CS6510
Applied Machine Learning

Time Series Analysis, HMMs

11+25 Nov 2017

Vineeth N Balasubramanian

आई आई टी हैदराबाद
IIT Hyderabad

Time Series Analysis

- Time series: A collection of observations made sequentially in time
- Many application fields:
  - Monthly closings of the stock exchange index
  - Malaria incidence or deaths over calendar years
  - Daily maximum temperatures
  - Hourly records of babies born at a maternity hospital
- Characteristics:
  - successive observations are NOT independent (i.i.d. assumption does not hold)
  - The order of observation is crucial

Modeling Time Series

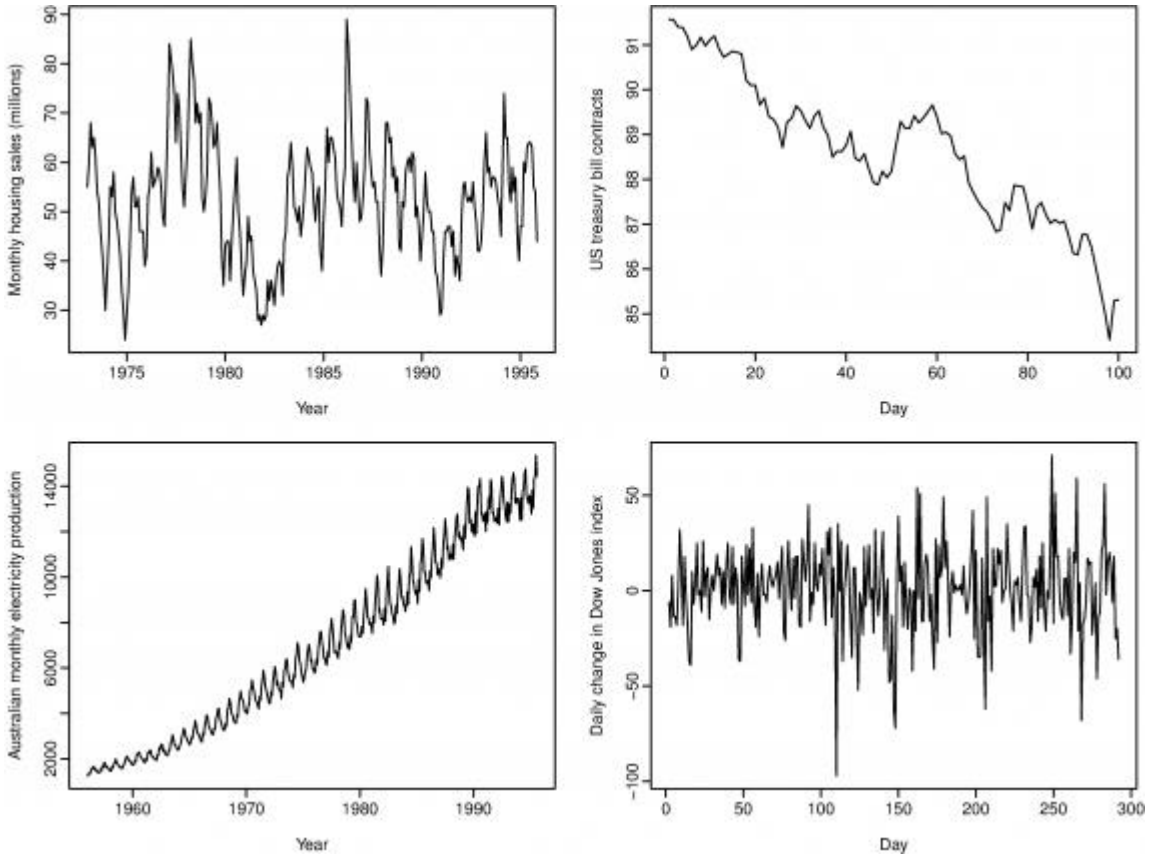
- Time series: data are correlated; data are realizations of stochastic processes
- Standard machine learning methods are often difficult to directly apply
  - Do not exploit temporal correlations
  - Computation & storage requirements typically scale poorly to realistic applications
- Stochastic linear discrete input-output models
- Often, assumption of stationarity (the mean and variance of the process generating the data do not change over time)
- Key difference: no causal variables

Time Series Analysis: Categorization

- In terms of variables assessed
  - **Univariate**: Analysis of a single sequence of data
  - **Multivariate**: Analysis of several sets of data for the same sequence of time periods
- In terms of assumptions
  - Stationary
  - Non-stationary
  - Weakly stationary

Kinds of Time Series Data

- **Trends**
  - Gradual, long-term movement (up or down) of demand.
  - Easiest to detect
- **Cyclical**
  - An up-and-down repetitive movement in demand.
  - repeats itself over a long period of time
- **Seasonal**
  - An up-and-down repetitive movement within a trend occurring periodically.
  - Often weather related but could be daily or weekly occurrence
- **Random/Irregular**
  - Erratic movements that are not predictable because they do not follow a pattern



Approaching Time Series Analysis

- There are many, many different time series techniques
- It is usually impossible to know which technique will be best for a particular data set
- It is customary to try out several different techniques and select the one that seems to work best
- To be an effective time series modeler, you need to keep several time series techniques in your "tool box"

Performance Metrics

- Mean absolute deviation: $\mathrm{MAD} = \sum_{i=1}^{n} \frac{\left|Y_i - \hat{Y}_i\right|}{n}$
- Mean absolute percent error: $\mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{\left|Y_i - \hat{Y}_i\right|}{Y_i}$
- Mean square error: $\mathrm{MSE} = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{n}$
- Root mean square error: $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$

Notations

- Observations may be denoted by

$$Y_1, Y_2, Y_3, \ldots \quad Y_t, \ldots, Y_T$$

observation at time t

since data are usually collected at discrete points in time

- The interval between observations can be any time interval (hours within days, days, weeks, months, years, etc).

Different Time Series Processes

- **White Noise Process**
- A series is called white noise if it is purely random in nature.
- Let $\{\epsilon_t\}$ denote such a series, such that it has:
  - Zero mean: $E(\epsilon_t) = 0$
  - Constant variance: $V(\epsilon_t) = \sigma^2$
  - Uncorrelated: $E(\epsilon_t \epsilon_s) = 0$
- Scatter plot of such a series across time will indicate no pattern, and hence, forecasting the future values of such a series is not possible.

Autoregressive (AR) Model

- AR(p) is a regression model that regresses each point on the previous p time points
- Thus: $Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots \epsilon_t)$
- Common representation of AR model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots.. + \beta_p Y_{t-p} + \epsilon_t$$

- Can be learned with linear estimation algorithm

Moving Average (MA) Model

- In the Moving Average model (MA(p)), Yt depends only on the random error terms which follow a white noise process, i.e.

$$Y_t = f(\epsilon_t , \epsilon_{t-1} , \epsilon_{t-2} , \epsilon_{t-3} , \ldots)$$

- It propagates over time the effect of the random fluctuations
- Common representation of MA(q) model:

$$Y_t = \beta_0 + \epsilon_t + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_3\epsilon_{t-3} + \ldots + \phi_q\epsilon_{t-q}$$

- An iterative estimation process is needed

ARMA Model

- AutoRegressive Moving Average model (ARMA)
- Data may sometimes require modeling including both kinds of values
- General form of ARMA(p,q) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots.. + \beta_p Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \phi_3 \epsilon_{t-3} + \ldots. + \phi_q \epsilon_{t-q}$$

Stationarity

- A series is said to be **strictly stationary** if the marginal distribution of Y at time t, $p(Y_t)$, is the same as at any other point in time.
- Therefore, $p(Y_t) = p(Y_{t+k})$, and $p(Y_t, Y_{t+k})$ does not depend on t.
- Another way of saying this: mean, variance and covariance of series $\{Y_t\}$ are time-invariant

