Human-Oriented Robotics

Basics of Probabilistic Reasoning

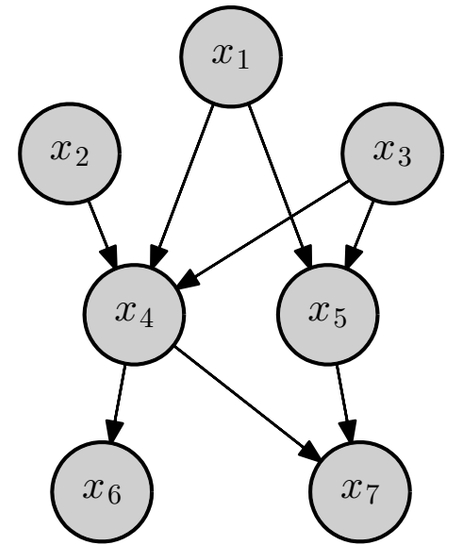
Kai Arras

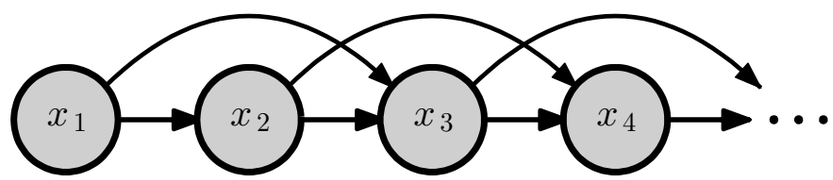
Social Robotics Lab, University of Freiburg

Human-Oriented Robotics
Prof. Kai Arras
Social Robotics Lab

Contents

- Probabilistic Reasoning
- Joint distribution
- Probabilistic Graphical Models
- Bayesian Networks
- Markov Models
- State Space Models





What is Reasoning?

- Reasoning is taking **available information** and reaching a **conclusion**
- A conclusion can be about what **might be true** in the world or about **how to act**
- The former is typically an **estimation** problem, the latter is typically a **decision** and **planning** problem
- Examples
  - A doctor takes information about a patient's symptoms to reach a conclusion about both his/her disease and treatment
  - A mobile robot senses its surrounding to reach a conclusion about the state of the environment and of itself and the next motion commands

What is Probabilistic Reasoning?

- **Reasoning under uncertainty** using probability theory as a framework

What is Probabilistic Reasoning?

- In probabilistic reasoning we focus on **models for complex systems** that involve a significant amount of **uncertainty**
- Such models can be acquired either through **learning** from data or from domain knowledge of **human experts**
- They typically involve sets of **random variables**
  - Example: a medical diagnosis domain may involve dozens or hundreds of symptoms, possible diseases, patient dispositions, and other influences. Each of those factor will be described by a **discrete** (e.g. disease A, B, C, ...) or **continuous** (e.g. fever temperature) **random variable**
- The task is then to reason probabilistically about the **values** of one or more of the variables given **observations** about some others
- In order to do so, we estimate a **joint probability distribution** over the involved random variables

Joint Distributions

- Joint probability distributions are very powerful models as they describe the **entire domain** and allow for a **broad range of interesting queries**, for instance, via marginalization
- For example, we can observe that variable x_i takes on value x_i* and ask what the probability distribution is over values of another variable x_j

- Example

  Consider a simple diagnosis system with two diseases (flu and hay fever), a 4-valued variable season, and two symptoms (running nose and muscle pain). Diseases and symptoms are either present or absent. Thus, our probability space has 2 x 2 x 4 x 2 x 2 = **64 values**

  Using a joint distribution over this space, we can, for example, ask questions such as how likely a patient with running nose but no muscle pain is to have flu in autumn

  Formally: $p(F = \text{true} | S = \text{autumn}, R = \text{true}, M = \text{false}) = ?$

Basics of Probabilistic Reasoning

- Specifying a joint distribution of 64 possible values seems feasible but what about a **larger, more realistic** diagnosis problem with dozens or hundreds of relevant attributes?
- With n variables that each can take m possible values, the joint distribution requires the specification of $m^n - 1$ values
- The explicit representation of such joint distributions is **unmanageable:**
  - **Computationally**, inference in such distributions is extremely expensive, if not intractable
  - **Cognitively** when defined from domain knowledge of human experts. It is impossible to acquire so many numbers from people
  - **Statistically** when such models are learned from data. We would need a huge amount of training data to estimate this many parameters robustly
- This was the **main barrier** to the adoption of probabilistic methods for expert systems until the development of probabilistic graphical models in the 1980s and 90s

