DCP 1206: Probability
Lecture 27 — Markov Chain

Ping-Chun Hsieh

December 25, 2019

A Few Words About MAP

- **Maximum a Posteriori Estimation (MAP)**: Given observed data D and prior distribution $P(\theta)$, choose θ that maximizes the posterior probability:

Likelihood prior

$$\theta_{\text{MAP}} = \arg\max_{\theta \in \Theta} \underbrace{P(\theta \mid D)}_{} = \arg\max_{\theta \in \Theta} \log \underbrace{P(D \mid \theta)}_{} \underbrace{P(\theta)}_{}$$

true mean parameter

1. Difference between SLLN/CLT and MLE/MAP in terms of estimation?

2. Both MLE and MAP are generic tools for learning unknown parameters (not necessarily the "mean")

choice of prior

3. MLE and MAP are different views of parameter estimation

4. Bernstein–von Mises theorem: MAP will converge to MLE (under some technical conditions about the prior)

A Few Words About HW6

estimate (Training data) $\prod_{i=1}^{n} p(x_i \mid spam)$

$$\theta_{\text{MAP}} = \arg\max_{\theta\in\Theta} P(\theta \mid D) = \arg\max_{\theta\in\Theta} \log P(D \mid \theta)P(\theta)$$

Text Message
May 12, 2017, 4:15 AM

Custom Essay & Research Writing Services. Let us take care of your writing needs. 100% Secure & Money-back guarantee.
Visit:
www.essaysondemand.com

$$p(\text{spam} \mid x_1, x_2, \cdots, x_n) \propto \underbrace{p(\text{spam})}_{\text{prior}} \cdot \underbrace{p(x_1, x_2, \cdots, x_n \mid \text{spam})}_{\text{likelihood}}$$

$$p(\text{ham} \mid x_1, x_2, \cdots, x_n) \propto p(\text{ham}) \cdot p(x_1, x_2, \cdots, x_n \mid \text{ham})$$

$x_1\ x_2\ x_3 \cdots x_n$ → spam $=0$, ham $=1$ $\quad \Theta = \{0, 1\}$

$P(\theta=0 \mid D=\{x_1, \cdots, x_n\})$

$P(\theta=1 \mid D=\{x_1, \cdots, x_n\})$

posterior distribution

Conditional independence:

$$p(x_1, x_2, \cdots, x_n \mid \text{label}) = \prod_{i=1}^{n} p(x_i \mid \text{label})$$

$x_1, x_2, \cdots, x_n$ are i.i.d. given the label ("spam" or "ham")

Multinomial NB

This Lecture

1. Stochastic Process

2. Markov Chain

- Reading material: Chapter 12.1-12.4

1. Stochastic Process

Randomness Beyond "i.i.d."?

- We have been studying i.i.d. settings for a while
- Question: Is this general enough?
- Example: Non-i.i.d. time series

