lecture 1

def a **stochastic process** is an indexed collection of random variables $\{X_t : \Omega \to S,\ t \in I\}$ with common codomain S (index set)

terminology:

- X_t: "state" of the process at t
- S: "state space", represents possible values / states of process
- $\{X_t(\omega)\}$: "sample path"
- $\{x_t\}_{t \in J}$, $J \subset I$: "realization" of the process

remarks:

- the index set I can be countable (e.g., $\mathbb{N}, \mathbb{Z}$) or uncountable (e.g., $\mathbb{R}$)
- the state space can be discrete or continuous (or neither)
- often, t indexes time

examples MONOPOLY: X_t = board position after t turns

$S = \{1, 2, \dots, 40\}$

$I = \mathbb{N}$

$P(X_0 = 1) = 1$

$$P(X_t = x \mid X_{t-1} = s) = \begin{cases} P[Y = (x-s) \bmod 40], & (x-s) \bmod 40 \in \{2, \dots, 12\} \\ 0, & \text{o/w} \end{cases}$$

$$P(Y = k) = \begin{cases} \frac{k-1}{36} & 2 \le k \le 7 \\ \frac{12-k+1}{36} & 7 < k \le 12 \end{cases}$$

assume $\{X_t\}$ is a discrete-space, discrete-time process with $I = \mathbb{N}$

def $\{X_t\}$ is a **markov chain** if

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, \dots, X_0 = x_0) = P(X_{t+1} = x_{t+1} \mid X_t = x_t)$$ for all $t \in I$ and $x_0, \dots, x_t \in S$

i.e., the past and future are independent given the present

def a markov chain is **time-homogeneous** if for every $i, j \in S$ and $t, s \in I$:

$$P(X_t = j \mid X_{t-1} = i) = P(X_s = j \mid X_{s-1} = i)$$

for time-homogeneous markov chains, we call

$$p_{ij} = P(X_t = j \mid X_{t-1} = i)$$

the **transition probability** of moving from state i to state j

and we call $P = \{p_{ij}\}$ the **transition matrix** for the process

remark for any transition matrix, row sums are $\sum_j p_{ij} = 1$.

def a matrix P is **stochastic** if $\sum_j p_{ij} = 1$ for each i

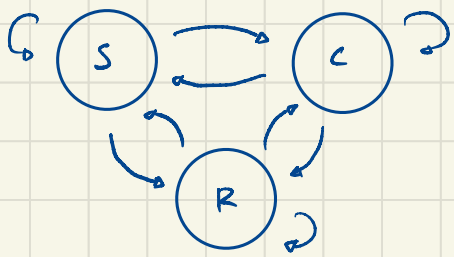
example X_t = weather in SLO on day t

$S = \{s, c, r\}$

sunny cloudy raining

suppose X_t is a time-homogeneous Markov chain with

$$P = \begin{array}{c} \\ s \\ c \\ r \end{array} \begin{array}{c} \begin{array}{ccc} s & c & r \end{array} \\ \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.2 & 0.7 & 0.1 \\ ? & 0.3 & 0.4 \end{bmatrix} \end{array}$$



check understanding: what's the probability it's sunny tomorrow?

what about the day after tomorrow?

see review on handout

review law of total probability:

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

in conditional form:

$$P(A \mid C) = \sum_i P(A \mid C \cap B_i) P(B_i \mid C)$$

in particular:

$$P(X=x \mid Y=y) = \sum_z P(X=x \mid Y=y, Z=z) P(Z=z \mid Y=y)$$

review recall that for a square matrix, the matrix powers are

$$P^2 = PP = \begin{bmatrix} & \\ & \end{bmatrix}\begin{bmatrix} & \\ & \end{bmatrix} \rightarrow (P^2)_{ij} = p_i^T p_j = \sum_k P_{ik} P_{kj}$$

(inner prod. of i^{th} row i, j^{th} col)

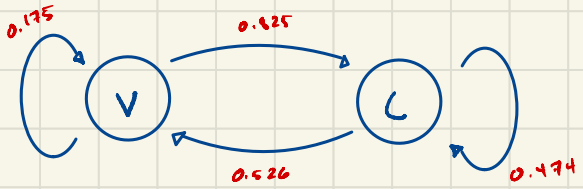
$$P^n = P P^{n-1} \Rightarrow (P^n)_{ij} = p_i^T p_j^{(n-1)} = \sum_k P_{ik} P_{kj}^{(n-1)}$$

example $X_t = \mathbb{1}\{t^{th} \text{ letter is consonant}\}$

$\mathcal{S} = \{0, 1\}$

$\mathcal{I} = \{1, 2, \ldots, 20000\}$

$$P = \begin{bmatrix} 0.175 & 0.825 \\ 0.526 & 0.474 \end{bmatrix} \qquad \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$



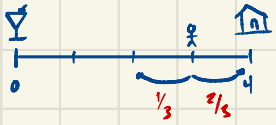
■ **Example 2.2 (Poetry and dependent sequences)** Andrei Andreyevich Markov, the Russian mathematician who introduced Markov chains over 100 years ago, first applied them in the analysis of the poem *Eugene Onegin* by Alexander Pushkin. In the first 20,000 letters of the poem, Markov counted (by hand!) 8,638 vowels and 11,362 consonants. He also tallied pairs of successive letters. Of the 8,638 pairs that start with vowels, 1,104 pairs are vowel–vowel. Of the 11,362 pairs that start with consonants, 3,827 are consonant–consonant. Markov treated the succession of letter types as a random sequence. The resulting transition matrix is

$$\boldsymbol{P} = \begin{array}{c} \\ \text{v} \\ \text{c} \end{array} \begin{array}{c} \begin{array}{cc} \text{vowel} & \text{consonant} \end{array} \\ \begin{pmatrix} 1,104/8,638 & 7,534/8,638 \\ 7,535/11,362 & 3,827/11,362 \end{pmatrix} \end{array} = \begin{pmatrix} 0.175 & 0.825 \\ 0.526 & 0.474 \end{pmatrix}.$$

Markov showed that the succession of letter types was not an independent sequence. For instance, if letter types were independent, the probability of two successive consonants would be $(11,362/20,000)^2 = 0.323$, whereas from Pushkin's poem the probability is $P_{cc} = 0.474$.

Markov's work was a polemic against a now obscure mathematician who argued that the law of large numbers only applied to independent sequences. Markov disproved the claim by showing that the Pushkin letter sequence was a dependent sequence for which the law of large numbers applied. ■

exercise bounded 1-D random walk on $\mathcal{S} = \{0, 1, 2, 3, 4\}$

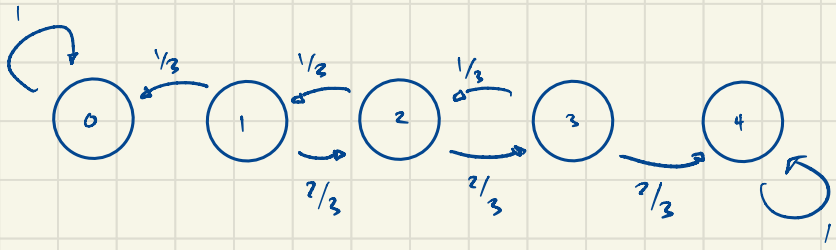


stop at 0 or 4.

i. find P

ii. draw the transitions as a graph

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



more practice: examples 2.6, 2.8, 2.9

lecture 2

to check whether $\{X_n\}$ is a homogeneous Markov chain:

1. $P(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1} \mid X_n = i_n)$

2. $P(X_{n+1} = j \mid X_n = i)$ does not depend on n for any n, $\{i_n\}$

example let $\{\varepsilon_n, n \in \mathbb{N}\}$ be IID $P(\varepsilon_n = 1) = P(\varepsilon_n = -1) = 1/2$ and

$X_n = \varepsilon_n + \varepsilon_{n-1}$ for $n = 1, 2, \dots$ so $X_n : \Omega \to \{-2, 0, 2\}$

Now: $P(X_{n+1} = 2 \mid X_n = 0) = P(X_{n+1} = 2 \mid X_n = 0, \varepsilon_n = 1)\, P(\varepsilon_n = 1 \mid X_n = 0)$

($P(\varepsilon_{n+1} = 1) = 1/2$; $P(\varepsilon_n = 1 \mid X_n = 0) = 1/2$)

$+ P(X_{n+1} = 2 \mid X_n = 0, \varepsilon_n = -1)\, P(\varepsilon_n = -1 \mid X_n = 0) = 1/4$