Probabilistic Graphical Models

- Probabilistic graphical models provide a framework for **exploiting structure** in joint distributions by using a **graph-based** representation
- Let us start by considering a joint distribution over three random variables x_1, x_2, x_3. By application of the chain rule, we can write

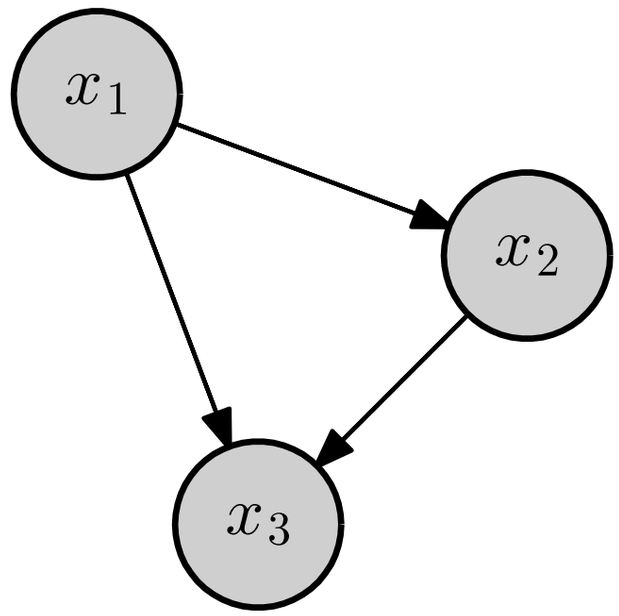
$$p(x_1, x_2, x_3) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2)$$

- To represent this decomposition in terms of a simple **graphical model,** we proceed as follows:

1. We introduce a **node** for each of the random variables and associate each node with the corresponding conditional distribution
2. For each conditional distribution we add directed **edges** (arrows) from the nodes of the corresponding conditioning variables

Probabilistic Graphical Models

- The result is a directed graphical model representing the joint distribution over x_1, x_2, x_3



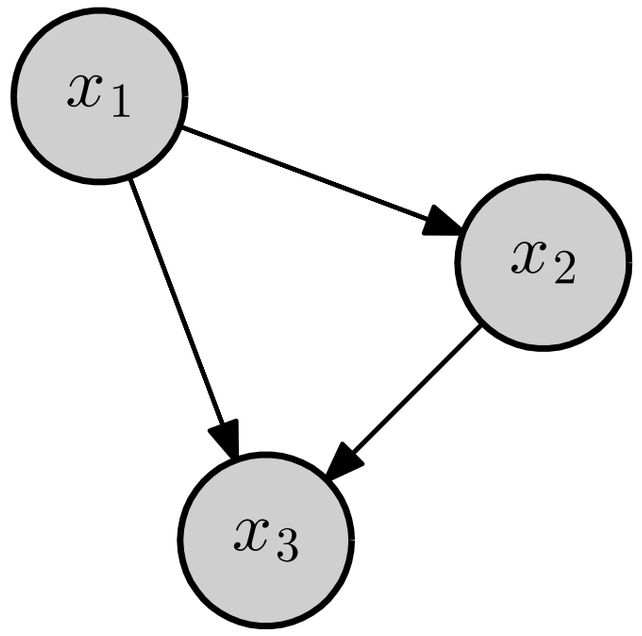
$$p(x_1, x_2, x_3) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2)$$

- If there is a link going from node x_1 to node x_2, then we say that the node x_1 is the **parent** of node x_2, and we say that node x_2 is the **child** of node x_1