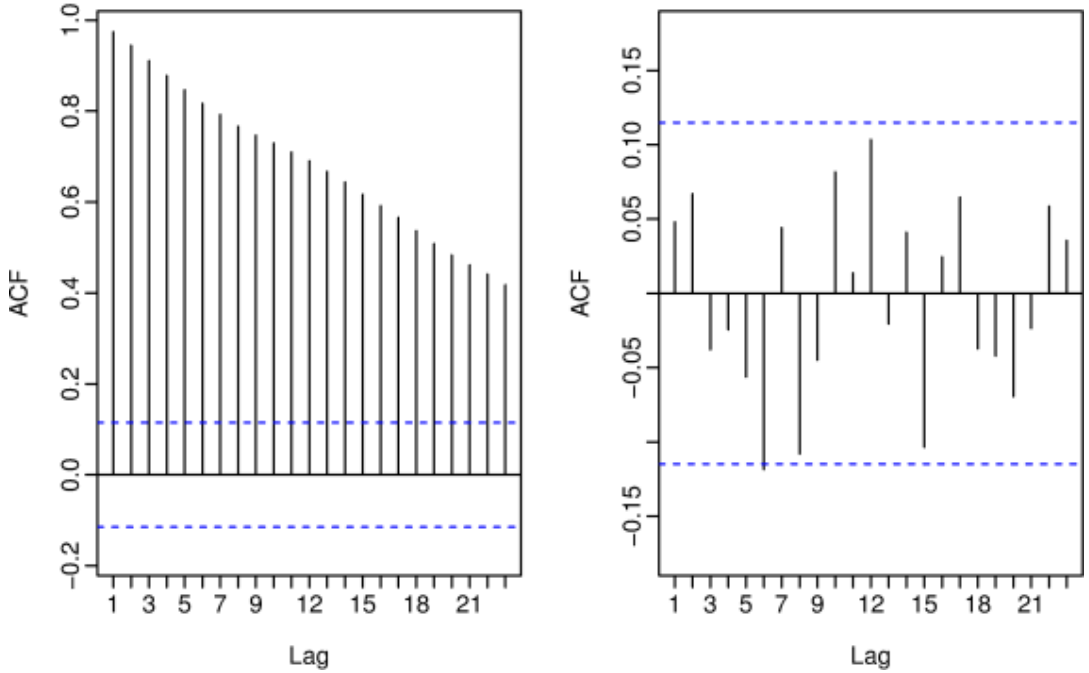
- A series is said to be **non-stationary** if it is not stationary

Weak Stationarity

- A series is said to be **weakly stationary** or **covariance stationary** if the following conditions are met:
- $E(Y_1) = E(Y_2) = \ldots = E(Y_t) = \text{constant}$
- $Var(Y_1) = Var(Y_2) = \ldots = Var(Y_t) = \text{constant}$
- $Cov(Y_1, Y_{1+k}) = Cov(Y_2, Y_{2+k}) = \ldots = Cov(Y_p, Y_{p+k}) = \text{constant}$, depends only on lag k

Making a series stationary: ARIMA

- A series which is non-stationary can be made stationary by **differencing**
  - There are other methods too: transforming, etc.
- A series which is stationary after being differentiated once is said to be integrated of order 1 and is denoted by I(1)
  - If integrated of order d, denoted by I(d)
  - Series which is stationary to begin with is I(0)
- ARMA with differencing = ARIMA



Why Stationarity?

- Results of classical econometric theory are derived under the stationarity assumption
- Standard techniques often invalid when series is non-stationary
- Non-stationary time series regressions can result in spurious regressions, i.e. regression expression shows relationship between two variables when so such relationship exists

Auto-Correlation Function (ACF)

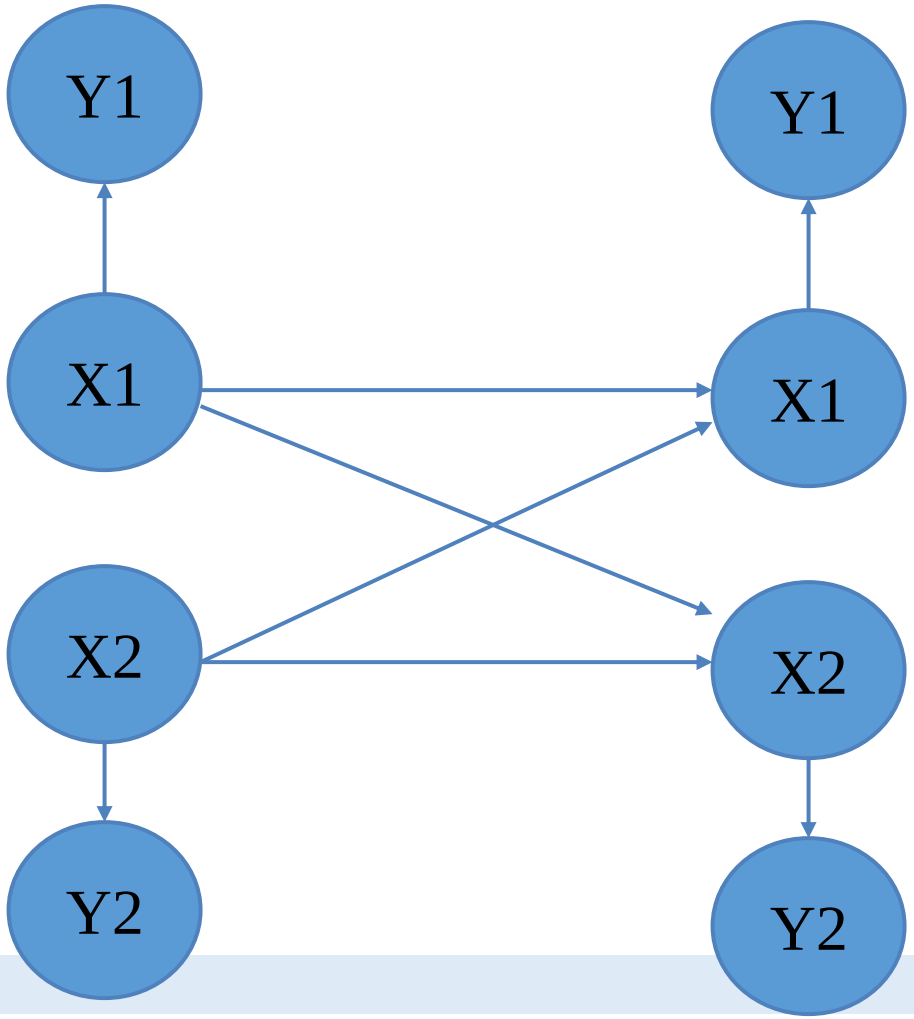
- ACF refers to the way observations in a time series are related to each other
- Measured by simple correlation between current observation, Y_t and previous with lag p, Y_{t-p}

  $Corr(Y_t, Y_{t-p}) = Cov(Y_t, Y_{t-p}) / sqrt(Var(Y_t))\ sqrt(Var(Y_{t-p}))$

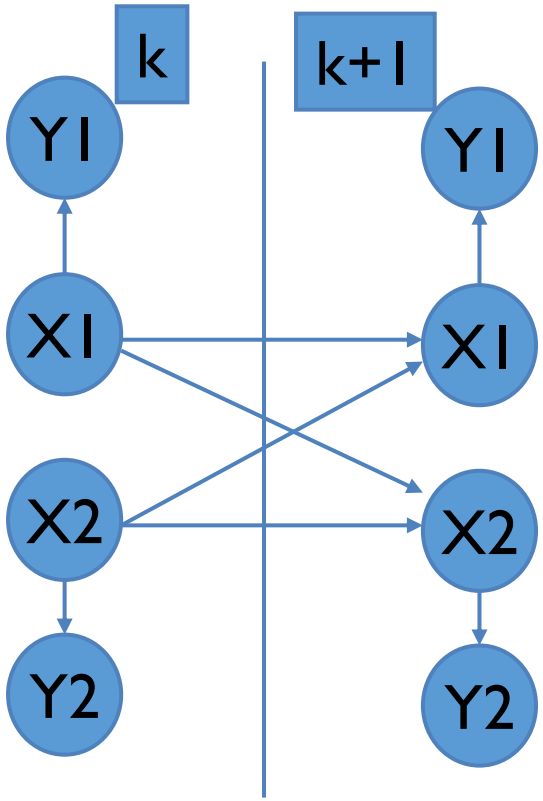
- **Correlogram** plotted with different ps to identify choice of p and q in AR, MA, ARMA, ARIMA