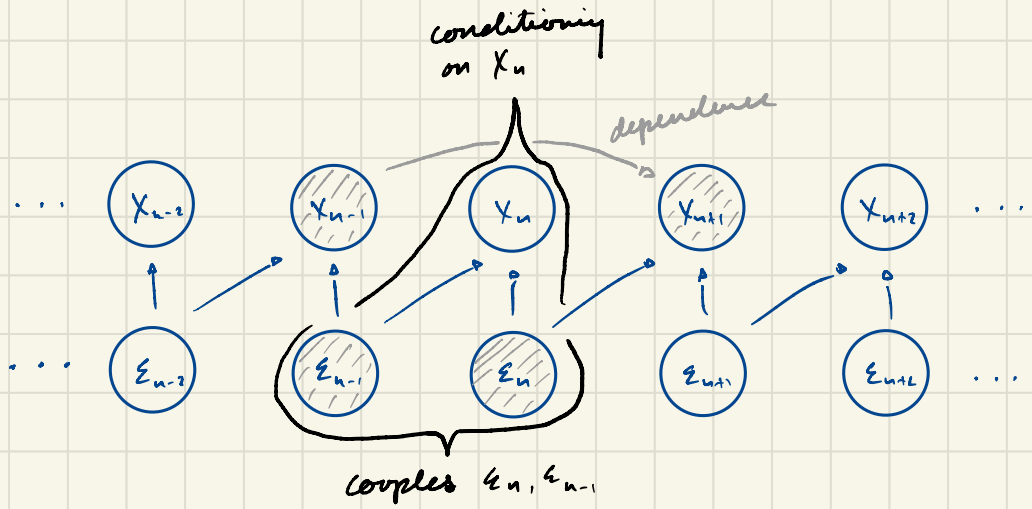
(first term: 0; $P(\varepsilon_n = -1 \mid X_n = 0) = 1/2$)

But: $P(X_{n+1} = 2 \mid X_n = 0, X_{n-1} = 2) = P(\varepsilon_{n+1} = 3) = 0$

$\Updownarrow$

$\varepsilon_n = -1, \varepsilon_{n-1} = 1, \varepsilon_{n-2} = 1$

so $\{X_n\}$ is NOT a Markov chain



"coupled" process

$\{(-1,-1), (-1,1), (1,-1), (1,1)\}$

but now let $Y_n = (\varepsilon_n, \varepsilon_{n-1}) : \Omega \to \{-1,1\}^2$. then for $i, j, \{i_n\}$ and n:

$$P[Y_{n+1} = j \mid Y_n = i_n, Y_{n-1} = i_{n-1}, \dots] = P[\varepsilon_{n+1} = j_2, \varepsilon_n = j_1 \mid \varepsilon_n = e_n, \varepsilon_{n-1} = e_{n-1}, \varepsilon_{n-2} = e_{n-2}, \dots]$$

$$= P[\varepsilon_{n+1} = j_2, \varepsilon_n = j_1 \mid \varepsilon_n = e_n, \varepsilon_{n-1} = e_{n-1}]$$

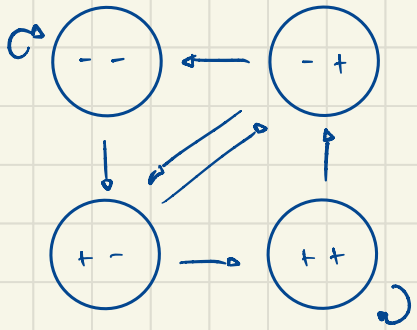
$$= P[Y_{n+1} = j \mid Y_n = i_n] \implies \text{Markov chain}$$

$$P[Y_{n+1} = (j_1, j_2) \mid Y_n = (i_1, i_2)] = P[\varepsilon_{n+1} = j_2, \varepsilon_n = j_1 \mid \varepsilon_n = i_2, \varepsilon_{n-1} = i_1]$$

$$= P[\varepsilon_{n+1} = j_2, \varepsilon_n = j_1 \mid \varepsilon_n = i_2]$$

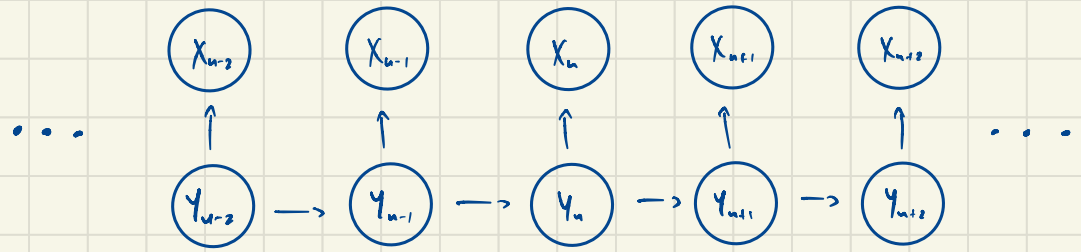
$$= \begin{cases} 1/2, & j_1 = i_2 \\ 0, & j_1 \neq i_2 \end{cases} \implies \text{homogeneous}$$

$$P = \begin{array}{c} \\ -- \\ -+ \\ +- \\ ++ \end{array} \begin{array}{c} \begin{array}{cccc} -- & -+ & +- & ++ \end{array} \\ \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix} \end{array}$$



so: $X_n = \varepsilon_n + \varepsilon_{n-1}$ is not a Markov chain but $Y_n = (\varepsilon_n, \varepsilon_{n-1})$ is

notice X_n is a function of Y_n: $X_n = 1^T Y_n$ emission function

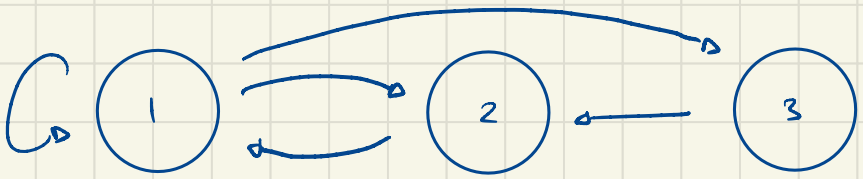


emissions

latent states

this is an example of a hidden Markov model

question how can you get from 1 to 3?



in 1: $1 \to 3$

in 2: $1 \to 1 \to 3$

in 3: $1 \to 1 \to 1 \to 3$

$1 \to 2 \to 1 \to 3$

def the probability of transitioning from i to j in n steps is

$$p_{ij}^{(n)} = P(X_t = j \mid X_{t-n} = i)$$

we call $p_{ij}^{(n)}$ the **n-step transition probability**

lemma $p_{ij}^{(n)} = (P^n)_{ij}$. proof is by induction.

for $n = 2$,

$$p_{ij}^{(2)} = P(X_t = j \mid X_{t-2} = i)$$

$$= \sum_k P(X_t = j \mid X_{t-1} = k, X_{t-2} = i) \, P(X_{t-1} = k \mid X_{t-2} = i)$$

(markov property; $P(X_t = j \mid X_{t-1} = k, X_{t-2} = i) = p_{kj}$, $P(X_{t-1} = k \mid X_{t-2} = i) = p_{ik}$)

$$= \sum_k p_{kj} p_{ik}$$

$$= (P^2)_{ij}$$

for n,

$$p_{ij}^{(n+1)} = P(X_t = j \mid X_{t-n-1} = i)$$

$$= \sum_k P(X_t = j \mid X_{t-1} = k, X_{t-n-1} = i) \, P(X_{t-1} = k \mid X_{t-n-1} = i)$$

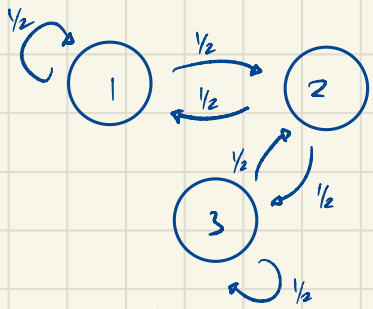
(markov property; $P(X_t = j \mid X_{t-1} = k, X_{t-n-1} = i) = p_{kj}$, $P(X_{t-1} = k \mid X_{t-n-1} = i) = p_{ik}^{(n)}$)

$$= \sum_k p_{kj} p_{ik}^{(n)}$$

$$= (P P^n)_{ij} = (P^{n+1})_{ij}$$

so by induction, P^n gives n-step transition probabilities. □

example $P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}$



$P^2 = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$ ← $P(1 \to 3 \text{ in 2 steps})$

lemma since $P^n = P^{n-m} P^m$, n-step transition probabilities can also be obtained by averaging over intermediate states reachable in m steps:

$$(P^n)_{ij} = \sum_k (P^{n-m})_{kj} (P^m)_{ik}$$

$$\Rightarrow P(X_t = j \mid X_{t-n} = i) = \sum_k \underbrace{P(X_t = j \mid X_{t-n+m} = k)}_{k \text{ to } j \text{ in } n-m \text{ steps}} \underbrace{P(X_{t-n+m} = k \mid X_{t-n} = i)}_{i \text{ to } k \text{ in } m \text{ steps}}$$

this is referred to as the Chapman-Kolmogorov equation(s)

example (contd) to get from $1 \to 3$, you must pass through 2. only one two-step path: $1 \to 2 \to 3$, with probability

$$P(X_{n+2} = 3 \mid X_n = 1) = \sum_{x=1}^{3} P(X_{n+2} = 3 \mid X_{n+1} = x) P(X_{n+1} = x \mid X_n = 1)$$

