MATH2105: Lecture Note on Stochastic Processes

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2024 年 4 月 1 日

Contents

1 Review of Probability Theory **4**

1.1 Probability Space . . . 4

1.1.1 Sample space . . . 4

1.1.2 Events . . . 4

1.1.3 Probability space . . . 5

1.2 Conditional Prob. . . . 5

1.3 Independent events . . . 6

1.4 Bayes' formula . . . 6

1.5 Discrete random variables . . . 7

1.5.1 the Bernoulli random vari. . . . 7

1.5.2 the Binomial random vari. . . . 7

1.5.3 the geometric random vari. . . . 8

1.5.4 the Poisson random vari. . . . 8

1.6 Cont. Random vari. . . . 9

1.6.1 Uniform random variable . . . 10

1.6.2 Exponential random vari. . . . 10

1.6.3 Gamma ranodm vari. . . . 11

1.6.4 Normal random vari. . . . 11

1.6.5 Inverse Gamma Random Variable . . . 11

1.6.6 History of Normal distribution . . . 12

1.7 Jointly distributed random variables . . . 12

1.7.1 independent random variables . . . 12

1.7.2 Covariance and Variance of Sums of Random Variables . . . 13

1.7.3 Sum of two independent variables . . . 14

1.8 Moment Generating Functions . . . 15

1.9 Limit Theorems . . . 17

2 Conditional Probability **20**

2.1 The Discrete Case . . . 20

2.2 The Continuous Case . . . 21

2.3 Computing Expectatoins by Conditioning . . . 22

2.4 Computing Probabilities by Conditioning . . . 26

2.5 Computing Conditional Expectation and Conditional Probability by Conditioning 28

3 Markov Chain **31**

3.1 Introduction 31
3.2 Chapman-Kolmogorov Equations 33
3.3 Unconditional Distribution of the State 35
3.4 Classification of States 36
3.4.1 recurrent and transient states and classes 37
3.4.2 Periodicity 43
3.4.3 positive and null recurrent states and classes 44
3.4.4 ergodicity 45
3.4.5 Classification 45
3.5 Stationary Distributions, Limiting Probabilities, Limiting Distributions 49
3.5.1 Stationary Distributions 49
3.5.2 The limiting behavior of transtion probability matrix 52
3.5.3 limiting probabilities and limiting distributions 52
3.5.4 Examples for stationary distributions and limiting probabilities 53
3.6 Mean Time Spent in Transient States 55
3.7 Time Reversible Markov Chains 60
3.8 Hidden Markov Model 64
3.8.1 Introduction to HMM 64
3.8.2 Key ingredients to HMM 66
3.8.3 Three Basic Problems for HMMs 66
3.8.4 Forward and Backward Approaches for Problem 1 67

4 Markov Chain Monte Carlo (MCMC) **69**

4.1 Monte Carlo Methods 69
4.1.1 deterministic vs. stochastic 69
4.1.2 background 69
4.1.3 difference bw some terminology 70
4.1.4 statistical mechanics 70
4.1.5 applications 70
4.2 Generation of Random Numbers 73
4.2.1 generation of random numbers from $U(0,1)$ 73
4.2.2 generation of random numbers with general distributions 74
4.2.3 technique for reducing the variance 76
4.3 Introduction to Markov Chain Monte Carlo (MCMC) 79
4.3.1 Application of MCMC 79
4.3.2 Metropolis-Hasting Algorithm 81
4.3.3 Gibbs sampling 85
4.4 Parameter estimation problems 87
4.4.1 The Bayesian approach 87

4.4.2 Applications of Bayesian statistics 89

5 Poisson Process **91**

6 Continuous-Time Markov Chain **92**

7 Brownian Motion and Stationary Process **93**

8 Time Series **94**

Chapter 1

Review of Probability Theory

Randomness should be taken into account in data, model, equations (PDEs), etc. This can be realized by allowing the model to be probabilistic in nature, which is referred to as a probability model. The reference book is [10].

1.1 Probability Space

A probability theory is made up of three part, $(\Omega, \mathcal{F}, P)$.

1.1.1 Sample space

Ω is a sample space.

Example 1.1.1 $\Omega = \{H(ead), T(ail)\}$ *for a coin flipping.*

Example 1.1.2 $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ *for flipping two coins.*

Example 1.1.3 $\Omega = \{1, 2, \ldots, 6\}$ *for rolling of a die.*

Example 1.1.4 $\Omega = \left\{ \begin{array}{ccc} (1,1) & \cdots & (1,6) \\ \vdots & & \vdots \\ (6,1) & \cdots & (6,6) \end{array} \right\}$ *for rolling of two dice.*

1.1.2 Events

Subset E of Ω is known as an event.

Example 1.1.5 $E = \{H\}$

Example 1.1.6 $E = \{2,4,6\}$. *Even number appears.*

- union of events. Given $E_1 = \{1,3,5\}$ and $E_2 = \{1,2,3\}$, then $E_1 \cup E_2 = \{1,2,3,5\}$.
- intersection of events. Given above, then $E_1E_2 := E_1 \cap E_2 = \{1,3\}$.
- complement. $E_1^c = \{2,4,6\}$.
- mutually exclusive. E, F, G are called mutually exclusive if $EF = \varnothing, EG = \varnothing, FG = \varnothing$. For example, $E = \{1,2\}, F = \{3,4\}, G = \{5,6\}$ are mutually exclusive.

Definition 1.1.7 *$\mathcal{F}$ is a family of subsets of Ω satisfying:*
(1) $\Omega \in \mathcal{F}$.
(2) $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$.
(3) $E_j \in \mathcal{F} \Rightarrow \bigcup_{j=1}^{\infty} E_j \in \mathcal{F}$.
Then $\mathcal{F}$ is a σ-albegra of Ω. $\Omega, \mathcal{F}$ is called a measurable space.

1.1.3 Probability space

Definition 1.1.8 *P is a function defined on satisfying:*
(1) non-negative. $0 \leq P(E) \leq 1, \forall E \in \mathcal{F}$.
(2) completeness. $P(\Omega) = 1$.
(3) For any countable mutually exclusive sets in $\mathcal{F}$, $P\big(\bigcup_{j=1}^{\infty} E_j\big) = \sum_{j=1}^{\infty} P(E_j)$.
Then $P(E)$ is the prob. of E. $(\Omega, \mathcal{F}, P)$ is the triple elements of a prob. space.

Example 1.1.9 $P(\{H\}) = P(\{T\}) = \frac{1}{2}$.

Example 1.1.10 $P(\{1\}) = \cdots = P(\{6\}) = \frac{1}{6}$.
$P(\{1,3,5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{2}$.

Example 1.1.11 $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1E_2) - P(E_1E_3) - P(E_2E_3) + P(E_1E_2E_3)$.

1.2 Conditional Prob.

Definition 1.2.1

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Example 1.2.2 *Choose one number from 1-10. The number is at least five, then what is the cond. prob. that it is ten?*
Sol: Let $E = \{10\}$ and $F = \{\geq 5\}$. Then

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{\frac{1}{10}}{\frac{6}{10}} = \frac{1}{6}.$$

Example 1.2.3 *An urn contains 7 black balls and 5 white balls. Draw two balls without replacement. Each ball is equally drawn. What is prob. that both drawn balls are black?*
Sol: F={first ball is black}, E={2nd ball is black}. Since $P(E|F) = \frac{6}{11}, P(F) = \frac{7}{12}$, then

$$P(EF) = P(F)P(E|F) = \frac{7 * 6}{12 * 11}.$$

Another solution is given directly by

$$P(\text{both black}) = \frac{C_7^2}{C_{12}^2} = \frac{7 * 6}{12 * 11}.$$

1.3 Independent events

$$\begin{aligned} E, F \text{ areindependent} &\Leftrightarrow P(EF) = P(E)P(F) \\ &\Leftrightarrow P(E|F) = P(E), \quad P(F) \neq 0 \\ &\Leftrightarrow P(F|E) = P(F), \quad P(E) \neq 0 \end{aligned}$$

Example 1.3.1 *Let a ball be drawn from an urn containing 4 balls {1,2,3,4}. Let* $E = \{1, 2\}$, $F = \{1, 3\}$, $G = \{1, 4\}$. *Then*

$$\begin{aligned} P(EF) = P(E)P(F) = \frac{1}{4}, \\ P(EG) = P(E)P(G) = \frac{1}{4}, \\ P(FG) = P(F)P(G) = \frac{1}{4}. \end{aligned}$$

However,

$$\frac{1}{4} = P(EFG) \neq P(E)P(F)P(G) = \frac{1}{8}.$$

E, F, G *are not jointly independ.*

1.4 Bayes' formula

- sub-additive.

$$P(\bigcup_{j=1}^{\infty} E_j) \leq \sum_{j=1}^{\infty} P(E_j).$$

- multiplicity formula.

$$P(B_1 B_2 \dots B_n) = P(B_1)P(B_2|B_1) \cdots P(B_n|B_1 B_2 \dots B_{n-1}).$$

• formula of total probability. Suppose that $F_1, \dots, F_n$ are mutually exclusive events such that $\cup_{i=1}^n F_i = \Omega$ (sample space) or $E \subset \cup_{i=1}^n F_i$, then $E = \cup_{i=1}^n EF_i$. Noticing EF_i are mutually exclusive, we obtain that

$$P(E) = \sum_{j=1}^{n} P(EF_j) = \sum_{j=1}^{n} P(E|F_j)P(F_j).$$

In particular, if $n = 2$, then $E = EF \cup EF^c$ and

$$P(E) = P(EF) + P(EF^c) = P(E|F)P(F) + P(E|F^c)P(F^c).$$

Another formula is: suppose that $\cup_{i=1}^n C_i = \Omega$ (sample space) or $B \subset \cup_{i=1}^n C_i$, then if $P(A) > 0$,

$$P(B|A) = \sum_{j=1}^{n} P(C_j|A)P(B|AC_j).$$

Pf.

$$\begin{aligned} \sum_{j=1}^{n} P(C_j|A)P(B|AC_j) &= \sum_{j=1}^{n} \frac{P(C_j A)}{P(A)} \frac{P(BAC_j)}{P(AC_j)} = \sum_{j=1}^{n} \frac{P(BAC_j)}{P(A)} \\ &= \frac{P(BA)}{P(A)} = P(B|A). \end{aligned}$$

• Bayes' formula.

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{j=1}^{n} P(E|F_j)P(F_j)}.$$

Example 1.4.1 *a multiple-choice test. Let p be the prob. that the student knows the answer. Assume that a student who guesses at the answer will be correct with prob. $1/m$, where there are m choices. What is the prob. that a student knew the answer given that she answered it correctly?*
Sol: Let C and K be "correct" and "know", respectively.

$$\begin{aligned} P(K|C) &= \frac{P(KC)}{P(C)} = \frac{P(C|K)P(K)}{P(C|K)P(K) + P(C|K^c)P(K^c)} \\ &= \frac{p}{p + (1/m)(1-p)}. \end{aligned}$$

If $m = 5$ and $p = \frac{1}{2}$, $P(K|C) = 5/6$.

1.5 Discrete random variables

Definition 1.5.1 *If X is discrete with prob. mass function $p(x)$, then for any real-valued function g, the expectation is defined as*

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

1.5.1 the Bernoulli random vari.

An experiment, whose outcome is either a success or a failure. $X = 1$ is a success and $X = 0$ is a failure. Then X is denoted as $X \sim B(1, p)$ and the pmf is

$$\begin{aligned} p(0) &= P(X = 0) = 1 - p, \\ p(1) &= P(X = 1) = p. \end{aligned}$$

Its expect and var is

$$\begin{aligned} EX &= 1 \cdot p + 0 \cdot q = p, \\ Var(X) &= EX^2 - (EX)^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p(1-p). \end{aligned}$$

1.5.2 the Binomial random vari.

Suppose there are n trials of Bernoulli experiments. That is, If $X_1, \ldots, X_n$ are samples from $B(1, p)$, then
(1) $Y = X_1 + \cdots + X_n \sim B(n, p)$.
(2) pdf is given by $p(i) = C_n^i p^i (1-p)^{n-i}$, $i = 0, \ldots, n$.
(3) $EY = \sum_{i=0}^{n} ip(i) = \sum_{i=0}^{n} iC_n^i p^i q^{n-i} = \sum_{i=1}^{n} i\frac{n!}{i!(n-i)!}p^i q^{n-i} = np\sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!}p^{i-1}q^{n-i} = np$.
$Var(Y) = npq$.

Example 1.5.2 *Suppose each independent engine of an airplane will fail, when in flight, with prob.* $1-p$. *Suppose that the airplane will make a successful flight if at least 50 percent of its engines remain operative. For what values of* p *is a four-engine plane preferable to a two-engine plane?*
Sol: A four-engine plane makes a successful flight with prob.

$$\begin{aligned} & C_4^2 p^2(1-p)^2 + C_4^3 p^3(1-p)^1 + C_4^4 p^4(1-p)^0 \\ = \;& 6p^2(1-p)^2 + 4p^3(1-p) + p^4. \end{aligned}$$

The prob. for a two-engine plane is

$$C_2^1 p^1(1-p)^1 + C_2^2 p^2(1-p)^0 = 2p(1-p) + p^2.$$

Hence a four-engine plane is safer if

$$\begin{aligned} 6p^2(1-p)^2 + 4p^3(1-p) + p^4 &\geq 2p(1-p) + p^2 \\ 6p^3 - 12p^2 + 6p + 4p^2 - 4p^3 + p^3 &\geq 2 - p \\ 3p^3 - 8p^2 + 7p - 2 &\geq 0 \\ (p-1)^2(3p-2) &\geq 0. \\ p &\geq \frac{2}{3}. \end{aligned}$$

1.5.3 the geometric random vari.

Suppose that independent trials, each having prob. p of being a success, are performed until a success occurs. Let X be the number of trials required until the first success. The pmf is given by

$$p(n) = P(X = n) = (1-p)^{n-1}p, \quad n = 1, 2, \ldots$$

To check it is a pmf

$$\sum_{n=1}^{\infty} p(n) = p\sum_{n=1}^{\infty}(1-p)^{n-1} = \frac{p}{1-(1-p)} = 1.$$

The expect. and var is

$$\begin{aligned} EX &= \sum_{i=1}^{\infty} iq^{i-1}p = p\sum_{i=1}^{\infty}\frac{dq^i}{dq} = p\left(\frac{q}{1-q}\right)' = \frac{p}{(1-q)^2} = \frac{1}{p}. \\ Var(X) &= \frac{1-p}{p^2}. \end{aligned}$$

1.5.4 the Poisson random vari.

X is said to be a Poisson random vari. with parameter λ, denoted by $X \sim P(\lambda)$,

$$p(i) = P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}, i = 0, 1,$$

Check it is pmf

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1.$$

(1) $EX = \lambda, Var(X) = \lambda$.
(2) If $X_1, \ldots, X_n$ are indepen, $X_i \sim P(\lambda_i)$, then

$$X_1 + \cdots + X_n \sim P(\lambda_1 + \cdots + \lambda_n).$$

Pf. By induction.

$$\begin{aligned} P(X_1 + X_2 &= i) = \sum_{k=0}^{i} P(X_1 = k)P(X_2 = i-k) \\ &= \sum_{k=0}^{i} e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{i-k}}{(i-k)!} = e^{-(\lambda_1+\lambda_2)}\frac{1}{i!}\sum_{k=0}^{i}\frac{i!}{k!(i-k)!}\lambda_1^k\lambda_2^{i-k} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{1}{i!}(\lambda_1+\lambda_2)^i \sim P(\lambda_1+\lambda_2). \end{aligned}$$

Example 1.5.3 *Suppose that the number of typo errors on a single page of a book has a Poisson distr. with parameter $\lambda = 1$. Calculate the prob. that there is at least one error on this page.*
Sol: $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1} = 0.633$.

1.6 Cont. Random vari.

Definition 1.6.1 *The cumulative distribution function (cdf) (or sometimes just distribution function) $F(\cdot)$ is defined by, $F(b) = P(X \leq b)$, satisfying (i) $F(b)$ is a nondecreasing function of b, (ii) $\lim_{b\to\infty} F(b) = F(\infty) = 1$, (iii) $\lim_{b\to-\infty} F(b) = F(-\infty) = 0$.*

One can see obviously that

$$P(a \leq X \leq b) = F(b) - F(a), \quad \text{for all } a < b.$$

Definition 1.6.2 *If there exists a nonnegative function $f(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B,*

$$P(X \in B) = \int_B f(x)dx.$$

The function $f(x)$ is called the prob. density function (pdf) of X.

The relation bw the cdf F and the pdf f is

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx,$$

and

$$\frac{dF(a)}{da} = f(a).$$

Definition 1.6.3 *If X is cont. random vari. with pdf $f(x)$, then for any real-valued function g, its expectation is defined by*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

1.6.1 Uniform random variable

Definition 1.6.4 *A random vari X is said to be uniformly distributed over $(0,1)$, if its pdf is given by*

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

Denote by $X \sim \mathcal{U}(0,1)$. Its expectation and variance are

$$EX = \int_0^1 xf(x)dx = \frac{1}{2}, \quad Var(X) = EX^2 - (EX)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

1.6.2 Exponential random vari.

Definition 1.6.5 *A cont. random vari. whose pdf is given, for some $\lambda > 0$, by*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \geq 0, \\ 0, & if\ x < 0. \end{cases}$$

is said to be an exponential random variable with rate parameter λ. Denote by $X \sim \mathcal{E}(\lambda)$.

The cdf can be calculated by

$$\begin{aligned} F(a) &= \int_0^a \lambda e^{-\lambda x}dx = 1 - e^{-\lambda a}, \quad a \geq 0. \\ F(\infty) &= \int_0^\infty \lambda e^{-\lambda x}dx = 1. \end{aligned}$$

(1) $EX = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$.
(2) $P(X > t) = e^{-\lambda t}, t \geq 0$.
(3) Y is an exponential random vari. if and only if $EY > 0$ and for $\forall s, t > 0$, such that

$$P(Y > s + t | Y > s) = P(Y > t),$$

where this condition is called memoryless. Denote $\overline{F}(t) = P(Y > t)$, then above Eq. is equivalent to

$$\overline{F}(t + s) = \overline{F}(t)\overline{F}(s).$$

Pf. One can easily check $\Rightarrow$ by noticing that $\overline{F}(t) = e^{-\lambda t}$. For the opposite direction, we want to prove $\overline{F}(t)$ is an exponential function. We first prove if $f(t + s) = f(t) + f(s)$, then f is linear function. For integers t and s, we have $f(n) = nf(1)$. For rational numbers t and s, $qf(p/q) = f(p) = pf(1)$, then $f(p/q) = p/qf(1)$. Since rational numbers are dense in real numbers, one can show $f(x) = xf(1)$ for all x real. Finally, $\overline{F}(t) = e^{f(t)} = e^{tf(1)}$, which is an expon. function.

Example 1.6.6 *Suppose a clock or a watch has a lifetime with exponential distribution with expectation 1 year. If it already works for 2 months, what's its remaining lifetime? (1 year since memoryless).*

Example 1.6.7 *Assume that the customer comes with interarrival time being exponential dist. If a cashier wants to go washroom, he/she goes right now or later on? (Right now since memoryless).*

1.6.3 Gamma ranodm vari.

Definition 1.6.8 *A cont. random vari. whose pdf is given, for some $\lambda > 0, \alpha > 0$, by*

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

is said to be a gamma random variable with rate parameters λ and α. Denote by $X \sim \Gamma(\alpha, \lambda)$. A Gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx.$$

The expectation and var of gamma vari is given by

$$EX = \frac{\alpha}{\lambda}, Var(X) = \frac{\alpha}{\lambda^2}.$$

1.6.4 Normal random vari.

Definition 1.6.9 *X is normal random vari. with parameters μ and σ^2 if the density of X is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty.$$

The density is bell-shaped curve that is symmetric around μ.

Definition 1.6.10 *multivariate normal distribution. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^T$. If $X = \mu + B\varepsilon$, then*

$$X \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma = BB^T$ is the covariance matrix of X.

(1) $X = (X_1, \ldots, X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ if and only if $\forall a_1, \ldots, a_n, \sum_{j=1}^n a_j X_j$ is normally distributed.
(2) Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then $X_1, \ldots, X_n$ are independent if and only if they are uncorrelated, i.e., $Cov(X_i, X_j) = 0$ for $i \neq j$. The proof can be found following.

Many real-world quantities tend to be normally distributed—for instance, human heights and other body measurements, cumulative hydrologic measures such as annual rainfall or monthly river discharge, errors in astronomical or physical observations, and diffusion of a substance in a liquid or gas. Some things are products of many independent variables (rather than sums), and in such cases the logarithm will be approximately normal since it is a sum of many independent variables—this is often the case for economic quantities such as stock market indices, due to the effect of compound interest.

1.6.5 Inverse Gamma Random Variable

If X is Gamma distributed then the distribution of $1/X$ is called the Inverse Gamma distribution. More precisely, if $X \sim \text{Gamma}(a, b)$ and $Y = 1/X$ then $Y \sim \text{InvGamma}(a, b)$, and the p.d.f. of Y is

$$\text{InvGamma}(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-b/y). \tag{1.1}$$

So, putting a $\text{Gamma}(a, b)$ prior on the precision λ is equivalent to putting an $\text{InvGamma}(a, b)$ prior on the variance $\sigma^2 = 1/\lambda$. The Inverse Gamma can be used to define a NormalInvGamma distribution for use as a prior on (μ, σ^2), which is sometimes more convenient than (but equivalent to) using a NormalGamma prior on (μ, λ).

1.6.6 History of Normal distribution

Carl Friedrich Gauss

James Clerk Maxwell

Adolphe Quetelet

In 1809, Carl Friedrich Gauss (1777–1855) proposed the normal distribution as a model for the errors made in astronomical measurements, as a formal way of justifying the use of the sample mean, by showing it to be the most likely estimate—that is, the maximum likelihood estimate—of the true value (and more generally, to justify the method of least squares in linear regression). With astonishing speed, following Gauss' proposal, Laplace proved the central limit theorem in 1810. Laplace also calculated the normalization constant of the normal distribution, which is not a trivial task. James Clerk Maxwell (1831–1879) showed that the normal distribution arose naturally in physics, particularly in thermodynamics. Adolphe Quetelet (1796–1874) pioneered the use of the normal distribution in the social sciences. (See Fig. 1.6.6.)

1.7 Jointly distributed random variables

1.7.1 independent random variables

Definition 1.7.1 *The random variables X and Y are said to be independent if, for all a, b,*

$$P(X \le a, Y \le b) = P(X \le a)P(Y \le b).$$

In terms of the joint distribution function F, we have that

$$F(a, b) = F_X(a)F_Y(b) \quad \text{for all } a, b.$$

Corollary 1.7.2 *When X and Y are discrete, the condition of indep. reduces to*

$$p(x, y) = p_X(x)p_Y(y).$$

If X and Y are jointly continuous, independence reduces to

$$f(x, y) = f_X(x)f_Y(y).$$

Pf.

$$\begin{aligned} P(X \le a, Y \le b) &= \sum_{y \le b} \sum_{x \le a} p(x, y) = \sum_{y \le b} \sum_{x \le a} p_X(x)p_Y(y) \\ &= \sum_{y \le b} p_Y(y) \sum_{x \le a} p_X(x) = P(Y \le b)P(X \le a). \end{aligned}$$

If X and Y are independ., then for any h and g

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Pf.

$$E[g(X)h(Y)] = \sum_y \sum_x g(x)h(y)p(x,y) = \sum_y \sum_x g(x)h(y)p_X(x)p_Y(y) = E[g(X)]E[h(Y)].$$

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy = E[g(X)]E[h(Y)]. \end{aligned}$$

Example 1.7.3 ***(Variance of a Binomial Random Variable)*** *Compute the Variance of a Binomial Random Variable. Sol. Binomial is the sum of n indep. Bernoulli.*

$$Var(X) = Var(X_1) + \cdots + Var(X_n) = npq,$$

since $Var(X_i) = pq$ for each Bernoulli distribution.

1.7.2 Covariance and Variance of Sums of Random Variables

Definition 1.7.4 *The covariance of any two random vari. is*

$$Cov(X,Y) = E[(X - EX)(Y - EY)] = E[XY] - E[X]E[Y].$$

If X and Y are independent, then $Cov(X,Y) = 0$.

Corollary 1.7.5 *Property of Covariance*

(1) $Cov(X,X) = Var(X)$,

(2) $Cov(X,Y) = Cov(Y,X)$,

(3) $Cov(cX + dZ, Y) = cCov(X,Y) + dCov(Z,Y)$.

A useful expression for the variance can be found as follows:

$$\begin{aligned} Var(\sum_{i=1}^n X_i) &= Cov(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j) = \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j) \\ &= \sum_{i=1}^n Cov(X_i, X_i) + 2\sum_{i<j} Cov(X_i, X_j). \end{aligned}$$

Moreover, if X_i are indep. random variables, then above equation reduces to

$$Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n Var(X_i).$$

Definition 1.7.6 *If $X_1, \ldots, X_n$ are i.i.d., then the random variable $\overline{X} = \sum_{i=1}^n X_i/n$ is called the* ***sample mean****.*

Proposition 1.7.7 *Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. Then*

(a) $E[\overline{X}] = \mu$.

(b) $Var(\overline{X}) = \sigma^2/n$.

(c) $Cov(\overline{X}, X_i - \overline{X}) = 0, i = 1, \ldots, n$.

Pf. Parts (a) and (b) are easy:

$$\begin{aligned} E[\overline{X}] &= \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu, \\ Var[\overline{X}] &= \left(\frac{1}{n}\right)^2 Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}. \end{aligned}$$

To prove (c), we follow

$$\begin{aligned} Cov(\overline{X}, X_i - \overline{X}) &= Cov(\overline{X}, X_i) - Cov(\overline{X}, \overline{X}) = \frac{1}{n} Cov(X_i + \sum_{j \neq i} X_j, X_i) - Var[\overline{X}] \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0. \end{aligned}$$

Proposition 1.7.8 *The **sample variance** is given by*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then it is unbiased, that is,

$$ES^2 = \sigma^2.$$

Pf. Notice that

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \overline{X})^2. \end{aligned}$$

Then we obtain

$$\begin{aligned} E[(n-1)S^2] &= \sum_{i=1}^{n} E(X_i - \overline{X})^2 = \sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\mu - \overline{X})^2 \\ &= n\sigma^2 - nVar[\overline{X}] = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \end{aligned}$$

1.7.3 Sum of two independent variables

Let us derive the formula first. Suppose that X and Y are continuous and independent, X having pdf f and Y having pdf g. Letting $F_{X+Y}(a)$ be the cdf of $X + Y$, we have

$$\begin{aligned} F_{X+Y}(a) &= P(X + Y \leq a) = \iint_{x+y \leq a} f(x)g(y)dxdy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f(x)dx\right) g(y)dy = \int_{-\infty}^{\infty} F_X(a-y)g(y)dy. \end{aligned}$$

By differentiating above, we obtain the pdf $f_{X+Y}(a)$ of $X+Y$ given by

$$f_{X+Y}(a) = \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)g(y)dy = \int_{-\infty}^{\infty} f(a-y)g(y)dy.$$

Thus f_{X+Y} is the **convolution** of functions f and g.

Example 1.7.9 *Two uniform random vari. If X and Y are indepdt. both uniformly distributed on $(0,1)$, then calculate the pdf of $X+Y$.*

Sol. The pdf's are

$$f(a) = g(a) = \begin{cases} 1, & 0 < a < 1, \\ 0, & \text{otherwise.} \end{cases}$$

we obtain

$$f_{X+Y}(a) = \int_0^1 f(a-y)g(y)dy.$$

For $0 \le a \le 1$, this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

since $0 \le a-y \le 1$ and $0 \le y \le 1 \Rightarrow 0 \le y \le a$. And for $1 \le a \le 2$, this yields

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2-a$$

since $0 \le a-y \le 1$ and $0 \le y \le 1 \Rightarrow a-1 \le y \le 1$. Hence,

$$f_{X+Y}(a) = \begin{cases} a, & 0 < a < 1, \\ 2-a & 1 < a < 2 \\ 0, & \text{otherwise.} \end{cases}$$

1.8 Moment Generating Functions

Definition 1.8.1 *The moment generating function (MGF) $\phi(t)$ of random variables X is defined by*

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx}p(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx}f(x)dx, & \text{if } X \text{ is continuous} \end{cases}$$

It is called MGF because all moments of X can be obtained by successively differentiating $\phi(t)$. For example,

$$\phi'(t) = \frac{d}{dt}E[e^{tX}] = E[\frac{d}{dt}e^{tX}] = E[Xe^{tX}].$$

Hence $\phi'(0) = EX$. Similarly,

$$\phi''(t) = \frac{d}{dt}E[Xe^{tX}] = E[\frac{d}{dt}Xe^{tX}] = E[X^2e^{tX}],$$

so that $\phi''(0) = EX^2$. One can show that $\phi^{(n)}(0) = E[X^n]$ for $n \ge 1$.

