

Mathematical Methods for Computer Science

Probability methods

Dr R.J. Gibbens

Computer Laboratory
University of Cambridge

Computer Science Tripos, Part IB
Michaelmas Term 2013/14

Revision: 47e24e0 (2013-10-15)

Outline

- ▶ Probability methods (6 lectures, Dr R.J. Gibbens)
 - ▶ Limits and inequalities (3 lectures)
 - ▶ Markov chains (3 lectures)
- ▶ Fourier and related methods (6 lectures, Professor J. Daugman)

Reference books (Probability methods)

- ▶ (*) Ross, Sheldon M.
Probability Models for Computer Science.
Harcourt/Academic Press, 2002
- ▶ Mitzenmacher, Michael & Upfal, Eli.
Probability and Computing: Randomized Algorithms and Probabilistic Analysis.
Cambridge University Press, 2005

Some notation

RV	random variable
IID	independent, identically distributed
PGF	probability generating function $G_X(z)$
mgf	moment generating function $M_X(t)$
$X \sim U(0, 1)$	RV X has the distribution $U(0, 1)$, etc
$\mathbb{I}(A)$	indicator function of the event A
$\mathbb{P}(A)$	probability that event A occurs, e.g. $A = \{X = n\}$
$\mathbb{E}(X)$	expected value of RV X
$\mathbb{E}(X^n)$	n^{th} moment of RV X , for $n = 1, 2, \dots$
$F_X(x)$	distribution function, $F_X(x) = \mathbb{P}(X \leq x)$
$f_X(x)$	density of RV X given, when it exists, by $F'_X(x)$

Limits and inequalities

Limits and inequalities

We are familiar with limits of real numbers. If $x_n = 1/n$ for $n = 1, 2, \dots$ then $\lim_{n \rightarrow \infty} x_n = 0$ whereas if $x_n = (-1)^n$ no such limit exists.

Behaviour **in the long-run** or **on average** is an important characteristic of everyday life.

In this section we will be concerned with these notions of limiting behaviour when the real numbers x_n are replaced by random variables X_n . As we shall see there are several distinct notions of convergence that can be considered.

To study these forms of convergence and the limiting theorems that emerge we shall on the way also gather a potent collection of concepts and tools for the probabilistic analysis of models and systems.

Probabilistic inequalities

To help assess how close RVs are to each other it is useful to have methods that provide upper bounds on probabilities of the form

$$\mathbb{P}(X > a)$$

for fixed constants a , and where, for example, $X = |X_1 - X_2|$.

We shall consider several such bounds and related inequalities.

- ▶ Markov's inequality
- ▶ Chebyshev's inequality
- ▶ Lyapunov's inequality

Theorem (Markov's inequality)

If $\mathbb{E}(X) < \infty$ then for any $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}.$$

Proof.

We have that

$$\mathbb{I}(|X| \geq a) = \begin{cases} 1 & |X| \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$|X| \geq a \mathbb{I}(|X| \geq a)$$

hence

$$\mathbb{E}(|X|) \geq \mathbb{E}(a \mathbb{I}(|X| \geq a)) = a \mathbb{P}(|X| \geq a)$$

which yields the result.



Theorem (Chebyshev's inequality)

Let X be a RV with mean $\mu = \mathbb{E}(X)$ and finite variance $\sigma^2 = \text{Var}(X)$ then for all $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof.

Consider, for example, the case of a continuous RV X and put $Y = |X - \mu|$ so that $\mathbb{E}(Y^2) = \mathbb{E}((X - \mu)^2) = \sigma^2$ and then

$$\sigma^2 = \mathbb{E}(Y^2) = \int y^2 f_Y(y) dy = \int_{0 \leq y < a} y^2 f_Y(y) dy + \int_{y \geq a} y^2 f_Y(y) dy$$

so that

$$\sigma^2 \geq 0 + a^2 \mathbb{P}(Y \geq a).$$



Theorem (Lyapunov's inequality)

If $r \geq s > 0$ then $\mathbb{E}(|X|^r)^{1/r} \geq \mathbb{E}(|X|^s)^{1/s}$.

Proof.

Omitted.



Moment generating function

Definition

The **moment generating function** (mgf) of a RV X is given by

$$M_X(t) = \mathbb{E}(e^{tX})$$

and is defined for those values of $t \in \mathbb{R}$ for which this expectation exists.

Using the power series $e^x = 1 + x + x^2/2! + x^3/3! + \dots$ we see that

$$M_X(t) = \mathbb{E}(e^{tX}) = 1 + \mathbb{E}(X)t + \mathbb{E}(X^2)t^2/2! + \mathbb{E}(X^3)t^3/3! + \dots$$

and so the n^{th} moment of X , $\mathbb{E}(X^n)$, is given by the coefficient of $t^n/n!$ in the power series expansion of the mgf $M_X(t)$.

Note that for every RV, X , we have that $M_X(0) = 1$ since

$$M_X(0) = \mathbb{E}(e^{0X}) = \mathbb{E}(1) = 1.$$

Elementary properties of the mgf

1. If X has mgf $M_X(t)$ then $Y = aX + b$ has mgf $M_Y(t) = e^{bt}M_X(at)$.
2. If X and Y are **independent** then $X + Y$ has mgf $M_{X+Y}(t) = M_X(t)M_Y(t)$.
3. $\mathbb{E}(X^n) = M_X^{(n)}(0)$ where $M_X^{(n)}$ is the n^{th} derivative of M_X .
4. If X is a discrete RV taking values $0, 1, 2, \dots$ with **probability generating function** $G_X(z) = \mathbb{E}(z^X)$ then $M_X(t) = G_X(e^t)$.

Fundamental properties of the mgf

- 1. Uniqueness:** to each mgf there corresponds a unique distribution function having that mgf.
In fact, if X and Y are RVs with the **same** mgf in some region $-a < t < a$ where $a > 0$ then X and Y have the **same** distribution.
- 2. Continuity:** if distribution functions $F_n(x)$ converge pointwise to a distribution function $F(x)$, the corresponding mgf's (where they exist) converge to the mgf of $F(x)$. Conversely, if a sequence of mgf's $M_n(t)$ converge to $M(t)$ which is continuous at $t = 0$, then $M(t)$ is a mgf, and the corresponding distribution functions $F_n(x)$ converge to the distribution function determined by $M(t)$.