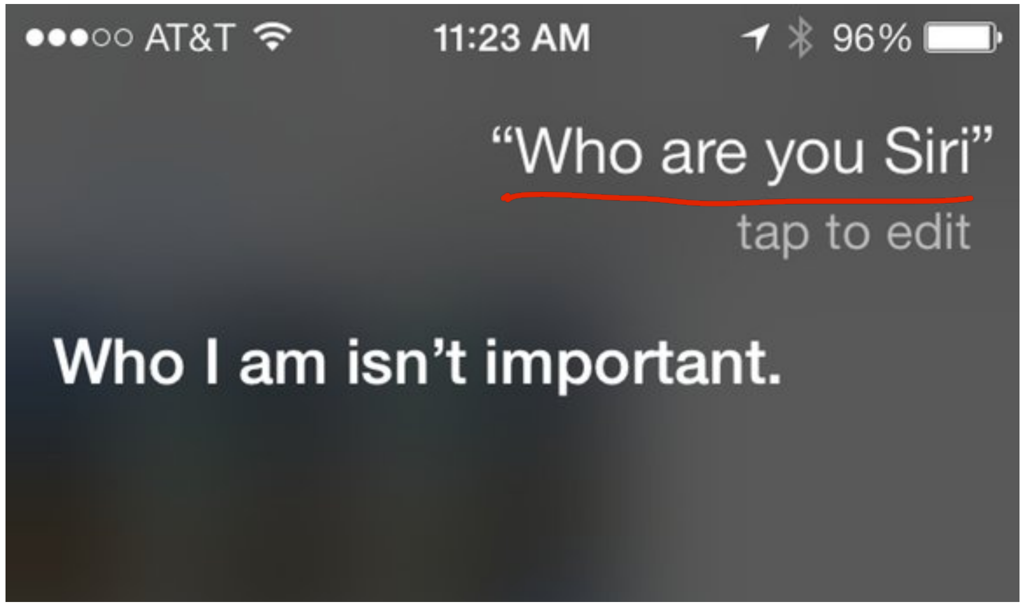
Who are

Stock price

Is the price i.i.d.?

are

Speech sequences



Are the words in a speech sequence i.i.d.?

Stochastic Process (Formally)

$I = \{0, 1, 2, 3, \cdots, 100, \cdots\}$

$X_0, X_1, X_2, \cdots,$

- **Question**: A general way to describe such time series? index set

- **Stochastic Process (also called random process)**: A stochastic process is an indexed collection of random variables $\{X_t, t \in I\}$, for an index set I (can be finite, countable, or uncountable). Moreover, the set of all possible values of $\{X_t\}$ is called the state space.

  random

  $X_t \in \{-1, 1\}$

- **Remark**: We focus on discrete state space in this class

- **Remark**: Two main categories of stochastic processes
  - Discrete-time: the index set I is discrete
  - Continuous-time: " " is continuous

Example: Stochastic Process

- Example: $X_1, X_2, \cdots$, are i.i.d. Bernoulli random variables
  - Question: Is $\{X_n\}$ a stochastic process? Yes
  - Question: What's the index set? Time is discrete or continuous?
  - Question: What's the state space?