From Black-box to Structural Stochastic Models

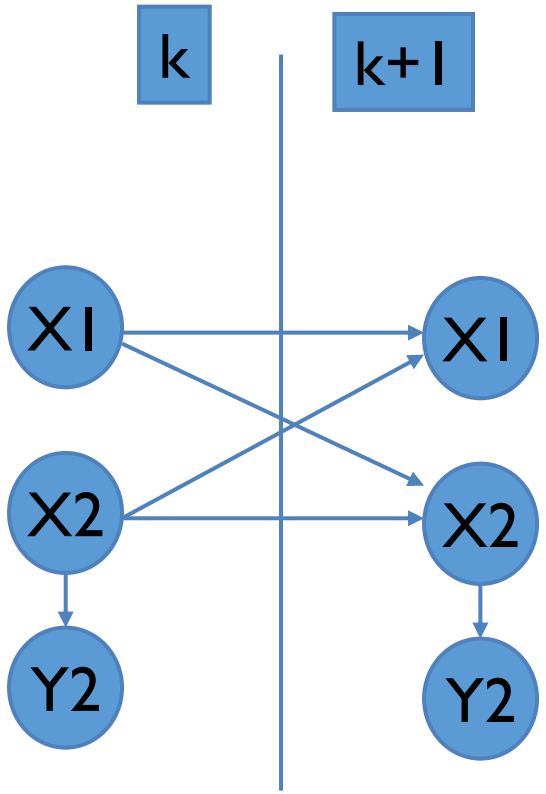
- Examples:
  - Kalman filters
  - Dynamic BNs
  - Hidden Markov Models



Observable and partially observable models



Fully observable



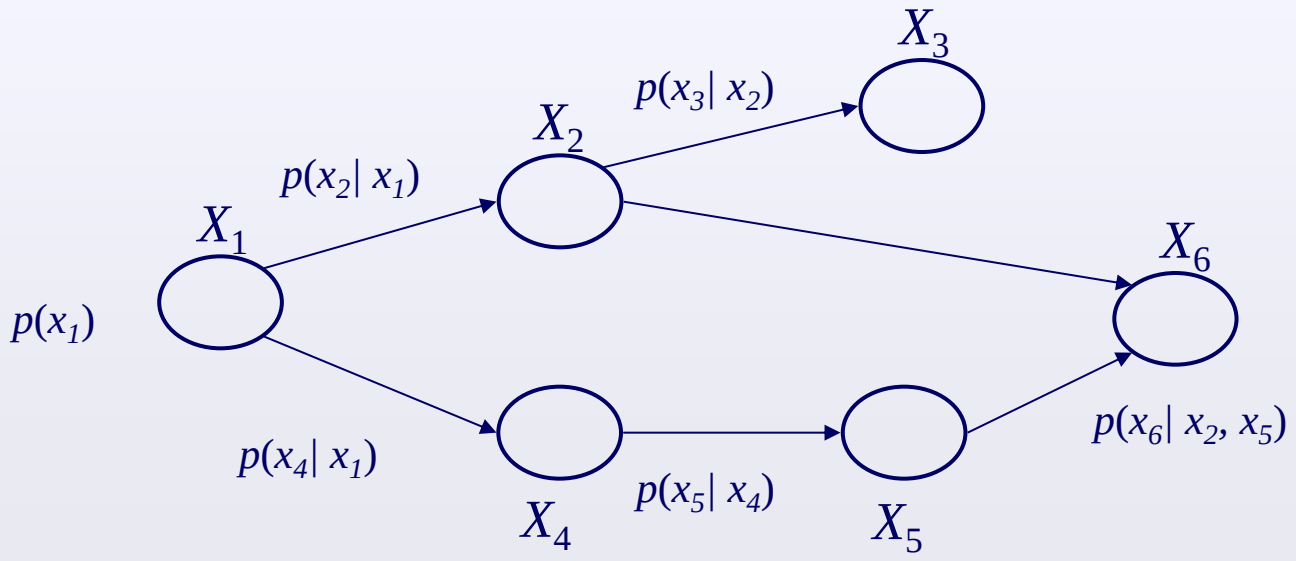
Partially observable

Acknowledgements

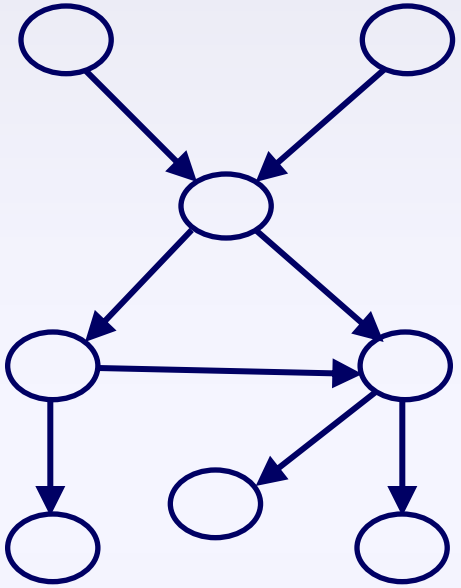
- Next few slides borrowed from:
  - Erik Suddherth, U Mass-Amherst
  - Eric Xing, CMU

Graphical Models – A Quick Intro

- A way of specifying conditional independences.
- Directed Graphical Modes: a DAG
- Nodes are random variables.
- A node's distribution depends on its parents.
- Joint distribution: $p(x) = \Pi_i p(x_i|\mathrm{Parents}_i)$
- A node's value conditional on its parents is independent of other ancestors.

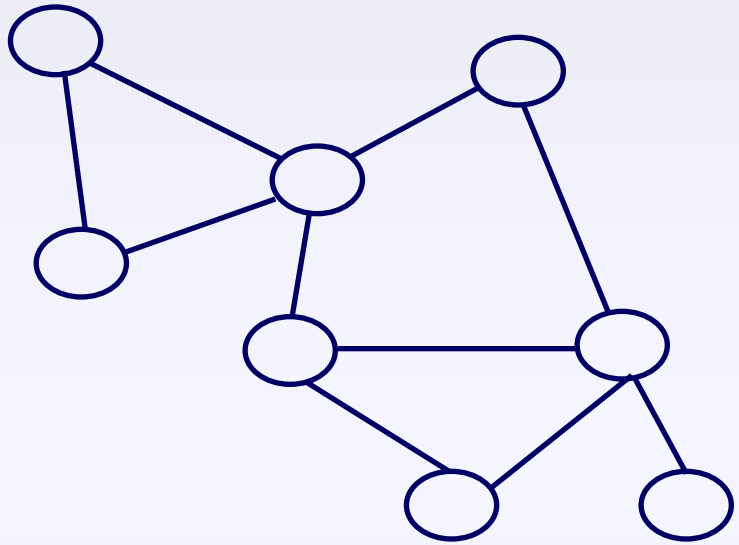


Types of Graphical Models



Nodes ⟷ Random Variables

Edges ⟷ Probabilistic (Markov) Relationships



Directed Graphs

Specify a hierarchical, causal generative process (child nodes depend on parents)

$$p(x) = \Pi_i p(x_i|\text{Parents}_i)$$

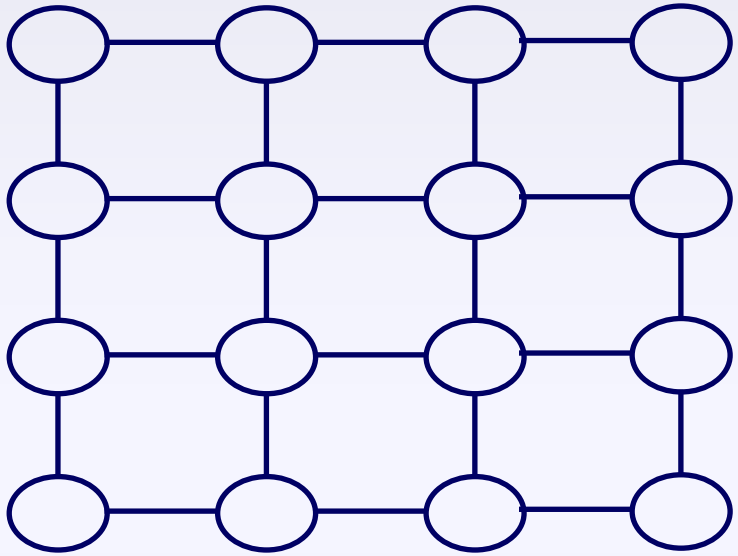
Undirected Graphs

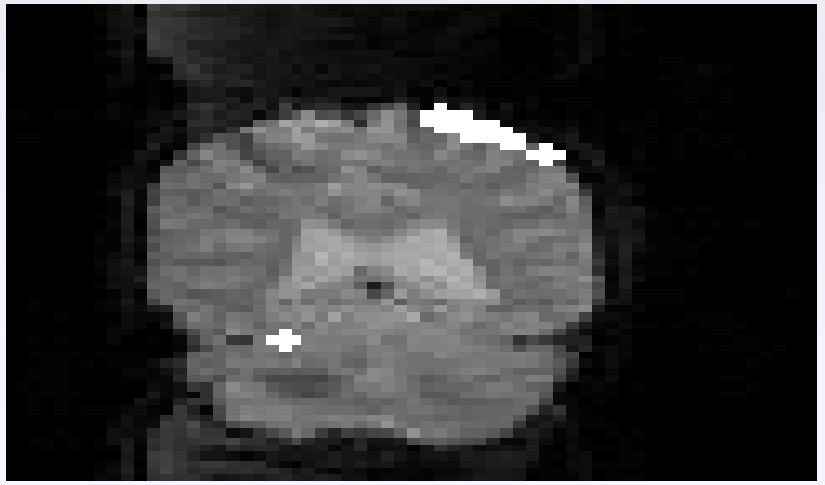
Specific symmetric, non-causal dependencies (soft or probabilistic constraints)

$$p(x) = \prod_{\text{cliques}} \Psi(x_{\text{clique}})$$

Markov Random Fields in Vision

Idea: Nearby pixels are similar.





