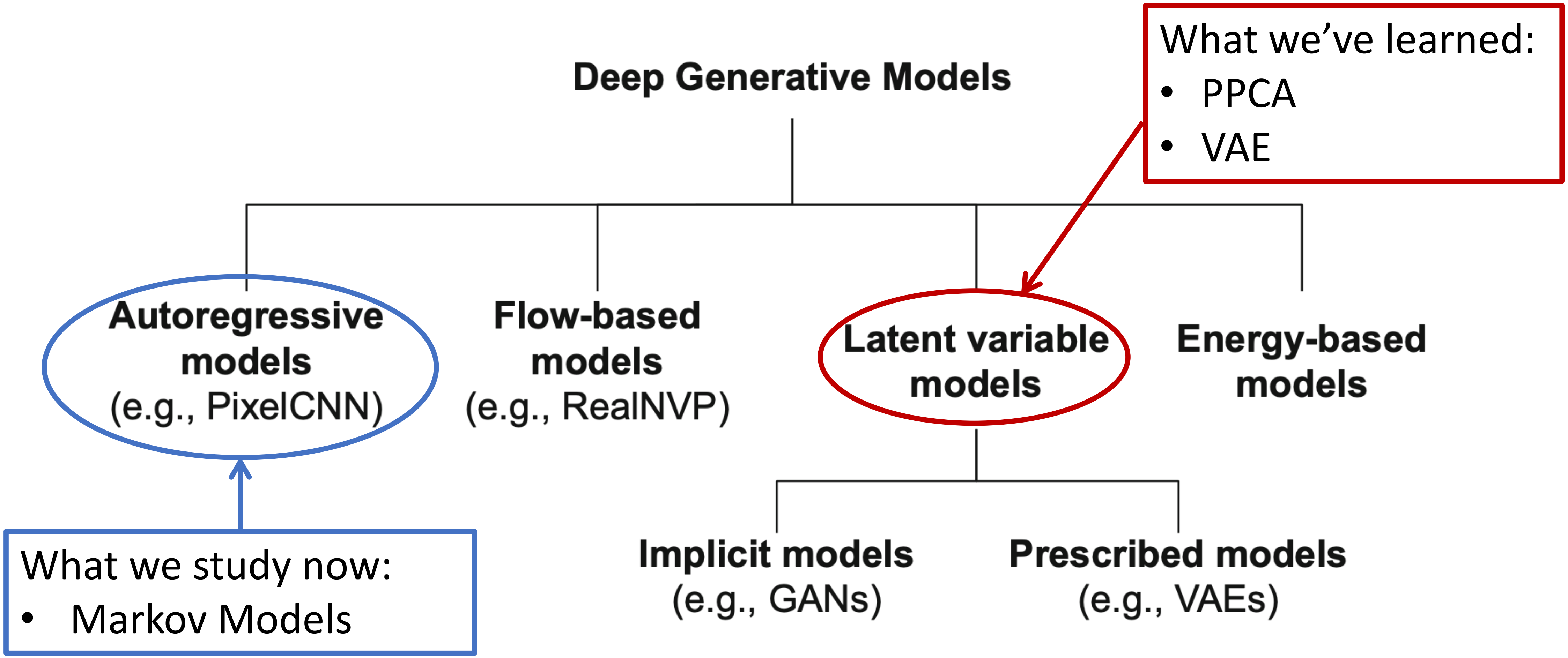
Deep Generative Models: Markov Models

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 Matrix
  - Examples
- Inference via Matrix Multiplication
- Learning via Maximum Likelihood

Stochastic Process

- Definition: A *stochastic process* $(X_0, X_1, \ldots, X_T)$ is a sequence of random variables where each X_t takes values on the same sample space Ω (state space)

- Example (Bernoulli Process): X_t has 2 states and $\Omega \coloneqq \{0,1\}$

$$X_t \sim \mathrm{Bernouli}(q), \qquad t = 0, \ldots, T$$

  - How many states does X_t has? What is Ω?