index set $I = \{1, 2, \cdots \} \Rightarrow$ discrete-time

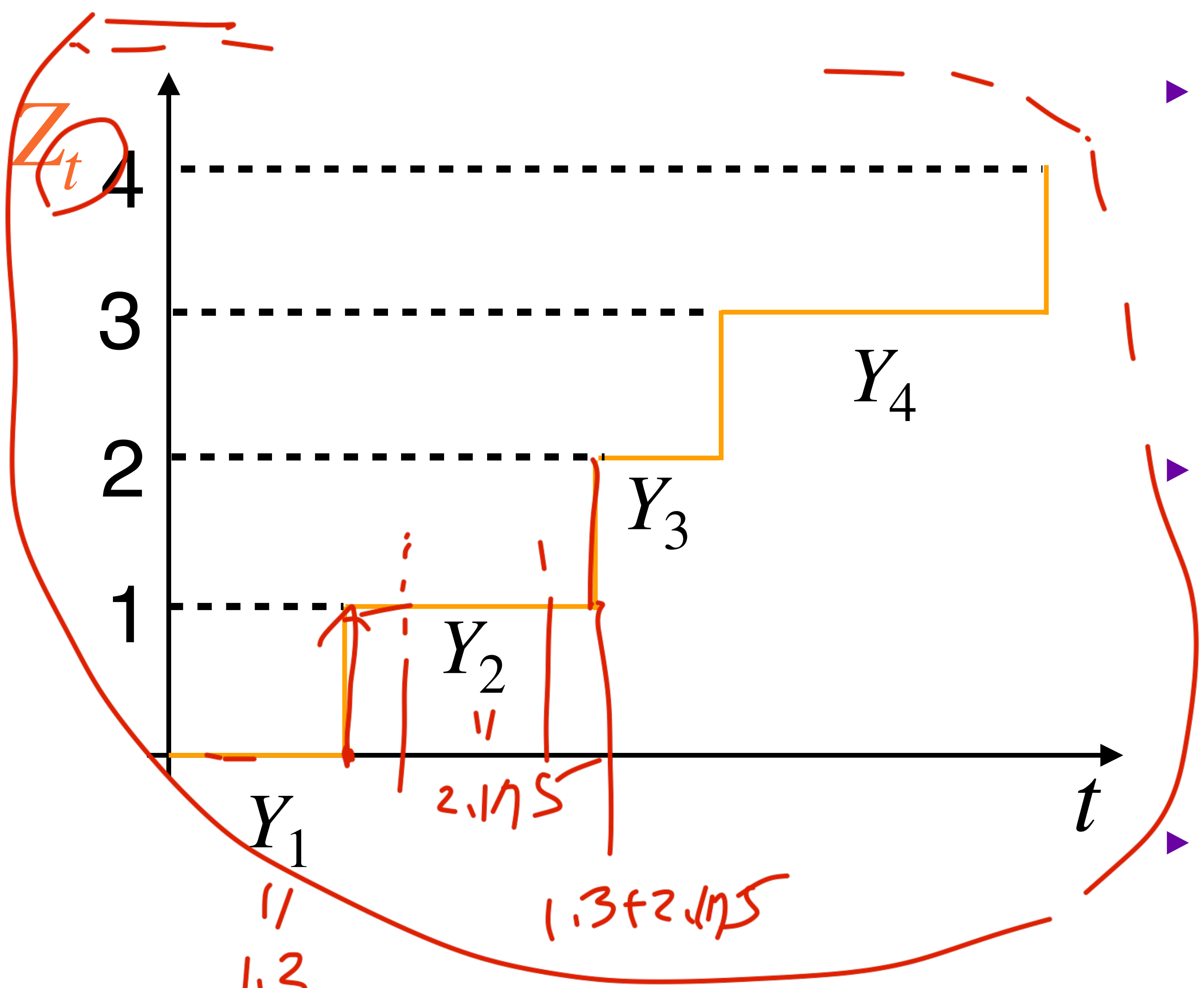
State space $S = \{0, 1\}$

(Bernoulli process)

Example: Stochastic Process (Cont.)

(Counting Process)
(Poisson Process)

- Example: $Y_1, Y_2, \cdots,$ are i.i.d. exponential random variables and Z_t is constructed based on Y_n as follows:



- Question: Is $\{Z_t\}$ a stochastic process? What is the index set?

  Yes $I = [0, \infty) = \mathbb{R}_+$

- Question: What is the index set? Time is discrete or continuous?

