Course 495: Advanced Statistical Machine Learning/Pattern Recognition

- Lecturer: Stefanos Zafeiriou

- Goal (Lectures): To present discrete and continuous valued probabilistic linear dynamical systems (HMMs & Kalman Filters).

- Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the models.

Materials

- Chapter 13: Pattern Recognition & Machine Learning, Christopher M. Bishop.
- Chapter 17: Machine Learning a Probabilistic Perspective, Kevin Murphy
- Rabiner, Lawrence. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE* 77.2 (1989): 257-286.

Linear Dynamical Systems

Applications of probabilistic linear dynamical systems

- Language modelling
- Object/Face tracking
- Speech/Gesture recognition
- Finance
- Bioinformatics

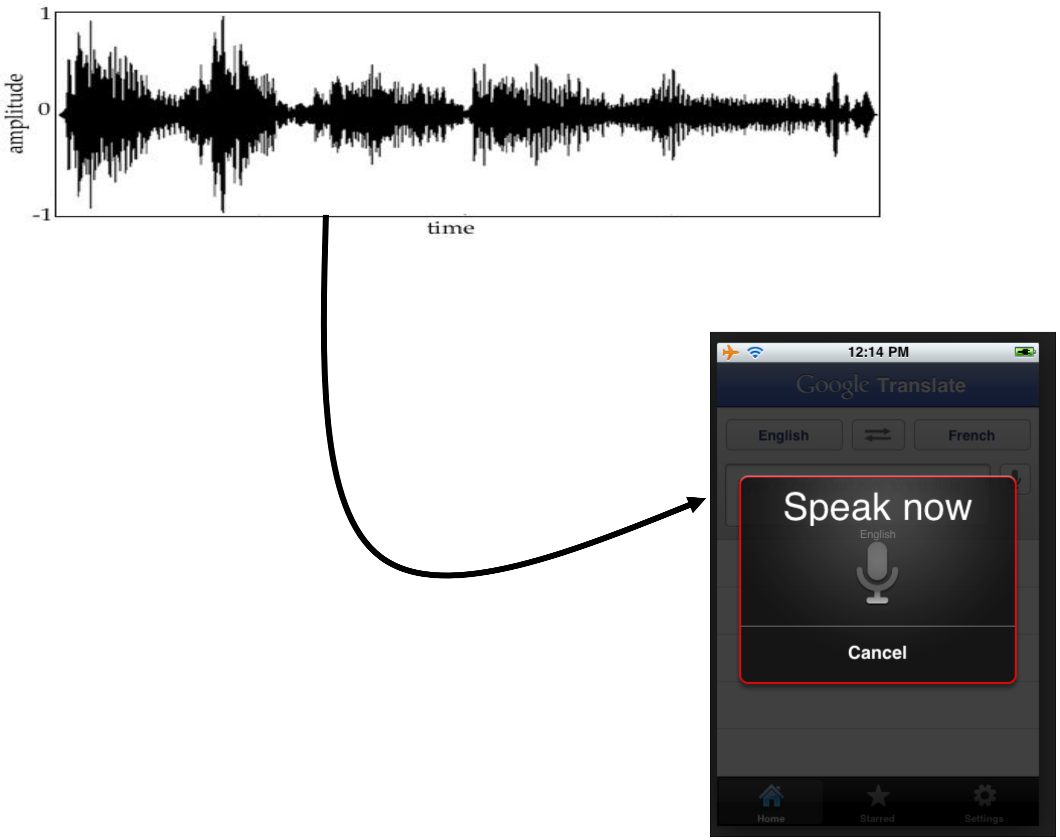
Applications

Object-target tracking

Applications

Speech Recognition (voice Google search)