fMRI Analysis *(Kim et. al. 2000)*

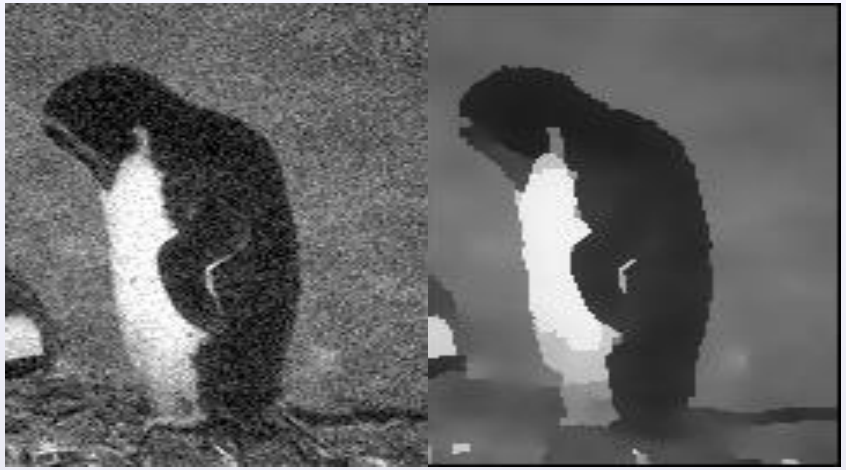
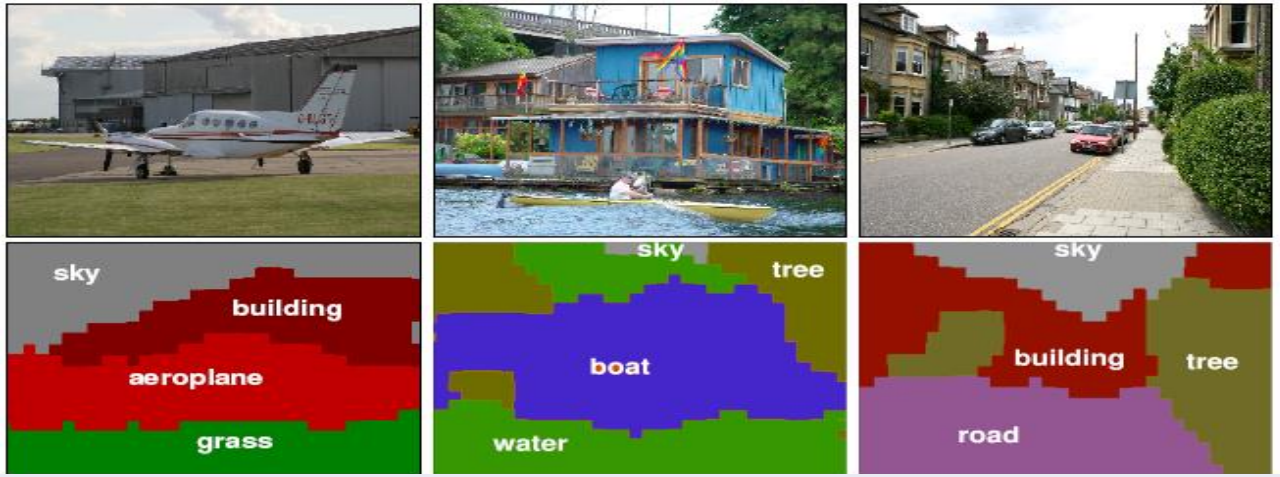


Image Denoising

(Felzenszwalb & Huttenlocher 2004)

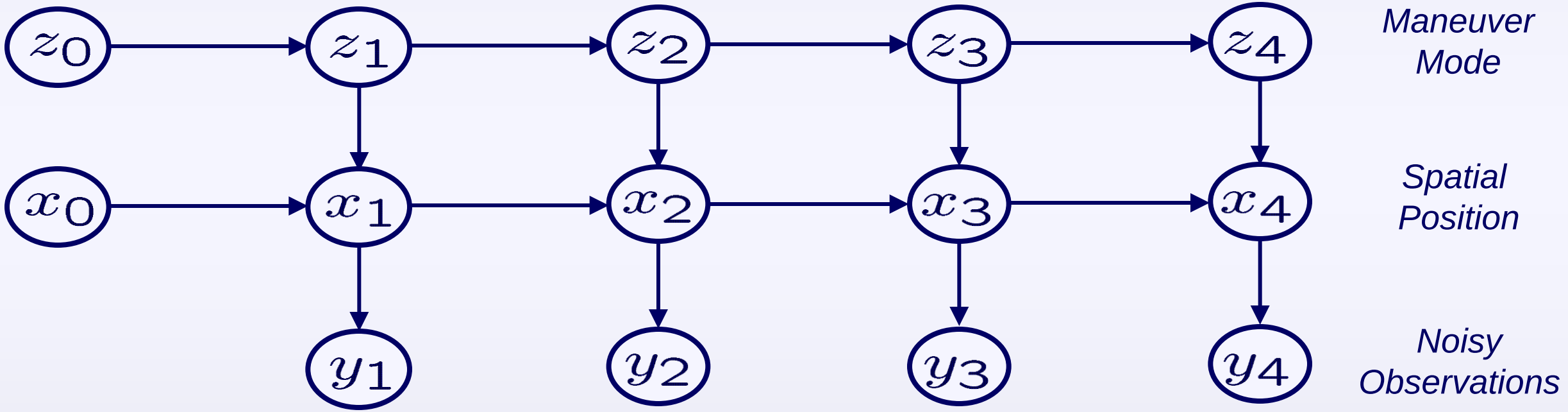


Segmentation & Object Recognition

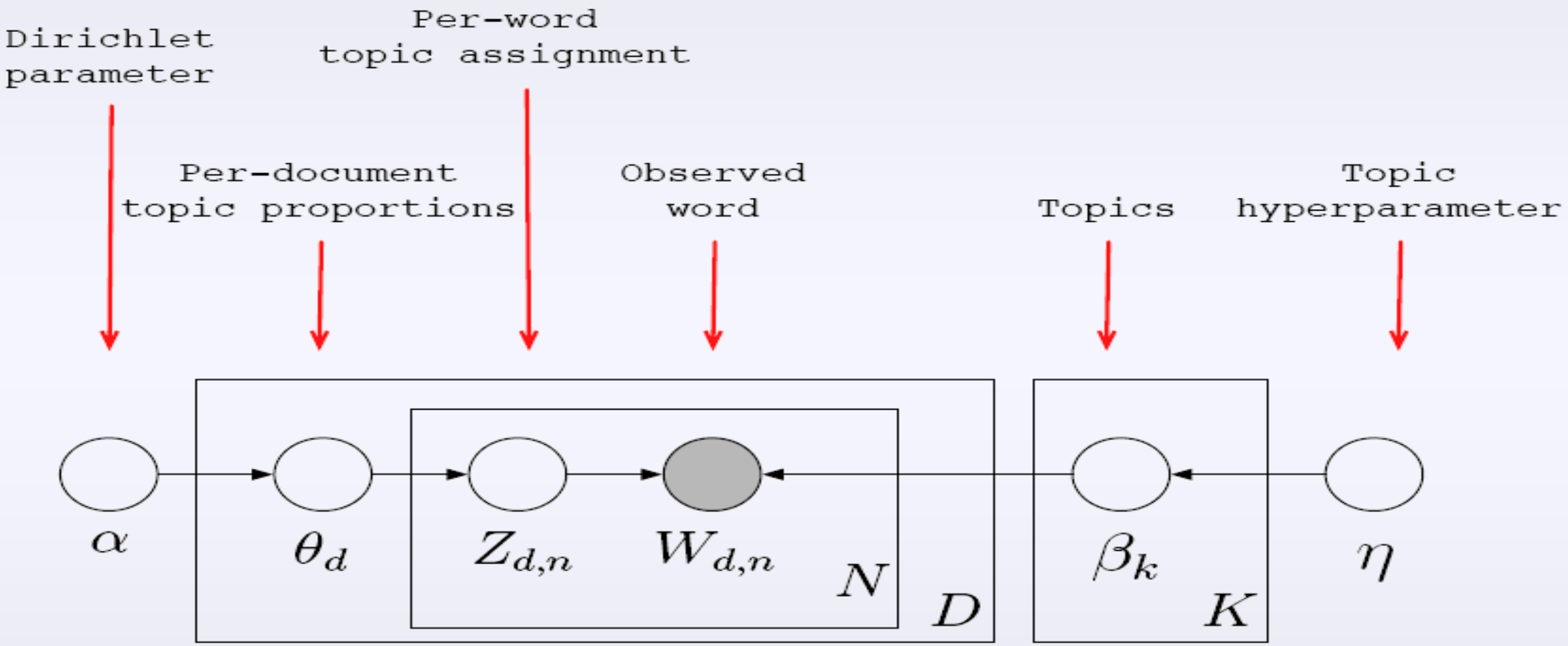
(Verbeek & Triggs 2007)