Human-Oriented Robotics
Prof. Kai Arras
Social Robotics Lab

Probabilistic Graphical Models

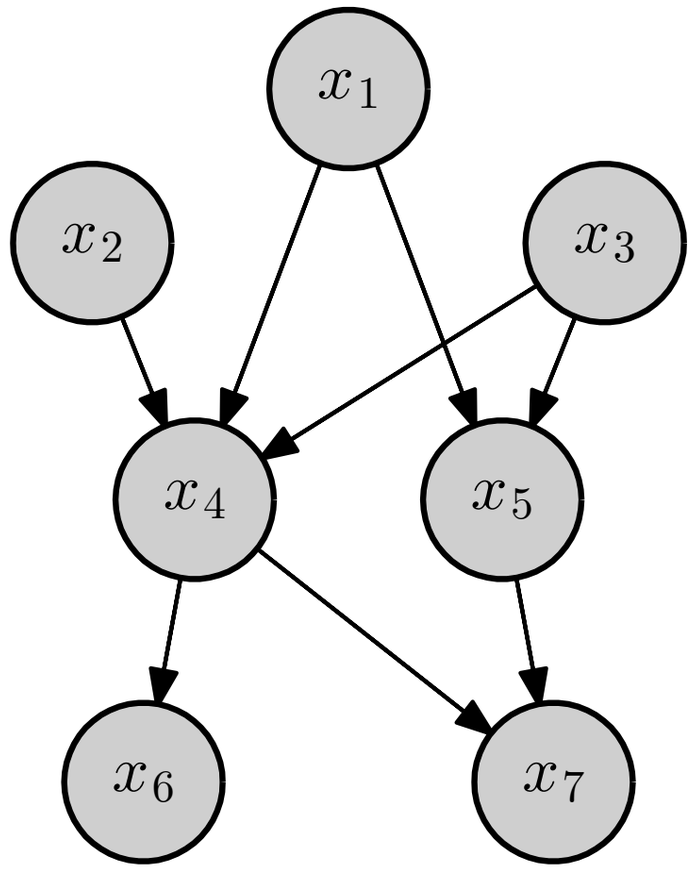
- This procedure scales to joint distributions over **arbitrary** numbers of variables
- Such distributions can be written as a product of conditional probabilities, one for each variable, obtained by **repeated applications** of the chain rule
- The resulting graphs are said to be **fully connected** because there is a link between every pair of nodes



- It is the **absence of links** in the graph that conveys the interesting information about the properties of the joint distribution

Probabilistic Graphical Models

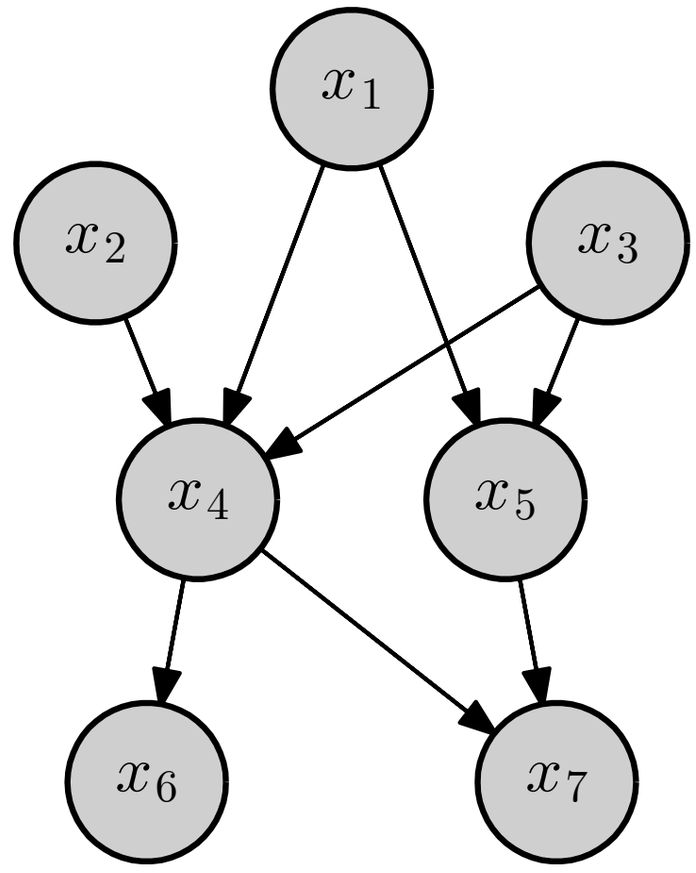
- To illustrate this, let us consider a directed graph describing the joint distribution over variables x_1 to x_7 which is **not** fully connected. There is no link , for example, from x_1 to x_2 or from x_3 to x_7
- We now go **backwards** and derive the joint probability distribution from the graph
- There will be **seven factors**, one for each node in the graph. Each factor is a conditional distribution, conditioned **only on its parents**



Probabilistic Graphical Models

- With $\mathbf{x} = \{x_1, \ldots, x_7\}$ we find

$$
\begin{aligned}
p(\mathbf{x}) \quad = \quad & p(x_1)\, p(x_2)\, p(x_3) \\
& \cdot\, p(x_4|x_1, x_2, x_3)\, p(x_5|x_1, x_3) \\
& \cdot\, p(x_6|x_4)\, p(x_7|x_4, x_5)
\end{aligned}
$$