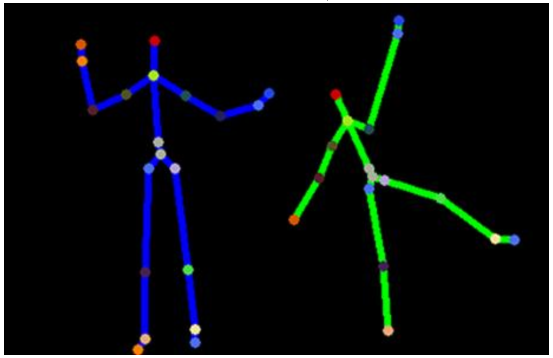
Waveform



Hello world

Applications

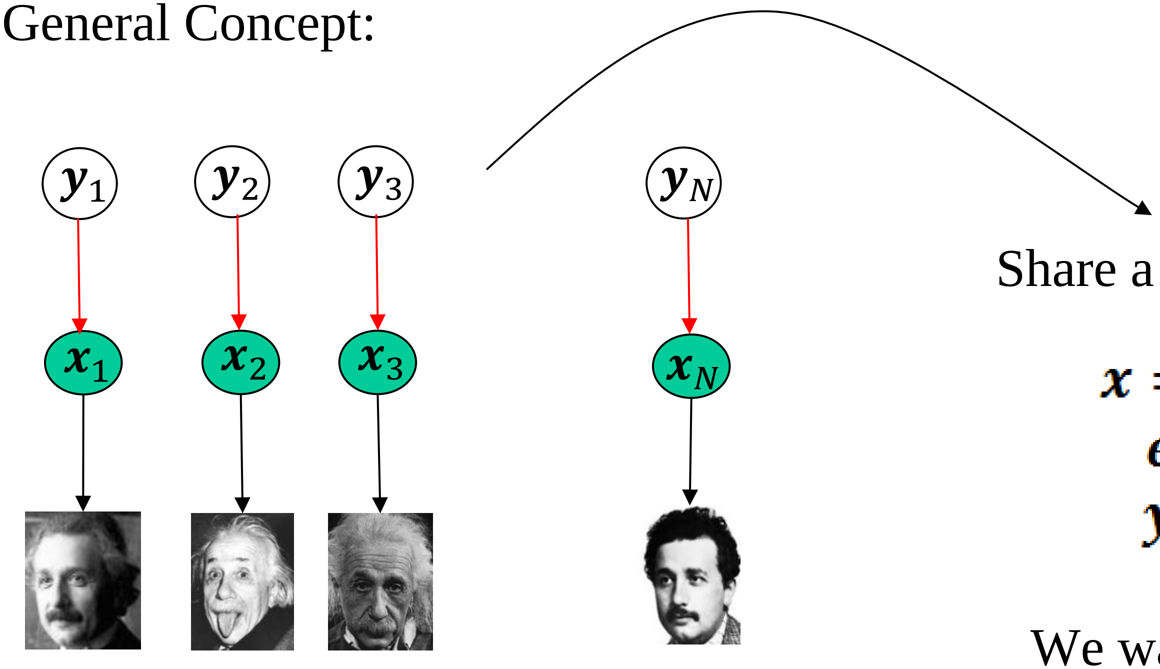
Gesture recognition (Kinect games)



Gestures

Latent Variable Models (Static)

General Concept:



Share a common linear structure

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{e}|\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$
$$\boldsymbol{y} \sim N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I})$$

We want to find the parameters:

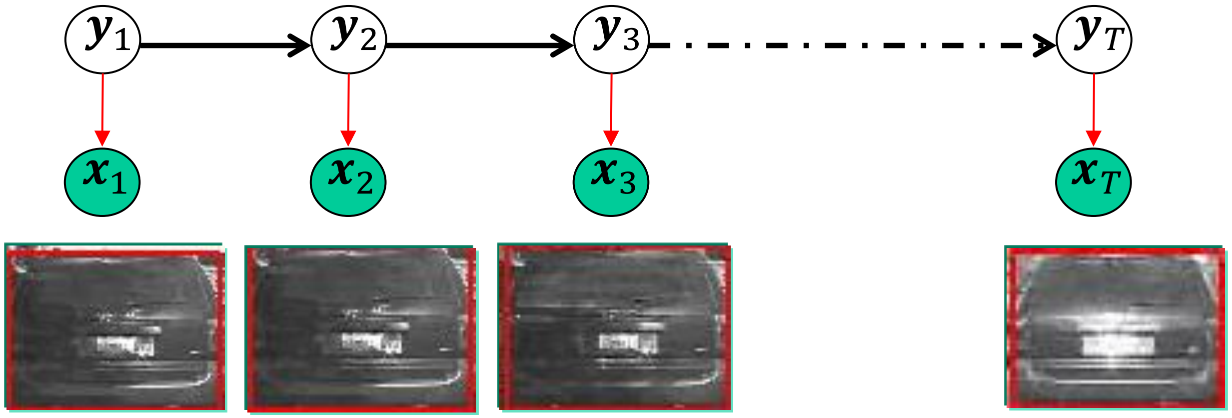
$$\theta = \{\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2\}$$

Joint likelihood maximization:

$$\mathrm{p}(\boldsymbol{x}_1, .., \boldsymbol{x}_N, \boldsymbol{y}_1, .., \boldsymbol{y}_N|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{W}, \boldsymbol{\mu}, \sigma) \prod_{i=1}^{N} \boldsymbol{p}(\boldsymbol{y}_i)$$

Latent Variable Models (Dynamic, Continuous)

Latent Variable Models (Dynamic, Continuous)



Generative Model

$$\boldsymbol{x}_n = \mathbf{W}\boldsymbol{y}_n + \boldsymbol{e}_n$$

$$\boldsymbol{y}_1 = \boldsymbol{\mu}_0 + \boldsymbol{u}$$

$$\boldsymbol{y}_n = \boldsymbol{A}\boldsymbol{y}_{n-1} + \boldsymbol{v}_n$$

Noise distribution

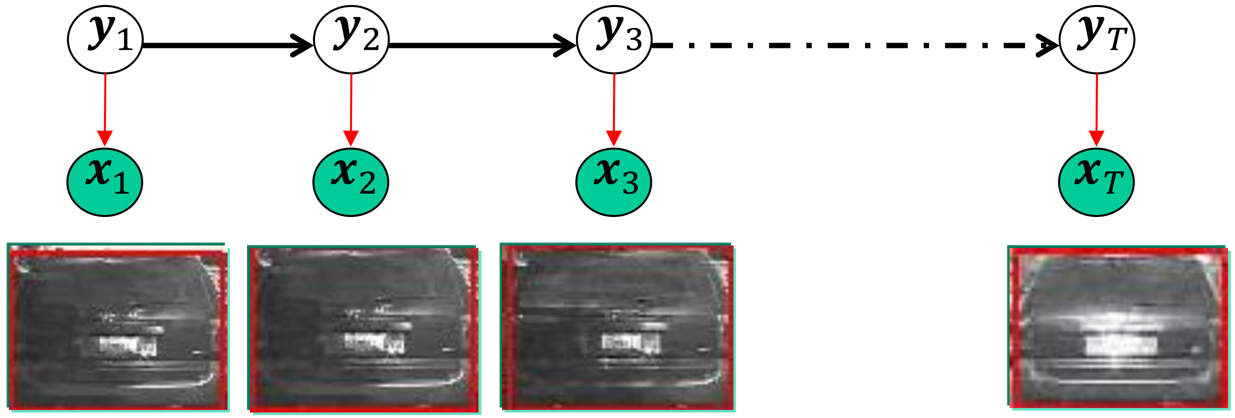
$$\mathbf{e} \sim N(\boldsymbol{e}|\mathbf{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{u} \sim N(\boldsymbol{u}|\mathbf{0}, \boldsymbol{P}_0)$$

