

# Deep Generative Models: Markov Models

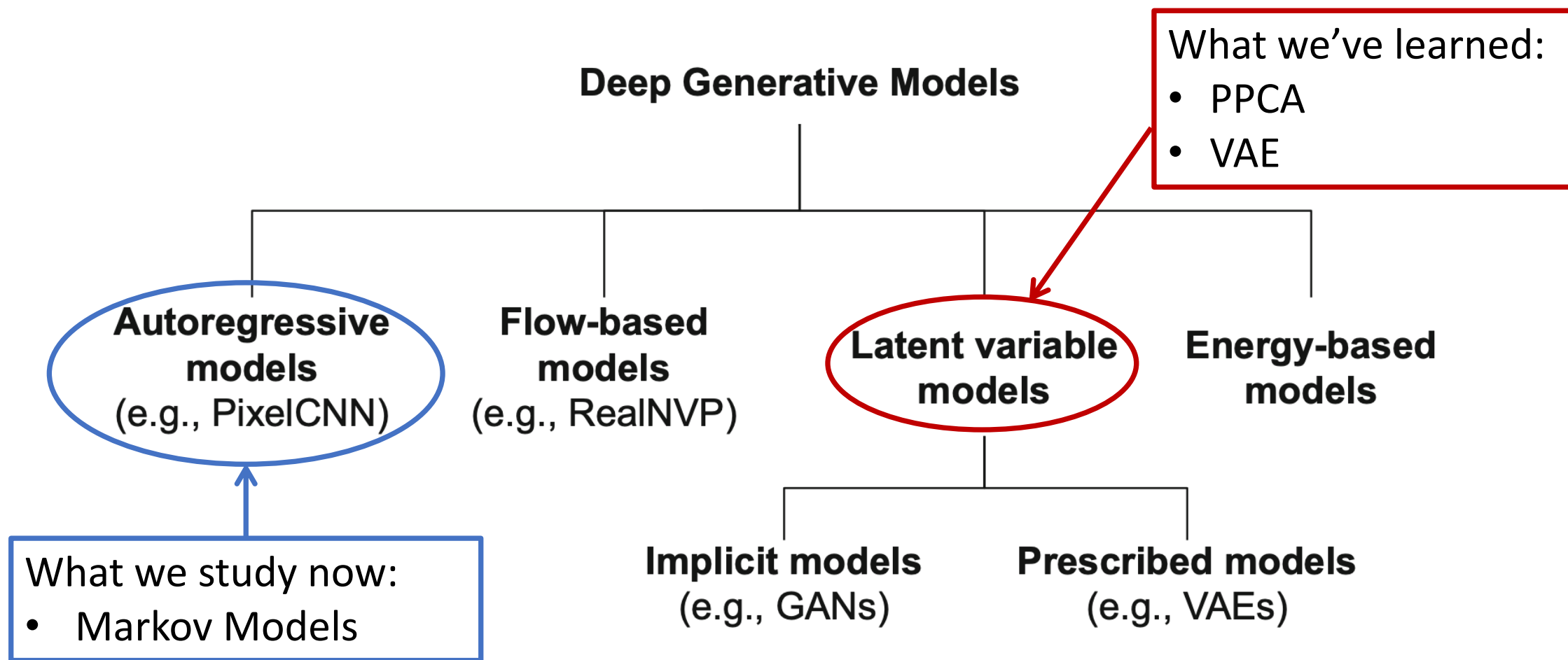
Fall Semester 2025

René Vidal

Director of the Center for Innovation in Data Engineering and Science (IDEAS),  
Rachleff University Professor, University of Pennsylvania  
Amazon Scholar & Chief Scientist at NORCE



# Taxonomy of Generative Models



# Lecture Outline

- Stochastic Processes
  - Definition and Examples
- Markov Models and Markov Chains
  - Definition
  - Transition Probability and Transition Matrix
  - Examples
  - Stationarity and Convergence
- Maximum Log-Likelihood for Markov Chains

