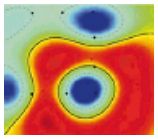
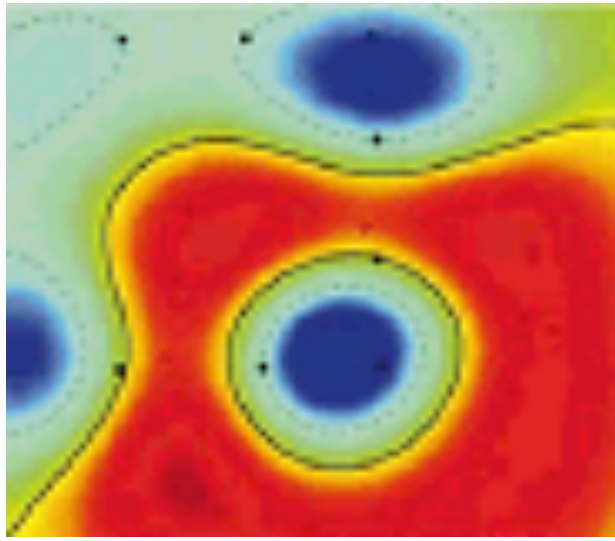


SoSe 2024

Machine Learning 2/2-X

Technische Universität Berlin



Lecture 7

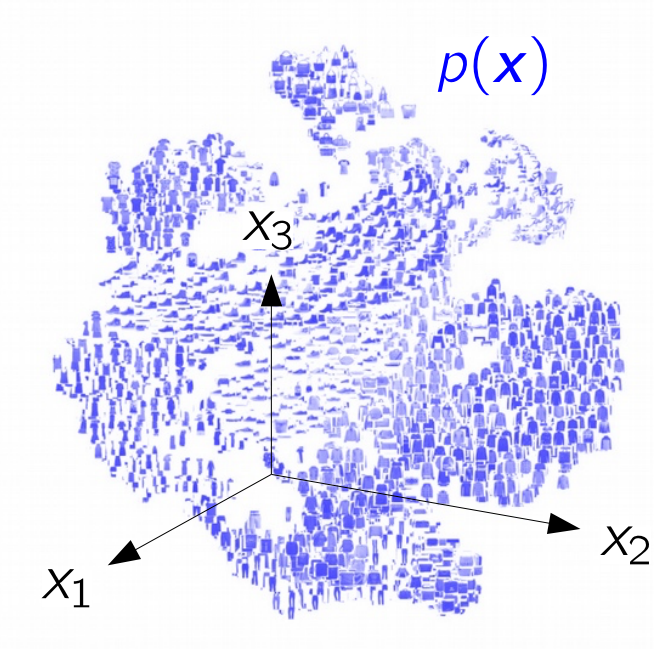
Hidden Markov Models

Introduction

World is represented by a vector of variables

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

governed by a probability function $p(\boldsymbol{x})$

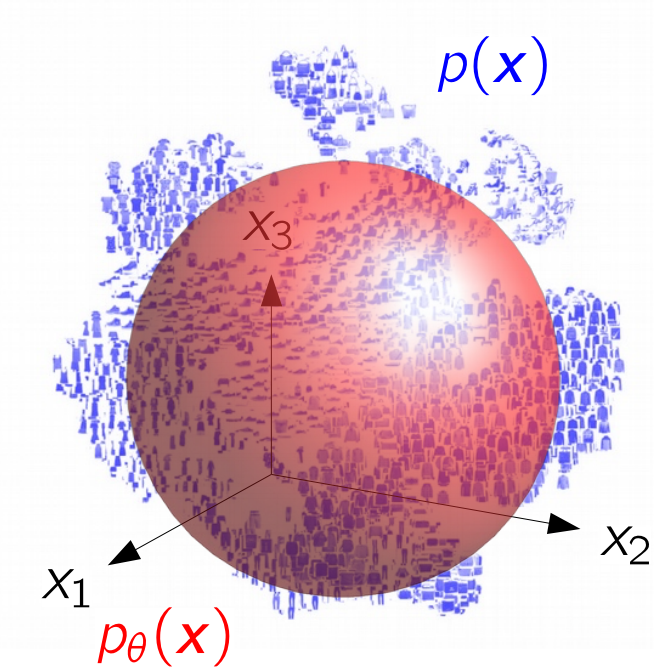


Introduction

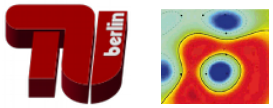
World is represented by a vector of variables

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

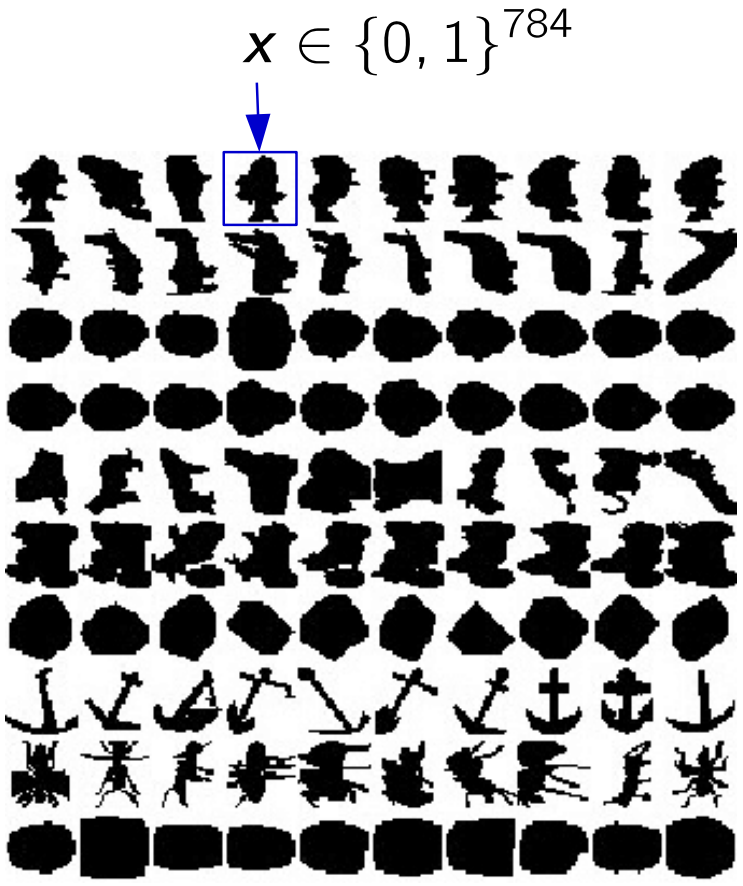
governed by a probability function $p(\boldsymbol{x})$