Dynamic Bayesian Networks

Specify and exploit internal structure in the hidden states underlying a time series. Generalizes HMMs



Topic Models for Documents



D. Blei, 2007

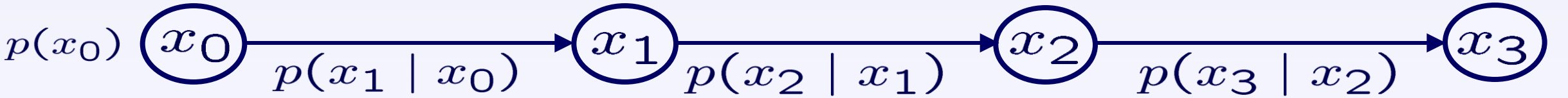
Topics Learned from *Science*

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

D. Blei, 2007

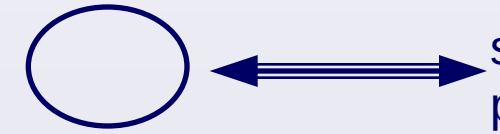
Markov Chains: Graphical Models

$$p(x_0, x_1, \ldots, x_T) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1})$$



$$Q = \begin{bmatrix} 0.5 & 0.1 & 0.0 \\ 0.3 & 0.0 & 0.4 \\ 0.2 & 0.9 & 0.6 \end{bmatrix}$$

- Graph interpretation differs from state transition diagrams:



state values at particular times

nodes

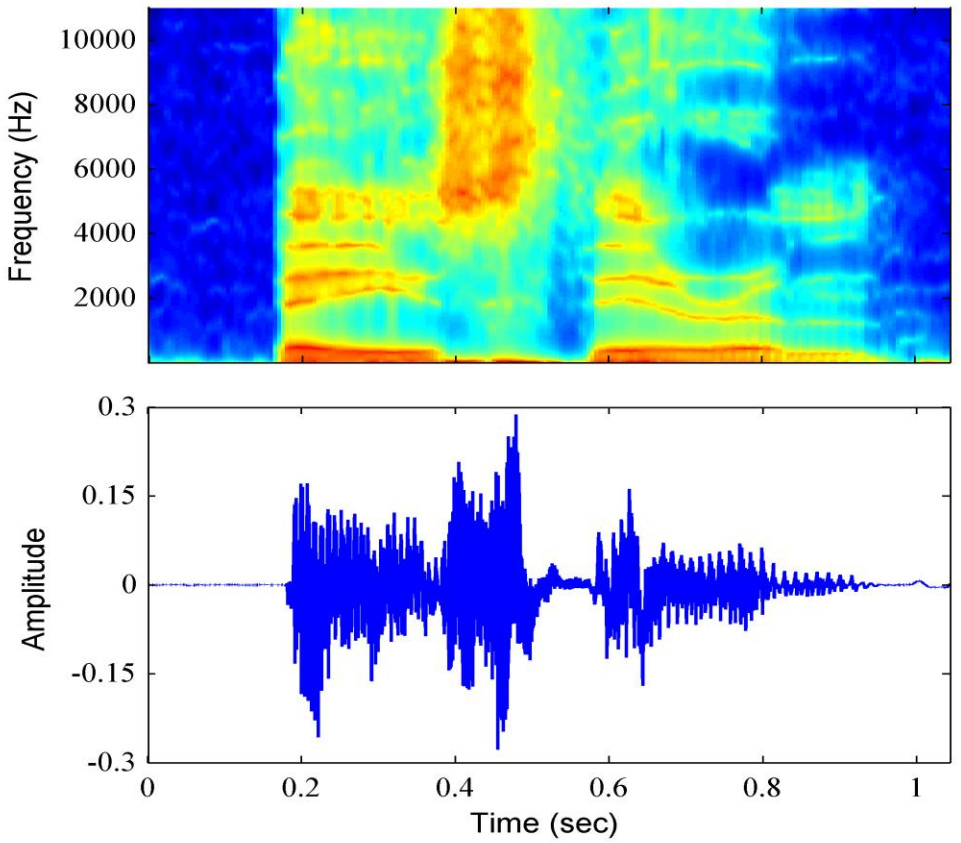
Markov properties

edges

i.i.d to sequential data

- So far we assumed independent, identically distributed data $\{X_i\}_{i=1}^n \overset{iid}{\sim} p(\mathbf{X})$
- Sequential data
  - Time-series data

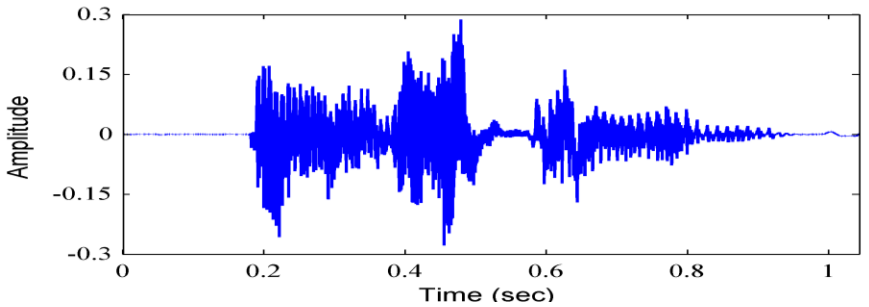
    E.g. Speech



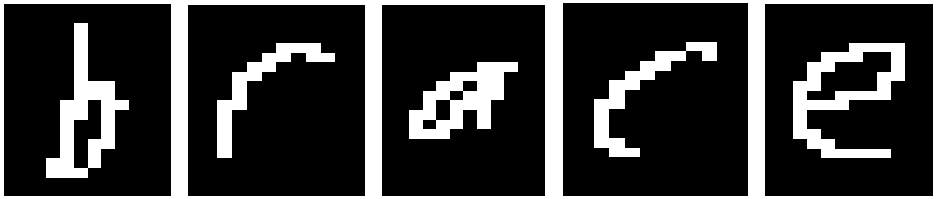
i.i.d to sequential data

- So far we assumed independent, identically distributed data $\{X_i\}_{i=1}^n \overset{iid}{\sim} p(\mathbf{X})$
- Sequential data
  - Time-series data

    E.g. Speech



  - Characters in a sentence



  - Base pairs along a DNA strand

Markov Models

- Joint Distribution

$$
\begin{aligned}
p(\mathbf{X}) &= p(X_1, X_2, \ldots, X_n) \\
&= p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)\ldots p(X_n|X_{n-1}, ..., X_1) \\
&= \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_1) \qquad \text{Chain rule}
\end{aligned}
$$

- Markov Assumption (m^{th} order)

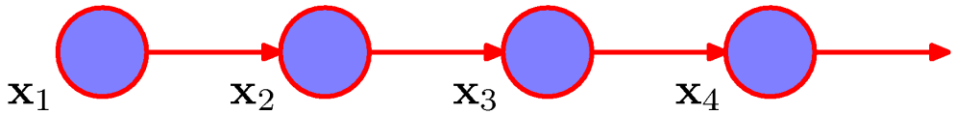
$$
p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1}, \ldots, X_{n-m})
$$

