < \infty$且令$\sigma^2 = Var(X_j)$,那么

$$\frac{S_n - n\eta}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0,1).$$

在此有更一般的结果，如果令$G_n(x), \Phi(x)$分别为$S_n/\sqrt{n}$和 $N(0,1)$的分布函数，$\phi(x)$ 为$N(0,1)$的概率密度函数，那么有

$$G_n(x) = \Phi(x) - \phi(x)\left(\frac{\gamma(x^2-1)}{6\sqrt{n}}\right) + O\left(\frac{1}{n}\right).$$

其中$\gamma = \mathbb{E}X_j^3$.

因此中心极限定理给了大数法则收敛速率的估计，有一下结果

$$\frac{S_n}{n} - \eta \approx \frac{\sigma}{\sqrt{n}} N(0,1),$$

因此$\frac{S_n}{n}$收敛到η的速率为$\mathcal{O}(n^{-1/2})$。这也就是Monte Carlo方法的收敛速率。

对于方差无界的随机变量的中心极限定理需要用到稳定律的结果，这里不再展开。

2.3 大偏差原理

让$\{X_j\}_{j=1}^n$是 i.i.d.的随机变量序列，且令$\eta = \mathbb{E}X_1$。大数法则告诉我们对于任意的 $\epsilon > 0$，对于充分大的n，有$|S_n/n - \eta| < \epsilon$的概率趋近于1。反之，如果$y \neq \eta$，那么随着$n \to \infty$，$S_n/n$趋近于$y$的概率趋近于0。对于$y \neq \eta$，事件$\{|S_n/n - y| < \epsilon\}$被称为大偏差事件。

为了准确估计$\mathbb{P}(|S_n/n - y| < \epsilon)$趋于0的速率，我们假设具有有限指数矩的随机变量$X_j$的分布为$\mu$。定义矩母函数和累积量生成函数为

$$M(\lambda) = \mathbb{E}e^{\lambda X_j}, \Lambda(\lambda) = \log M(\lambda).$$

并且定义$\Lambda(t)$的Legendre-Fenchel变换为

$$I(x) = \sup_{\lambda}\{x\lambda - \Lambda(\lambda)\}.$$

让$\mu_n(\Gamma) = \mathbb{P}(S_n/n \in \Gamma)$. 接下来重点对$\mu_n$进行讨论。

定理 2.4. *(Cramér 定理)* $\{\mu_n\}$满足大偏差原理：

- 对于任何的闭集$F \in \mathcal{R}$，有
$$\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(F) \leq -\inf_{x\in F} I(x).$$
- 对于任何的开集$G \in \mathcal{R}$，有
$$\liminf_{n\to\infty} \frac{1}{n}\log \mu_n(G) \geq -\inf_{x\in G} I(x).$$

这两个界，定理暗含了

$$\mu_n(\Gamma) \asymp \exp(-n \inf_{x\in\Gamma} I(x)).$$

这里的符号 $\asymp$ 表示在对数级别上的等价关系，也即如果$c_\epsilon \asymp d_\epsilon$,表示 $\lim_{\epsilon\to 0} \log c_\epsilon / \log d_\epsilon = 1$ 。

这是对于独立同分布的随机变量的大偏差定理，大致表明大偏差事件的概率以指数速率收敛到0，且速率函数为 $I(x)$。

引理 2.1. 假设 $f : \mathbb{R}^d \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ 是一个下半连续的凸函数。让 F 是它的共轭函数*(*的*Legendre-Fenchel*变换*)*：

$$F(y) = \sup_{\mathbf{x}}\{(\mathbf{x}, \mathbf{y}) - f(\mathbf{x})\}.$$

那么：

- F也是下半连续的凸函数。

- *Fenchel*不等式成立

$$(\mathbf{x},\mathbf{y}) \le f(\mathbf{x}) + F(\mathbf{y}).$$

- 共轭关系保持

$$f(\mathbf{x}) = \sup_{\mathbf{y}}\{(\mathbf{x},\mathbf{y}) - F(\mathbf{y})\}$$

引理 2.2. *速率函数 $I(x)$ 有以下性质:*

- *$I(x)$ 是一个下半连续的凸函数。*
- *$I(x)$ 非负且 $I(\eta) = 0$*
- *$I(x)$ 在 $[\eta,\infty)$上不减，在 $(-\infty,\eta]$ 上不增。*
- *如果 $x > \eta$，那么 $I(x) = \sup_{\lambda>0}\{\lambda x - \Lambda(\lambda)\}$，如果 $x < \eta$，那么 $I(x) = \sup_{\lambda<0}\{\lambda x - \Lambda(\lambda)\}$。*

当把大偏差理论应用到Bernoulli分布时，速率函数对应相对熵或Kullback-Leibler 散度以及Sanov 定理，这里不在进一步讨论。

在统计力学的应用 大偏差理论与平衡态热力学息息相关，例如可以直接把速率函数与Boltzmann熵联系起来。

考虑一个含有n个独立自旋的系统，每个自旋向上和向下的概率都为 1/2。如果向上，我们记作1，反之记作0. 那么我们能定义微观状态集为

$$\Omega = \{\omega : \omega = (s_1, s_2, \ldots, s_n), s_i = 1 \text{ or } 0\}.$$

对于每个微观状态，我们定义它的平均能量为

$$h_\omega = \frac{1}{n}\sum_{i=1}^{n} s_i.$$

在热力学中，熵是宏观能量的函数。在统计力学中，Boltzmann 熵被定义为

$$S(E) = k_B \log W(E),$$

其中 k_B表示Boltzmann常数且 $W(E)$ 表示固定能量 E 对应的微观状态数。

从大偏差理论我们知道

$$I(E) = \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(h_n \in [E, E+dE]),$$

其中 dE 是一个无穷小量，所以

$$\begin{aligned} I(E) &= \lim_{n\to\infty} -\frac{1}{n}\log\frac{W(h_n \in [E, E+dE])}{2^n} \\ &= \log 2 - \frac{1}{k_B}\lim_{n\to\infty}\frac{1}{n}S_n(E). \end{aligned}$$

这里 $S_n(E)$ 表示 n个自旋系统的Boltzmann熵，记 $S(E)$ 为 $n\to\infty$时，$S_n(E)/n$ 的极限。我们可以得到

$$k_B I(E) = k_B \log 2 - \frac{1}{k_B}S(E).$$

我们会发现速率函数是负的熵（差一个因子 $1/k_B$ ）加一个常数。

对于上面有一般的结果，在一个正则系综的统计力学中（自旋数目 n,和温度 T 固定不变），下面讨论 Λ 的物理意义。能量 $H_n(\omega) = nh_n(\omega)$ 的对数矩生成函数,且归一化 $1/n$之后为

$$\Lambda(\lambda) = \lim_{x\to\infty} \frac{1}{n} \log \mathbb{E} e^{\lambda H_n(\omega)},$$

这里使用 H_n 而不是 nh_n是因为的描述也可以解释一般情况。让 $\lambda = -\beta = -(k_B T)^{-1}$,我们有

$$\Lambda(\lambda) = \lim_{x\to\infty} \frac{1}{n} \log \left(\sum_{\omega} e^{-\beta H_n(\omega)} \right) - \log 2.$$

定义配分函数

$$Z_n(\beta) = \sum_{\omega} e^{-\beta H_n(\omega)},$$

和 Helmholtz 自由能

$$F_n(\beta) = -\frac{1}{\beta} \log Z_n(\beta).$$

我们有

$$\Lambda(-\beta) = -\beta \lim_{x\to\infty} \frac{1}{n} F_n(\beta) - \log 2 = -\beta F(\beta) - \log 2.$$

因此自由能函数 $F(\beta)$ 是负对数矩生成函数加熵一个常数。

根据大偏差理论我们有

$$-\beta F(\beta) - \log 2 = \sup_E \{-\beta E - \log 2 + k_B^{-1} S(E)\};$$

也即

$$F(\beta) = \inf_E \{\beta E - TS(E)\}.$$

这个最小值在驻点 E^* 取得使得

$$\frac{\partial S(E)}{\partial E}\Big|_{E=E^*} = \frac{1}{T},$$

这里与与热力学中的概率一致，且 E^*被称为内能。

2.4 极值统计

在此讨论独立同分布随机变量序列的最大值或最小值。让 $\{X_j\}_{j=1}^n$ 是一列独立同分布的随机变量。最大值 $M_n = \max\limits_{1\le j\le n} X_j$,我们想要了解当 $n \to \infty$ 时，M_n 的分布是怎么样的。最小值的情况可以能被

$$\min_{1\le j\le n} X_j = -\max_{1\le j\le n} (-X_j).$$

定理 2.5. *(Fisher-Tippett-Gnedenko定理) 如果存在 $\{a_n\}$和 $\{b_n\}$使得对于任意的*

$$\mathbb{P}(a_n(M_n - b_n) \le x) \to G(x), \quad asn \to \infty,$$

那么 $G(x)$ 一定有下面形式:

- *Type I (Gumbel):* $G(x) = e^{-e^{-x}}$.
- *Type II (Fréchet):* $G(x) = \begin{cases} 0 & x \le 0, \\ e^{-(x^{-\alpha})} & x > 0, \alpha > 0. \end{cases}$
- *Type III (Weibull):* $G(x) = \begin{cases} e^{-|x^{\alpha}|} & x \le 0, \alpha > 0. \\ 1 & x > 0, \end{cases}$

3 Part3 马尔科夫链

3.1 离散时间有限马尔可夫链

让 $(\Omega, \mathcal{F}, \mathbb{P})$ 是一个概率空间。一个马尔可夫链是一族带参数的随机变量 $\{X_t\}_{t\in\mathbf{T}}$，满足马尔可夫性质(下面引入)，其中 $\mathbf{T}$是指标集合。当 $\mathbf{T}=\mathbb{N}$ 时，记作这族随机变量为 $\{X_n\}_{n\in\mathbb{N}}$,这也是本章所考虑的。假设 X_n 在状态空间 S中取值。

定义 3.1. 如果

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}),$$

则称 $\{X_n\}$ 满足马尔可夫性质，即未来状态只依赖于当前状态，与过去无关。

例子（Symmetric random walk, Ehrenfest's diffusion model, Autoregressive model）我们这里不在描述。

对于离散时间的马尔科夫链有两个关键要素，即状态空间(马尔科夫链的取值空间)和转移概率。为了简单起见，我们假设状态空间 S 是有限的。不失一般性，假设状态空间 $S=\{1,2,\ldots,N\}$，则考虑时间 n 时刻在状态 k 条件下，时间 $n+1$ 时刻在状态 j 的条件概率为

$$p_{kj}^{(n)} = \mathbb{P}(X_{n+1} = j | X_n = k),$$

称之为转移概率。如果 $p_{kj}^{(n)}$ 不依赖于 n，称马尔科夫链是平稳的。在此我们仅仅讨论平稳的马尔科夫链，让$p_{kj}^{(n)} = p_{kj}$，若记 $\mathbf{P}=(p_{ij})_{i,j\in S}$为转移概率矩阵，则有

$$p_{ij} \geq 0, \quad \sum_{j\in S} p_{ij} = 1.$$

根据Bayes公式和马尔可夫性容易得到

命题 3.1. *(Chapman-Kolmogorov 方程)*

$$\mathbb{P}(X_n = j | X_0 = i) = \sum_{k\in S} \mathbb{P}(X_n = j | X_m = k)\mathbb{P}(X_m = k | X_0 = i), \quad 1 \leq m \leq n-1.$$

使用Chapman-Kolmogorov 方程，我们能得到

$$\mathbb{P}(X = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mathbb{P}(X_0 = i_0) \prod_{j=0}^{n-1} p_{i_j i_{j+1}}.$$

且又有

$$\mathbb{P}(X_n = i_n | X_0 = i_0) = \sum_{i_1,\ldots,i_{n-1}\in S} \prod_{j=0}^{n-1} p_{i_j i_{j+1}} = (\mathbf{P}^n)_{i_0 i_n},$$

最后一项是矩阵 $\mathbf{P}$ 的 n 次幂的 (i_0, i_n) 元素。

3.2 不变分布

给定马尔科夫链的初始分布 μ_0,则 X_n的分布为

$$\mu_n = \mu_0 \mathbf{P}^n.$$

如果存在一个概率分布 π，使得

$$\pi = \pi \mathbf{P}, \tag{1}$$

则称 π 为马尔科夫链的平稳分布（stationary distribution）。如果初始分布是平稳分布，则马尔科夫链的分布是不随时间变化的。满足Eq. (1)的测度 π 还满足 $\pi_i \geq 0, \forall i \in S$,称为马尔科夫链的平衡分布。(π 不需要归一化)

下面考虑平稳分布何时存在和唯一。这与概率转移矩阵 $\mathbf{P}$ 的性质有关。问题归结到 $\mathbb{P}$是否存在非负的左特征向量且对应的特征值为1。

引理 3.1. $\mathbb{P}$的谱半径等于*1*:

$$\rho(\mathbb{P}) = \max_{\lambda} |\lambda| = 1,$$

其中最大值是取在 $\mathbb{P}$ 的特征值上。

定义 3.2. *(不可约性)* 如果存在一个置换矩阵 $\mathbf{R}$，使得 $\mathbf{R}^T\mathbf{P}\mathbf{R} = \begin{pmatrix} \mathbf{A_1} & \mathbf{B} \\ 0 & \mathbf{A_2} \end{pmatrix}$，则称 $\mathbf{P}$ 是可约的,否则是不可约的。

不可约性除了通过上面的置换矩阵定义外，还可以通过状态之间的可通信定义。如果存在 $s \geq 1$和 $(k_1, \cdots, k_{s-1})$使得

$$p_{k_0 k_1} p_{k_1 k_2} \cdots p_{k_{s-1} k_s} > 0, \quad k_0 = i, k_s = j,$$

则称 j是从 i 可达的。如果 j是从 i 可达的且 i是从 j 可达的，则称 (i, j) 是可通信的。

引理 3.2. *(不可约性与可通信)* 如果 $\mathbf{P}$ 是不可约的，等价于状态空间中任意两个状态 i, j 都是可通信的。

定理 3.1. *(Perron-Frobenius 定理)* 如果 $\mathbf{A}$ 是不可约的且有正的转移概率，且让 $\rho(A)$是其谱半径：$\rho(A) = \max_\lambda |\lambda(A)|$，那么

- $\lambda = \rho(A)$ 是重数唯*1*的特征值。
- *A*存在正的左右特征向量 x 和 y ，使得

$$x^T A = \lambda x^T, x_j > 0 \quad Ay = \lambda y, y_j > 0.$$

特别地，不可约性不是存在唯一的平稳分布的必要条件。

3.3 有限马尔科夫链的遍历性

不可约性保证了时间平均收敛到空间平均是相等的，但是不可约性不足以保证强遍历性，也即

$$\mu_n = \mu_0 \mathbf{P}^n \to \pi,$$

对任意的初始分布 μ_0。

定义 3.3. 一个马尔科夫链是原始链，如果存在 $s > 0$ 对每一个对 (i, j) 使得 $(\mathbf{P}^s)_{ij} > 0$ 。

定理 3.2. 如果一个马尔科夫链是原始链，那么对于任意的初始分布 μ_0，都有

$$\mu_n = \mu_0 \mathbf{P}^n \to \pi, \quad \textit{exponential fast as } n \to \infty,$$

其中 π 是唯一的平稳分布。

定理 3.3. *(Ergodic 定理)* 如果一个$\{X_m\}$ 是不可约的有限马尔科夫链，让 f为 S到 $\mathbb{R}$ 的有界函数，那么

$$\frac{1}{n} \sum_{m=0}^{n-1} f(X_m) \to \langle f \rangle_\pi = \sum_{i \in S} f(i)\pi_i, \quad n \to \infty \ a.s.$$

其中 π 是唯一的平稳分布。

3.4 Poisson 过程

在此考虑 $\mathbf{T} = \mathbb{R}_+$的情况。在此考虑有限状态的情况，先可数状态的只考虑常数速率的Possion过程，并且使用 X_t表示在时间 t 时刻的状态。

定义 3.4. *(Poisson 过程)* 让 $\{N_t\}_{t\geq 0}$ 是一个增的右连续的整数值过程，满足

- $N_0 = 0$;
- N_t有独立平稳增量，也即：对于任意的 $0 \leq t_1 < t_2 < \cdots < t_n$, 有

$$N_{t_2} - N_{t_1}, N_{t_3} - N_{t_2}, \cdots, N_{t_n} - N_{t_{n-1}},$$

是独立的，且对于任意的 $t \geq 0, s \geq 0$，

$$N_{t+s} - N_t,$$

的分布独立于 t。

- 对于任意的 $t \geq 0, h > 0$，当 $h \to 0_+$时，有

$$\begin{aligned}\mathbb{P}(N_{t+h} = N_t + 1|N_t) &= \lambda h + o(h)\\ \mathbb{P}(N_{t+h} = N_t|N_t) &= 1 - \lambda h + o(h)\\ \mathbb{P}(N_{t+h} \geq N_t + 2) &= o(h),\end{aligned}$$