# Stochastic Process

- **Definition:** A *stochastic process* refers to a sequence of random variables  
 $(X_1, X_2, \dots, X_T)$

- Each  $X_t$  takes values from the same sample space  $\Omega$  (state space)
  - You can assume  $X_t$  has  $K$  states and  $\Omega := \{1, \dots, K\}$

- **Example** (Bernoulli Process):

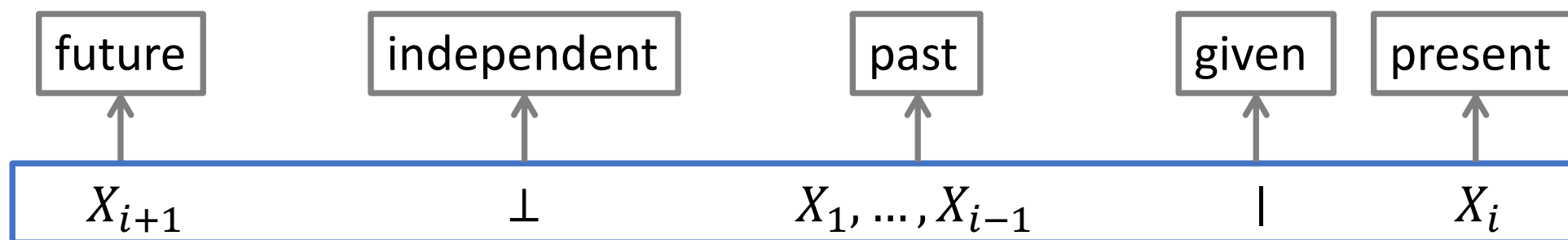
$$X_t \sim \text{Bernouli}(p), \quad t = 1, \dots, T$$

- How many states does  $X_t$  has? What is  $\Omega$ ?

# Markov Property Revisited

- **Issue:** Modeling the joint distribution  $\mathbb{P}(X_1, X_2, \dots, X_T)$  might require **exponentially many parameters** in the absence of any assumptions on  $P$

- **Conditional Independence Assumption** (Markov property):



- **Consequence:** We now only need **linearly many parameters**:

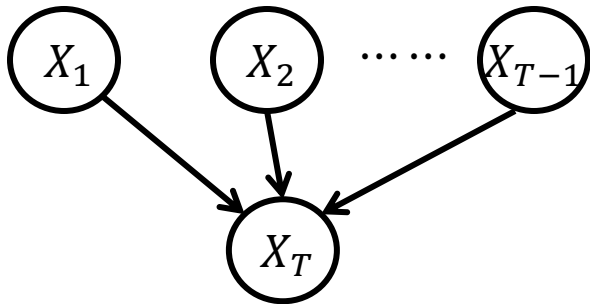
$$\begin{aligned}\mathbb{P}(x_1, \dots, x_T) &= \mathbb{P}(x_1) \mathbb{P}(x_2 \mid x_1) \mathbb{P}(x_3 \mid \cancel{x_1}, x_2) \cdots \mathbb{P}(x_T \mid \cancel{x_1, \dots, x_{T-2}}, x_{T-1}) \\ &= \mathbb{P}(x_1) p(x_2 \mid x_1) \mathbb{P}(x_3 \mid x_2) \cdots \mathbb{P}(x_T \mid x_{T-1})\end{aligned}$$

# Markov Chains

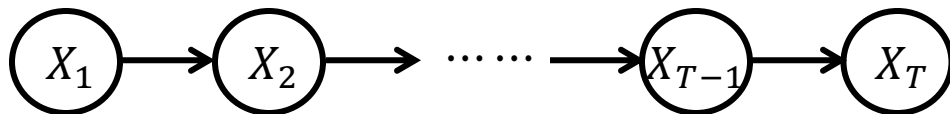
- **Definition:** A (discrete time) *Markov chain* is a stochastic process  $(X_1, X_2, \dots, X_T)$  with the *Markov property*

$$\mathbb{P}(x_1, \dots, x_T) = \mathbb{P}(x_1) \mathbb{P}(x_2 \mid x_1) \mathbb{P}(x_3 \mid x_2) \cdots \mathbb{P}(x_T \mid x_{T-1})$$