Current observation only depends on past m observations

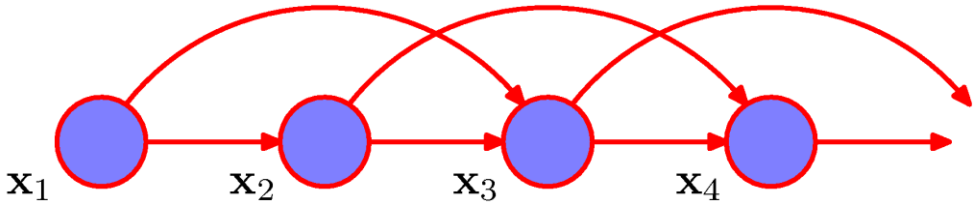
Markov Models

- Markov Assumption

1st order $$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1})$$



2nd order $$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1}, X_{n-2})$$



Markov Models

parameters in stationary model K-ary variables

- Markov Assumption

1st order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1})$ $O(K^2)$

mth order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1},\ldots,X_{n-m})$ $O(K^{m+1})$

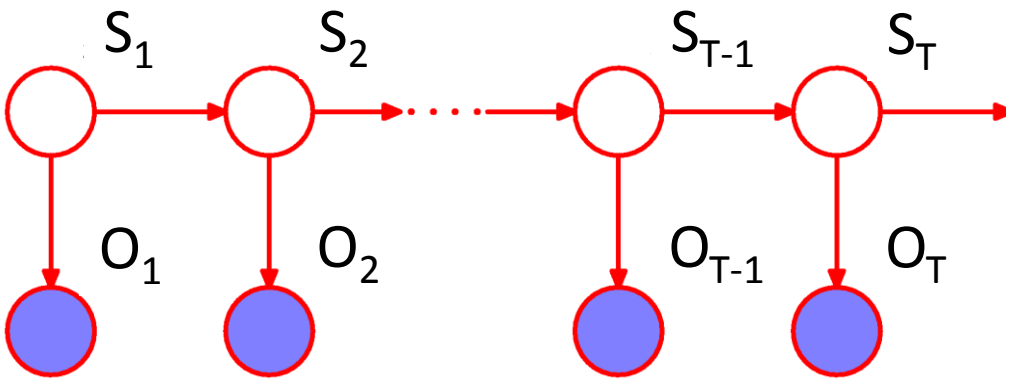
n-1th order $p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n|X_{n-1},\ldots,X_1)$ $O(K^n)$

≡ no assumptions – complete (but directed) graph

Homogeneous/stationary Markov model (probabilities don't depend on n)

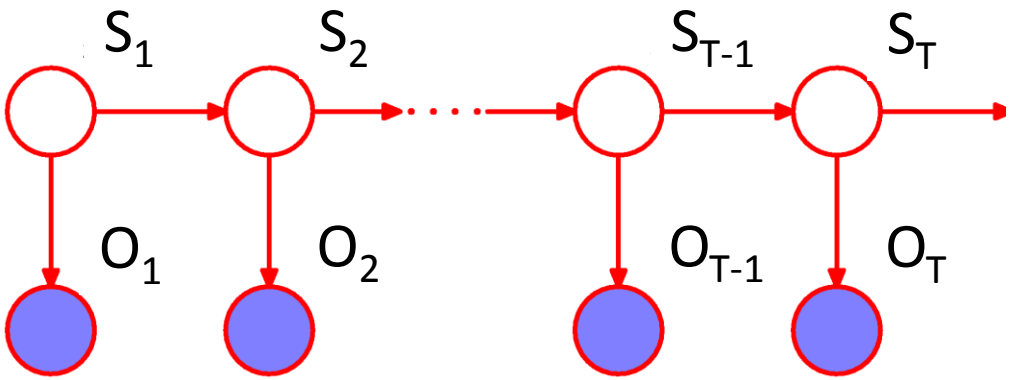
Hidden Markov Models

- Distributions that characterize sequential data with few parameters but are not limited by strong Markov assumptions.



Observation space	$O_t \in \{y_1, y_2, \ldots, y_K\}$
Hidden states	$S_t \in \{1, \ldots, I\}$

Hidden Markov Models



$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

1st order Markov assumption on hidden states $\{S_t\}$ t = 1, …, T
(can be extended to higher order).

Note: O_t depends on all previous observations $\{O_{t-1},\ldots O_1\}$

Hidden Markov Models

- Parameters – stationary/homogeneous markov model (independent of time t)

Initial probabilities

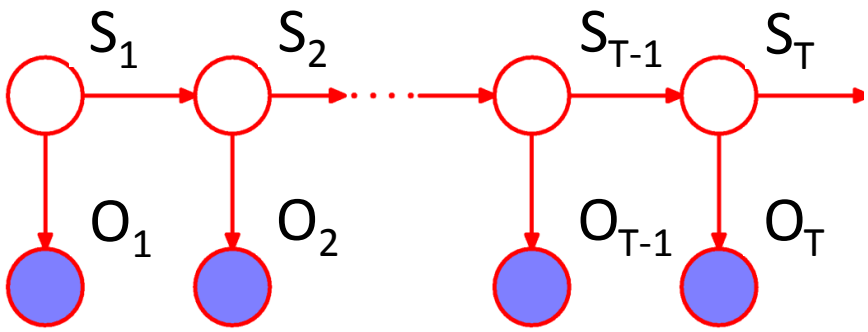
$p(S_1 = i) = \pi_i$

Transition probabilities

$p(S_t = j | S_{t-1} = i) = p_{ij}$

Emission probabilities

$p(O_t = y | S_t = i) = q_i^y$



$$p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = p(S_1) \prod_{t=2}^{T} p(S_t|S_{t-1}) \prod_{t=1}^{T} p(O_t|S_t)$$

HMM Example

- The Dishonest Casino

A casino has two die:

Fair dice

$P(1) = P(2) = P(3) = P(5) = P(6) = 1/6$

Loaded dice

$P(1) = P(2) = P(3) = P(5) = 1/10$

$P(6) = ½$

Casino player switches back-&-forth between fair and loaded die once every 20 turns

HMM Problems

GIVEN: A sequence of rolls by the casino player

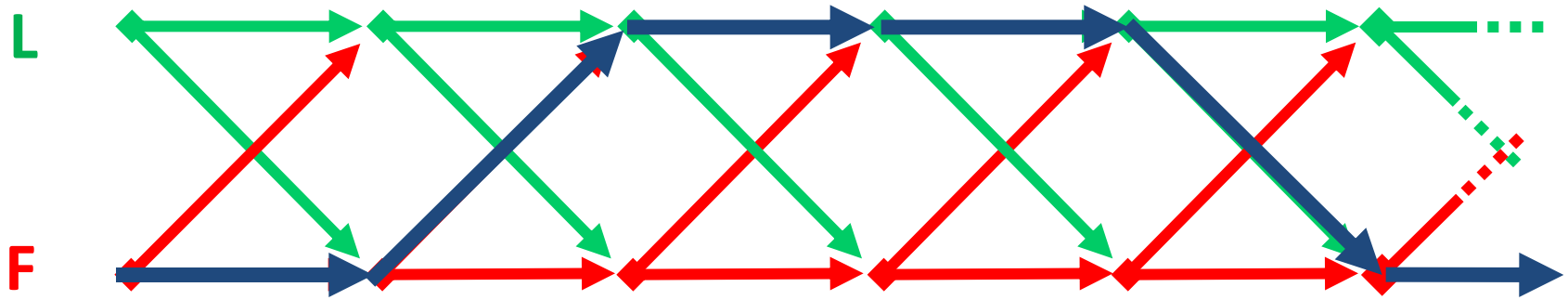
12455264621461461361366616646616366163661636161515615115146123562344

QUESTION

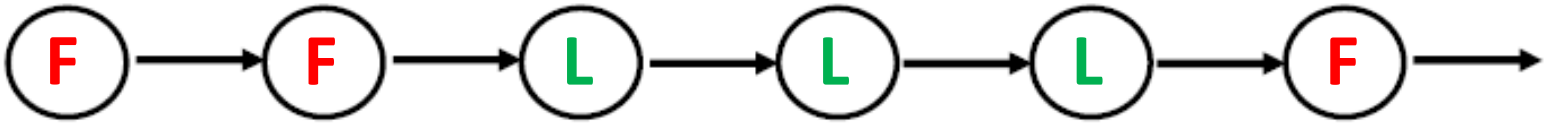
- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem in HMMs
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question in HMMs
- How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question in HMMs