Example: exponential distribution

If X has an exponential distribution with parameter $\lambda > 0$ then $f_X(x) = \lambda e^{-\lambda x}$ for $0 < x < \infty$. Hence, for $t < \lambda$,

$$\begin{aligned}M_X(t) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \int_0^{\infty} \lambda e^{-(\lambda-t)x} dx \\ &= \left[-\frac{\lambda}{(\lambda-t)} e^{-(\lambda-t)x} \right]_0^{\infty} = \frac{\lambda}{\lambda-t}.\end{aligned}$$

For $t < \lambda$

$$\frac{\lambda}{(\lambda-t)} = \left(1 - \frac{t}{\lambda}\right)^{-1} = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \dots$$

and hence $\mathbb{E}(X) = 1/\lambda$ and $\mathbb{E}(X^2) = 2/\lambda^2$ so that

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1/\lambda^2.$$

Example: normal distribution

Consider a normal RV $X \sim N(\mu, \sigma^2)$ then $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
so that

$$\begin{aligned}M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(-2tx\sigma^2 + (x-\mu)^2)/2\sigma^2} dx.\end{aligned}$$

So, by completing the square,

$$\begin{aligned}M_X(t) &= e^{\mu t + \sigma^2 t^2/2} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-(\mu+t\sigma^2))^2/2\sigma^2} dx \right\} \\ &= e^{\mu t + \sigma^2 t^2/2}.\end{aligned}$$

Example: uniform distribution

Consider a uniform RV $X \sim U(a, b)$. Then

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

Hence, for $t \neq 0$,

$$\begin{aligned} M_X(t) &= \int_a^b \frac{e^{tx}}{b-a} dx \\ &= \left[\frac{e^{tx}}{(b-a)t} \right]_a^b \\ &= \frac{e^{bt} - e^{at}}{(b-a)t}. \end{aligned}$$

and $M_X(0) = 1$.

Theorem (Chernoff's bound)

Suppose that X has mgf $M_X(t)$ and $a \in \mathbb{R}$ then for all $t > 0$

$$\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t).$$

Proof.

Using Markov's inequality, we have that

$$\begin{aligned} \mathbb{P}(X \geq a) &= \mathbb{P}(e^{tX} \geq e^{ta}) \\ &\leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} \\ &= e^{-ta} M_X(t) \end{aligned}$$



Note that the above bound holds for all $t > 0$ so we can select the **best** such bound by choosing t to minimize $e^{-ta} M_X(t)$.

Notions of convergence: $X_n \rightarrow X$ as $n \rightarrow \infty$

For a sequence of RVs $(X_n)_{n \geq 1}$, we shall define several distinct notions of convergence to some RV X as $n \rightarrow \infty$.

Definition (Convergence in distribution)

$X_n \xrightarrow{D} X$ if $F_{X_n}(x) \rightarrow F_X(x)$ for all points x at which F_X is continuous.

Definition (Convergence in probability)

$X_n \xrightarrow{P} X$ if $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$ for all $\varepsilon > 0$.

Definition (Convergence almost surely)

$X_n \xrightarrow{a.s.} X$ if $\mathbb{P}(X_n \rightarrow X) = 1$.

Definition (Convergence in r^{th} mean)

$X_n \xrightarrow{r} X$ if $\mathbb{E}(|X_n - X|^r) \rightarrow 0$.

Convergence theorems

Theorem

If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{P} X$.

Theorem

If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$.

Theorem

If $r > s \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{s} X$.

Theorem

If $r \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{P} X$.

Theorem

If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{P} X$.

Proof.

Omitted.



Theorem

If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$.

Proof

We prove this theorem as follows. Fix, $\varepsilon > 0$ then

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x \cap X > x + \varepsilon) + \mathbb{P}(X_n \leq x \cap X \leq x + \varepsilon)$$

since $X > x + \varepsilon$ and $X \leq x + \varepsilon$ form a partition. But if $X_n \leq x$ and $X > x + \varepsilon$ then $|X_n - X| > \varepsilon$ and $\{X_n \leq x \cap X \leq x + \varepsilon\} \subset \{X \leq x + \varepsilon\}$. Therefore,

$$F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \varepsilon) + F_X(x + \varepsilon).$$

Similarly,

$$\begin{aligned} F_X(x - \varepsilon) &= \mathbb{P}(X \leq x - \varepsilon \cap X_n > x) + \mathbb{P}(X \leq x - \varepsilon \cap X_n \leq x) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon) + F_{X_n}(x). \end{aligned}$$

The proof is completed by noting that together these inequalities show that

$$F_X(x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \varepsilon) + F_X(x + \varepsilon).$$

But $X_n \xrightarrow{P} X$ implies that $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$. So, as $n \rightarrow \infty$, $F_{X_n}(x)$ is squeezed between $F_X(x - \varepsilon)$ and $F_X(x + \varepsilon)$.

Hence, if F_X is continuous at x , $F_{X_n}(x) \rightarrow F_X(x)$ and so $X_n \xrightarrow{D} X$. \square

Theorem

If $r > s \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{s} X$.

Proof.

Set $Y_n = |X_n - X| \geq 0$ then by Lyapunov's inequality

$$\mathbb{E}(Y_n^r)^{1/r} \geq \mathbb{E}(Y_n^s)^{1/s}.$$

Hence, if $\mathbb{E}(Y_n^r) \rightarrow 0$ then $\mathbb{E}(Y_n^s) \rightarrow 0$.



Theorem

If $r \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{P} X$.

Proof.

By Markov's inequality, for all $\varepsilon > 0$

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}(|X_n - X|)}{\varepsilon}.$$

But $X_n \xrightarrow{r} X$ implies $X_n \xrightarrow{1} X$ and so the right hand side tends to zero and as required $X_n \xrightarrow{P} X$. □

Limit theorems

Given a sequence of RVs $(X_n)_{n \geq 1}$, let

$$S_n = X_1 + X_2 + \cdots + X_n \quad \text{and} \quad \bar{X}_n = S_n/n.$$

What happens to \bar{X}_n for large n ?

Theorem (Weak Law of Large Numbers/WLLN)

Suppose $(X_n)_{n \geq 1}$ are IID RVs with finite mean μ (and finite variance σ^2) then $\bar{X}_n \xrightarrow{P} \mu$.

Theorem (Strong Law of Large Numbers/SLLN)

Suppose $(X_n)_{n \geq 1}$ are IID RVs with finite mean μ (and finite fourth moment) then $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.

Note that convergence to μ in the WLLN and SLLN actually means convergence to a **degenerate** RV, X , with $\mathbb{P}(X = \mu) = 1$.

WLLN

Theorem (Weak Law of Large Numbers/WLLN)

Suppose $(X_n)_{n \geq 1}$ are IID RVs with finite mean μ and finite variance σ^2 then $\bar{X}_n \xrightarrow{P} \mu$.

Proof.

Recall that $\mathbb{E}(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$. Hence, by Chebyshev's inequality, for all $\varepsilon > 0$

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

and so, letting $n \rightarrow \infty$,

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0$$

hence $\bar{X}_n \xrightarrow{P} \mu$ as required. □

SLLN

Theorem (Strong Law of Large Numbers/SLLN)

Suppose $(X_n)_{n \geq 1}$ are IID RVs with finite mean μ (and finite fourth moment) then $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.

Proof.