Goal: learn from a few observations a model $p_\theta(\boldsymbol{x})$ that that is close to the true probability function $p(\boldsymbol{x})$



Example: caltech101 silhouettes

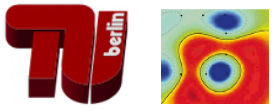


Each image consists of 28x28 pixels either black or white. There are 2^{28x28} possible images.

Question: can we simply estimate the frequency of each of these possible images?

Answer: Unfeasible, because data (and memory) are nite.

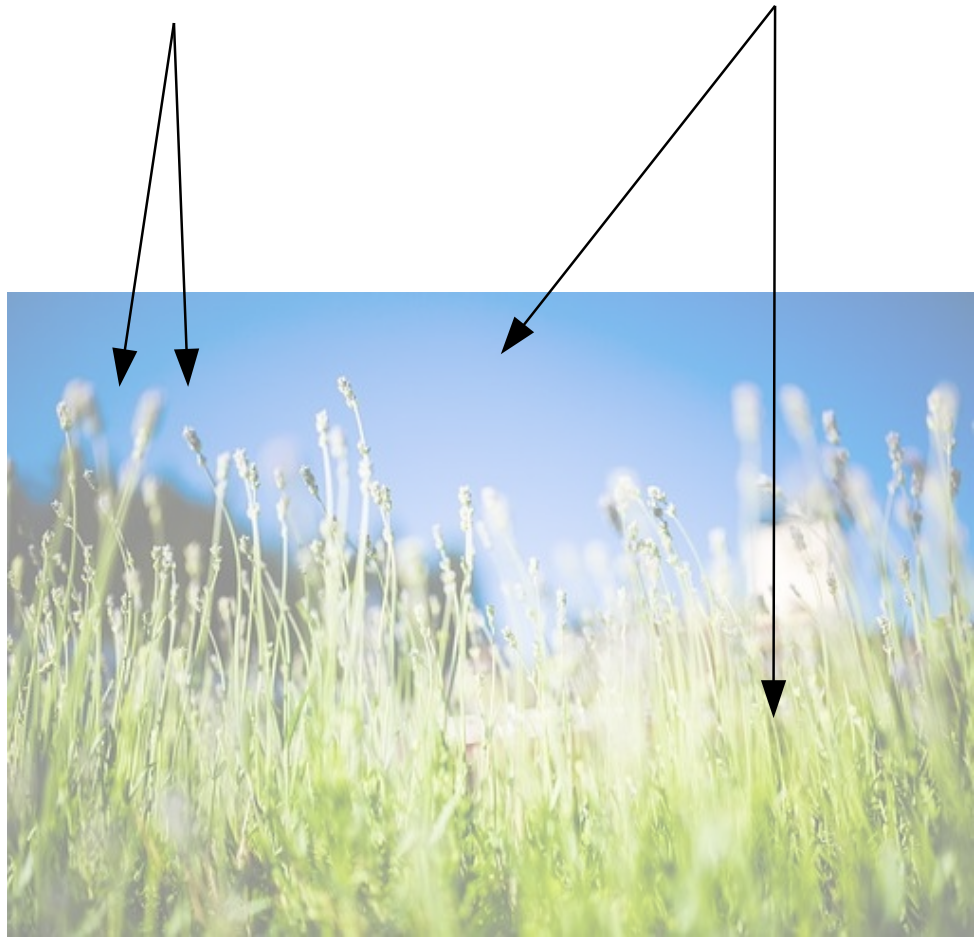
Therefore, we need to impose a **structure** to our probability function.



What Kind of Structure?

nearby pixel activities
are strongly dependent

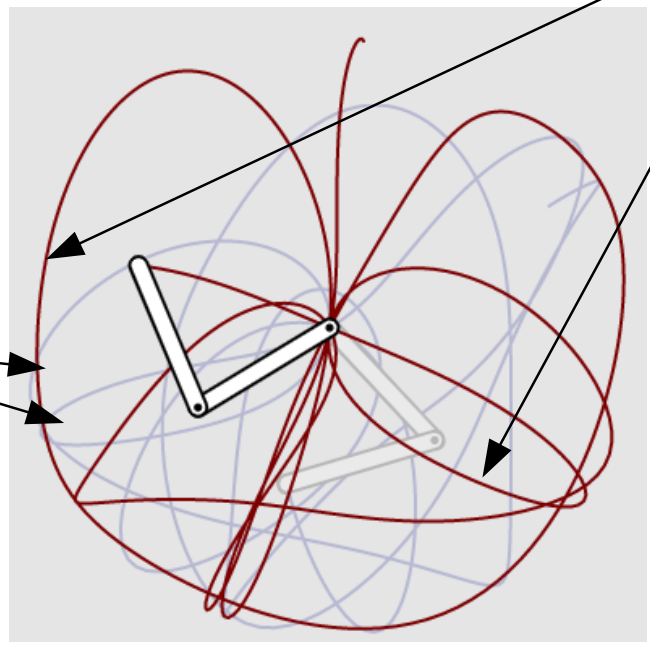
distant pixel activities
are less dependent



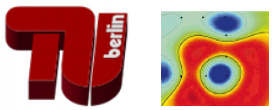
What Kind of Structure?