其中 $\lambda > 0$ 是一个正常数，称为速率。

那么 $\{N_t\}_{t\geq 0}$ 是一个有速率 λ 的 *Poisson*过程。

定理 3.4. 对任意固定的时间 t，N_t服从参数为 λt 的 *Poisson* 分布。

等待时间分布 对于 Poisson 过程，定义等待时间为

$$\tau = \inf\{t : t > 0 : N_t \neq N_0\},$$

且等待时间的分布为

$$\nu(t) = \mathbb{P}(\tau \leq t).$$

那么 $\nu(0) = 1$且令

$$\nu(t) - \nu(t + h) = \nu(t)\lambda h + o(h), h \ll 1.$$

令 $h \to 0$, 我们能得到

$$\frac{d}{dt}\nu(t) = -\nu(t)\lambda,$$

因此可以得到

$$\nu(t) = e^{-\lambda t},$$

这意味着等待时间为速率为 λ 的指数分布。

3.5 Q过程

在此部分我们讲考虑一类连续时间的马尔可夫链，称为Q过程。让 $\{X_t\}_{t\geq 0}$ 是一个右连续的时间马尔可夫链，且状态空间为有限状态空间 $S(S=\{1,2,\cdots,N\}$)。让转移概率表示为

$$p_{ij}(t)=\mathbb{P}(X_{t+s}=j|X_s=i).$$

这里假设马尔可夫链是平稳的，即 $p_{ij}(t)=p_{ij}$ 不依赖于 t。通过上面定义，我们有

$$p_{ij}(t)\geq 0,\quad \sum_{j\in S}p_{ij}(t)=1.$$

另外，我们要求当 $h\to 0_+$时，满足

$$\begin{aligned}p_{ii}&=1-\lambda_i h+o(h),\quad \lambda_i>0,\\ p_{ij}&=\lambda_{ij}h+o(h),\quad j\neq i.\\ p_{jj}&=1.\end{aligned}$$

从上面的非负性和归一化条件，我们可以得到

$$\lambda_{ij}\geq 0,\quad \sum_{j\in S,i\neq j}\lambda_{ij}=\lambda_i.$$

这个过程的马尔可夫性可以得到Chapman-Kolmogorov 方程

$$p_{ij}(t+s)=\sum_{k\in S}p_{ik}(t)p_{kj}(s),\quad \forall i,j\in S,t,s\geq 0.$$

当引入矩阵形式时，

$$\mathbf{P}(t)=(p_{ij}(t))_{i,j\in S},$$

则Chapman-Kolmogorov 方程可以写成矩阵形式

$$\mathbf{P}(t+s)=\mathbf{P}(t)\mathbf{P}(s)=\mathbf{P}(s)\mathbf{P}(t).$$

在此定义

$$\mathbf{Q}=\lim_{h\to 0_+}\frac{\mathbf{P}(h)-I}{h},$$

且记作 $\mathbf{Q}=(q_{ij})$,根据 $\mathbf{P}(h)$ 的定义，我们有

$$q_{ii}=-\lambda_i,\quad q_{ij}=\lambda_{ij}(i\neq j),\quad \sum_{j\in S}q_{ij}=0.$$

我们称 $\mathbf{Q}$ 为马尔科夫链的生成元矩阵(Generator matrix)。为了方便，我们也定义 $q_i=-q_{ii}\geq 0$。因为

$$\frac{\mathbf{P}(t+h)-\mathbf{P}(t)}{h}=\mathbf{P}(t)\frac{\mathbf{P}(h)-\mathbf{I}}{h},$$

当 $h\to 0_+$ 时, 可以得到 $\mathbf{P}(t)$ 的前向方程为

$$\frac{d\mathbf{P}(t)}{dt}=\mathbf{P}(t)\mathbf{Q}.$$

由于 $\mathbf{P}(h)$和$\mathbf{P}(t)$可交换，我们后向方程

$$\frac{d\mathbf{P}(t)}{dt}=\mathbf{Q}\mathbf{P}(t).$$

由于 $\mathbf{P}(0)=\mathbf{I}$，我们可以得到 $\mathbf{P}(t)$ 的解为

$$\mathbf{P}(t)=e^{\mathbf{Q}t}\mathbf{P}(0)=e^{\mathbf{Q}t}.$$

下面一部分我们讨论马尔科夫链的分布是随时间如何演化的。让 $\mu(t)$ 是 X_t 的分布，那么对于无穷小量 dt，我们有

$$\begin{aligned}\mu_j(t+dt)&=\sum_{i\neq j}\mu_i(t)p_{ij}(dt)+\mu_j(t)p_{jj}(dt)\\&=\sum_{i\neq j}\mu_i(t)q_{ij}dt+\mu_j(t)(1-q_{jj})dt+o(dt).\end{aligned}$$

也就是

$$\frac{d\mu(t)}{dt}=\mu(t)\mathbf{Q},$$

这是对于分布的前向Kolmogorov 方程。它的解为

$$\mu_j(t)=\sum_{i\in S}\mu_i(0)p_{ij}(t),$$

或者矩阵形式有

$$\mu(t)=\mu(0)e^{\mathbf{Q}t}.$$

因此下面考虑Q过程的不变分布。让 π 是平稳分布，则有

$$\pi\mathbf{Q}=0,\quad \pi\cdot\mathbf{1}=1,$$

其中 $\mathbf{1}$ 是全1向量，且不变分布也满足

$$\pi\mathbf{P}(t)=\pi,\quad \forall t\geq 0.$$

等待时间分布 考虑对于每个初始状态 j，等待时间分布为

$$\nu_j(t)=\mathbb{P}(\tau\geq t|X_0=j),$$

其中 $\tau=\inf\{t>0:X_t\neq j\}$。那么有与前面的Poisson过程类似的结果

$$\frac{d\nu_j(t)}{dt}=q_{jj}\nu_j(t),\quad \nu_j(0)=1.$$

因此状态 j 的等待时间分布为速率为 $q_j=-q_{jj}$ 的指数分布。由于指数分布的无记忆性，等待时间能从任何时间开始计算。

我们感兴趣下面概率

$$p(\theta,j|0,i):=\mathbb{P}^i(\text{the jump time }\tau\text{ is in }[\theta,\theta+d\theta]\text{ and }X_\tau=j),$$

其中 $\mathbb{P}^i$ 表示从状态 i 出发的分布。我们有

$$\begin{aligned}p(\theta,j|0,i)d\theta&=\mathbb{P}(\text{no jump occurs in }[0,\,\theta)\text{ given }X_0=i)\\&\quad\times\mathbb{P}(\text{one jump occurs from }i\text{ to }j\text{ in }[\theta,\theta+d\theta))\\&=\nu_i(\theta)q_{ij}d\theta=\exp(q_{ii}\theta)q_{ij}d\theta,\end{aligned}$$

因此有

$$\mathbb{P}^i(X_\tau=j)=p(j|0,i)=\frac{q_{ij}}{q_i}=\frac{q_{ij}}{\sum\limits_{j\neq i}q_{ij}}$$

其中 τ 为等待时间。这些结果对于Q过程的随机模拟非常有用，这也是化学反应动力学中常用的随机模拟方法（SSA或 Gillespie算法）。

3.6 嵌入链与不可约性

一个有生成元矩阵 $\mathbf{Q}$ 的Q过程连接到一个具有转移概率矩阵 $\tilde{\mathbf{Q}} = (\tilde{q}_{ij})$ 的离散时间的马尔科夫链，其中

$$\tilde{q}_{ij} = \begin{cases} q_{ij}/q_i, & \text{if } i \neq j, \text{and } q_i > 0 \\ 0, & \text{if } i \neq j, \text{and } q_i = 0, \end{cases}$$

$$\tilde{q}_{ii} = \begin{cases} 0, & \text{if } q_i > 0 \\ 1, & \text{if } q_i = 0, \end{cases}$$

$\tilde{\mathbf{Q}}$ 被称为跳矩阵，且这个马尔科夫链被称为原始Q矩阵的嵌入链或跳链。

定义原始Q过程 $\{X_t\}_{t\geq 0}$ 的跳时间为

$$J_0 = 0, \quad J_{n+1} = \inf\{t : t > J_n : X_t \neq X_{J_n}\}, \quad n \in \mathbb{N},$$

其中我们使用惯例 $\inf\{\emptyset\} = \infty$. 定义持续时间

$$H_n = \begin{cases} J_n - J_{n-1} & \text{if} J_{n-1} < \infty \\ \infty & \text{otherwise}, \end{cases}$$

其中 $n = 1, 2, \ldots$。如果 $J_{n+1} = \infty$, 我们让 $X_\infty = X_{J_n}$。那么由 X_t 诱导的跳链为

$$Y_n = X_{J_n}, n \in \mathbb{N}.$$

定理 3.5. *这个跳链 $\{Y_n\}$ 的转移概率矩阵为 $\tilde{\mathbf{Q}}$, 且有持续时间 $H_1, H_2, \ldots$ 是独立的指数随机变量有参数 $q_{Y_0}, q_{Y_1}, \ldots$。*

如果存在 $t > 0$使得 $p_{ij}(t) > 0$, 我们称状态 j 从状态 i 可达的。类似离散时间的马尔科夫链，我们可以定义可通信关系和不可约性。

定理 3.6. *Q过程 X 是不可约的当且仅当它的嵌入链 $\{Y_n\}$ 是不可约的。*

3.7 Q过程的遍历性

对于Q过程，不可约性足以保证收敛到稳态。

定理 3.7. *(稳态收敛性) 假设 $\mathbf{Q}$ 是不可约的，那么对于任何的初始分布 $\mu(0)$，都有*

$$\mu(t) = \mu(0)\mathbf{P}(t) \to \pi, \quad \textit{exponentially fast as } t \to \infty,$$