Example 1.8.2 ***(Poisson Distribution with mean λ).***

$$\begin{aligned} \phi(t) &= E[e^{tX}] = \sum_{n=0}^{\infty} \frac{e^{tn}e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(e^t\lambda)^n}{n!} = e^{-\lambda}e^{\lambda e^t} = \exp[\lambda(e^t-1)]. \\ \phi'(t) &= \lambda e^t \exp[\lambda(e^t-1)], \\ \phi''(t) &= \left(\lambda e^t\right)^2 \exp[\lambda(e^t-1)] + \lambda e^t \exp[\lambda(e^t-1)], \end{aligned}$$

Thus,

$$\begin{aligned} EX &= \phi'(0) = \lambda, EX^2 = \phi''(0) = \lambda^2 + \lambda, \\ Var(X) &= \lambda. \end{aligned}$$

Example 1.8.3 ***Exponential Distribution with parameter*** λ.

$$\begin{aligned} \phi(t) &= E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - t} \text{ for } t < \lambda \\ \phi'(t) &= \frac{\lambda}{(\lambda - t)^2}, \phi''(t) = \frac{2\lambda}{(\lambda - t)^3}. \end{aligned}$$

Hence,

$$EX = \frac{1}{\lambda}, EX^2 = \frac{2}{\lambda^2}, Var(X) = \frac{1}{\lambda^2}.$$

Example 1.8.4 ***Normal distribution with mean*** μ ***and variance*** σ^2. *Compute the MGF of a standard normal random variable* Z *as follows.*

$$E[e^{tZ}] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{tx}e^{-x^2/2}dx = e^{t^2/2}.$$

If $X = \sigma Z + \mu$ *is normal, then*

$$\begin{aligned} \phi(t) &= E[e^{tX}] = e^{t\mu}E[e^{t\sigma Z}] = \exp\{\frac{\sigma^2 t^2}{2} + \mu t\}. \\ \phi'(t) &= (\mu + \sigma^2 t)\exp\{\frac{\sigma^2 t^2}{2} + \mu t\} \\ \phi''(t) &= (\mu + \sigma^2 t)^2\exp\{\frac{\sigma^2 t^2}{2} + \mu t\} + \sigma^2\exp\{\frac{\sigma^2 t^2}{2} + \mu t\} \end{aligned}$$

Hence $EX = \mu, Var(X) = \sigma^2$.

Theorem 1.8.5 *A MGF uniquely determines a probability distribution.*

Theorem 1.8.6 $X_1, \ldots, X_n$ *are independent if and only if their MGFs satisfy*

$$Ee^{\vec{t}\cdot\vec{X}} := \phi(t_1, \ldots, t_n) = \prod_{i=1}^n \phi_i(t_i) := \prod_{i=1}^n Ee^{t_i X_i}.$$

Theorem 1.8.7 X_i *are independent and have MGFs* $\phi_i(t)$. *Then* $Y = X_1 + \cdots + X_n$ *has the MGF*

$$\phi_Y(t) = \prod_{i=1}^n \phi_i(t).$$

Example 1.8.8 *Using the results from above the example to show that* $Cov(X_i, X_j) = 0$ *for multivariate Gaussian distribution* $\Rightarrow$ *they are independent.*
Sol. Let $\vec{X} \sim \mathcal{N}(\mu, \Sigma)$, *where* $\vec{X} = (X_1, \ldots, X_n) = \vec{\mu} + B\vec{\epsilon}$ *and* $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. *(Think about why there is such a linear transformation? since rank of* Σ *is* n.*) One has the pdf*

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

Since we have computed that $\phi_{\vec{\epsilon}}(\vec{t}) = E[e^{\vec{t}\cdot\vec{\epsilon}}] = \prod_{i=1}^n e^{t_i^2/2} = e^{\vec{t}\cdot\vec{t}/2}$, *so that*

$$\begin{aligned} \phi_{\vec{X}}(\vec{t}) &= E[e^{\vec{t}\cdot\vec{X}}] = Ee^{\vec{t}\cdot\vec{\mu} + \vec{t}^T B\vec{\epsilon}} = e^{\vec{t}^T\vec{\mu} + \frac{1}{2}t^T BB^T t} = e^{\vec{t}^T\vec{\mu} + \frac{1}{2}t^T\Sigma t} \\ &= \prod_{i=1}^n e^{t_i\mu_i + \frac{1}{2}t_i^2\sigma_i^2} = \prod_{i=1}^n \phi_i(t_i). \end{aligned}$$

Example 1.8.9 *Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and X and Y are independent. Then*

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = e^{t\mu_1 + \frac{1}{2}t^2\sigma_1^2} e^{t\mu_2 + \frac{1}{2}t^2\sigma_2^2} = e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2},$$

so that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

1.9 Limit Theorems

Proposition 1.9.1 ***(Markov's Inequality)*** *If X is a random variable that takes only nonnegative values, then for any $a > 0$*

$$P\{X \geq a\} = \frac{E[X]}{a}.$$

Pf. We give a proof for the case where X is continuous with denfity f,

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx = \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\ &\geq \int_a^\infty xf(x)dx \geq a\int_a^\infty f(x)dx = aP\{X \geq a\}. \end{aligned}$$

Proposition 1.9.2 ***(Chebyshev's Inequality)*** *If X is a random variable with mean μ and variance σ^2, then, for any $k > 0$,*

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}.$$

Pf. We apply Markov's inequality to the nonnegative $(X - \mu)^2$,

$$P\{(X-\mu)^2 \geq k^2\} \leq \frac{E[(X-\mu)^2]}{k^2}.$$

Remark 1.9.3 *The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probs. when only the mean, or both the mean and the variance, are known. Of course, if the true distribution were known, then the desired probs. could be exactly computed, and we would not need to resort to bounds.*

Theorem 1.9.4 ***(Strong Law of Large Numbers)*** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables have a commom distribution, and let $E[X_t] = \mu$. Then, with probability 1, or almost surely,*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty.$$

Definition 1.9.5 *If $P(\lim_{n\to\infty} X_n = X) = 1$, then we say $X_n \to X$, a.s. (almost surely) or $X_n \to X$, w.p.1 (with probability 1).*

Theorem 1.9.6 ***(Central Limit Theorem)*** *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed (i.i.d.) random variables, each with mean μ and variance σ^2. Then the distribution of*

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

goes to the standard normal as $n \to \infty$. That is,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^a e^{-x^2/2}dx = \Phi(a),$$

as $n \to \infty$

Pf. Note that the theorem holds for any distribution of the X_is; herein lies its power.

We now present a heuristic proof the CLT. Suppose first that the X_i have mena 0 and variance 1, and then the MGF can be computed,

$$E\left[\exp\left\{t\frac{X_1+\cdots+X_n}{\sqrt{n}}\right\}\right] = E[e^{tX_1/\sqrt{n}}\cdots e^{tX_n/\sqrt{n}}] = (Ee^{tX_i/\sqrt{n}})^n \text{ by independence.}$$

For large n, we obtain by Taylor expansion,

$$e^{tX_i/\sqrt{n}} = 1 + \frac{tX_i}{\sqrt{n}} + \frac{(tX_i)^2}{2n} + O(n^{-3/2}),$$

that is the reason for the **central** word. Taking expectations shows that when n is large,

$$E[e^{tX_i/\sqrt{n}}] = 1 + \frac{t^2}{2n} + O(n^{-3/2}), \text{ since } EX = 0 \text{ and } EX^2 = 1.$$

Therefore, we obtain

$$E\left[\exp\left\{t\frac{X_1+\cdots+X_n}{\sqrt{n}}\right\}\right] \approx \left(1+\frac{t^2}{2n}\right)^n \to e^{t^2/2}.$$

Thus, the MGF of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the moment generating function of a standard normal random variable with mean 0 and variance 1. Notice that for $X \sim N(0,1)$, its MGF $\phi(t) = e^{t^2/2}$. Hence, it can be proven that the distribution function of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the distribution function of a standard normal Φ. When X_i have mean μ and variance σ^2, the random variables $\frac{X_i-\mu}{\sigma}$ have mean 0 and variance 1. Done.

Proposition 1.9.7 *The convergence has the following relations:*

$$\left.\begin{array}{c}\textit{conv. in moments or } L^p \textit{ converges}\\ \textit{a.s. or w.p.1}\end{array}\right\} \Rightarrow \textit{conv. in prob.} \Rightarrow \textit{conv. in distribution.}$$

Lemma 1.9.8 ***(Levy-Crammer)*** *$\{F_n\}$ is a set of distributions. If $\hat{F}_n \to \phi(t)$ conv. pointwisely, then $F_n \to F$ converges weakly, where ϕ is the character function of F and $\hat{F}_n$ is the character function of F_n.*

Theorem 1.9.9 ***(Linderberg-Levy CLT)*** *check https://zhuanlan.zhihu.com/p/69862244. character function and Lindberg-Levy central limit theorem.*

Example 1.9.10 *If X is binomially distributed with parameters n and p, then X is the sum of n independent Bernoulli random variables, each with parmeter p. Hence, the distribution of*

$$\frac{X-E[X]}{\sqrt{Var(X)}} = \frac{\sum X_i - n\mu}{\sqrt{n}\sigma} = \frac{X-np}{\sqrt{np(1-p)}}$$

approaches the standard normal distribution as n approaches ∞. The normal approximation will be quite good for $np(1-p) \geq 10$ or $\sqrt{Var(X)} \geq \sqrt{10}$.

Example 1.9.11 ***(Normal approximation to the Binomial)*** *Let X be the number of times that a fair coin, flipped 40 times, lands heads. Find the prob. that $X = 20$.*
Sol.

$$\begin{aligned} P\{X=20\} &= P\{19.5 < X < 20.5\} \\ &= P\left\{\frac{19.5-20}{\sqrt{10}} < \frac{X-20}{\sqrt{10}} < \frac{20.5-20}{\sqrt{10}}\right\} \\ &= P\left\{-0.16 < \frac{X-20}{\sqrt{10}} < 0.16\right\} = \Phi(0.16) - \Phi(-0.16) \\ &= 0.1272. \end{aligned}$$

The exact result is

$$P\{X=20\}=C_{40}^{20}\left(\frac{1}{2}\right)^{20}\left(\frac{1}{2}\right)^{20}=0.1268.$$

Example 1.9.12 *The lifetime of a battery is a random variable with mean* 40 *hours and standard deviation* 20 *hours. Assume a stockpile of* 25 *such batteries, approximate the probability that over* 1100 *hours of use can be obtained.*

Sol.

$$\begin{aligned} P\{X_1+\cdots+X_{25} &> 1100\}=P\left\{\frac{X_1+\cdots+X_{25}-25\times 40}{20\sqrt{25}}>\frac{1100-25\times 40}{20\sqrt{25}}\right\} \\ &= P\{N(0,1)>1\}=1-\Phi(1)=0.1587. \end{aligned}$$

Chapter 2

Conditional Probability

2.1 The Discrete Case

It is natural to define the conditional prob. mass function of X given that $Y = y$, by

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)},$$

for all values of y such that $P\{Y = y\} > 0$. The conditional expectation of X given that $Y = y$ is defined by

$$E[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

Example 2.1.1 *Suppose that the joint prob. mass function of X and Y is given by*

$$p(1,1) = 0.5, p(1,2) = 0.1, p(2,1) = 0.1, p(2,2) = 0.3.$$

Calculate the conditional prob. mass function of X given that $Y = 1$.
Sol. We first compute

$$p_Y(1) = \sum_x p(x,1) = p(1,1) + p(2,1) = 0.6.$$

Hence

$$p_{X|Y}(1|1) = \frac{p(1,1)}{p_Y(1)} = \frac{5}{6}, p_{X|Y}(2|1) = \frac{p(2,1)}{p_Y(1)} = \frac{1}{6}.$$

Example 2.1.2 *If X_1 and X_2 are independent binomial random variables with respective parameters (n_1, p) and (n_2, p), calculate the conditional prob. mass function of X_1 given that $X_1 + X_2 = m$.*
Sol. We first compute

$$\begin{aligned} P\{X_1 &= k|X_1 + X_2 = m\} = \frac{P\{X_1 = k, X_2 = m - k\}}{P\{X_1 + X_2 = m\}} \\ &= \frac{C_{n_1}^k p^k q^{n_1-k} C_{n_2}^{m-k} p^{m-k} q^{n_2-m+k}}{C_{n_1+n_2}^m p^m q^{n_1+n_2-m}} \\ &= \frac{C_{n_1}^k C_{n_2}^{m-k}}{C_{n_1+n_2}^m}. \end{aligned}$$

where we used $X_1 + X_2$ is a binomial with parameters $(n_1 + n_2, p)$.

Example 2.1.3 *Three possible outcomes with prob. p_i satisfying $p_1+p_2+p_3=1$. Suppose that n independent replications are performed, and let X_i be the number of times outcome i occurs. Determine the conditional expectation of X_1 given that $X_2=m$.*
Sol. For $k \le n-m$,

$$\begin{aligned} P\{X_1 &= k|X_2=m\} = \frac{P\{X_1=k, X_2=m\}}{P\{X_2=m\}} \\ &= \frac{P\{X_1=k, X_2=m, X_3=n-m-k\}}{P\{X_2=m, X_1+X_3=n-m\}} \\ &= \frac{\frac{n!}{k!m!(n-m-k)!}p_1^k p_2^m p_3^{n-m-k}}{\frac{n!}{m!(n-m)!}p_2^m(1-p_2)^{n-m}} \\ &= \frac{(n-m)!}{k!(n-m-k)!}\left(\frac{p_1}{1-p_2}\right)^k\left(\frac{p_3}{1-p_2}\right)^{n-m-k}, \end{aligned}$$

which is a Binomial with parameters $n-m$ and $\frac{p_1}{1-p_2}$. Thus, the conditional expectation is

$$E[X_1|X_2=m] = (n-m)\frac{p_1}{1-p_2}$$

2.2 The Continuous Case

If X and Y have a joint density function $f(x,y)$, then the conditional prob. density function of X given that $Y=y$, is defined by

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)},$$

for all values of y such that $f_Y(y)>0$. The conditional expectation of X given that $Y=y$ is defined by

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx.$$

Example 2.2.1 *Suppose that the joint density of X and Y is given by*

$$f(x,y) = \begin{cases} 4y(x-y)e^{-(x+y)}, & 0<x<\infty, 0\le y\le x, \\ 0, & \text{otherwise.} \end{cases}$$

Compute $E[X|Y=y]$.
Sol. The conditional density of X, given that $Y=y$, is given by

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} = \frac{4y(x-y)e^{-(x+y)}}{\int_y^\infty 4y(x-y)e^{-(x+y)}dx}, x>y \\ &= \frac{(x-y)e^{-x}}{\int_y^\infty (x-y)e^{-x}dx} = \frac{(x-y)e^{-x}}{\int_0^\infty we^{-(y+w)}dw}, w=x-y \\ &= (x-y)e^{-(x-y)} \end{aligned}$$

where we used that $w\sim\mathcal{E}(1)$ and $\int_0^\infty we^{-w}dw$ is the expected value of an exponential random variable with mean 1. Thus, using $EW=1$, $Var(W)=1$,

$$\begin{aligned} E[X|Y &= y] = \int_y^\infty x(x-y)e^{-(x-y)}dx = \int_0^\infty (w+y)we^{-w}dw \\ &= E[W^2]+yE[W] = Var(W)+(EW)^2+y = 2+y. \end{aligned}$$

Example 2.2.2 *The joint density of X and Y is*

$$f(x,y) = \begin{cases} \frac{1}{2}ye^{-xy}, & 0 < x < \infty, 0 < y < 2, \\ 0, & \text{otherwise.} \end{cases}$$

What is $E[e^{X/2}|Y=1]$?
Sol. The conditional density of X, given that $Y=1$, is given by

$$f_{X|Y}(x|1) = \frac{f(x,1)}{f_Y(1)} = \frac{\frac{1}{2}e^{-x}}{\int_0^\infty \frac{1}{2}e^{-x}dx} = e^{-x}.$$

Hence

$$E[e^{X/2}|Y=1] = \int_0^\infty e^{x/2} f_{X|Y}(x|1)dx = \int_0^\infty e^{-x/2}dx = 2.$$

2.3 Computing Expectatoins by Conditioning

Double expectation formula is very important:

$$E[X] = E[E[X|Y]],$$

The conditional expectation's expectation is unconditional expectation. If Y is discrete, then

$$E[X] = \sum_y E[X|Y=y]P(Y=y).$$

If Y is continuous with density $f_Y(y)$, then

$$E[X] = \int_{-\infty}^\infty E[X|Y=y]f_Y(y)dy.$$

Proof for **both discrete X and Y.**

$$\begin{aligned} \sum_y E[X|Y &= y]P(Y=y) = \sum_{x,y} xP[X|Y=y]P(Y=y) \\ &= \sum_{x,y} xP[X=x|Y=y]P(Y=y) = \sum_{x,y} xP[X=x,Y=y] \\ &= \sum_x xP[X=x] = E[X]. \end{aligned}$$

Proof for **both continuous X and Y.**

$$\begin{aligned} \int_{-\infty}^\infty E[X|Y &= y]f_Y(y)dy = \int_{-\infty}^\infty dx \int_{-\infty}^\infty xf_{X|Y}(x|y)f_Y(y)dy \\ &= \int_{-\infty}^\infty dx \int_{-\infty}^\infty xf(x,y)dy = \int_{-\infty}^\infty xf_X(x)dx = E[X]. \end{aligned}$$

Example 2.3.1 *Sam will choose either prob. book or history book with equal probability. If the number of misprints in prob. book is Poisson distributed with mean 2 and if the number of misprints in history book is Poisson distributed with mean 5. What is the expected number of misprints that Sam will com across?*
Sol. Let X be the number of misprints and

$$Y = \begin{cases} 1, & \textit{history book} \\ 2, & \textit{prob. book} \end{cases}$$

then

$$\begin{aligned} EX &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} \\ &= 5\left(\frac{1}{2}\right) + 2\left(\frac{1}{2}\right) = 7/2. \end{aligned}$$

Example 2.3.2 ***(The Expectation of the Sum of a Random Number of Random Variables)*** *Suppose that an industrial plant has the expected number of four accidents. Also suppose that the number of workers injured in each accident are independent random variables with a common mean 2. What is the expected number of injuries during a week?*
Sol. Let N be the number of accidents and X_i be the number injured in the ith accident. Then the total number of injuries is $\sum_{i=1}^{N} X_i$

$$E\left[\sum_{i=1}^{N} X_i\right] = E\left[E\left[\sum_{i=1}^{N} X_i|N\right]\right].$$

The enclosed quantity can be computed by

$$E\left[\sum_{i=1}^{N} X_i|N=n\right] = E\left[\sum_{i=1}^{n} X_i\right] = nEX_i \Rightarrow E\left[\sum_{i=1}^{N} X_i|N\right] = NE[X_i].$$

Thus

$$E\left[\sum_{i=1}^{N} X_i\right] = E[NE[X_i]] = E[N]E[X_i] = 4 \times 2 = 8.$$

Definition 2.3.3 *The random number N is independent of i.i.d. random variables X_i, then $\sum_{i=1}^{N} X_i$ is said to be a* ***compound random variable****.*

Example 2.3.4 *(The Mean of a Geometric Distribution) A coin, having prob. p of coming up heads, is to be successively flipped until the first head appears. What is the expected number of flips required?*
Sol. Let N be the number of flips required, and let

$$Y = \begin{cases} 1, & \text{if the first flip in a head,} \\ 0, & \text{if the first flip in a tail.} \end{cases}$$

Now,

$$\begin{aligned} E[N] &= E[N|Y=1]P\{Y=1\} + E[N|Y=0]P\{Y=0\} \\ &= pE[N|Y=1] + (1-p)E[N|Y=0]. \end{aligned}$$

Notice that

$$E[N|Y=1] = 1, E[N|Y=0] = E[N] + 1.$$

Then

$$E[N] = p + (1-p)(E[N]+1) \Rightarrow E[N] = 1/p,$$

which is the same with the mean of a geometric distribution.

Example 2.3.5 *A miner is trapped in a mine containing three doors and only one door is out. The 1st door takes him to safety after 2 hours of travel. The 2nd door returns him to the mine after 3 hours of travel. The 3rd door returns him to the mine after 5 hours. Assume that the miner is at all times equally likely to*

choose one of the doors, what is the expected length of time until the miner reaches safety?
Sol. Let X denote the time until the miner reaches safety, and let Y denote the door he intially chooses.

$$\begin{aligned} EX &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} + E[X|Y=3]P\{Y=3\} \\ &= \frac{1}{3}(E[X|Y=1] + E[X|Y=2] + E[X|Y=3]). \end{aligned}$$

Since $E[X|Y=1] = 2, E[X|Y=2] = 3 + EX, E[X|Y=3] = 5 + EX$, hence

$$EX = \frac{1}{3}(10 + 2EX) \Rightarrow E[X] = 10.$$

Example 2.3.6 *Indepedent trials, each success with prob. p, are performed until there k consecutive successes. What is the mean number of necessary trials?*
Sol. Let N_k denote the number necessary trials to obtain k consecutive successes, and let $M_k = EN_k$. Then compute by conditioning on N_{k-1},

$$M_k = E[N_k] = E[E[N_k|N_{k-1}]].$$

Notice that

$$E[N_k|N_{k-1}] = N_{k-1} + \{p \cdot 1 + (1-p)(1 + E[N_k])\}.$$

Taking expectations at both sides,

$$M_k = M_{k-1} + 1 + (1-p)M_k \Rightarrow M_k = \frac{1}{p} + \frac{M_{k-1}}{p}.$$

Since N_1 is geometric with parameter 1, then $M_1 = \frac{1}{p}$, and recursively

$$M_2 = \frac{1}{p} + \frac{1}{p^2},$$

In general,

$$M_k = \frac{1}{p} + \cdots + \frac{1}{p^k}.$$

Another way to use conditioning is to obtain the variance of a random variable by applying the conditional variance formula. The conditional variance is defined by

$$\begin{aligned} Var(X|Y &= y) = E[(X - E[X|Y=y])^2|Y=y] \\ &= E[X^2|Y=y] - (E[X|Y=y])^2. \end{aligned}$$

Proposition 2.3.7 ***(The Conditional Variance Formula)***

$$Var(X) = E[Var(X|Y)] + Var(E[X|Y]).$$

Pf.

$$\begin{aligned} E[Var(X|Y)] &= E\{E[X^2|Y] - (E[X|Y])^2\} = E\{E[X^2|Y]\} - E\{(E[X|Y])^2\} \\ &= E[X^2] - E\{(E[X|Y])^2\}. \end{aligned}$$

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

Thus

$$E[Var(X|Y)] + Var(E[X|Y]) = E[X^2] - (E[X])^2 = Var(X).$$

Example 2.3.8 ***(The Variance of a Compound Random Variable)*** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Assume they are independent of the nonnegative random integer N. The random variable $S = \sum_{i=1}^{N} X_i$ is a compound random variable. Find its variance.*
Sol. Let us use the conditional variance formula. Compute

$$Var(S|N=n) = Var(\sum_{i=1}^{N} X_i | N=n) = Var(\sum_{i=1}^{n} X_i) = n\sigma^2.$$

$$E(S|N=n) = E(\sum_{i=1}^{N} X_i | N=n) = E(\sum_{i=1}^{n} X_i) = n\mu.$$

Therefore,

$$Var(S|N) = N\sigma^2, \quad E(S|N) = N\mu.$$

Using conditional variance formula,

$$\begin{aligned} Var(S) &= E[Var(S|N)] + Var[E(S|N)] = E[N]\sigma^2 + Var(N\mu) \\ &= \sigma^2 E[N] + \mu^2 Var(N). \end{aligned}$$

Another way is to compute the variance directly.

$$\begin{aligned} Var(S) &= ES^2 - (ES)^2, \\ ES &= E[E[S|N]] = \mu E[N], \\ ES^2 &= E[E[S^2|N]] = E[N^2\mu^2 + N\sigma^2] = \mu^2 E[N^2] + \sigma^2 E[N], \end{aligned}$$

where

$$\begin{aligned} E[S^2|N] &= E[(\sum_{i=1}^{N} X_i)^2] = E[\sum_{i=1}^{N} X_i^2 + 2\sum_{i<j} X_i X_j] \\ &= NEX_i^2 + (N^2 - N)EX_iEX_j = N(\mu^2 + \sigma^2) + (N^2 - N)\mu^2 \\ &= .N^2\mu^2 + N\sigma^2. \end{aligned}$$

Therefore,

$$\begin{aligned} Var(S) &= ES^2 - (ES)^2 = \mu^2 E[N^2] + \sigma^2 E[N] - (\mu E[N])^2 \\ &= \mu^2 Var(N) + \sigma^2 E[N]. \end{aligned}$$

Example 2.3.9 *If N is a Poisson random variable, then $S = \sum_{i=1}^{N} X_i$ is called a* ***compound Poisson random variable****. Since*

$$E[N] = Var(N) = \lambda,$$

then

$$Var(S) = \mu^2 Var(N) + \sigma^2 E[N] = \mu^2\lambda + \sigma^2\lambda = \lambda E[X^2].$$

We should point out that ***the following derivation is wrong****:*

$$Var(S) = Var(\sum_{i=1}^{N} X_i) = E[N]Var(X_i) = \lambda Var(X).$$

2.4 Computing Probabilities by Conditioning

We may also use conditioning approach to compute probabilities. Let E denote an arbitrary event and define the indicator random variable X by

$$X = \begin{cases} 1, & \text{if } E \text{ occurs,} \\ 0, & \text{if } E \text{ does not occur.} \end{cases}$$

It follows from the definition of X that

$$\begin{aligned} E[X] &= 1 \cdot P(E) + 0 \cdot P(E^c) = P(E), \\ E[X|Y &= y] = 1 \cdot P(E|Y=y) + 0 \cdot P(E^c|Y=y) = P(E|Y=y), \end{aligned}$$

for any random variable Y. Then from double expectation formula,

$$P(E) = \begin{cases} \sum\limits_y P(E|Y=y)P(Y=y), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} P(E|Y=y)f_Y(y)dy, & \text{if } Y \text{ is continuous.} \end{cases}$$

Example 2.4.1 *Suppose X and Y are independent continuous random variables having densities f_X and f_Y. Compute $P\{X < Y\}$.*
Sol. Conditioning on the value of Y yields

$$\begin{aligned} P\{X &< Y\} = \int_{-\infty}^{\infty} P(X < Y|Y=y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P(X < y) f_Y(y) dy = \int_{-\infty}^{\infty} F_X(y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} dy \int_{-\infty}^{y} f_X(x) f_Y(y) dx. \end{aligned}$$

Example 2.4.2 *An insurance company supposes that (1) each policyholder has number of accidents with Poisson distributed, with the random mean. (2) The mean of Poisson has a gamma distribution with density*

$$g(\lambda) = \lambda e^{-\lambda}, \quad \lambda \geq 0.$$

What is the prob. that a randomly chosen policyholder has exactly n accidents next year?
Sol. Let X denote the number of accidents that a randomly chosen policyholder has next year. Let Y be the Poisson mean number of accidents for this policyholder. Then, conditioning on Y yields

$$\begin{aligned} P\{X &= n\} = \int_0^{\infty} P\{X = n|Y=\lambda\} g(\lambda) d\lambda \\ &= \int_0^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \lambda e^{-\lambda} d\lambda = \int_0^{\infty} \frac{1}{n!} \lambda^{n+1} e^{-2\lambda} d\lambda. \end{aligned}$$

By induction,

$$\begin{aligned} P\{X &= 0\} = \int_0^{\infty} \frac{1}{0!} \lambda^1 e^{-2\lambda} d\lambda = -\frac{1}{2} \int_0^{\infty} \lambda de^{-2\lambda} \\ &= -\frac{1}{2} \left(\lambda e^{-2\lambda}|_0^{\infty} - \int_0^{\infty} e^{-2\lambda} d\lambda \right) = \frac{1}{2} \int_0^{\infty} e^{-2\lambda} d\lambda \\ &= -\frac{1}{4} e^{-2\lambda}|_0^{\infty} = \frac{1}{4}. \end{aligned}$$

$$\begin{aligned} P\{X &= 1\} = \int_0^\infty \lambda^2 e^{-2\lambda} d\lambda = -\frac{1}{2}\int_0^\infty \lambda^2 de^{-2\lambda} \\ &= -\frac{1}{2}\left(\lambda^2 e^{-2\lambda}|_0^\infty - 2\int_0^\infty \lambda e^{-2\lambda} d\lambda\right) = \int_0^\infty \lambda e^{-2\lambda} d\lambda = \frac{1}{4} = \frac{2}{8}. \end{aligned}$$

$$\begin{aligned} P\{X &= 2\} = \frac{1}{2}\int_0^\infty \lambda^3 e^{-2\lambda} d\lambda = -\frac{1}{4}\int_0^\infty \lambda^3 de^{-2\lambda} \\ &= -\frac{1}{4}\left(\lambda^3 e^{-2\lambda}|_0^\infty - 3\int_0^\infty \lambda^2 e^{-2\lambda} d\lambda\right) = \frac{3}{4}\int_0^\infty \lambda^2 e^{-2\lambda} d\lambda = \frac{3}{16}. \end{aligned}$$

$$\begin{aligned} P\{X &= 3\} = \frac{1}{6}\int_0^\infty \lambda^4 e^{-2\lambda} d\lambda = -\frac{1}{12}\int_0^\infty \lambda^4 de^{-2\lambda} \\ &= -\frac{1}{12}\left(\lambda^4 e^{-2\lambda}|_0^\infty - 4\int_0^\infty \lambda^3 e^{-2\lambda} d\lambda\right) = \frac{1}{3}\int_0^\infty \lambda^3 e^{-2\lambda} d\lambda = \frac{1}{3}\frac{3}{8} = \frac{1}{8} = \frac{4}{32}. \end{aligned}$$