Omitted. □

Applications: estimating probabilities

Suppose we wish to estimate the probability, p , that we succeed when we play some game. For $i = 1, \dots, n$, let

$$X_i = \mathbb{I}(\{i^{\text{th}} \text{ game is success}\}).$$

So $\bar{X}_n = m/n$ if we succeed m times in n attempts. We have that $\mu = \mathbb{E}(X_i) = \mathbb{P}(X_i = 1) = p$ so then

$$m/n \xrightarrow{\text{a.s.}} p$$

by the SLLN.

Thus we have shown the important result that the empirical estimate of the probability of some event by its observed sample frequency converges to the correct value as the number of samples grows.

This result forms the basis of all simulation methods.

Applications: Shannon's entropy

Theorem (Asymptotic Equipartition Property/AEP)

If X_n is a sequence of IID discrete RV with probability distribution given by $\mathbb{P}(X_i = x) = p(x)$ for each $x \in I$ then

$$-\frac{1}{n} \log_2 p(X_1, X_2, \dots, X_n) \xrightarrow{P} H(X)$$

where Shannon's *entropy* is defined by

$$H(X) = H(X_1) = \dots = H(X_n) = - \sum_{x \in I} p(x) \log_2 p(x)$$

and

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i)$$

is the joint probability distribution of the n IID RVs X_1, X_2, \dots, X_n .

Proof.

Observe that $p(X_i)$ is a RV taking the value $p(x)$ with probability $p(x)$ and similarly $p(X_1, X_2, \dots, X_n)$ is a RV taking a value $p(x_1, x_2, \dots, x_n)$ with probability $p(x_1, x_2, \dots, x_n)$. Therefore,

$$\begin{aligned} -\frac{1}{n} \log_2 p(X_1, X_2, \dots, X_n) &= -\frac{1}{n} \log_2 \prod_{i=1}^n p(X_i) \\ &= -\frac{1}{n} \sum_{i=1}^n \log_2 p(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n (-\log_2 p(X_i)) \\ &\xrightarrow{P} \mathbb{E}(-\log_2 p(X_i)) && \text{by WLLN} \\ &= -\sum_{x \in I} p(x) \log_2 p(x) \\ &= H(X) \end{aligned}$$



AEP implications

By the AEP, for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left| -\frac{1}{n} \log_2 p(X_1, X_2, \dots, X_n) - H(X) \right| \leq \varepsilon\right) = 1$$

$$\lim_{n \rightarrow \infty} \mathbb{P}(H(X) - \varepsilon \leq -\frac{1}{n} \log_2 p(X_1, X_2, \dots, X_n) \leq H(X) + \varepsilon) = 1$$

$$\lim_{n \rightarrow \infty} \mathbb{P}(-n(H(X) - \varepsilon) \geq \log_2 p(X_1, X_2, \dots, X_n) \geq -n(H(X) + \varepsilon)) = 1$$

$$\lim_{n \rightarrow \infty} \mathbb{P}(2^{-n(H(X)+\varepsilon)} \leq p(X_1, X_2, \dots, X_n) \leq 2^{-n(H(X)-\varepsilon)}) = 1$$

Thus, the sequences of outcomes (x_1, x_2, \dots, x_n) for which

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

have a high probability and are referred to as **typical sequences**. An efficient (optimal) coding is to assign short codewords to such sequences leaving longer codewords for any non-typical sequence. Such long codewords must arise only rarely in the limit.

Central limit theorem

Theorem (Central limit theorem/CLT)

Let $(X_n)_{n \geq 1}$ be a sequence of IID RVs with mean μ , variance σ^2 and whose moment generating function converges in some interval $-a < t < a$ with $a > 0$. Then

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0, 1).$$

Proof of CLT

Set $Y_i = (X_i - \mu)/\sigma$ then $\mathbb{E}(Y_i) = 0$ and $\mathbb{E}(Y_i^2) = \text{Var}(Y_i) = 1$ so

$$M_{Y_i}(t) = 1 + \frac{t^2}{2} + o(t^2)$$

where $o(t^2)$ refers to terms of higher order than t^2 which will therefore tend to 0 as $t \rightarrow 0$. Also,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i.$$

Hence,

$$\begin{aligned} M_{Z_n}(t) &= \left(M_{Y_i} \left(\frac{t}{\sqrt{n}} \right) \right)^n \\ &= \left(1 + \frac{t^2}{2n} + o \left(\frac{t^2}{n} \right) \right)^n \\ &\rightarrow e^{t^2/2} \quad \text{as} \quad n \rightarrow \infty. \end{aligned}$$

But $e^{t^2/2}$ is the mgf of the $N(0, 1)$ distribution so, together with the continuity property, the CLT now follows as required.

CLT example

Suppose X_1, X_2, \dots, X_n are the IID RVs showing the n sample outcomes of a 6-sided die with common distribution

$$\mathbb{P}(X_i = j) = p_j, \quad j = 1, 2, \dots, 6$$

Set $S_n = X_1 + X_2 + \dots + X_n$, the total score obtained, and consider the two cases

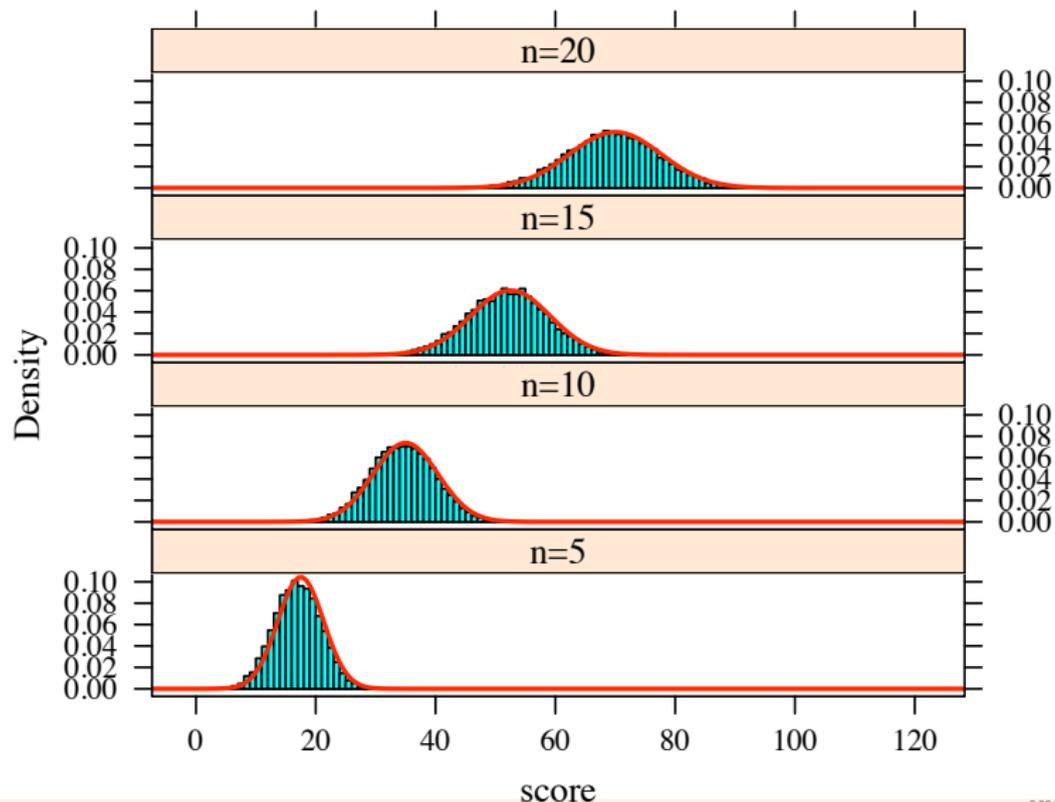
- ▶ **symmetric**: $(p_j) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ so that $\mu = \mathbb{E}(X_i) = 3.5$ and $\sigma^2 = \text{Var}(X_i) \approx 2.9$
- ▶ **asymmetric**: $(p_j) = (0.2, 0.1, 0.0, 0.0, 0.3, 0.4)$ so that $\mu = \mathbb{E}(X_i) = 4.3$ and $\sigma^2 = \text{Var}(X_i) \approx 4.0$