- Without Markov Property:



- With Markov Property:



# Parameters of Markov Chains

- Initial Probability  $\pi_1, \dots, \pi_K$ :  $\pi_i := \mathbb{P}(X_1 = i)$ .
- Transition Probability  $a_{ij}$ :

$$a_{ij} := \mathbb{P}(X_{t+1} = j \mid X_t = i) \quad \forall i, j \in \Omega = \{1, \dots, K\}$$

- This is the probability that  $X_t$  transitions from state  $i$  to state  $j$
- Matrix and Vector Notations:

$$A := \begin{bmatrix} a_{11} & \dots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \dots & a_{KK} \end{bmatrix} \in \mathbb{R}^{K \times K},$$

$$\pi := [\pi_1, \dots, \pi_K] \in \mathbb{R}^{1 \times K}$$

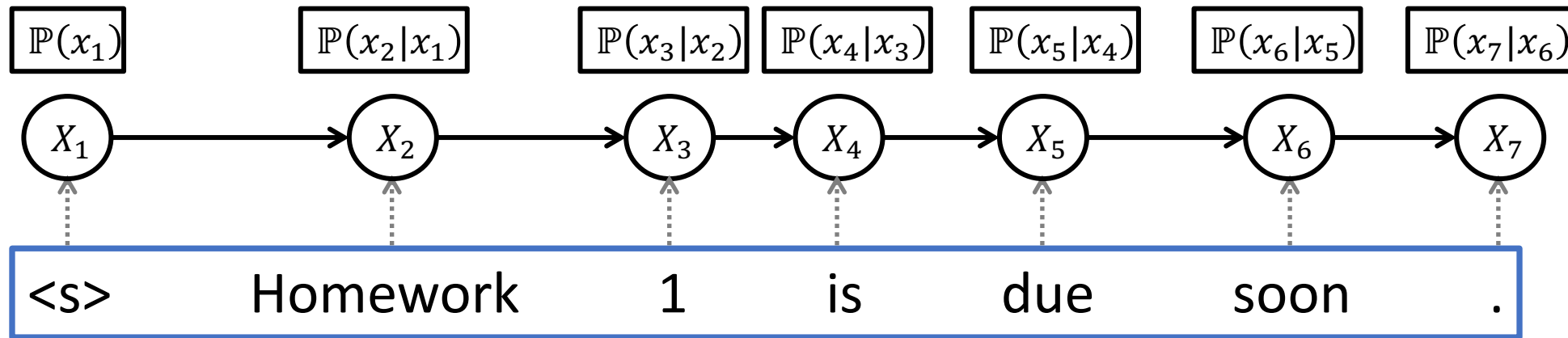
Row Vector



A Markov chain is fully specified by its parameters  $\theta := (\pi, A)$

# Example: Markov Sentence Model

- State space  $\Omega = \{\text{all possible words}\}$ 
  - The following are viewed as words and included in the state space
    - $\langle s \rangle$ : the start of the sentence
    - Digits
    - Punctuations
- Each sentence is a Markov chain where the words are random variables:



- Meaning of  $\mathbb{P}(x_{t+1}|x_t)$ : Given that the current word is  $x_t$ , what is the probability that the next word is  $x_{t+1}$ ?



# Example: DNA Sequencing

- State Space  $\Omega = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$

- Transition Matrix  $A$ :

	$\mathcal{A}$	$\mathcal{C}$	$\mathcal{G}$	$\mathcal{T}$
$\mathcal{A}$	0.359	0.143	0.167	0.331
$\mathcal{C}$	0.384	0.156	0.023	0.437
$\mathcal{G}$	0.306	0.199	0.150	0.345
$\mathcal{T}$	0.284	0.182	0.177	0.357

Initial Probability Vector  $\pi$ :

$\mathcal{A}$	0.25
$\mathcal{C}$	0.25
$\mathcal{G}$	0.25
$\mathcal{T}$	0.25

- **Question 1:** Given  $\mathcal{C}$ , what is the probability of getting DNA sequence  $\mathcal{CTGAC}$ ?