- Example (Categorical Process): X_t has K states and $\Omega \coloneqq \{1, \ldots, K\}$

$$X_t \sim \mathrm{Cat}(\pi), \qquad t = 0, \ldots, T$$

  - How many states does X_t has? What is Ω?

Markov Property Revisited

- **Issue**: In the absence of any assumptions on $\mathbb{P}$, modeling the joint distribution $\mathbb{P}(X_0, X_1, \dots, X_T)$ might require exponentially many parameters

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

$$\overset{\text{future}}{X_{i+1}} \quad \overset{\text{independent}}{\perp} \quad \overset{\text{past}}{X_0, \dots, X_{i-1}} \quad \overset{\text{given}}{\mid} \quad \overset{\text{present}}{X_i}$$

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

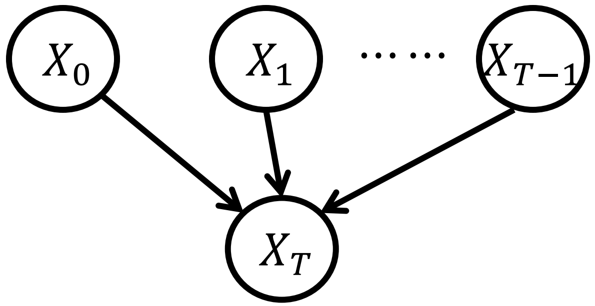
$$\begin{aligned}\mathbb{P}(x_0, \dots, x_T) &= \mathbb{P}(x_0)\mathbb{P}(x_1 \mid x_0)\mathbb{P}(x_2 \mid \cancel{x_0}, x_1) \cdots \mathbb{P}(x_T \mid \cancel{x_0, \cdots, x_{T-2}}, x_{T-1}) \\ &= \mathbb{P}(x_0)p(x_1 \mid x_0)\mathbb{P}(x_2 \mid x_1) \cdots \mathbb{P}(x_T \mid x_{T-1})\end{aligned}$$

Markov Chains

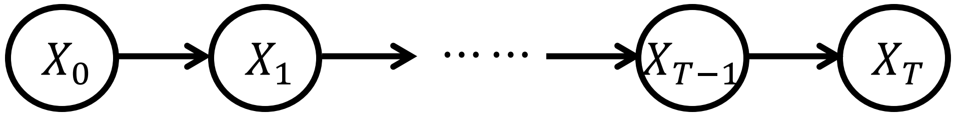
- Definition: A (discrete time) *Markov chain* is a stochastic process $(X_0, X_1, \ldots, X_T)$ with the *Markov property*

$$\mathbb{P}(x_0, \ldots, x_T) = \mathbb{P}(x_0)\mathbb{P}(x_1 \mid x_0)\mathbb{P}(x_2 \mid x_1) \cdots \mathbb{P}(x_T \mid x_{T-1})$$

  - Without the Markov Property:



  - With the Markov Property:



Parameters of Markov Chains

- Initial Probability: π_i is the probability that X_0 starts at state i

$$\pi_i \coloneqq \mathbb{P}(X_0 = i) \qquad \forall i \in \Omega = \{1, \dots, K\}$$

- Transition Probability: a_{ij} is the probability that X_t transitions from state i to j

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

- Matrix and Vector Notations:

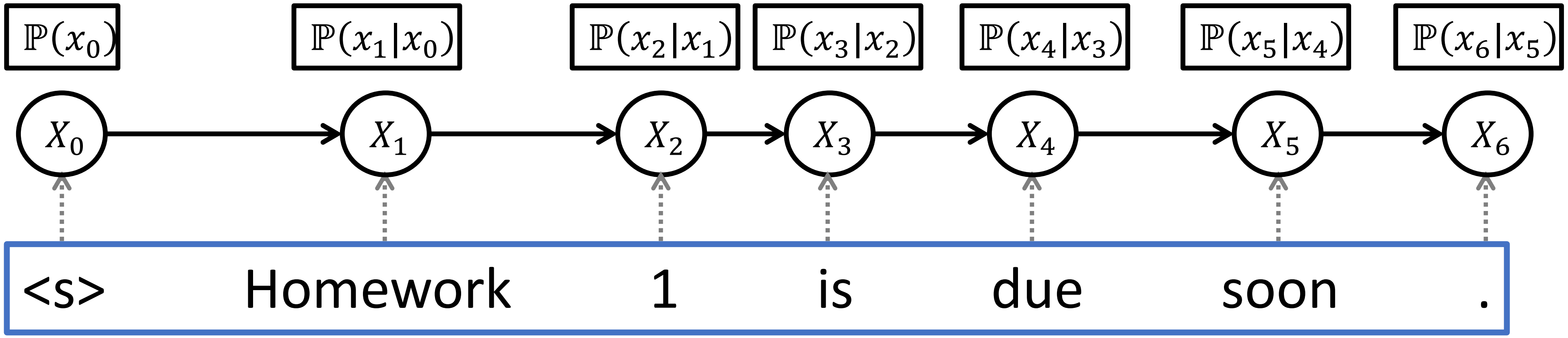
Row Vector

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

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
    - <s>: 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 π:

$\mathcal{A}$	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:

$$\mathbb{P}(\mathcal{CTGAC} \mid X_0 = \mathcal{C}) = \underbrace{\mathbb{P}(\mathcal{T} \mid \mathcal{C})}_{\boxed{0.437}} \cdot \underbrace{\mathbb{P}(\mathcal{G} \mid \mathcal{T})}_{\boxed{0.177}} \cdot \underbrace{\mathbb{P}(\mathcal{A} \mid \mathcal{G})}_{\boxed{0.306}} \cdot \underbrace{\mathbb{P}(\mathcal{C} \mid \mathcal{A})}_{\boxed{0.143}} \approx 0.00338$$

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 $X_2 = \mathcal{A}$ given $X_0 = \mathcal{C}$?

- Answer 2: The state transition is $\mathcal{C} \to x_1 \to \mathcal{A}$ for all possible $x_1 \in \Omega$:

$$\mathbb{P}(X_2 = \mathcal{A} \mid X_0 = \mathcal{C}) = \sum_{x_1 \in \Omega} \mathbb{P}(X_2 = \mathcal{A} \mid X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = \mathcal{C})$$

$$= [0.384, 0.156, 0.023, 0.437] \begin{bmatrix} 0.359 \\ 0.384 \\ 0.306 \\ 0.284 \end{bmatrix}$$

- 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
$\mathcal{C}$	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 3: What's the probability of $X_2 = \mathcal{A}$?

- Answer 3: The state transition is $x_0 \to x_1 \to \mathcal{A}$ for all possible $x_0, x_1 \in \Omega$:

$$\mathbb{P}(X_2 = \mathcal{A}) = \sum_{x_0 \in \Omega} \underbrace{\mathbb{P}(X_2 = \mathcal{A} \mid X_0 = x_0)}_{\text{Question 2}} \cdot \underbrace{\mathbb{P}(X_0 = x_0)}_{\text{Initial Probability}}$$

  - This is the inner product of the vector π and the first column of A^2: $\pi A^2 e_1$

Transition Matrix and Distribution of Future States

- 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 π :

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

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

$$\mathbb{P}(X_s = j) = \sum_{i \in \Omega} \mathbb{P}(X_s = j \mid X_0 = i) \cdot \mathbb{P}(X_0 = i) \underset{\boxed{\text{Claim 1}}}{=} \sum_{i \in \Omega} (A^s)_{ij} \cdot \pi_i = \pi (A^s)_{:j} = (\pi A^s)_j$$

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

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 π :