By induction, we can show that $P\{X = n\} = \frac{n+1}{2^{n+2}}$. *That is,*

$$\begin{aligned} P\{X &= n\} = \frac{1}{n!}\int_0^\infty \lambda^{n+1} e^{-2\lambda} d\lambda = -\frac{1}{2n!}\int_0^\infty \lambda^{n+1} de^{-2\lambda} \\ &= -\frac{1}{2n!}\left(\lambda^{n+1} e^{-2\lambda}|_0^\infty - (n+1)\int_0^\infty \lambda^n e^{-2\lambda} d\lambda\right) = \frac{n+1}{2n}\int_0^\infty \frac{1}{(n-1)!}\lambda^n e^{-2\lambda} d\lambda \\ &= \frac{n+1}{2n}\frac{n}{2^{n+1}} = \frac{n+1}{2^{n+2}}. \end{aligned}$$

Another way is to notice that

$$h(\lambda) = \frac{2e^{-2\lambda}(2\lambda)^{n+1}}{(n+1)!}$$

is the density function of a gamma $(n+2, 2)$ *random variable, its integral is* 1. *Therefore,*

$$1 = \int_0^\infty \frac{2e^{-2\lambda}(2\lambda)^{n+1}}{(n+1)!} d\lambda = \frac{2^{n+2}}{n+1}\int_0^\infty \frac{1}{n!}\lambda^{n+1} e^{-2\lambda} d\lambda$$

Thus

$$P\{X = n\} = \int_0^\infty \frac{1}{n!}\lambda^{n+1} e^{-2\lambda} d\lambda = \frac{n+1}{2^{n+2}}.$$

Example 2.4.3 *Suppose that the number of people who visit a yoga studio each day is a Poisson random variable with mean* λ. *Suppose further that each persion is independently femail with prob.* p *or mail with prob.* $1-p$. *Find the joint prob. that exactly* n *women and* m *men visit the academy today.*
Sol. Let N_1 *denote the number of women and* N_2 *the number of men, so that* $N = N_1 + N_2$ *is the total number of people. Conditioning on* N *gives*

$$\begin{aligned} P\{N_1 &= n, N_2 = m\} = \sum_{i=0}^{\infty} P\{N_1 = n, N_2 = m | N = i\} P\{N = i\} \\ &= P\{N_1 = n, N_2 = m | N = n+m\} P\{N = n+m\} \\ &= P\{N_1 = n, N_2 = m | N = n+m\} e^{-\lambda}\frac{\lambda^{n+m}}{(n+m)!}, \ \textit{Poisson} \\ &= C_{n+m}^n p^n (1-p)^m e^{-\lambda}\frac{\lambda^{n+m}}{(n+m)!}, \quad \textit{Binomial} \\ &= e^{-\lambda p}\frac{(\lambda p)^n}{n!} e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!}. \end{aligned}$$

Since

$$\begin{aligned} P\{N_1 &= n\} = \sum_{m=0}^{\infty} P\{N_1 = n, N_2 = m\} = e^{-\lambda p}\frac{(\lambda p)^n}{n!}, \\ P\{N_2 &= m\} = \sum_{n=0}^{\infty} P\{N_1 = n, N_2 = m\} = e^{-\lambda(1-p)}\frac{(\lambda(1-p))^m}{m!}. \end{aligned}$$

so that $P\{N_1 = n, N_2 = m\} = P\{N_1 = n\}P\{N_2 = m\}$ *and we can conclude that* N_1 *and* N_2 *are independent Poisson random variables with means* λp *and* $\lambda(1-p)$*. Poisson diversion and confluence.*

Example 2.4.4 *(The Ballot Problem)* （投票问题） *In an election, candidate A receives* n *votes, and candidate B receives* m *votes where* $n > m$*. Assuming that all orderings are equally likely, show that the prob. that A is always ahead in the count of votes is* $(n-m)/(n+m)$*.*
Sol. Let $P_{n,m}$ *denote the desired prob. By conditioning on which candidate receives the last vote, we have*

$$\begin{aligned} P_{n,m} &= P\{A \text{ always ahead} | A \text{ receives last vote}\}\frac{n}{n+m} \\ &\quad + P\{A \text{ always ahead} | B \text{ receives last vote}\}\frac{m}{n+m}. \end{aligned}$$

Notice that

$$\begin{aligned} P\{A \text{ always ahead} | A \text{ receives last vote}\} &= P_{n-1,m}, \\ P\{A \text{ always ahead} | B \text{ receives last vote}\} &= P_{n,m-1}. \end{aligned}$$

Then one obtains the recursive formula,

$$P_{n,m} = P_{n-1,m}\frac{n}{n+m} + P_{n,m-1}\frac{m}{n+m}, \quad \text{as } n-1 \geq m.$$

Notice that

$$\begin{aligned} P_{1,0} &= 1, P_{2,0} = 1, \ldots, P_{n,0} = 1, \\ P_{1,1} &= 0, P_{2,2} = 0, \ldots, P_{n,n} = 0. \end{aligned}$$

We now can prove the result by induction

$$\begin{aligned} P_{n,m} &= \frac{n-1-m}{n-1+m}\frac{n}{n+m} + \frac{n-m+1}{n+m-1}\frac{m}{n+m} \\ &= \frac{n^2 - n - mn + mn - m^2 + m}{(n-1+m)(n+m)} = \frac{(n-m)(n+m-1)}{(n-1+m)(n+m)} \\ &= \frac{n-m}{n+m}. \end{aligned}$$

2.5 Computing Conditional Expectation and Conditional Probability by Conditioning

The analog of

$$E[X] = \begin{cases} \sum_y E[X|Y=y]P(Y=y), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)dy, & \text{if } Y \text{ is continuous.} \end{cases}$$

is

$$E[X|Y=y]=\begin{cases}\sum_w E[X|Y=y,W=w]P(W=w|Y=y), & \text{if } W \text{ is discrete,}\\ \int_{-\infty}^{\infty} E[X|Y=y,W=w]f_{W|Y}(w|y)dw, & \text{if } W \text{ is continuous.}\end{cases}$$

That is, in another form,

$$E[X|Y]=E\big[E[X|Y,W]|Y\big].$$

Here is a proof for the above formula. For discrete W,

$$\begin{aligned}\sum_w E[X|Y &= y,W=w]P(W=w|Y=y)=\sum_w\sum_x xP[X=x|Y=y,W=w]P(W=w|Y=y)\\ &= \sum_w\sum_x \frac{xP[X=x,Y=y,W=w]}{P\{Y=y,W=w\}}\frac{P(W=w,Y=y)}{P(Y=y)}\\ &= \sum_x \frac{xP[X=x,Y=y]}{P(Y=y)}=\sum_x xP[X=x|Y=y]=E[X|Y=y].\end{aligned}$$

For continuous W,

$$\begin{aligned}\int_w E[X|Y &= y,W=w]f_{W|Y}(w|y)dw=\int_w\int_x xf_{X|Y,W}(x|y,w)f_{W|Y}(w|y)dwdx\\ &= \int_w\int_x x\frac{f_{X,Y,W}(x,y,w)}{f_{Y,W}(y,w)}\frac{f_{Y,W}(y,w)}{f_Y(y)}dwdx\\ &= \int_w\int_x xf_{X,W|Y}(x,w|y)dwdx=\int_x xf_{X|Y}(x|y)dx=E[X|Y=y].\end{aligned}$$

Example 2.5.1 *(1) An insurance company classifies each policyholder as being one of k types. (2) Type i has the number of accidents with Poisson distribution with mean λ_i. (3) A newly policyholder is type i with prob. $p_i, \sum_{i=1}^k p_i=1$. (4) Given that a policyholder had n accidents in her first year, the question is [1] what is the expected number that she has in her second year? [2] What is the conditional prob. that she has m accidents in her second year?*

Sol. Let N_i denote the number of accidents the policyholder has in ith $(i=1,2)$ year. Conditioning on her risk type T gives

$$\begin{aligned}E[N_2|N_1 &= n]=\sum_{j=1}^k E[N_2|T=j,N_1=n]P\{T=j|N_1=n\}\\ &= \sum_{j=1}^k E[N_2|T=j]P\{T=j|N_1=n\}=\sum_{j=1}^k \lambda_j P\{T=j|N_1=n\},\end{aligned}$$

where the last follows the Poission mean λ_j. Using Bayes' formula

$$\begin{aligned}P\{T &= j|N_1=n\}=\frac{P\{T=j,N_1=n\}}{P\{N_1=n\}}=\frac{P\{N_1=n|T=j\}P\{T=j\}}{\sum_{j=1}^k P\{N_1=n|T=j\}P\{T=j\}}\\ &= \frac{e^{-\lambda_j}\lambda_j^n/n!p_j}{\sum_{j=1}^k e^{-\lambda_j}\lambda_j^n/n!p_j},\end{aligned}$$

one arrives at

$$\begin{aligned}E[N_2|N_1 &= n]=\sum_{j=1}^k \lambda_j P\{T=j|N_1=n\}=\sum_{j=1}^k \lambda_j\frac{e^{-\lambda_j}\lambda_j^n/n!p_j}{\sum_{j=1}^k e^{-\lambda_j}\lambda_j^n/n!p_j}\\ &= \frac{\sum_{j=1}^k e^{-\lambda_j}\lambda_j^{n+1}p_j}{\sum_{j=1}^k e^{-\lambda_j}\lambda_j^n p_j}.\end{aligned}$$

The conditional prob. can also be obtained by conditioning on her type:

$$\begin{aligned} P\{N_2 &= m|N_1 = n\} = \sum_{j=1}^{k} P[N_2 = m|T = j, N_1 = n]P\{T = j|N_1 = n\} \\ &= \sum_{j=1}^{k} \frac{e^{-\lambda_j}\lambda_j^m}{m!} \frac{e^{-\lambda_j}\lambda_j^n/n!p_j}{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n/n!p_j} = \frac{\sum_{j=1}^{k} e^{-2\lambda_j}\lambda_j^{m+n}p_j}{m!\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n p_j}. \end{aligned}$$

The second approach is to calculate the prob. $P\{N_2 = m|N_1 = n\}$ first,

$$\begin{aligned} P\{N_2 &= m|N_1 = n\} = \frac{P\{N_2 = m, N_1 = n\}}{P\{N_1 = n\}} \\ &= \frac{\sum_{j=1}^{k} P\{N_2 = m, N_1 = n|T = j\}p_j}{\sum_{j=1}^{k} P\{N_1 = n|T = j\}p_j} = \frac{\sum_{j=1}^{k} \frac{e^{-\lambda_j}\lambda_j^m}{m!}\frac{e^{-\lambda_j}\lambda_j^n}{n!}p_j}{\sum_{j=1}^{k} \frac{e^{-\lambda_j}\lambda_j^n}{n!}p_j} \\ &= \frac{1}{m!}\frac{\sum_{j=1}^{k} e^{-2\lambda_j}\lambda_j^{m+n}p_j}{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n p_j}. \end{aligned}$$

Then conditional expectation can be calculated by

$$\begin{aligned} E\{N_2|N_1 &= n\} = \sum_{m=0}^{\infty} mP\{N_2 = m|N_1 = n\} = \sum_{m=1}^{\infty} \frac{m}{m!}\frac{\sum_{j=1}^{k} e^{-2\lambda_j}\lambda_j^{m+n}p_j}{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n p_j} \\ &= \frac{\sum_{j=1}^{k} \lambda_j^1 e^{-\lambda_j}\lambda_j^n p_j}{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n p_j} = \frac{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^{n+1}p_j}{\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^n p_j}. \end{aligned}$$

where made use of $\sum_{m=1}^{\infty} m\frac{e^{-\lambda_j}\lambda_j^m}{m!} = \lambda_j$.

Chapter 3

Markov Chain

3.1 Introduction

• Stochastic Process $\{X(t), t \in T\}$ is a collection of random variables. For each $t \in T$, $X(t)$ is a random variable.

• t is the time, either discrete or continuous.

• $X(t)$ is the state at time t. For example, (1) number of custumers in a supermarket (2) number of accidents on a highway.

• T is the index set of the process.

(1) T is finite or countable $\rightarrow$ discrete-time process.

(2) T is an interval of a real line $\rightarrow$ continuous-time process.

• $\{X_n, n = 0, 1, 2, \cdots\}$ is a discrete-time SP (Stochastic Process). $\{X(t), t \geq 0\}$ is a continuous-time SP.

• State space of a SP is the set of all possible values that $X(t)$ can take, either subset of $\mathbb{R}^m$ or discrete spaces like $\mathbb{Z}^m$.

continuous
$\begin{cases} \text{temperature every day} \\ \text{the time spent for study} \end{cases}$
discrete
$\begin{cases} \text{number of custumers in a supermarket} \\ \text{machine/engine works or not, 0 or 1.} \end{cases}$

• State space. $I = \{0, 1, 2, \ldots\}$ finite or countable

$X_n = i$, in state i at time n.

• $\{X_n\}_{n=0}^{\infty}$ is a time-dependent process. X_{n+1} may depend on X_n and even earlier.

Example 3.1.1 *The followings are Stochastic Processes. (1) Brownian Motion, pollen （花粉） , Jean Perrin. (2) Poisson Process which is a counting process. (3) The meters one walked every day. (4) number of students in the classroom.*

Definition 3.1.2 *The following stochastic process is known as a* ***homogeneous Markov Chain****:*

$$\begin{aligned} P(X_{n+1} &= j|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j|X_n = i) \\ &= P(X_1 = j|X_0 = i) := p_{ij}, \quad i, j \in I, \end{aligned}$$

where the time-independent p_{ij} is the ***one-step transition probability*** *（一步转移概率）, and* $\mathbf{P} = (p_{ij})_{i,j\in I}$ *is the* ***one-step transition prob. matrix*** *（转移概率矩阵）. Obviously, one has* $p_{ij} \geq 0$ *and by mutually exclusive property,*

$$\sum_{j=0}^{\infty} p_{ij} = \sum_{j=0}^{\infty} P(X_1 = j|X_0 = i) = P(\bigcup_{j=0}^{\infty}\{X_1 = j\}|X_0 = i) = 1.$$

The row sum of a stochastic matrix （随机矩阵） is 1:

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Definition 3.1.3 *The* ***nonhomogeneous Markov Chain*** *（非时齐） still has the Markovian property. In non homogeneous chains, transition probabilities can vary across time. That is to say, the Markov property is retained but the transition probabilities may depend on time.*

$$P(X_{n+1} = j|X_n = i) \neq P(X_1 = j|X_0 = i),$$

or $p_{ij}(n)$ *depends on time* n *or* $p_{ij}(t)$ *depends on time* t*.*

Example 3.1.4 X_n *depends on* $X_{n-1}, X_{n-2}, \ldots, X_{n-p}$*. Markov Process,* $AR(1)$ *model,*

$$X_n = a_1 X_{n-1} + \varepsilon.$$

$AR(p)$ *model* $(p \geq 2)$*. time series. non-Markov process.*

$$X_n = a_1 X_{n-1} + \cdots + a_p X_{n-p} + \varepsilon.$$

Example 3.1.5 *(Forecasting the weather) Suppose that if it rains today, then it will rain tomorrow with prob.* α*; and if it does not rain today, then it will rain tomorrow with prob.* β*. Find the transition prob. matrix.*

Sol. Let

state 1, rain,
state 2, not rain.

Then

$$\mathbf{P} = \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}.$$

Example 3.1.6 *(Transform a non-Markov process to a Markov Chain) Suppose that whether it rains today depends on the last two days. Suppose that it has rained for the past two days, then it will rain tomorrow with prob. 0.7; if it rained today but not yesterday, then it will rain tomorrow with prob. 0.5; if it rained yesterday but not today, then it will rain tomorrow with prob. 0.4; if it has not rained in the past two days, then it will rain tomorrow with prob. 0.2. Give the transition prob. matrix.*

Sol. The condition is

yesterday	*today*	*tomorrow*
✓	✓	✓ = 0.7, × = 0.3
×	✓	✓ = 0.5, × = 0.5
✓	×	✓ = 0.4, × = 0.6
×	×	✓ = 0.2, × = 0.8

.

Let

$$\begin{array}{ll} \textit{state 0,} & \textit{(rain,rain)} \\ \textit{state 1,} & \textit{(not,rain)} \\ \textit{state 2,} & \textit{(rain,not)} \\ \textit{state 3,} & \textit{(not,not)} \end{array}.$$

The four-state Markov Chain has a transition prob. matrix,

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{bmatrix}.$$

Example 3.1.7 ***(A random walk model)*** （随机游走模型） *A Markov chain whose state space is given by the integers $i = 0, \pm 1, \pm 2, \ldots$ is said to be a random walk if, for some $0 < p < 1$,*

$$p_{i,i+1} = p, \quad p_{i,i-1} = 1 - p, \quad i = 0, \pm 1, \pm 2, \ldots$$

Note that this is an infinite-state chain.

Example 3.1.8 ***(A Gambling model)*** *Consider a gambler who, at each play, either wins 1 dollar with prob p or loses 1 dollar with prob. $1 - p$. Suppose that the gambler quits playing either when he goes broke or he attains a fortune of N dollars. Then the Markov chain has the transition prob.*

$$\begin{aligned} p_{i,i+1} &= p, \quad p_{i,i-1} = 1 - p, \quad i = 1, 2, \ldots, N-1, \\ p_{00} &= p_{NN} = 1. \end{aligned}$$

State 0 and N are called absorbing states （吸收态，吸收壁） *since once entered they are never left. Note that this is a finite-state chain with absorbing barriers.*

Proposition 3.1.9 *I is the state space of the homogeneous Markov Chain $\{X_n\}$. $A, A_j \subset I$.*
(1) When $X_n = i$ is given, the future $\{X_m : m \geq n+1\}$ is independent of the past $\{X_j : j \leq n-1\}$.
(2) $P(X_{n+k} = j | X_n = i) = P(X_k = j | X_0 = i)$.
(3) $P(X_{n+k} = j | X_n = i, X_{n-1} \in A_{n-1}, \ldots, X_0 \in A_0) = P(X_k = j | X_0 = i)$.
(4) $P(X_{n+k} \in A | X_n = i, X_{n-1} \in A_{n-1}, \ldots, X_0 \in A_0) = P(X_k \in A | X_0 = i)$.

3.2 Chapman-Kolmogorov Equations

（强调我的记号）

We now define the n-**step transition prob.** $p_{ij}^{(n)}$ to be the prob. that a process in state i will be in state j after n additional transitions. That is

$$p_{ij}^{(n)} = P\{X_{n+k} = j | X_k = i\}, \quad n \geq 0, i, j \geq 0.$$

Obviously, $p_{ij}^{(1)} = p_{ij}$. In particular,

$$p_{ij}^{(0)} = P\{X_0 = j | X_0 = i\} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

or say $\mathbf{P}^{(1)} = \mathbf{P}$, and $\mathbf{P}^{(0)} = \mathbf{I}$ (identity matrix).

Theorem 3.2.1 ***(Chapman-Kolmogorov Equation)*** *For any $m, n \geq 0$,*

$$\begin{aligned} p_{ij}^{(n+m)} &= \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)}, \\ \mathbf{P}^{(n+m)} &= \mathbf{P}^{n+m}. \end{aligned}$$

Pf.

$$\begin{aligned} p_{ij}^{(n+m)} &= P\{X_{n+m} = j | X_0 = i\} \\ &= \sum_{k \in I} P\{X_{n+m} = j | X_n = k, X_0 = i\} P\{X_n = k | X_0 = i\} \\ &= \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)}. \end{aligned}$$

Corollary 3.2.2 *(1)* $p_{ij}^{(n+m)} \geq p_{il}^{(n)} p_{lj}^{(m)}$;
(2) $p_{ii}^{(n+k+m)} \geq p_{ij}^{(n)} p_{jl}^{(k)} p_{li}^{(m)}$;
(3) $p_{ii}^{(nk)} \geq \left(p_{ii}^{(n)}\right)^k$.

Pf.

$$p_{ij}^{(n+m)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)} \geq p_{il}^{(n)} p_{lj}^{(m)}.$$

Example 3.2.3 *Consider Example 3.1.5. If $\alpha = 0.7$ and $\beta = 0.4$, then calculate the prob. that it will rain in four days given that it is raining today.*
Sol. The one-step transition prob. matrix is given by

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

Hence,

$$\begin{aligned} \mathbf{P}^2 &= \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}^2 = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix}, \\ \mathbf{P}^4 &= (\mathbf{P}^2)^2 = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}, \end{aligned}$$

so that the desired prob. $p_{00}^{(4)} = 0.5749$. (notice that row sum is always 1.)

Example 3.2.4 *Consider Example 3.1.6. Given that it rained on Monday and Tuesday, what is the prob. that it will rain on Thursday?*
Sol. The two-step transition matrix is given by

$$\mathbf{P}^{(2)} = \mathbf{P}^2 = \begin{bmatrix} 0.7 & 0 & 0.3 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.2 & 0 & 0.8 \end{bmatrix}^2 = \begin{bmatrix} 0.49 & 0.12 & 0.21 & 0.18 \\ 0.35 & 0.20 & 0.15 & 0.30 \\ 0.20 & 0.12 & 0.20 & 0.48 \\ 0.10 & 0.16 & 0.10 & 0.64 \end{bmatrix}.$$

The desired prob. is $p_{00}^{(2)} + p_{01}^{(2)} = 0.49 + 0.12 = 0.61$. (Notice that row sum is always 1.)

3.3 Unconditional Distribution of the State

So far, all prob. are conditional prob. If the unconditional distribution of the state at time n is desired, it is necessary to specify the prob. distribution of the initial state.

Definition 3.3.1 *X_0 has the prob. distribution*

$$\pi_j^{(0)} = P(X_0 = j), \quad j \in I = \{1, 2, \ldots\}.$$

The initial distribution of $\{X_n\}$ is denoted by

$$\pi^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \cdots).$$

Moreover, let the distribution at time n be

$$\begin{aligned} \pi_j^{(n)} &= P(X_n = j), \quad j \in I, \\ \pi^{(n)} &= (\pi_1^{(n)}, \pi_2^{(n)}, \cdots). \end{aligned}$$

In fact, $\pi^{(n)}$ can be uniquely determined by the initial distribution $\pi^{(0)}$ and the transition matrix $\mathbf{P}$.

Theorem 3.3.2 *Assume Markov chain $\{X_n\}$ has initial dist. $\pi^{(0)}$ and trans. prob. matrix $\mathbf{P}$. Then For any $0 \le n_0 < n_1 < \cdots < n_m$,*

$$P(X_{n_0} = i_0, X_{n_1} = i_1, \cdots, X_{n_m} = i_m) = \pi_{i_0}^{(n_0)} p_{i_0 i_1}^{(n_1 - n_0)} p_{i_1 i_2}^{(n_2 - n_1)} \cdots p_{i_{m-1} i_m}^{(n_m - n_{m-1})}.$$

(2) For any $\forall n \ge 1$,

$$\begin{cases} \pi^{(n+1)} = \pi^{(n)}\mathbf{P}, & \text{where } \pi^{(n)} \text{ is a row vector,} \\ \pi^{(n)} = \pi^{(0)}\mathbf{P}^n, & \text{or say } \pi_j^{(n)} = \sum_{i \in I} \pi_i^{(0)} p_{ij}^{(n)}, \\ \pi^{(n)} = \pi^{(k)}\mathbf{P}^{n-k}, & 0 \le k \le n. \end{cases}$$

Proof. (1) Using the product formula,

$$P(B_1 B_2 \ldots B_n) = P(B_1) P(B_2|B_1) \cdots P(B_n|B_1 B_2 \ldots B_{n-1}),$$

one has

$$\begin{aligned} P(X_{n_0} &= i_0, X_{n_1} = i_1, \cdots, X_{n_m} = i_m) \\ &= P(X_{n_0} = i_0) P(X_{n_1} = i_1 | X_{n_0} = i_0) \cdots P(X_{n_m} = i_m | X_{n_{m-1}} = i_{m-1}) \\ &= \pi_{i_0}^{(n_0)} p_{i_0 i_1}^{(n_1 - n_0)} p_{i_1 i_2}^{(n_2 - n_1)} \cdots p_{i_{m-1} i_m}^{(n_m - n_{m-1})}. \end{aligned}$$

(2)

$$\begin{aligned} \pi_j^{(n+1)} &= P(X_{n+1} = j) = \sum_{i \in I} P(X_{n+1} = j | X_n = i) P(X_n = i) \\ &= \sum_{i \in I} \pi_i^{(n)} p_{ij}, \end{aligned}$$

that is $\pi^{(n+1)} = \pi^{(n)}\mathbf{P}$. ■

Example 3.3.3 *An urn always contains 2 balls possibly with red and blue colors. (1) At each stage a ball is randomly chosen. (2) Then the chosen ball is replaced by the same color with prob.* 0.8 *and opposite color with prob.* 0.2*. If initially both balls are red, find the prob. that the third ball selected is red.*
Sol. Define X_n *to be the number of red balls in the urn after the* n*th section. Then the transition matrix is*

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.2 & 0.8 \end{bmatrix}.$$

Then

$$\mathbf{P}^2 = \begin{bmatrix} 0.66 & 0.32 & 0.02 \\ 0.16 & 0.68 & 0.16 \\ 0.02 & 0.32 & 0.66 \end{bmatrix},$$

and the desired prob. is

$$\begin{aligned} P\{3rd\ section\ is\ red|X_0 &= 2\} = \sum_i P\{3rd\ section\ is\ red|X_2 = i, X_0 = 2\}P\{X_2 = i|X_0 = 2\} \\ &= (0)p_{20}^{(2)} + (0.5)p_{21}^{(2)} + (1)p_{22}^{(2)} = (0.5)(0.32) + (1)(0.66) \\ &= 0.16 + 0.66 = 0.82. \end{aligned}$$

3.4 Classification of States

See **more references** in [1, 12].
吸收态　i通j　i,j相通或互通　等价类　不可约链

Definition 3.4.1 *I is the state space of Markov chain $\{X_n\}$.*
(a) If $p_{ii} = 1$, then i is called an absorbing state.
(b) If $\exists n \geq 0$, s.t. $p_{ij}^{(n)} > 0$, then i is said to be accessible to state j. (denoted by $i \to j$)
(c) If $i \to j$ and $j \to i$, then they are said to communicate ($i \leftrightarrow j$).
(d) Two states that communicate are said to be in the same class.
(e) The Markov chain is said to be irreducible if there is one class in the chain, that is, all states communicate with each other.

Remark 3.4.2 *If i not accessible to j, then*

$$\begin{aligned} P(ever\ enter\ j|start\ in\ i) &= P(\bigcup_{n=0}^{\infty}\{X_n = j\}|X_0 = i) \\ &\leq \sum_{n=0}^{\infty} P(X_n = j|X_0 = i) = \sum_{n=0}^{\infty} p_{ij}^{(n)} = 0. \end{aligned}$$

Proposition 3.4.3 *The relation of communication satisfies the following three properties:*
(i) State i communicates with state i, all $i \geq 0$.
(ii) If state i communicates with state j, then state j communicates with state i.
(iii) If i communicates with j, j communicates with k, then state i communicates with state k.

Proof. (i)

$$p_{ii}^{(0)} = P(X_0 = i | X_0 = i) = 1.$$

(iii) We prove that if $i \to j, j \to k$, then $i \to k$. There exist n and m s.t. $p_{ij}^{(n)} > 0, p_{jk}^{(m)} > 0$, then

$$p_{ik}^{(n+m)} = \sum_r p_{ir}^{(n)} p_{rk}^{(m)} \geq p_{ij}^{(n)} p_{jk}^{(m)} > 0.$$

■

Example 3.4.4 *$S = \{1, 2, 3\}$, where 1 is good, 2 is normal, and 3 is wrong.*

$$\mathbf{P} = \begin{bmatrix} 17/20 & 2/20 & 1/20 \\ 0 & 9/10 & 1/10 \\ 0 & 0 & 1 \end{bmatrix}.$$

One can see that 3 is different from 1 and 2. Each state is a class.

Example 3.4.5 *How many classes?*
(1) $S = \{1, 2, 3\}$.

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 0 & 3/4 \\ 0 & 2/3 & 1/3 \end{bmatrix}.$$

(2) $S = \{1, 2, 3, 4, 5\}$.

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.1 & 0 & 0.3 & 0 \\ 0.2 & 0.5 & 0.1 & 0.2 & 0 \\ 0.2 & 0.2 & 0.4 & 0.1 & 0.1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

(3) $S = \{1, 2, 3, 4, 5\}$.

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Sol. (1) irreducible chain. One class.
(2) {1,2,3} {4} {5}
(3) {1,2} {3,4,5}.

3.4.1 recurrent and transient states and classes

常返与非常返

- For state i, f_{ii} denotes the prob. that, starting in state i, the process will ever re-enter state i.
- Def. State i is recurrent if $f_{ii} = 1$.

State i is transient if $f_{ii} < 1$.