其中 π 是唯一的不变分布。

定理 3.8. *(遍历性) 假设 $\mathbf{Q}$ 是不可约的，那么对于任意有界函数 $f : S \to \mathbb{R}$，都有*

$$\frac{1}{T}\int_0^T f(X_s)ds \to \langle f \rangle_\pi = \sum_{i\in S} f(i)\pi_i, \quad T \to \infty \ a.s.$$

其中 π 是唯一的不变分布。

3.8 时间反演

首先让我们考虑一个不可约的离散时间马尔科夫链 $\{X_n\}_{n\in\mathbb{N}}$，且有转移概率矩阵 $\mathbf{P}$，让 π 是唯一的平稳分布。假设初始分布也为 π。定义时间逆过程 $\{\hat{X}_n\}_{0\le n\le N}, \hat{X}_n = X_{N-n}$, 其中 N 是一个固定的正整数。我们想要研究过程 $\{\hat{X}_n\}$ 的转移规则。

注意到对任何的 $i_0, i_1, \ldots, i_N \in S$，都有

$$\begin{aligned}
&\mathbb{P}(\hat{X}_0 = i_0, \hat{X}_1 = i_1, \ldots, \hat{X}_N = i_N) \\
&\quad = \mathbb{P}(X_N = i_0, X_{N-1} = i_1, \ldots, X_0 = i_N) \\
&\quad = \pi_{i_N} p_{i_N i_{N-1}} \cdots p_{i_1 i_0} \\
&\quad = \pi_{i_0} \hat{p}_{i_0 i_1} \cdots \hat{p}_{i_{N-1} i_N},
\end{aligned}$$

其中矩阵 $\hat{\mathbf{P}} = (\hat{p}_{ij})_{i,j\in S}$ 的元素定义为

$$\hat{p}_{ij} = \frac{\pi_j p_{ji}}{\pi_i}, \quad i, j \in S.$$

经过计算显示 $(\hat{X}_n)_{0\le n\le N}$ 也是一个马尔科夫链，且其转移概率矩阵为 $\hat{\mathbf{P}}$, 不变分布为 π。

对于上面的结构可以应用到Q过程。对一个不可约的Q过程 $\{X_t\}_{0\le t\le T}$有生成元矩阵 $\mathbf{Q}$，且有唯一的不变分布 π。初始分布也假设为 π。定义时间逆过程$\{\hat{X}_t\}_{0\le t\le T}$,$\hat{X}_t = X_{T-t}$。对于 $\hat{X}$ 在时间点 $0 \le t_0 < t_1 < \ldots < t_N \le T$ 的联合分布为

$$\begin{aligned}
&\mathbb{P}(\hat{X}_{t_0} = i_0, \hat{X}_{t_1} = i_1, \ldots, \hat{X}_{t_N} = i_N) \\
&\quad = \mathbb{P}(X_{T-t_N} = i_0, X_{T-t_{N-1}} = i_1, \ldots, X_{T-t_0} = i_N) \\
&\quad = \pi_{i_N} p_{i_N i_{N-1}}(s_N) \cdots p_{i_1 i_0}(s_1) \\
&\quad = \pi_{i_0} \hat{p}_{i_0 i_1}(s_1) \cdots \hat{p}_{i_{N-1} i_N}(s_N),
\end{aligned}$$

其中 $s_k = t_k - t_{k-1}, k = 1, 2, \ldots, N$。像前面一样，时间依赖的转移概率矩阵 $\hat{\mathbf{P}}(t)$ 的元素定义为

$$\hat{p}_{ij}(t) = \frac{\pi_j p_{ji}(t)}{\pi_i}, \quad i, j \in S.$$

那么 $\hat{\mathbf{P}}(t)$ 所满足的Chapman-Kolmogorov 方程为

$$\frac{d\hat{\mathbf{P}}(t)}{dt} = \hat{\mathbf{Q}}\hat{\mathbf{P}}(t) = \hat{\mathbf{P}}(t)\hat{\mathbf{Q}}.$$

其中 $\hat{\mathbf{Q}}$ 是时间逆过程的生成元矩阵，其元素为

$$\hat{q}_{ij} = \frac{\pi_j q_{ji}}{\pi_i}, \quad i, j \in S.$$

忽略$\hat{X}$路径的左连续性，我们可以得到 $\hat{X}$ 是一个Q过程，有不变分布 π，和生成元矩阵 $\hat{\mathbf{Q}}$。

一类特殊的马尔可夫链是满足细致平衡条件的，也即 $\hat{X}$与 X 有相同的统计特征。对于离散时间和连续时间的不同，细致平衡条件不同，有下面形式

$$\pi_i p_{ij} = \pi_j \hat{p}_{ji} \quad \text{(离散时间)}, \qquad \pi_i q_{ij} = \pi_j q_{ji} \quad \text{(连续时间)}.$$

在这个情况下，我们有 $\hat{p}_{ij} = p_{ij}$或 $\hat{q}_{ij} = q_{ij}$。我们称这样的马尔科夫链是可逆的。一个可逆的链具有变分结构和良好的谱性质。在离散时间情况下，定义Laplacian矩阵为

$$\mathbf{L} = \mathbf{P} - \mathbf{I},$$

且它对任意函数的作用为

$$(\mathbf{L}f)(i) = \sum_{j\in S} p_{ij}(f(j) - f(i)).$$

让 L^2_π 是有 π权重标量积

$$(f, g)_\pi = \sum_{i\in S} \pi_i f(i)g(i).$$

的平方可积的函数空间。定义Dirichlet形式为, 或函数 f 的能量，

$$D(f) = \frac{1}{2}\sum_{i,j\in S} \pi_i p_{ij}(f(j) - f(i))^2.$$

因此可以验证有 $D(f) = (f, -\mathbf{L}f)_\pi$。相似的结构能对Q过程以及一般的随机过程。这些公式在马尔科夫链的位势理论中特别有用。

3.9 隐马尔科夫模型

隐马尔科夫链的基本问题是基于观测数据来推断马尔科夫规则。下面我们将考虑已知马尔科夫链生成状态序列的反问题。我们会限制在离散时间环境。我们可获得的时间序列为 $\mathbf{Y} = (Y_{1:N}) = (Y_1, \cdots, Y_N)$,这是一个马尔科夫链的轨迹 $\mathbf{X} = (X_{1:N}) = (X_1, \cdots, X_N)$,的部分观测，该马尔科夫链的初始分布为 $\pi = (\mu_i)_{i\in S}$，转移概率矩阵为 $\mathbf{P} = (p_{ij})_{i,j\in S}$。我们假设观测是在状态空间 O 和观测概率通过所谓的emission矩阵 $\mathbf{R} = (r_{ij})_{i\in S, j\in O}$来定义，其中 $r_{ij} = \mathbb{P}(Y = j|X = i)$表示当隐藏状态是 i 时，观测为 j。在此我们假设 $\mathbf{P}$ 和 $\mathbf{R}$ 是时间独立的。那么隐马尔可夫模型的所有参数为：

$$\theta = (\mathbf{P}, \mathbf{R}, \mu).$$

在此形式上，在隐藏状态 $\mathbf{X}$ 条件下，观测 $\mathbf{Y}$的条件概率为

$$\mathbb{P}(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{n=1}^{N} \mathbb{P}(Y_n|X_n, \theta) = \prod_{n=1}^{N} r_{X_n Y_n},$$

并且在参数 θ 下，生成隐藏状态序列 $\mathbf{X}$ 的概率为

$$\mathbb{P}(\mathbf{X}|\theta) = \mu_{X_1} \prod_{n=1}^{N-1} p_{X_n X_{n+1}}.$$

因此在给定参数 θ 下，状态序列 $\mathbf{X}$ 和观测序列 $\mathbf{Y}$ 的联合概率为

$$\mathbb{P}(\mathbf{X}, \mathbf{Y}|\theta) = \mathbb{P}(\mathbf{Y}|\mathbf{X}, \theta)\mathbb{P}(\mathbf{X}|\theta) = \mu_{X_1} r_{X_1 Y_1} \prod_{n=1}^{N-1} p_{X_n X_{n+1}} r_{X_{n+1} Y_{n+1}}.$$

这个联合概率分布统计上也被称为似然函数。我们主要关注以下三种类型的问题：

- **模型选择**：观测序列 $\mathbf{Y}$和可能的参数集合 $\theta_k = (\mathbf{P}_k, \mathbf{R}_k, \mu_k)$，寻找生成该观测数据的最高效可能模型。
- **最优预测**：观测序列 $\mathbf{Y}$和参数 θ，寻找最可能的状态序列$\mathbf{X}$解释观测序列。
- **参数估计**：给定观测序列 $\mathbf{Y}$，发现模型参数参数 $\theta = (\mathbf{P}, \mathbf{R}, \mu)$,和最大似然函数 $\mathbb{P}(\mathbf{X}, \mathbf{Y}|\theta)$。