for varying sample sizes $n = 5, 10, 15$ and 20 .

The CLT tells us that for large n , S_n is approximately distributed as $N(n\mu, n\sigma^2)$ where μ and σ^2 are the mean and variance, respectively, of X_i .

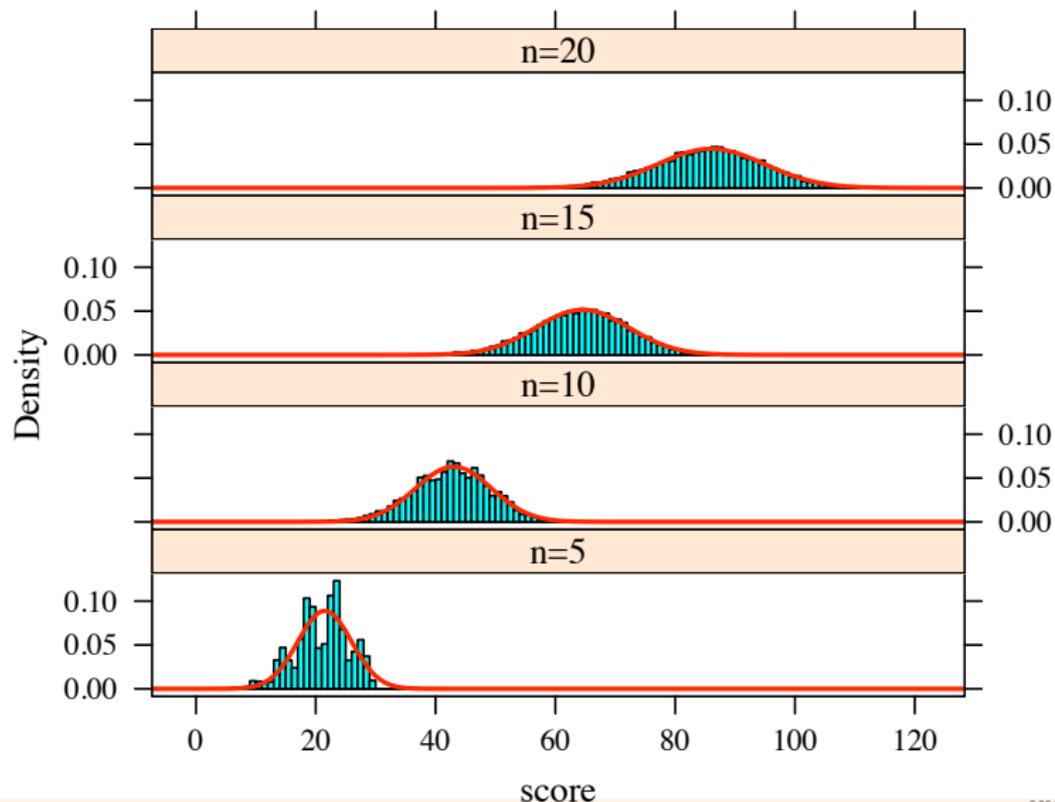
CLT example: symmetric

10,000 replications



CLT example: asymmetric

10,000 replications



Confidence intervals I

One of the major statistical applications of the CLT is to the construction of **confidence intervals**. The CLT shows that

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is asymptotically distributed as $N(0, 1)$. If, the true value of σ^2 is unknown we may estimate it by the **sample variance** given by

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

For instance, it can be shown that $\mathbb{E}(S^2) = \sigma^2$ and then

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

is approximately distributed as $N(0, 1)$ for large n .

Confidence intervals II

Define z_α so that $\mathbb{P}(Z > z_\alpha) = \alpha$ where $Z \sim N(0, 1)$ and so

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Hence,

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$
$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha.$$

The interval $\bar{X}_n \pm z_{\alpha/2} S/\sqrt{n}$ is thus an (approximate) $100(1 - \alpha)$ percent **confidence interval** for the unknown parameter μ .

Confidence intervals: example

Consider a collection of n IID RVs, X_i , with common distribution $X_i \sim \text{Pois}(\lambda)$. Hence,

$$\mathbb{P}(X_i = j) = \frac{\lambda^j e^{-\lambda}}{j!} \quad j = 0, 1, \dots$$

with mean $\mathbb{E}(X_i) = \lambda$.

Then a 95% confidence interval for the (unknown) mean value λ is given by

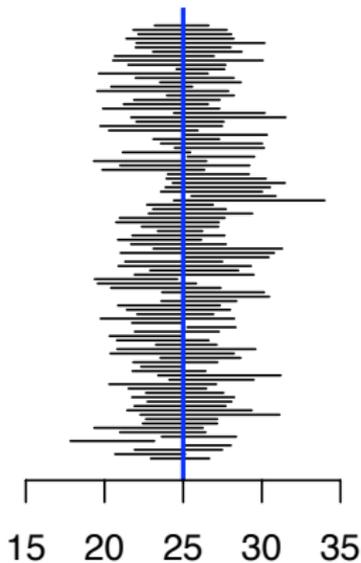
$$\bar{X}_n \pm 1.96S/\sqrt{n}$$

where $z_{0.025} = 1.96$.

Alternatively, to obtain 99% confidence intervals replace 1.96 by $z_{0.005} = 2.58$.