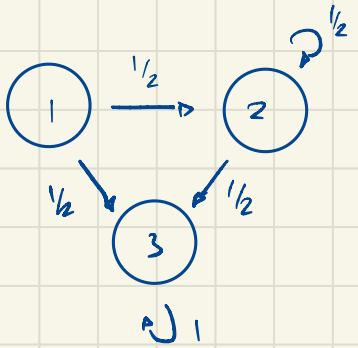
$$= \underbrace{0 \cdot 1/2}_{x=1} + \underbrace{1/2 \cdot 1/2}_{x=2} + \underbrace{1/2 \cdot 0}_{x=3}$$

$$= 1/4$$

note this is exactly what C-K gives. let $n = 1$ so:

$$(P^2)_{13} = \sum_{x=1}^{3} P_{1x} P_{x3} = 0 \cdot 1/2 + 1/2 \cdot 1/2 + 1/2 \cdot 0$$

example $P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$



two 2-step paths from 1 to 3:

$$P[1 \to 2 \to 3] = P_{12} P_{23} = 1/2 \cdot 1/2 = 1/4$$

$$P[1 \to 3 \to 3] = P_{13} P_{33} = 1/2 \cdot 1 = 1/2$$

$$P_{13}^{(2)} = 1/4 + 1/2 = 3/4$$

lecture 3

lemma (Markov property) if $\{X_n\}$ is a time-homogeneous Markov chain then

$$P(X_{n+1}=j \mid X_{n-m}=i, X_{n-m-1}=x_{n-m-1}, \dots) = P(X_{n+1}=j \mid X_{n-m}=i) = p_{ij}^{(m+1)} \quad \text{for } m<n$$

when $m=1$,

$$P(X_{n+1}=j \mid X_{n-1}=i, X_{n-2}=x_{n-2}, \dots) = \sum_k^1 P(X_{n+1}=j \mid X_n=k, X_{n-1}=i, \dots) P(X_n=k \mid X_{n-1}=i, \dots)$$
$$= \sum_k^1 p_{kj} p_{ik}$$
$$= p_{ij}^{(2)}$$

if true for $m-1$,

$$P(X_{n+1}=j \mid X_{n-m}=i, \dots) = \sum_k^1 P(X_{n+1}=j \mid X_{n-m+1}=k) P(X_{n-m+1}=k \mid X_{n-m}=i)$$
$$= \sum_k^1 p_{kj}^{(m)} p_{ik}$$
$$= p_{ij}^{(m+1)} \qquad \square$$

to obtain marginal distributions, one needs an initial distribution, i.e., a distribution, say α, of X_0. then if $\alpha_i = P(X_0=i)$:

$$P(X_t=j) = \sum_i \underbrace{P(X_t=j \mid X_0=i)}_{p_{ij}^n} \underbrace{P(X_0=i)}_{\alpha_i} = (\alpha P^t)_j$$

for any (finite) subcollection of times $t_1 < t_2 < \dots < t_k$, the joint distribution of $X_{t_1}, \dots, X_{t_k}$ is:

$$P(X_{t_1}=i_1, X_{t_2}=i_2, \dots, X_{t_k}=i_k) = (\alpha P^{t_1})_{i_1} p_{i_1 i_2}^{t_2-t_1} p_{i_2 i_3}^{t_3-t_2} \cdots p_{i_{k-1} i_k}^{t_k-t_{k-1}}$$

thus the joint probability of a realization $\{i_t\}$ is

$$P(X_0=i_0, \dots, X_t=i_t) = \alpha_{i_0} \cdot \prod_{k=1}^{t} p_{i_{k-1} i_k}$$

question what happens as $t \to \infty$? [examples in R]

def a Markov chain has a **limiting distribution** if $\lim_{n\to\infty} p_{ij}^{(n)}$ exists for each $i, j \in S$ and defines a valid probability measure on the state space for each i. we say that such a limit is:

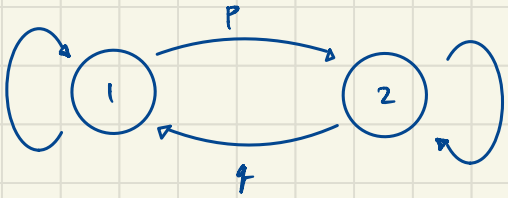
(i) **weak** if $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_{ij}$ depends on i

(ii) **strong** if $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ depends only on j

usually "limiting distribution" is used in the strong sense.

example consider a two-state homogeneous Markov chain with

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$



now $P^n = P^{n-1}P = \begin{bmatrix} p_{11}^{(n-1)} & 1-p_{11}^{(n-1)} \\ 1-p_{22}^{(n-1)} & p_{22}^{(n-1)} \end{bmatrix} \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$

$$\Rightarrow \quad p_{11}^{(n)} = p_{11}^{(n-1)}(1-p) + (1-p_{11}^{(n-1)})q$$

$$= q + (1-p-q)p_{11}^{(n-1)}$$

$$= q + (1-p-q)\left[q + (1-p-q)p_{11}^{(n-2)}\right]$$

$$\vdots$$

$$= q\left[1 + (1-p-q) + (1-p-q)^2 + \cdots + (1-p-q)^{n-1}\right] + (1-p-q)^n$$

$$= q\sum_{k=0}^{n-1}(1-p-q)^k + (1-p-q)^n$$

$$= q\,\frac{1-(1-p-q)^n}{p+q} + (1-p-q)^n$$

$$\sum_{i=0}^{k} a^i = \frac{1-a^{k+1}}{1-a}$$

$$= \frac{q}{p+q} + \frac{p}{p+q}(1-p-q)^n$$

$$p_{22}^{(n)} = (1-p_{22}^{(n-1)})p + p_{22}^{(n-1)}(1-q) \qquad \text{(same as above w/ } p \leftrightarrow q \text{ reversed)}$$

$$= \frac{p}{p+q} + \frac{q}{p+q}(1-p-q)^n$$

note $1 - \left[\frac{p}{p+q} + \frac{q}{p+q}(1-p-q)^n\right] = \frac{p+q}{p+q} - \frac{p}{p+q} - \frac{q}{p+q}(1-p-q)^n = \frac{q}{p+q} - \frac{q}{p+q}(1-p-q)^n$

so $P^n = \frac{1}{p+q}\begin{bmatrix} q + p(1-p-q)^n & p - p(1-p-q)^n \\ q - q(1-p-q)^n & p + q(1-p-q)^n \end{bmatrix}$

if $|1-p-q| < 1$ (ie, NOT $p=q=0$ or 1) then $(1-p-q)^n \to 0$ and

$\lim_{n\to\infty} P^n = \frac{1}{p+q}\begin{bmatrix} q & p \\ q & p \end{bmatrix}$ so $\pi = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}^T$ is the (strong) limiting dist.

if $p=q=0$ then $P = I$ so $P^n \to I$ and

$\left.\begin{aligned} \pi^{(1)} &= (1 \;\; 0)^T \\ \pi^{(2)} &= (0 \;\; 1)^T \end{aligned}\right\}$ are each (weak) limiting distributions

if $p=q=1$ then $P^n = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^n = \begin{cases} I, & n \text{ odd} \\ P, & n \text{ even} \end{cases}$

= so there is no limiting distribution

lemma if a limiting distribution exists then so does $\lim_n P(X_n = j)$ for every initial distribution.

pf/ if $\alpha_k = P(X_0 = k)$ and $\pi_{ij} = \lim_n p_{ij}^{(n)}$ then

$$P(X_n = j) = \sum_k P(X_n = j \mid X_0 = k) P(X_0 = k) = \sum_k p_{kj}^{(n)} \alpha_k \to \sum_k \pi_{kj} \alpha_k < \infty \quad \text{as } n \to \infty$$

the sum is finite since $\sum_k \pi_{kj} \alpha_k \le \left(\sum_k \pi_{kj}\right)\left(\sum_k \alpha_k\right) = 1$ □

note in the proof that:

- π is a strong limit $\Rightarrow \lim_n P(X_n = j) = \pi_j \sum_k \alpha_k = \pi_j$ independent of α
- π is a weak limit $\Rightarrow \lim_n P(X_n = j) = \sum_k \pi_{kj} \alpha_k$ depends on α

certain initial distributions may still produce stable long-run behavior even without a limit

example in the two-state chain let $p = q = 1$. if $\alpha^T = (1/2 \quad 1/2)$ then:

$$P(X_1 = j) = \alpha^T P = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix} = \alpha^T$$

and in general,

$$P(X_n = j) = \alpha^T P^n = \alpha^T (P P^{n-1}) = (\alpha^T P) P^{n-1} = \alpha^T P^{n-1} = \cdots = \alpha^T$$

so for $\alpha = (1/2 \quad 1/2)^T$:
- still no limit since $p_{ij}^{(n)}$ does not converge
- but $P(X_n = j)$ is constant
- so $\lim_n P(X_n = j)$ exists

def π is a **stationary distribution** if $\sum_{i,k} \pi_k p_{kj} = \pi_j$

remark when $\mathcal{S}$ is finite this can be expressed as $\pi^T P = \pi^T$

lemma if π is a limiting distribution then it is also a stationary distribution.

pf/
$$\pi_{ij} = \lim_n p_{ij}^{(n)} = \lim_n \sum_{i,k} p_{ik}^{(n-1)} p_{kj} = \sum_{i,k} \pi_{ik} p_{kj} \Rightarrow \pi_i \text{ is a stationary distribution } \square$$

note if $\{X_n\}$ has a strong limiting distribution then the stationary distribution is unique by uniqueness of limits.

questions
1. when do stationary distributions exist?
2. when do limits exist?

lecture 4

lemma let $\{X_n\}$ be a homogeneous markov chain. if the state space S is finite then a stationary distribution exists.