$$\boldsymbol{v} \sim N(\boldsymbol{v}|\mathbf{0}, \boldsymbol{\Gamma})$$

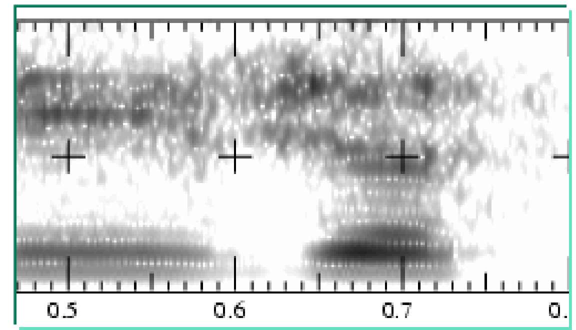
Parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \boldsymbol{P}_0\}$

Latent Variable Models (Dynamic, Continuous)



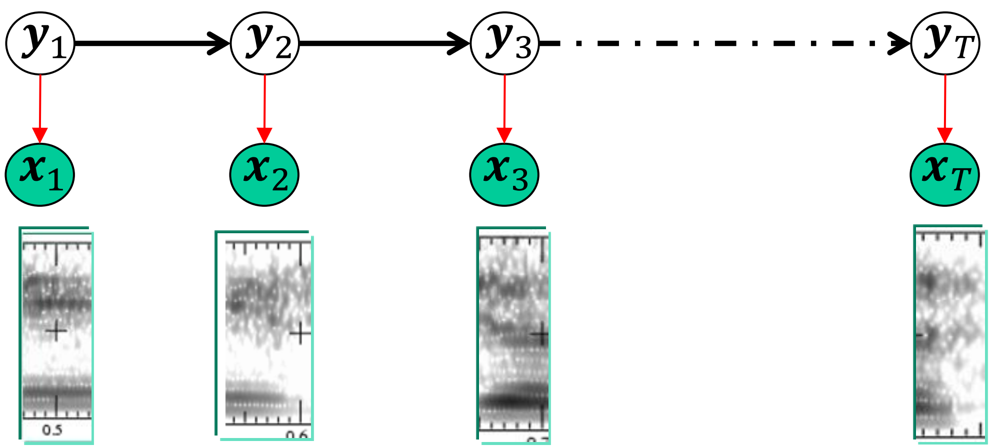
Markov Property: $p(\boldsymbol{y}_i, |\boldsymbol{y}_1, .., \boldsymbol{y}_{i-1}) = p(\boldsymbol{y}_i|\boldsymbol{y}_{i-1})$

Latent Variable Models (Dynamic, Discrete)



Word: need

Phonemes: n iy d

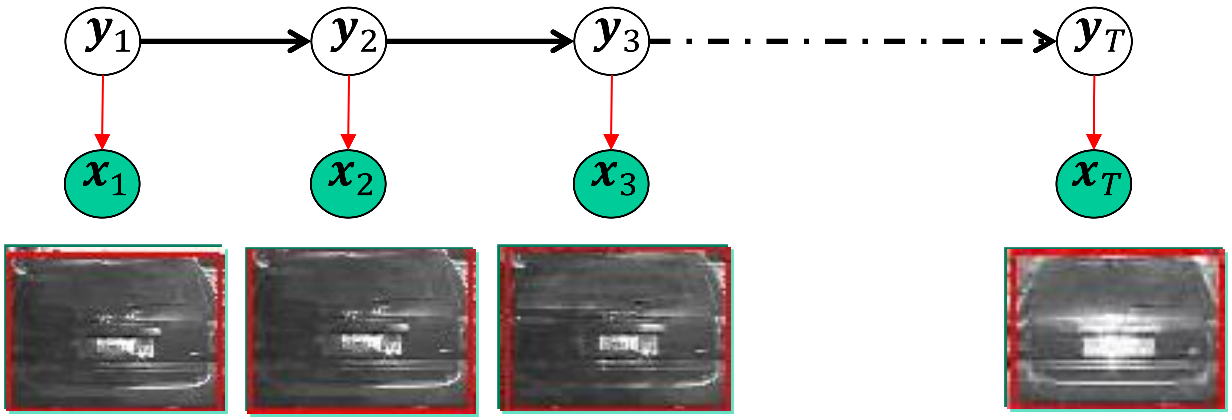


Latent structure takes discrete values:

$y_t \in \{\text{start}, \text{n}, \text{iy}, \text{d}, \text{end}\}$

Summarize what we will study?

