4. Markov Chains, Part (i)

- A discrete time process $\{X_n, n = 0, 1, 2, \ldots\}$ with discrete **state space** $X_n \in \{0, 1, 2, \ldots\}$ is a **Markov chain** if it has the **Markov property**:

$$\boxed{\begin{aligned} &\mathbb{P}[X_{n+1}{=}j \,|\, X_n{=}i, X_{n-1}{=}i_{n-1}, \ldots, X_0{=}i_0] \\ &\qquad = \mathbb{P}[X_{n+1}{=}j \,|\, X_n{=}i] \end{aligned}}$$

- In words, "the past is conditionally independent of the future given the present state of the process" or "given the present state, the past contains no additional information on the future evolution of the system."

- The Markov property is common in probability models because, by assumption, one supposes that the important variables for the system being modeled are all included in the state space.

- We consider **homogeneous** Markov chains for which $\mathbb{P}[X_{n+1}{=}j \,|\, X_n{=}i] = \mathbb{P}[X_1{=}j \,|\, X_0{=}j]$.

Example: **physical systems**. If the state space contains the masses, velocities and accelerations of particles subject to Newton's laws of mechanics, the system in Markovian (but not random!)

Example: **speech recognition**. Context can be important for identifying words. Context can be modeled as a probability distrubtion for the next word given the most recent k words. This can be written as a Markov chain whose state is a vector of k consecutive words.

Example: **epidemics**. Suppose each infected individual has some chance of contacting each susceptible individual in each time interval, before becoming removed (recovered or hospitalized). Then, the number of infected and susceptible individuals may be modeled as a Markov chain.

- Define $\boxed{P_{ij} = \mathbb{P}\big[X_{n+1}{=}j \,|\, X_n{=}i\big]}$.

  Let $P = [P_{ij}]$ denote the (possibly infinite) **transition matrix** of the **one-step transition probabilities**.

- Write $P^2_{ij} = \sum_{k=0}^{\infty} P_{ik}P_{kj}$, corresponding to standard matrix multiplication. Then

$$
\begin{aligned}
P^2_{ij} &= \sum_k \mathbb{P}\big[X_{n+1}{=}k \,|\, X_n{=}i\big]\, \mathbb{P}\big[X_{n+2}{=}j \,|\, X_{n+1}{=}k\big] \\
&= \sum_k \mathbb{P}\big[X_{n+2}{=}j, X_{n+1}{=}k \,|\, X_n{=}i\big] \\
&\qquad\qquad \text{(via the Markov property. Why?)} \\
&= \mathbb{P}\Big[\bigcup_k \{X_{n+2}{=}j, X_{n+1}{=}k\} \Big| X_n{=}i\Big] \\
&= \mathbb{P}\big[X_{n+2}{=}j \,|\, X_n{=}i\big]
\end{aligned}
$$

- Generalizing this calculation:

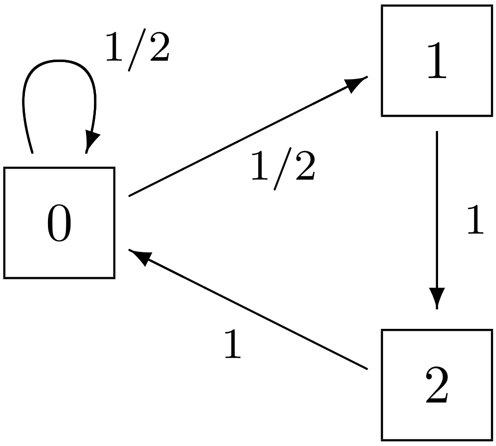
  The matrix power P^n_{ij} gives the n-step transition probabilities.

- The matrix multiplication identity
  $P^{n+m} = P^n P^m$

  corresponds to the **Chapman-Kolmogorov equation**

  $P^{n+m}_{ij} = \sum_{k=0}^{\infty} P^n_{ik} P^m_{kj}$.
- Let $\nu^{(n)}$ be the (possibly infinite) row vector of probabilities at time n, so $\nu^{(n)}_i = \mathbb{P}\big[X_n = i\big]$. Then $\nu^{(n)} = \nu^{(0)} P^n$, using standard multiplication of a vector and a matrix. Why?

Example. Set $X_0 = 0$, and let X_n evolves as a Markov chain with transition matrix



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

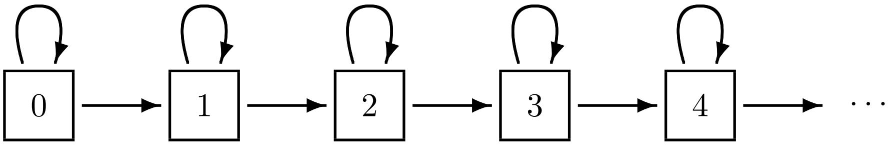
Find $\nu_0^{(n)} = \mathbb{P}[X_n = 0]$ by

(i) using a probabilistic argument

(ii) using linear algebra.