- **Answer 1:**

$$\mathbb{P}(\mathcal{CTGAC} | X_1 = \mathcal{C}) = \mathbb{P}(\mathcal{C} | \mathcal{T}) \cdot \mathbb{P}(\mathcal{T} | \mathcal{G}) \cdot \mathbb{P}(\mathcal{G} | \mathcal{A}) \cdot \mathbb{P}(\mathcal{A} | \mathcal{C}) \approx 0.00403$$

↑  

0.182

↑  

0.345

↑  

0.167

↑  

0.384

# Example: DNA Sequencing

- State Space  $\Omega = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$

	$\mathcal{A}$	$\mathcal{C}$	$\mathcal{G}$	$\mathcal{T}$
$\mathcal{A}$	0.359	0.143	0.167	0.331
$\mathcal{C}$	0.384	0.156	0.023	0.437
$\mathcal{G}$	0.306	0.199	0.150	0.345
$\mathcal{T}$	0.284	0.182	0.177	0.357

$\mathcal{A}$	0.25
$\mathcal{C}$	0.25
$\mathcal{G}$	0.25
$\mathcal{T}$	0.25

- Question 2: What's the probability of  $X_3 = \mathcal{A}$  given  $X_1 = \mathcal{C}$ ?
- Question 3: What's the probability of  $X_3 = \mathcal{A}$ ?

# Example: DNA Sequencing

- State Space  $\Omega = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$

	$\mathcal{A}$	$\mathcal{C}$	$\mathcal{G}$	$\mathcal{T}$	
$\mathcal{A}$	0.359	0.143	0.167	0.331	$\mathcal{A}$ 0.25
$\mathcal{C}$	0.384	0.156	0.023	0.437	$\mathcal{C}$ 0.25
$\mathcal{G}$	0.306	0.199	0.150	0.345	$\mathcal{G}$ 0.25
$\mathcal{T}$	0.284	0.182	0.177	0.357	$\mathcal{T}$ 0.25

- Question 2:** What's the probability of  $X_3 = \mathcal{A}$  given  $X_1 = \mathcal{C}$ ?
- Question 3:** What's the probability of  $X_3 = \mathcal{A}$ ?
- Answer 2:** The state transition is  $\mathcal{C} \rightarrow x_2 \rightarrow \mathcal{A}$  for all possible  $x_2 \in \Omega$ :

$$\begin{aligned} \mathbb{P}(X_3 = \mathcal{A} | X_1 = \mathcal{C}) &= \sum_{x_2 \in \Omega} \mathbb{P}(X_3 = \mathcal{A} | X_2 = x_2) \cdot \mathbb{P}(X_2 = x_2 | X_1 = \mathcal{C}) \\ &= [0.384, 0.156, 0.023, 0.437] \begin{bmatrix} 0.359 \\ 0.384 \\ 0.306 \\ 0.284 \end{bmatrix} \end{aligned}$$

- This is the inner product of **the second row** and **first column** of the transition matrix  $A$
- This is the (2,1)-th entry of  $A^2$

# Example: DNA Sequencing

- State Space  $\Omega = \{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$

	$\mathcal{A}$	$\mathcal{C}$	$\mathcal{G}$	$\mathcal{T}$		
$\mathcal{A}$	0.359	0.143	0.167	0.331	$\mathcal{A}$	0.25
$\mathcal{C}$	0.384	0.156	0.023	0.437	$\mathcal{C}$	0.25
$\mathcal{G}$	0.306	0.199	0.150	0.345	$\mathcal{G}$	0.25
$\mathcal{T}$	0.284	0.182	0.177	0.357	$\mathcal{T}$	0.25