Dependence over large time spans is difficult to keep track of (e.g. chaotic systems).

successive posititions depedent



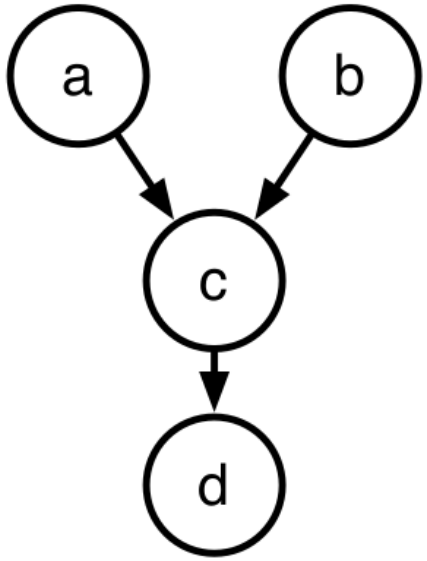
http://rocs.hu-berlin.de/explorables/explorables/double-trouble/



Directed Graphical Models

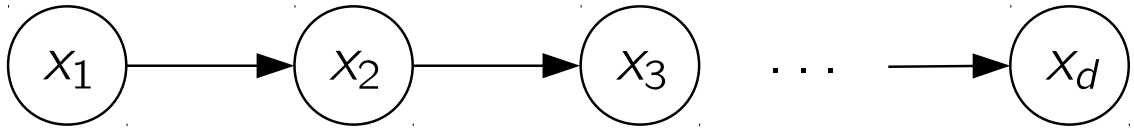
Pictorial way of representing independence assumptions in the data.
Each graphical model can be mapped to a given distribution.

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$$



Factored models of probabilities

Example: Markov Chain



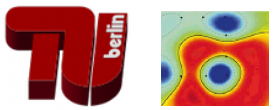
$\boldsymbol{x} \in \{0,1\}^d$

$p(x_1, x_2, \ldots, x_d) = p(x_d|x_{d-1}) \cdot \ldots \cdot p(x_2|x_1) \cdot p(x_1)$

x_{i-1}	x_i	$p(x_i\|x_{i-1})$
0	0	μ
0	1	$1-\mu$
1	0	ν
1	1	$1-\nu$

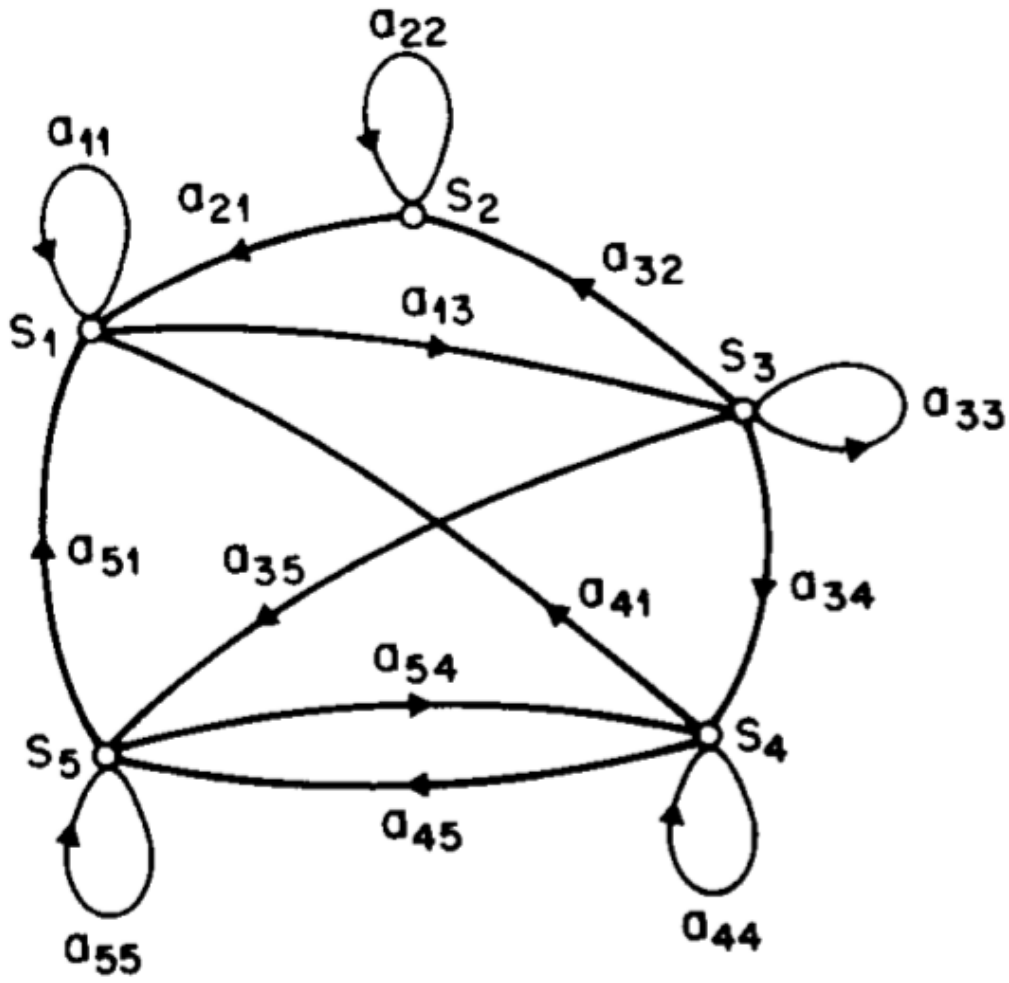
Number of parameters kept low by imposing the Markov property and stationarity.