• Recurrent means that with prob. 1, the process will re-enter state i. By Markov chain, the process will start over again. （离开回来）

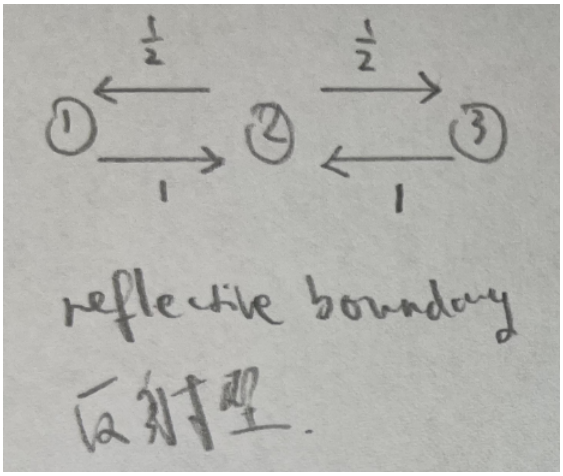
• If i is recurrent, then the process will re-enter state i again and again and again. **Infinitely many times !!!**

• Transient: With prob. f_{ii}, the process will re-enter state i, and with prob. $1 - f_{ii}$, it will never again enter state i.

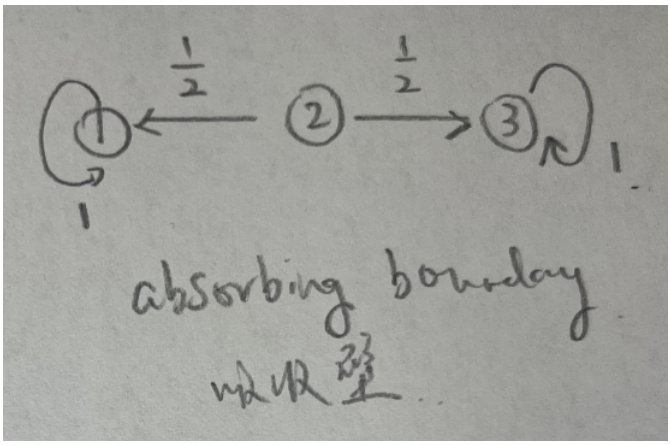
• The prob. that the process in state i for exactly n times equals $f_{ii}^{n-1}(1 - f_{ii})$, $n \geq 1$.

• If i is transient, the number of time periods that the process will be in state i has a geometric distribution with a finite mean $\frac{1}{1-f_{ii}}$.

Example 3.4.6 *All states 1,2,3 are recurrent.*



Example 3.4.7 *States 1 and 3 are recurrent, and state 2 is transient.*



Definition 3.4.8 *The first passage time probability* （首达概率） *is defined as*

$$\begin{aligned} f_{ij}^{(1)} &= P(X_1 = j | X_0 = i), \\ f_{ij}^{(n)} &= P(X_n = j, X_k \neq j, 1 \leq k \leq n-1 | X_0 = i), n \geq 1. \end{aligned}$$

Definition 3.4.9 $f_{ij} = \sum_{n=1*}^{\infty} f_{ij}^{(n)}$ *represents the prob. that one starts from* i *and first-time arrives at* j *after fnite-time steps. In fact, let* $A_1 = \{X_1 = j\}, A_n = \{X_n = j, X_k \neq j, 1 \leq k \leq n-1\}$ *be mutually exclusive events.* $\bigcup_{n=1}^{\infty} A_n$ *means that one ever enters* j*. Then*

$$f_{ij} = \sum_{n=1*}^{\infty} f_{ij}^{(n)} = \sum_{n=1*}^{\infty} P(A_n|X_0 = i) = P(\bigcup_{n=1}^{\infty} A_n|X_0 = i) \leq 1,$$

satisfies the property of probability.

Definition 3.4.10 *State* i *is recurrent if* $f_{ii} = 1$*, and state* i *is transient if* $f_{ii} < 1$*.*

Remark 3.4.11 当$f_{ii} = 1$时，质点从i出发以概率1回到i，再出发，再回到i，必然回到i无穷次。当$f_{ii} < 1$时，质点从i出发以正概率$1 - f_{ii}$不再回到i，如果回到i，则再次从i出发以正概率$1 - f_{ii}$不再回到i，所以只要$f_{ii} < 1$，质点不回到i总会发生，因而最终离开i。

Example 3.4.12 $S = \{1, 2, 3\}$*, where 1 is good, 2 is normal, and 3 is wrong.*

$$\mathbf{P} = \begin{bmatrix} 17/20 & 2/20 & 1/20 \\ 0 & 9/10 & 1/10 \\ 0 & 0 & 1 \end{bmatrix}.$$

One can show that

$$\lim_{n\to\infty} \mathbf{P}^{(n)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

by seeting that $\mathbf{P} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$*, where*

$$\mathbf{V} = \begin{bmatrix} 1 & 0.89 & 0.58 \\ 0 & 0.45 & 0.58 \\ 0 & 0 & 0.58 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 0.85 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The absorbing state i *satisfying* $f_{33} = f_{33}^{(1)} = 1$ *is recurrent. State 1 with* $f_{11} = f_{11}^{(1)} + 0 + \cdots = \frac{17}{20}$ *and state 2 with* $f_{22} = f_{22}^{(1)} + 0 + \cdots = \frac{9}{10}$ *are transient.*

Theorem 3.4.13 *For* $\forall i, j \in I, n \geq 1$*,*

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{3.1}$$

Proof. Let $A_n = \{X_n = j, X_k \neq j, 1 \leq k \leq n-1\}$ be mutually exclusive events. Since $\{X_n = j\} \subset \bigcup_{n=1}^{\infty} A_k$, then

$$p_{ij}^{(n)} = P(X_n = j|X_0 = i) = \sum_{k=1}^{n} P(X_n = j|A_k, X_0 = i)P(A_k|X_0 = i) = \sum_{k=1}^{n} p_{jj}^{(n-k)} f_{ij}^{(k)}.$$

■

Theorem 3.4.14 *Markov chain* $\{X_n\}$*, then*

(1)

$$\sum_{n=0*}^{\infty} p_{ii}^{(n)} = \frac{1}{1 - f_{ii}}.$$

(2) i *is recurrent* $\Leftrightarrow \sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$*.* i *is transient* $\Leftrightarrow \sum_{n=0}^{\infty} p_{ii}^{(n)} = \frac{1}{1-f_{ii}} < \infty$*.*

(3) If i *is recurrent and* $i \to j$*, then* $i \leftrightarrow j$*, and* j *is also recurrent. Recurrence is a class property.*

Proof. (1) Take $j = i$ in Eq. (3.1), one has

$$p_{ii}^{(n)}\rho^n = \sum_{k=1}^{n} f_{ii}^{(k)}\rho^k p_{ii}^{(n-k)}\rho^{n-k},$$

where $\rho \in (0,1), 1 \le k \le n, 1 \le n \le \infty$. By summing over n, we obtain

$$\begin{aligned} G(\rho) &: = \sum_{n=0}^{\infty} p_{ii}^{(n)}\rho^n = 1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\rho^n = 1 + \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ii}^{(k)}\rho^k p_{ii}^{(n-k)}\rho^{n-k} \\ &= 1 + \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} f_{ii}^{(k)}\rho^k p_{ii}^{(n-k)}\rho^{n-k} = 1 + \left(\sum_{k=1}^{\infty} f_{ii}^{(k)}\rho^k\right)\left(\sum_{n-k=0}^{\infty} p_{ii}^{(n-k)}\rho^{n-k}\right) \\ &= 1 + F(\rho)G(\rho), \end{aligned}$$

where $F(\rho) = \sum_{k=1}^{\infty} f_{ii}^{(k)}\rho^k$. Then

$$G(\rho) = \frac{1}{1 - F(\rho)}.$$

Let $\rho \to 1$, done.

(2) Direct corollary.

(3) $i \to j$说明从i到j的概率是正数，每次质点从i出发达到j，由于i常返，所以质点一定会回到i，因此j通i，因此$i \leftrightarrow j$。 Since i communicates with j, there exist integers k and m such that $p_{ij}^{(k)} > 0, p_{ji}^{(m)} > 0$. Notice that for any integer n,

$$p_{jj}^{(m+n+k)} \ge p_{ji}^{(m)} p_{ii}^{(n)} p_{ij}^{(k)}.$$

Summing over all n,

$$\sum_{n=1}^{\infty} p_{jj}^{(m+n+k)} \ge p_{ji}^{(m)} p_{ij}^{(k)} \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty,$$

from which we can conclude that j is recurrent. ■

Remark 3.4.15 *$\sum_{n=0}^{\infty} p_{ii}^{(n)}$ is the expected number of times that one starts from i and re-enter i (or say the expected number of time periods that the process is in state i). To see this, let*

$$I_n = \begin{cases} 1, & X_n = i \\ 0, & X_n \neq i \end{cases}$$

$\sum_{n=0}^{\infty} I_n$ represents the nubmer of times that the process in state i.

$$E[\sum_{n=0}^{\infty} I_n | X_0 = i] = \sum_{n=0}^{\infty} E[I_n | X_0 = i] = \sum_{n=0}^{\infty} P[X_n = i | X_0 = i] = \sum_{n=0}^{\infty} p_{ii}^{(n)}.$$

Remark 3.4.16

(1) recurrent $\Leftrightarrow$ i can be visited infinitely many times.

(2) transient $\Leftrightarrow$ i can be visited finite times.

Remark 3.4.17 *For finite-state MC, not all states can be transient, so that at least one of the states must be recurrent. Assume that all M states are transient, then say after T_1, state 1 will never be visited. So for each state i. Then, after $T = \max\{T_1, \ldots, T_M\}$, no state will be visited. However, the process must be in some state after T. Contradiction!*

Remark 3.4.18 *Transience is a class property. If state i is transient and communicates with state j, then state j must also be transient. This is a corollary of the conclusion that recurrence is a class property.*

Remark 3.4.19 *The result that not all states in a finite MC can be transient leads to the conclusion that all states of a finite irreducible Markov chain are recurrent.*

Example 3.4.20 *Let the Markov chain consisting of the states, 0,1,2,3 have the transition matrix*

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Determine which states are transient and which are recurrent.
Sol. All states communicate, and hence since this is a finite chain, all states must be recurrent.

Example 3.4.21 *Consider the Markov chain with states 0,1,2,3,4 and*

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ & & \frac{1}{2} & \frac{1}{2} & & \\ & & \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{4} & \frac{1}{4} & & & \frac{1}{2} & \end{bmatrix},$$

then the chain has three classes {0,1}, {2,3}, and {4}. The first two classes are recurrent and the third transient. For state 1, $f_{11} = \sum_{i\geq 1} f_{11}^{(i)} = \sum_{i\geq 1} \left(\frac{1}{2}\right)^i = 1$ since a particle once leaves state 1 and then can stay at state 2 for any times. For state 5, $f_{44} = f_{44}^{(1)} = \frac{1}{2}$.

Example 3.4.22 *In the chain with $\{0, 1, \ldots, n\}$ with two-sided absorbing boundaries. $\{1, 2, \ldots, n-1\}$ is in the transient class, and $\{0\}$ and $\{n\}$ are in the recurrent class. For transient class, one can see that if particle starts from state 1 and then never comes back to 1 with prob. $1 - f_{11} > P\{X_1 = 0 | X_0 = 1\} = q > 0$.*

Example 3.4.23 *For the chain with $\{0, 1, \ldots, n\}$ with two-sided reflection boundaries. Since all state communicate, they are in the same class. Moreover, due to finite number of states, thus all states are recurrent.*

Example 3.4.24 ***(A Random Walk)*** *Consider a Markov chain with infinite length $i = 0, \pm 1, \pm 2, \ldots$ and has transition prob. by*

$$p_{i,i+1} = p = 1 - p_{i,i-1}, \text{ for all } i.$$

One colorful interpretation of thsi process is that it represents the ***wanderings of a drunken man*** *as he walks along a straight line. Since all states communicate, they are all either recurrent or transient. Let us consider state 0 and determine if $\sum_{n=1}^{\infty} p_{00}^{(n)}$ is finite or infinite. It is obvious for odd steps that $p_{00}^{(2n-1)} = 0$ for all n. For even steps,*

$$p_{00}^{(2n)} = C_{2n}^n p^n (1-p)^n.$$

Using the Stirling formula

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi},$$

we obtain

$$p_{00}^{(2n)} \sim \frac{(4p(1-p))^n}{\sqrt{\pi n}}.$$

Hence, $\sum_{n=1}^{\infty} p_{00}^{(n)}$ *will converge if and only if*

$$\sum_{n=1}^{\infty} \frac{(4p(1-p))^n}{\sqrt{\pi n}}$$

does. Using Abel-Dirichlet determinant method, one can see that $\sum_{n=1}^{\infty} p_{00}^{(n)} = \infty$ *if and only if* $p = 1/2$ *(recurrent). We call it a* ***symmetric random walk****. When* $p \neq 1/2$*, the chain is transient.*

Example 3.4.25 *For 2D symmetric random walk, the transition prob is given by*

$$p_{up} = p_{down} = p_{left} = p_{right} = \frac{1}{4}.$$

The chain is irreducible. All states will be recurrent if state $(0,0)$ *is recurrent. Then*

$$\begin{aligned} p_{00}^{(2n)} &= \sum_{i=0}^{n} \frac{(2n)!}{i!i!(n-i)!(n-i)!} \left(\frac{1}{4}\right)^{2n} \\ &= \left(\frac{1}{4}\right)^{2n} \sum_{i=0}^{n} \frac{(2n)!}{n!n!} \frac{n!}{i!(n-i)!} \frac{n!}{i!(n-i)!} \\ &= \left(\frac{1}{4}\right)^{2n} \frac{(2n)!}{n!n!} C_{2n}^{n}. \end{aligned}$$

Now

$$C_{2n}^{n} \sim \frac{4^n}{\sqrt{\pi n}},$$

so that

$$p_{00}^{(2n)} \sim \frac{1}{\pi n}.$$

The series is divergent, and thus all states are recurrent.

Example 3.4.26 *For dimension greater than or equal to 3 symmetric random walk, all states are transient. Take* $K_n = \{j,k | j,k \geq 0, j+k \leq n\}$*, then*

$$\begin{aligned} p_{00}^{(2n)} &= \sum_{j,k \in K_n} \frac{(2n)!}{(j!k!(n-j-k)!)^2} \left(\frac{1}{6}\right)^{2n} \\ &= \frac{1}{6^{2n}} C_{2n}^{n} \sum_{j,k \in K_n} \frac{n!n!}{(j!k!(n-j-k)!)^2} \\ &\leq \frac{1}{6^{2n}} C_{2n}^{n} \frac{n!}{\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!} \sum_{j,k \in K_n} \frac{n!}{j!k!(n-j-k)!} \\ &= \frac{1}{6^{2n}} C_{2n}^{n} \frac{n!}{\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!} (1+1+1)^n. \end{aligned}$$

Using Stirling,

$$\frac{n!}{\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!\left(\frac{n}{3}\right)!} = O\left(\frac{n^{n+1/2}e^{-n}}{[(n/3)^{n/3+1/2}e^{-n/3}]^3}\right) = O\left(\frac{3^{n+3/2}}{n}\right),$$

then

$$p_{00}^{(2n)} = O\left(\frac{1}{6^{2n}} \frac{4^n}{\sqrt{n}} \frac{3^{n+3/2}}{n} 3^n\right) = O(n^{-3/2}),$$

eventually gives a convergent series. The symmetric random walk is transient. In general, for $d \geq 3$, $\sum_n n^{-d/2}$ *is convergent, which gives transient results. However, intuitively, it is hard to think about why the behavior is quite different between* $d \leq 2$ *and* $d \geq 3$. *This result is known as Polya theorem.*

3.4.2 Periodicity

Definition 3.4.27 *For Markov chain* $\{X_n\}$, *define the periodicity:*
(a) If $\sum_{n=1}^{\infty} p_{ii}^{(n)} = 0$, *then the particle can never re-enter* i *once leaving* i *and* i *is said to have* ∞ *period.*
(b) State i *is said to have period* d *if* $p_{ii}^{(n)} = 0$ *whenever* n *is not divisible by* d ($d \nmid n$), *and* d *is the largest integer with this property. For instance, starting in* i, *it may be possible for the process to enter state* i *only at times 2,4,6,8,. . . , in which case state* i *has period* 2.
(c) A state with period 1 *is said to be aperiodic.*

非周期

Remark 3.4.28 *Let* d *be the period of state* i. *If* $p_{ii}^{(n)} > 0$, *then* $d|n$ *and* d *must be the largest integer with this property. Moreover, when state* i *has period* $d < \infty$, *then* $p_{ii}^{(nd)} > 0$ *is satisfied only for some* n, *instead for all* n.

Remark 3.4.29 *Let* $d < \infty$ *be the period of the state* i. *Then* d *is the maximum common divisor of the set* $D = \{n | p_{ii}^{(n)} > 0, n \geq 1\}$.

最大公约数

Example 3.4.30 *On a line, if a particle moves forward 1-step with prob.* 1/3, *backward 1-step with prob.* 1/3, *moves forward 2-step with prob.* 1/3. *Then all states are aperiodic. The reason is*

$$\begin{aligned} p_{ii}^{(2)} &\geq p_{i,i+1}p_{i+1,i} > 0, \\ p_{ii}^{(3)} &\geq p_{i,i-2}p_{i-2,i-1}p_{i-1,i} > 0, \end{aligned}$$

so that $2 = nd, 3 = md \Rightarrow d = 1$.

Example 3.4.31 *On a line, if a particle moves either forward 1-step with prob.* p *or backward 5-step with prob.* $1-p$. *Then the period is* 6 *for each state. If* $p_{ii}^{(n)} > 0$, *then assume move forward* k *times and backward* m *times resulting in* $k = 5m$ *so that* $n = k + m = 6m$. *Since* $p_{ii}^{(6)} > p^5 q > 0$, *so* $6 = md$ *and thus* $d \leq 6$. *The period of state* i *is* $d = 6$.

Theorem 3.4.32 *Periodicity is a class property. That is, if state* i *has period* d, *and state* i *and* j *communicate, then state* j *also has period* d.

Proof. There exist r and s such that $p_{ji}^{(r)} > 0$ and $p_{ij}^{(s)} > 0$. Let d and t be the periods of states i and j, respect. Then $\exists n = mt$, s.t. $p_{jj}^{(n)} > 0$. Then

$$p_{ii}^{(s+n+r)} \geq p_{ij}^{(s)} p_{jj}^{(n)} p_{ji}^{(r)} > 0.$$

Thus

$$d | s + n + r.$$

Also

$$p_{ii}^{(r+s)} \geq p_{ij}^{(s)} p_{ji}^{(r)} > 0 \Rightarrow d|s+r.$$

Therefore,

$$d|n \Rightarrow d|mt.$$

• If $m = 1$, then $d|t$.

• If there exist $(m_1, m_2) = 1$ among all m_1 and m_2, then $d|m_1 t$ $d|m_2 t \Rightarrow d|t$.

• If $(m_1, m_2) > 1$ for all m_1 and m_2, then state j has period $(m_1, m_2)t$, contradiction.

Similarly, $t|d$. Thus, $t = d$. ■

3.4.3 positive and null recurrent states and classes

正常返和零常返

Definition 3.4.33 *State i is said to be positive recurrent, if, starting in i, expected time until the process returns to state i is finite.*

Definition 3.4.34 *$T_i = n$ represents that one arrives at state i at nth step for the first time, that is,*

$$T_i = \begin{cases} \min\{n|X_n = i, n \geq 1\}, & \text{if } \bigcup_{n=1}^{\infty} \{X_n = i\} \text{ occurs,} \\ \infty, & \text{else.} \end{cases}$$

Then,

$$f_{ii}^{(n)} = P(T_i = n|X_0 = i) = P(X_n = i, X_k \neq i, 1 \leq k \leq n-1|X_0 = i).$$

平均回转时间

Definition 3.4.35 *Let $f_{ii} = 1$ and $X_0 = i$. Denote μ_i the expected time until the process return to state i or say the first return time. Then the **mean recurrence time** or **mean first-passage time** is*

$$\mu_i = E[T_i|X_0 = i] = \sum_{n=1}^{\infty} nP(T_i = n|X_0 = i) = \sum_{n=1}^{\infty} nf_{ii}^{(n)}.$$

State i is said to be positive recurrent if $\mu_i < \infty$; State i is said to be null recurrent if $\mu_i = \infty$.

Lemma 3.4.36 *Let state i have period d and mean recurrence time $\mu_i = E(T_i|X_0 = i)$. Then*

$$\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{\mu_i}.$$

Proof is hard.

Theorem 3.4.37 *Let state i be recurrent.*
(1) i is null recurrent $\Leftrightarrow \lim_{n\to\infty} p_{ii}^{(n)} = 0$.
(2) i is positive recurrent and aperiodic (ergodic) $\Leftrightarrow \lim_{n\to\infty} p_{ii}^{(n)} = \frac{1}{\mu_i} > 0$.

Proof. (1) "⇒" Since i is recurrent, then there exists some $n < \infty$ such that $p_{ii}^{(n)} > 0$. Thus, we have $d < n < \infty$.
i is null recurrent ⇒ by lemma $\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{\infty} = 0$. That is,

• when $d|m = nd \Rightarrow \lim_{m\to\infty} p_{ii}^{(m)} = 0$

• when $d \nmid m \Rightarrow p_{ii}^{(m)} = 0$ based on the definition of the period. Thus, one always has $\lim_{n\to\infty} p_{ii}^{(n)} = 0$.

"$\Leftarrow$" Assume that i is positive recurrent, then $\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{\mu_i} > 0$. Contradiction!

(2) "$\Rightarrow$" ergodic so that $d = 1$. By lemma, done.

"$\Leftarrow$" Since $\lim_{n\to\infty} p_{ii}^{(n)} = \frac{1}{\mu_i} > 0$, i is positive recurrent. Compare condition with lemma, then we see that $d = 1$. ■

Theorem 3.4.38 *Let state i be recurrent.*
(1) If i is null recurrent, and $i \to j$, then j is null recurrent. That is, the null recurrence is a class property.
(2) If i is positive recurrent, and $i \to j$, then j is positive recurrent. That is, the positive recurrence is a class property.

Proof. (1) First, since i recurrent and $i \to j$ so that $i \leftrightarrow j$. There exist integers m and n such that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$. Using the fact that $p_{ii}^{(n+k+m)} \geq p_{ij}^{(n)} p_{jj}^{(k)} p_{ji}^{(m)}$ and $\lim_{n\to\infty} p_{ii}^{(n)} = 0$, we have that

$$p_{jj}^{(k)} \leq \frac{1}{p_{ij}^{(n)} p_{ji}^{(m)}} p_{ii}^{(n+k+m)} \to 0, \text{ as } k \to \infty.$$

(2) Disproof. Assume j is null recurrent, then i is also null recurrent. However, i is positive recurrent. Contradiction! ■

3.4.4 ergodicity

Definition 3.4.39 *Positive recurrent, aperiodic states are called ergodic.*

Example 3.4.40 *Consider the Markov chain. See Fig. 3.1. Then, states 3 and 4 are transient. States 1*

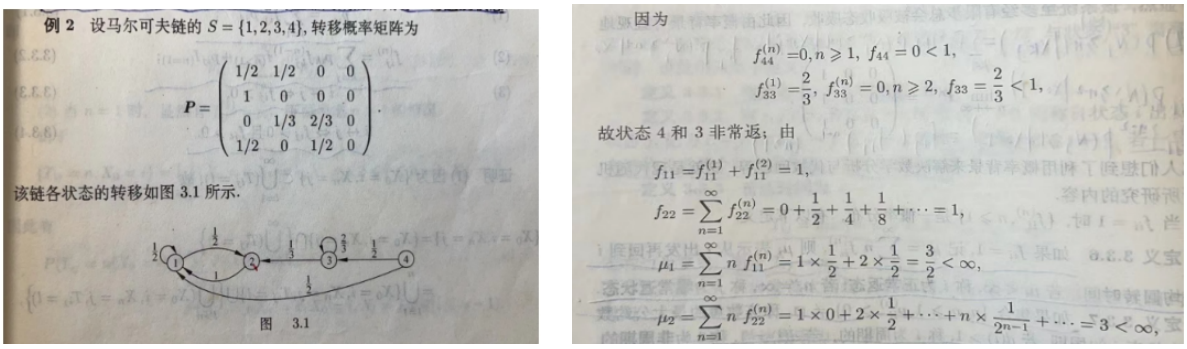


Figure 3.1: Example for classification of transient and ergodic classes.

and 2 are positive recurrent, in addition, they are aperiodic, and thus they are ergodic.

Here is the summary of this subsection (see Fig. 3.2).

3.4.5 Classification

Definition 3.4.41 *I is state space of Markov chain $\{X_n\}$ and $B \subset I$. If any state $b \in B$ is not accessible to $B^c = I - B$, then B is said to be a closed set.*

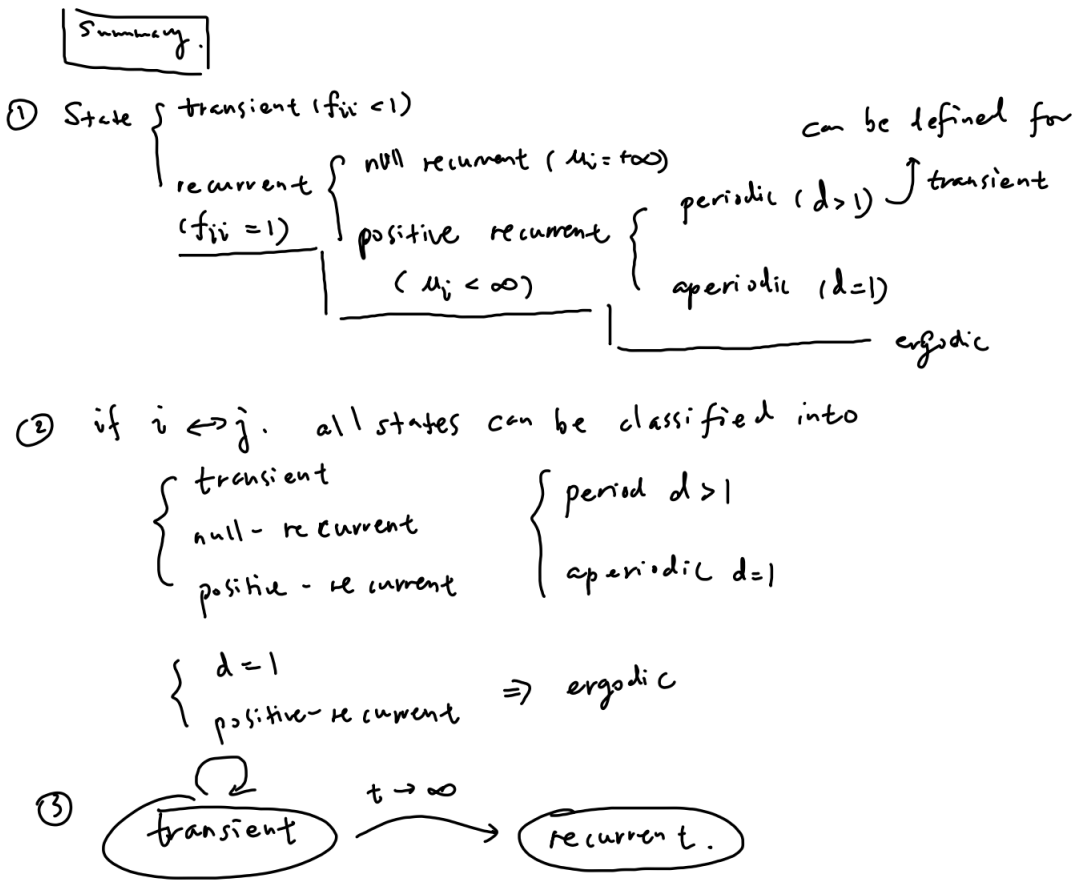


Figure 3.2: Summary of classification.

Theorem 3.4.42 *For Markov chain,*

(1) There is no intersection between two different classes.

(2) Each state belongs to one of the following class, transient, null recurrent, or positive recurrent. Moreover, all states in one class have the same properties.

(3) Recurrent class is a closed set: all particles in the recurrent class cannot leave it.

(4) Null recurrent class has infinitely many states.

(5) If transient class is closed, then it contains infinitely many states. (see. e.g., non-symmetric random walk with drift.)