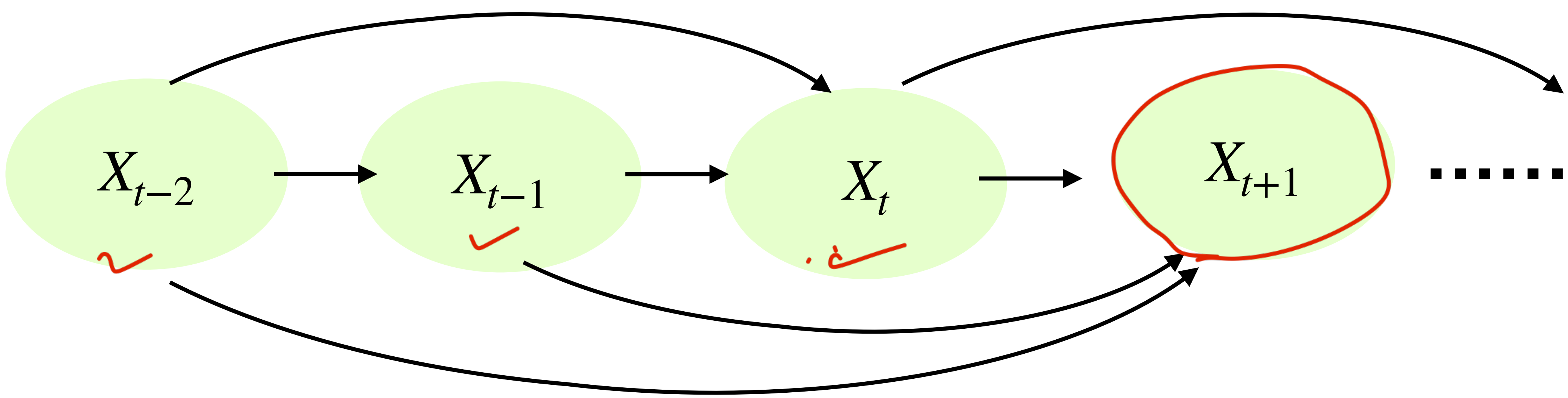
  continuous-time

- Question: What's the state space?

  State space $= \{0, 1, 2, \cdots\}$

History-Dependent Stochastic Processes

- (Discrete-time) history-dependent stochastic process:



- X_{t+1} is dependent on the past history $\{X_1, \cdots, X_{t-1}\}$
- Such dependency can be described using conditional distribution

$$\underbrace{P(X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \cdots, X_1)}$$

- Question: Any issue with this? Difficult (Curse of dimensionality)

History Fully Captured by the Present?

- Issue: Usually intractable to study general history-dependent stochastic processes
- Question: Any realistic and tractable assumption?

> Assume “history is fully captured by the present”

- Example: Monopoly

X_t = the location after t throws

X_t stochastic process

X_{10}

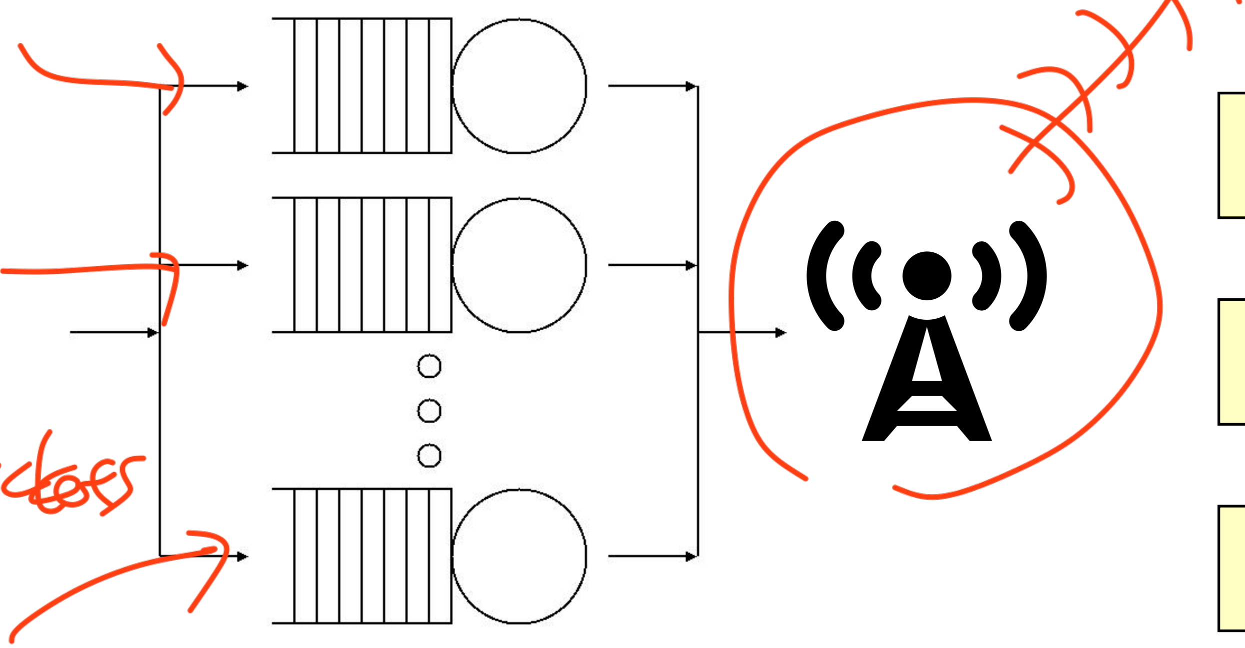
Other Examples: History Captured by the Present