We can sample from the model in a forward pass. Learning algorithm: parameters are the transition counts obtained from the data.



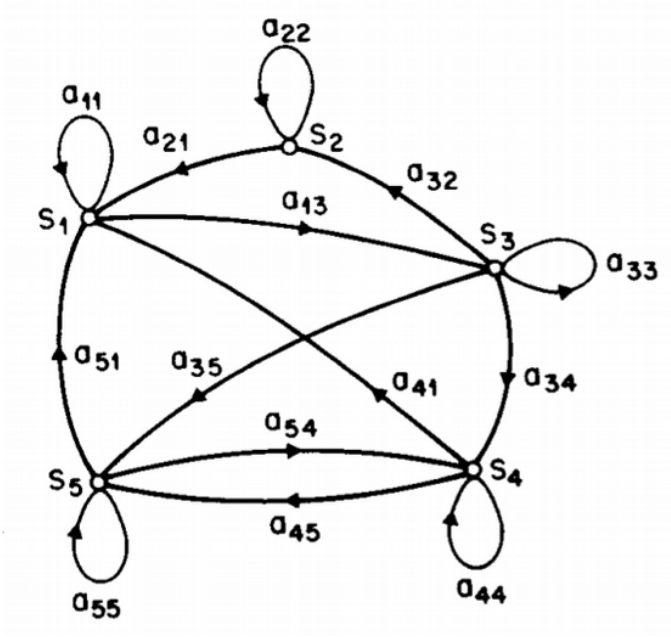
Discrete Markov Process

Consider a system which may be described at any time as being in one of a set of N distinct states, $S_1, S_2, \cdots, S_N$,



Source: L. Rabiner. A Tutorial on HMMs (1989).

Discrete Markov Process

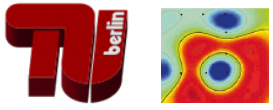


we denote the actual state at time t as q_t

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \cdots] = P[q_t = S_j | q_{t-1} = S_i].$$

Furthermore we only consider those processes in which the right-hand side of (1) is independent of time, thereby leading to the set of state transition probabilities a_{ij} of the form

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i], \qquad 1 \leq i, j \leq N$$



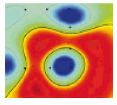
Source: L. Rabiner. A Tutorial on HMMs (1989).

Discrete Markov Process: Example

State 1: rain or (snow)
State 2: cloudy
State 3: sunny.

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

What is the probability (according to the model) that the weather for the next 7 days will be "sun-sun-rain-rain-sun-cloudy-sun $\cdots$"? Stated more formally, we define the observation sequence O as $O = \{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$ corresponding to $t = 1, 2, \cdots, 8$, and we wish to determine the probability of O, given the model.



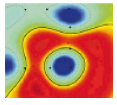
Source: L. Rabiner. A Tutorial on HMMs (1989).

Discrete Markov Process: Example

State 1: rain or (snow)
State 2: cloudy
State 3: sunny.

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

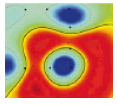
$$\begin{aligned} P(O|\text{Model}) &= P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|\text{Model}] \\ &= P[S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_3] \cdot P[S_1|S_3] \\ &\quad \cdot P[S_1|S_1] \cdot P[S_3|S_1] \cdot P[S_2|S_3] \cdot P[S_3|S_2] \\ &= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} \end{aligned}$$



Source: L. Rabiner. A Tutorial on HMMs (1989).

Extension for Hidden Markov Models

So far we have considered Markov models in which each state corresponded to an observable (physical) event. This model is too restrictive to be applicable to many problems of interest. In this section we extend the concept of Markov models to include the case where the observation is a probabilistic function of the state



Source: L. Rabiner. A Tutorial on HMMs (1989).

Elements of an Hidden Markov Model

1) N, the number of states in the model.

2) M, the number of distinct observation symbols per state, i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the system being modeled.

3) The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N.$$

4) The observation symbol probability distribution in state j, $B = \{b_j(k)\}$, where

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j], \quad 1 \leq j \leq N$$

$$1 \leq k \leq M.$$

5) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N.$$

Source: L. Rabiner. A Tutorial on HMMs (1989).

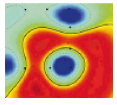
Generating with an HMM

Given appropriate values of N, M, A, B, and π, the HMM can be used as a generator to give an observation sequence

$$O = O_1 O_2 \cdots O_T \tag{10}$$

(where each observation O_t is one of the symbols from V, and T is the number of observations in the sequence) as follows:

1) Choose an initial state $q_1 = S_i$ according to the initial state distribution π.
2) Set $t = 1$.
3) Choose $O_t = v_k$ according to the symbol probability distribution in state S_i, i.e., $b_i(k)$.
4) Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, i.e., a_{ij}.
5) Set $t = t + 1$; return to step 3)if $t < T$; otherwise terminate the procedure.