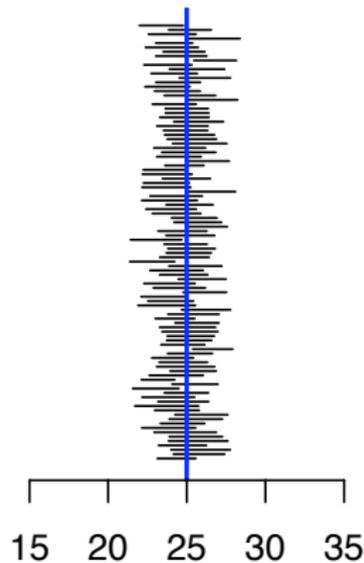
95% confidence intervals: illustration with $\lambda = 25$ and $\alpha = 5\%$

100 runs, $n = 10$



confidence interval

100 runs, $n = 40$



confidence interval

Monte Carlo simulation

Suppose we wish to estimate the value of π . One way to proceed is to perform the following experiment. Select a point $(X, Y) \in [-1, 1] \times [-1, 1]$, the square of side 2 and area 4 units, with X and Y chosen independently and uniformly in $[-1, 1]$. Now consider those points within unit distance of the origin then

$$\mathbb{P}((X, Y) \text{ lies in unit circle}) = \mathbb{P}(X^2 + Y^2 \leq 1) = \frac{\text{area of circle}}{\text{area of square}} = \frac{\pi}{4}.$$

Suppose we have access to a stream of random variables $U_i \sim U(0, 1)$ then $2U_i - 1 \sim U(-1, 1)$. Now set $X_i = 2U_{2i-1} - 1$, $Y_i = 2U_{2i} - 1$ and $H_i = \mathbb{I}(\{X_i^2 + Y_i^2 \leq 1\})$ so that

$$\mathbb{E}(H_i) = \mathbb{P}(X_i^2 + Y_i^2 \leq 1) = \frac{\pi}{4}.$$

Then by the SLLN the proportion of points (X_i, Y_i) falling within the unit circle converges almost surely to $\pi/4$. Furthermore, the CLT can be used to form confidence intervals.

Markov Chains

Markov chains

Definition (Markov chain)

Suppose that $(X_n)_{n \geq 0}$ is a sequence of discrete random variables taking values in some countable state space S . The sequence (X_n) is a **Markov chain** if

$$\mathbb{P}(X_n = x_n | X_0 = x_0, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$$

for all $n \geq 1$ and for all $x_0, x_1, \dots, x_n \in S$.

Since, S is countable we can always choose to label the possible values of X_n by integers and say that when $X_n = i$ the Markov chain is in the “ **i^{th} state at the n^{th} step**” or “**visits i at time n** ”.

Transition probabilities

The dynamics of the Markov chain are governed by the **transition probabilities** $\mathbb{P}(X_n = j | X_{n-1} = i)$.

Definition (time-homogeneous MC)

A Markov chain (X_n) is **time-homogeneous** if

$$\mathbb{P}(X_n = j | X_{n-1} = i) = \mathbb{P}(X_1 = j | X_0 = i)$$

for all $n \geq 1$ and states $i, j \in S$.

- ▶ We shall assume that our MCs are time-homogeneous unless explicitly stated otherwise.

Transition matrix

Definition (Transition matrix)

The **transition matrix**, P , of a MC (X_n) is given by $P = (p_{ij})$ where for all $i, j \in S$

$$p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i).$$

- ▶ Note that P is a **stochastic matrix**, that is, it has non-negative entries ($p_{ij} \geq 0$) and the row sums all equal one ($\sum_j p_{ij} = 1$).
- ▶ The transition matrix completely characterizes the dynamics of the MC.

Example

Suppose the states of the MC are $S = \{1, 2, 3\}$ and that the transition matrix is given by

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 2/3 & 0 & 1/3 \end{pmatrix}.$$

- ▶ Thus, in state 1 we are equally likely to be in any of the three states at the next step.
- ▶ In state 2, we can move with equal probabilities to 1 or 3 at the next step.
- ▶ Finally in state 3, we either move to state 1 with probability $2/3$ or remain in state 3 at the next step.

n -step transition matrix

Definition (n -step transition matrix)

The n -step transition matrix is $P^{(n)} = (p_{ij}^{(n)})$ where

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i).$$

Thus $P^{(1)} = P$ and we also set $P^{(0)} = I$, the identity matrix.

Chapman-Kolmogorov equations

Theorem (Chapman-Kolmogorov)

For all states i, j and for all steps m, n

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)}.$$

Hence, $P^{(m+n)} = P^{(m)} P^{(n)}$ and $P^{(n)} = P^n$, the n^{th} power of P .

Proof.

$$\begin{aligned} p_{ij}^{(m+n)} &= \mathbb{P}(X_{m+n} = j | X_0 = i) = \sum_k \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k p_{kj}^{(n)} p_{ik}^{(m)} \end{aligned}$$

The Chapman-Kolmogorov equations tell us how the long-term evolution of the MC depends on the short-term evolution specified by the transition matrix.

If we let $\lambda_i^{(n)} = \mathbb{P}(X_n = i)$ be the elements of a row vector $\lambda^{(n)}$ specifying the distribution of the MC at the n^{th} time step then the follow holds.

Lemma

$$\lambda^{(m+n)} = \lambda^{(m)} P^{(n)}$$

and so,

$$\lambda^{(n)} = \lambda^{(0)} P^{(n)}$$

where $\lambda^{(0)}$ is the initial distribution $\lambda_i^{(0)} = \mathbb{P}(X_0 = i)$.

Proof.

$$\begin{aligned}\lambda_j^{(m+n)} &= \mathbb{P}(X_{m+n} = j) = \sum_i \mathbb{P}(X_{m+n} = j | X_m = i) \mathbb{P}(X_m = i) \\ &= \sum_i \lambda_i^{(m)} p_{ij}^{(n)} = \left(\lambda^{(m)} P^{(n)} \right)_j\end{aligned}$$

Classification of states

Definition (Accessibility)

If, for some $n \geq 0$, $p_{ij}^{(n)} > 0$ then we say that state j is **accessible** from state i , written $i \rightsquigarrow j$.

If $i \rightsquigarrow j$ and $j \rightsquigarrow i$ then we say that i and j **communicate**, written $i \leftrightarrow j$.

Observe that the relation **communicates** \leftrightarrow is

- ▶ reflexive
- ▶ symmetric
- ▶ transitive

and hence is an equivalence relation. The corresponding equivalence classes partition the state space into subsets of states, called **communicating classes**, that communicate with each other.

Irreducibility

- ▶ A communicating class, C , that once entered can not be left is called **closed**, that is $p_{ij} = 0$ for all $i \in C, j \notin C$.
- ▶ A closed communicating class consisting of a single state is called **absorbing**.
- ▶ When the state space forms a single communicating class, the MC is called **irreducible** and is called **reducible** otherwise.