下面我们将介绍如何解决这些问题。

模型选择 从 $\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta)$, 我们得到边缘分布

$$\mathbb{P}(\mathbf{Y}|\theta)=\sum_{\mathbf{x}}\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta) \tag{2}$$

并且对于第 k 个模型 θ_k 在观测 $\mathbf{Y}$ 的后验分布为

$$\mathbb{P}(\theta_k|\mathbf{Y})=\frac{1}{\mathbb{P}(\mathbf{Y})}\mathbb{P}(\mathbf{Y}|\theta_k)\cdot\mathbb{P}(\theta_k),\quad k=1,\cdots,K$$

其中

$$\mathbb{P}(\mathbf{Y})=\sum_{k}\mathbb{P}(\mathbf{Y}|\theta_k)\cdot\mathbb{P}(\theta_k)$$

. 因为观测序列 $\mathbf{Y}$ 是固定的，我们有

$$\mathbb{P}(\theta_k|\mathbf{Y})\propto\mathbb{P}(\mathbf{Y}|\theta_k)\cdot\mathbb{P}(\theta_k).$$

先验的，我们假设对于模型的选择是无偏的，因此

$$\mathbb{P}(\theta_k)=\frac{1}{K}.$$

并且

$$\mathbb{P}(\theta_k|\mathbf{Y})\propto\mathbb{P}(\mathbf{Y}|\theta_k).$$

因此确定哪个模型是最适合数据的最自然方法是在所有 K 个候选模型中选择 $\mathbb{P}(\theta_k|\mathbf{Y})$ 最大值。然后通过(2)可以知道我们需要经过 I^N 次操作，甚至需要更多次的计算。这在实际中是不可行的，因此我们需要使用近似方法来计算。

隐马尔科夫链的前向算法是基于动态规划的思想。考虑前向量 $\alpha_n(i)$ 定义为

$$\alpha_n(i)=\mathbb{P}(Y_{1:n}=y_{1:n},X_n=i|\theta),$$

表示给定模型 θ下，观测序列 $Y_{1:n}$ 的概率且在时间 n处的隐藏状态。在此 $\{\alpha_n(i)\}$满足动态规划原理：

$$\alpha_n(i)=\left(\sum_{j\in S}\alpha_{n-1}(j)p_{ji}\right)r_{iy_n},n=2,\cdots,N;i\in S. \tag{3}$$

所以分布 $\mathbb{P}(\mathbf{Y}|\theta)$能通过前向算法计算得到：

算法 3.1. *(前向算法)对于每个模型的后验概率。*

- 初始化：计算 $\alpha_1(i)=\mu_i r_{iY_1},i\in S.$
- 递推：通过(3),$\{\alpha_n(i)\}$.
- 终止：计算最后的后验概率分布 $\mathbb{P}(\mathbf{Y}|\theta)=\sum_{i\in S}\alpha_N(i).$

注意前向算法的时间复杂度为 $O(IN)$. 前向算法的基本思想是通过利用递归关系来避免枚举所有可能路径。使用前向算法，可以计算每个模型的后验概率，并选择概率最高的模型。在贝叶斯统计中称为最大后验概率（MAP）估计。

最优预测 从联合分布 $\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta)$, 我们有条件概率

$$\mathbb{P}(\mathbf{X}|\mathbf{Y},\theta)=\frac{1}{\mathbb{P}(\mathbf{Y}|\theta)}\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta)\propto\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta)$$

这里是因为参数 θ 和观测 $\mathbf{Y}$ 是已知的。所以给定参数 θ 和观测 $\mathbf{Y}$，最优状态序列 $\mathbf{X}$ 能被定义为使得概率分布 $\mathbb{P}(\mathbf{X},\mathbf{Y}|\theta)$在路径空间 S^N,上的最大。再一次，枚举法不太行，这里用到 Viterbi 算法。

Viterbi 算法也是使用动态规划的思想把最优问题拆解为子问题。首先，让我们定义部分概率 $\delta_n(i)$:

$$\delta_n(i)=\max_{x_{1:n}}\{\mathbb{P}(X_{1:n-1}=x_{1:n-1},X_n=i,Y_{1:n}=y_{1:n}|\theta)\},$$

到时间 n 的最后状态 $X_n=i$ 且部分观察序列。通过迭代我们有

$$\delta_{n+1}(j)=\left(\max_{i\in S}(\delta_n(i)p_{ij})\right)r_{jy_{n+1}},n=1,\cdots,N-1;j\in S.$$

为了检索状态序列，我们需要对于每个 n 和 j使得 $\delta_n(j)$最大化。可以通过上式中第 n 步最大化记录在一个数值中 ψ_n.

算法 3.2. *(Viterbi 算法)* 给定 θ和 $\mathbf{Y}$,寻找最优状态序列 $\hat{\mathbf{X}}=(\hat{X}_{1:N})$.

- 初始化：计算 $\delta_1(i)=\mu_i r_{iY_1},i\in S.$
- 递推：计算

$$\psi_n(j)=\arg\max_{i\in S}\{\delta_n(i)p_{ij}\},n=1,\cdots,N-1;j\in S.$$

$$\delta_n(j)=\left(\max_{i\in S}\{\delta_{n-1}(i)p_{ij}\}\right)r_{jy_n},n=2,\cdots,N;j\in S.$$

- 终止：得到最大概率和记录在最后步的最大值

$$P^*=\max_{i\in S}\{\delta_N(i)\},\hat{X}_N=\arg\max_{i\in S}\{\delta_N(i)\}.$$

- 回溯：通过回溯得到最优状态序列

$$\hat{X}_n=\psi_n(\hat{X}_{n+1}),n=N-1,\cdots,1.$$

除了回溯，Viterbi 算法是相似的于前向算法。最主要的差异是对先前状态的最大化，而非求和。

参数估计 在统计上，参数估计的通用方法是最大化似然函数，也即给定观测 $\mathbf{Y}$, 最优化

$$\max_{\theta}\mathbb{P}(\mathbf{Y}|\theta),$$

其中 $\mathbb{P}(\mathbf{Y}|\theta)$ 定义在(2)。这种最大化并非总是容易的，通常是非凸的且涉及隐藏变量。下面的简单迭代方法 Baum-Welch 算法。

首先我们引入后向过程。我们定义后向变量 $\beta_n(i)$, 表示从 $n+1$ 到 N 的观测序列的概率，给定时间 n 的状态为 i 和模型 θ:

$$\beta_n(i)=\mathbb{P}(Y_{n+1:N}=y_{n+1:N}|X_n=i,\theta).$$

动态规划原理能简化 $\{\beta_n(i)\}$的计算，

$$\beta_n(i)=\sum_{j\in S}p_{ij}r_{jy_{n+1}}\beta_{n+1}(j),n=N-1,N-2,\ldots,1;i\in S. \tag{4}$$

类似前向过程，我们有

算法 3.3. *(后向算法)* 后向变量 $\{\beta_n(i)\}$ 的计算。

- 初始化. 设置 $\beta_N(i) = 1, i \in S$.
- 递推. 通过(4)计算 $\{\beta_n(i)\}$.

而且，让我们定义

$$\xi_n(i,j) = \mathbb{P}(X_n = i, X_{n+1} = j|\mathbf{Y}, \theta), n = 1, \cdots, N-1,$$

为给定观测序列 $\mathbf{Y}$ 和模型 θ 下，在时间 n下，状态 i 到 j 的转移概率。通过前向和后向变量，我们表示马尔可夫性

$$\begin{aligned}\xi_n(i,j) &= \frac{\mathbb{P}(X_n = i, X_{n+1} = j, \mathbf{Y}|\theta)}{\mathbb{P}(\mathbf{Y}|\theta)} \\ &= \frac{\alpha_n(i)p_{ij}r_{jy_{n+1}}\beta_{n+1}(j)}{\sum\limits_{i,j\in S}\alpha_n(i)p_{ij}r_{jy_{n+1}}\beta_{n+1}(j)}\end{aligned}$$

而且，我们定义 ξ_n 在时间 n 的边际分布为

$$\gamma_n(i) = \mathbb{P}(X_n = i|\mathbf{Y}, \theta) = \sum_{j\in S}\xi_n(i,j), n = 1, \cdots, N-1.$$

对 $\gamma_n(i)$ 和 $\xi_n(i,j)$ 在 n处的求和：

$$\sum_{n=1}^{N-1}\gamma_n(i),$$

给定 $(\mathbf{Y}, \theta)$从状态 i 的转移数量。

$$\sum_{n=1}^{N-1}\xi_n(i,j),$$

给定 $(\mathbf{Y}, \theta)$从状态 i 到状态 j 的转移数量。注意到上述等价关系相差一个常数情况下成立。

算法 3.4. *(Baum-Welch 算法)* 隐马尔科夫链模型。

- 初始化：给定初始参数 $\theta = (\mathbf{P}, \mathbf{R}, \mu)$.
- 更新参数 θ:

$$\mu^* = \gamma_1(i), \tag{5}$$

$$p_{ij}^* = \frac{\sum\limits_{n=1}^{N-1}\xi_n(i,j)}{\sum\limits_{n=1}^{N-1}\gamma_n(i)}, \tag{6}$$

$$r_{ij}^* = \frac{\sum\limits_{n=1,Y_n=k}^{N}\gamma_n(j)}{\sum\limits_{n=1}^{N}\gamma_n(j)}. \tag{7}$$