Sequential data (2 weeks):



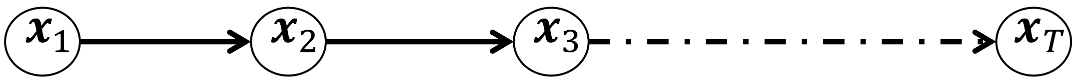
What are the models?:

- *The Markov & Hidden Markov Models (1 week).*
- *The Kalman Filter (1 week).*

What we will learn?:

- *How to formulate probabilistically the problems and learn parameters.*

Markov Chains with Discrete Random Variables



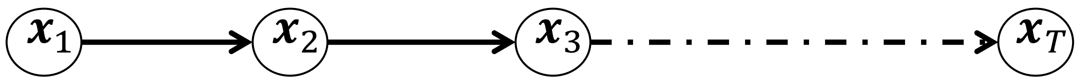
Let’s assume we have discrete random variables (e.g., taking 3 discrete values $\boldsymbol{x}_t = \{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}\}$)

Markov Property: $p(\boldsymbol{x}_t|\boldsymbol{x}_1, .., \boldsymbol{x}_{t-1}) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$

e.g. $p(\boldsymbol{x}_t = \begin{bmatrix}1\\0\\0\end{bmatrix} | \boldsymbol{x}_{i-1} = \begin{bmatrix}0\\1\\0\end{bmatrix})$

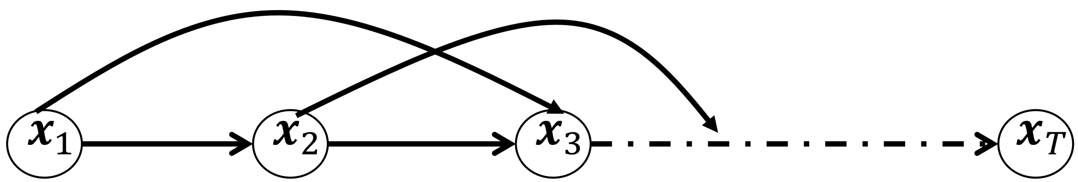
Stationary, Homogeneous or Time-Invariant if the distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ does not depend on t

Markov Chains with Discrete Random Variables



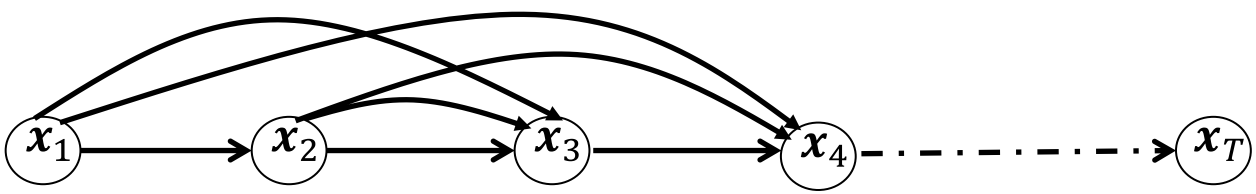
$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$

bigram model



$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_{t-2})$

Tri-gram model



$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_{t-2},\boldsymbol{x}_{t-3})$

4-gram model

Markov Chains with Discrete Random Variables

Joint distribution in the first order case:

$$p(\boldsymbol{x}_1,..,\boldsymbol{x}_T) = p(\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_2,..,\boldsymbol{x}_T|\boldsymbol{x}_{\mathbf{1}})$$

$$= p(\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_2|\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_3,..,\boldsymbol{x}_T|\boldsymbol{x}_{\mathbf{1}},\boldsymbol{x}_2)$$

$$= p(\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_2|\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_3,..,\boldsymbol{x}_T|\boldsymbol{x}_2)$$

$$= p(\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_2|\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_3|\boldsymbol{x}_{\mathbf{2}})p(\boldsymbol{x}_4,..,\boldsymbol{x}_T|\boldsymbol{x}_2,\boldsymbol{x}_3)$$

$$= p(\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_2|\boldsymbol{x}_{\mathbf{1}})p(\boldsymbol{x}_3|\boldsymbol{x}_{\mathbf{2}})p(\boldsymbol{x}_4,..,\boldsymbol{x}_T|\boldsymbol{x}_3)$$

$$= p(\boldsymbol{x}_{\mathbf{1}})\prod_{i=2}^{\mathrm{T}} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$$

First Order Markov Chains

$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ can be represented as a KxK transition matrix $\boldsymbol{A} = [a_{ij}]$

which is the probability of going from state i to state j

$p(\boldsymbol{x}_t \vert \boldsymbol{x}_{t-1})$		$\boldsymbol{x}_t$ = 1	2	3
$\boldsymbol{x}_{t-1}$	1	a_{11}	a_{12}	a_{13}
	2	a_{21}	a_{22}	a_{23}
	3	a_{31}	a_{32}	a_{33}

$\boldsymbol{A}$ is a stochastic matrix, i.e.,

$$\sum_{k=1}^{3} a_{ik} = 1$$

First Order Markov Chains

If we make use of our vector notation of discrete random variable

then if

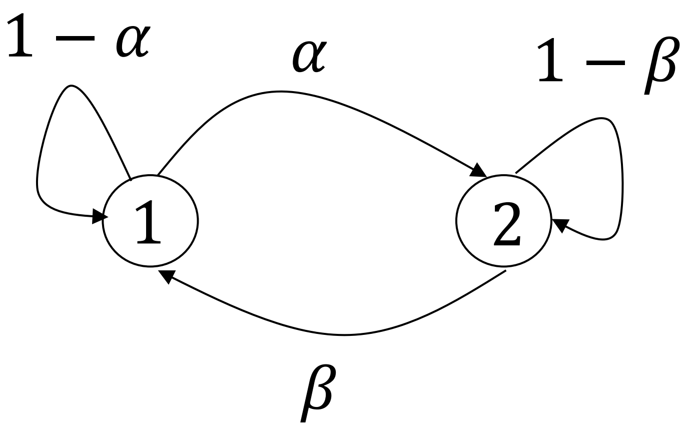
$$\boldsymbol{x}_t = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$ has only its j-th element "activated"