Classification of States

- State j is **accessible** from i if $p_{ij}^k > 0$ for some $k \geq 0$.
- The transition matrix can be represented as a **directed graph** with arrows corresponding to positive one-step transition probabilities j is accessible from i if there is a path from i to j. For example,



Here, 4 is accessible from 0, but not vice versa.

- i and j **communicate** if they are accessible from each other. This is written $i \leftrightarrow j$, and is an **equivalence relation**, meaning that
(i) $i \leftrightarrow i$ [reflexivity]
(ii) If $i \leftrightarrow j$ then $j \leftrightarrow i$ [symmetry]
(iii) If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$ [transitivity]

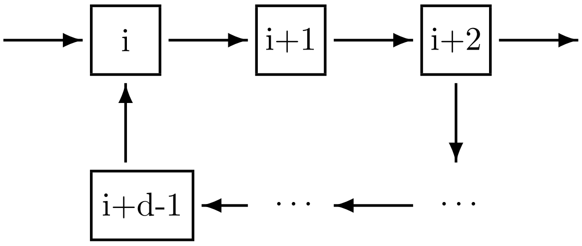
- An equivalence relation divides a set (here, the state space) into disjoint classes of equivalent states (here, called **communication classes**).
- A Markov chain is **irreducible** if all the states communicate with each other, i.e., if there is only one communication class.
- The communication class containing i is **absorbing** if $P_{jk} = 0$ whenever $i \leftrightarrow j$ but $i \not\leftrightarrow k$ (i.e., when i communicates with j but not with k). An absorbing class can never be left. A partial converse is $\ldots$

Example: Show that a communication class, once left, can never be re-entered.

- State i has **period** d if $P_{ii}^n = 0$ when n is not a multiple of d and if d is the greatest integer with this property. If $d = 1$ then i is **aperiodic**.

Example: Show that all states in the same communication class have the same period.

Note: This is "obvious" by considering a generic directed graph for a periodic state:



- State i is **recurrent** if $\mathbb{P}\big[\text{re-enter } i \,|\, X_0{=}i\big] = 1$, where $\{\text{re-enter } i\}$ is the event $\bigcup_{n=1}^{\infty} \{X_n = i, X_k \neq i \text{ for } k = 1, \ldots, n-1\}$.
If i is not recurrent, it is **transient**.

- Let $S_0 = 0$ and $S_1, S_2, \ldots$ be times of successive returns to i, with $N_i(t) = \max\{n : S_n \leq t\}$ being the corresponding counting process.

- If i is recurrent, then $N_i(t)$ is a renewal process, since the Markov property gives independence of interarrival times $X_n^A = S_n - S_{n-1}$.
Letting $\mu_{ii} = E[X_1^n]$, the **expected return time** for i, we then have the following from renewal theory:
  - ◇ $\mathbb{P}\big[\lim_{t\to\infty} N_i(t)/t = 1/\mu_{ii} \,|\, X_0{=}i\big] = 1$
  - ◇ If $i \leftrightarrow j$, $\lim_{n\to\infty} \sum_{k=1}^{n} P_{ij}^{k}/n = 1/\mu_{jj}$
  - ◇ If i is aperiodic, $\lim_{n\to\infty} P_{ij}^{n} = 1/\mu_{jj}$ for $j \leftrightarrow i$
  - ◇ If i has period d, $\lim_{n\to\infty} P_{ii}^{nd} = d/\mu_{ii}$

- If i is transient, then $N_i(t)$ is a **defective renewal process**. This is a generalization of renewal processes where $X_1^A, X_2^A, \ldots$ are still iid, abut we allow $\mathbb{P}[X_1^A = \infty] > 0$.

Proposition: i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^n = \infty$.

Example: Show that if i is recurrent and $i \leftrightarrow j$ then j is recurrent.

Example: If i is recurrent and $i \leftrightarrow j$, show that $\mathbb{P}\big[\text{never enter state } j \,|\, X_0 = i\big] = 0$.

- If i is transient, then $\mu_{ii} = \infty$. If i is recurrent and $\mu_{ii} < \infty$ then i is said to be **positive recurrent**. Otherwise, if $\mu_{ii} = \infty$, i is **null recurrent**.

Proposition: If $i \leftrightarrow j$ and i is recurrent, then either i and j are both positive recurrent, or both null recurrent (i.e., positive/null recurrence is a property of communication classes).

Random Walks

- The **simple random walk** is a Markov chain on the integers, $\mathbb{Z} = \{\ldots, -1, 0, 1, 2, \ldots\}$ with $X_0 = 0$ and $\mathbb{P}\big[X_{n+1}{=}X_n + 1\big] = p$, $\mathbb{P}\big[X_{n+1}{=}X_n - 1\big] = 1 - p$.

Example: If X_n counts the number of successes minus the number of failures for a new medical procedure, X_n could be modeled as a random walk, with p the success rate of the procedure. When should the trial be stopped?