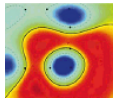
Source: L. Rabiner. A Tutorial on HMMs (1989).

Parameters of an HMM

It can be seen from the above discussion that a complete specification of an HMM requires specification of two model parameters (N and M), specification of observation symbols, and the specification of the three probability measures A, B, and π. For convenience, we use the compact notation

$$\lambda = (A, B, \pi) \tag{11}$$

to indicate the complete parameter set of the model.



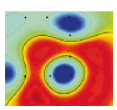
Source: L. Rabiner. A Tutorial on HMMs (1989).

Three Basic Problems for HMMs

Problem 1: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

Problem 2: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense (i.e., best "explains" the observations)?

Problem 3: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 1

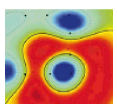
We wish to calculate the probability of the observation sequence, $O = O_1 O_2 \cdots O_T$, given the model λ, i.e., $P(O|\lambda)$. The most straightforward way of doing this is through enumerating every possible state sequence of length T (the number of observations). Consider one such fixed state sequence

$$Q = q_1 q_2 \cdots q_T \tag{12}$$

where q_1 is the initial state. The probability of the observation sequence O for the state sequence of (12) is

$$P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda) \tag{13a}$$

$$= b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdots b_{q_T}(O_T)$$



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 1

The joint probability of O and Q, i.e., the probability that O and Q occur simultaneously, is simply the product of the above two terms, i.e.,

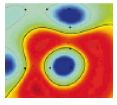
$$P(O, Q|\lambda) = P(O|Q, \lambda)\, P(Q, \lambda). \tag{15}$$

The probability of O (given the model)is obtained by summing this joint probability over all possible state sequences q giving

$$P(O|\lambda) = \sum_{\text{all } Q} P(O|Q, \lambda)\, P(Q|\lambda)$$

$$= \sum_{q_1, q_2, \cdots, q_T} \pi_{q_1} b_{q_1}(O_1)\, a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(Q_T).$$

involves on the order of $2T \cdot N^T$ calculations



Source: L. Rabiner. A Tutorial on HMMs (1989).

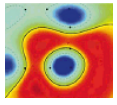
Solution of Problem 1

Clearly a more efficient procedure is required to solve Problem 1. Fortunately such a procedure exists and is called the forward-backward procedure.

The Forward–Backward Procedure [2], [3][6]: Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1\, O_2 \cdots O_t,\, q_t = S_i|\lambda) \tag{18}$$

i.e., the probability of the partial observation sequence, O_1 $O_2 \cdots O_t$, (until time t) *and* state S_i at time t, given the model λ.



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 1

$$\alpha_t(i) = P(O_1\, O_2 \cdots O_t,\, q_t = S_i|\lambda) \tag{18}$$

We can solve for $\alpha_t(i)$ inductively, as follows:

1) Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N.$$

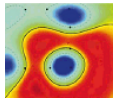
2) Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1$$

$$1 \le j \le N.$$

3) Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

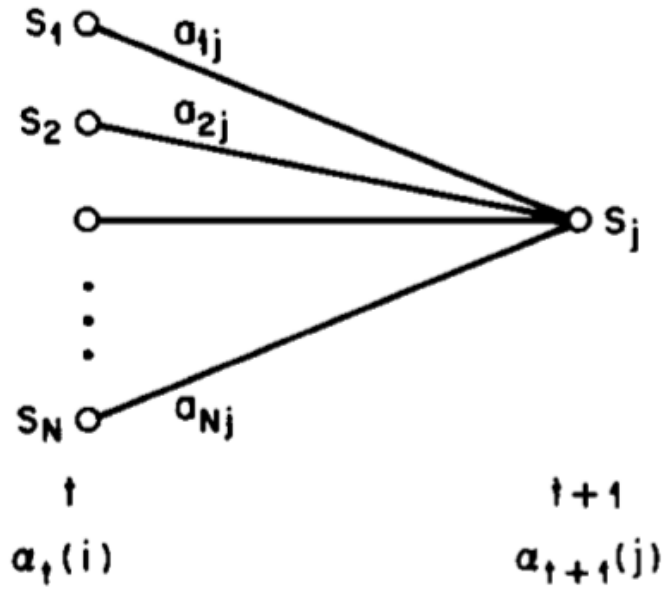


Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 1

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1$$

$$1 \le j \le N.$$



on the order of N^2T calculations, rather than $2TN^T$ as required by the direct calculation. (Again, to be precise, we

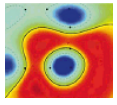
Source: L. Rabiner. A Tutorial on HMMs (1989).

Three Basic Problems for HMMs

Problem 1: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

Problem 2: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense (i.e., best "explains" the observations)?

Problem 3: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?



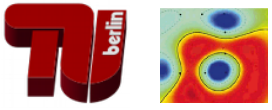
Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 2

Viterbi Algorithm [21], [22]: To find the single best state sequence, $Q = \{q_1\, q_2 \cdots q_T\}$, for the given observation sequence $O = \{O_1\, O_2 \cdots O_T\}$, we need to define the quantity

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P[q_1\, q_2 \cdots q_t = i, O_1\, O_2 \cdots O_t | \lambda]$$

i.e., $\delta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state S_i. By induction we have



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 2

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P[q_1\, q_2 \cdots q_t = i, O_1\, O_2 \cdots O_t | \lambda]$$

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N$$

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T$$

$$1 \le j \le N$$

3) Termination:

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 2

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T$$

$$\psi_t(j) = \operatorname{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 1 \le j \le N$$

3) Termination:

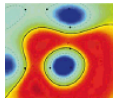
$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \operatorname{argmax}_{1 \le i \le N} [\delta_T(i)].$$

add pointers to recover the most likely sequence.