- 迭代：重复步骤 *2*，直到收敛。

3.10 网络和马尔科夫链

对于一个有向或无向的网络 G，其节点集为 S, 权重集为 W,记为 $G=(S,W)$. 我们假设网络有 I个节点即，$S=\{1,2,\ldots,I\}$，权重矩阵为 $W=\{e_{ij}\}_{i,j\in S}$, 其中 e_{ij}为节点 i到节点 j的边权重。最简单的权重矩阵为邻接矩阵。在此部分我们仅考虑无向图即 W是对称的。

对于给定的网络，我们能自然的定义离散时间的马尔科夫链，概率转移矩阵为 $\mathbf{P}=(p_{ij})_{i,j\in S}$，其中

$$p_{ij}=\frac{e_{ij}}{d_i},\quad d_i=\sum_{k\in S}e_{ik}.$$

其中 d_i 为节点 i 的度。让

$$\pi_i=\frac{d_i}{\sum_{k\in S}d_k},$$

那么很容易有结果

$$\sum_{i\in S}\pi_i p_{ij}=\pi_j,$$

也即 π 为马尔科夫链的不变分布。进一步也有细致平衡条件

$$\pi_i p_{ij}=\pi_j p_{ji}.$$

对于定义在 S 上的 $\mathbf{u}=(u_i)_{i\in S},\mathbf{v}=(v_i)_{i\in S}$, 引入内积

$$(\mathbf{u},\mathbf{v})_\pi=\sum_{i\in S}\pi_i u_i v_i.$$

对于该内积， $\mathbf{P}$ 是自伴的，也即

$$(\mathbf{Pu},\mathbf{v})_\pi=(\mathbf{u},\mathbf{Pv})_\pi.$$

对于 $\mathbf{P}$ 的特征值我们记为 $\{\lambda_k\}_{k=0,\cdots,I-1}$，且与之对应的左右特征向量为 $\{\phi_k\}_{k=0}^{I-1}$ 和 $\{\psi_k\}_{k=0}^{I-1}$. 我们有 $(\phi_j,\phi_j)_\pi=\delta_{jk}$(内积的定义),$\psi_{k,i}=\phi_{k,i}\pi_i$,(把这个带入左特征向量，用到细致平衡条件，发现满足)和

$$\mathbf{P}\phi_k=\lambda_k\phi_k,\quad \psi_k^T\mathbf{P}=\lambda_k\psi_k^T,\quad k=0,1,\cdots I-1.$$

更进一步，我们有 1 是特征值，且其他特征值都在区间 $[-1,1]$ 之内（ $\mathbf{P}$的每一行和为 1 ）。我们会把特征值排序 $1=\lambda_0\geq|\lambda_1|\geq\cdots\geq|\lambda_{I-1}|$. 注意到 $\psi_0=\pi$ 并且 ϕ_0 是一个常向量。且对于矩阵 $\mathbf{P}^t$的谱分解为

$$(\mathbf{P}^t)_{ij}=\sum_{k=0}^{I-1}\lambda_k^t\phi_{k,i}\psi_{k,j}.$$

一个有趣的问题是从网络上的随机游走来推断网络的结构和动力学性质。例如我们这里讨论网络的结构和马尔科夫链的可聚合性（lumpability）。一个网络具有社区结构是说，如果网络能被划分为不同子网络的并，且子网络间没有边连接。马尔科夫链是可聚合的（lumpable）是说，如果状态空间能被分解为子集的并，且该马尔科夫链可以等效为子集之间的链，则称马尔科夫链是可聚合的。这两个描述之间相互对应，然而这两个都太严苛。因此在应用中，我们会考虑近似社区结构和近似可聚合性。下面我们开始更准确的描述可聚合性。

给定一个状态空间 S 的马尔科夫链和对应的划分 $S:S=\cup_{k=1}^K C_k$，且如果 $k\neq l$，则 $C_k\cap C_l$, 通过联合概率分布

$$p_{i_0i_1\cdots i_m}=\mathbb{P}\{X_m\in C_{i_m}|X_{m-1}\in C_{i_{m-1}},\cdots,X_0\in C_{i_0}\},\tag{8}$$

定义聚合过程。

定义 3.5. *(可聚合性)* 一个马尔科夫链是关于划分 $S = \cup_{k=1}^{K} C_k$ 可聚合的, 如果对每一个初始分布 μ_0，聚合过程可以通过(8)定义的概率分布来描述, 是一个马尔科夫链且转移概率不依赖于初始分布 μ_0。

定理 3.9. 一个马尔科夫链的转移概率 $\mathbf{P}$ 是关于划分 $S = \cup_{k=1}^{K} C_k$ 可聚合的，当且仅当对于任意的 $k, l \in \{1, 2, \ldots, K\}$，都有

$$p_{i,C_l} := \sum_{j \in C_l} p_{ij}$$

关于划分$S = \cup_{k=1}^{K} C_k$是分段常数。这些常数的元素为 $(\hat{p}_{kl}), k, l \in \{1, 2, \ldots, K\}$,为可聚合概率转移矩阵的元素。

定理 3.10. 假设转移概率矩阵 $\mathbf{P}$ 有 I 个独立的特征向量。令 $S = \cup_{k=1}^{K} C_k$ 为 S 的划分。那么马尔科夫链关于划分是可聚合的关于聚合的矩阵 $\{\hat{p}_{kl}\}_{k,l=1}^{K}$是非奇异的，当且仅当 $\mathbf{P}$ 有 K 个独立的特征向量关于划分是分段常数的且相关的特征值是非零的。

下面关于近似马尔科夫链的粗粒化过程。为此目的，我们定义了一个在状态 S 有不变分布 π 上的马尔科夫链的空间（随机空间）。让 $\mathbf{Q} = (q_{ij})_{i,j \in S}$是不变分布 π 上的生成元矩阵。我们定义它的模为

$$\|\mathbf{Q}\|_{\pi}^{2} = \sum_{i,j \in S} \frac{\pi_i}{\pi_j} q_{ij}^{2}.$$

如果 $\mathbf{Q}$ 满足细致平衡条件，那么这个模就是 $\mathbf{Q}$ 的特征值平方的和。

给定 $S : S = \cup_{k=1}^{K} C_k$ 的分割，如果 $k \neq l$, $C_k \cup C_l$, 让 $\hat{\mathbf{P}} = (\hat{p}_{kl})_{k,l=1}^{K}$ 是在状态空间的 $C = \{C_1, \cdots, C_K\}$ 上的随机矩阵。那么对应于原始空间上的随机矩阵有

$$\tilde{p}_{ij} = \frac{\pi_j \hat{p}_{kl}}{\sum_{k \in C_l} \pi_k}, \quad \text{if } i \in C_k, j \in C_l. \tag{9}$$

公式(9)描述了从任何在 C_k 的状态跳跃的概率是相同的，且随机游走者根据不变分布进入 C_l的状态。这于将原始状态粗粒化到新状态空间 $C = \{C_1, \cdots, C_K\}$ 并忽略集合 C_k 内的细致状态。注意如果$\hat{\mathbf{P}}$是在 C上有不变分布 $\hat{\pi}$,那么$\tilde{\mathbf{P}}$ 是随机矩阵是在 S 上的不变分布 π。

给定马尔科夫链 $\mathbf{P}$ 和预定义的划分 C, 发现最优粗粒链 $\tilde{\mathbf{P}}$ 是一个有用的问题。研究这个最优预测框架下关于模

$$\min_{\hat{\mathbf{P}} \cdot \mathbf{1} = \mathbf{1}, \hat{\mathbf{P}} \geq 0} \left\| \tilde{\mathbf{P}} - \mathbf{P} \right\|_{\pi}.$$

进一步可以问最优划分，那个划分使得上面误差更小。从代数的角度来看，网络的社区结构存在或聚类特性和 $\mathbf{P}$的特征值的谱隙的存在有关，也即 $\lambda_k = 1 - \eta_k \delta, k = 0, \cdots, K-1$, 且 $|\lambda_k| < \lambda^* < 1, k = K, \cdots, I-1$, 其中 $0 < \delta \ll 1$, $\eta_k > 0$.