- Question 2:** What's the probability of  $X_3 = \mathcal{A}$  given  $X_1 = \mathcal{C}$ ?
- Question 3:** What's the probability of  $X_3 = \mathcal{A}$ ?
- Answer 2:** The state transition is  $\mathcal{C} \rightarrow x_2 \rightarrow \mathcal{A}$  for all possible  $x_2 \in \Omega$ :

$$\mathbb{P}(X_3 = \mathcal{A} | X_1 = \mathcal{C}) = \sum_{x_2 \in \Omega} \mathbb{P}(X_3 = \mathcal{A} | X_2 = x_2) \cdot \mathbb{P}(X_2 = x_2 | X_1 = \mathcal{C})$$

- Answer 3:** The state transition is  $x_1 \rightarrow x_2 \rightarrow \mathcal{A}$  for all possible  $x_1, x_2 \in \Omega$ :

$$\mathbb{P}(X_3 = \mathcal{A}) = \sum_{x_1 \in \Omega} \mathbb{P}(X_3 = \mathcal{A} | X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$$

Question 2

Initial Probability

# Generalizing the DNA Sequencing Example

- State Space  $\Omega = \{1, \dots, K\}$
- $(A^s)_{ij}$ : the  $(i, j)$ -th entry of  $A^s$
- $(A^s)_{:j}$ : the  $j$ -th column of  $A^s$
- $(\cdot)_j$ : the  $j$ -th entry of a vector

Transition Matrix  $A$  and initial probability distribution  $\pi$  :

$$A := \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{bmatrix} \in \mathbb{R}^{K \times K}, \quad \pi := [\pi_1, \dots, \pi_K] \in \mathbb{R}^{1 \times K}$$

- Claim 1:  $\mathbb{P}(X_{t+s} = j \mid X_t = i) = (A^s)_{ij} \quad (\forall s, t, i, j)$
- Claim 2:  $\mathbb{P}(X_{s+1} = j) = (\pi A^s)_j \quad (\forall s, j)$
- Proof of Claim 2:

$$\mathbb{P}(X_{s+1} = j) = \sum_{i \in \Omega} \mathbb{P}(X_{s+1} = j \mid X_1 = i) \cdot \mathbb{P}(X_1 = i) = \sum_{i \in \Omega} (A^s)_{ij} \cdot \pi_i = \pi(A^s)_{:j} = (\pi A^s)_j$$

- Proof of Claim 1: By induction (next page)

Claim 1

# Proof of Claim 1

- State Space  $\Omega = \{1, \dots, K\}$
- $(A^s)_{ij}$ : the  $(i, j)$ -th entry of  $A^s$
- $(A^s)_{:j}$ : the  $j$ -th column of  $A^s$
- $(\cdot)_j$ : the  $j$ -th entry of a vector

Transition Matrix  $A$  and initial probability distribution  $\pi$  :

$$A := \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{bmatrix} \in \mathbb{R}^{K \times K}, \quad \pi := [\pi_1, \dots, \pi_K] \in \mathbb{R}^{1 \times K}$$

- **Claim 1:**  $\mathbb{P}(X_{t+s} = j \mid X_t = i) = A_{ij}^s$
- **Proof of Claim 1 (Induction):**
  - $\forall s, t$ , it is easy to prove **shift invariance**:  $\mathbb{P}(X_{t+s} = j \mid X_t = i) = \mathbb{P}(X_{1+s} = j \mid X_1 = i)$
  - Next we prove  $\mathbb{P}(X_{1+s} = j \mid X_1 = i) = A_{ij}^s$  by induction on  $s$ :
    - The base case  $s = 1$  follows from the definition of  $A$
    - Suppose we have  $\mathbb{P}(X_s = j \mid X_1 = i) = A_{ij}^{s-1}$  then:

$$\mathbb{P}(X_{1+s} = j \mid X_1 = i) = \sum_{k \in \Omega} \mathbb{P}(X_{s+1} = j \mid X_s = k) \cdot \mathbb{P}(X_s = k \mid X_1 = i) = \sum_{k \in \Omega} a_{kj} \cdot (A^{s-1})_{ik} = (A^s)_{ij}$$