- We can now state the **general** relation-ship between a given directed graph and the corresponding distribution
- With $\mathbf{x} = \{x_1, \ldots, x_K\}$ and pa_k being the parents of x_k, we have

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Probabilistic Graphical Models

- Coming back to our medical diagnosis domain with variables flu *(F)*, hay fever *(H)*, season *(S)*, running nose *(R)*, and muscle pain *(M)*. From our own "expertise" we can state the following conditional independencies
  - Flu only depends on season which contains all relevant information for flu. Given season, flu is independent on anything else: $p(F|S)$
  - The same applies for hay fever, it only depends on season: $p(H|S)$
  - Muscle pain is only caused by flu. Given flu, muscle pain is independent on anything else: $p(M|F)$
  - Season itself does not depend on anything: $p(S)$
  - Running nose depends on flu and hay fever. These variables contain all relevant information: $p(R|F,H)$
- Repeated application of the **chain rule** (in a good ordering) yields

$$p(S,F,H,R,M) = p(S)\,p(F|S)\,p(H|S,F)\,p(R|S,F,H)\,p(M|S,F,H,R)$$

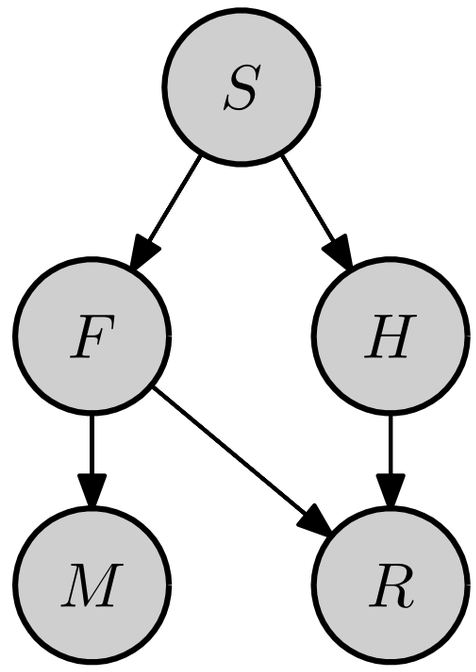
Basics of Probabilistic Reasoning

- Now let's simplify

$$p(S,F,H,R,M) = p(S)\, p(F|S)\, p(H|S,F)\, p(R|S,F,H)\, p(M|S,F,H,R)$$

and we obtain the following factorization and graphical model

$$p(S,F,H,R,M) = p(S)\, p(F|S)\, p(H|S)\, p(R|F,H)\, p(M|F)$$



What have we **gained**?

- This parametrization is **significantly more compact**, requiring only $4 + 4 + 4 + 4 + 2 = 18$ values as opposed to 64 values

  (Note that the numbers of non-redundant parameters are 17 and 63, as the sum over all entries in the joint distribution must sum to 1)

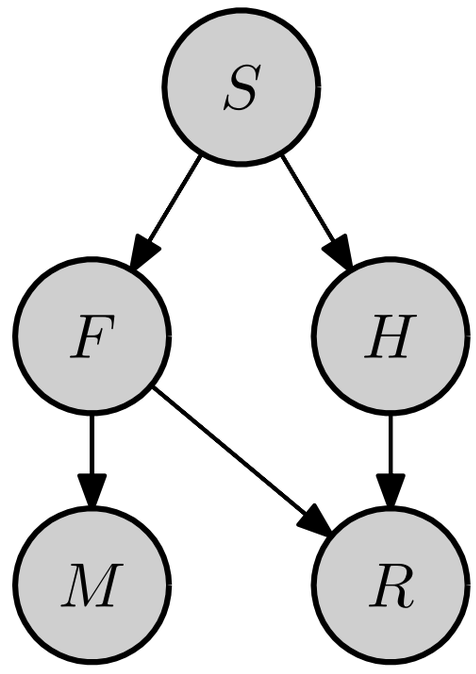
Basics of Probabilistic Reasoning

- Now let's simplify

$$p(S,F,H,R,M) = p(S)\, p(F|S)\, p(H|S,F)\, p(R|S,F,H)\, p(M|S,F,H,R)$$

and we obtain the following factorization and graphical model

$$p(S,F,H,R,M) = p(S)\, p(F|S)\, p(H|S)\, p(R|F,H)\, p(M|F)$$



What have we **gained**?

