


# Machine Learning for Graphs and Sequential Data

## *Sequential Data – Markov Chains*

Lecturer: Prof. Dr. Stephan Günnemann  
[cs.cit.tum.de/daml](https://cs.cit.tum.de/daml)

---

Summer Term 2023

Data Analytics and  
Machine Learning 

# Roadmap

---

- Chapter: Temporal Data / Sequential Data
  1. Autoregressive Models
  - 2. Markov Chains**
  3. Hidden Markov Models
  4. Neural Network Approaches
  5. Temporal Point Processes

# Markov Chains - Definition

- Definition: A **Markov Chain** is a sequence of r.v.  $X_1, X_2, \dots, X_T$  which fulfills the **Markov property** :

$$P(X_t | X_1, \dots, X_{t-1}) = P(X_t | X_{t-1})$$

- The values taken by the time index  $t$  are discrete i.e.  $t \in \{1, 2, \dots, T\}$
- We assume that the r.v.  $X_t$  are discrete i.e.  $X_t \in \{1, 2, \dots, K\}$
- The joint distribution of a Markov Chain is:

$$P(X_1 = i_1, \dots, X_T = i_T) = P(X_1 = i_1) \prod_{t=1}^{T-1} P(X_{t+1} = i_{t+1} | X_t = i_t)$$

# Markov Chain – General case

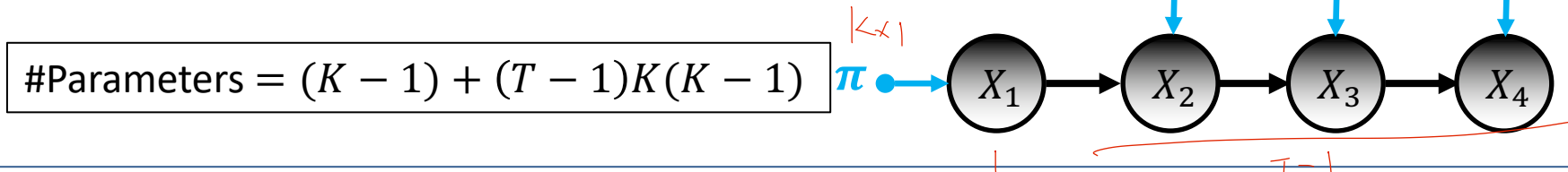
- In the general case, the distribution of each r.v. can be different:

$$P(X_1 = i) = \pi_i \text{ and } P(X_{t+1} = j | X_t = i) = A_{ij}^{(t+1)}$$

where  $\pi \in \mathbb{R}^K$  is a **prior probability** on the initial state, and  $A^{(t)} \in \mathbb{R}^{K \times K}$  are the **transition matrices**.

- Consequently the joint probability and the graphical model are:

$$P(X_1 = i_1, \dots, X_T = i_T) = \pi_{i_1} \times A_{i_1, i_2}^{(2)} \times \dots \times A_{i_{T-1}, i_T}^{(T)}$$



# Markov Chain – Stationary case

- To simplify, we assume a **time-homogeneous** or **stationary** Markov Chain:

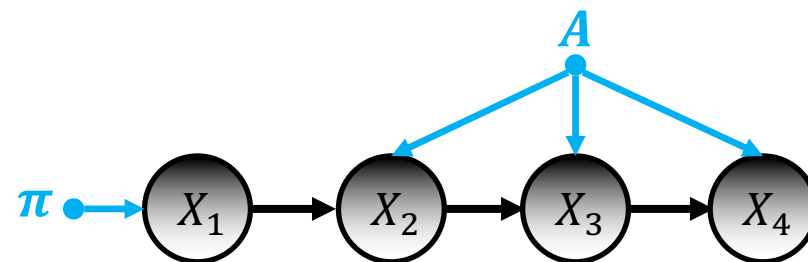
$$P(X_1 = i) = \pi_i \text{ and } P(X_{t+1} = j | X_t = i) = A_{ij}$$

- The transition matrix  $A^{(t)} = A$  does not depend on  $t$ .  
All r.v.  $X_2, \dots, X_T$  follow the same conditional distribution.