# Limiting Behavior of Markov Chains

- We have just proved

$$\mathbb{P}(X_{t+s} = j \mid X_t = i) = (A^s)_{ij} \quad (\forall s, t, i, j)$$

$$\mathbb{P}(X_{s+1} = j) = (\pi A^s)_j \quad (\forall s, j)$$

- Our next goal is to understand the limits  $\lim_{s \rightarrow \infty} A^s, \lim_{s \rightarrow \infty} \pi A^s$ .
- The two limits are related to the eigenvalues of  $A$ :
  - Assume  $A$  is diagonalizable and write  $A = U\Lambda U^{-1}$  with eigenvalues  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$ 
    - The diagonalizability assumption is not necessary but to simplify the exposition...
  - Then we have
$$\lim_{s \rightarrow \infty} A^s = U \left( \lim_{s \rightarrow \infty} \Lambda^s \right) U^{-1}, \quad \lim_{s \rightarrow \infty} \pi A^s = \pi U \left( \lim_{s \rightarrow \infty} \Lambda^s \right) U^{-1}$$
- Hence, a necessary condition for the limits to exist is that  $|\lambda_k| \leq 1$  for all  $k$ .

# Eigenvalues of Transition Matrix

$$A := \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{bmatrix} \in \mathbb{R}^{K \times K}$$

- **Proposition.** Let  $\lambda_1, \dots, \lambda_K$  be eigenvalues of  $A$ . Then

$$\max_{k=1, \dots, K} |\lambda_k| = 1.$$

- **Proof.** We first show  $|\lambda_k| \leq 1$ . Let  $(\lambda, u)$  be an eigen-pair with  $Au = \lambda u$ ,  $\|u\|_2 = 1$  and  $u = [u_1, \dots, u_K]^\top$ . Let  $i$  be the index such that  $|u_i|$  is maximized, i.e.,  
$$i = \operatorname{argmax}_j |u_j|.$$

Then  $Au = \lambda u$  implies  $\sum_j a_{ij} u_j = \lambda u_i$ , which furthermore gives

$$|\lambda| \leq \left| \frac{\sum_j a_{ij} u_j}{u_i} \right| \leq \sum_j |a_{ij}| \cdot \left| \frac{u_j}{u_i} \right| \leq \sum_j |a_{ij}| = \sum_j a_{ij} = 1.$$

Finally,  $A$  always has an eigenvalue 1:  $A \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$



# $|\lambda_k| \leq 1$ is not sufficient for convergence

- Intuition:

- Assume  $A$  is diagonalizable and write  $A = U\Lambda U^{-1}$  with eigenvalues  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_K)$ 
  - The diagonalizability assumption is not necessary but to simplify the exposition...

- Then  $\lim_{s \rightarrow \infty} A^s = U \left( \lim_{s \rightarrow \infty} \Lambda^s \right) U^{-1} = U \left( \text{diag} \left( \lim_{s \rightarrow \infty} \lambda_1^s, \dots, \lim_{s \rightarrow \infty} \lambda_K^s \right) \right) U^{-1}$

- And  $\lim_{s \rightarrow \infty} \lambda_k^s$  ...
  - is equal to 0 if  $|\lambda_k| < 1$
  - is equal to 1 if  $\lambda_k = 1$
  - does not exist if  $\lambda_k = -1$

- $\lambda_k$  can even be a complex eigenvalue with  $|\lambda_k| = 1$

# Lesson

- **Existence.** In order for  $\lim_{s \rightarrow \infty} A^s, \lim_{s \rightarrow \infty} \pi A^s$  to exist, we need to make assumptions such that:
  - $A$  has no eigenvalues of magnitude 1 other than 1 itself.
- **Uniqueness.** In order for  $\lim_{s \rightarrow \infty} \pi A^s$  to be the same for different initial distribution  $\pi$ , we need to make assumptions such that:
  - 1 is the eigenvalue of  $A$  of geometric/algebraic multiplicity 1
- The assumptions should be “interpretable” in terms of Markov chains or states
  - e.g., assuming  $A$  to be diagonalizable is not interpretable