- This parametrization is **significantly more compact**, requiring only $4 + 4 + 4 + 4 + 2 = 18$ values as opposed to 64 values

  (Note that the numbers of non-redundant parameters are 17 and 63, as the sum over all entries in the joint distribution must sum to 1)

Probabilistic Graphical Models

- In general, factored representations may have **exponentially fewer parameters** than the joint distribution. The result is
  - Lower **sample complexity** (less data for learning)
  - Lower **time complexity** (less time for inference)

- Benefits of the **graph representation** include
  - **Modular** representation of knowledge makes it easier e.g. to specify complex models
  - **Local**, distributed algorithms for inference and learning
  - **Intuitive** interpretation and visualization of a model's structure

- One way to think about conditional independence relations is to consider them as **redundancies** in the joint probability distribution, another way is to consider them as **structure** in the distribution

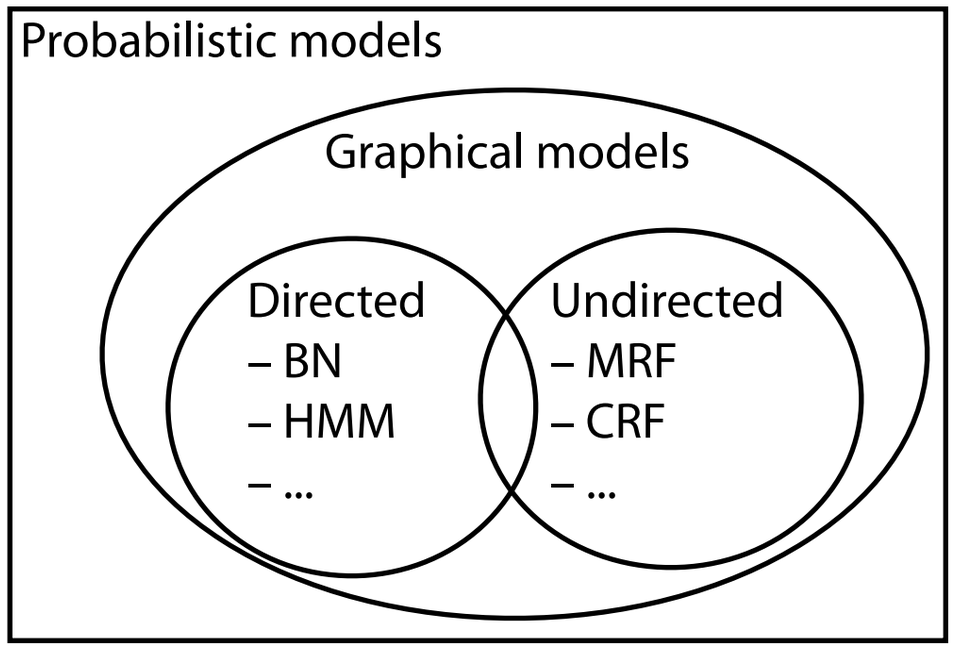
Probabilistic Graphical Models

- A joint distribution can be expanded by the chain rule using **any order of variables**, the result will be the same. However, each ordering produces a different graph with varying numbers of links/probabilities to be specified
- Which ordering should we choose?
- One rule is that **higher-numbered variables** may correspond to terminal nodes that represent observations (e.g. symptoms), **lower-numbered variables** may correspond to latent or hidden variables
- The problem of finding an optimal ordering can be hard in general, human domain knowledge and heuristics are used in practice

- Notice that while so far, nodes corresponded to scalar random variables, they can also stand for a **group of variables** such as a random vector

Probabilistic Graphical Models

- We have considered **directed** graphical models whose links have a particular direction indicated by arrows
- Such models are called **Bayesian Networks** (BN)
- The other major class of graphical models are **Markov Networks**, also known as **undirected** graphical models, in which links have **no** direction. A prominent example are Markov Random Fields (MRF)
- Directed graphs are useful for expressing **causal relationships**, whereas undirected graphs are better at expressing **soft constraints** between variables



Probabilistic Graphical Models

- **Directed** graphs are subject to an important restriction: there must be **no directed cycles**. It should **not** be possible to move along the links from node to node and ending up back at the start node
- This is why such models are also called **directed acyclic graphs** (DAG)
- In this course, we will only consider Bayesian Networks

- Application example

  One of the earliest applications of Bayesian Networks was medical diagnosis. They were quickly found to outperform non-probabilistic expert systems in the 1980s and 90s. A prominent example is the Pathfinder project [2, p.67] which evolved over several generations into a powerful diagnosis system for more than 60 different diseases. Evaluations showed that diagnostic accuracy of Pathfinder was at least as good as that of the medical experts who designed the system and significantly better than less expert pathologists