Recurrence and transience

Write for $n \geq 1$

$$f_{ij}^{(n)} = \mathbb{P}(X_1 \neq j, \dots, X_{n-1} \neq j, X_n = j | X_0 = i)$$

so that $f_{ij}^{(n)}$ is the probability starting in state i that we visit state j for the **first time** at time n . Also, let

$$f_{ij} = \sum_{n \geq 1} f_{ij}^{(n)}$$

the probability that we ever visit state j , starting in state i .

Definition

- ▶ If $f_{ij} < 1$ then state i is **transient**
- ▶ If $f_{ij} = 1$ then state i is **recurrent**.

Recurrence and transience, ctd

- ▶ Observe that if we return to a state i at some time n then the evolution of the MC is independent of the path before time n . Hence, the probability that we will return at least N times is f_{ii}^N .
- ▶ Now, if i is recurrent $f_{ii}^N = 1$ for all N and we are sure to return to state i infinitely often.
- ▶ Conversely, if state i is transient then $f_{ii}^N \rightarrow 0$ as $N \rightarrow \infty$ and so there is zero probability of returning infinitely often.

Theorem

- ▶ i is transient $\Leftrightarrow \sum_{n \geq 1} p_{ii}^{(n)}$ converges
- ▶ i is recurrent $\Leftrightarrow \sum_{n \geq 1} p_{ii}^{(n)}$ diverges

If i and j belong to the same communicating class then they are either both recurrent or both transient — the **solidarity property**.

Proof

First, define generating functions

$$P_{ii}(z) = \sum_{n=0}^{\infty} p_{ii}^{(n)} z^n \quad \text{and} \quad F_{ii}(z) = \sum_{n=0}^{\infty} f_{ii}^{(n)} z^n$$

where we take $p_{ii}^{(0)} = 1$ and $f_{ii}^{(0)} = 0$.

By examining the first time, r , that we return to i , we have for $m = 1, 2, \dots$ that

$$p_{ii}^{(m)} = \sum_{r=1}^m f_{ii}^{(r)} p_{ii}^{(m-r)}.$$

Now multiply by z^m and summing over m we get

$$\begin{aligned} P_{ii}(z) &= 1 + \sum_{m=1}^{\infty} z^m p_{ii}^{(m)} \\ &= 1 + \sum_{m=1}^{\infty} z^m \sum_{r=1}^m f_{ii}^{(r)} p_{ii}^{(m-r)} \\ &= 1 + \sum_{r=1}^{\infty} f_{ii}^{(r)} z^r \sum_{m=r}^{\infty} p_{ii}^{(m-r)} z^{m-r} \\ &= 1 + F_{ii}(z) P_{ii}(z) \end{aligned}$$

Thus, $P_{ii}(z) = 1/(1 - F_{ii}(z))$. Now let $z \nearrow 1$ then $F_{ii}(z) \rightarrow F_{ii}(1) = f_{ii}$ and $P_{ii}(z) \rightarrow \sum_n p_{ii}^{(n)}$.

If i is transient then $f_{ii} < 1$ so $\sum_n p_{ii}^{(n)}$ converges. Conversely, if i is recurrent then $f_{ii} = 1$ and $\sum_n p_{ii}^{(n)}$ diverges.

Furthermore, if i and j are in the same class then there exist m and n so that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Now, for all $r \geq 0$

$$p_{ii}^{(m+r+n)} \geq p_{ij}^{(m)} p_{jj}^{(r)} p_{ji}^{(n)}$$

so that $\sum_r p_{ij}^{(r)}$ and $\sum_k p_{ji}^{(k)}$ diverge or converge together. □

Mean recurrence time

First, let

$$T_j = \min\{n \geq 1 : X_n = j\}$$

be the time of the first visit to state j and set $T_j = \infty$ if no such visit ever occurs.

Thus, $\mathbb{P}(T_j = \infty | X_0 = i) > 0$ if and only if i is transient in which case $\mathbb{E}(T_j | X_0 = i) = \infty$.

Definition (Mean recurrence time)

The **mean recurrent time**, μ_i , of a state i is defined as

$$\mu_i = \mathbb{E}(T_i | X_0 = i) = \begin{cases} \sum_n n f_{ii}^{(n)} & \text{if } i \text{ is recurrent} \\ \infty & \text{if } i \text{ is transient.} \end{cases}$$

- ▶ Note that μ_i may still be infinite when i is recurrent.

Positive and null recurrence

Definition

A recurrent state i is

- ▶ **positive recurrent** if $\mu_i < \infty$ and
- ▶ **null recurrent** if $\mu_i = \infty$.

Example: simple random walk

Recall the **simple random walk** where $X_n = \sum_{i=1}^n Y_i$ where (Y_n) are IID RVs with $\mathbb{P}(Y_i = 1) = p = 1 - \mathbb{P}(Y_i = -1)$. Thus X_n is the position after n steps where we take unit steps up or down with probabilities p and $1 - p$, respectively.

It is clear that return to the origin is only possible after an even number of steps. Thus the sequence $(p_{00}^{(n)})$ alternates between zero and a positive value.

Periodicity

Let d_i be the greatest common divisor of $\{n : p_{ii}^{(n)} > 0\}$.

Definition

- ▶ If $d_i = 1$ then i is **aperiodic**.
- ▶ If $d_i > 1$ then i is **periodic** with period d_i .

- ▶ It may be shown that the period is a class property, that is, if $i, j \in C$ then $d_i = d_j$.

We will now concentrate on irreducible and aperiodic Markov chains.

Stationary distributions

Definition

The vector $\pi = (\pi_j; j \in S)$ is a **stationary distribution** for the MC with transition matrix P if

1. $\pi_j \geq 0$ for all $j \in S$ and $\sum_{j \in S} \pi_j = 1$
2. $\pi = \pi P$, or equivalently, $\pi_j = \sum_{i \in S} \pi_i p_{ij}$.

Such a distribution is stationary in the sense that $\pi P^2 = (\pi P)P = \pi P = \pi$ and for all $n \geq 0$

$$\pi P^n = \pi.$$

Thus if X_0 has distribution π then X_n has distribution π for all n . Moreover, π is the **limiting distribution** of X_n as $n \rightarrow \infty$.

Markov's example

Markov was lead to the notion of a Markov chain by study the patterns of vowels and consonants in text. In his original example, he found a transition matrix for the states {vowel, consonant) as

$$P = \begin{pmatrix} 0.128 & 0.872 \\ 0.663 & 0.337 \end{pmatrix}.$$

Taking successive powers of P we find

$$P^2 = \begin{pmatrix} 0.595 & 0.405 \\ 0.308 & 0.692 \end{pmatrix} \quad P^3 = \begin{pmatrix} 0.345 & 0.655 \\ 0.498 & 0.502 \end{pmatrix} \quad P^4 = \begin{pmatrix} 0.478 & 0.522 \\ 0.397 & 0.603 \end{pmatrix}.$$

As $n \rightarrow \infty$,

$$P^n \rightarrow \begin{pmatrix} 0.432 & 0.568 \\ 0.432 & 0.568 \end{pmatrix}.$$

Check that $\pi = (0.432, 0.568)$ is a stationary distribution, that is $\pi P = \pi$.