Lemma 3.4.43 *If j is transient or null recurrent, then for any $i \in I$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0. \tag{3.2}$$

Proof. Proof for the Lemma. If j is null recurrent, then $\lim_{n\to\infty} p_{jj}^{(n)} = 0$. If j is transient, then $\sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty$ gives also $\lim_{n\to\infty} p_{jj}^{(n)} = 0$. For any $i \in I$,

$$p_{ij}^{(n)} = \sum_{k=1}^{m} f_{ij}^{(k)} p_{jj}^{(n-k)} + \sum_{k=m+1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} \le \sum_{k=1}^{m} p_{jj}^{(n-k)} + \sum_{k=m+1}^{n} f_{ij}^{(k)}.$$

Let $n \to \infty$, then the first term on RHS $\sum_{k=1}^{m} p_{jj}^{(n-k)} \to 0$. Thus,

$$\lim_{n\to\infty} p_{ij}^{(n)} \le \sum_{k=m+1}^{\infty} f_{ij}^{(k)} \le 1, \text{ (bounded)}$$

Let $m \to \infty$, then

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

■

Proof. We only prove for (4) and (5) in Theorem 3.4.42. Let C be a class of either null recurrent or closed transient states. Based on (3), null recurrent is closed. Thus, C is always closed, that is, if $i, j \in C$, then i, j can never leave the closed C. Notice that for any $i \in C$,

$$\sum_{j\in C} p_{ij}^{(n)} = \sum_{j\in I} p_{ij}^{(n)} = 1, \quad \text{for any } n.$$

We now use the result $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for transient or null recurrent j from above Lemma. If C contains only finite states, then

$$1 = \lim_{n\to\infty} \sum_{j\in C} p_{ij}^{(n)} = \sum_{j\in C} \lim_{n\to\infty} p_{ij}^{(n)} = \sum_{j\in C} 0 = 0.$$

Contradiction! We conclude that C contains infinitely many states. ■

Theorem 3.4.44 *State space I can be decomposed as*

$$I = \bigcup_{j=1}^{m} C_j + T, \quad m \le \infty$$

where C_j is recurrent class and T is transient class.

Remark 3.4.45 *For a transient class T,*

(1) A particles can be in T forever, or can move from T to some C_j and then stay in C_j forever.

(2) If T has finite states, then the particle will always leave T, enter some closed C_j and stay in C_j forever.

On the other hand, for a recurrent class C_j,

(1) C_j has finite states $\Rightarrow$ all states are positive recurrent.

(2) C_j is null recurrent $\Rightarrow$ C_j contains infinitely many states.

Remark 3.4.46 *For Markov chain,*

(1) For finite-state Markov chain, either positive recurrent or transient, impossible to be null recurrent. (e.g., random walk model with absorbing boundaries)

(2) In a finite-state Markov chain, all recurrent states are positive recurrent.

(3) For irreducible finite-state Markov chain, all states are positive recurrent. (e.g., random walk model with reflection)

(4) For infinite-length irreducible Markov chain, all states can be

$$\begin{cases} \textit{transient}, & \textit{random walk } p \neq \frac{1}{2}, \\ \textit{null recurrent}, & \textit{random walk } p = \frac{1}{2}, \\ \textit{positive recurrent}, & \textit{e.g., see following example.} \end{cases}$$

(5) Each absorbing state is positive recurrent since $\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} = 1 \cdot f_{ii}^{(1)} = 1 < \infty$. Infinite-length can have positive recurrent states. (e.g., each state is an absorbing state)

Example 3.4.47 *Consider the Markov chain with infinitely many states (see Fig. 3.3). Then, all states are*

例 4 设马尔可夫链 $S=\{1,2,3,\cdots\}$, 转移概率为 $p_{11}=\frac{1}{2}, p_{ii+1}=\frac{1}{2}, p_{i1}=\frac{1}{2}, i\in S$. 分析状态 1. 由如下状态转移图 3.5 易知:

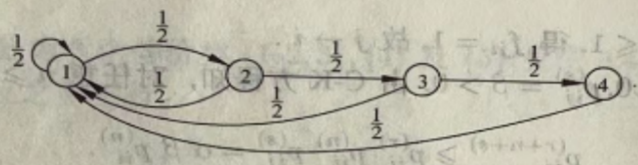


图 3.5

$$f_{11}^{(1)}=\frac{1}{2},\quad f_{11}^{(2)}=\left(\frac{1}{2}\right)^2,\quad f_{11}^{(3)}=\left(\frac{1}{2}\right)^3,\quad \cdots,\quad f_{11}^{(n)}=\left(\frac{1}{2}\right)^n.$$

故 $f_{11}=\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n=1$, 所以 "1" 是常返状态. 又 $\mu_1=\sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n<\infty$, 所以 "1" 是正常返状态. 再由 $p_{11}^{(1)}=\frac{1}{2}>0$, 知 "1" 是非周期的. 从而 "1" 是遍历状态. 对其他 $i\neq 1$, 因 $i\leftrightarrow 1$, 故 i 也是遍历状态 (若求 $f_{ii}^{(n)}$ 则较麻烦)

Figure 3.3: All states are ergodic (positive recurrent and aperiodic) for a Markov chain with infinitely many states.

ergodic.

3.5 Stationary Distributions, Limiting Probabilities, Limiting Distributions

For the two-state Markov chain with the transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

It turns out to be

$$\mathbf{P}^{(4)} = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}, \mathbf{P}^{(8)} = \begin{bmatrix} 0.572 & 0.428 \\ 0.570 & 0.430 \end{bmatrix}.$$

They are almost identical. In fact it seems that $p_{ij}^{(n)}$ converges to some value as $n \to \infty$. In other words, there seems to exist a limiting prob. that the process will be in state j after a large number of transitions, and this value is independent of the initial state.

The question is:

(a) If stationary distribution exists and if it is unique? What is the value of the stationary distribution?

(b) If limiting probability $\lim_{n\to\infty} p_{ij}^{(n)}$ exists? If exists, what is its value?

(c) For each i, if $\lim_{n\to\infty} \pi_i^{(n)}$ exists? If $\lim_{n\to\infty} \pi^{(n)}$ exists, where $\pi^{(n)} = (\pi_1^{(n)}, \pi_2^{(n)}, \ldots)$? What is its value?

3.5.1 Stationary Distributions

Definition 3.5.1 *If a probability distribution $\pi = (\pi_1, \pi_2, \ldots)$ satisfies*

$$\sum_{j\in I} \pi_j = 1, \pi = \pi\mathbf{P},$$

or equivalently

$$\sum_{j\in I} \pi_j = 1, \pi_j = \sum_{k\in I} \pi_k p_{kj} \geq 0, j \in I,$$

then π_j is called stationary probability and π is called the stationary prob. distribution. Obviously,

$$\pi = \pi\mathbf{P} = \cdots = \pi\mathbf{P}^n.$$

See reference in [2].

Theorem 3.5.2 *Let C^+ contain all positive recurrent states of the Markov chain $\{X_n\}$ and $i \in C^+$.*
Part *(1) If C^+ has only one ergodic class, then the stationary prob. π_j satisfies*

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}, \quad j \in I,$$

is the unique stationary distribution.

Proof. Define

$$\mu_j = \begin{cases} \infty, & j \text{ is null recurrent or transient}, \\ \text{some positive finite number}, & j \text{ is positive recurrent}. \end{cases}$$

ergodic $\Rightarrow d = 1$. Based on the result in some example (think about why for i and j, the following holds true?),

$$\lim_{n\to\infty} p_{ij}^{(n)} = \begin{cases} 0, & j \notin C^+, \\ \frac{d}{\mu_j} = \frac{1}{\mu_j}, & j \in C^+. \end{cases}$$

A particle starting from $i \in C^+$ can never leave C^+, thus

$$\sum_{j\in I} p_{ij}^{(n)} = \sum_{j\in C^+} p_{ij}^{(n)} = 1. \tag{3.3}$$

Define

$$\pi_j := \lim_{n\to\infty} p_{ij}^{(n)},$$

we would like to show that π_j is a stationary probability, that is, π_j satisfies the equations in Definition 3.5.1.
Existence of stationary distribution. Let $n \to \infty$ in Eq. (3.3), one has

$$\sum_{j\in I} \pi_j = \sum_{j\in C^+} \pi_j = 1.$$

(think about why can be interchanged?) Using Chapman-Kolmogorov equation,

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{k\in C^+} p_{ik}^{(n-1)} p_{kj} = \sum_{k\in C^+} \pi_k p_{kj} = \sum_{k\in I} \pi_k p_{kj},$$

where made use of Dominated Convergence Theorem. Thus, π_j is the stationary distribution.
Uniqueness. Suppose that $\{v_j\}$ is another stationary distribution. If $j \notin C^+$, using the property of stationary distribution $v{=}v\mathbf{P}^n$ and assume that $n \to \infty$ and $\sum_{k\in I}$ can be interchanged, we obtain:

$$v_j = \lim_{n\to\infty} \sum_{k\in I} v_k p_{kj}^{(n)} = \sum_{k\in I} \lim_{n\to\infty} v_k p_{kj}^{(n)} = 0 \Rightarrow v_j = 0 = \lim_{n\to\infty} p_{ij}^{(n)} := \pi_j,$$

where made use of the fact that for any $j \notin C^+$, one has $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for any $i \in I$ (see equation (3.2)). If $j \in C^+$,

$$v_j = \sum_{k\in I} v_k p_{kj}^{(n)} = \sum_{k\in C^+} v_k p_{kj}^{(n)} \to \left(\sum_{k\in C^+} v_k \right) \pi_j = \pi_j, \text{ as } n \to \infty,$$

where the second equality holds true since $v_k = 0$ if $k \notin C^+$, and the last equality holds true since there is only one C^+. Therefore, $v_j = \pi_j$. ■

Theorem 3.5.3 Part *(2) If C^+ is a class with period d, then*

$$\pi_j = \frac{1}{d} \lim_{n\to\infty} p_{jj}^{(nd)} = \lim_{n\to\infty} \frac{1}{d} \sum_{s=1}^{d} p_{ij}^{(nd+s)} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)}, j \in I,$$

is the unique stationary distribution. Moreover, $\pi_j = \frac{1}{\mu_j}$.

Theorem 3.5.4 Part *(3) $\{X_n\}$ has a unique stationary distribution. $\Leftrightarrow$ C^+ is an equivalent class. $\Leftrightarrow$ there exists a unique class of positive recurrent states.*

Theorem 3.5.5 Part *(4) Stationary distribution exists for $\{X_n\} \Leftrightarrow C^+ \neq \varnothing$.*

Theorem 3.5.6 Part *(5) Corollary of (4): A Markov chain with finite states has at least one stationary distribution.*

Proof. $C^+ \neq \varnothing$ since at least one state is recurrent and further positive recurrent (null recurrence has infinitely many states). ■

Theorem 3.5.7 Part *(6) Corollary of (1): An irreducible and aperiodic MC with finite states has a unique stationary distribution.*

Proof. finite states + irreducible ⇒positive recurrent. positive recurrent+aperiodic⇒ergodic. Thus, ∃!stationary distribution. ■

Theorem 3.5.8 Part *(7) Structure of* $\mathbf{P}$*. The transition matrix is*

$$\mathbf{P} = \begin{array}{c} \begin{array}{ccccc} C_1 & C_2 & \cdots & C_m & T \end{array} \\ \begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{P}_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{P}_m & 0 \\ \mathbf{R}_1 & \mathbf{R}_2 & \cdots & \mathbf{R}_m & \mathbf{Q}_T \end{bmatrix} \end{array} \begin{array}{c} C_1 \\ C_2 \\ \vdots \\ C_m \\ T \end{array}$$

Since each C_j *is a closed set, row sum of each* $\mathbf{P}_j$ *is* 1*. That is to say each* C_j *is a irreducible MC, particle starting from* C_j *or move from* T *to* C_j *will always stay in* C_j*. One also has*

$$\mathbf{P}^{(n)} = \begin{array}{c} \begin{array}{ccccc} C_1 & C_2 & \cdots & C_m & T \end{array} \\ \begin{bmatrix} \mathbf{P}_1^n & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{P}_2^n & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{P}_m^n & 0 \\ \mathbf{R}_1^{(n)} & \mathbf{R}_2^{(n)} & \cdots & \mathbf{R}_m^{(n)} & \mathbf{Q}_T^n \end{bmatrix} \end{array} \begin{array}{c} C_1 \\ C_2 \\ \vdots \\ C_m \\ T \end{array}$$

after n*-step transition. Row sum of each* $\mathbf{P}_j^n$ *is* 1*. Moreover,* $\lim_{n\to\infty} \mathbf{Q}_T^n = \lim_{n\to\infty} (q_{ij}^{(n)})_{i,j\in T} = 0$*, any transient state can not be stayed for long time.*

任何一个非常返态不可久留

Theorem 3.5.9 Part *(8) Corollary of (3) and (4). If* C^+ *has at least two positive recurrent classes,* C_1 *and* C_2*, then there are infinitely many stationary distribution.*

Proof. Assume that C_1 and C_2 correspond to transition matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. Then there are stationary distributions, π_1 and π_2, s.t.

$$\pi_1 = \pi_1 \mathbf{P}_1, \pi_2 = \pi_2 \mathbf{P}_2.$$

For any $r = 1 - s \in [0, 1]$, define

$$\pi = [r\pi_1, s\pi_2, 0, \cdots, 0].$$

Then,

$$\begin{aligned}
\pi\mathbf{P} &= [r\pi_1, s\pi_2, 0, \cdots, 0]\begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{P}_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{P}_m & 0 \\ \mathbf{R}_1 & \mathbf{R}_2 & \cdots & \mathbf{R}_m & \mathbf{Q}_T \end{bmatrix} \\
&= [r\pi_1\mathbf{P}_1, s\pi_2\mathbf{P}_2, 0, \cdots, 0] \\
&= [r\pi_1, s\pi_2, 0, \cdots, 0] = \pi.
\end{aligned}$$

■

3.5.2 The limiting behavior of transtion probability matrix

Here we restate the results for $\lim_{n\to\infty} p_{ij}^{(n)}$.

Theorem 3.5.10 *If j is transient or null recurrent, then for any $i \in I$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Corollary 3.5.11

(1) Finite-state Markov chain has not null recurrent states.

(2) All states of a finite-state irreducible Markov chain are positive recurrent.

(3) If a Markov chain has a null recurrent state, then it has infinitely many null recurrent states.

Lemma 3.5.12 *If j is positive recurrent, then $\lim_{n\to\infty} p_{ij}^{(n)}$ may not exist. Even it exists, it may be related to state i.*

Theorem 3.5.13 *If j is ergodic, then for $i \in I$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}.$$

Theorem 3.5.14 *For a irreducible ergodic chain, for any $i, j \in I$,*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}.$$

3.5.3 limiting probabilities and limiting distributions

Definition 3.5.15 *If $\lim_{n\to\infty} \pi_j^{(n)} = \pi_j^*$ $(j \in I)$ exists, then $\pi^* = (\pi_1^*, \pi_2^*, \ldots)$ is called the limiting distribution of the Markov chain.*

Theorem 3.5.16 *For an aperiodic irreducible Markov chain, all states are positive recurrent if and only if the chain has a stationary distribution. Moreover, the stationary distribution is the limiting distribution.*

Proof. $\Leftarrow$ Let the stationary distribution be $\pi = (\pi_1, \pi_2, \ldots)$, so that $\pi = \pi\mathbf{P} = \cdots = \pi\mathbf{P}^n$, that is,

$$\pi_j = \sum_{i\in I} \pi_i p_{ij}^{(n)}.$$

Using Dominated Convergence Theorem, one can interchange the limitation and the summation for each $j \in I$,

$$\pi_j = \lim_{n\to\infty} \sum_{i\in I} \pi_i p_{ij}^{(n)} = \sum_{i\in I} \pi_i \lim_{n\to\infty} p_{ij}^{(n)} = \sum_{i\in I} \pi_i \cdot 0 = 0,$$

if we assume that all states are either null recurrent or transient. However, since $\sum_{j\in I} \pi_j = 1$, at least one $\pi_j > 0$ is strictly positive. Contradiction! Thus, at least one state is positive recurrent. Since the Markov chain is irreducible, all states are positive recurrent.

$\Rightarrow$ Proved in Theorem 3.5.2.

Moreover, the conclusion that the stationary distribution is the limiting distribution is also proved in Theorem 3.5.2. ■

3.5.4 Examples for stationary distributions and limiting probabilities

Example 3.5.17 *Let the state space of a Markov chain be $I = \{1, 2\}$ and the transition prob. matrix*

$$\mathbf{P} = \begin{bmatrix} 3/4 & 1/4 \\ 5/8 & 3/8 \end{bmatrix}.$$

(1) Compute the stationary distribution π and the limiting $\lim_{n\to\infty} \mathbf{P}^n$;

Sol. Since $\pi = \pi\mathbf{P}$, compute the eigenvector of $\mathbf{P}^T$ corresponding to eigenvalue 1:

$$(\mathbf{P}^T - \mathbf{I}) \to \begin{pmatrix} 1 & -5/2 \\ 0 & 0 \end{pmatrix}.$$

Thus

$$\pi \propto \begin{pmatrix} 5/2 \\ 1 \end{pmatrix} = \begin{pmatrix} 5/7 \\ 2/7 \end{pmatrix},$$

and

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix} = \begin{pmatrix} 5/7 & 2/7 \\ 5/7 & 2/7 \end{pmatrix}.$$

(2) Compute the mean recurrence time μ_1 and μ_2.

Sol. Based on the formula, $\pi_j = 1/\mu_j$, so that $\mu_1 = 7/5$ and $\mu_2 = 7/2$.

Example 3.5.18 ***Ehrenfest model.*** *Suppose that $2a$ molecules are distributed among two urns; and each time point one of the molecules is chosen at random, removed from its urn, and placed in the other one. Let X_n be the number of molecules in urn 1. The number of molecules in urn 1 is the Markov chain having transition prob.*

$$p_{ij} = \begin{cases} \frac{2a-i}{2a}, & 0 \le i \le 2a-1, j = i+1. \\ \frac{i}{2a}, & 1 \le i \le 2a, j = i-1. \\ 0, & \text{else.} \end{cases}$$

Compute the stationary distribution π.

Sol. This is a positive recurrent Markov chain with period 2, and thus there exists a unique stationary distribution. Define $\pi_{-1} = \pi_{2a+1} = 0$. Based on $\pi = \pi\mathbf{P}$,

$$\pi_i = \pi_{i-1} p_{i-1,i} + \pi_{i+1} p_{i+1,i}, 0 \le i \le 2a.$$

Then

$$\pi_{i+1} = \frac{\pi_i - \pi_{i-1}p_{i-1,i}}{p_{i+1,i}}.$$

Based on computation recursively

$$\begin{aligned}
\pi_1 &= \frac{\pi_0}{p_{10}} = 2a\pi_0 = C_{2a}^1\pi_0, \\
\pi_2 &= (\pi_1 - \pi_0)2a/2 = (2a-1)a\pi_0 = C_{2a}^2\pi_0, \\
&\cdots \\
\pi_{2a} &= C_{2a}^{2a}\pi_0.
\end{aligned}$$

By induction,

$$\begin{aligned}
\pi_{i+1} &= \frac{\pi_i - \pi_{i-1}p_{i-1,i}}{p_{i+1,i}} = \frac{C_{2a}^i\pi_0 - C_{2a}^{i-1}\pi_0\frac{2a-(i-1)}{2a}}{\frac{i+1}{2a}} \\
&= \frac{\pi_0}{i+1}\left(2aC_{2a}^i - (2a-i+1)C_{2a}^{i-1}\right) = \frac{\pi_0}{i+1}\left(\frac{(2a)\cdot(2a)!}{i!(2a-i)!} - \frac{(2a)!\cdot(2a+1-i)}{(i-1)!(2a+1-i)!}\right) \\
&= \frac{\pi_0}{i+1}\frac{(2a)!}{(i-1)!(2a-i)!}\left(\frac{2a}{i} - 1\right) = \frac{\pi_0(2a)!(2a-i)}{(i+1)(i-1)!(2a-i)!i} = C_{2a}^{i+1}\pi_0.
\end{aligned}$$

Using $\pi_0 + \cdots + \pi_{2a} = 2^{2a}\pi_0 = 1$, *we obtain*

$$\pi_i = C_{2a}^i\left(\frac{1}{2}\right)^i\left(\frac{1}{2}\right)^{2a-i}, 0 \le i \le 2a.$$

This is a Binomial with $B(2a, 1/2)$ *since* $X_1 + \cdots + X_{2a} \sim B(2a, 1/2)$.

Example 3.5.19 *5. Example 4.21 ignored. too simple.*

Example 3.5.20 *6.* **The Hardy-Weinberg Law and a Markov chain in Genetics.** *Consider a large population of individuals, where each individual gene is either type A or type a. Assume that the proportions of individuals whose gene pairs are AA, aa, or Aa are* p_0, q_0, r_0 *(*$p_0 + q_0 + r_0 = 1$*). When two individuals mate, each contributes one of his or her genes, chosen at random, to the offspring.*
By conditioning on the gene pair of the parent, we see that for the first generation, a randomly chosen gene will be type A with prob.

$$\begin{aligned}
P\{A\} &= P\{A|AA\}P\{AA\} + P\{A|aa\}P\{aa\} + P\{A|Aa\}P\{Aa\} \\
&= p_0 + \frac{1}{2}r_0.
\end{aligned}$$

Similarly, it will be type a with prob.

$$P\{a\} = q_0 + \frac{1}{2}r_0.$$

Thus, under random mating a randomly chosen member of the next generation will be type AA with prob., where

$$p = P\{A\}P\{A\} = \left(p_0 + \frac{1}{2}r_0\right)^2.$$

The prob. for aa is

$$q = P\{a\}P\{a\} = \left(q_0 + \frac{1}{2}r_0\right)^2.$$

The prob. for Aa is

$$r = 2P\{A\}P\{a\} = 2\left(p_0 + \frac{1}{2}r_0\right)\left(q_0 + \frac{1}{2}r_0\right),$$

where we notice that $p + q + r = 1$.

We notice that ***the proportion of*** A ***will be unchanged from the previous generation****:*

$$\begin{aligned} P\{A^{new}\} &= p + \frac{1}{2}r = \left(p_0 + \frac{1}{2}r_0\right)^2 + \left(p_0 + \frac{1}{2}r_0\right)\left(q_0 + \frac{1}{2}r_0\right) \\ &= p_0 + \frac{1}{2}r_0 = P\{A^{old}\}. \end{aligned}$$

Similarly,

$$P\{a^{new}\} = q + \frac{1}{2}r = q_0 + \frac{1}{2}r_0 = P\{a^{old}\}.$$

From this it follows that, ***under random mating, in all successive generations after the initial one, the percentages of the population having gene pairs*** AA,aa, ***and*** Aa ***will remain fixed at the values*** $p, q,$ ***and*** r***. This is known as the Hardy-Weinberg law****.*

For instance, see the following example:

	1*st generation*	2*nd generation*
$p_0 = 0.1$	$p_1 = 0.25$	$p_2 = 0.25$
$q_0 = 0.1$	$q_1 = 0.25$	$q_2 = 0.25$
$r_0 = 0.8$	$r_1 = 0.5$	$r_2 = 0.5$

Let us now see the problem from another point of view. For a given individual, let X_n *denote the genetic state of her descendant in the* n*th generation. The transition prob. matrix of the Markov chain is*

$$\begin{array}{c} \\ AA \\ aa \\ Aa \end{array} \begin{array}{c} \begin{array}{ccc} AA & aa & Aa \end{array} \\ \begin{bmatrix} p + r/2 & 0 & q + r/2 \\ 0 & q + r/2 & p + r/2 \\ p/2 + r/4 & q/2 + r/4 & p/2 + q/2 + r/2 \end{bmatrix} \end{array},$$

where $AA \to AA$ *since* AA *must contribute one* A *and another* A *comes from* $P\{A\} = p + r/2$, *and* $Aa \to AA$ *since* $\frac{1}{2}P\{A\} = p/2 + r/4$. *Let us verify that the stationary limiting prob. of this Markov chain is* p, q, r. *It suffices to show that*

$$\begin{aligned} p &= p(p + \frac{r}{2}) + r(\frac{p}{2} + \frac{r}{4}) = (p + \frac{r}{2})^2 = (p_0 + \frac{r_0}{2})^2, \\ q &= q(q + \frac{r}{2}) + r(\frac{q}{2} + \frac{r}{4}) = (q + \frac{r}{2})^2 = (q_0 + \frac{r_0}{2})^2, \\ p + q + r &= 1. \end{aligned}$$

Example 3.5.21 *7. 4.5.1 The gambler's ruin problem ignored. Introduced in the following.*

Example 3.5.22 *8. 4.5.2 ignored since too hard.*

3.6 Mean Time Spent in Transient States

See reference in [7].

$\mathbf{P}$ transition prob. matrix for MC $\{X_n\}$ and $\mathbf{Q}$ is submatrix of $\mathbf{P}$ which includes only transient states.

$$\mathbf{P} = \begin{bmatrix} \tilde{\mathbf{P}}_0 & 0 \\ \mathbf{S} & \mathbf{Q} \end{bmatrix}, \quad \mathbf{P}^n = \begin{bmatrix} \tilde{\mathbf{P}}_0^n & 0 \\ \mathbf{S}_n & \mathbf{Q}^n \end{bmatrix}.$$

Example 3.6.1 *5 states with two absorbing boundaries.* $S = \{0, 4, 1, 2, 3\}$.

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 4 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 4 & 1 & 2 & 3 \end{array} \\ \left[\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right] \end{array}$$

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \end{array}.$$

$\mathbf{Q}$ is called a substochastic matrix, i.e., a mtrix with nonnegative entries whose row sums are less than or equal to 1.

$\mathbf{Q}$ is transient $\Rightarrow \mathbf{Q}^n = 0$ as $n \to \infty$. $\Rightarrow$All eigenvalues of $\mathbf{Q}$ have absolutely values strictly less than 1. $\Rightarrow \mathbf{I} - \mathbf{Q}$ is an invertible matrix. $\Rightarrow \mathbf{S} := (\mathbf{I} - \mathbf{Q})^{-1}$ is well-defined.

Let j be a T state and consider Y_j the total number of visits to j,

$$Y_j = \sum_{n=0}^{\infty} I\{X_n = j\}.$$

这些I事件可以相容 Since j is T, $Y_j < \infty$ with prob. 1. Suppose $X_0 = i$, where is i another T. Then

$$\begin{aligned} s_{ij} &= E(Y_j | X_0 = i) = E(\sum_{n=0}^{\infty} I\{X_n = j\} | X_0 = i) \\ &= \sum_{n=0}^{\infty} P(X_n = j | X_0 = i) = \sum_{n=0}^{\infty} p_{ij}^{(n)}, \quad 0 \text{ included.} \end{aligned}$$

That is, $E(Y_j | X_0 = i)$ is the (i, j) entry of matrix

$$\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots,$$

which is the same as the (i, j) entry of matrix

$$\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots \quad \text{(because Block property)}.$$

However, a simple calculation shows that

$$\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots = (\mathbf{I} - \mathbf{Q})^{-1} = \mathbf{S}.$$

Theorem 3.6.2 *Let i, j be T. Then*
(1) $s_{ij} = [\mathbf{S}]_{ij} = [(\mathbf{I} - \mathbf{Q})^{-1}]_{ij}$ is the expected number of visits to j starting at i.
(2) The expected number of steps until the chain enters a recurrent class can be computed by summing s_{ij} over all transient j in one T class.

Example 3.6.3 *cont.*

$$\mathbf{S} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix}.$$

The expected number of visits from 1 to 3 is $1/2$. *The expected number of visits from 1 until absorption is* $\frac{3}{2} + 1 + \frac{1}{2} = 3$.