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

- Claim 1: $\mathbb{P}(X_{t+s} = j \mid X_t = i) = (A^s)_{ij}$
- Proof of Claim 1 (Induction):
  - $\forall s, t$, it is easy to prove shift invariance: $\mathbb{P}(X_{t+s} = j \mid X_t = i) = \mathbb{P}(X_s = j \mid X_0 = i)$
  - Next we prove $\mathbb{P}(X_s = j \mid X_0 = i) = A_{ij}^s$ by induction on s:
    - The base case $s = 1$ follows from the definition of A
    - Suppose we have $\mathbb{P}(X_{s-1} = j \mid X_0 = i) = A_{ij}^{s-1}$ then:

$$\mathbb{P}(X_s = j \mid X_0 = i) = \sum_{k \in \Omega} \mathbb{P}(X_s = j \mid X_{s-1} = k) \cdot \mathbb{P}(X_{s-1} = k \mid X_1 = i) = \sum_{k \in \Omega} a_{kj} \cdot (A^{s-1})_{ik} = (A^s)_{ij}$$

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 $\lambda_1, \ldots, \lambda_K$ be eigenvalues of A. Then

$$\max_{k=1,\ldots,K} |\lambda_k| = 1 \qquad \text{and} \qquad A \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Proof: Let (λ, u) be an eigen-pair, i.e., $Au = \lambda u$, $u = [u_1, \ldots, u_K]^\top$ and $||u||_2 = 1$. Let i^* be the index such that $|u_i|$ is maximized, i.e., $i^* = \mathrm{argmax}_i \, |u_i|$. Then, $Au = \lambda u$ implies $\sum_j a_{i^*j} u_j = \lambda u_{i^*}$, which furthermore gives

$$|\lambda| \leq \left| \frac{\sum_j a_{i^*j} u_j}{u_{i^*}} \right| \leq \sum_j |a_{i^*j}| \cdot \left| \frac{u_j}{u_{i^*}} \right| \leq \sum_j |a_{i^*j}| = \sum_j a_{i^*j} = 1.$$

Finally, since $\sum_j a_{ij} = 1$ for all $i = 1, \ldots K$, then $A1 = 1$, hence $\max_{k=1,\ldots,K} |\lambda_k| = 1$.

Learning the Parameters θ 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

$$X_0 \to X_1 \to \cdots\cdots \to X_{T-1} \to X_T$$

- Assume we have N i.i.d. samples $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ from distribution $p_\theta(\boldsymbol{x})$
  - $\boldsymbol{x} \coloneqq (x_0, \dots, x_T)$ each x_t can take on K different values
  - $\theta = (A, \pi)$: unknown transition matrix and initial probability distribution
- MLE:

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

Markov Property

$$(\hat{A}_{ML}, \hat{\pi}_{ML}) = \operatorname{argmax}_{A,\pi} \prod_{n=1}^{N} p_\pi\left(x_0^{(n)}\right) \prod_{t=1}^{T} p_A\left(x_t^{(n)} \mid x_{t-1}^{(n)}\right)$$

Variables are separable

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

$$\hat{\pi}_{ML} = \operatorname{argmax}_{\pi} \prod_{n=1}^{N} p_\pi\left(x_0^{(n)}\right)$$

Our focus next

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

Simplifying The MLE

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

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

$$N_{ij} \coloneqq \sum_{n=1}^{N} \sum_{t=0}^{T} \mathbb{I}\left(x_t^{(n)} = j, x_{t-1}^{(n)} = i \right)$$

Then we have:

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

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

$$= \prod_{i=1}^{K} \prod_{j=1}^{K} \left(a_{ij}\right)^{N_{ij}}$$

This gives:

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

Simplifying The MLE

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

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

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

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

Variables are separable

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

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

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

$$\widehat{a_{ij}}_{ML} = \frac{N_{ij}}{\sum_{j=1}^{K} N_{ij}}$$

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
  - Modeling text sequences
  - Modeling gene sequences

- The ML estimate $\widehat{a_{ij}}_{ML}$ of the transition matrix is given by

$$\widehat{a_{ij}}_{ML} = \frac{N_{ij}}{\sum_{j=1}^{K} N_{ij}}$$

Number of transitions from state i to j

- Similarly, the ML estimate $\widehat{\pi}_{i_{ML}}$ of the initial probability is given as

$$\widehat{\pi}_{i_{ML}} = \frac{N_{i0}}{\sum_{j=1}^{K} N_{j0}}$$

Number of times we start in state i