4 Monte Carlo 方法

从实际角度来看，概率论的一个重要方向是：如何对概率分布进行采样。抽样对以下方面有很大应用。

- 计算积分。
- 优化。（模拟退火）
- 参数估计。（期望最大算法）

一个是采样，一个是基于数据推断分布。

4.1 数值积分

让 f 是 $[0,1]$上的连续函数，下面要通过概率方法计算积分 $I(f)=\int_0^1 f(x)dx$。基于大数法则，如果 $\{X_i\}_{i=1}^N$ 是在 $[0,1]$ 上独立同分布的随机变量，那么有

$$\int_0^1 f(x)dx=\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N f(X_i).$$

因此对于 $I(f)$ 有下面近似

$$I_N(f)=\frac{1}{N}\sum_{i=1}^N f(X_i).$$

对于误差 $e_N=I_N(f)-I(f)$，我们有估计

$$\mathbb{E}\left|e_N\right|^2=\frac{1}{N^2}\sum_{i,j=1}^N\mathbb{E}\left[(f(X_i)-I(f))(f(X_j)-I(f))\right]=\frac{1}{N}\mathbb{E}\left[(f(X_1)-I(f))^2\right]=\frac{1}{N}\mathrm{Var}(f).$$

因此有

$$I_N(f)-I(f)\approx\frac{\sigma_f}{\sqrt{N}},\tag{10}$$

其中 $\sigma_f=\sqrt{\mathrm{Var}(f)}$ 是标准差。我们将此与 N 个点的梯形误差界限比较为

$$e_N=\frac{1}{12}h^2\max_{x\in[0,1]}|f''(x)|,$$

其中 h 为网格间距大小. 在一维情况下，$h=\frac{1}{N}$，因此有如果 f 足够光滑，梯形法则显然有优势。

概率方法的真正意义在于高维问题。让 d 是维度， f 是在 $[0,1]^d$ 上的连续函数. 如果我们选取 N 个在 $[0,1]^d$ 上均匀分布的随机点且定义 $I_N(f)$ 与上面一致，那么误差界依然成立。特别地，对充分大的 N, 误差依然以 $N^{-1/2}$ 下降。而对于梯形方法为 $N^{-2/d}$, 因此当 $d>4$ 时Monte Carlo 更有优势。（虽然其他高阶方法能提高精度，但是对于高维问题，Monte Carlo 依然更有优势）

一个存在在统计物理中的重要高维积分为

$$\langle f\rangle=\int_{\mathbb{R}^{6n}}f(x)\pi(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

其中

$$\pi(\mathbf{x})=\frac{1}{Z}e^{-\beta H(\mathbf{x})}$$

是系统的平衡分布（Gibbs分布），Z 是配分函数，$H(\mathbf{x})$ 是哈密顿量，β 是温度的倒数。这个积分在计算期望值时非常重要，尤其是在量子场论和统计力学中。

4.2 随机变量的生成

均匀分布 对均匀分布$\mathcal{U}[0,1]$上伪随机数的生成的最常用算法为线性同余法（LCG），它有简单形式

$$X_{n+1}\equiv aX_n+b\mod m,$$

其中 a,b 和 m 是精心挑选的自然数，X_0 是种子（初始值）。序列 $\{X_n/m\}$ 给出了 $[0,1]$ 上的伪随机数。特别的这种递归是周期的，且有下面结果。

定理 4.1. 一个线性同余生成器的周期为 m，当且仅当

- b 和 m 互质。
- m 的每个质因子都能整除 $a-1$。
- 如果 $4|m$，则 $4|a-1$。

一般形式为

$$X_{n+1} = (a_0 X_n + a_1 X_{n-1} + \cdots + a_{j-1} X_{n-j+1} + b) \mod m.$$

对于线性同余生成器对s-元组（s-tuples）的效果是差的。s-元组是

$$(X_n, X_{n+1}, \ldots, X_{n+s-1})$$

的集合,他们是 s 个连续的伪随机数。

定理 4.2. 让 $\{Y_n\}$是由

$$X_{n+1} \equiv aX_n + b \mod m,$$

生成的序列，且 $X_n = Y_n/m$。那么 s-元组$(X_n, X_{n+1}, \ldots, X_{n+s-1})$位于 $[0,1]^s$ 上的超立方体内最多 $(s!m)^{1/s}$ 个等距平行的超平面上。

因此对于 $[0,1]^s$ 上的均匀分布，当 s 较大时，线性同余生成器的效果是差的。这可以通过使用更复杂的算法来解决，例如梅森旋转算法（Mersenne Twister）或其他伪随机数生成器。

各种统计检验被提出来测试伪随机数的质量。主要通过检验假设

(H0): 伪随机数序列 $\{X_1, X_2, \ldots, X_N\}$是独立同分布的满足$\mathcal{U}[0,1]$。通常由三种检验方法来检验伪随机数的质量：

- 等分布检验：检验伪随机数序列是否在 $[0,1]$ 上均匀分布。通常使用 χ^2 检验。
- 串行检验：基于4.2，检验 s-元组是否均匀分布。通常使用 s-元组的 χ^2 检验。
- 运行检验：根据上升段是服从正态分布。

逆变换方法 逆变换方法是从均匀分布 $\mathcal{U}[0,1]$ 上的随机变量生成其他分布的随机变量的一种方法。

命题 4.1. *(逆变换方法)* 设 F 是 X 的分布函数，也即 $F(x) = \mathbb{P}(X \le x)$。让 U 是在 $[0,1]$ 上均匀分布的随机变量。定义 F 的广义逆

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \ge u\}, \quad u \in (0,1).$$

那么 $X =: F^{-1}(U)$ 是一个随机变量，且有分布函数 F，对于 $U \in (0,1)$ 。

该方法对高斯分布效果不好，因为高斯分布的逆变换没有显式的形式。通常把高斯分布转化为均匀分布的方法是 Box-Muller 方法，（主要是使用极坐标）。

有时候对复杂的分布可以进行分解，例如将其表示为多个简单分布的组合。

接受拒绝方法 逆变化方法只对某些分布有效。接受拒绝方法是一个更通用的采样方法。它的基本思想是从一个容易采样的分布中生成样本，然后根据目标分布的形状来接受或拒绝这些样本。

算法 4.1. *(接受拒绝方法)*

Step 1 生成 $X_i \sim \mathcal{U}[0,1]$。

Step 2 生成一个决策变量 $Y_i \sim \mathcal{U}[0,d]$。

Step 3 如果 $0 \le Y_i \le p(X_i)$，则接受样本 X_i；否则拒绝样本。

Step 4 返回*Step 1*。

对于无界随机变量，使用更一般的比较函数法。

算法 4.2. *(接受拒绝方法)*

Step 1 生成 $X_i = F^{-1}(ZW_i)$,其中$W_i \sim \mathcal{U}[0,1]$。

Step 2 生成一个决策变量 $Y_i \sim \mathcal{U}[0,f(X_i)]$。

Step 3 如果 $0 \le Y_i \le p(X_i)$，则接受样本 X_i；否则拒绝样本。

Step 4 返回*Step 1*。

4.3 方差缩减

从式(10)可以看到，Monte Carlo 方法的误差是与 σ_f 成正比的。方差缩减方法是通过减少 σ_f 来提高 Monte Carlo 方法的效率。

重要抽样 重要抽样是通过改变采样分布来减少方差的一种方法。假设我们有一个目标分布 $p(x)$，我们希望从中采样。我们选择一个容易采样的分布 $q(x)$，并使用以下公式来计算期望值：

$$I(f) = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \mathbb{E}_q\left[f(X)\frac{p(X)}{q(X)}\right],$$

其中 $X \sim q(x)$。因此我们可以通过从 $q(x)$ 中采样来计算期望值：

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{p(X_i)}{q(X_i)},$$

其中 $X_i \sim q(x)$。重要抽样的关键是选择一个合适的 $q(x)$，使得 $\frac{p(x)}{q(x)}$ 的方差较小。

特别地，当 $p(x)$ 的归一化常数未知时，我们可以使用

$$\hat{I}_N(f) = \frac{\sum_{i=1}^{N} f(X_i)\frac{p(X_i)}{q(X_i)}}{\sum_{i=1}^{N}\frac{p(X_i)}{q(X_i)}}.$$

控制变量 控制变量方法是通过引入一个已知分布的随机变量来减少方差。假设我们有一个目标分布 $p(x)$，并且我们知道一个与 $p(x)$ 相关的分布 $g(x)$，其期望值 $\mathbb{E}_g[f(X)] = I(g)$. 我们可以使用以下公式来计算期望值：

$$I(f) = \mathbb{E}_p[f(X)] = \mathbb{E}_p[f(X) - g(X)] + I(g).$$

因此我们可以通过从 $p(x)$ 中采样来计算期望值：

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N}[f(X_i) - g(X_i)] + I(g),$$

其中 $X_i \sim p(x)$。控制变量方法的关键是选择一个合适的 $g(x)$，使得 $\mathbb{E}_p[f(X) - g(X)]$ 的方差较小。

Rao-Blackwellization Rao-Blackwellization 是一种通过引入条件期望来减少方差的方法。假设我们有 n 个独立的样本 $X_1, X_2, \ldots, X_n$来自分布 $\pi(\mathbf{x})$, 我们的兴趣是计算 $\int f(\mathbf{x})\pi(\mathbf{x})$. 直接方法是