# Irreducibility and Strongly Connected Graph

- **Definition.** A directed graph is called *strongly connected* if there is a path in each direction between each pair of vertices of the graph.
- **Definition.** A transition matrix  $A$  is called *irreducible* if every state can be reached from any other state, i.e., for any  $i, j$ , there is some  $t$  such that

$$\mathbb{P}(X_t = j \mid X_1 = i) > 0.$$

- **Remark.** Each state can be denoted by a vertex and, if  $a_{ij} > 0$  then we add a directed edge from vertex  $i$  to vertex  $j$ . This way, we obtain a directed graph. We can see that  **$A$  is irreducible if and only if the graph is strongly connected**

# Limiting Behavior of Markov Chains

- **Theorem.** Assume  $A$  is irreducible, then there is some  $v = [v_1, \dots, v_K]$  such that

- For any initial distribution  $\pi$  we have:

$$\lim_{s \rightarrow \infty} \frac{I + A + \dots + A^{s-1}}{s} = ev,$$

$$\lim_{s \rightarrow \infty} \pi \left( \frac{I + A + \dots + A^{s-1}}{s} \right) = v$$

$$e := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- If furthermore there is some  $a_{ii} > 0$ , then for any initial distribution  $\pi$  we have

$$\lim_{s \rightarrow \infty} A^s = ev,$$

$$\lim_{s \rightarrow \infty} \pi A^s = v$$

- **Remark.** In the latter case,  $v$  is called the **stationary distribution** as it is the unique vector that satisfies:

$$vA = v, \quad v_i > 0 \ (\forall i), \quad \sum_i v_i = 1$$

- **Remark on Proof.** This result is related to Perron–Frobenius Theory (Google search it). For its proof, see Chapter 7 (Perron–Frobenius Theory) of “*Matrix Analysis and Applied Linear Algebra*”, Second Edition (Carl D. Meyer, 2023).

# Example

• Let  $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$  and  $v = [0.5, 0.5]$  and  $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ . Note that  $a_{11} = a_{22} = 0$

•  $A$  has two eigenvalues, 1 and  $-1$ .

• We have  $A^{2t} = I$  and  $A^{2t+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$  for any  $t$ , so  $\lim_{s \rightarrow \infty} A^s$  does not exist.

• We have  $vA = [0.5, 0.5] \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = [0.5, 0.5] = v$ , so  $\lim_{s \rightarrow \infty} vA^s = [0.5, 0.5]$ . However, for any initial distribution  $\pi$  different from  $v$ ,  $\lim_{s \rightarrow \infty} \pi A^s$  does not exist.

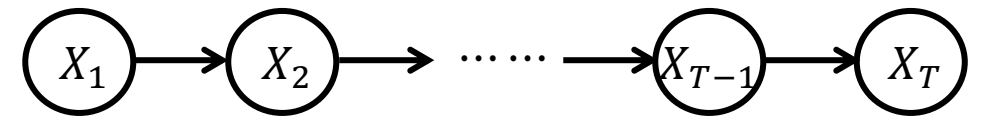
• On the other hand, we have  $\left(\frac{I+A}{2}\right)^2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} = \frac{I+A}{2}$ , which implies

$$\lim_{s \rightarrow \infty} \left(\frac{I+A}{2}\right)^s = \frac{I+A}{2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [0.5, 0.5] = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \pi$$

# Estimate Transition Parameters $\theta$ from Data

- We have derived some results based on the transition matrix ...
- In practice, we are given data samples rather than the transition matrix
- We will assume the data are sampled from a Markov chain, and then compute the transition matrix from data via **maximum likelihood estimation (MLE)**

# MLE of Markov Chains



- Assume we have  $N$  i.i.d. samples  $\{\mathbf{x}^{(n)}\}_{n=1}^N$  from distribution  $p_{\theta}(\mathbf{x})$ 
  - $\mathbf{x} := (x_1, \dots, x_T)$  each  $x_t$  has  $K$  status
  - $\theta = (A, \pi)$ : unknown transition matrix and initial probability distribution