---

review let A be an $m \times m$ matrix. λ is an **eigenvalue** of A if for some nonzero $v \in \mathbb{R}^m$,

$$Av = \lambda v$$

if so, v is a corresponding (right) **eigenvector**.

since $Av = \lambda v \iff (A - \lambda I)v = 0$ and the RHS has a solution just in case $(A - \lambda I)$ is singular, the eigenvalues are the solutions of

$$\det(A - \lambda I) = 0$$

or the roots of the "**characteristic polynomial**" $p(\lambda) = \det(A - \lambda I)$

thus there are at most m distinct eigenvalues.

if $\det(A - tI) \propto (t - \lambda)^k$, then $(t - \lambda)$ can be factored out k times and the root λ is "repeated".

λ has **algebraic multiplicity** k if $p(t) \propto (t - \lambda)^k$

λ has **geometric multiplicity** q if $\dim\left(\ker(A - \lambda I)\right) = q$

there are at most q linearly independent eigenvectors associated with λ.

it is conventional to index the eigenvalues in descending order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_m$$

note that A & A^T have the same eigenvalues since

$$\det(A^T - \lambda I) = \det\left[(A - \lambda I)^T\right] = \det(A - \lambda I)$$

we say u is a **left eigenvector** of A (corresponding to λ) if it is an eigenvector of A^T, i.e.,

$$A^T u = \lambda u$$

note that, equivalently, $u^T A = \lambda u^T$.

left & right eigenvectors may differ since $Av \neq A^T v$ unless $A = A^T$.

the **spectral radius** of A is $\rho(A) = \max_i |\lambda_i|$

(Perron-Frobenius theorem) if A is non-negative (i.e., $a_{ij} \geq 0$) then $\rho(A)$ is an eigenvalue of A and there exists a corresponding nonnegative right eigenvector.

[for a proof, see any standard matrix analysis text, e.g., Horn & Johnson]

=

proof of the lemma: $|S| < \infty \Rightarrow \pi$ exists //

if $|S| = k$, then the transition matrix P is finite-dimensional ($k \times k$) and since P is stochastic ($\sum_j p_{ij} = 1$ for each i),

$$P\,\mathbf{1} = \begin{bmatrix} & p_{ij} & \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 1 \quad \Rightarrow \quad \lambda_1 = 1$$

($\mathbf{1}$: vector of ones; λ_1: largest eigenvalue)

1 is the largest eigenvalue since for any $x \in \mathbb{R}^k$, $\|Px\|_\infty \le \|x\|_\infty$ ($\|x\|_\infty = \max_i |x_i|$)
so for any eigenvector v with eigenvalue λ,

$$\|Pv\|_\infty = \|\lambda v\|_\infty = |\lambda| \|v\|_\infty \le \|v\|_\infty \quad \Rightarrow \quad |\lambda| \le 1$$

so 1 is also an eigenvalue of P^T, and thus $\rho(P^T) = 1$.

by the Perron-Frobenius theorem, there exists a nonnegative eigenvector π with eigenvalue $\rho(P^T)$, i.e.,

$$P^T \pi = \rho(P^T)\,\pi = \pi \quad \overset{\text{take transposes}}{\Leftrightarrow} \quad \pi^T P = \pi^T$$

so π is a stationary distribution □

remark the proof shows that for finite Markov chains, the stationary distributions are the left eigenvectors of P corresponding to $\lambda = 1$.

thus, eigendecomposition provides an easy way to find stationary distributions. also:

- geometric multiplicity of 1 for $P^T = 1 \Rightarrow$ stationary distribution unique
- " " $> 1 \Rightarrow$ " " not unique

algorithm to find the stationary distribution of a finite Markov chain:

1 compute eigenvalues $\lambda_1, \dots, \lambda_m$ of P^T and corresponding eigenvectors
(via spectral decomposition / diagonalization / Jordan decomposition / Schur decomposition)

2 for each nonnegative eigenvector v_j with $\lambda_j = 1$, normalize:

$$\pi_j = \frac{v_j}{\mathbb{1}^T v_j}$$

3 check for redundant π_j.

lecture 5

to characterize when limits exist we distinguish

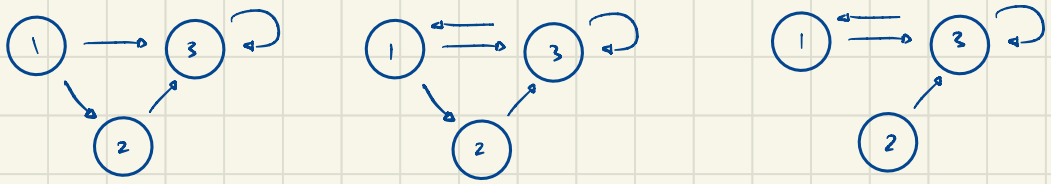
"recurrent" states are visited repeatedly
"transient" states may never be visited

def if $p_{ij}^{(n)} > 0$ for some $n \geq 0$ (important! includes "zeroth" transition), we say that

- state j is **accessible** from state i
- states i & j **communicate** if i is accessible from j and j is accessible from i

communication is an equivalence relation (reflexive, symmetric, transitive); the corresponding equivalence classes are called communication classes

example



communication classes: $\{1\}$ $\{2\}$ $\{3\}$ | $\{1,2,3\}$ | $\{1,3\}$ $\{2\}$

def the **first passage time** to state j is $T_j = \inf\{n > 0 : X_n = j\}$ (greatest lower bound)

the **hitting time** to state j is $\tau_j = \inf\{n \geq 0 : X_n = j\}$

these are examples of "stopping times".

exercise 3.22

$$P = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \end{array}$$

(a) $P(T_1 \geq n \mid X_0 = 1) = P(X_{n-1} = X_{n-2} = \cdots = X_1 = 2 \mid X_0 = 1)$

$$= \prod_{k=1}^{n-2} \underbrace{P(X_{k+1} = 2 \mid X_k = 2)}_{P_{22} = (1-q)} \underbrace{P(X_1 = 2 \mid X_0 = 1)}_{P_{12} = p}$$

$$= p(1-q)^{n-2} \longrightarrow 0 \quad \text{as } n \to \infty \qquad (q > 0)$$

so X_n returns to 1 almost surely (i.e., w.p. 1)

(b) $\mathbb{E}(T_1 \mid X_0 = 1) = \sum_{n=1}^{\infty} P(T_1 \geq n \mid X_0 = 1)$

$P(T_1 \geq 1 \mid X_0 = 1) = 1$

$$= 1 + \sum_{n=2}^{\infty} P(T_1 \geq n \mid X_0 = 1)$$

$$= 1 + p \sum_{j=0}^{\infty} (1-q)^j$$

$$= 1 + \frac{p}{q}$$

$$= \frac{q+p}{q}$$

$$= \frac{1}{\pi_1}$$

expected recurrence time

the expected hitting time from state 2 is:

condition on X_1

$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)]$

$$\mathbb{E}(T_1 \mid X_0 = 2) = 1 \cdot P(X_1 = 1 \mid X_0 = 2) + \left[1 + \mathbb{E}(T_1 \mid X_1 = 2)\right] P(X_1 = 2 \mid X_0 = 2)$$

$$= q + (1-q)\left[1 + \mathbb{E}(T_1 \mid X_0 = 2)\right]$$

$$= 1 + (1-q)\mathbb{E}(T_1 \mid X_0 = 2)$$

$$\Rightarrow \quad \mathbb{E}(T_1 \mid X_0 = 2) = \frac{1}{q}$$

shorter when chain mixes fast $(p + q > 1)$

now denote $f_j = P(T_j < \infty \mid X_0 = j)$: probability of returning to j eventually

$h_{ij} = P(\tau_j < \infty \mid X_0 = i)$: probability of hitting j eventually, starting from i

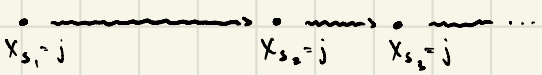
def state j is : { **recurrent** if $f_j = 1$ (return to j "almost surely") ; **transient** if $f_j < 1$ (chance of never returning) }

consider the number of visits to state j : $N_j = \sum_{n=0}^{\infty} \mathbb{1}\{X_n = j\}$ (I_n for short)

define $S_1, S_2, \dots, S_k, \dots$ to be successive visit times.