Inference in Probabilistic Graphical Models

- So far, we have introduced the **representation** of probabilistic graphical models. What about **inference** and **learning**?
- Let us exemplify our statement that joint probability distributions are powerful because they allow for a broad range of interesting queries. The **most relevant two** query types are as follows:

- **Probability query:**
  with q being a set of **query variables** and $e = e^*$ being the **evidence** (a set of instantiated variable-value pairs), we can compute the **posterior probability distribution** $p(q|e = e^*)$ over the query variables given the evidence. Examples:
  - **Robot localization:** e = camera image of the environment, q = robot pose
  - **Medical diagnosis:** e = set of symptoms, q = diseases
  - **Speech recognition:** e = sequence of acoustical signals, q = spoken word

Inference in Probabilistic Graphical Models

- **Maximum a posteriori (MAP) query:**
  finding the most likely values of a variable given evidence $e = e^*$

$$\arg\max_q \; p(q|e = e^*)$$

- The result can also be seen as the **most probable explanation**
- There might be **more than one solution** to this query in cases of multiple modes of the underlying posterior distribution

- **All variables in the domain** can be query or evidence variables
- The process of answering queries is called **inference**

Inference in Probabilistic Graphical Models

- One of the key advantages of graphical models is that by **leveraging the structure** of the joint distribution, **inference algorithms are particularly efficient** and scale much better than brute force approaches
- Generally, the complexity of inference algorithms in graphical models is inversely proportional to the **sparsity** of the graph (exploiting the absence of links)

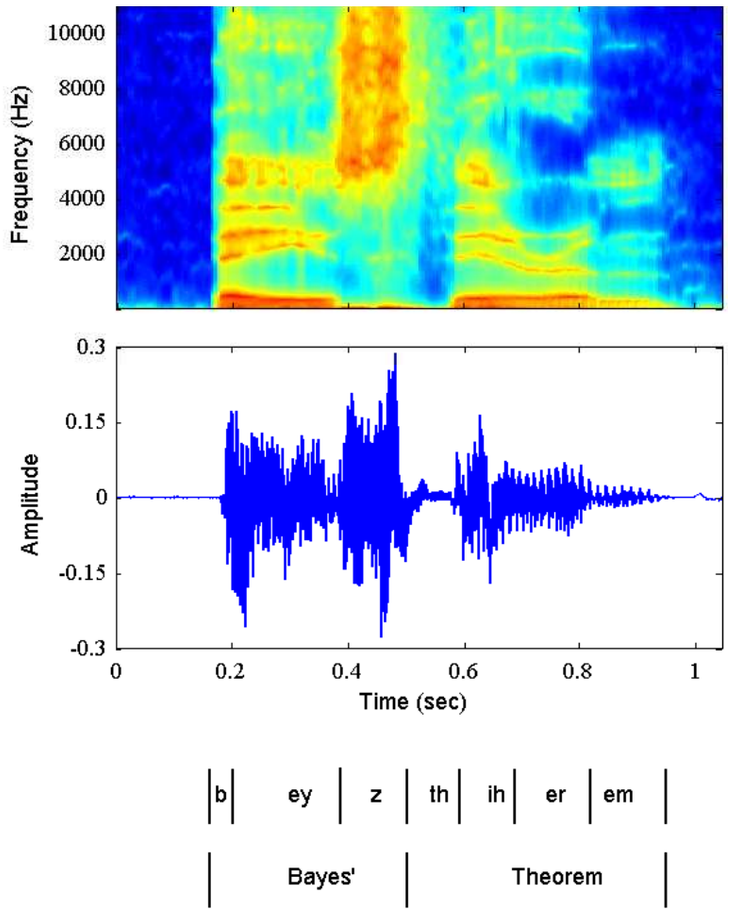
- Here we will consider inference for graphical models in particular for two important Bayesian network types that describe **sequential data: hidden Markov models** and **linear dynamical systems**
- These are examples of **temporal models**

Sequential Data

- Sequential data often arise through **measurements of time series**, for example:
  - Rainfall measurements on successive days at a particular location
  - Daily currency exchange rates
  - Acoustic features at successive time frames used for speech recognition
  - A human's arm and hand movements used for sign language understanding
- **Other forms of sequential data** e.g. over space exist as well. The models considered here equally apply to them
- In applications, we typically wish to be able to **predict the next value** given observations of the previous values (think of financial forecasting)
- We expect that **recent observations** are likely to be **more informative** than **more historical observations**