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i).$$

假设 $\mathbf{x}$ 能分解为两部分 $\mathbf{x} = (\mathbf{x^{(1)}}, \mathbf{x^{(2)}})$，且如果条件期望 $\mathbb{E}(f(\mathbf{X})|\mathbf{x^{(2)}})$能被低成本高精度的近似。我们可以引入无偏估计：

$$\tilde{I} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[f(\mathbf{X})|\mathbf{X}_i^{(2)}].$$

如果直接计算和取条件计算的成本相似，我们能倾向于使用后者，因为

$$\mathrm{Var}(f(\mathbf{X})) = \mathbb{E}[\mathrm{Var}(f(\mathbf{X})|\mathbf{X}^{(2)})] + \mathrm{Var}(\mathbb{E}[f(\mathbf{X})|\mathbf{X}^{(2)}]),$$

因此

$$\mathrm{Var}(\hat{I}) \geq \mathrm{Var}(\tilde{I}).$$

4.4 Metropolis 算法

Metropolis 算法是一个重要的采样方法，特别是在统计物理和贝叶斯推断中。它是基于马尔科夫链蒙特卡洛（MCMC）方法的一种特殊情况。Metropolis 算法的基本思想是通过构造一个满足细致平衡条件的马尔科夫链，使其平稳分布为目标分布，从而实现对目标分布的采样。包括两个步骤：提议分布和接受-拒绝步骤。

算法 4.3. *(Metropolis 算法)*

Step 1 选择初始状态 X_0。

Step 2 生成候选状态 $X' \sim q(X'|X_t)$。

Step 3 计算接受概率

$$\alpha = \min\left(1, \frac{p(X')q(X_t|X')}{p(X_t)q(X'|X_t)}\right).$$

Step 4 生成均匀随机数 $U \sim \mathcal{U}[0,1]$，如果 $U < \alpha$，则接受候选状态 $X' = X_t$；否则拒绝，X_t 保持不变。

Step 5 返回 *Step 2*。

应用于贝叶斯推断 在贝叶斯推断中，我们通常需要从后验分布 $p(\theta|\mathbf{D})$ 中采样，其中 θ 是参数，$\mathbf{D}$ 是观测数据。Metropolis 算法可以用于从后验分布中采样。

动力学Monte Carlo方法

以Ising 模型为例，动力学Monte Carlo方法是通过模拟系统的时间演化来进行采样的一种方法。他是一种无拒绝的算法。首先把任意的configuation 根据每个spin的邻居的状态分为六类，分别是： 如果一个格点上的spin被反转，新构型和旧构型之间的能量差仅于spin的类别有关。我们定义翻转概率为

$$P_j = \min(1, \exp(-\beta\Delta H_j)), j = 1, 2, \cdots, 6,$$

Class	Spin	Number of neighboring spin ups
1	↑	2
2	↑	1
3	↑	0
4	↓	0
5	↓	1
6	↓	2

表 1: Ising 模型的状态分类

其中 ΔH_j 是类别 j 的spin翻转前后的能量差。

对于每个给定的状态 $\mathbf{x}$, 让 n_j 是类别 j 的格点数目。我们定义一个状态 $\mathbf{x}$ 的翻转率为

$$R(\mathbf{x}) = \sum_{j=1}^{6} n_j P_j,$$

且为了后面方便记 $Q_0 = 0$. 那么BKL算法的步骤为

算法 4.4. *(BKL算法)*

Step 1 生成 $R \sim \mathcal{U}[0, Q_6)$.

Step 2 确定翻转类别 $j(j = 1, 2, \cdots, 6\ )$，使得 $Q_{j-1} \le R < Q_j$。

Step 3 在第 j 类别中随机选择一个格点 i 翻转。

对于每个配置的时间增量，可以通过

$$\tau = -\frac{T_a}{Q_6} \log R, \quad R \sim \mathcal{U}[0, 1],$$

其中 T_a 为每个位点尝试翻转一次的物理时间。当我们通过上面算法求某个热力学平均时

$$\langle f \rangle_\pi = \frac{1}{\sum\limits_j \tau_j} \sum_j f(X_j)\tau_j,$$

由于对上面式子计算时， T_a 被消掉，因此我们可以简单选择 $T_a = 1$.

动力学Monte Carlo方法从数学上是通过连续时间的Q过程来替换离散时间马尔科夫链的结果。

Simulated Tempering

当我们对 Gibbs 分布 $\pi(x) \propto e^{-\beta H(x)}$ 进行采样时，或通过MCMC算法计算系综平均时，在低温 T下攀登能量景观中的势垒时，接受的概率通常会很小。Simulated Tempering 方法通过引入温度参数 β，在不同温度下进行采样，使系统能够在能量空间中更好地探索。具体做法是将温度作为一个附加的马尔科夫链状态，交替更新系统状态和温度，从而实现更高效的采样。

算法 4.5. *(Simulated Tempering)* 设置混合权重 $\alpha_0 \in (0, 1)$.

Step 1 给定 $(x_n, i_n) = (x, i)$, 采样 $u \sim \mathcal{U}[0, 1]$.

Step 2 如果 $u < \alpha_0$, 让 $i_{n+1} = i$且根据当前状态 x_n 采样 $x_{n+1} \sim \pi_{st}(x|i)$。

Step 3 如果 $u \ge \alpha_0$, 让 $x_{n+1} = x$且根据当前状态 i_n 采样 $i_{n+1} \sim \pi_{st}(i|x)$。

Simulated Annealing

Simulated Annealing 是一种基于随机采样的全局优化算法，灵感来自于物理学中的退火过程。其基本思想是在搜索空间中引入温度参数，通过控制温度的逐渐降低，允许算法在初始阶段接受较差的解，从而跳出局部最优，最终收敛到全局最优解。

让 H 是目标函数，模拟退火的想法是对于最优问题

$$\min_{x\in\mathcal{X}} H(x),$$

有对应的参数化的 Gibbs 分布

$$\pi_\beta(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)}, \quad Z(\beta) = \sum_{x\in\mathcal{X}} e^{-\beta H(x)} dx,$$

其中 β 是温度的倒数。定义 H 的全局最小值的集合

$$\mathcal{M} = \{x_0 : H(x_0) = \min_{x\in\mathcal{X}} H(x)\}.$$

定理 4.3. *对于这里Gibbs分布有以下极限*

$$\lim_{\beta\to\infty} \pi_\beta(x) = \begin{cases} \frac{1}{|\mathcal{M}|}, & x \in \mathcal{M}, \\ 0, & x \notin \mathcal{M}, \end{cases}$$

其中 $|\mathcal{M}|$ 是集合 $\mathcal{M}$ 的元素个数。如果 $x \in \mathcal{M}$，那么对于充分大的 β，$\pi_\beta(x)$ 关于 β 单调递增，如果 $x \notin \mathcal{M}$，那么对于充分大的 β，$\pi_\beta(x)$ 关于 β 单调递减。

在物理上，$\beta \to \infty$ 时，系统的温度趋近于零。这个过程称之为退火。一个直觉上的思考是，对于高温，系统的能量分布是平坦的，因此系统可以在整个状态空间中自由探索。随着温度的降低，系统的能量分布变得越来越集中，最终收敛到全局最优解。

让

$$P_\beta(x,y) = \begin{cases} Q(x,y)\frac{\pi_\beta(y)}{\pi_\beta(x)}, & \text{if } \pi_\beta(y) < \pi_\beta(x) \text{and } x \neq y, \\ Q(x,y), & \text{if } \pi_\beta(y) > \pi_\beta(x) \text{and } x \neq y, \\ 1 - \sum_{z\neq x} P_\beta(x,z), & \text{if } x = y, \end{cases}$$

其中 $Q(x,y)$ 是提议矩阵，假设是对称的。在退火过程中，温度逐渐降低，在第 n步 $\beta = \beta(n)$, 一个问题是如果选择 $\beta(n)$。

定理 4.4. *(模拟退火的收敛性) 假设 H 定义在有限集合 $\mathcal{X}$ 且Q 是一个对称不可约的提议矩阵。如果退火过程是选择使得 $\beta(n) \le C\log n$,其中 C仅依赖于 Q 和 H，那么对于任意的初始分布 μ_0, 有*

$$\lim_{n\to\infty} \left\|\mu_0 P_{\beta(1)} \cdots P_{\beta(n)} - \pi_\infty\right\| = 0,$$

其中 $\|\mu_1 - \mu_2\| =: \sum_x |\mu_1(x) - \mu_2(x)|$是总变差距离，且 π_∞ 是 $\beta \to \infty$ 时的平稳分布。

因为对于定理中，退火过程，如果 $\beta(n) = N_0 \gg 1$,那么 $n \sim O(\exp(N_0))$.这通常很难实现。一个更实用的选择是 $\beta(n) \sim p^{-n}, p \lesssim 1$。

5 随机过程

参考文献

[1] Weinan E, Tiejun Li, and Eric Vanden-Eijnden. *Applied Stochastic Analysis*. Number 199 in Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island, 2019.