- If $p = 1/2$, the random walk is **symmetric**.

- The symmetric random in d dimensions is a vector valued Markov chain, with state space $\mathbb{Z}^d$, $X_0^{(d)} = (0, \ldots, 0)$. Two possible definitions are
  (i) Let $X_{n+1}^{(d)} - X_n^{(d)}$ take each of the 2^d possibilities $(\pm 1, \pm 1, \ldots, \pm 1)$ with equal probability.
  (ii) Let $X_{n+1}^{(d)} - X_n^{(d)}$ take one of the $2d$ values $(\pm 1, 0, \ldots, 0)$, $(0, \pm 1, 0, \ldots, 0), \ldots$ with equal probability. This is harder to analyze.

Example: A biological signaling molecule becomes separated from its receptor. It starts diffusing, due to themal noise. Suppose the diffusion is well modeled by a random walk. Will the molecule return to the receptor? If the molecule is constrained to a one-dimensional line? A two-dimensional surface? Three-dimensional space?

Proposition: The symmetric random walk is null recurrent when $d = 1$ and $d = 2$, but transient for $d \geq 3$.

Proof: The method is to employ stirling's formula, $n! \sim n^{n+1/2}e^{-n}\sqrt{2\pi}$ where $a_n \sim b_n$ means $\lim_{n\to\infty}(a_n/b_n) = 1$, to approximate $\mathbb{P}\big[X_n^{(d)}\!=X_0^{(d)}\big]$.

Note: This is in HW 5. You are expected to solve the simpler case (i), though you can solve (ii) if you want a bigger challenge.

Stationary Distributions

- $\pi = \{\pi_i, i = 0, 1, \ldots\}$ is a **stationary distribution** for $P = [P_{ij}]$ if $\boxed{\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}}$ with $\pi_i \geq 0$ and $\sum_{i=0}^{\infty} \pi_i = 1$.
- In matrix notation, $\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}$ is $\pi = \pi P$ where π is a row vector.

Theorem: An irreducible, aperiodic, positive recurrent Markov chain has a unique stationary distribution, which is also the **limiting distribution**

$\boxed{\pi_j = \lim_{n\to\infty} P_{ij}^n}$.

- Such Markov chains are called **ergodic**.

Proof

Proof continued

- Irreducible chains which are transient or null recurrent have no stationary distribution. Why?

- Chains which are periodic or which have multiple communicating classes may have $\lim_{n\to\infty} P_{ij}^n$ not existing, or depending on i.

- A chain started in a stationary distribution will remain in that distribution, i.e., will result in a **stationary** process.

- If we can find any probability distribution solving the stationarity equations $\pi = \pi P$ and we check that the chain is irreducible and aperiodic, then we know that
  (i) The chain is positive recurrent.
  (ii) π is the unique stationary distribution.
  (iii) π is the limiting distribution.

Example: **Monte Carlo Markov Chain**

- Suppose we wish to evaluate $\mathbb{E}\big[h(X)\big]$ where X has distribution π (i.e., $\mathbb{P}\big[X{=}i\big] = \pi_i$). The **Monte Carlo** approach is to generate $X_1, X_2, \ldots X_n \sim \pi$ and estimate $\mathbb{E}\big[h(X)\big] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)$

- If it is hard to generate an iid sample from π, we may look to generate a sequence from a Markov chain with limiting distribution π.

- This idea, called **Monte Carlo Markov Chain (MCMC)**, was introduced by Metropolis and Hastings (1953). It has become a fundamental computational method for the physical and biological sciences. It is also commonly used for Bayesian statistical inference.

Metropolis-Hastings Algorithm

(i) Choose a transition matrix $Q = [q_{ij}]$

(ii) Set $X_0 = 0$

(iii) for $n = 1, 2, \ldots$

◇ generate Y_n with $\mathbb{P}\big[Y_n{=}j \,|\, X_{n-1}{=}i\big] = q_{ij}$.

◇ If $X_{n-1} = i$ and $Y_n = j$, set

$$X_n = \begin{cases} j & \text{with probability } \min(1, \pi_j q_{ji}/\pi_i q_{ij}) \\ i & \text{otherwise} \end{cases}$$

- Here, Y_n is called the **proposal** and we say the proposal is **accepted** with probability $\min(1, \pi_j q_{ji}/\pi_i q_{ij})$. If the proposal is not accepted, the chain stays in its previous state.

Proposition: Set

$$P_{ij} = \begin{cases} q_{ij} \min(1, \pi_j q_{ji}/\pi_i q_{ij}) & j \neq i \\ q_{ii} + \sum_{k \neq i} q_{ik} \{1 - \min(1, \pi_k q_{ki}/\pi_i q_{ik})\} & j = i \end{cases}$$

Then π is a stationary distribution of the Metropolis-Hastings chain $\{X_n\}$. If P_{ij} is irreducible and a periodic, then π is also the limiting distribution.

Proof.