Sequential Data

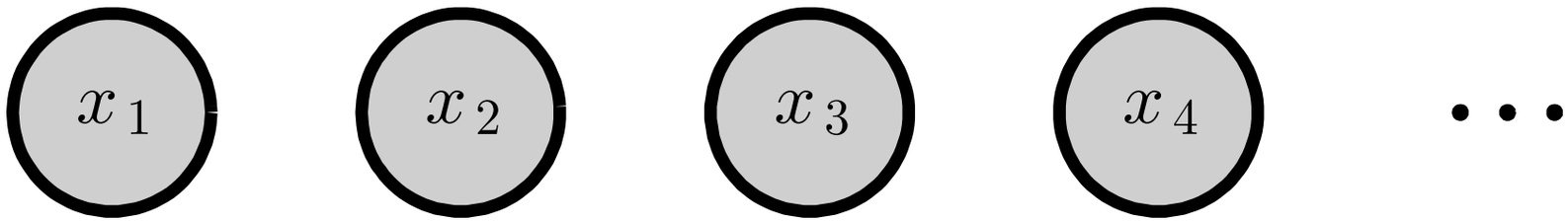
- This is the case when **successive values** in time series are **correlated**



- **Example:** spectrogram of the spoken word “Bayes’ theorem”

Sequential Data

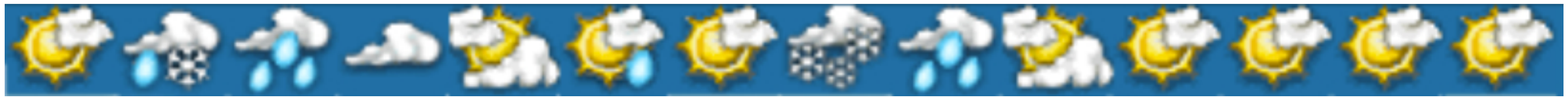
- The easiest way to treat sequential data would be simply to ignore the sequential aspect and consider the **observations as i.i.d. random variables** (independent and identically distributed)
- This would lead to the following graphical model



- Such a model **fails to exploit the sequential patterns** in the data
- An example of such **sequential patterns** is our weather

Sequential Data

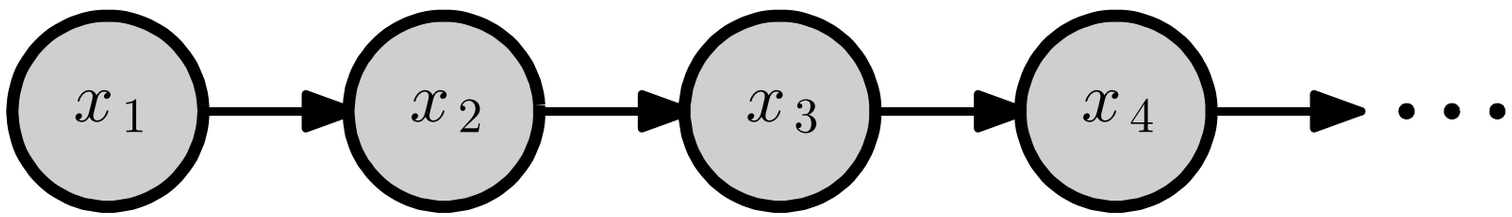
- Suppose we observe a binary variable "sunshine" and we wish to predict whether or not the sun will shine on the next day. If we treat the data as i.i.d., then the only information that we can extract from the data is the (a priori) relative frequency of sunny days
- However, we know that weather often exhibits **trends that may last for several days**. Thus, **observing today's weather** is of significant help in predicting if the sun will shine tomorrow



- Is there a model that allows us to exploit those correlation or trends?

Markov Models

- Consider a model that postulates dependencies of future observations **on all previous observations**. Such a model would be impractical because its complexity would **grow without limits** as the number of observations increases
- This leads us to consider **Markov Models**



- Markov models assume that future **predictions** are independent of all **but the most recent observations**