$$\boldsymbol{x}_{t-1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$ has only its i-th element "activated"

then $a_{ij} = p(x_{tj} = 1 \mid x_{t-1i} = 1)$

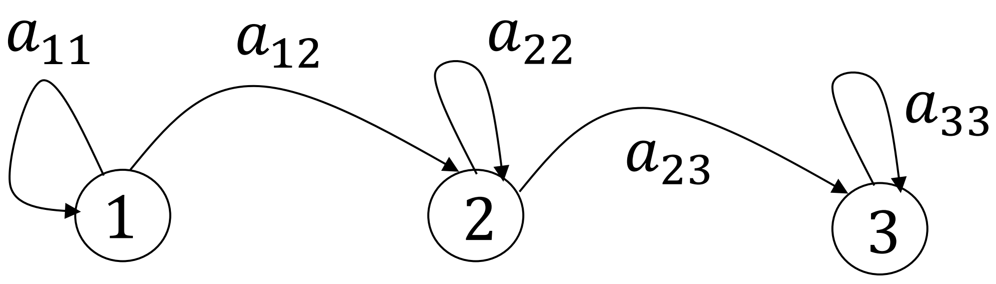
Transition Matrices

A stationary finite-state Markov chain is equivalent to a stochastic automaton.



$$\boldsymbol{A} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

Transition Matrices



$$A = \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{aligned} a_{12} &= 1 - a_{11} \\ a_{23} &= 1 - a_{22} \end{aligned}$$

Transition Matrices

- Transition matrix $\boldsymbol{A}$ specifies the probability of getting from i to j in one step.

- How can we compute the probability of i to j in exactly n-steps?

$$a_{ij}(n) = p\left(x_{t+nj} = 1 \middle| x_{ti} = 1\right)$$

Probability of getting from i to k in one step and then from k to j in $n-1$ steps and summing for all k

$$= \sum_{k=1}^{K} p(x_{t+1k} = 1|x_{ti} = 1) p\left(x_{t+nj} = 1 \middle| x_{t+1k} = 1\right)$$

$$= \sum_{k=1}^{K} a_{ik} a_{kj}(n-1)$$

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

$$\Rightarrow \boldsymbol{A}(n) = \boldsymbol{A}^n$$

Stationary Distribution of the Markov Chain

- Markov model are used to define joint probability distributions over sequences.
- But can be also interpreted as stochastic dynamical systems, where we “hop” from one state to another over time.
- We are interested long term distribution over states, known as stationary distribution of the chain.
- Important application: Google’s Page Rank

Stationary Distribution of the Markov Chain

Assume a Markov Chain.

$$\boldsymbol{A} = [a_{ij}] = [p(x_{tj} = 1 \mid x_{t-1i} = 1)\ ]$$

$$\boldsymbol{\pi}_0 = [\boldsymbol{p}(x_{0i} = 1)]$$

then

$$\boldsymbol{p}(x_{1i} = 1) = \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} p(x_{1i} = 1, x_{0k} = 1)$$

$$= \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} p(x_{0k} = 1) p(x_{1i} = 1 | x_{0k} = 1)$$

$$\Rightarrow \pi_{1\mathrm{j}} = \sum_{k=1}^{K} \pi_{0k}\, a_{kj} \Rightarrow \boldsymbol{\pi}_1{}^T = \boldsymbol{\pi}_0{}^T \boldsymbol{A}$$

Stationary Distribution of the Markov Chain

- We image iterating these equations. If we ever reach a stage where:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \boldsymbol{A}$$

we have reached the stationary distribution (also called the invariant distribution or equilibrium distribution)

- In case of three states the above is written:

$$(\pi_1 \pi_2 \pi_3) = (\pi_1 \pi_2 \pi_3) \begin{pmatrix} 1 - a_{12} - a_{13} & a_{12} & a_{13} \\ a_{21} & 1 - a_{21} - a_{23} & a_{23} \\ a_{31} & a_{32} & 1 - a_{31} - a_{32} \end{pmatrix}$$

Stationary Distribution of the Markov Chain

so $\pi_1 = \pi_1(1 - a_{12} - a_{13}) + \pi_2 a_{21} + \pi_3 a_{31}$

or $\pi_1(a_{12} + a_{13}) = \pi_2 a_{21} + \pi_3 a_{31}$

similarly $\pi_2(a_{21} + a_{23}) = \pi_1 a_{12} + \pi_3 a_{13}$

and $\pi_3(a_{31} + a_{32}) = \pi_1 a_{31} + \pi_2 a_{32}$

In general, we have $\pi_i \sum_{j \neq i} a_{ij} = \sum_{j \neq i} \pi_j a_{ji}$ and $\sum_j \pi_j = 1$

The probability of being in state i times the net flow out of the state i must equal the probability of being in each other state j times the net flow from that state into i.