Another derivation of the above result. For $i, j \in T$, let s_{ij} be the expected number of visits that the MC is in j, starting from i. Let

$$\delta_{ij} = 1, \quad i = j$$
$$\delta_{ij} = 0, \quad i \neq j$$

Then

$$\begin{aligned} s_{ij} &= E(Y_j | X_0 = i) = \delta_{ij} + \sum_{k \in I} E(Y_j | X_1 = k, X_0 = i) P(X_1 = k | X_0 = i) \\ &= \delta_{ij} + \sum_{k \in I} p_{ik} s_{kj} = \delta_{ij} + \sum_{k \in T} p_{ik} s_{kj} \quad (\text{since if } k \to j \in T, \text{ then } k \in T). \end{aligned}$$

In matrix form,

$$\mathbf{S} = (\mathbf{I} - \mathbf{Q})^{-1}. \text{ (notice that } s_{jj} \geq 1).$$

In the following, we derive another useful formula. For $i, j \in T$, the quantity f_{ij}, equal to the prob. that the MC ever makes a transition into state j given that it starts in state i. Derive an expression for s_{ij} by conditioning on whether state j is ever entered,

$$A_1 = \{\text{ever transit to } j\}, A_2 = \{\text{never transit to } j\}.$$

$$\begin{aligned} s_{ij} &= E(\text{times in } j | \text{start in } i, \text{ ever transit to } j) f_{ij} \\ &\quad + E(\text{times in } j | \text{start in } i, \text{ never transit to } j)(1 - f_{ij}) \\ &= (\delta_{ij} + s_{jj}) f_{ij} + \delta_{ij}(1 - f_{ij}) \\ &= \delta_{ij} + f_{ij} s_{jj}. \end{aligned}$$

Thus

$$f_{ij} = \frac{s_{ij} - \delta_{ij}}{s_{jj}}.$$

Example 3.6.4 ***(Gambler Ruin Problem)*** *Consider a gambler who at each play has prob. p of winning one unit and prob. $q = 1 - p$ of losing one unit. Assuming that successive plays are independent. Now given $p = 0.4$ and $N = 7$. Starting with 3 units, what is the prob. that the gambler ever has a fortune of 1?*

Sol. We have a 6×6 *matrix*

$$\mathbf{Q} = \begin{bmatrix} & 0.4 & & & & \\ 0.6 & & 0.4 & & & \\ & 0.6 & & 0.4 & & \\ & & 0.6 & & 0.4 & \\ & & & 0.6 & & 0.4 \\ & & & & 0.6 & \end{bmatrix},$$

then one obtains $\mathbf{S} = (\mathbf{I} - \mathbf{Q})^{-1}$, *and hence* $s_{3,1} = 1.4206$ *and* $s_{1,1} = 1.6149$. *Then*

$$f_{3,1} = \frac{s_{3,1}}{s_{1,1}} = 0.8797.$$

Another way is, it is the prob. that the gambler's fortune will go down 2 *before going up* 4, *which is the prob. that a gambler starting with 2 will go broke before reaching 6. Therefore,*

$$f_{3,1} = 1 - \frac{1 - (0.6/0.4)^2}{1 - (0.6/0.4)^6} = 0.8797.$$

or

$$f_{3,1} = \frac{1 - (0.4/0.6)^4}{1 - (0.4/0.6)^6} = 0.8797.$$

Example 3.6.5 ***(Gambler Ruin Problem)*** *Consider a gambler who at each play has prob.* p *of winning one unit and prob.* $q = 1 - p$ *of losing one unit. Assuming that successive plays are independent, what is the prob. that, starting with* i *units, the gambler's fortune will reach* N *before reaching 0?*
Sol. Let X_n *denote the fortune at time* n, *then* $\{X_n\}$ *is the MC process with transition prob.*

$$\begin{aligned} p_{00} &= p_{NN} = 1, \\ p_{i,i+1} &= p = 1 - p_{i,i-1}, i = 1, 2, \ldots, N-1. \end{aligned}$$

Let

$$Y = \bigcup_{n=0}^{\infty} A_n = \bigcup_{n=0}^{\infty} \{X_n = N, X_{n-1}, \ldots, X_0 \neq N\}.$$

Since $\{1, 2, \ldots, N-1\}$ *is transient and* $\{0\}, \{N\}$ *are recurrent, the gambler will either attain goal* N *or go broke. Let* P_i *denote the prob. that, starting in* i, *the gambler's fortune will eventually reach* N. *By conditioning on the outcome of the initial play of the game we obtain*

$$\begin{aligned} P(Y|X_0 = i) &= P(Y|X_0 = i, X_1 = i+1)P(X_1 = i+1|X_0 = i) \\ &+ P(Y|X_0 = i, X_1 = i-1)P(X_1 = i-1|X_0 = i). \end{aligned}$$

That is

$$\begin{aligned} P_i &= pP_{i+1} + qP_{i-1}, \quad i = 1, 2, \ldots, N-1. \\ P_{i+1} - P_i &= \frac{q}{p}(P_i - P_{i-1}), \quad i = 1, 2, \ldots, N-1. \end{aligned}$$

Hence, since $P_0 = 0$, *we obtain*

$$\begin{aligned} P_2 - P_1 &= \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1, \\ P_N - P_{N-1} &= \frac{q}{p}(P_{N-1} - P_{N-2}) = \left(\frac{q}{p}\right)^{N-1} P_1. \end{aligned}$$

Adding together yields

$$P_i - P_1 = P_1 \left[\left(\frac{q}{p}\right) + \left(\frac{q}{p}\right)^2 + \cdots + \left(\frac{q}{p}\right)^{i-1} \right].$$

or

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)} P_1, & q/p \neq 1, \\ iP_1, & q/p = 1. \end{cases}$$

Using the fact that $P_N = 1$,

$$P_1 = \begin{cases} \frac{1-(q/p)}{1-(q/p)^N}, & q/p \neq 1, \\ 1/N, & q/p = 1. \end{cases}$$

and hence

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N}, & q/p \neq 1, \\ i/N, & q/p = 1. \end{cases}$$

Note that as $N \to \infty$,

$$P_i \to \begin{cases} 1-(q/p)^i, & p > 1/2, \\ 0, & p \leq 1/2. \end{cases}$$

Example 3.6.6 *Determine the expected number of steps that an irreducible Markov chain takes to go from one state* i *to another state* j. *We change* j *to the first site*

$$\mathbf{P} = \begin{bmatrix} p(j,j) & R \\ S & Q \end{bmatrix}.$$

We then change j *to an absorbing state,*

$$\tilde{\mathbf{P}} = \begin{bmatrix} 1 & 0 \\ S & Q \end{bmatrix}.$$

Let T_j *be the number of steps needed to reach state* j. *In other words,* T_j *is the smallest time* n *such that* $X_n = j$. *For any other state* k, *let* $T_{k,j}$ *be the number of visits to* k *before reaching* j *(if we start at state* k, *we include this as one visit to* k*).*

$$E(T_j | X_0 = i) = E(\sum_{k \neq j} T_{k,j} | X_0 = i) = \sum_{k \neq j} s_{ik}.$$

$S\mathbf{1}$ *gives a vector whose* i*th component is the number of steps starting at* i *until reaching* j.

Example 3.6.7 *A random walk with reflecting boundary,* $\{0,1,2,3,4\}$.

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \left[\begin{array}{cc|ccc} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}.$$

If set finally at $j = 0$, *then*

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array},$$

$$\mathbf{S} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 2 & 4 & 6 & 4 \end{bmatrix}.$$

Then

$$S\mathbf{1} = \begin{bmatrix} 7 \\ 12 \\ 15 \\ 16 \end{bmatrix}.$$

Hence, the expected number of steps to get from 4 *to* 0 *is* 16.

3.7 Time Reversible Markov Chains

Consider a *stationary* ergodic Markov chain (that is, the chain has run for a long time) having transition prob. p_{ij} and stationary prob. π_i. Suppose that starting at some time n, we trace the states going backward in time. It turns out that the sequence of states is itself a Markov chain with transition prob. q_{ij} defined by

$$\begin{aligned} q_{ij} &= P\{X_m = j | X_{m+1} = i\} = \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \\ &= \frac{P\{X_m = j\} P\{X_{m+1} = i | X_m = j\}}{P\{X_{m+1} = i\}} = \frac{\pi_j p_{ji}}{\pi_i}. \end{aligned}$$

To see the reversed process is Markov, we must verify that

$$P\{X_m = j | X_{m+1} = i, X_{m+2}, X_{m+3}, \ldots\} = P\{X_m = j | X_{m+1} = i\}.$$

The reason is

$$P\{X_{m+2}, X_{m+3}, \ldots | X_{m+1}, X_m, \ldots\} = P\{X_{m+2}, X_{m+3}, \ldots | X_{m+1}\} \Rightarrow X_{m+k}(X_{m+1}), k \geq 2,$$

thus

$$P\{X_m = j | X_{m+1} = i, X_{m+2}(X_{m+1}), X_{m+3}(X_{m+1}), \ldots\} = P\{X_m = j | X_{m+1} = i\}.$$

Therefore, the reversed process is a Markov chain with transition prob. given by q_{ij}.

Definition 3.7.1 *If $q_{ij} = p_{ij}$ for all i, j, then the Markov chain is said to be **time reversible**. The condition for time reversibility can also be expressed as a **detailed balance condition**,*

$$\pi_i p_{ij} = \pi_j p_{ji}, \text{ for all } i, j. \tag{3.4}$$

The above condition can also be stated that, for all i, j, the rate at which the process goes from i to j (numely, $\pi_i p_{ij}$) is equal to the rate at which it goes from j to i (numely, $\pi_j p_{ji}$).

Proposition 3.7.2 *If equation (3.4) has solution, then the solution is the limiting stationary prob. π_i. However, it is possible that (3.4) has no solution. This is so since if*

$$x_i p_{ij} = x_j p_{ji}, \text{ for all } i, j, \sum_i x_i = 1,$$

then summing over i yields

$$\sum_i x_i p_{ij} = x_j \sum_i p_{ji} = x_j,$$

and, becuase the limiting stationary prob. π_i is the unique solution of the preceding, it follows that $x_i = \pi_i$ for all i.

Example 3.7.3 *Consider a random walk with states $0, 1, \ldots, M$ and transition prob.*

$$\begin{aligned} p_{i,i+1} &= \alpha_i = 1 - p_{i,i-1}, \quad i = 1, \ldots, M-1, \\ p_{0,1} &= \alpha_0 = 1 - p_{0,0}, \\ p_{M,M} &= \alpha_M = 1 - p_{M,M-1}. \end{aligned}$$

By observation, the MC is time reversible. This follows by noting that the number of transitions form i to $i+1$ must at all times be within 1 of the number from $i+1$ to i. This is so because between any two transitions from i to $i+1$ there must be one from $i+1$ to i. It follows that the rate of transitions from i to $i+1$ equals the rate from $i+1$ to i. Thus

$$\begin{aligned} \pi_0\alpha_0 &= \pi_1(1-\alpha_1), \\ \pi_1\alpha_1 &= \pi_2(1-\alpha_2), \\ &\vdots \\ \pi_i\alpha_i &= \pi_{i+1}(1-\alpha_{i+1}), \quad i = 0, 1, \ldots, M-1. \end{aligned}$$

Solve to get

$$\pi_i = \frac{\alpha_{i-1}\cdots\alpha_0}{(1-\alpha_i)\cdots(1-\alpha_1)}\pi_0, \quad i = 1, \ldots, M.$$

Since $\sum_{i=0}^{M} \pi_i = 1$,

$$\pi_0 = \left[1 + \sum_{j=1}^{M} \frac{\alpha_{j-1}\cdots\alpha_0}{(1-\alpha_j)\cdots(1-\alpha_1)}\right]^{-1}$$

Example 3.7.4 *One special case of above example is the Ehrenfest model. Suppose that M molecules are distributed among two urns; and each time point one of the molecules is chosen at random, removed from its urn, and placed in the other one. The number of molecules in urn 1 is a special case of the Markov chain having*

$$\alpha_i = \frac{M-i}{M}, i = 0, 1, \ldots, M.$$

Hence

$$\begin{aligned} \pi_0 &= \left[1 + \sum_{j=1}^{M} \frac{(M-j+1)\cdots(M-1)M}{j(j-1)\cdots 1}\right]^{-1} \\ &= \left[\sum_{j=0}^{M} C_M^j\right]^{-1} = \left(\frac{1}{2}\right)^M, \end{aligned}$$

and then

$$\pi_i = C_M^i \left(\frac{1}{2}\right)^M, \quad i = 0, 1, \ldots, M.$$

Example 3.7.5 *Consider an undirected graph having a weight w_{ij} associated with (i,j) for each arc. If at any time the particle resides at node i, then it will next move to node j with prob. where*

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}},$$

and where w_{ij} is 0 if (i,j) is not an arc. For instance, in Fig. 3.4, $p_{12} = 3/(3+1+2) = 1/2$. The time reversibility equation or the detailed balance condition reduces to

$$\pi_i \frac{w_{ij}}{\sum_j w_{ij}} = \pi_j \frac{w_{ji}}{\sum_i w_{ji}}.$$

Since $w_{ij} = w_{ji}$,

$$\frac{\pi_i}{\sum_j w_{ij}} = \frac{\pi_j}{\sum_i w_{ji}} = c.$$

Since $1 = \sum_i \pi_i$,

$$\pi_i = \frac{\sum_j w_{ij}}{\sum_i \sum_j w_{ij}}.$$

For the graph in Fig. 3.4, we have that

$$\pi_1 = \frac{6}{32}, \pi_1 = \frac{3}{32}, \pi_1 = \frac{6}{32}, \pi_1 = \frac{5}{32}, \pi_1 = \frac{12}{32}.$$

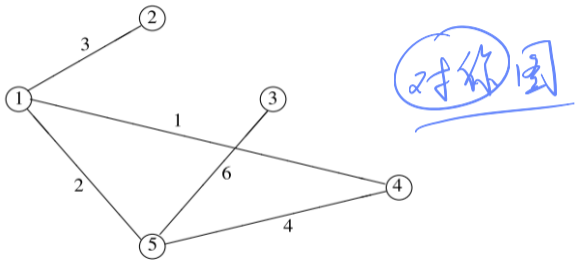


Figure 4.1 A connected graph with arc weights.

Figure 3.4: A connected graph with arc weights.

More about detailed balance condition. If we try to solve detailed balance condition for an arbitrary Markov chain with states $0, \ldots, M$, it will usually turn out that no solution exists. For example,

$$\begin{aligned} x_i p_{ij} &= x_j p_{ji}, \\ x_k p_{kj} &= x_j p_{jk}, \end{aligned}$$

implies that

$$\frac{x_i}{x_k} = \frac{p_{ji} p_{kj}}{p_{ij} p_{jk}},$$

which in general not equal to p_{ki}/p_{ik}. Thus, we see that a necessary condition for time reversibility is that

$$p_{ik} p_{kj} p_{ji} = p_{ij} p_{jk} p_{ki}, \quad \text{for all } i, j, k.$$

See counterexample in Fig. 3.5.

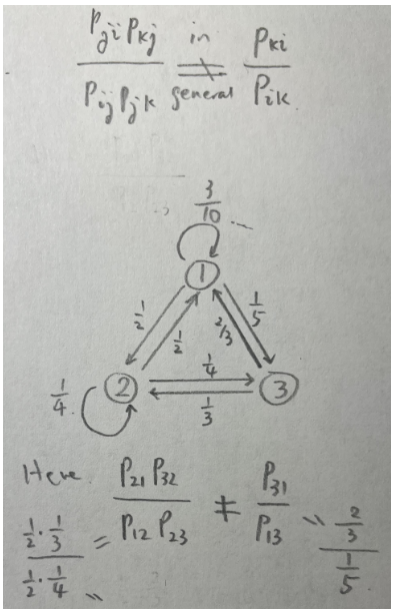


Figure 3.5: Detailed balance condition has no solution.

Theorem 3.7.6 *An ergodic Markov chain for which $p_{ij} = 0$ whenever $p_{ji} = 0$ is time reversible if and only if starting in state i, any path back to i has the same prob. as the reversed path. That is, if*

$$p_{i,i_1}p_{i_1i_2}\cdots p_{i_k,i} = p_{i,i_k}p_{i_k,i_{k-1}}\cdots p_{i_1,i},$$

for all states $i, i_1, \ldots, i_k$.

Proof. We have already proven necessity. To prove sufficiency, fix states i and j and rewrite above equation as

$$p_{i,i_1}p_{i_1i_2}\cdots p_{i_k,j}p_{ji} = p_{ij}p_{j,i_k}p_{i_k,i_{k-1}}\cdots p_{i_1,i}.$$

Summing the preceding over all states $i_1, \ldots, i_k$ yields

$$p_{ij}^{(k+1)}p_{ji} = p_{ij}p_{ji}^{(k+1)}.$$

Letting $k \to \infty$ yields

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

which proves the theorem. ■

Example 3.7.7 ***(PageRank)*** *With the development of Internet, many companies are seeking the methodology of judging the popularity of all websites. This is called PageRank, which is proposed by Page and his colleagues in 1998. They first labelled all the websites to obtain the state space $I = \{1, 2, \ldots, n\}$. When ith website has link to the other $m(i)$ number of websites, then we define the transition prob. from i to j as*

$$p_{ij} = \begin{cases} \frac{1}{m(i)}, & \text{when } i \text{ is accessible to } j, \\ 0, & \text{else.} \end{cases}$$

Obviously,

$$\sum_{j\in I} p_{ij} = 1.$$

Since MC has finite states, then all states are positive recurrent when all states communicate. Assume aperiodic. Thus ergodic finite-state Markov chain gives a unique stationary distribution.
For the stationary distribution $\pi = \{\pi_j\}$, π_j reflects the visiting prob. for the website j. When j th website becomes more popular, it will be visited more times, and thus π_j reflects the popularity of the website j. One can rank the website based on the each π_j.
Not all websites communicate or some garbage websites construct many links to themselves via virus, and thus the model needs to be improved. For example, the garbage website usually is visited for a short period, so that one can design a more reasonable prob. model. Also for example, for those websites having no links to others, one can modify the transition prob. matrix to be a weighted one, $\tilde{\mathbf{P}} = \alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{E}$.

3.8 Hidden Markov Model

3.8.1 Introduction to HMM

• Let $\{X_n, n = 1, 2, \ldots\}$ be a Markov chain with transition prob. p_{ij} and initial state prob. $\pi_i^{(0)} = P(X_0 = i)$ $i \geq 0$.

• A signal from Φ is emitted each time the MC enters a state. See Fig. 3.6.

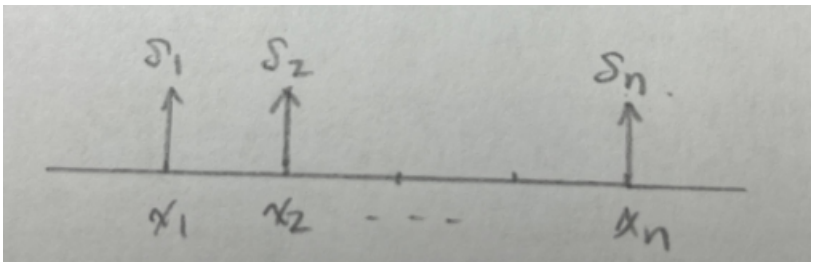


Figure 3.6: Hidden Markov Model.

• We have

$$\begin{aligned} P(S_1 &= s|X_1 = j) = q(s|j), \quad \sum_{s\in\Phi} q(s|j) = 1, \\ P(S_n &= s|X_1, S_1, \ldots X_{n-1}, S_{n-1}, X_n = j) = P(S_n = s|X_n = j) = q(s|j). \end{aligned}$$

• $S_1, S_2, \ldots$ are observed while $X_1, X_2, \ldots$ are not.
This is called a Hidden Markov Model.

Example 3.8.1 *A machine has 1 good state and 2 poor state. The transition prob. is*

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0 & 1 \end{bmatrix} \end{array}.$$

Each item produced is of acceptable quality with prob. 0.99 when the process is in state 1.
Each item produced is of acceptable quality with prob. 0.96 when the process is in state 2.

If item accepted or not can be observed, while the state of the machine cannot be observed (HMM).

$$\begin{aligned} q(a|1) &= 0.99, \quad q(u|1) = 0.01, \\ q(a|2) &= 0.96, \quad q(u|2) = 0.04. \end{aligned}$$

- $\{S_n, n \geq 1\}$ is not MC.
- $\vec{S}^n = (S_1, \ldots, S_n)$ is a random variable.

  $\vec{s}_k = (s_1, \ldots, s_k)$ is one observation.

  Define $F_n(j) = P(\vec{S}^n = \vec{s}_n, X_n = j)$, then

$$P(X_n = j|\vec{S}^n = \vec{s}_n) = \frac{P(X_n = j, \vec{S}^n = \vec{s}_n)}{P(\vec{S}^n = \vec{s}_n)} = \frac{F_n(j)}{\sum_i F_n(i)}.$$

Now,

$$\begin{aligned} F_n(j) &= P(\vec{S}^{n-1} = \vec{s}_{n-1}, S_n = s_n, X_n = j) \\ &= \sum_i P(\vec{S}^{n-1} = \vec{s}_{n-1}, X_{n-1} = i, S_n = s_n, X_n = j) \\ &= \sum_i P(\vec{S}^{n-1} = \vec{s}_{n-1}, X_{n-1} = i) P(S_n = s_n, X_n = j|\vec{S}^{n-1} = \vec{s}_{n-1}, X_{n-1} = i) \\ &= \sum_i F_{n-1}(i) P(S_n = s_n, X_n = j|X_{n-1} = i). \end{aligned}$$

Notice that

$$\begin{aligned} P(S_n &= s_n, X_n = j|X_{n-1} = i) = P(X_n = j|X_{n-1} = i) P(S_n = s_n|X_n = j, X_{n-1} = i) \\ &= p_{ij} q(s_n|j). \end{aligned}$$

Then

$$F_n(j) = \sum_i F_{n-1}(i) p_{ij} q(s_n|j) = q(s_n|j) \sum_i F_{n-1}(i) p_{ij}. \tag{3.5}$$

Starting with

$$\begin{aligned} F_1(i) &= P(S_1 = s_1, X_1 = i) = P(S^1 = s_1|X_1 = i) P(X_1 = i) \\ &= \pi_i^{(0)} q(s_1|i). \end{aligned} \tag{3.6}$$

We can use Eq. (3.5) to recursively determine $F_2(i), F_3(i), \ldots$ up to $F_n(i)$.

Example 3.8.2 *(Cont.)* $P(X_1) = 0.8$. *The first 3 items are a,u,a.*

(1) What is the prob. that the process was in good state when the 3rd item was produced?

(2) What is the prob. that X_4 *is 1?*

(3) What is the prob. that the next item produced is accepted?

Sol. $\vec{s}_3 = \{a, u, a\}$. *We have*

$$\begin{aligned}
F_1(i &= 1) = \pi^{(0)}_{i=1} q(a|i=1) = (0.8)(0.99) = 0.792. \\
F_1(i &= 2) = \pi^{(0)}_{i=2} q(a|i=2) = (0.2)(0.96) = 0.192. \\
F_2(i &= 1) = q(u|1)[F_1(1)p_{11} + F_1(2)p_{21}] = (0.01)[(0.792)(0.9) + (0.192)(0)] \\
&= 0.007128. \\
F_2(i &= 2) = q(u|2)[F_1(1)p_{12} + F_1(2)p_{22}] = (0.04)[(0.792)(0.1) + (0.192)(1)] \\
&= 0.010848. \\
F_3(i &= 1) = q(a|1)[F_2(1)p_{11} + F_2(2)p_{21}] = (0.99)[(0.007128)(0.9) + (0.010848)(0)] \\
&= 0.006351. \\
F_3(i &= 2) = q(a|2)[F_2(1)p_{12} + F_2(2)p_{22}] = (0.96)[(0.007128)(0.1) + (0.010848)(0)] \\
&= 0.011098.
\end{aligned}$$

(a) We have

$$\begin{aligned}
P(X_3 &= 1|\vec{s}_3) = \frac{P(X_3 = 1, \vec{s}_3)}{\sum_i P(X_3 = i, |\vec{s}_3)} = \frac{F_3(1)}{F_3(1) + F_3(2)} = \frac{0.006351}{0.006351 + 0.011098} \\
&= 0.364.
\end{aligned}$$

(b) Conditioning on X_3, *we have*

$$\begin{aligned}
P(X_4 &= 1|\vec{s}_3) = P(X_4 = 1|X_3 = 1, \vec{s}_3)P(X_3 = 1|\vec{s}_3) + P(X_4 = 1|X_3 = 2, \vec{s}_3)P(X_3 = 2|\vec{s}_3) \\
&= (p_{11})(0.364) + (p_{21})(1 - 0.364) = 0.3276.
\end{aligned}$$

(c) Conditioning on X_4,

$$\begin{aligned}
P(S_4 &= a|\vec{s}_3) = P(S_4 = a|X_4 = 1, \vec{s}_3)P(X_4 = 1|\vec{s}_3) + P(S_4 = a|X_4 = 2, \vec{s}_3)P(X_4 = 2|\vec{s}_3) \\
&= P(S_4 = a|X_4 = 1)(0.3276) + P(S_4 = a|X_4 = 2)(1 - 0.3276) \\
&= (0.99)(0.3276) + (0.96)(1 - 0.3276) = 0.9698.
\end{aligned}$$

See reference in [9].

3.8.2 Key ingredients to HMM

(1) transition prob. matrix $p_{ij} = P(X_{n+1} = j|X_n = i)$.
(2) observation prob. distribution. $q(s|j) = P(S_n = s|X_n = j)$.
(3) initial state distribution. $\pi^{(0)}_i = P(X_0 = i)$.
Compact notation: $\lambda = (p, q, \pi)$.

3.8.3 Three Basic Problems for HMMs

Compared to the traditional Hidden Markov Model (HMM), the RNN, LSTM become more popular nowadays and receive more attentions. In order for HMM model useful in real-world application [9], we present 3 basis problems:

(1) (probability computation problem) Given observation $\vec{s}_n = (s_1, \ldots, s_n)$ and a model $\lambda = (p, q, \pi)$, how to efficiently compute $P(\vec{s}_n|\lambda)$, the prob. of the observation sequence given the model?

(2) (decoding problem) Given observation $\vec{s}_n = (s_1, \ldots, s_n)$ and a model $\lambda = (p, q, \pi)$, how to choose a state sequence $X = (X_1, \ldots, X_n)$ which is optimal in some meaningful sense?

(3) (learning problem or inverse problem) Given observation $\vec{s}_n = (s_1, \ldots, s_n)$, how do we adjust or estimate the model parameter $\lambda = (p, q, \pi)$ to maximize the likelihood $P(\vec{s}_n|\lambda)$?

3.8.4 Forward and Backward Approaches for Problem 1

Forward approach

Compute $P(\vec{S}^n = \vec{s}_n)$ by $= \sum_i F_n(i)$ for $i = 1, \ldots, N$ and recursive formula,

$$F_n(j) = q(s_n|j) \sum_i F_{n-1}(i) p_{ij}.$$

Notice that $F_1(i = 1), \ldots, F_1(i = N)$ is $O(N)$, $F_2(i = 1), \ldots, F_2(i = N)$ is $O(N)$, gives total complexity is $O(nN^2)$.

Direct computation

by conditioning on the first n states of the Markov chain.

$$\begin{aligned} P(\vec{S}^n = \vec{s}_n) &= \sum_{i_1 \cdots i_n} P(\vec{S}^n = \vec{s}_n | X_1 = i_1, \ldots, X_n = i_n) P(X_1 = i_1, \ldots, X_n = i_n) \\ &= \sum_{i_1 \cdots i_n} q(s_1|i_1) \cdots q(s_n|i_n) \pi^{(0)}_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \end{aligned}$$

where made use of

$$\begin{aligned} P(X_1, \ldots, X_n) &= P(X_n|X_{n-1}) P(X_{n-1}|X_{n-2}) \cdots P(X_2|X_1) P(X_1), \\ P(S_1 \ldots S_n | \underbrace{X_1 \ldots X_n}_{A}) &= P(S_n|S_1 \ldots S_{n-1} A) P(S_{n-1}|S_1 \ldots S_{n-2} A) \cdots P(S_1|A). \end{aligned}$$

The complexity is $O(N^n)$.

Backward approach

Define the quantity

$$B_k(i) = P(S_{k+1} = s_{k+1}, \cdots, S_n = s_n | X_k = i).$$

A recursive formula for $B_k(i)$ can be obtained by conditioning on X_{k+1},

$$\begin{aligned} B_k(i) &= \sum_j P(S_{k+1} = s_{k+1}, \cdots, S_n = s_n | X_k = i, X_{k+1} = j) P(X_{k+1} = j | X_k = i) \\ &= \sum_j P(S_{k+1} = s_{k+1}, \cdots, S_n = s_n | X_{k+1} = j) p_{ij} \\ &= \sum_j P(S_{k+1} = s_{k+1} | X_{k+1} = j) P(S_{k+2} = s_{k+2}, \cdots, S_n = s_n | S_{k+1} = s_{k+1}, X_{k+1} = j) p_{ij} \\ &= \sum_j q(s_{k+1}|j) P(S_{k+2} = s_{k+2}, \cdots, S_n = s_n | X_{k+1} = j) p_{ij} \\ &= \sum_j q(s_{k+1}|j) B_{k+1}(j) p_{ij}. \end{aligned}$$

The initial condition is

$$\begin{aligned} B_{n-1}(i) &= P(S_n = s_n | X_{n-1} = i) = \sum_j P(S_n = s_n | X_{n-1} = i, X_n = j) P(X_n = j | X_{n-1} = i) \\ &= \sum_j p_{ij} q(s_n | j). \end{aligned}$$

Determine the funtions $B_{n-2}(i), B_{n-3}(i), \ldots$ subsequently. Finally,

$$\begin{aligned} P(\vec{S}^n &= \vec{s}_n) = \sum_i P(\vec{S}^n = \vec{s}_n | X_1 = i) \pi_i^{(0)} \\ &= \sum_i P(S_1 = s_1 | X_1 = i) P(S_2 = s_2, \ldots, S_n = s_n | S_1 = s_1, X_1 = i) \pi_i^{(0)} \\ &= \sum_i q(s_1 | i) P(S_2 = s_2, \ldots, S_n = s_n | X_1 = i) \pi_i^{(0)} \\ &= \sum_i q(s_1 | i) B_1(i) \pi_i^{(0)}. \end{aligned}$$

Both Forward and Backward approaches

Suppose that we have computed $F_k(j)$ and $B_k(j)$ for some k,

$$\begin{aligned} P(\vec{S}^n &= \vec{s}_n, X_k = j) = P(\vec{S}^k = \vec{s}_k, X_k = j) P(S_{k+1}, \ldots, S_n | \vec{S}^k = \vec{s}_k, X_k = j) \\ &= P(\vec{S}^k = \vec{s}_k, X_k = j) P(S_{k+1}, \ldots, S_n | X_k = j) \\ &= F_k(j) B_k(j) \end{aligned}$$

Thus,

$$P(\vec{S}^n = \vec{s}_n) = \sum_j P(\vec{S}^n = \vec{s}_n, X_k = j) = \sum_j F_k(j) B_k(j), \text{ for any } k.$$

One can first parallel compute $F_k(j)$ and $B_k(j)$, recursively.