HMM Example

- Observed sequence: $\{O_t\}_{t=1}^{T}$

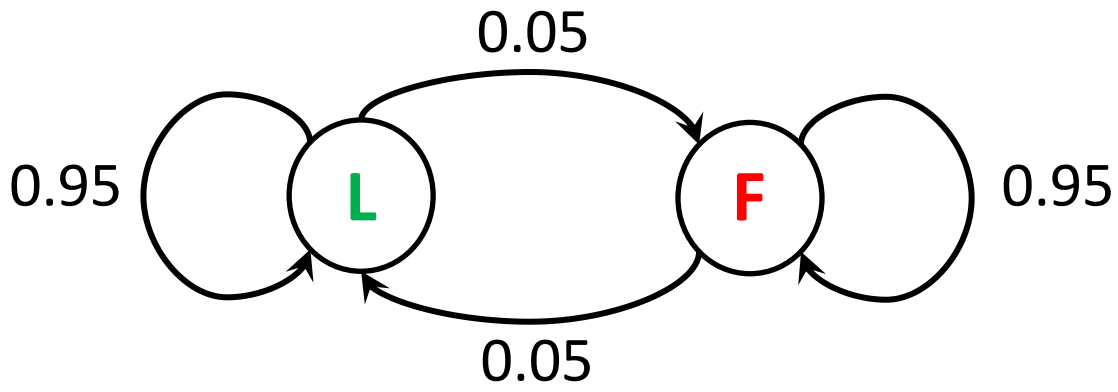


- Hidden sequence $\{S_t\}_{t=1}^{T}$ or segmentation):



State Space Representation

- Switch between **F** and **L** once every 20 turns (1/20 = 0.05)



- HMM Parameters

Initial probs	$P(S_1 = \mathbf{L}) = 0.5 = P(S_1 = \mathbf{F})$
Transition probs	$P(S_t = \mathbf{L}/\mathbf{F} \mid S_{t-1} = \mathbf{L}/\mathbf{F}) = 0.95$
	$P(S_t = \mathbf{F}/\mathbf{L} \mid S_{t-1} = \mathbf{L}/\mathbf{F}) = 0.05$
Emission probabilities	$P(O_t = y \mid S_t = \mathbf{F}) = 1/6$ $\quad y = 1,2,3,4,5,6$
	$P(O_t = y \mid S_t = \mathbf{L}) = 1/10$ $\quad y = 1,2,3,4,5$
	$= 1/2$ $\quad y = 6$

Three main problems in HMMs

- Evaluation – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $p(\{O_t\}_{t=1}^T)$ prob of observed sequence

- Decoding – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $\arg \max_{s_1,\ldots,s_T} p(\{S_t\}_{t=1}^T | \{O_t\}_{t=1}^T)$ most probable

  sequence of hidden states

- Learning – Given HMM with unknown parameters and $\{O_t\}_{t=1}^T$ observation sequence

  find $\arg \max_{\theta} p(\{O_t\}_{t=1}^T | \theta)$ parameters that maximize

  likelihood of observed data

HMM Algorithms

- Evaluation – What is the probability of the observed sequence? Forward Algorithm

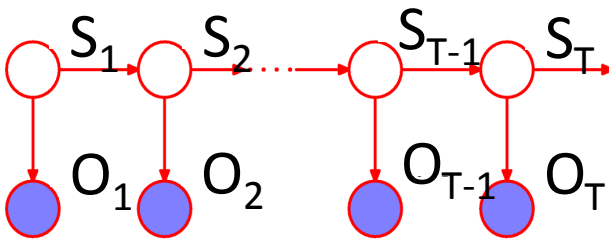
- Decoding – What is the probability that the third roll was loaded given the observed sequence? Forward-Backward Algorithm

  – What is the most likely die sequence given the observed sequence? Viterbi Algorithm

- Learning – Under what parameterization is the observed sequence most probable? Baum-Welch Algorithm (EM)

Evaluation Problem

- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$



find probability of observed sequence

$$p(\{O_t\}_{t=1}^T) = \sum_{S_1,\ldots,S_T} p(\{O_t\}_{t=1}^T, \{S_t\}_{t=1}^T)$$

$$= \sum_{S_1,\ldots,S_T} p(S_1) \prod_{t=2}^T p(S_t|S_{t-1}) \prod_{t=1}^T p(O_t|S_t)$$

requires summing over all possible hidden state values at all times – K^T exponential # terms!

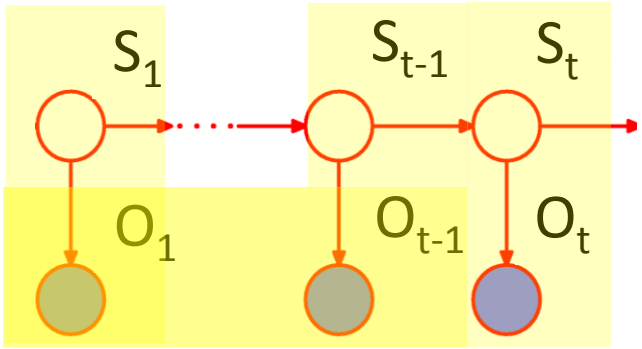
Instead: $p(\{O_t\}_{t=1}^T) = \sum_k \underbrace{p(\{O_t\}_{t=1}^T, S_T = k)}_{\alpha_T^k}$ Compute recursively

Forward Probability

$$p(\{O_t\}_{t=1}^T) = \sum_k p(\{O_t\}_{t=1}^T, S_T = k) = \sum_k \alpha_T^k$$

Compute forward probability α_t^k recursively over t

$$\alpha_t^k := p(O_1, \ldots, O_t, S_t = k)$$



Introduce S_{t-1}

Chain rule

Markov assumption

$$= p(O_t|S_t = k) \sum_i \alpha_{t-1}^i p(S_t = k|S_{t-1} = i)$$

Forward Algorithm

Can compute α_t^k for all k, t using dynamic programming:

- Initialize: $\alpha_1^k = p(O_1 | S_1 = k)\, p(S_1 = k)$ for all k
- Iterate: for t = 2, …, T

$$\alpha_t^k = p(O_t | S_t = k) \sum_i \alpha_{t-1}^i \, p(S_t = k | S_{t-1} = i) \quad \text{for all k}$$

- Termination: $p(\{O_t\}_{t=1}^T) = \sum_k \alpha_T^k$

Decoding Problem 1

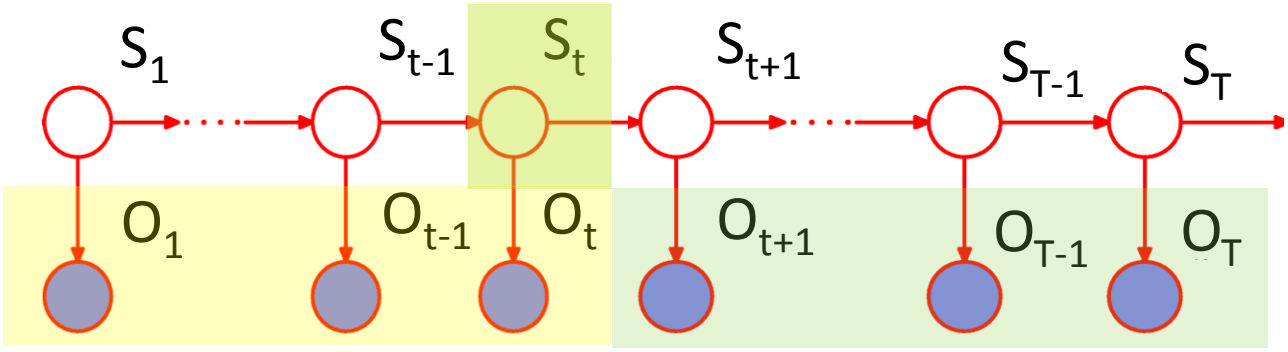
- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

  find probability that hidden state at time t was k $p(S_t = k|\{O_t\}_{t=1}^T)$

$$
\begin{aligned}
p(S_t = k, \{O_t\}_{t=1}^T) &= p(O_1, \ldots, O_t, S_t = k, O_{t+1}, \ldots, O_T) \\
&= p(O_1, \ldots, O_t, S_t = k) p(O_{t+1}, \ldots, O_T | S_t = k)
\end{aligned}
$$

Compute recursively

α_t^k β_t^k

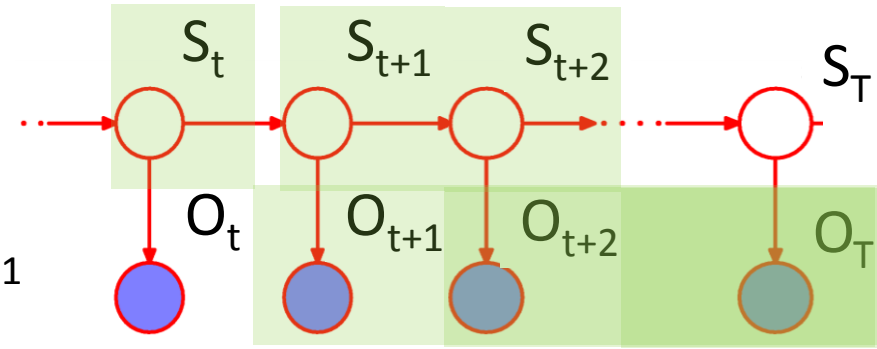


Backward Probability

$$p(S_t = k, \{O_t\}_{t=1}^T) = p(O_1, \ldots, O_t, S_t = k) p(O_{t+1}, \ldots, O_T | S_t = k) = \alpha_t^k \beta_t^k$$

Compute forward probability β_t^k recursively over t

$$\beta_t^k := p(O_{t+1}, \ldots, O_T | S_t = k)$$



Introduce S_{t+1}

.
.
.