4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1.$$



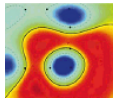
Source: L. Rabiner. A Tutorial on HMMs (1989).

Three Basic Problems for HMMs

Problem 1: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

Problem 2: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense (i.e., best "explains" the observations)?

Problem 3: How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

Baum-Welch method

probability of being in state

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

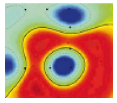
$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda).$$

reestimation of the parameters

$$\bar{\pi}_i = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

Baum-Welch method

probability of being in state

how to compute it?

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda).$$

reestimation of the parameters

$$\bar{\pi}_i = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

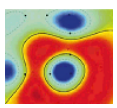
can be expressed

$$= \frac{\alpha_t(i)\,\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\,\beta_t(i)}$$

with

$$\alpha_t(i) = P(O_1\, O_2 \cdots O_t,\, q_t = S_i|\lambda) \qquad \text{forward variable}$$

$$\beta_t(i) = P(O_{t+1}\, O_{t+2} \cdots O_T | q_t = S_i,\, \lambda) \qquad \text{backward variable}$$



Source: L. Rabiner. A Tutorial on HMMs (1989).

The Forward-Backward Model

forward variable

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1$$

$$1 \le j \le N.$$

backward variable

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j),$$

$$t = T-1, T-2, \cdots, 1, \quad 1 \le i \le N. \qquad (25)$$

Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

Baum-Welch method

probability of being in state

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

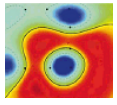
$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda).$$

reestimation of the parameters

$$\bar{\pi}_i = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$



Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

Baum-Welch method

probability of being in state

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda).$$

reestimation of the parameters

$$\bar{\pi}_i = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

Is it optimal?

Source: L. Rabiner. A Tutorial on HMMs (1989).

Solution of Problem 3

A

$$\pi_i = \frac{\pi_i \frac{\partial P}{\partial \pi_i}}{\sum_{k=1}^{N} \pi_k \frac{\partial P}{\partial \pi_k}}$$

$$a_{ij} = \frac{a_{ij} \frac{\partial P}{\partial a_{ij}}}{\sum_{k=1}^{N} a_{ik} \frac{\partial P}{\partial a_{ik}}} \qquad b_j(k) = \frac{b_j(k) \frac{\partial P}{\partial b_j(k)}}{\sum_{\ell=1}^{M} b_j(\ell) \frac{\partial P}{\partial b_j(\ell)}}.$$

B

$$\bar{\pi}_i = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad \bar{b}_j(k) = \frac{\sum_{t=1 \text{ s.t. } O_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

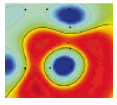
By appropriate manipulation of A the right-hand sides of each equation can be readily converted to be *identical* to the right-hand sides of each part of (40a)–(40c) B, thereby showing that the reestimation formulas are indeed exactly correct at critical points of *P*. In fact the form of is essen-

Source: L. Rabiner. A Tutorial on HMMs (1989).

Beyond HMMs

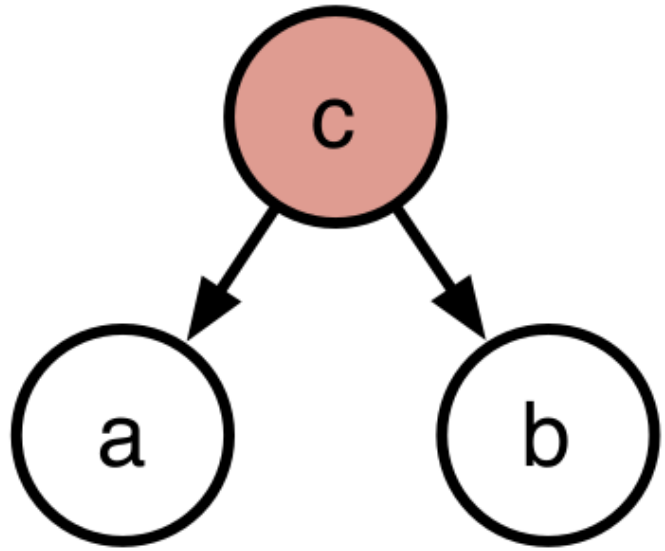
- HMMs are directed graphical models that match prior knowledge about the modeled task (latent states generates observations, current latent state generates next latent state).

- For more general models, the causality may be unknown. Setting the causality wrong may introduce a modeling bias.



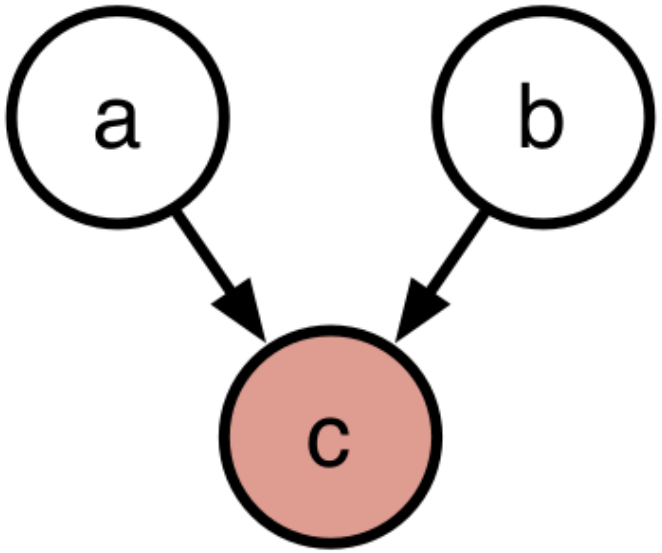
Conditional Independence

$$
\begin{aligned}
p(a,b,c) &= p(a|c)p(b|c)p(c) \\
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= \frac{p(a|c)p(b|c)p(c)}{p(c)} \\
&= p(a|c)p(b|c) \\
&\Rightarrow a \perp\!\!\!\perp b|c
\end{aligned}
$$