$S_1 = \inf\{n \geq 0 : X_n = j\}$ $\quad (= \tau_j)$

$S_{k+1} = \inf\{n \geq S_k : X_n = j\}$



now $\{N_j \geq n\} \iff \{S_n < \infty\}$.

for $n = 1$: $P(S_1 < \infty \mid X_0 = i) = h_{ij}$

for $n > 1$: $P(S_n < \infty \mid S_{n-1} < \infty) = P(S_n < \infty \mid X_{S_{n-1}} = j)$

$= P(T_j < \infty \mid X_0 = j)$ (homogeneity)

$= f_j$

so $P(S_n < \infty \mid X_0 = i) = \prod_{k=2}^{n} \underbrace{P(S_n < \infty \mid S_{n-1} < \infty)}_{f_j} \underbrace{P(S_1 < \infty \mid X_0 = i)}_{h_{ij}} = h_{ij} f_j^{n-1}$

and $\mathbb{E}(N_j \mid X_0 = i) = \sum_{n=1}^{\infty} P(N_j \geq n \mid X_0 = i) = h_{ij} \sum_{n=1}^{\infty} f_j^{n-1}$ ← geometric series

$$= \begin{cases} \frac{h_{ij}}{1 - f_j}, & f_j < 1 \quad \text{(transient)} \\ \infty, & f_j = 1 \text{ and } h_{ij} > 0 \\ 0, & f_j = 1 \text{ and } h_{ij} = 0 \end{cases} \quad \text{(recurrent)}$$

if visited eventually, recurrent states are visited infinitely often!

in particular, starting from state j, the expected number of returns is :

$$\mathbb{E}(N_j \mid X_0 = j) = \begin{cases} \frac{1}{1 - f_j}, & j \text{ transient} \\ \infty, & j \text{ recurrent} \end{cases}$$

now since $\mathbb{E}(N_j \mid X_0 = j) = \sum_{n=1}^{\infty} f_j^{n-1}$ diverges $\Leftrightarrow f_j = 1$
converges $\Leftrightarrow f_j < 1$

and $\mathbb{E}(N_j \mid X_0 = i) = \mathbb{E}\left[\sum_{n=0}^{\infty} I_n \mid X_0 = i\right] = \sum_{n=0}^{\infty} \mathbb{E}(I_n \mid X_0 = i) = \sum_{n=0}^{\infty} P(X_n = j \mid X_0 = i)$
$= \sum_{n=0}^{\infty} p_{ij}^{(n)}$

we obtain an alternate characterization of recurrence / transience:

lemma state j is: { transient just in case $\sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty$; recurrent just in case $\sum_{n=0}^{\infty} p_{jj}^{(n)} = \infty$ }

lemma the states in a communication class are either all transient or all recurrent.

pf/ suppose i communicates with j so that $p_{ij}^{(r)} > 0$ and $p_{ji}^{(s)} > 0$ for some r, s.
then by chapman-kolmogorov ($P^n = P^{n-m} P^m$):

$$p_{ii}^{(r+n+s)} = \sum_k \sum_\ell p_{ik}^{(r)} p_{k\ell}^{(n)} p_{\ell j}^{(s)} \geq p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)}$$

$$\Rightarrow \sum_{n=0}^{\infty} p_{ii}^{(n)} \geq \sum_{n=r+s}^{\infty} p_{ii}^{(n)} = \sum_{n=0}^{\infty} p_{ii}^{(n+r+s)} \geq p_{ij}^{(r)} \left(\sum_{n=0}^{\infty} p_{jj}^{(n)}\right) p_{ji}^{(s)}$$

so: (i) if j is recurrent, so is i
(ii) if i is transient, so is j
} i, j are either both recurrent or both transient. □

def a Markov chain is irreducible if all states communicate.

lemma if an irreducible Markov chain is finite, then all states are recurrent.

pf/ suppose all states are transient. then $\mathbb{E}(N_j \mid X_0 = i) = \frac{h_{ij}}{1 - f_j} < \infty$, but:

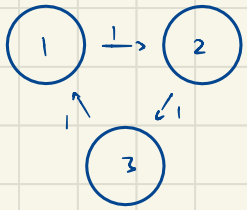
$$\sum_j N_j = \sum_j \sum_n \mathbb{1}\{X_n = j\} = \sum_n \overbrace{\sum_j \mathbb{1}\{X_n = j\}}^{\text{\# states @ time } n} = \sum_{n=0}^{\infty} 1 = \infty \quad \text{w.p. } 1$$

$$P\left[\sum_i \mathbb{1}\{X_n = j\} = 1\right] = P\left[\bigcup_j \{X_n = j\}\right] = 1$$

$$\Rightarrow \mathbb{E}\left[\sum_j N_j \mid X_0 = i\right] = \sum_j \mathbb{E}(N_j \mid X_0 = i) = \infty \quad *$$

question: is recurrence enough to get limits?

consider:



finite + irreducible $\Rightarrow$ all recurrent

but $p_{12}^{(n)} = \mathbb{1}\{n \bmod 3 = 1\}$ has no limit:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	...
$p_{12}^{(n)}$	0	1	0	0	1	0	0	1	0	0	1	0	0	...

so recurrence is not enough — you also can't have periodic behavior.

def the **period** of state j is $d(j) = \begin{cases} \gcd\{n>0 : p_{jj}^{(n)} > 0\} & , \ f_j > 0 \\ \infty & , \ f_j = 0 \end{cases}$

if $d(i) = 1$ we say i is **aperiodic**

informally, possible return times are all multiples of $d(i)$. when $1 < d(i) < \infty$, passing through state i is periodic, which inhibits limiting behavior unless i is transient.

lemma: if i communicates with j then $d(i) = d(j)$.

pf/ by hypothesis $p_{ij}^{(r)} > 0$ and $p_{ji}^{(s)} > 0$ for some $r, s > 0$, so

$$p_{ii}^{(r+s)} = \sum_k p_{ik}^{(r)} p_{ki}^{(s)} \geq p_{ij}^{(r)} p_{ji}^{(s)} > 0 \Rightarrow d(i) \text{ divides } r+s$$

if $p_{jj}^{(n)} > 0$ for some $n > 0$ then:

$$p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)} > 0 \Rightarrow d(i) \text{ divides } r+s+n$$

so $d(i)$ divides n and $d(i) \leq d(j)$; but by the same argument with $i \leftrightarrow j$ switched, $d(j) \leq d(i)$. so $d(i) = d(j)$. $\square$

lemma L1 for a finite Markov chain, a strong limit exists if the chain is irreducible & aperiodic.

pf sketch

$\{X_n\}$ is aperiodic & irreducible $\Leftrightarrow$ (takes some work) P is "primitive", i.e., $(P^n)_{ij} > 0$ for some n

when P is primitive:

i. the Perron-Frobenius eigenvalue $\rho(P) = 1$ has algebraic multiplicity 1 (→ stationary dist. is unique)

ii. $|\lambda| < 1$ for all other eigenvalues λ

now let π be the stationary distribution (exists per earlier lemma) and

$$L = 1\pi^T = \begin{bmatrix} \pi^T \\ \vdots \\ \pi^T \end{bmatrix}$$

(projection onto 1D right eigenspace spanned by $(1 \cdots 1)$ and assoc. w/ $\lambda = 1$)

so $PL = P1\pi^T = 1\pi^T = L$ and $LP = 1\pi^T P = 1\pi^T = L$.

let $R = P - L$, and note $\begin{cases} R1 = P1 - L1 = 0 \Rightarrow RL = 0 \\ \pi^T R = \underbrace{\pi^T P}_{\pi^T} - \underbrace{\pi^T L}_{\text{also } \pi^T} = 0 \Rightarrow LR = 0 \end{cases}$

now P & R have the same eigenvalues except for $\lambda = 1$, since L is a projection onto the associated eigenspace, so $\rho(R) < 1$. it follows that $R^n \to 0$.

then $\left.\begin{array}{l} P = L + R \\ LR = RL = 0 \end{array}\right\} \Rightarrow P^n = (L+R)^n = L^n + R^n = L + R^n \to L$

so $P^n \to 1\pi^T$ ∎

lecture 6

irreducibility is not a necessary condition for a strong limit:

$$P = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \quad \text{gives} \quad P^n \to \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

all you need is that eventually the chain ends up staying in one (recurrent) communication class with probability 1.

def a communication class C is **closed** if no other states are accessible from within the class, i.e.,

$$p_{ij} = 0 \quad \text{for each } i \in C \text{ and } j \notin C$$

recurrent classes are always closed since if $i \in C$, $j \notin C$, $p_{ij} > 0$, then

$$P(T_i < \infty \mid X_0 = i) = \sum_k p_{ik} \underbrace{P(T_i < \infty \mid X_1 = k)}_{=0 \text{ when } k = j} = \sum_{k \neq j} p_{ik} P(T_i < \infty \mid X_1 = k) \leq \sum_{k \neq j} p_{ik} = 1 - p_{ij} < 1 \quad \#$$

lemma for a finite chain, a communication class is closed iff it is recurrent.

pf/ apply the argument showing irreducibility $\Rightarrow$ recurrence restricted to C.

def for any finite Markov chain, the **canonical decomposition** is

$$S = T \cup C_1 \cup \cdots \cup C_k$$

(T: transient states; $C_1, \ldots, C_k$: closed (recurrent) classes)

thus by permuting the rows & columns of P (i.e., relabeling states) one has:

$$P = \begin{array}{c} \\ T \\ C_1 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} T & C_1 & \cdots & C_k \end{array} \\ \left[\begin{array}{c|ccc} Q & R_1 & \cdots & R_k \\ \hline 0 & P_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{array}\right] \end{array} = \begin{bmatrix} Q & R \\ 0 & B \end{bmatrix} \quad \text{(canonical form)}$$

$$\Rightarrow \quad P^n = \begin{bmatrix} Q^n & \sum_{k=0}^{n-1} Q^k R B^{n-k-1} \\ 0 & B^n \end{bmatrix}$$

lemma L2 P^n converges just in case every closed class is aperiodic.