Limiting behaviour as $n \rightarrow \infty$

Theorem (Erdős-Feller-Pollard)

For all states i and j in an irreducible, aperiodic MC,

1. if the chain is transient, $p_{ij}^{(n)} \rightarrow 0$
2. if the chain is recurrent, $p_{ij}^{(n)} \rightarrow \pi_j$, where
 - 2.1 (null recurrent) either, every $\pi_j = 0$
 - 2.2 (positive recurrent) or, every $\pi_j > 0$, $\sum_j \pi_j = 1$ and π is the unique probability distribution solving $\pi P = \pi$.
3. In case (2), let T_i be the time to return to i then $\mu_i = \mathbb{E}(T_i) = 1/\pi_i$ with $\mu_i = \infty$ if $\pi_i = 0$.

Proof.

Omitted. □

Remarks

- ▶ The limiting distribution, π , is seen to be a stationary one. Suppose the current distribution is given by π and consider the evolution of the MC for a further period of T steps. Since π is stationary, the probability of being in any state i remains π_i , so we will make around $T\pi_i$ visits to i . Consequently, the mean time between visits to i would be $T/(T\pi_i) = 1/\pi_i$.
- ▶ Using $\lambda_j^{(n)} = \mathbb{P}(X_n = j)$ and since $\lambda^{(n)} = \lambda^{(0)}P^n$
 1. for transient or null recurrent states $\lambda^{(n)} \rightarrow 0$, that is, $\mathbb{P}(X_n = j) \rightarrow 0$ for all states j
 2. for a positive recurrent state, $p^{(n)} \rightarrow \pi > 0$, that is, $\mathbb{P}(X_n = j) \rightarrow \pi_j > 0$ for all j , where π is the unique probability vector solving $\pi P = \pi$.
- ▶ Note the distinction between a transient and a null recurrent chain is that in a transient chain we might never make a return visit to some state i and there is zero probability that we will return infinitely often. However, in a null recurrent chain we are sure to make infinitely many return visits but the mean time between consecutive visits is infinite.

Time-reversibility

Suppose now that $(X_n : -\infty < n < \infty)$ is an irreducible, positive recurrent MC with transition matrix P and unique stationary distribution π . Suppose also that X_n has the distribution π for all $-\infty < n < \infty$. Now define the **reversed chain** by

$$Y_n = X_{-n} \quad \text{for } -\infty < n < \infty$$

Then (Y_n) is also a MC and where Y_n has the distribution π .

Definition (Reversibility)

A MC (X_n) is **reversible** if the transition matrices of (X_n) and (Y_n) are equal.

Theorem

A MC (X_n) is reversible if and only if

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j \in S.$$

Proof.

Consider the transition probabilities q_{ij} of the MC (Y_n) then

$$\begin{aligned} q_{ij} &= \mathbb{P}(Y_{n+1} = j | Y_n = i) \\ &= \mathbb{P}(X_{-n-1} = j | X_{-n} = i) \\ &= \mathbb{P}(X_m = i | X_{m-1} = j) \mathbb{P}(X_{m-1} = j) / \mathbb{P}(X_m = i) \quad \text{where } m = -n \\ &= p_{ji} \pi_j / \pi_i. \end{aligned}$$

Hence, $p_{ij} = q_{ij}$ if and only if $\pi_i p_{ij} = \pi_j p_{ji}$. □

Theorem

For an irreducible chain, if there exists a vector π such that

1. $0 \leq \pi_i \leq 1$ and $\sum_i \pi_i = 1$
2. $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j \in S$

then the chain is reversible and positive recurrent, with stationary distribution π .

Proof.

Suppose that π satisfies the conditions of the theorem then

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

and so $\pi = \pi P$ and the distribution is stationary. □

The conditions $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j \in S$ are known as the **local balance** conditions.

Ehrenfest model

Suppose we have two containers A and B containing a total of m balls. At each time step a ball is chosen uniformly at random and switched between containers. Let X_n be the number of balls in container A after n units of time. Thus, (X_n) is a MC with transition matrix given by

$$p_{i,i+1} = 1 - \frac{i}{m}, \quad p_{i,i-1} = \frac{i}{m}.$$

Instead of solving the equations $\pi = \pi P$ we look for solutions to

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

which yields $\pi_i = \binom{m}{i} \left(\frac{1}{2}\right)^m$, a binomial distribution with parameters m and $\frac{1}{2}$.

Random walk on an undirected graph

Consider a **graph** G consisting of a countable collection of vertices $i \in N$ and a finite collection of edges $(i, j) \in E$ joining (unordered) pairs of vertices. Assume also that G is connected. A natural way to construct a MC on G uses a random walk through the vertices. Let v_i be the number of edges incident at vertex i . The random walk then moves from vertex i by selecting one of the v_i edges with equal probability $1/v_i$. So the transition matrix, P , is

$$p_{ij} = \begin{cases} \frac{1}{v_i} & \text{if } (i, j) \text{ is an edge} \\ 0 & \text{otherwise.} \end{cases}$$

Since G is connected, P is irreducible. The local balance conditions for $(i, j) \in E$ are

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

$$\pi_i \frac{1}{V_i} = \pi_j \frac{1}{V_j}$$

$$\frac{\pi_i}{\pi_j} = \frac{V_j}{V_i}.$$

Hence,

$$\pi_i \propto V_i$$

and the normalization condition $\sum_{i \in N} \pi_i = 1$ gives

$$\pi_i = \frac{V_i}{\sum_{j \in N} V_j}$$

and P is reversible.

Ergodic results

Ergodic results tell us about the limiting behaviour of averages taken over time. In the case of Markov Chains we shall consider the long-run proportion of time spent in a given state.

Let $V_i(n)$ be the number of visits to i before time n then

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{I}(\{X_k = i\}).$$

Thus, $V_i(n)/n$ is the proportion of time spent in state i before time n .

Theorem (Ergodic theorem)

Let (X_n) be a MC with irreducible transition matrix P then

$$\mathbb{P} \left(\frac{V_i(n)}{n} \rightarrow \frac{1}{\mu_i} \text{ as } n \rightarrow \infty \right) = 1$$

where $\mu_i = \mathbb{E}(T_i | X_0 = i)$ is the expected return time to state i .

Proof

If P is transient then the total number of visits, V_i , to i is finite with probability one, so

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \rightarrow 0 = \frac{1}{\mu_i} \quad n \rightarrow \infty.$$

Alternatively, if P is recurrent let $Y_i^{(r)}$ be the r^{th} duration between visits to any given state i . Then $Y_i^{(1)}, Y_i^{(2)}, \dots$ are non-negative IID RVs with $\mathbb{E}(Y_i^{(r)}) = \mu_i$.