• MLE:

$$(\hat{A}_{ML}, \hat{\pi}_{ML}) = \operatorname{argmax}_{A, \pi} \prod_{n=1}^N p_{A, \pi}(\mathbf{x}^{(n)})$$

Markov Property

$$(\hat{A}_{ML}, \hat{\pi}_{ML}) = \operatorname{argmax}_{A, \pi} \prod_{n=1}^N p_{\pi}(x_1^{(n)}) \prod_{t=2}^T p_A(x_t^{(n)} | x_{t-1}^{(n)})$$

Variables are separable

Similar to  
estimating  
 $\hat{A}_{ML}$ , so left as  
an exercise

$$\hat{\pi}_{ML} = \operatorname{argmax}_{\pi} \prod_{n=1}^N p_{\pi}(x_1^{(n)})$$

Our focus next

$$\hat{A}_{ML} = \operatorname{argmax}_A \prod_{n=1}^N \prod_{t=2}^T p_A(x_t^{(n)} | x_{t-1}^{(n)})$$

# Simplifying The MLE

- $\mathbb{I}(\cdot)$ : indicator function
- $N_{ij}$ : the number of samples with transitions from state  $i$  to state  $j$ , i.e.,

$$N_{ij} := \sum_{n=1}^N \sum_{t=2}^T \mathbb{I} \left( x_t^{(n)} = j, x_{t-1}^{(n)} = i \right)$$

Then we have:

1.  $p_A \left( x_t^{(n)} \mid x_{t-1}^{(n)} \right) = \prod_{i=1}^K \prod_{j=1}^K (a_{ij})^{\mathbb{I}(x_t^{(n)}=j, x_{t-1}^{(n)}=i)}$
2.  $\prod_{n=1}^N \prod_{t=2}^T p_A \left( x_t^{(n)} \mid x_{t-1}^{(n)} \right) = \prod_{n=1}^N \prod_{t=2}^T \prod_{i=1}^K \prod_{j=1}^K (a_{ij})^{\mathbb{I}(x_t^{(n)}=j, x_{t-1}^{(n)}=i)}$   
 $= \prod_{i=1}^K \prod_{j=1}^K (a_{ij})^{N_{ij}}$

This gives:

$$\hat{A}_{ML} = \operatorname{argmax}_A \prod_{i=1}^K \prod_{j=1}^K (a_{ij})^{N_{ij}}$$

$$\hat{A}_{ML} = \operatorname{argmax}_A \prod_{n=1}^N \prod_{t=2}^T p_A \left( x_t^{(n)} \mid x_{t-1}^{(n)} \right)$$



# Simplifying The MLE

$$\hat{A}_{ML} = \operatorname{argmax}_A \prod_{i=1}^K \prod_{j=1}^K (a_{ij})^{N_{ij}}$$

- $N_{ij}$ : the number of samples with transitions from state  $i$  to state  $j$

Taking logarithm and adding constraints  $\sum_j a_{ij} = 1$ :

$$\hat{A}_{ML} = \operatorname{argmax}_A \sum_{i=1}^K \sum_{j=1}^K N_{ij} \log a_{ij} \quad \text{subject to} \quad \sum_{j=1}^K a_{ij} = 1 \quad (\forall i)$$



Variables are separable

Solve the following for every  $i = 1, \dots, K$ :

$$\hat{a}_{ij_{ML}} = \operatorname{argmax}_{a_{ij}} \sum_{j=1}^K N_{ij} \log a_{ij} \quad \text{subject to} \quad \sum_{j=1}^K a_{ij} = 1$$

Remark: We have seen how to solve it using Lagrangian multipliers (recall *EM for Gaussian Mixture Models*)

$$\hat{a}_{ij_{ML}} = \frac{N_{ij}}{\sum_{j=1}^K N_{ij}}$$

Remark: The optimal transition matrix can be found by simply counting and classifying the number of the transitions of the sample states!

# Conclusion

- Markov chains have several applications
- For irreducible transition matrix with at least one positive entry, the Markov chain will eventually reach a stationary distribution
- The transition matrix can be learned from data via maximum likelihood estimation