- The joint probability and the graphical model become:

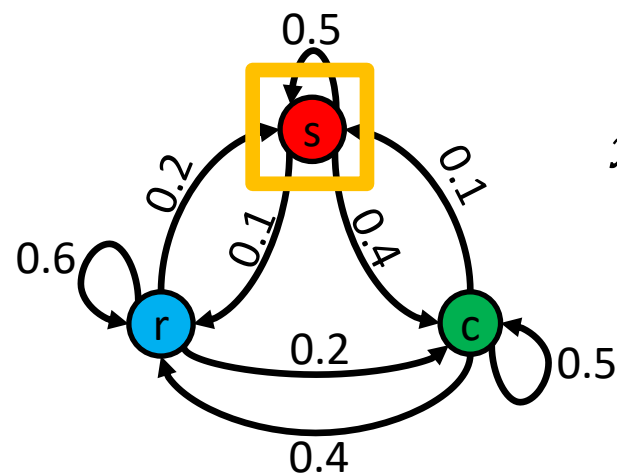
$$P(X_1 = i_1, \dots, X_T = i_T) = \pi_{i_1} \times A_{i_1, i_2} \times \dots \times A_{i_{T-1}, i_T}$$

$\text{\#Parameters} = (K - 1) + K(K - 1)$



# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



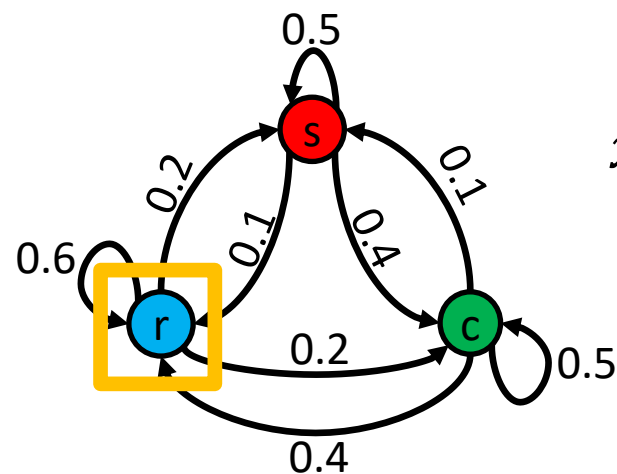
$$x_{1:T} = \boxed{s} \ r \ c \ r \ r \ c \ s$$

$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



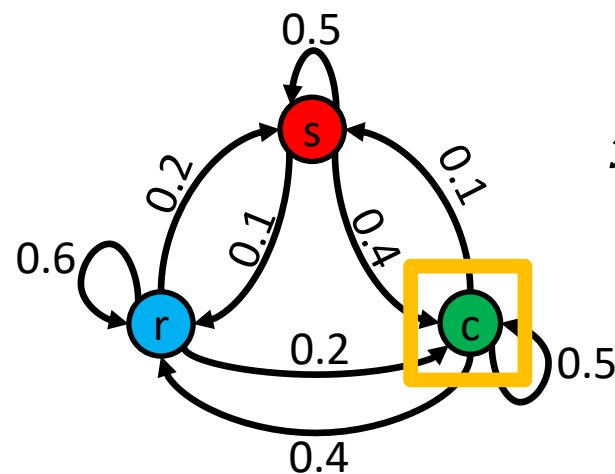
$$x_{1:T} = s \quad \boxed{r} \quad c \quad r \quad r \quad c \quad s$$

$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



$x_{1:T} = s \ r \ c \ r \ r \ c \ s$

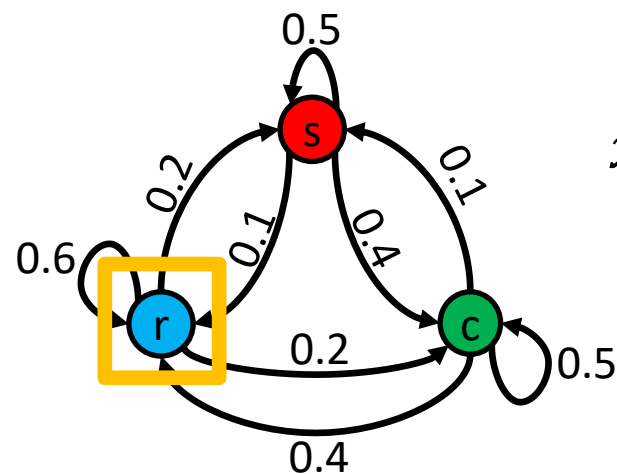
$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$



# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



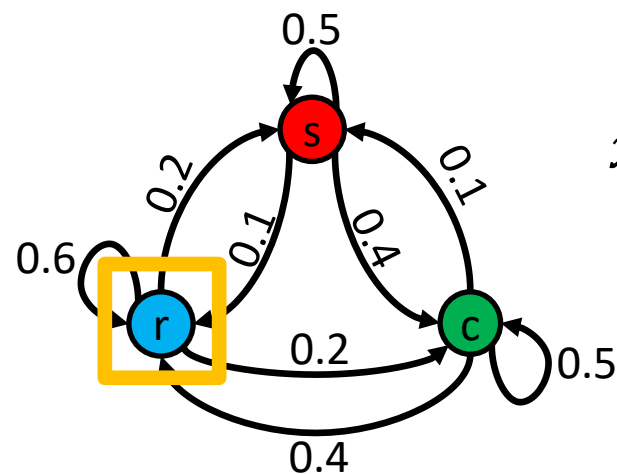
$x_{1:T} = s \ r \ c \ r \ r \ c \ s$

$$A = \begin{matrix} & \begin{matrix} rainy & sunny & cloudy \end{matrix} \\ \begin{matrix} rainy \\ sunny \\ cloudy \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



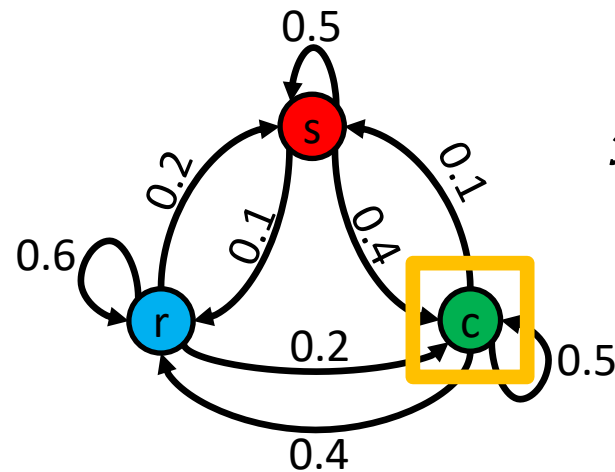
$x_{1:T} = s \ r \ c \ r \ r \ c \ s$

$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



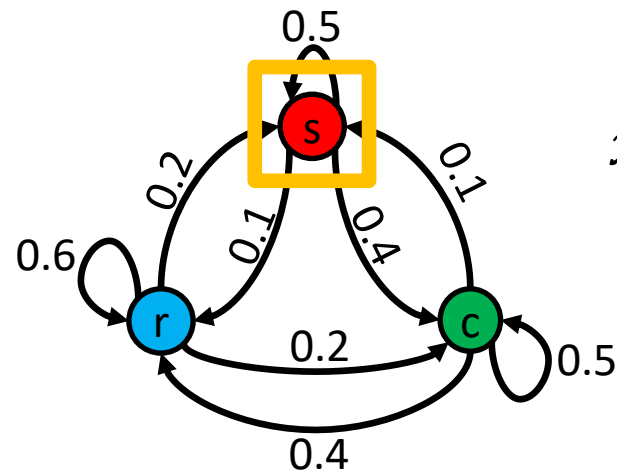
$$x_{1:T} = s \ r \ c \ r \ r \ c \ s$$

$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – As a Random Walk

- Time-homogeneous discrete MCs can be interpreted as state machines
- Example: a model for weather condition
  - $X_t \in \{\text{rainy}, \text{sunny}, \text{cloudy}\}$  weather condition on  $t$ -th day
  - We can think of a sequence (i.e. a sample from the MC) as a random walk.



$x_{1:T} = s \ r \ c \ r \ r \ c$

$s$

$$A = \begin{matrix} & \begin{matrix} \text{rainy} & \text{sunny} & \text{cloudy} \end{matrix} \\ \begin{matrix} \text{rainy} \\ \text{sunny} \\ \text{cloudy} \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}$$

$$P(X_{1:T} = x_{1:T}) = P(X_1 = s) \times 0.1 \times 0.2 \times 0.4 \times 0.6 \times 0.2 \times 0.1$$

# Markov Chain – Learning of Model Parameters

- Given a set  $\{X_{1:T_n}^{(n)}\}$  of  $N$  observed sequences, we can learn  $\boldsymbol{\pi}$  and  $\mathbf{A}$  using maximum-likelihood.