- Example:

transmitting

Queues at a base station

N-gram language model



packets

Winter	break	is	coming	soon
Winter	break	is	coming	soon
Winter	break	is	coming	soon

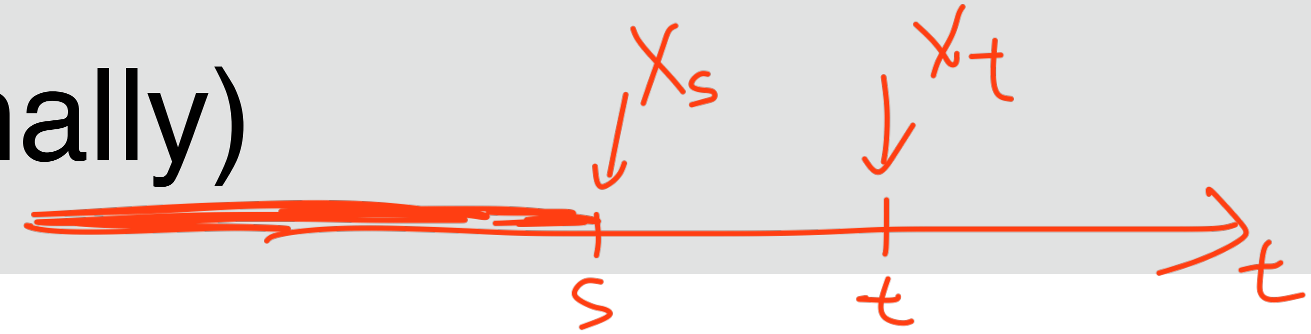
X_1 X_2 X_3 X_4 X_5

X_t = # of packets in the queue

(don't care about $X_{t-1}, X_{t-2}, \cdots$)

$$\underbrace{P(X_5 \mid X_4)}_{\text{bi-gram}}$$

Markov Property (Formally)



- **Markov Property**: A stochastic process $\{X_t\}$ with index set I and state space S is said to satisfy Markov property if the conditional distribution of future states depends only on the present state. That is, for any $s, t \in I$ with $s < t$, for any set $A \subseteq S$,

$$P(X_t \in A \mid \{X_\tau, \tau \leq s\}) = P(X_t \in A \mid X_s)$$

- Interpretation: Given the present state (X_s), future state (X_t) is independent of the past states ($\{X_\tau, \tau \leq s\}$)

- Remark: For discrete-time cases, this is called Markov chain

2. Markov Chain

Markov Chain (Formally)

discrete index set

- **Markov Chain**: A stochastic process $\{X_t : t = 0,1,2,\cdots\}$ with a state space S (finite or countable infinite) is said to be a Markov chain, if for all $i, j, i_0, \cdots, i_{t-1} \in S$ and for all $t = 0,1,\cdots$,

$$\checkmark \; P(X_{t+1} = j \mid X_t = i, X_{t-1} = i_{t-1}, \cdots, X_0 = i_0)$$

$$= P(X_{t+1} = j \mid X_t = i) \longleftarrow \text{one-step transition probability}$$

9·t

- **Question**: To specify a Markov chain up to time t, how many one-step transition probabilities do we need?

$S=\{0,1,2\}$.