Stationary Distribution of the Markov Chain

$\boldsymbol{A}^T\boldsymbol{\pi} = \boldsymbol{\pi}$ looks like an eigen-analysis problem

i.e., $\boldsymbol{\pi}$ is an eigenvector with eigenvalue 1

Such an eigenvector always exists since $\boldsymbol{A}$ is row-stochastic $\boldsymbol{A}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{A}$ and $\boldsymbol{A}^T$ have the same eigenvalues

But the eigenvectors of $\boldsymbol{A}$ are real-valued only when $a_{ij} > 0$

What happens in the case that a_{ij}=0?

Stationary Distribution of the Markov Chain

$\boldsymbol{\pi}^T(\boldsymbol{I}-\boldsymbol{A})=\boldsymbol{0} \Rightarrow K$ constraints

$\boldsymbol{\pi}^T\boldsymbol{1}=\boldsymbol{1} \Rightarrow$ 1 extra constraint

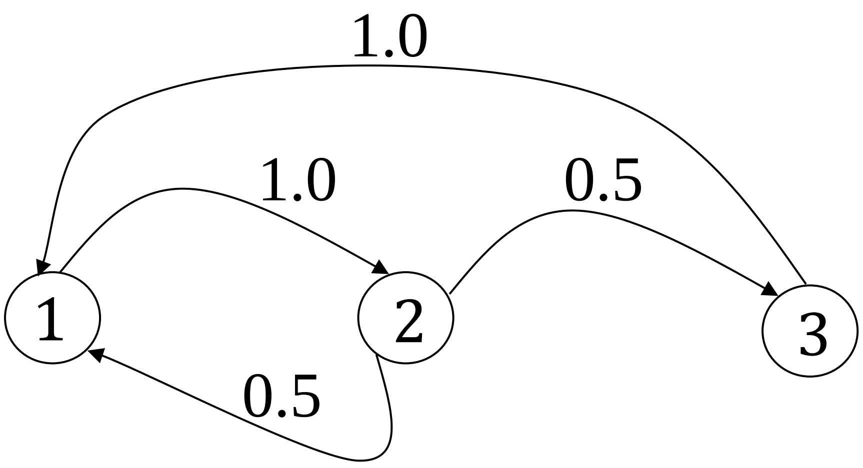
⇒Problem is over constrained

Define matrix $\boldsymbol{M}=\boldsymbol{I}-\boldsymbol{A}$

and replace one column with 1s

$$(\pi_1 \ \pi_2 \ \pi_3)\begin{pmatrix} 1-a_{11} & -a_{12} & 1 \\ -a_{21} & 1-a_{22} & 1 \\ -a_{31} & -a_{32} & 1 \end{pmatrix} = (0\ 0\ 1)$$

Stationary Distribution of the Markov Chain

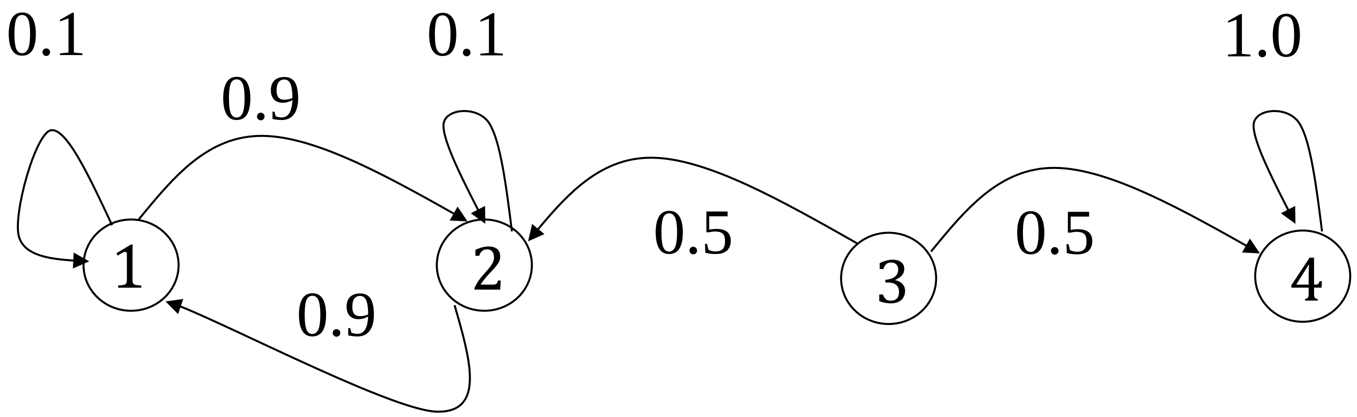


$$(\pi_1\ \pi_2\ \pi_3)\begin{pmatrix} 1 & -1 & 1 \\ -0.5 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = (0\ 0\ 1)$$

$$(\pi_1\ \pi_2\ \pi_3) = (0.4\ 0.4\ 0.2)$$

Stationary Distribution of the Markov Chain

- When does a stationary distribution exists

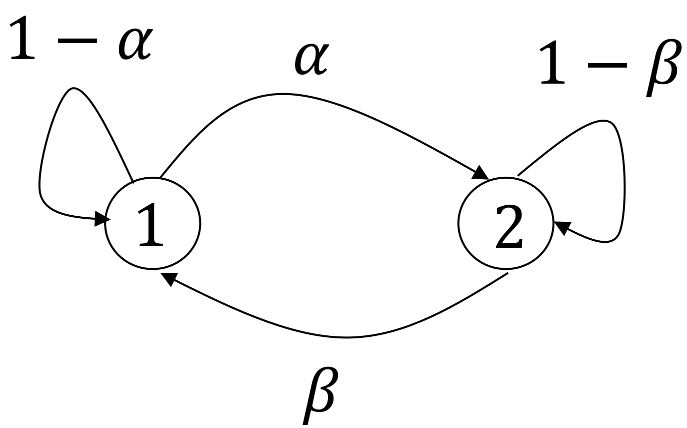


State 4 is an absorbing state hence $\boldsymbol{\pi} = (0,0,0,1)$ is a possible stationary distribution

so is $\boldsymbol{\pi} = (0.5,0.5,0,0)$

Stationary Distribution of the Markov Chain

- First necessary condition to have a unique stationary distribution is that the state transition diagram be a singly connected component.
- Such chains are called irreducible (i.e., you can go from any state to any other state).