$$L(k) = \#(X_1 = k)$$

$$N(i, j) = \#(X_t = i, X_{t+1} = j)$$

$i \Rightarrow j$

$$P(all) = \prod_{n=1}^N P(X_1^{(n)}) \prod_{t=1}^{T_n-1} \Pr(X_{t+1}^{(n)} | X_t^{(n)}) = \left( \prod_{k=1}^K \pi_k^{L(k)} \right) \left( \prod_{i=1}^K \prod_{j=1}^K A_{ij}^{N(i,j)} \right)$$

$$\Rightarrow \log P(all) = \sum_{k=1}^K L(k) \log(\pi_k) + \sum_{i=1}^K \sum_{j=1}^K N(i, j) \log(A_{ij})$$

- Maximizing  $\log P(all)$  subject to  $\sum_k \pi_k = 1$  and  $\sum_j A_{ij} = 1$ , we get:

$$A_{ij} = \frac{N(i, j)}{\sum_{j'} N(i, j')} \quad \pi_k = \frac{L(k)}{\sum_{k'} L(k')}$$

# Markov Chain – More Insights

- Task 1: Determine  $A_{ij}(n) = P(X_{t+n} = j | X_t = i)$ 
  - In words,  $A_{ij}(n)$  = probability of getting from state  $i$  to state  $j$  in  $n$  steps
- How to compute  $A_{ij}(n)$  ?

$$\begin{aligned}
 P(X_{t+n} = j | X_t = i) &= \sum_{k=1}^K P(X_{t+n} = j, \underbrace{X_{t+n-1} = k}_{\text{n jump}} | X_t = i) \\
 &= \sum_{k=1}^K P(X_{t+n} = j | X_{t+n-1} = k, \underbrace{X_t = i}_{\text{red}}) P(X_{t+n-1} = k | X_t = i) \\
 &= \sum_{k=1}^K P(X_{t+n} = j | X_{t+n-1} = k) P(X_{t+n-1} = k | X_t = i) = \sum_{k=1}^K A_{kj} A_{ik}(n-1)
 \end{aligned}$$

$$\Rightarrow A(n) = A(n-1)A \xrightarrow{A(1) = A}$$

$$A(n) = A^n$$

$$A^{m+n} = A^m \cdot A^n$$

- Chapman-Kolmogorov equations:

$$A_{ij}(m+n) = \sum_{k=1}^K A_{ik}(m) A_{kj}(n) \Rightarrow$$

$$A(m+n) = A(m)A(n)$$

# Markov Chain – More Insights

- Task 2: Determine  $\pi_j(t) = \Pr(X_t = j)$ 
  - In words,  $\pi_j(t)$  = probability of reaching state  $j$  in step  $t$ .
- How to compute  $\pi_j(t)$  ?

$$\Pr(X_t = j) = \sum_{i=1}^K \Pr(X_t = j | X_{t-1} = i) \Pr(X_{t-1} = i) = \sum_{i=1}^K A_{ij} \pi_i(t-1)$$

$$\Rightarrow \underline{\pi(t) = \pi(t-1)A}$$

$$\pi(t-1) = \pi(t-2) \cdot A$$

$$\Rightarrow \pi(t) = \pi A^{(t-1)}$$

$\pi(t)$  and  $\pi$  are row vectors

# Questions – MC

1. We assume that  $X_t \in \{1, 2, 3\}$ . We consider  $\pi = \begin{bmatrix} 0.0 \\ 0.5 \\ 0.5 \end{bmatrix}$  and  $A = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}$ .

- What is the probability to observe the sequence  $\mathbf{X}^{(1)} = [1, 2, 3]$ ? 0
- What is the probability to observe the sequence  $\mathbf{X}^{(2)} = [2, 2, 3]$ ?  $0.5 \cdot 0.5 \cdot 0.4$

2. We assume that  $X_t \in \{1, 2, 3\}$  and we observed three sequences:

- $\mathbf{X}^{(1)} = [1, 3, 2]$
- $\mathbf{X}^{(2)} = [3]$
- $\mathbf{X}^{(3)} = [1, 1, 3, 2]$

$2 \rightarrow \frac{1}{3}$

$$A = \begin{pmatrix} \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \end{pmatrix}$$

What is the MLE of the transition matrix  $A \in \mathbb{R}^{3 \times 3}$ ?



# Reading Material

---

- [1] Pattern Recognition and Machine Learning, section 13.1:  
<https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>

# Machine Learning for Graphs and Sequential Data

## *Sequential Data – Markov Chains*

Lecturer: Prof. Dr. Stephan Günnemann  
[cs.cit.tum.de/daml](https://cs.cit.tum.de/daml)

---

Summer Term 2023

Data Analytics and  
Machine Learning 