But

$$Y_i^{(1)} + \dots + Y_i^{(V_i(n)-1)} \leq n - 1$$

since the time of the last visit to i before time n occurs no later than time $n - 1$ and

$$Y_i^{(1)} + \dots + Y_i^{(V_i(n))} \geq n$$

since the time of the first visit to i after time $n - 1$ occurs no earlier than time n .

Hence,

$$\frac{Y_i^{(1)} + \dots + Y_i^{(V_i(n)-1)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{Y_i^{(1)} + \dots + Y_i^{(V_i(n))}}{V_i(n)}.$$

However, by the SLLN,

$$\mathbb{P}\left(\frac{Y_i^{(1)} + \dots + Y_i^{(n)}}{n} \rightarrow \mu_i \text{ as } n \rightarrow \infty\right) = 1$$

and for P recurrent we know that $\mathbb{P}(V_i(n) \rightarrow \infty \text{ as } n \rightarrow \infty) = 1$. So,

$$\mathbb{P}\left(\frac{n}{V_i(n)} \rightarrow \mu_i \text{ as } n \rightarrow \infty\right) = 1$$

which implies

$$\mathbb{P}\left(\frac{V_i(n)}{n} \rightarrow \frac{1}{\mu_i} \text{ as } n \rightarrow \infty\right) = 1.$$



Example: random surfing on web graphs

Consider a web graph, $G = (V, E)$, with vertices given by a finite collection of web pages $i \in V$ and (directed) edges given by (i, j) whenever there is a hyperlink from page i to page j .

Random walks through the web graph have received much attention in the last few years.

Consider the following model, let $X_n \in V$ be the location (that is, web page visited) by the surfer at time n and suppose we choose X_{n+1} uniformly from the, $L(i)$, outgoing links from i , in the case where $L(i) > 0$ and uniformly among all pages in V if $L(i) = 0$ (the **dangling page** case).

Hence, the transition matrix, \hat{P}_{ij} , say, is given by

$$\hat{p}_{ij} = \begin{cases} \frac{1}{L(i)} & \text{if } (i,j) \in E \\ \frac{1}{|V|} & \text{if } L(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $|V|$ is the number of pages (that is, vertices) in the web graph. A potential problem remains in that \hat{P} may not be irreducible or may be periodic.

We make a further adjustment to ensure irreducibility and aperiodicity as follows.

For $0 < \alpha \leq 1$ set

$$p_{ij} = (1 - \alpha)\hat{p}_{ij} + \frac{\alpha}{|V|}.$$

We can interpret this as an “easily bored web surfer” model and see that the transitions take the form of a mixture of two distributions. With probability $1 - \alpha$ we follow the randomly chosen outgoing link (unless the page is dangling in which case we move to a randomly chosen page) while with probability α we jump to a random page selected uniformly from the entire set of pages V .

PageRank

Brin *et al* (1999) used this approach to define PageRank through the limiting distribution of this Markov Chain, that is π_i where the vector π satisfies

$$\pi = \pi P$$

They report typical values for α of between 0.1 and 0.2.

The ergodic theorem now tells us that the random surfer in this model spends a proportion π_i of the time visiting page i — a notion in some sense of the **importance** of page i .

Thus, two pages i and j can be ranked according to the total order defined by

$$i \geq j \quad \text{if and only if} \quad \pi_i \geq \pi_j.$$

See, “The PageRank Citation Ranking: Bring Order to the Web” Sergey Brin, Lawrence Page, Rajeev Motwani and Terry Winograd (1999) Technical Report, Computer Science Department, Stanford University.

<http://dbpubs.stanford.edu:8090/pub/1999-66>

Computing PageRank: the power method

We seek a solution to the system of equations

$$\pi = \pi P$$

that is, we are looking for an eigenvector of P (with corresponding eigenvalue of one). Google's computation of PageRank is one of the world's largest matrix computations.

The power method starts from an initial distribution $\pi^{(0)}$, updating $\pi^{(k-1)}$ by the iteration

$$\pi^{(k)} = \pi^{(k-1)} P = \dots = \pi^{(0)} P^k$$

Advanced methods from linear algebra can be used to speed up convergence of the power method and there has been much study of related MCs, to include web browser back buttons and many other properties and alternative notions of the “importance” of a web page.

Hidden Markov Models

An extension of Markov Chains is provided by **Hidden Markov Models** (HMM) where a statistical model of observed data is constructed from an underlying but usually hidden Markov Chain.

Such models have proved very popular in a wide variety of fields including

- ▶ speech and optical character recognition
- ▶ natural language processing
- ▶ bioinformatics and genomics.

We shall not consider these applications in any detail but simply introduce the basic ideas and questions that Hidden Markov Models address.

A Markov model with hidden states

Suppose we have a MC with transition matrix P but that the states i of the chain are not directly observable. Instead, we suppose that on visiting any state i at time n there is a randomly chosen output value or token, Y_n , that is observable.

The probability of observing the output token t when in state i is given by some distribution b_i , depending on the state i that is visited.

Thus,

$$\mathbb{P}(Y_n = t | X_n = i) = (b_i)_t$$

where $(b_i)_t$ is the t^{th} component of the distribution b_i .

For an excellent introduction to HMM, see “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition” Lawrence R. Rabiner. Proceedings of the IEEE, Vol 77, No 2, February 1988.

Three central questions

There are many variants of this basic setup but three central problems are usually addressed.

Definition (Evaluation problem)

Given a sequence y_1, y_2, \dots, y_n of observed output tokens and the parameters of the HMM (namely, P , b_i and the distribution for the initial state X_0) how do we compute

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \text{HMM parameters})$$

that is, the probability of the observed sequence given the model?

Such problems are solved in practice by the **forward algorithm**.

A second problem that may occur in an application is the **decoding problem**.

Definition (Decoding problem)

Given an observed sequence of output tokens y_1, y_2, \dots, y_n and the full description of the HMM parameters, how do we find the best fitting corresponding sequence of (hidden) states i_1, i_2, \dots, i_n of the MC?

Such problems are solved in practice by a dynamic programming approach called the **Viterbi algorithm**.

The third important problem is the **learning problem**.

Definition (Learning problem)

Given an observed sequence of output tokens y_1, y_2, \dots, y_n , how do we adjust the parameters of the HMM to maximize

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \text{HMM parameters})$$

The observed sequence used to adjust the model parameters is called a **training sequence**. Learning problems are crucial in most applications since they allow us to create the “**best**” models in real observed processes.

Iterative procedures, known as the **Baum-Welch method**, are used to solve this problem in practice.

Applications of Markov Chains

These and other applications of Markov Chains are important topics in a variety of Part II courses, including

- ▶ Artificial Intelligence II
- ▶ Bioinformatics
- ▶ Computer Systems Modelling