$$\alpha = \beta = 1$$

Let's start from state 2

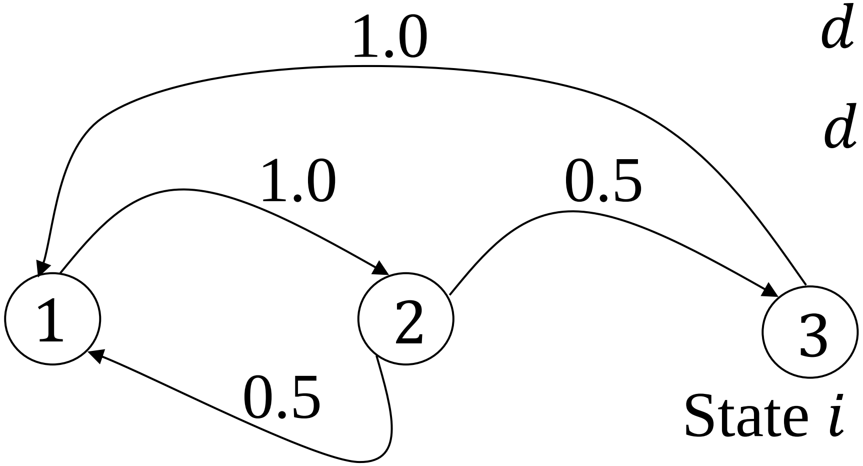
$t = 2b + 1$ $\quad state\ 1$

$t = 2b$ $\quad state\ 2$

$\Rightarrow$oscillates

Stationary Distribution of the Markov Chain

$$d(i) = \gcd\{t: a_{ii}(t) > 0\}$$



$$d(1) = \gcd\{2,3,4,6,..\} = 1$$

$$d(2) = \gcd\{2,3,4,6,..\} = 1$$

$$d(3) = \gcd\{3,5,6,..\} = 1$$

State i is aperiodic if $d(i) = 1$

Markov Chain is aperiodic if $d(i) = 1$ for all i

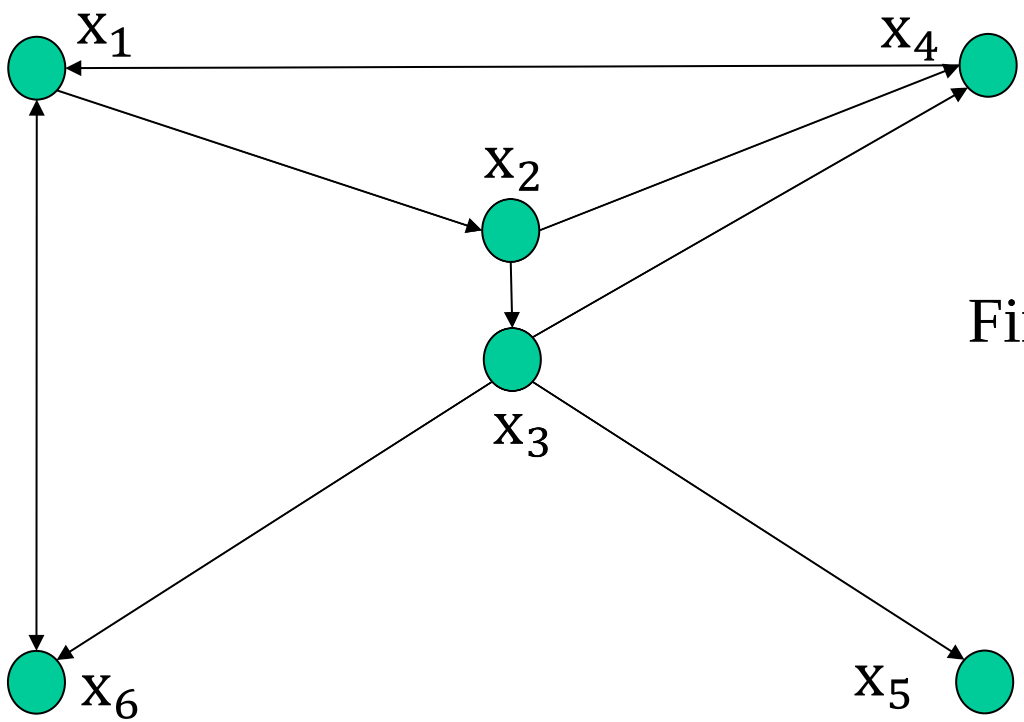
Stationary Distribution of the Markov Chain

- Every irreducible (singly connected), aperiodic finite state Markov chain has a limiting distribution, which is equal to $\boldsymbol{\pi}$, its unique stationary distribution.

- Special cases and sufficient conditions: Every regular finite state chain has a unique stationary distribution (i.e., $a_{ij}(t) > 0$).