Proof sketch

owing to transience, $Q^n 1 < 1$ entrywise for all $n > n^*$. it follows $\rho(Q) < 1$ and therefore $Q^n \to 0$ always.

now B^n converges $\Leftrightarrow$ P_k^n converges and P_k^n converges $\Leftrightarrow$ C_k aperiodic:

($\Leftarrow$) if C_k are aperiodic, then P_k are primitive and by lemma L1 $P_k^n \to 1\pi_k^\top$.

($\Rightarrow$) assume P_k^n converges but $d(j) > 1$ for $j \in C_k$.
then $(P_k^n)_{jj} = 0$ for n not d but $(P_k^{nd})_{jj} > 0$ for all $n > n^*$ ✳
(✳ two subsequences w/ differing limits)
so $d(j) = 1$ and C_k is aperiodic.

lastly, if $B^n \to L$ and $Q^n \to 0$ then $\sum_{k=0}^{n-1} Q^k R B^{n-k-1}$ converges:

$$\sum_{k=0}^{n-1} Q^k R B^{n-k-1} = \underbrace{\sum_{k=0}^{n-1} Q^k R (B^{n-k-1} - L)}_{\text{error term}} + \underbrace{\sum_{k=0}^{n-1} Q^k R L}_{\text{limit term}} \to (I-Q)^{-1} R L$$

error term $\to 0$ (takes some work); limit term $\to (I-Q)^{-1} R L$

so P^n converges iff $C_1, \dots, C_k$ are aperiodic and the limit is

$$P^n \to \begin{bmatrix} 0 & (I-Q)^{-1} R \Lambda \\ 0 & \Lambda \end{bmatrix} \quad \text{where } \Lambda = \operatorname{diag}\left(1\pi_1^\top, \dots, 1\pi_k^\top\right)$$

□

theorem T1 for a finite Markov chain $\{X_n\}$:

(i) a weak limiting distribution exists if & only if all closed classes are aperiodic

(ii) a strong limiting distribution exists if & only if there is exactly one closed class and it is aperiodic.

(iii) a stationary distribution always exists

algorithm to find the limiting distributions (when it exists):

1 write P in canonical form
2 obtain the left eigenvector π_u^T for each $P_1, \dots, P_c$
3 compute $(I - Q)^{-1} R \Lambda$

Lecture 7

For any finite chain, the canonical decomposition gives:

$$P = \begin{array}{c} \\ T \\ C_1 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} T & C_1 & \cdots & C_k \end{array} \\ \left[\begin{array}{c|ccc} Q & R_1 & \cdots & R_k \\ \hline 0 & P_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{array}\right] \end{array} = \begin{bmatrix} Q & R \\ 0 & B \end{bmatrix}$$

canonical form

we showed P^n converges iff $C_1, \dots, C_k$ are aperiodic and the limit is

$$P^n \to \begin{bmatrix} 0 & (I-Q)^{-1} R \Lambda \\ 0 & \Lambda \end{bmatrix} \quad \text{where } \Lambda = \operatorname{diag}\left(\mathbb{1}\pi_1^T, \dots, \mathbb{1}\pi_k^T\right)$$

stationary dist. of $P_1, \dots, P_k$

when the limit exists, this representation facilitates computing several quantities of interest:

1. $F = (I-Q)^{-1}$ ("fundamental" matrix (?)) : $F_{ij} = \mathbb{E}(N_j \mid X_0 = i)$ expected number of visits to j

2. $(F\mathbb{1})_i = \mathbb{E}(\tau_C \mid X_0 = i)$ expected time to absorption into a closed class
   ($\tau_C = \inf\{n \geq 0 : X_n \notin T\}$)

3. $(F R_k \mathbb{1})_i = P(X_n \text{ enters } C_k \mid X_0 = i)$ absorption probability

4. $\mathbb{E}(T_j \mid X_0 = j) = \frac{1}{\pi_j}$ for $j \in \bigcup_k C_k$ expected return times

well state these without proof.

example 3.27 gambler's ruin w/ $p = 0.6$, $k = 5$.

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & & & \\ 0.4 & & 0.6 & & & \\ & 0.4 & & 0.6 & & \\ & & 0.4 & & 0.6 & \\ & & & 0.4 & & 0.6 \\ & & & & & 1 \end{bmatrix} \end{array} \xrightarrow{\text{permute}} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 0 & 5 \end{array} \\ \begin{bmatrix} & 0.6 & & & 0.4 & \\ 0.4 & & 0.6 & & & \\ & 0.4 & & 0.6 & & \\ & & 0.4 & & & 0.6 \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 0 \\ 5 \end{array}$$

here $S = T \cup C_1 \cup C_2$, $T = \{1, 2, 3, 4\}$, $C_1 = \{0\}$, $C_2 = \{5\}$

now $Q = \begin{bmatrix} & 0.6 & & \\ 0.4 & & 0.6 & \\ & 0.4 & & 0.6 \\ & & 0.4 & \end{bmatrix}$ $R = \begin{bmatrix} 0.4 & \\ & \\ & \\ & 0.6 \end{bmatrix}$ $B = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$

and $F = (I - Q)^{-1} = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1.54 & 1.35 & 1.07 & 0.64 \\ 0.90 & 2.25 & 1.78 & 1.09 \\ 0.47 & 1.18 & 2.25 & 1.35 \\ 0.19 & 0.47 & 0.90 & 1.54 \end{bmatrix} \end{array}$

if $X_0 = 2$ w.p. 1 (ie, $\alpha = [0 \ 0 \ 1 \ 0 \ 0 \ 0]^T$) ...

a. how many times will you have \$1 ?

$$\mathbb{E}(N_1 \mid X_0 = 2) = F_{21} = 0.9$$

see also examples in §3.8

b. how long will you play ?

$$\mathbb{E}(\tau_c \mid X_0 = 2) = (F1)_2 = 6$$

c. likelihood of winning ?

$$P(X_t \text{ enters } C_2 \mid X_0 = 2) = (FR)_{22} = 0.64$$

d. chance you get to $1 before $4?

trick: pretend 4, 1 are absorbing states & rewrite F

	2	3	1	4	0	5
2		0.6	0.4			
3	0.4			0.6		
1	0.6				0.4	
4		0.4				0.6
0					1	
5						1

$\Rightarrow \quad \tilde{Q} = \begin{bmatrix} & 0.6 \\ 0.4 & \end{bmatrix} \qquad \tilde{R} = \begin{bmatrix} 0.4 & \\ & 0.6 \end{bmatrix}$

now find $P(\text{absorb into } 1) = (\tilde{F}\tilde{R})_{11}$

$\tilde{F} = \begin{bmatrix} 1.32 & 0.79 \\ 0.53 & 1.32 \end{bmatrix}$ so $(\tilde{F}\tilde{R})_{11} = \begin{bmatrix} 1.32 & 0.79 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0 \end{bmatrix} = 0.528$

example 3.29

$P =$

	d	Fr	So	Jr	Se	G
d	1					
Fr	0.06	0.03	0.91			
So	0.06		0.03	0.91		
Jr	0.04			0.03	0.93	
Se	0.04				0.03	0.93
G						1

→

	Fr	So	Jr	Se	d	Gr
Fr	0.03	0.91			0.06	
So		0.03	0.91		0.06	
Jr			0.03	0.93	0.04	
Se				0.03	0.04	0.93
d					1	
Gr						1

$Q = \begin{bmatrix} 0.03 & 0.91 & & \\ & 0.03 & 0.91 & \\ & & 0.03 & 0.93 \\ & & & 0.03 \end{bmatrix}, \quad R = \begin{bmatrix} 0.06 & \\ 0.06 & \\ 0.04 & \\ 0.04 & 0.93 \end{bmatrix}$

$FR = (I - Q)^{-1} R = \begin{array}{c} \\ \text{Fr} \\ \text{So} \\ \text{Jr} \\ \text{Se} \end{array} \begin{array}{c} \begin{array}{cc} \text{d} & \text{Gr} \end{array} \\ \begin{bmatrix} 0.19 & 0.81 \\ 0.14 & 0.86 \\ 0.08 & 0.92 \\ 0.04 & 0.96 \end{bmatrix} \end{array}$

absorption probabilities

$\tilde{p}_{ij} = p_{ij} h_j / h_i$

expected grad. time, assuming graduated: condition on graduating

$\tilde{Q} = \begin{bmatrix} \frac{0.03 \cdot 0.81}{0.81} & \frac{0.91 \cdot 0.86}{0.81} & & \\ & \frac{0.03 \cdot 0.86}{0.96} & \frac{0.91 \cdot 0.92}{0.96} & \\ & & \ddots & \ddots \\ & & & \end{bmatrix} \rightarrow (\tilde{F}1)_1 = 4.12$