$P(X_{t+1}=0 \mid X_t=0)$ $P(X_{t+1}=0 \mid X_t=1)$ $P(\,'' \mid X_t=2)$

$P(X_{t+1}=1 \mid X_t=0)$ $P(X_{t+1}=1 \mid X_t=1)$ $P(\,'' \mid X_t=2)$

$P(X_{t+1}=2 \mid X_t=0)$ $P(X_{t+1}=2 \mid X_t=1)$ $P(\,'' \mid X_t=2)$

Stationary Markov Chain (Formally)

- **Stationary Markov Chain**: A stochastic process $\{X_t : t = 0,1,2,\cdots\}$ with a state space S (finite or countable infinite) is said to be a stationary Markov chain, if for all $i, j, i_0, \cdots, i_{t-1} \in S$ and for all $t = 0,1,\cdots$,

$$P(X_{t+1} = j \,|\, X_t = i, X_{t-1} = i_{t-1}, \cdots, X_0 = i_0)$$

$$= P(X_{t+1} = j \,|\, X_t = i) = p_{i,j}$$

One-step transition probability is independent of t

state j state i

- Question: What are the components needed to fully specify a stationary Markov chain?

$S = \{0,1,2\}$ 3×3

For simplicity, we assume finite state space $S = \{0,1,2,\cdots,M\}$ for the rest of the slides

Transition Probability Matrix

- Stationary Markov chain:

$$P(X_{t+1} = j \mid X_t = i, X_{t-1} = i_{t-1}, \cdots, X_0 = i_0)$$
$$= P(\underline{X_{t+1} = j} \mid \underline{X_t = i}) = p_{i,j}$$

One-step transition probability is independent of t

- Transition probability matrix: a more compact way to specify a stationary Markov chain

Suppose state space $S = \{0, 1, \cdots, M\}$

$$P = \begin{array}{cc} (M+1) & \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0M} \\ \vdots & & \ddots & \vdots \\ p_{M0} & p_{M1} & \cdots & p_{MM} \end{bmatrix} \\ & (M+1) \end{array}$$

transition matrix

Applications of Markov Chain

- Reinforcement learning (Markov decision process)
- Sampling methods (e.g. Markov chain Monte Carlo)
- Queueing theory
- Natural language processing (e.g. n-gram)

Some Historical Accounts of Markov Chain

- Markov chain was proposed by Andrey Markov

- Motivation: Can we have SLLN for a sequence of non-i.i.d. random variables?

Example: Gambler's Ruin

- Example: Bill is playing with a bandit machine
  - Initially he has \$3
  - In each round, he wins \$1 with probability 0.4 and loses \$1 with probability 0.6
  - He stops playing if he has \$0 or \$5
- Question: State space? Transition matrix?

X_t = # of dollars Bill has at time t

$X_0 = 3 \longrightarrow X_1 = 2$ or 4
(0.6) (0.4)

$(k=1,2,3,4)$

$X_t = k$ → $X_{t+1} = k+1$, w.p. 0.4
→ $X_{t+1} = k-1$, w.p. 0.6

$X_t = 5 \longrightarrow X_{t+1} = 5$

$X_t = 0 \longrightarrow X_{t+1} = 0$

$P = [\quad]$

Markov Chain State Diagram

- Note: A Markov chain can be represented in a state diagram

- Example: Bill is playing with a bandit machine
  - Initially he has \$3
  - In each round, he wins \$1 with probability 0.4 and loses \$1 with probability 0.6
  - He stops playing if he has \$0 or \$5

Example: TCP Congestion Control

- Example: TCP enforces congestion control as follows
  - Available rates = $\{1, 2, 3, \cdots, 32\}$ Mbps
  - If congestion happens, decrease rate by half (with floor function)
  - Otherwise, increase the rate by 1 Mbps
  - When rate = k Mbps, P(congestion occurs) = $k/100$
- Question: What is the Markov chain of interest? State space? Transition matrix?