Chapter 4

Markov Chain Monte Carlo (MCMC)

Key concepts include:

• important sampling,

• Metropolis-Hasting (MH) sampling,

• Hamilton MH sampling

• 1D data sampling

• Monte Carlo sum, see [13] 数值分析书

• The algorithm and realization of MH sampling (discrete and continuous, Metroplis algorithm, q is symmetric)

• Gibbs sampling, see chapter 6 of [5],

• MCMC, advantage and disadvantage, see chapter 6 of [5],

• other notification, including burn-in. see Handbook [3],

• Hamiltonian Monte Carlo (HMC), see chapter 5 of Handbook [3],

• Vanilla MCMC, see zhihu PR II

• Choice of transition q, paper in John Harlim's book see [4].

4.1 Monte Carlo Methods

4.1.1 deterministic vs. stochastic

(1) deterministic is preferred for low dimension problems, numerical differentiation or integral for ODE and PDE problems.

(2) stochastic is preferred for high dimensional problems.

4.1.2 background

Monte Carlo方法是对任何利用随机数序列来做随机模拟的这类数值方法的统称，该类方法的使用已达数世纪之久， 但是只在半个多世纪前才开始成长为应用于复杂性高的问题的数值方法之一。

Monte Carlo本是摩纳哥（Monaco）的著名赌城，第二次世界大战时期N. Metropolis在曼哈顿计划中取其博彩游戏和随机模拟算法二者的相似之处，首次借用其名来作为随机模拟算法的名称。

4.1.3 difference bw some terminology

Markov chain is a concept for the property of a time series or a continuous-time stochastic process.

Monte Carlo is terminology for all stochastic simulation, in particular, referred to the summation of an integral using stochastic approaches.

所有随机数模拟的统称

MCMC is a sampling approach.

采样方法

4.1.4 statistical mechanics

Stochastic Monte Carlo approach is different from the deterministic numerical discretizaiton approach since the latter one is often used for ODEs and PDEs to describe the physics or mathematical systems. Monte Carlo studies the system from microscopic point of view based on the concept of pdf for many particles. In around 1950, people use pdf of many particles to calculate various macroscopic quantities, including temperature and heat capacity, which are usually high dimensional problems. Since too many particles or say the dimension of phase space is very high, the computation of expectations or numerical integrals for high dimensional space becomes extremely important.

4.1.5 applications

statistical mechanics, then quantum chemistry, material science, biological mathematics, financial mathematics, deep learning.

Example 4.1.1 *(Deterministic) Consider the integral,*

$$I(f) = \int_0^1 f(x)dx.$$

For equi-distance partition (see Fig. 4.1), that is, $0, 1/N, \ldots, (N-1)/N, 1$*, one can apply the Simpson method*

$$I(f) \approx (\frac{1}{2}f(x_0) + \sum_{i=1}^{N-1} f(x_i) + \frac{1}{2}f(x_N))h,$$

which is of order $h^2 = O(N^{-2})$*. One can also apply the Darboux sum*

$$I(f) \approx \sum_{i=0}^{N-1} f(x_i)h, \quad \text{or } I(f) \approx \sum_{i=1}^{N} f(x_i)h,$$

which is of order $h = O(N^{-1})$*.*

（相合估计）

Example 4.1.2 *(Stochastic)* $I(f) = Ef(x)$ *when* $X \sim U(0,1)$*. Law of large number. We have the Monte Carlo method for numerical integration:*

$$I(f) \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) := I_N(f),$$

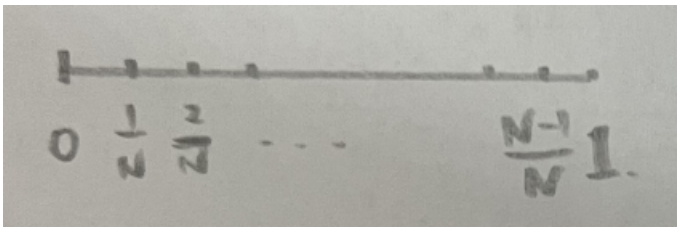


Figure 4.1: Equi-distance partition of the interval $[0, 1]$.

where X_i $(i = 1, 2, \ldots, N)$ are i.i.d. uniformly distributed $\sim U(0, 1)$. Obviously, $I_N(f)$ is unbiased and consistent (converges in prob.)

Pf. First show that $I_N(f)$ is unbiased,

$$EI_N(f) = E\left[\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 f(x)dx = I(f).$$

Denote

$$e_N = I_N(f) - I(f)$$

as a random variable with $Ee_N = 0$. Then

$$\begin{aligned} E|e_N|^2 &= E(I_N(f) - I(f))^2 = E[\frac{1}{N}\sum_{i=1}^{N}\underbrace{(f(x_i) - I(f))}_{:=a_i}]^2 \\ &: = E[\frac{1}{N}\sum_{i=1}^{N} a_i]^2 = \frac{1}{N^2}E\left(\sum_{i=1}^{N} a_i\right)^2 \\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N} Ea_i^2 + \sum_{i<j} 2\underbrace{Ea_ia_j}_{=0}\right) = \frac{1}{N}Ea_i^2 \\ &= \frac{1}{N}E(f(x_i) - I(f))^2 = \frac{1}{N}Var(f) \to 0. \end{aligned}$$

Therefore, by Chebyshev inequality, one can obtain the consistency:

$$P(|I_N f - If| > \varepsilon) \le \frac{Var(e_N)}{\varepsilon^2} = \frac{E|e_N|^2}{\varepsilon^2} \to 0.$$

By Schwartz inequality,

$$E|e_N| \le \sqrt{E|e_N|^2} = \sqrt{\frac{Var(f)}{N}} \sim N^{-1/2}.$$

If $Var(f) < \infty$, then the convergence rate for Monte Carlo is $1/2$.

Example 4.1.3 ***(Project 1)*** *Consider the integral*

$$\int_0^{\pi/2} \sin(x)dx = 1.$$

The Monte Carlo approximation is

$$\int_0^{\pi/2} \sin(x)dx \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\pi}{2}\sin(\frac{\pi}{2}x_i),$$

where x_i is i.i.d. $\sim U(0,1)$. Then fix $m = 100$ independent trials. Take many different $N = 10, 20, 40, 80, \ldots, 640, 1280, 2560, 5120, 10240$. Let e_N^j be the error for the jth trial under a given N. Define

$$e_N = \frac{1}{m} \sum_{j=1}^{m} e_N^j.$$

Then plot $\log e_N$ vs. $\log N$ (similar to the follows in Fig. 4.2).

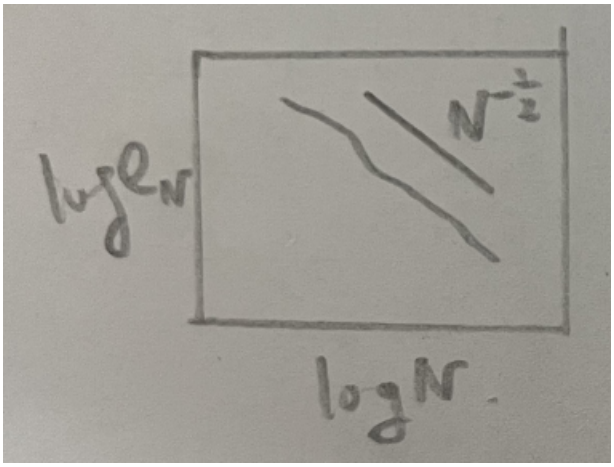


Figure 4.2: Error e_N as a function of the number of particles N.

Example 4.1.4 *For general integral,*

$$I(f) = \int_0^1 f(x)p(x)dx,$$

where $p(x)$ is a pdf s.t. $\int_0^1 p(x)dx = 1, p(x) \geq 0$. Then

$$I_N(f) = \frac{1}{N} \sum_{i=1}^{N} f(x_i), \text{ with } x_i \text{ i.i.d. with pdf } p(x).$$

Example 4.1.5 *For $I(f) = \int_0^1 f(x)dx$, then*

$$I_N(f) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{p(x_i)}, \text{ where } x_i \text{ i.i.d. with pdf } p(x).$$

Usually, $p(x)$ can be approximated by kernel density estimation or histogram.

Example 4.1.6 *Multivariate case. Consider the hyper-cubic $\Omega = [0,1]^d$ in $\mathbb{R}^d$ and the integral*

$$I(f) = \int \cdots \int_{\Omega} f(\vec{x})p(\vec{x})d\vec{x}, \quad \vec{x} = (x_1, \ldots, x_d),$$

where $p(\vec{x})$ is a pdf s.t. $\int_0^1 p(\vec{x})d\vec{x} = 1, p(\vec{x}) \geq 0$. For determinisitc approach, assume that $[0,1]$ is equi-distance partitioned into n intervals for each dimension (see Fig. 4.3). Then the accuracy is still $O(n^{-2}) = O(N^{-\frac{2}{d}})$. However, the complexity requires $O(N = n^d)$ amounts of computation.
On the other hand, for Monte Carlo, $\{x_i\}_{i=1}^M$ i.i.d. pdf is $p(x)$. Let

$$I_M(f) := \frac{1}{M} \sum_{i=1}^{M} f(x_i) \to I(f),$$

with the convergence rate of $M^{-1/2}$ and the complexity of $O(M)$. If we let the same accuracy $M^{-1/2} = O(n^{-2}) = O(N^{-2/d})$, then we obtain that $M = O(N^{4/d})$. Thus when $d > 4$, we have $M < N$, which means that the complexity of Monte Carlo is smaller than the Simposon method.

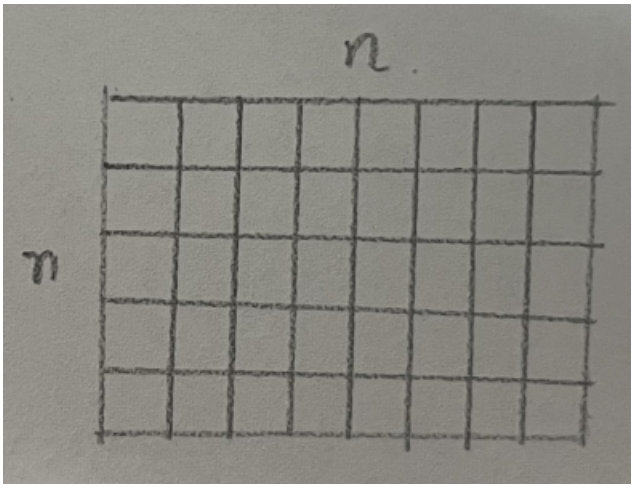


Figure 4.3: Equi-distance partition of a rectangular domain.

Example 4.1.7 *In Statistical Mechanics,*

$$\langle A \rangle = \frac{1}{Z} \int_{\mathbb{R}^{6N}} A(\vec{c}) e^{-\beta H(\vec{c})} d\vec{c},$$

where $\beta = (k_B T)^{-1}$ is the inverse temperature, k_B is the Boltzman constant, T temperature, $d\vec{c} = d\vec{x}_1 \cdots d\vec{x}_N$ $d\vec{p}_1 \cdots d\vec{p}_N$, N is number of particles, and

$$Z = \int_{\mathbb{R}^{6N}} e^{-\beta H(\vec{c})} d\vec{c},$$

is the partition function. 如果模拟中取值*1000*个粒子，则*Monte Carlo*方法做*100*个运算量*(*暂且忽略产生随机数所需要的运算量*)*达到的数值精度基本上需要梯形公式做10^3000次运算量才可匹敌，后者的运算量对目前威力最大的计算机而言也是鞭长莫及的，可见，虽然*Monte Carlo*方法的半阶收敛性是非常糟糕的结果，但是在处理维数很高的问题的时候几乎是唯一的选择，因为这种方法受空间维数的制约远没有其他经典的数值离散方法大。

4.2 Generation of Random Numbers

random numbers usually are not i.i.d. in computer, but called "**pseudo random**".

4.2.1 generation of random numbers from $U(0,1)$

1. Midsquare method

（平方取中法）

In the early development of computers, Von Neumann and others generated pseudo random numbers using midsquare method. For example, first take a 4-digit number 3333 and square it to obtain 11108889.

Take the mid 4 digits 1088 and square it to obtain 1183744, and so forth. Every number 3333, 1088, 1837, divided by 10^4 to obtain a pseudo random number between $[0,1]$ with uniform distribution. However, the maximum cycle length of the approach is less than 10^4, and its statistics is not very good while this algorithm was early used in the computation of nuclear reaction.

核反应计算 循环长度

2. Linear congruential algorithm

（线性同余法）

In the random number generator of $U(0,1)$, one early popular method is called linear congruential algorithm. The algorithm is as follows:

$$X_{n+1} = aX_n + b(\text{mod m}),$$

where a, b, m are given integers. One important criterion for judging a random number generator is the called maximum cycle length. In the same period, the longer the maximum cycle length, the better the performance of the generator. For example,

$$m = 2^k, \quad a = 4c + 1, \quad b \text{ is odd.}$$

3. Magic "16807"

（神奇的"16807"）

1969年，Lewis, Goodman和Miller提出了如下发生器：

$$X_{n+1} = aX_n \ (\text{mod } m),$$

并且取$a = 7^5 = 16807$, $m = 2^{31} - 1 = 2147483647$. Shrage 给出了一个在计算机上高效实现上述乘法同余的算法，这样得到的伪随机数发生器最大循环长度可达到2.1×10^9. 这个发生器通过了当时的所有理论测试，被称为最小标准发生器(Minimal standard generator)(意指其他的发生器如果要被接受，至少要能达到这一发生器的质量).

后来，基于这一方法，L'Ecuyer 采用所谓Bays-Durham 洗牌算法(见文献[6]) 给出了一个更为强大的随机数发生器，其最大循环长度达到约2.3×10^{18}. 在数值计算的著名书籍[8]中，给出了这一算法的具体实现程序ran2().该书作者声称，如果有人能给出使用上述算法而导致系统性失败的案例，将付款1000美元!

随机数发生器是进行随机模拟的基石，如果没有一个可靠的随机数发生器，一切的计算结果都不再可信，笔者强烈建议读者使用经过大量测试的、成熟的随机数发生器程序包，而不提倡自行编写此类程序.另外，对程序的选取也要慎重，我们推荐 文献[8]中的程序以及www.netlib.org上的随机数发生器.

4.2.2 generation of random numbers with general distributions

1. Transformation method

（变换法）

Proposition 4.2.1 *Let a random variable Y with the distribution function $F(y)$, that is*

$$P(Y \le y) = F(y).$$

If another random variable $X \sim U(0,1)$, then $Y = F^{-1}(X)$ satisfies the desired distribution.

Proof. Since $X \sim U(0,1)$ and $Y = F^{-1}(X)$, so that

$$\begin{aligned} P\{Y &\leq y\} = P\{F^{-1}(X) \leq y\} \\ &= P\{X \leq F(y)\} = F(y). \end{aligned}$$

■

Based on above prop., if we already have random variables $X_i (i = 1, 2, \cdots)$ uniformly distributed on $(0,1)$, then $Y_i = F^{-1}(X_i)$ is randomly distributed with distribution function $F(y)$. The larger the pdf is, the steeper the cdf is. The smaller the pdf is, the flatter the cdf is. See Fig. 4.4.

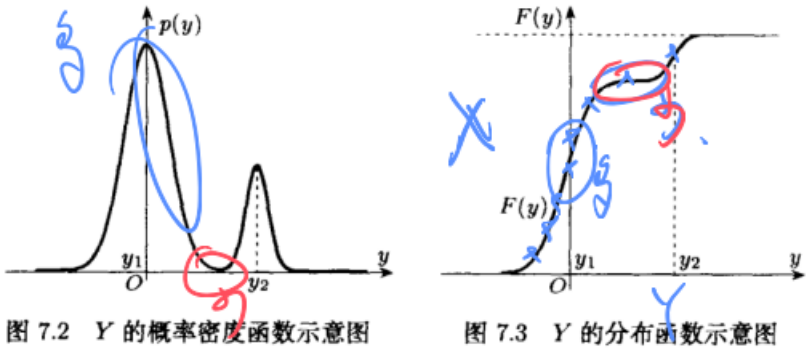


图 7.2 Y 的概率密度函数示意图　　图 7.3 Y 的分布函数示意图

Figure 4.4: The probability density function (pdf) and cumulative distribution function(cdf) of a random variable Y.

Example 4.2.2 ***(Project 2)*** *(Exponential Distribution) Let the exponential distribution with pdf*

$$p(y) = \begin{cases} 0, & y \leq 0, \\ \lambda e^{-\lambda y}, & y > 0, \end{cases}$$

Its cdf is $F(y) = 1 - e^{-\lambda y}$. *Then*

$$F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x), \quad x \in (0,1).$$

Based on the transformation method, the exponentially distributed random variable can be obtained by

$$Y_i = -\frac{1}{\lambda}\ln(1-X_i), \quad i = 1, 2, \cdots$$

where $X_i \sim U(0,1)$. *Check the pdf usign histogram or Kernel Density Estimation (KDE). Or check the statistics based on p-value using* χ^2*-test.*

Example 4.2.3 *(Normal Distribution) The normally distributed variable has the pdf*

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and its cdf is

$$F(x) = \int_{-\infty}^{x} p(y)dy = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{x}{\sqrt{2}}\right),$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ *is called the error function. Thus,* $F^{-1}(x) = \sqrt{2}\text{erf}^{-1}(2x-1)$. *However, it is directly realized during numerical implementation since* erf^{-1} *is difficult to compute. This means that the transformation method has limitation.*

Box-Muller method

To generate normally distributed random variables, we use the following well-known Box-Muller method. Notice that

$$\left(\int_{-\infty}^{+\infty} e^{-x^2}dx\right)^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}dxdy = \int_0^{+\infty}\int_0^{2\pi} e^{-r^2}rdrd\theta = \pi,$$

on which the Box-Muller method is based. Let $(x_1, x_2) = (r\cos\theta, r\sin\theta)$, then

$$\frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}dx_1dx_2 = \frac{1}{2\pi}e^{-\frac{r^2}{2}}rdrd\theta = \left(\frac{1}{2\pi}d\theta\right)\left(e^{-\frac{r^2}{2}}rdr\right).$$

In θ direction, $\frac{1}{2\pi}$ is the density for the uniform distribution $U(0, 2\pi)$. In r direction, $e^{-\frac{r^2}{2}}r$ is the density corresponding to the distribution function $F(r) = \int_0^r e^{-\frac{s^2}{2}}sds = 1 - e^{-\frac{r^2}{2}}$. Thus, random variables (Y_1, Y_2) for the two-dimensional normal distribution can be generated through

$$\begin{cases} Y_1 = \sqrt{-2\ln X_1}\cos(2\pi X_2), \\ Y_2 = \sqrt{-2\ln X_1}\sin(2\pi X_2), \end{cases}$$

where X_1 and X_2 are independent random variables satisfying $U(0, 1)$. Notice that

$$F^{-1}(r) = \sqrt{-2\ln(1-r)},$$

where $1 - r \sim U(0, 1)$ and $r \sim U(0, 1)$. Then the above expression is written in this way since $Y_1 = r\cos\theta$ and $Y_2 = r\sin\theta$, where $\theta = 2\pi X_2$ and $r = \sqrt{-2\ln X_1}$.

Acceptance-rejection method

Not introduced at this moment. The idea is kind of similar to MCMC.

4.2.3 technique for reducing the variance

The error of Monte Carlo is $\sigma/\sqrt{N}$, where $\sigma = (Var(f))^{1/2}$.

The rate $1/\sqrt{N}$ usually cannot be improved!

But the constant σ can!

Importance sampling (Project 2)

If i.i.d. $x_i \sim U(0, 1)$, then

$$If \approx I_X(f) := \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

On the other hand,

$$If = \int_0^1 f(y)dy = \int_0^1 \frac{f(y)}{p(y)}p(y)dy,$$

where $p(y)$ is a pdf. Then

$$I(f) \approx I_Y(f) := \frac{1}{N}\sum_{i=1}^{N}\frac{f(y_i)}{p(y_i)},$$

where i.i.d. $y_i \sim p(y)$.

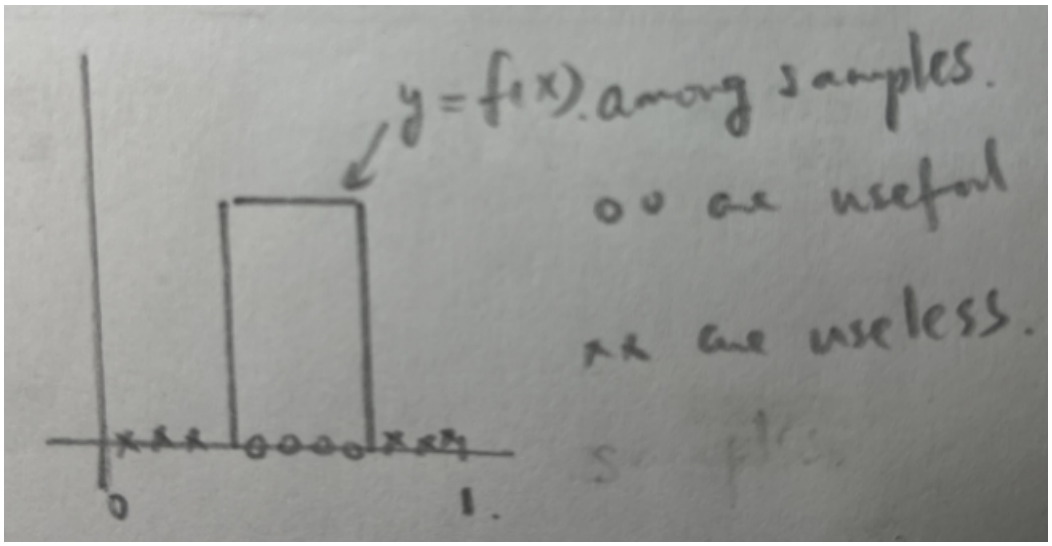


Figure 4.5: Sketch for importance sampling.

Let's see Fig. 4.5 for the sketch for importance sampling. The nonzero values of the integrand function $f(x)$ are focused in the interval (a, b). If using $x_i \sim U(0, 1)$, then many samples of x_i lie out of the interval (a, b) and these samples are used for computing the integral $\int_0^a f(x)dx + \int_b^1 f(x)dx = 0$. This means that we spend amounts of computation to calculate the integral part that does not contribute to the final result. Thus, the efficiency is low and also the accuracy is low. On the other hand, if we take $p(y)$ proportional to $f(y)$, most samples are used to compute the dominant part of the integral. This is called the importance sampling and $p(y)$ reflects the important part of $f(x)$.

Theoretically, we can analyze the error from variance point of view:

$$\begin{aligned} Var_X(f) &= \int_0^1 [f - I(f)]^2 dx = \int_0^1 f^2 dx - I^2(f), \\ Var_Y(\frac{f}{p}) &= \int_0^1 \left(\frac{f}{p}\right)^2 p dy - I^2(f) = \int_0^1 \frac{f^2}{p} dy - I^2(f). \end{aligned} \tag{4.1}$$

Here, we notice that since

$$e_X = I_X(f) - I(f), \quad e_Y = I_Y(f) - I(f),$$

their means are

$$Ee_X = 0, \quad Ee_Y = 0,$$

and variances are

$$\begin{aligned} E|e_X|^2 &= \frac{Var_X(f)}{N}, \\ E|e_Y|^2 &= E(I_Y(f) - I(f))^2 = E[\frac{1}{N}\sum_{i=1}^N \underbrace{(\frac{f(y_i)}{p(y_i)} - I(f))}_{:=b_i}]^2 \\ &= \frac{1}{N^2} E\left(\sum_{i=1}^N b_i\right)^2 = \frac{1}{N^2}\left(\sum_{i=1}^N Eb_i^2 + \sum_{i<j} 2\underbrace{Eb_i b_j}_{=0}\right) = \frac{1}{N} Eb_i^2 \\ &= \frac{1}{N} E\left(\frac{f(y_i)}{p(y_i)} - I(f)\right)^2 = \frac{1}{N} Var_Y(\frac{f}{p}). \end{aligned}$$

In equation (4.1), if we take an appropriate $p(y)$, s.t. $\int_0^1 \frac{f^2}{p}dy < \int_0^1 f^2 dx$, then the variance reduces $Var_Y(\frac{f}{p}) < Var_X(f)$. In particular, if $p(y) = \frac{f(y)}{I(f)} \propto f(y)$, then

$$Var_Y(\frac{f}{p}) = \int_0^1 \frac{f^2}{p}dy - I^2(f) = I(f)\int_0^1 f dy - I^2(f) = I^2(f) - I^2(f) = 0.$$

This means that when the importance of $p(x)$ is the same with that of $f(x)$, the variance becomes 0 and we obtain the exact integral value. However, in practice this is difficult to realize due to two reasons. First, for $p(y) = \frac{f(y)}{I(f)}$, it is impossible for us to prior get the value of $I(f)$ in the computation of $I_Y(f) := \frac{1}{N}\sum_{i=1}^N \frac{f(y_i)}{p(y_i)}$. Second, it is not trivial to get samples with density $p(y) \propto f(y)$ (while can be done using MCMC).

重要性抽样对实际计算有根本思想性的指导作用，它依赖于我们对一个问题先验的了解程度并以此构造适当的抽样所依据的概率密度。

Advantages of importance sampling over simple Monte Carlo (uniform sampling)

- Can significantly improve performance, by reducing the variance
- Can use samples from a different distribution, say q, to compute expectations with respect to p
- Can compute the normalization constant of p, as well as Bayes factors

Disadvantages

- Need to be able to evaluate the pdf/pmf $p(x)$ and $q(x)$, at least up to proportionality constants
- It might not be obvious how to choose a good q

Control Variate method, modification of importance sampling

控制变量法

The basic idea here is to use a random variable with given statistics (such as the mean) to control another random variable with unknown statistics. For example,

$$\int_0^1 f(x)dx = \int_0^1 [f(x) - g(x)]dx + \int_0^1 g(x)dx,$$

with $\int_0^1 g(x)dx$ is already known. Then one can use Monte Carlo to obtain

$$I_N(f) := \frac{1}{N}\sum_{i=1}^N [f(x_i) - g(x_i)] + I(g), \text{ with } x_i \sim \text{i.i.d. } U(0,1).$$

If $Var[f-g] \le Var[f]$, then the constant coefficient is reduced while the convergence rate is kept the same. In extreme case, when $f = g$, then $Var[f-g] = 0$.

Example 4.2.4 *Consider the integral,*

$$I(f) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(1+r)^{-1}e^{-\frac{x^2}{2}}dx,$$

where $r = e^{\sigma x}$ $(\sigma > 0)$. *Notice that*

$$(1+r)^{-1} \approx \begin{cases} 1, & x \le 0, \\ 0, & x > 0, \end{cases} := h(x).$$

Then $I(f)$ *can be approximated by*

$$I(f) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} [(1+r)^{-1} - h(x)]e^{-\frac{x^2}{2}}dx + \frac{1}{2}.$$

4.3 Introduction to Markov Chain Monte Carlo (MCMC)

4.3.1 Application of MCMC

(1) MCMC can be used to generate samples from a given pdf, and then possibly be used for **Monte Carlo sum**.

(2) Statistics have two popular methods, **frequentist statistics** and **Bayesian statistics**. The main differences are listed in Table 4.1. Once you understand the Bayesian approach, it seems so natural that it is hard to imagine any alternative. In fact, there was no satisfying alternative until the early 1900's, when Karl Pearson, Jerzy Neyman, Egon Pearson, and Ronald Fisher initiated what is now called frequentist statistics.

Both can be used for parameter estimation. For **Bayesian inference**,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta},$$

where y is data and θ is model (or paramters of model), MCMC is one popular approach, which can be used for uncertainty quantification. However, MCMC is not appropriate for too large amount of data or too complicated model. Another Bayesian approach is variational inference which involves Kullback-Leibler (KL) divergence.

(3) **The development of MCMC and its connection to statistical mechanics.**

The popular application of MCMC starts from the development of statistical mechanics. In statistical mechanics, an ensemble average is thought of as a concept that the macro state we see is the prob. average of many micro states in equilibrium (see Fig. 4.6). The prob. here usually corresponds to Gibbs measure (a more general used terminology for both finite and infinite systems) or Boltzmann distribution (a term for finite systems). For examle, here is a box containing many gas particles in equilibrium and we can measure their temperature, pressure, etc. In SM, this macro system, in fact, corresponds to many micro systems, and its macro quantity is the statistical average of the quantities of these many micro systems. In mathematics, the discrete and continuous cases are

$$\begin{aligned}\langle A(\sigma)\rangle &= \sum_{\sigma} \frac{\exp\{-\beta H(\sigma)\}}{Z} A(\sigma),\\ \langle A(\sigma)\rangle &= \int_{\sigma} \frac{\exp\{-\beta H(\sigma)\}}{Z} A(\sigma)d\sigma,\end{aligned}$$

where $Z = \sum_{\sigma} \exp\{-\beta H(\sigma)\}$ and $Z = \int_{\sigma} \exp\{-\beta H(\sigma)\}d\sigma$ are partition functions.

The following is the statement of equilibrium statistical mechanics. Each of the m subsystems are independent. each one at a time corresponds a micro state but at any time the prob. distributions of these microstates are invariant. Metropolis has another point of view of this problem. Since macro quantities, such as temperature, are invariant in equilibrium with respect to time, this means that the **time average** should be the same as the **ensemble average**, that is

$$\langle A(\sigma)\rangle \approx \frac{1}{N}\sum_{i=1}^{N} A(\sigma^{(i)}),$$

where $\{\sigma^{(i)}\}_{i=1}^{N}$ is the time series of the states evolving under the physics law. We notice that usually a physics law gives us a Makovian process. The equality between the the ensemble average and the time average is referred to as ergodicity in SM. In mathematics, the Metropolis algorithm is based on the **ergodic theory**.