Chain rule

Markov assumption

$$= \sum_i p(S_{t+1} = i | S_t = k) p(O_{t+1} | S_{t+1} = i) \beta_{t+1}^i$$

Backward Algorithm

Can compute β_t^k for all k, t using dynamic programming:

- Initialize: $\beta_T^k = 1$ for all k
- Iterate: for t = T-1, …, 1

$$\beta_t^k = \sum_i p(S_{t+1} = i | S_t = k) p(O_{t+1} | S_{t+1} = i) \beta_{t+1}^i \quad \text{for all k}$$

- Termination: $p(S_t = k, \{O_t\}_{t=1}^T) = \alpha_t^k \beta_t^k$

$$p(S_t = k | \{O_t\}_{t=1}^T) = \frac{p(S_t = k, \{O_t\}_{t=1}^T)}{p(\{O_t\}_{t=1}^T)} = \frac{\alpha_t^k \beta_t^k}{\sum_i \alpha_t^i \beta_t^i}$$

Most likely state vs. Most likely sequence

- Most likely state assignment at time t

$$\arg\max_k p(S_t = k|\{O_t\}_{t=1}^T) = \arg\max_k \alpha_t^k \beta_t^k$$

E.g. Which die was most likely used by the casino in the third roll given the observed sequence?

- Most likely assignment of state sequence

$$\arg\max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T|\{O_t\}_{t=1}^T)$$

E.g. What was the most likely sequence of die rolls used by the casino given the observed sequence?

Not the same solution !

MLA of x?
MLA of (x,y)?

x	y	P(x,y)
0	0	0.35
0	1	0.05
1	0	0.3
1	1	0.3

Decoding Problem 2

- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

  find most likely assignment of state sequence

$$\arg \max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T | \{O_t\}_{t=1}^T) = \arg \max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T)$$

$$= \arg \max_k \underbrace{\max_{\{S_t\}_{t=1}^{T-1}} p(S_T = k, \{S_t\}_{t=1}^{T-1}, \{O_t\}_{t=1}^T)}_{V_T^k}$$

Compute recursively

V_T^k - probability of most likely sequence of states ending at state $S_T = k$

Viterbi Decoding

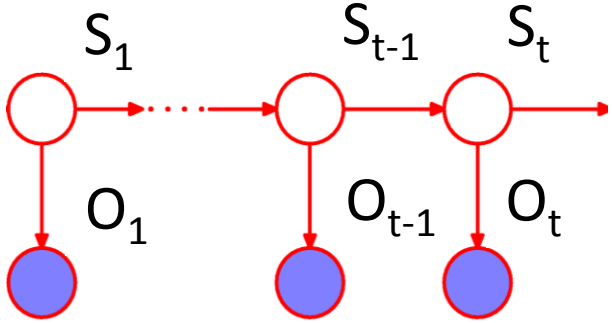
$$\max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = \max_k V_T^k$$

Compute probability V_t^k recursively over t

$$V_t^k := \max_{S_1,\ldots,S_{t-1}} p(S_t = k, S_1, \ldots, S_{t-1}, O_1, \ldots, O_t)$$

. Bayes rule

. Markov assumption



$$= p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i) V_{t-1}^i$$

Viterbi Algorithm

Can compute V_t^k for all k, t using dynamic programming:

- Initialize: $V_1^k = p(O_1|S_1=k)p(S_1 = k)$ for all k

- Iterate: for t = 2, …, T

$$V_t^k = p(O_t|S_t = k) \max_i p(S_t = k|S_{t-1} = i)V_{t-1}^i \quad \text{for all k}$$

- Termination: $\max_{\{S_t\}_{t=1}^T} p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) = \max_k V_T^k$

Traceback:

$$S_T^* = \arg\max_k V_T^k$$

$$S_{t-1}^* = \arg\max_i p(S_t^*|S_{t-1} = i)V_{t-1}^i$$

Computational complexity

- What is the running time for Forward, Forward-Backward, Viterbi?

$$\alpha_t^k = q_k^{O_t} \sum_i \alpha_{t-1}^i \; p_{i,k}$$

$$\beta_t^k = \sum_i p_{k,i} \; q_i^{O_{t+1}} \; \beta_{t+1}^i$$

$$V_t^k = q_k^{O_t} \; \max_i p_{i,k} \; V_{t-1}^i$$

$O(K^2T)$ linear in T instead of $O(K^T)$ exponential in T!

Learning Problem

- Given HMM with unknown parameters $\theta = \{\{\pi_i\}, \{p_{ij}\}, \{q_i^k\}\}$ and observation sequence $\mathbf{O} = \{O_t\}_{t=1}^T$

  find parameters that maximize likelihood of observed data

$$\arg\max_\theta p(\{O_t\}_{t=1}^T|\theta)$$

But likelihood doesn't factorize since observations not i.i.d.

  hidden variables – state sequence $\{S_t\}_{t=1}^T$

  EM (Baum-Welch) Algorithm:

  E-step – Fix parameters, find expected state assignments

  M-step – Fix expected state assignments, update parameters

Baum-Welch (EM) Algorithm

- Start with random initialization of parameters
- **E-step** – Fix parameters, find expected state assignments

$$\gamma_i(t) = p(S_t = i|O,\theta) = \frac{\alpha_t^i \beta_t^i}{\sum_j \alpha_t^j \beta_t^j}$$

Forward-Backward algorithm

$$\xi_{ij}(t) = p(S_{t-1}= i, S_t = j|O,\theta)$$

$$= \frac{p(S_{t-1} = i|O,\theta)p(S_t = j, O_t, \ldots, O_T|S_{t-1} = i,\theta)}{p(O_t, \ldots, O_T|S_{t-1} = i,\theta)}$$

$$= \frac{\gamma_i(t-1)\ p_{ij}\ q_j^{O_t}\ \beta_t^j}{\beta_{t-1}^i}$$

Baum-Welch (EM) Algorithm

- Start with random initialization of parameters
- **E-step**

$$\gamma_i(t) = p(S_t = i|O,\theta)$$

$$\xi_{ij}(t) = p(S_{t-1}= i, S_t = j|O,\theta)$$

$\sum_{t=1}^{T} \gamma_i(t)$ = expected # times in state i

$\sum_{t=1}^{T-1} \gamma_i(t)$ = expected # transitions from state i

$\sum_{t=1}^{T-1} \xi_{ij}(t)$ = expected # transitions from state i to j

- **M-step**

$$\pi_i = \gamma_i(1)$$

$$p_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}$$

$$q_i^k = \frac{\sum_{t=1}^{T} \delta_{O_t=k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$$

Some connections

- HMM vs Linear Dynamical Systems (Kalman Filters)

HMM:	States are Discrete Observations Discrete or Continuous
Linear Dynamical Systems:	Observations and States are multi-variate Gaussians whose means are linear functions of their parent states (see Bishop: Sec 13.3)

HMMs.. What you should know

- Useful for modeling sequential data with few parameters using discrete hidden states that satisfy Markov assumption
- Representation - initial prob, transition prob, emission prob,

  State space representation

- Algorithms for inference and learning in HMMs
  - Computing marginal likelihood of the observed sequence: forward algorithm
  - Predicting a single hidden state: forward-backward
  - Predicting an entire sequence of hidden states: viterbi
  - Learning HMM parameters: an EM algorithm known as Baum-Welch

Readings

- "Introduction to Machine Learning" by Ethem Alpaydin, Chapter 15
- Pattern Recognition and Machine Learning (Bishop), Chapter 13
- Rabiner, A Tutorial on HMMs and Selected Applications in Speech Recognition