Stationary Distribution of the Markov Chain

Small web (uniform distribution over all states it is connected to)



First step make it regular.

$$(\pi_1\ \pi_2\ \pi_3\ \pi_4\ \pi_5\ \pi_6) = (0.32\ 0.17\ 0.1\ 0.137\ 0.064\ 0.2)$$

Markov Chain for Language Modelling.

- One important application of Markov Models is to make statistical language models (i.e., probability distributions over sequences of words).
- Sentence Completion. Predict next word based on the previous one.
- Data compression. Any density model can be used to define an encoding scheme, by assigning short code-words to more probably strings.
- Text classification. Any density model can be used as a class-conditional density.
- Automatic essay writing. Sample from $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$

Simple Parameter Estimation

abbbcbbabcbbbabc $p(x_1{}^1, \ldots, x_T{}^1)$

bbcabbababcbbbaba $p(x_1{}^2, \ldots, x_T{}^2)$

$\vdots$ $\vdots$

abccabbabbcbbbab $p(x_1{}^N, \ldots, x_T{}^N)$

a b c

$$\boldsymbol{x} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

$$p(\boldsymbol{x}_1|\boldsymbol{\pi}) = \prod_{k=1}^{3} \pi_\kappa{}^{x_{1k}} \qquad p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}{}^{x_{t-1j}x_{tk}}$$

Maximum Likelihood for Markov Chains

- What are the parameters in this case?

$$\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}\}$$

- The problem is now formulated as:

Given a set of observations $D_l = \{x_1{}^l, .., x_T{}^l\}, l = 1, .., N$
find the parameters θ that maximize $p(D_1, .., D_N | \theta)$

$$p(D_1, .., D_N | \theta) = \prod_{l=1}^{N} p(D_l | \theta)$$

Maximum Likelihood for Markov Chains

$$p(D_l|\theta) = p(x_1{}^l, \dots, x_T{}^l|\theta) = p(\boldsymbol{x_1}{}^{\boldsymbol{l}}) \prod_{t=2}^{\mathrm{T}} p(\boldsymbol{x_t}{}^{\boldsymbol{l}}|\boldsymbol{x_{t-1}}{}^{\boldsymbol{l}})$$

$$= \prod_{k=1}^{3} \pi_\kappa{}^{x_{1k}{}^l} \prod_{t=2}^{\mathrm{T}} \prod_{j=1}^{3} \prod_{k=1}^{3} a_{jk}{}^{x_{t-1j}{}^l x_{tk}{}^l}$$

$$\Rightarrow p(D_1, \dots, D_N|\theta) = \prod_{l=1}^{N} \prod_{k=1}^{3} \pi_\kappa{}^{x_{1k}{}^l} \prod_{t=2}^{\mathrm{T}} \prod_{j=1}^{K} \prod_{k=1}^{K} a_{jk}{}^{x_{t-1j}{}^l x_{tk}{}^l}$$

$$\underset{ln}{\Rightarrow} \ln p(\theta) = \sum_{l=1}^{N} \sum_{k=1}^{3} x_{1k}{}^l \ln \pi_\kappa + \sum_{l=1}^{N} \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} x_{t-1j}{}^l x_{tk}{}^l \ln a_{jk}$$

Maximum Likelihood for Markov Chains

$$= \sum_{l=1}^{N} \sum_{k=1}^{3} x_{1k}{}^{l} \ln \pi_{\kappa} + \sum_{l=1}^{N} \sum_{t=2}^{T} \sum_{j=1}^{3} \sum_{k=1}^{3} x_{t-1j}{}^{l} x_{tk}{}^{l} \ln a_{jk}$$

$$= \sum_{k=1}^{3} \left(\sum_{l=1}^{N} x_{1k}{}^{l} \right) \ln \pi_{\kappa} + \sum_{j=1}^{3} \sum_{k=1}^{3} \left(\sum_{l=1}^{N} \sum_{t=2}^{T} x_{t-1j}{}^{l} x_{tk}{}^{l} \right) \ln a_{jk}$$

Let us define the counts

$$N_k{}^{1} \triangleq \sum_{l=1}^{N} x_{1k}{}^{l} \qquad N_{jk} = \sum_{l=1}^{N} \sum_{t=2}^{T} x_{t-1j}{}^{l} x_{tk}{}^{l}$$

Maximum Likelihood for Markov Chains

$$= \sum_{k=1}^{3} N_k{}^1 \ln \pi_\kappa + \sum_{j=1}^{3} \sum_{k=1}^{3} N_{jk} \ln a_{jk}$$

Solve the above subject to: $\sum_{k=1}^{3} \pi_\kappa = 1 \quad \sum_{k=1}^{3} a_{jk} = 1$

The Lagrangian is:

$$L(\boldsymbol{\pi}, \boldsymbol{A}) = \sum_{k=1}^{3} N_k{}^1 \ln \pi_\kappa + \sum_{j=1}^{3} \sum_{k=1}^{3} N_{jk} \ln a_{jk} \, 0$$

$$-\lambda \left(\sum_{k=1}^{3} \pi_\kappa - 1 \right) - \gamma \left(\sum_{k=1}^{3} a_{jk} - 1 \right)$$

Maximum Likelihood for Markov Chains

which gives us:

$$\pi_k = \frac{{N_k}^1}{\sum_{k=1}^{3} {N_k}^1} \qquad\qquad a_{jk} = \frac{N_{jk}}{\sum_{k=1}^{3} N_{jk}}$$