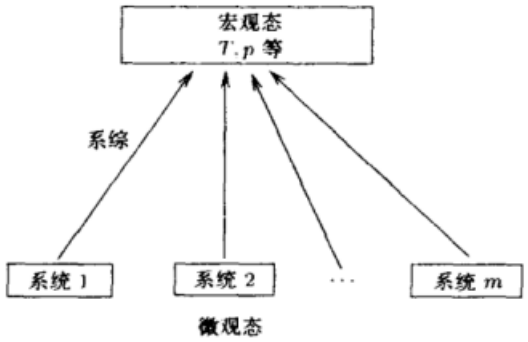


图 7.7　平衡态统计系综示意图

Figure 4.6: Ensembles in equilibrium.

Table 4.1: Comparison between frequentists and Bayesians.

	frequentists	Bayesians
Hypothesis testing	Set null and alternative hypotheses and use statistical tests to assess evidence against the null.	Consider **prior** beliefs when forming hypotheses.
Probability interpretation	Frame probability in terms of objective, long-term frequencies.	Interpret probabilities subjectively and update them as new data is collected.
Sampling	Emphasize random sampling and often require fixed sample sizes.	Can adapt well to varying sample sizes since Bayesians update their beliefs as more (observed) data comes in.
Assumption	Parameters that you estimate are fixed and are a single point while samples are random variables	There is a probability distribution around both the parameters and the samples.
The regime for application	Law of large number using a large amount of data.	Probability is degree of belief. Applicable when one has limited data, priors, and computing power.

4.3.2 Metropolis-Hasting Algorithm

Monte Carlo simulation

Let X be a discrete random vector whose set of possible values is x_j. Let the probability mass function of X be given by $P\{X = x_j\}$, and suppose that we are interested in calculating

$$\theta = E[h(X)] = \sum_j h(x_j)P\{X = x_j\},$$

for some specified function h. In situations where it is computationally difficult to evaluate the series, we often turn to simulation to approximate θ. The usual approach, called **Monte Carlo simulation**, is to use random numbers to generate a partial sequence of random vectors $X_1, X_2, ..., X_n$ having the mass function $P\{X = x_j\}$. Since the strong law of large numbers yields

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{h(X_i)}{n} = \theta,$$

it follows that we can estimate θ by letting n be large and using the average of the values of $h(X_i)$, $i = 1, \ldots, n$ as the estimator.

Discrete Case

The condition:

(1) X is a vector of dependent random variable whose samples cannot be generated via independent sampling technique.

(2) We only know $P(X = x_j) \propto b(j)$ but do not know the normalization factor. Assume that $B = \sum_j b(j)$ is finite so that the pmf is $P(X = x_j) = b(j)/B = \pi(j)$.

The MH algorithm is given as:

Step 1. Let Q be any specified irreducible Markov transition prob. matrix with entries $q(i, j)$.

Step 2. At the nth step, given $X_n = i$, draw a proposal $Y = j$ with prob. $P(Y = j) = q(i, j)$.

Step 3. Accept the proposal $X_{n+1} = Y = j$ with prob.

$$\alpha(i, j) = \min\left(\frac{b(j)q(j, i)}{b(i)q(i, j)}, 1\right).$$

Otherwise, reject the proposal and set $X_{n+1} = X_n = i$ with prob. $1 - \alpha(i, j)$.

Remark 4.3.1 *For the MH algorithm, we do not need to know the exact the distribution π, but only need to know $b(j)$ which is different from $\pi(j)$ by a normalization constant. The algorithm is OK when only $b(j)$ is given.*

Remark 4.3.2 *Numerically, draw a sample $z \sim U(0, 1)$, and set*

$$X_{n+1} = \begin{cases} Y = j, & \text{if } z < \alpha(i, j), \\ X_n = i, & \text{otherwise } z \geq \alpha(i, j). \end{cases}$$

Remark 4.3.3 *One can see that the sequence of states constitutes a Markov chian with transition prob. p_{ij} given by*

$$\begin{aligned} p_{ij} &= q(i, j)\alpha(i, j), \quad \text{if } j \neq i, \\ p_{ii} &= q(i, i) + \sum_{k\neq i} q(i, k)(1 - \alpha(i, k)). \end{aligned}$$

This Markov chain will be time reversible and have stationary prob. $\pi(j)$ if

$$\begin{aligned} \pi(i)p_{ij} &= \pi(j)p_{ji}, \quad \text{for } j \neq i, \\ \pi(i)q(i,j)\alpha(i,j) &= \pi(j)q(j,i)\alpha(j,i). \end{aligned} \tag{4.2}$$

We now check that the detailed balance equation (4.2) is satisfied in the MH algorithm. If

$$\alpha(i,j) = \frac{b(j)q(j,i)}{b(i)q(i,j)} < 1,$$

then $\alpha(j,i) = 1$ since $\frac{b(i)q(i,j)}{b(j)q(j,i)} > 1$ and equation (4.2) follows. If $\alpha(i,j) = 1$ then similarly again Eq. (4.2) follows.

Continuous Case

Consider a Markov transition kernel density $q : E \times E \to \mathbb{R}^+$ such that $\int_E q(x,y)dy = 1$, $\forall x \in E$. The MH method generates samples $\{x_i\}$ of a target density $\pi(x)$, as follows:

Step 1. At the ith step, given x_{i-1}, draw a proposal $u \sim q(x_{i-1}, \cdot)$.

Step 2. Accept the proposal, $x_i = u$ with probability $min(\alpha(x_{i-1}, u), 1)$. Otherwise set $x_i = x_{i-1}$. Here,

$$\alpha(x_{i-1}, u) = \frac{\pi(u)q(u, x_{i-1})}{\pi(x_{i-1})q(x_{i-1}, u)}.$$

Practically, we don't need to know the normalization constant for π since only the ratio of π is needed to determine the acceptance rate above. Numerically, this step can be realized by drawing a sample of the standard uniform distribution, $z \sim U[0,1]$, and then setting,

$$x_i = \begin{cases} u, & \text{if } z < \alpha(x_{i-1}, u), \\ x_{i-1}, & \text{otherwise } z \geq \alpha(x_{i-1}, u). \end{cases}$$

First, let's investigate why should we believe that x_i is a realization of the chain with stationary density $\pi(x)$. First of all, let x_i be a Markov chain with transition density $p(x_{i-1}, x_i)$ and we want to show that π is the stationary density. That is, we want to show that the chain generated by the MH method above satisfies the detailed balance condition. Notice that based on the MH method, we can write the transition kernel $p(x_{i-1}, u)$, which quantifies the probability of accepting u at the ith step given a realization of the chain x_{i-1} at the $(i-1)$th step, as a product of the probability of proposing u and the probability of accepting u,

$$p(x_{i-1}, u) = q(x_{i-1}, u) \min(\alpha(x_{i-1}, u), 1).$$

Thus,

$$\begin{aligned} \pi(x_{i-1})p(x_{i-1}, u) &= \pi(x_{i-1})q(x_{i-1}, u) \min(\frac{\pi(u)q(u, x_{i-1})}{\pi(x_{i-1})q(x_{i-1}, u)}, 1) \\ &= \pi(u)q(u, x_{i-1}) \min(1, \frac{\pi(x_{i-1})q(x_{i-1}, u)}{\pi(u)q(u, x_{i-1})}) \\ &= \pi(u)q(u, x_{i-1}) \min(\alpha(u, x_{i-1}), 1) \\ &= \pi(u)p(u, x_{i-1}), \end{aligned}$$

which means that x_i generated by MH method satisfies the detailed balanced condition (so that stationary) so $\pi(x)$ is the stationary density. To verify theo retically whether we obtain samples of π if we implement the MH method sufficiently large i is simply equivalent to asking whether the chain is recurrent and aperiodic (which is the hypothesis for the convergence). Detailed study of the convergence rates of some Metropolis-Hastings kernel densities were reported in [11].

Remark 4.3.4 *We should remark that if the proposal transition density q is chosen to be symmetric, $q(x, y) = q(y, x)$, then the resulting method is known as the* ***Metropolis scheme****, with an acceptance rate,*

$$\alpha(x_{i-1}, u) = \frac{\pi(u)}{\pi(x_{i-1})}. \quad \textit{This is Metropolis scheme.}$$

Intuitively, this rate compares the probability of the proposal u to that of the previous chain value, x_{i-1}. A popular choice of symmetric proposal density is Gaussian which yields the random walk Metropolis proposal,

$$q(x_{i-1}, u) = q(u|x_{i-1}) = N(x_{i-1}, C),$$

for some proposal covariance matrix C. Numerically, we realize the sample as follows,

$$u = x_{i-1} + C^{1/2}\xi_i, \quad \xi_i \sim N(0, I).$$

Remark 4.3.5 *There are many choice of q in history.*
(1) Tierney, 1994, Markov chain for exploring posterior distributions.
(2) Chib and Greenberg, 1995, Understanding the Metropolis-Hasting Algorithm.
(3) Hastings, 1970, $q(x, y) = q_2(y)$.

Remark 4.3.6 *Vanilla MCMC was abandoned long time ago since its acceptance rate is too low. For Vanilla MCMC,*

$$\alpha(i, j) = \pi(j)q(j, i), \quad \alpha(j, i) = \pi(i)q(i, j).$$

The low acceptance rate is caused by the too small values of above $\alpha(i, j)$ or $\alpha(j, i)$. Nowadays, it is still an important and popular topic to increase the acceptance rate, for instance, by using ***Hamiltonian MCMC*** *and* ***Langevin dynamics MCMC****.*

Remark 4.3.7 *(How useful MCMC is and why it works) In many real-world applications, we have to deal with complex probability distributions on complicated high-dimensional spaces. On rare occasions, it is possible to sample exactly from the distribution of interest, but typically exact sampling is difficult. Further, high dimensional spaces are very large, and distributions on these spaces are hard to visualize, making it difficult to even guess where the regions of high probability are located. As a result, it may be challenging to even design a reasonable proposal distribution to use with importance sampling.*

Markov chain Monte Carlo (MCMC) is a sampling technique that works remarkably well in many situations like this. Roughly speaking, my intuition for why MCMC often works well in practice is that
(a) the region of high probability tends to be "connected", that is, you can get from one point to another without going through a low-probability region, and
(b) we tend to be interested in the expectations of functions that are relatively smooth and have lots of "symmetries", that is, one only needs to evaluate them at a small number of representative points in order to get the general picture.

MCMC constructs a sequence of correlated samples $X_1, X_2, ...$ that meander through the region of high probability by making a sequence of incremental movements. Even though the samples are not independent, it turns out that under very general conditions, sample averages $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$ can be used to approximate expectations $Eh(X)$ just as in the case of simple ***Monte Carlo approximation****, and by a powerful result called* ***the ergodic theorem****, these approximations are guaranteed to converge to the true value.*

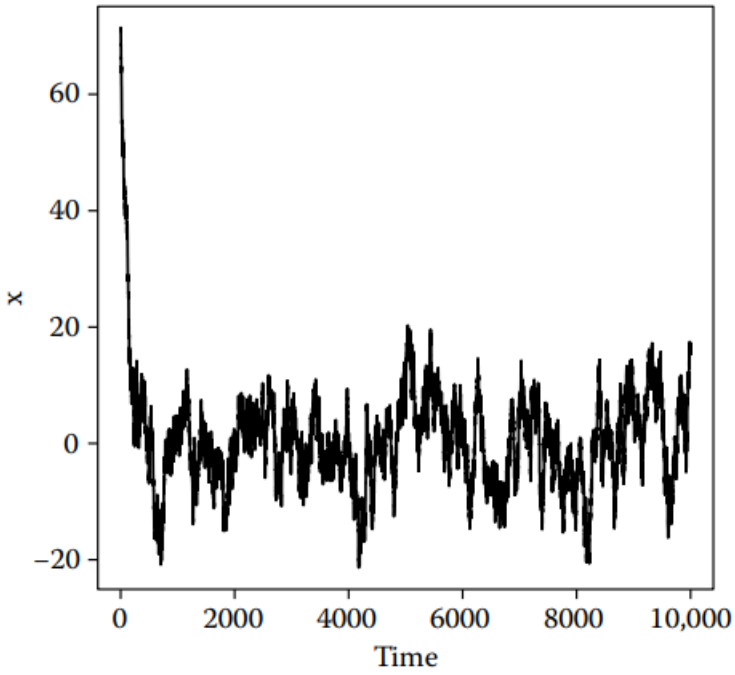


Figure 4.7: Burn-in is important for MCMC.

Remark 4.3.8 ***Burn-in*** *is important for MCMC so that at least one long run is required (see Fig. 4.7). For diagnostics, many relatively-long runs are preferred for detecting the pseudo-convergence. A Markov chain can appear to have converged to its equilibrium distribution when it has not. This happens when parts of the state space are* ***poorly connected*** *by the Markov chain dynamics: it takes many iterations to get from one part to another. When the time it takes to transition between these parts is much longer than the length of simulated Markov chain, then the Markov chain can appear to have converged but the distribution it appears to have converged to is the equilibrium distribution conditioned on the part in which the chain was started. We call this phenomenon* ***pseudo-convergence****. This phenomenon has also been called "****multimodality****" since it may occur when the equilibrium distribution is multimodal. But multimodality does not cause pseudo-convergence when the troughs between modes are not severe. Nor does pseudo-convergence only happen when there is multimodality. Some of the most alarming cases of pseudo-convergence occur when the state space of the Markov chain is discrete and "modes" are not well defined (Geyer and Thompson, 1995). Hence pseudo-convergence is a better term.*

Remark 4.3.9 *(Advantages and disadvantages for MCMC.)*
Advantages of MCMC:

• *applicable even when we can't directly draw samples. When we know nearly nothing about the model, we can apply MCMC for initial tests to give us something while not always give correct results or accurate results.*
• *works for complicated distributions in high-dimensional spaces, even when we don't know where the regions of high probability are (while MCMC is not guaranteed to give correct results). Notice that in high-dimensional spaces, the density $\pi(x)$ is hard for visualization, but on the other hand, the samples $\{x_i\}_{i=1}^N$ can be used to compute expectations of interested functions $Ef(x)$ or even for visualization.*
• *relatively easy to implement*
• *fairly reliable for sufficiently long runs.*
Disadvantages:
• *slower than simple Monte Carlo or importance sampling (i.e., requires more samples for the same level of accuracy)*
• *computationally expensive in* ***high dimensional spaces****, or* ***in multimodality cases****, or* ***when model is complicated****.*
• *can be very difficult to assess accuracy and evaluate convergence, even empirically. There is a great deal of theory about convergence of Markov chains. Unfortunately, none of it can be applied to get useful convergence information for most MCMC applications. Thus most users find themselves in the following situation we call black box MCMC:*

(1) You know nothing other than that. The Markov chain is a "black box" that you cannot see inside. When run, it produces output. That is all you know. You know nothing about the transition probabilities of the Markov chain, nor anything else about its dynamics. This Point 1 may seem extreme. You may know a lot about the particular Markov chain being used—for example, you may know that it is a Gibbs sampler—but if whatever you know is of no help in determining any convergence information about the Markov chain, then whatever knowledge you have is useless.

(2) You know nothing about the invariant distribution except what you may learn from running the Markov chain. This Point 2 may seem extreme. Many examples in the MCMC literature use small problems that can be done by independent and identically distributed (i.i.d.) Monte Carlo or even by pencil and paper and for which a lot of information about the invariant distribution is available, but in complicated applications point 2 is often simply true.

4.3.3 Gibbs sampling

The most widely used version of the Hastings–Metropolis algorithm is the Gibbs sampler.

Discrete Case

Let $\vec{X} = (X_1, ..., X_n)$ be a discrete random vector with probability mass function $p(\vec{x})$ that is only specified up to a multiplicative constant, and suppose that we want to generate a random vector whose distribution is that of $\vec{X}$. That is, we want to generate a random vector having mass function

$$p(\vec{x}) = Cg(\vec{x})$$

where $g(\vec{x})$ is known, but C is not.

Step 1. Choose an initial state $\vec{x} = (x_1, ..., x_n)$.

Step 2. For the current state $\vec{x}$, choose a coordinate which is equally likely to be any of the coordinates $1, ..., n$. If coordinate k is chosen, then generate a random variable Y whose probability distribution is given by

$$P\{Y = y\} = P\{X_k = y | X_j = x_j \text{ for } j = 1, ..., n \text{ with } j \neq k\},$$

where it is assumed that the above random variable Y can be generated having the above pmf. If $Y = y$, let the candidate state $\vec{y} = (x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n)$.

Step 3. The next state $\vec{x} = (x_1, ..., x_n)$ is set equal to $\vec{y}$. Repeat step 1 with this new state $\vec{x}$.

The Gibbs sampler uses the Metropolis-Hastings algorithm with the choice

$$\begin{aligned} q(\vec{x}, \vec{y}) &= \frac{1}{n} P\{X_k = y | X_j = x_j \text{ for } j = 1, ..., n \text{ with } j \neq k\} \\ &= \frac{1}{n} \frac{p(\vec{y})}{P\{X_j = x_j \text{ for } j = 1, ..., n \text{ with } j \neq k\}} \end{aligned}$$

for the Markov transition matrix Q. It is not difficult to verify that for this choice the acceptance probability $\alpha(\vec{x}, \vec{y})$ is given by

$$\begin{aligned} \alpha(\vec{x}, \vec{y}) &= \min\left(\frac{p(\vec{y})q(\vec{y}, \vec{x})}{p(\vec{x})q(\vec{x}, \vec{y})}, 1\right) \\ &= \min\left(\frac{p(\vec{y})p(\vec{x})}{p(\vec{x})p(\vec{y})}, 1\right) = 1. \end{aligned}$$

Hence, when utilizing the Gibbs sampler, the candidate state is always accepted as the next state of the Markov chain.

Continuous Case

Suppose $p(x, y)$ is a p.d.f. or p.m.f. that is difficult to sample from directly. Suppose, though, that we can easily sample from the conditional distributions $p(x|y)$ and $p(y|x)$. Roughly speaking, the Gibbs sampler proceeds as follows: set x and y to some initial starting values, then sample $x|y$, then sample $y|x$, then $x|y$, and so on. More precisely,

Step 0. Set (x_0, y_0) to some starting value.

Step 1. Sample $x_1 \sim p(x|y_0)$, that is, from the conditional distribution $X|Y = y_0$.
Sample $y_1 \sim p(y|x_1)$, that is, from the conditional distribution $Y|X = x_1$.

Step 2. Sample $x_2 \sim p(x|y_1)$, that is, from the conditional distribution $X|Y = y_1$.
Sample $y_2 \sim p(y|x_2)$, that is, from the conditional distribution $Y|X = x_2$.

$\vdots$

Each iteration $(1., 2., 3., \ldots)$ in the Gibbs sampling algorithm is sometimes referred to as a **sweep** or **scan**. The sampling steps within each iteration are sometimes referred to as **updates** or **Gibbs updates**. Note that when updating one variable, we always use the most recent value of the other variable (even in the middle of an iteration).

This procedure defines a sequence of pairs of random variables

$$(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), ...$$

which has the property of being a Markov chain—that is, the conditional distribution of (X_i, Y_i) given all of the previous pairs depends only on (X_{i-1}, Y_{i-1}).

Gibbs sampling with more than two variables is completely straightforward— roughly speaking, we cycle through the variables, sampling each from its conditional distributional given all the rest.

Remark 4.3.10 在以上算法中，坐标轴轮换采样不是必须的，可以在坐标轴轮换中引入随机性，这时候转移矩阵 Q 中任何两个点的转移概率中就会包含坐标轴选择的概率，而在通常的 *Gibbs Sampling* 算法中，坐标轴轮换是一个确定性的过程，也就是在给定时刻t，在一根固定的坐标轴上转移的概率是*1*。

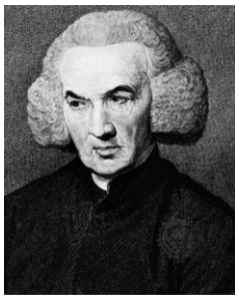
(Project 3) Show both the Metropolis-Hasting samplers and Gibbs samplers for 2D Gaussian distribution and the distribution with the density

$$p(x, y) \propto e^{-xy}\mathbf{1}(x, y \in (0, 1)).$$

4.4 Parameter estimation problems

4.4.1 The Bayesian approach

$\mathbb{P}(\text{this} = \text{Bayes} \mid \text{data}) < 1$



Richard Price

Pierre-Simon Laplace

Figure 4.8: Founders of Bayesian statistics.

• Thomas Bayes (1701–1761) was an ordained minister who was also a talented mathematician and a Fellow of the Royal Society. Bayes came up with an ingenious solution to this problem, but died before publishing it. Fortunately, his friend Richard Price carried his work further and published it in 1764. Apparently independently, Laplace rediscovered essentially the same idea in 1774, and developed it much further. (See Figure 4.8.)

• The idea is to assume a **prior** probability distribution for θ-that is, a distribution representing the plausibility of each possible value of θ before the data is observed. Then, to make inferences about θ, one simply considers the conditional distribution of θ given the observed data. This is referred to as the **posterior** distribution, since it represents the plausibility of each possible value of θ after seeing the data.

• Mathematically, this is expressed via **Bayes' theorem**,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta), \tag{4.3}$$

where x is the observed data (for example, $x = x_{1:n}$). In words, we say "the posterior is proportional to the likelihood times the prior". Bayes' theorem is essentially just the definition of conditional probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{4.4}$$

extended to conditional densities. (From the modern perspective, Bayes' theorem is a trivial consequence of the definition of a conditional density-however, when Bayes wrote his paper, the idea of a conditional probability density did not yet exist!)

• More generally, the Bayesian approach-in a nutshell-is to assume a prior distribution on any unknowns, and then just follow the rules of probability to answer any questions of interest. This provides a coherent framework for making inferences about unknown parameters θ as well as any future data or missing data, and for making rational decisions based on such inferences.

• Bayes, in a nutshell. The Bayesian approach can be summarized as follows: Assume a probability distribution on any unknowns (this the prior), assume the distribution of the knowns given the unknowns (this is the generating distribution or likelihood), and then just follow the rules of probability to answer any questions of interest.

An overarching theme of the Bayesian perspective is that uncertainty is quantified with probability distributions. Since essentially all statistical methods involve assuming the form of the generating distribution, it is the prior that distinguishes the Bayesian approach, and makes it possible to just follow the rules of probability.

What questions of interest often arise?

Here are some recurring examples:

• estimate some unknown parameter or property,

• infer hidden/latent variables or missing data,

• predict future data,

• test a hypothesis, or

• choose among competing models.

How is this done? What methods are employed?

In order to answer a question of interest, you usually have to get ahead of the posterior in one way or another, and compute one or more posterior expectations (integrals with respect to the posterior density). Three main categories of methods can be distinguished here: exact solution, deterministic approximation, and stochastic approximation.

1. Exact solution

In certain cases, it is computationally feasible to compute the posterior (and posterior expectations) exactly.

• Exponential families with conjugate priors often enable analytical solutions.

• Gaussians, in particular, are highly conducive to analytical solutions.

• For certain graphical models, dynamic programming can provide exact results.

2. Deterministic approximation

Methods include:

• numerical integration, a.k.a. quadrature/cubature

• quasi-Monte Carlo (QMC), low discrepancy sequences

• Laplace' s method / Laplace approximation

• expectation propagation (EP), variational Bayes (VB)

For low-dimensional integrals, numerical integration and QMC are superior to stochastic approximations. QMC can sometimes perform well in high-dimensional situations as well.

3. Stochastic approximation

For high-dimensional integrals, stochastic approximations are often the only option. The basic idea is that samples from the posterior can be used to approximate posterior expectations. Methods include:

• Monte Carlo approximation, importance sampling

• Markov chain Monte Carlo (MCMC) —Gibbs sampling, Metropolis algorithm, Metropolis–Hastings algorithm, slice sampling, Hamiltonian MCMC

• sequential importance sampling, sequential Monte Carlo, population Monte Carlo

• approximate Bayesian computation (ABC)

Overall recommendation: be pragmatic, not dogmatic

Overall, be pragmatic—that is, use what has been shown to work. As a default approach, the following will serve you well:

Design as a Bayesian, and evaluate as a frequentist.

In other words, construct models and procedures from a Bayesian perspective, and use frequentist tools to evaluate their empirical and theoretical performance. In the spirit of being pragmatic, it might seem unnecessarily restrictive to limit oneself to Bayesian procedures, and indeed, there are times when a non-Bayesian procedure may be preferable to a Bayesian one. However, typically, it turns out that there is no disadvantage in considering only Bayesian procedures—this has been shown formally via the "complete class theorems".

4.4.2 Applications of Bayesian statistics

• Tracking. For vehicle guidance, navigation, and control, it is essential to know the state of the vehicle (location, orientation, velocity) of the vehicle at any given time. Usually, an array of sensors provides various kinds of information of varying quality (e.g., compass, accelerometers, gyroscope, GPS, vision, laser scanner), and this must be combined with knowledge of the vehicle' s actions (e.g., wheels, propellors/turbines, rocket engines, ailerons), along with a physical model, in order to infer the state of the vehicle in real-time. In 1960, Rudolf Kalman proposed a solution using a Bayesian time-series model which became known as the Kalman filter. The Kalman filter and its successors have been extraordinarily successful-it is difficult to overstate their importance in the guidance systems of aircraft, spacecraft, and robotics.

• Phylogenetics. Understanding the evolutionary relationships among organisms—that is, the phylogenetic tree—is fundamental in nearly all biological research. Using genetic data from many organisms, along with models of how changes in the genome occur over time, researchers can infer the unknown evolutionary "family tree". Some of the dominant approaches use Bayesian inference (e.g., popular programs include MrBayes and BEAST) and these are widely used throughout biology.

• Computer science. Spam accounts for the majority of email traffic—typically between 60 to 70% of emails are spam. Yet, due to the sophisticated spam detection algorithms used by email service providers, very little spam gets through to your inbox—and only rarely is real mail classified as spam. For instance, in 2007, Gmail posted the chart, showing that the fraction of spam that gets through is very small indeed. Bayesian models are the most prominent methods for spam detection. A former Microsoft developer who moved to Google reportedly said, "Google uses Bayesian filtering the way Microsoft uses the if statement."

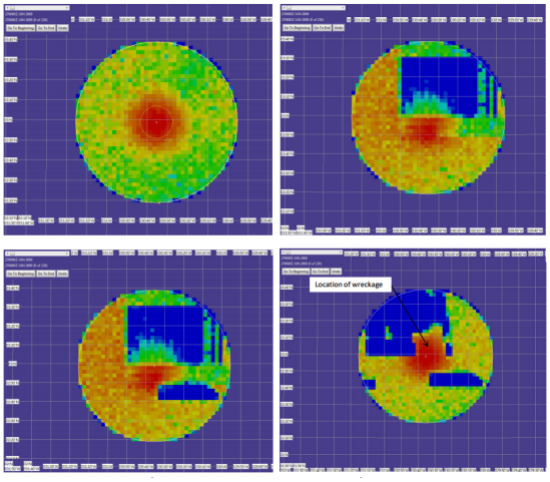


Figure 4.9: Prior and posteriors (after successive searches) for the location of the wreckage of Air France 447. (Stone et al. 2011).

• Search. On June 1, 2009, Air France Flight 447 crashed into the Atlantic Ocean, killing all aboard. Despite three intensive searches, the underwater wreckage had still not been found a year later. French authorities were eventually able to recover the wreckage with the help of a Bayesian search analysis provided by the Metron company (Figure 5). Bayesian search analysis involves formulating many hypothetical scenarios for what happened, constructing a probability distribution of the location under each scenario, and considering the posterior distribution on location given the searches conducted so far. It has also been used to find submarines and ships lost at sea.

Chapter 5

Poisson Process

Chapter 6

Continuous-Time Markov Chain

Chapter 7

Brownian Motion and Stationary Process

Chapter 8

Time Series

Bibliography

[1] 林元烈. 应用随机过程. 应用随机过程, 2002.

[2] 何书元. 随机过程. 北京大学出版社, 2008.

[3] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[4] J. Harlim. *Data-Driven Computational Methods: Parameter and Operator Estimations*. Cambridge University Press, UK, 2018.

[5] W. M. Jeffrey. Lecture notes on bayesian statistics. *Duke University, Durham, NC*, 2015.

[6] D. E. Knuth. *The Art of Computer Programming: Sorting and Searching*. Massachusetts: Addison-Wesley, 1973.

[7] G. F. Lawler. *Introduction to Stochastic Processes, Second Edition*. Introduction to Stochastic Processes, Second Edition, 2006.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and et al. *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1986.

[9] R. Rabiner, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

[10] S. M. Ross. Introduction to probability models, ninth edition. In *Academic Press, Inc.*, 2006.

[11] L. Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1728, 1994.

[12] C. Tijms, Henk. *A First Course in Stochastic Models*. Wiley, 2003.

[13] P. Zhang and T. Li. 数值分析. 北京大学数学教学系列丛书, 2007.