WHAT ABOUT COUNTABLY INFINITE STATE SPACES?

example consider a random walk on $\mathbb{Z}^d$:

$$p_{ij} = \begin{cases} \frac{1}{2d}, & j = i \pm e_k \text{ for } k \in \{1, \dots, d\} \\ 0, & \text{o/w} \end{cases}$$

e_k: $(0\,0 \dots 1 \dots 0)$ in position k

for example, $d = 1$: $p_{ij} = \frac{1}{2}$, $j = i \pm 1$

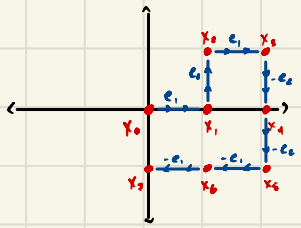
$d = 2$: $p_{ij} = \frac{1}{4}$, $j = i \pm (0,1), i \pm (1,0)$

questions:
1. is the SRW irreducible? [R script: simulation activity]
2. is it periodic?
3. is it recurrent?
4. does it have a limit?

lecture 8

SRW X_n can be represented as a convolution sum:

$$X_n = \sum_{i=0}^{n} \varepsilon_i \quad \text{where } \varepsilon_i \overset{iid}{\sim} \text{uniform}\{\pm e_1, \dots, \pm e_d\}$$



X_n is irreducible in any dimension but $d(j) = 2$. to remove periodicity, make it "lazy" by changing the innovation distribution

$$\varepsilon_i \overset{iid}{\sim} \text{uniform}\{0, \pm e_1, \dots, \pm e_d\} \quad \text{gives} \quad p_{ii} = \frac{1}{2d+1} > 0 \Rightarrow d(i) = 1$$

is it recurrent? let $N_0 = \sum_{n} \mathbb{1}\{X_n = 0\}$ be the number of visits to 0 and recall

$$\mathbb{E}(N_0 \mid X_0 = 0) = \sum_{n=0}^{\infty} P(X_n = 0 \mid X_0 = 0) = \sum_{n=0}^{\infty} p_{00}^{(n)}$$

now from before, X_n recurrent $\Leftrightarrow \mathbb{E}(N_0 \mid X_0 = 0) = \infty \Leftrightarrow \sum_{n=0}^{\infty} p_{00}^{(n)}$

one can show $p_{00}^{(2n)} \approx c\, n^{-d/2}$ for large n by a local CLT for lattice sums

Theorem let ε_i be iid on $\mathbb{Z}^d$, let $S_0 = 0$ & $S_t = \sum_{i=1}^{t} \varepsilon_i$, and let $X_t = x_0 + S_t$ so that $\{X_t\}$ is a random chain with $p_{xy}^{(n)} = P(S_n = y - x)$. denote

$$\mathbb{E}[\varepsilon] = \mu, \quad \text{var}[\varepsilon] = \Sigma, \quad \mathbb{E}[e^{i\theta\varepsilon}] = \varphi(\theta) \quad \text{for } \theta \in [-\pi, \pi]^d$$

then if $\mu = 0$, $\Sigma \succ 0$, and $|\varphi(\theta)| < 1 \ \ \forall \theta \neq 0$, one has

$$P(S_n = x) = (2\pi n)^{-d/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2n} x^T \Sigma^{-1} x\right\} + o(n^{-d/2})$$

see: Durrett; Lawler & Limic.

now at $x=0$, $\quad p_{00}^{(n)} = P(S_n = 0) = (2\pi n)^{-d/2} |\xi|^{-1/2} + o(n^{-d/2})$

so with $c := (2\pi)^{-d/2} |\xi|^{-1/2}$ one has $\quad \dfrac{P(S_n=0)}{c n^{-d/2}} = 1 + \dfrac{o(n^{-d/2})}{c n^{-d/2}} = 1 + o(1) \to 1$

and therefore $p_{00}^{(n)} \sim c n^{-d/2}$. applying the result yields

- for the SRW $\quad \mathbb{E}(N_0 | X_0 = 0) = \sum_{n=1}^{\infty} p_{00}^{(2n)} \approx c \sum_{n=1}^{\infty} n^{-d/2}$ ($p_{00}^{(2n+1)} = 0$)
- for the LSRW $\quad \mathbb{E}(N_0 | X_0 = 0) = \sum_{n=1}^{\infty} p_{00}^{(n)} \approx c \sum_{n=1}^{\infty} n^{-d/2}$

the series converges just in case $d > 2$ so X_t is recurrent for $d = 1, 2$ and transient otherwise.

what is the expected return time $\mathbb{E}(T_0 | X_0 = 0) = \mu_0$?

consider the successive return times $R_0 = 0$ and $R_{k+1} = \inf\{ n > R_k : X_n = 0 \}$

the by time homogeneity $R_{k+1} - R_k \overset{d}{=} T_0$ and are independent. (inter arrival times)

suppose $\mu_0 < \infty$. then we have

$$\mathbb{E}(R_{k+1} - R_k) = \mu_0 \implies \underbrace{\max_k \{ R_k \le n \}}_{\#\text{ visits by } n} \approx \frac{n}{\mu_0} \implies \underbrace{\frac{1}{n} N_0}_{\text{share of time at } 0} \to \pi_0 > 0$$

(π_0: stationary probability; $\frac{n}{\mu_0}$: linear rate)

in other words, there must be a positive rate of time spent at 0, but notice

$$\begin{aligned} \mathbb{E}\left(\tfrac{1}{n} N_0 \,|\, X_0 = 0\right) &= \tfrac{1}{n} \sum_{m=1}^{n} \mathbb{E}\left[\mathbb{1}\{X_m = 0\} \,|\, X_0 = 0 \right] \\ &= \tfrac{1}{n} \sum_{m=1}^{n} P(X_m = 0 \,|\, X_0 = 0) \\ &= \tfrac{1}{n} \sum_{m=1}^{n} p_{00}^{(m)} \\ &\to 0 \quad \text{since } p_{00}^{(m)} \approx c m^{-d/2} \quad \# \end{aligned}$$

so $\mu_0 = \infty$!

so X_n is recurrent for $d = 1, 2$ but expected return times are infinite

def state j is: (i) **positive recurrent** if $\mathbb{E}(T_j \mid X_0 = j) < \infty$

(ii) **null recurrent** if $\mathbb{E}(T_j \mid X_0 = j) = \infty$

(all states in a recurrent finite closed class are positive recurrent)

def a markov chain is **ergodic** if it is: i. irreducible

ii. aperiodic

iii. positive recurrent

theorem a strong limiting distribution exists (and is the unique stationary distribution) if X_n is ergodic. //

lecture 9

if $\{x_n\}_{n=0}^N$ is a sample path for a homogeneous markov chain $\{X_n\}$, the joint probability mass is:

$$P(X_0 = x_0, \dots, X_N = x_N) = \prod_{n=0}^{N-1} P(X_{n+1} = x_{n+1} \mid X_n = x_n)\, P(X_0 = x_0) = \underbrace{\alpha_{x_0}}_{\alpha(x_0)} \prod_{n=0}^{N-1} \underbrace{p_{x_{n+1}, x_n}}_{p(x_{n+1}, x_n)}$$

to avoid multilevel subscripts

so the log-likelihood is

$$\ell(p, \mu; x_0, \dots, x_N) = \log \alpha(x_0) + \sum_{n=0}^{N-1} \log\left[p(x_{n+1}, x_n)\right]$$

$$= \log \alpha(x_0) + \sum_i \sum_j N_{ij} \log p_{ij} \qquad N_{ij} = \#\{n : x_{n+1} = j,\ x_n = i\}$$

number of transitions

the standard approach to estimation is to treat α as unconstrained, which gives $\hat{\alpha} = \mathbb{1}\{X_0 = x_0\}$ and is thus equivalent to maximizing the conditional likelihood

$$\ell(p; x_1, \dots, x_N \mid x_0) = \sum_{i,j \in S} N_{ij} \log p_{ij} = \sum_i \underbrace{\left[\sum_j N_{ij} \log p_{ij}\right]}_{\text{multinomial likelihood}}$$

maximize the inner term in p_{ij} for fixed i subject to $\sum_j p_{ij} = 1$, i.e.,

$$\hat{p}_i = \operatorname*{argmin}_{p, \lambda} \underbrace{\left\{ -\sum_j N_{ij} \log p_j - \lambda\left[\sum_j p_j - 1\right] \right\}}_{\text{Lagrangian } \mathcal{L}(p, \lambda)}$$

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\frac{N_{ij}}{p_j} - \lambda = 0 \Rightarrow p_j = -\frac{N_{ij}}{\lambda}$$

$$1 = \sum_j p_j = -\frac{1}{\lambda} \sum_j N_{ij} \Rightarrow \lambda = -\underbrace{\sum_j N_{ij}}_{N_{i\cdot}}$$