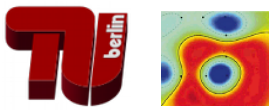


Reversing Causality

$$
\begin{aligned}
p(a,b,c) &= p(a)p(b)p(c|a,b) \\
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= \frac{p(a)p(b)p(c|a,b)}{p(c)} \\
&= p(a)p(b)\frac{p(c|a,b)}{p(c)}
\end{aligned}
$$

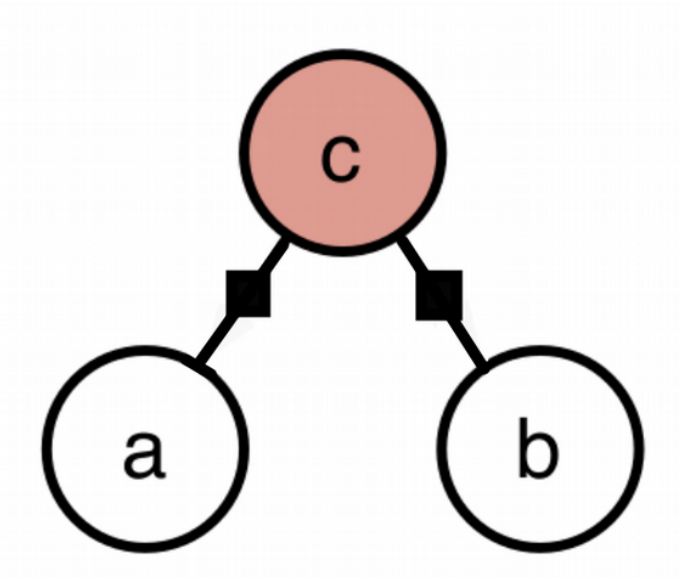


Variables a and b are no longer conditionally dependent. This effect is called “explaining away”.



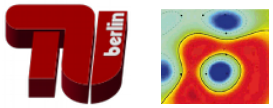
Removing Causality

$$p(a, b, c) = \frac{1}{Z}\Psi_1(a, b)\Psi_2(b, c)$$



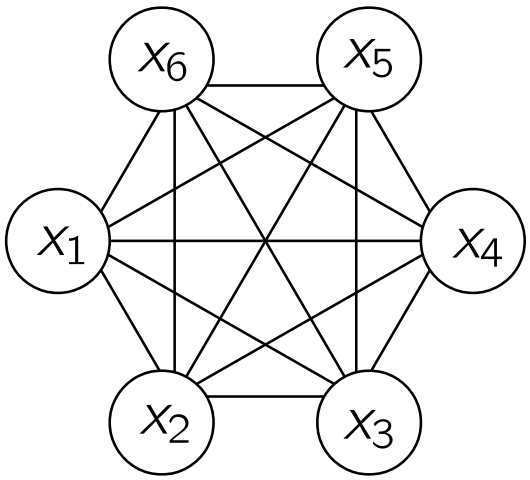
Probabilities are replaced by more general potential functions.

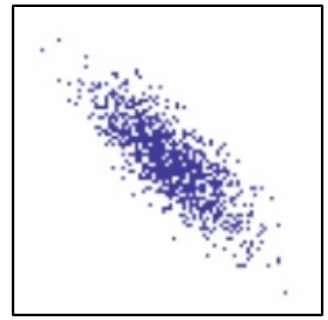
We introduce a normalization term Z.



Examples of Undirected Models

Gaussian Distribution:





$$\boldsymbol{x} \in \mathbb{R}^d$$

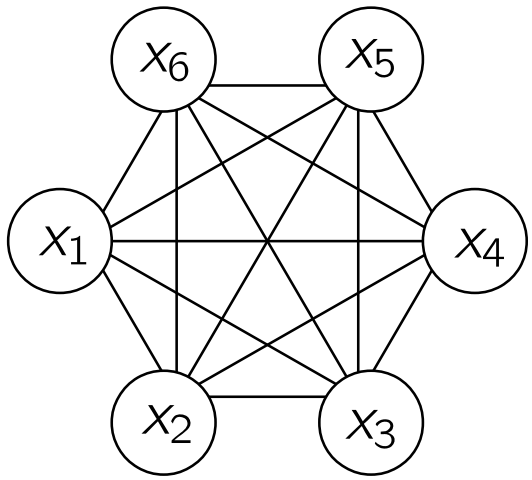
$$p_\theta = \mathcal{N}(0, \Sigma) \qquad S = \Sigma^{-1}$$

$$p_\theta(\boldsymbol{x}) = \frac{1}{Z(\theta)} \exp\left(-\frac{1}{2} \sum_{ij} x_i s_{ij} x_j\right)$$

$$= \frac{1}{Z(\theta)} \prod_{ij} \underbrace{\exp\left(-\frac{1}{2} x_i s_{ij} x_j\right)}_{\Psi_{ij}(x_i, x_j)}$$

Examples of Undirected Models

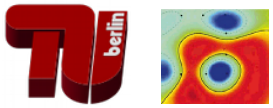
Boltzmann Machine:



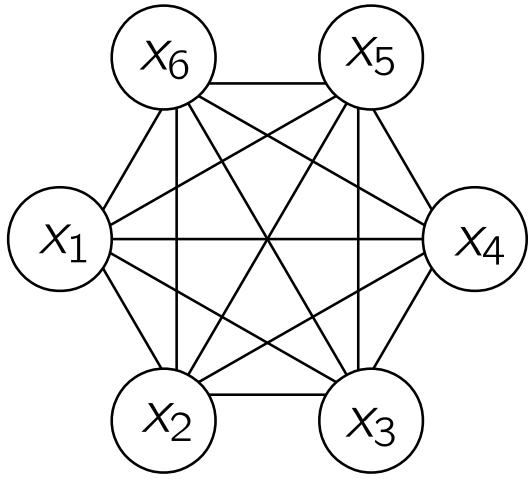
$$\boldsymbol{x} \in \{0,1\}^d$$

$$p(\boldsymbol{x}) = \frac{1}{Z(\theta)} \prod_{ij} \exp(x_i w_{ij} x_j)$$

Like for the Gaussian distribution, the model is composed of pairwise interactions. Normalization term is hard to evaluate, however, it remains easy to compute conditional probabilities.



Examples of Undirected Models



$$\boldsymbol{x} \in \{0,1\}^d$$

$$p(\boldsymbol{x}) \propto \prod_{ij} \exp(x_i w_{ij} x_j)$$

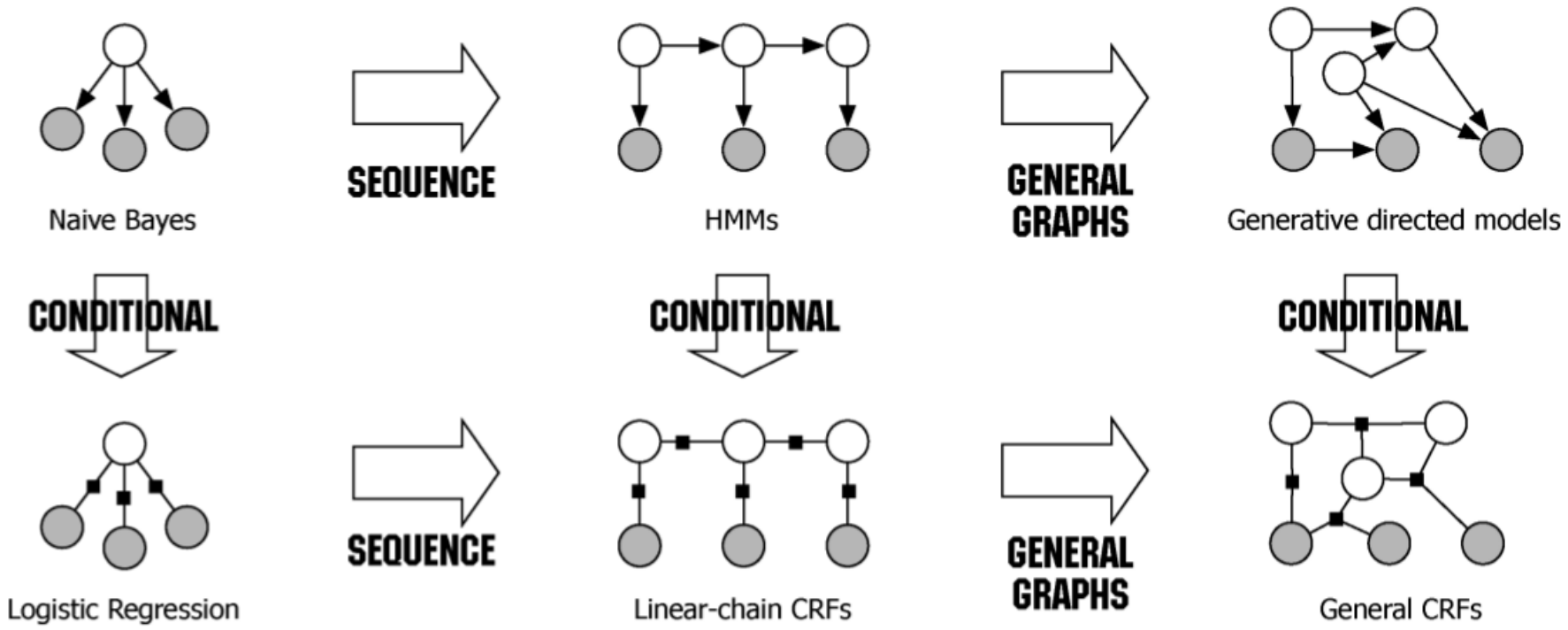
Conditional probabilities

$$p(x_i = 1|\boldsymbol{x}_{-i}) = \frac{p(x_i = 1, \boldsymbol{x}_{-i})}{\sum_{x_i \in \{0,1\}} p(x_i, \boldsymbol{x}_{-i})}$$

$$= \frac{\prod_j \exp(1 w_{ij} x_j) \prod_{k \in \{-i\}} \exp(x_k w_{kj} x_j)}{\sum_{i \in \{0,1\}} \prod_j \exp(x_i w_{ij} x_j) \prod_{k \in \{-i\}} \exp(x_k w_{kj} x_j)}$$

$$= \frac{\exp\left(\sum_j w_{ij} x_j\right)}{1 + \exp\left(\sum_j w_{ij} x_j\right)} = \text{sigm}\left(\sum_j w_{ij} x_j\right)$$

Directed vs. Undirected Models



Source: C. Sutton. An Introduction to CRFs (2011).