Find the State Distribution

- **Property**: Let S and $P = [p_{i,j}]_{i,j \in S}$ be the state space and transition matrix of the Markov chain, respectively. Let $\pi_t(i)$ be the probability that the chain is in state $i \in S$ at time t. For any $t = 0, 1, \cdots$ and $j \in S$, we have

$$\pi_{t+1}(j) = \sum_{i \in S} \pi_t(i) p_{i,j}$$

- Remark: Consider row vector $\pi_t = (\pi_t(0), \pi_t(1), \cdots, \pi_t(M))$

$$\pi_{t+1} =$$

Example: Gambler's Ruin (Cont.)

- Example: Bill is playing with a bandit machine
  - In each round, he wins \$1 with probability 0.4 and loses \$1 with probability 0.6
  - He stops playing if he has \$0 or \$5

- Question: Suppose at time 0, $P(\text{Bill has }\$1) = P(\text{Bill has }\$3) = 0.5$. What's the distribution of the chain at time 1?

n-Step Transition Probabilities

- Consider a stationary Markov chain:
  - The n-step transition probabilities can be written as

$$P(X_{t+n} = j \mid X_t = i) = P(X_n = j \mid X_0 = i) = p_{ij}^{(n)}$$

- Question: How to interpret this?

n-Step Transition Probabilities (Cont.)

- Consider a stationary Markov chain:
  - The n-step transition probabilities can be written as

$$P(X_{t+n} = j \,|\, X_t = i) = P(X_n = j \,|\, X_0 = i) = p_{ij}^{(n)}$$

- Question: How to connect transition matrix and $p_{i,j}^{(n)}$?

$$P^{(n)} :=$$

Example: Gambler's Ruin (Cont.)

- Example: Bill is playing with a bandit machine
  - In each round, he wins \$1 with probability 0.4 and loses \$1 with probability 0.6
  - He stops playing if he has \$0 or \$5

- Question: How to find $p_{12}^{(3)}$?

Chapman-Kolmogorov Equation

- **Chapman-Kolmogorov Equation**: Let $p_{i,j}^{(n)}$ be the n-step transition probability from state i to state j. For any $n = 0, 1, \cdots$ and for any $i, j \in S = \{0, 1, \cdots, M\}$

$$p_{i,j}^{(n)} = \sum_{k=0}^{M} p_{i,k}^{(v)} p_{k,j}^{(n-v)} \quad \text{for any } 0 \le v \le n$$

- Question: How to interpret this equation?

What happen if $n \to \infty$?

Does $\lim_{n\to\infty} p_{i,j}^{(n)}$ exist?

To study limiting distributions, we need to consider classification of states

Concept #1: Accessibility

- Accessible: State j is <u>accessible</u> from state i if $p_{i,j}^{(n)} > 0$, for some $n \geq 0$ (denoted by $j \leftarrow i$)
- Example: Consider a Markov chain with transition matrix as

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0 & 0 & 0.8 & 0.2 \end{bmatrix}$$

- Question: Which states are accessible from which other states?

Concept #2:Communicability and State Classes

- Communicate: State i and state j communicate if $j \leftarrow i$ and $i \leftarrow j$ (denoted by $i \leftrightarrow j$)
- Class: Two states are said to be in the same class if the two states communicate with each other

- Example: Consider a Markov chain with transition matrix as

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0 & 0 & 0.8 & 0.2 \end{bmatrix}$$

- Question: Which states communicate with each other state? How many classes are there?

Concept #3: Irreducibility

- Irreducible: A Markov chain is irreducible if all states belong to one class

- Example: Consider a Markov chain with transition matrix as

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0 & 0 & 0.8 & 0.2 \end{bmatrix}$$

- Question: Is this Markov chain irreducible?

Next Lecture

- Markov chain
  - Classification of states
  - Steady-state distribution

1-Minute Summary

1. Stochastic Process

- General stochastic process / history-dependent
- Markov property

2. Markov Chain

- State space / transition matrix
- n-step transition probabilities
- Chapman-Kolmogorov equation
- Accessibility / Communicate / Irreducibility