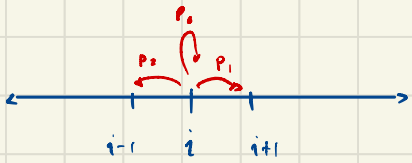
so $\hat{p}_{ij} = \dfrac{N_{ij}}{N_{i\cdot}}$ is the conditional MLE

this is still well-defined when S is countably infinite, but will have finite support since $N_{ij} = 0 \Rightarrow \hat{p}_{ij} = 0$

if the state space is countably infinite then one usually imposes some parametric structure on p_{ij}

example for the (lazy) random walk on $\mathbb{Z}$:

$$p_{ij} = \begin{cases} p_0 & , j = i \\ p_1 & , j = i+1 \\ p_2 & , j = i-1 \end{cases}$$



let $I_n = X_n - X_{n-1}$ denote the n^{th} increment. then:

$$P(X_0 = x_0, \dots, X_N = x_N) = \prod_{n=0}^{N-1} P(X_{n+1} = x_n \mid X_n = x_n)\, P(X_0 = x_0)$$

$$= \prod_{n=0}^{N-1} \left[p_0^{\mathbb{1}\{I_n = 0\}} \; p_1^{\mathbb{1}\{I_n = 1\}} \; p_2^{\mathbb{1}\{I_n = -1\}} \; \alpha_{x_0} \right]$$

$$\begin{cases} N_0 = \sum_i \mathbb{1}\{X_n = 0\} \\ N_1 = \sum_i \mathbb{1}\{X_n = 1\} \\ N_2 = \sum_i \mathbb{1}\{X_n = -1\} \end{cases}$$

$$\ell(p, q, r, \alpha; x_0, \dots, x_N) = \log \alpha(x_0) + N_0 \log p_0 + N_1 \log p_1 + N_2 \log p_2$$

$$= \log \alpha(x_0) + \underbrace{\sum_{j=} N_j \log p_j}_{\text{multinomial (again!)}}$$

$\Rightarrow$ conditional MLE is $\hat{p}_j = \frac{N_j}{N}$

the "drift" $\delta = \mathbb{E}\,\Delta_n$ (average increment) can be estimated using the plug-in estimator

$$\hat{\delta} = \hat{p}_1 - \hat{p}_2 \qquad \left(\text{equivalent to } \tfrac{1}{N}\textstyle\sum_n (x_{n+1} - x_n)\right)$$

the same estimation strategy generalizes to $\mathbb{Z}^d$, giving

$$\hat{p}_j = \frac{N_j}{N} \qquad \text{for } j = 0, 1, \dots, 2d$$

but even in the finite setting the MLE can break down:

- many states ($|S|$ is large)
- sparse transitions (zero sparsity in $\hat{P}$)
- rarely visited states ($\hat{p}_{ij} \approx 0$ or $N_{i\cdot}$ small)

so-called Laplace smoothing adds one to each count ($\tilde{N}_{ij} \leftarrow N_{ij} + 1$) giving

$$\hat{p}_{ij} = \frac{N_{ij} + 1}{N_{i\cdot} + k} \qquad \text{where } k := |S|$$

this ensures $\hat{p}_{ij} > 0$ (and thus irreducibility & aperiodicity):

$$\begin{cases} N_{ij} = 0 \Rightarrow \hat{p}_{ij} = \frac{1}{N_{i\cdot} + k} \\ N_{ij} = N_{i\cdot} = 0 \Rightarrow \hat{p}_{ij} = 1/k \end{cases}$$

so one always gets an ergodic chain with a strong limit & unique stationary distribution.

lecture 10

Laplace smoothing is equivalent to the posterior mean under a symmetric Dirichlet prior.

$$(p_{i1} \cdots p_{ik}) \sim \text{Dirichlet}(1, \ldots, 1) \iff f(p_i \mid \alpha_i) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} p_{ij}^{\alpha_{ij}-1} \quad \text{w/ } \alpha_{ij} = 1$$

$$B(\alpha) = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma[\sum_j \alpha_j]}$$

this gives the posterior

$$f(p \mid \{x_n\}) \overset{\text{proportional}}{\propto} \underbrace{\left[\prod_i \prod_j p_{ij}^{N_{ij}}\right]}_{p(\{x_n\} \mid p)} \underbrace{\left[\prod_i \prod_j p_{ij}^{\alpha_{ij}-1}\right]}_{f(p)} = \underbrace{\prod_i \prod_j p_{ij}^{N_{ij}+\alpha_{ij}-1}}_{\text{Dirichlet}(N_{i1}+\alpha_{i1}, \ldots, N_{ik}+\alpha_{ik}) \text{ kernel}}$$

fact: $p \sim \text{Dir}(\alpha_1, \ldots, \alpha_k) \Rightarrow \mathbb{E}\, p_i = \frac{\alpha_i}{\sum_j \alpha_j}$

so $\mathbb{E}_\alpha(p_i \mid \{x_n\}) = \frac{N_{ij}+\alpha_{ij}}{N_{i\cdot}+\alpha_{i\cdot}} \Rightarrow$

$$\begin{cases} \mathbb{E}_{\alpha=1}(p_i \mid \{x_n\}) = \frac{N_{ij}+1}{N_{i\cdot}+k} & \text{(Laplace smoothing)} \\ \mathbb{E}_{\alpha=1\alpha}(p_i \mid \{x_n\}) = \frac{N_{ij}+\alpha}{N_{i\cdot}+k\alpha} & \text{(symmetric prior)} \end{cases}$$

the symmetric prior smooths toward a uniform distribution, in the sense that

$$\hat{P}_{ij} = \frac{N_{ij}+\alpha}{N_{i\cdot}+\alpha k} \to \frac{1}{k} \quad \text{as} \quad \alpha \to \infty$$

consider instead shrinking toward $g = (g_1 \cdots g_k)$ via the prior

$$p_i \sim \text{Dirichlet}(\alpha g) \Rightarrow p_i \mid \{x_n\} \sim \text{Dir}(N_{i1}+\alpha g_1, \ldots, N_{ik}+\alpha g_k)$$

$$\Rightarrow \hat{p}_{ij} = \frac{N_{ij}+\alpha g_j}{N_{i\cdot}+\alpha} = \underbrace{\frac{N_{i\cdot}}{N_{i\cdot}+\alpha}}_{\text{weights for data}} \hat{p}_{ij}^{(mle)} + \underbrace{\frac{\alpha}{N_{i\cdot}+\alpha}}_{\text{prior depend on effective sample size}} g_j$$

an empirical Bayes estimate of g is:

$$\hat{g}_j = \frac{N_{\cdot j}}{\sum_i N_{ij}} \quad \text{(marginal destination probabilities)}$$

now $N_{i\cdot} \ll \alpha \Rightarrow \hat{p}_{ij} \approx \hat{q}_j$ to $\alpha \vec{q}$ shrinks $\hat{p}_{ij}$ toward $\hat{q}_j$.

$N_{i\cdot} \gg \alpha \Rightarrow \hat{p}_{ij} \approx \hat{p}_{ij}^{(mle)}$

fact $(p_1 \cdots p_k) \sim \text{Dir}(\alpha) \Rightarrow p_j \sim \text{beta}(\alpha_j, \sum_{i \neq j} \alpha_i)$

pointwise credible intervals for p_{ij} are quantiles from:

$$\text{beta}\left[N_{ij} + \alpha_j, N_{i\cdot} + \sum_j \alpha_j - (N_{ij} + \alpha_j)\right]$$

↑ 1, laplace
α, symmetric prior
$\alpha \hat{q}_j$, empirical bayes

a joint HPD region can be approximated by sampling from the posterior to estimate the level set

$$R = \{p : f(p \mid s, x) \geq c\} \quad \text{s.t.} \quad P(R) \geq 1 - \alpha$$

two options for intervals based on MLE:

(i) marginally $N_{ij} \sim b(N_{i\cdot}, p_{ij}) \rightarrow$ construct interval from binomial tails

$(\hat{p}_L, \hat{p}_u)$ with $\hat{p}_L = \inf\{p : 1 - F(N_{ij} - 1) \geq \alpha/2\}$
$\hat{p}_u = \sup\{p : F(N_{ij}) \geq \alpha/2\}$

(ii) asymptotically $\sqrt{N_{i\cdot}}(\hat{p}_i - p_i) \rightarrow N(0, \Sigma)$ where

$$\Sigma = \text{diag}(p_i) - p_i p_i^T, \text{ ie., } \begin{cases} \sigma_{jj} = p_{ij}(1 - p_{ij}) \\ \sigma_{jk} = -p_{ij} p_{ik} \end{cases}$$

so $\sqrt{N_{i\cdot}}(\hat{p}_{ij} - p_{ij}) \xrightarrow{d} N(0, p_{ij}(1 - p_{ij}))$ marginally

which gives the Wald interval

$$\hat{p}_{ij} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{N_{i\cdot}}}$$