Markov Models

- Formally, we recall the **chain rule**

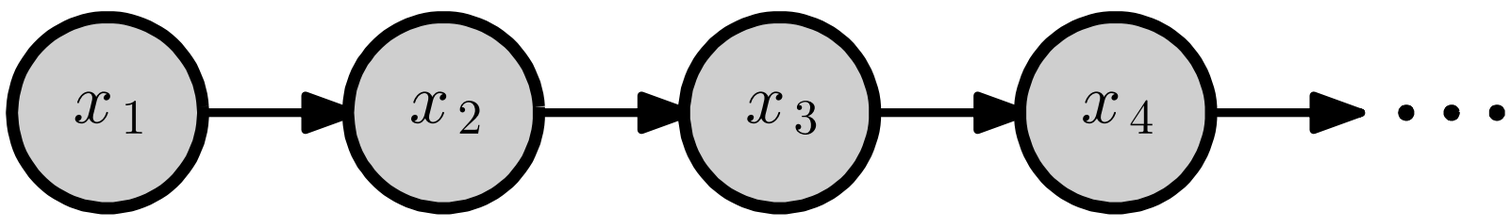
$$p(x_1, x_2, \ldots, x_K) = \prod_{i=1}^{K} p(x_i | x_1, \ldots, x_{i-1})$$

- If we now assume that each of the conditional distributions on the right hand side is independent of all previous observations **except the most recent one**,

$$\begin{aligned} p(x_1, x_2, \ldots, x_K) &= \prod_{i=1}^{K} p(x_i | x_1, \ldots, x_{i-1}) \\ &= p(x_1)\, p(x_2|x_1)\, p(x_3|\cancel{x_1}, x_2)\, p(x_4|\cancel{x_1}, \cancel{x_2}, x_3) \cdots \\ &\quad \cdot p(x_K|\cancel{x_1}, \cancel{x_2}, \ldots, x_{K-1}) \end{aligned}$$

Markov Models

- we obtain the **first-order Markov chain**

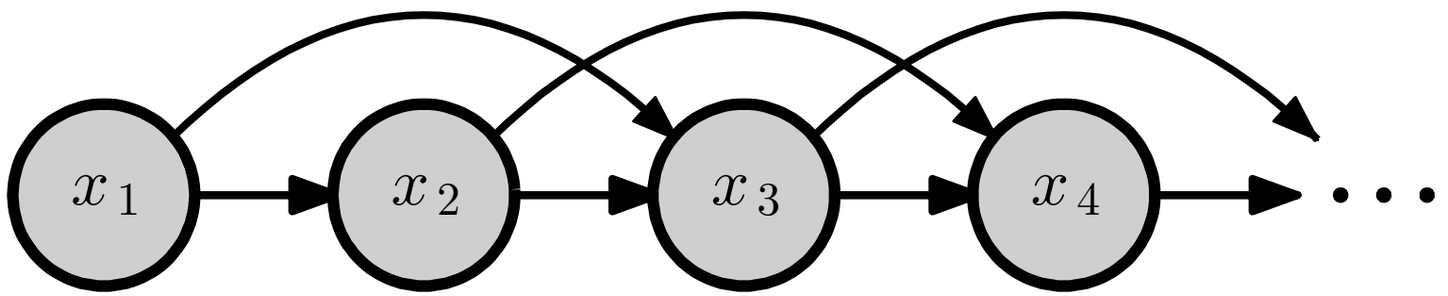


$$p(x_1, x_2, \ldots, x_K) = p(x_1) \prod_{i=2}^{K} p(x_i|x_{i-1})$$

- We further make the (weak) assumption that the conditional distributions $p(x_i|x_{i-1})$ are the same for all i, corresponding to the model of a **station-ary** time series. This is also known as a **homogeneous** Markov model

Markov Models

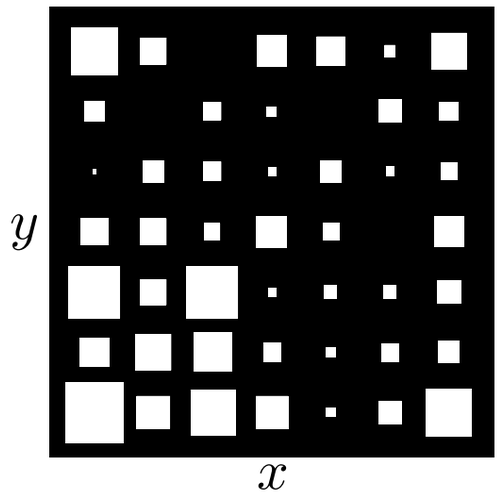
- A more flexible class of models that may be able to even better capture trends in the data, are **higher-order Markov models** in which earlier observations can also have an influence
- If we allow the predictions to depend on the **two** previous observations, we obtain the **second-order Markov chain**



$$p(x_1, x_2, \ldots, x_K) = p(x_1)\, p(x_2|x_1) \prod_{i=3}^{K} p(x_i|x_{i-2}, x_{i-1})$$

Markov Models

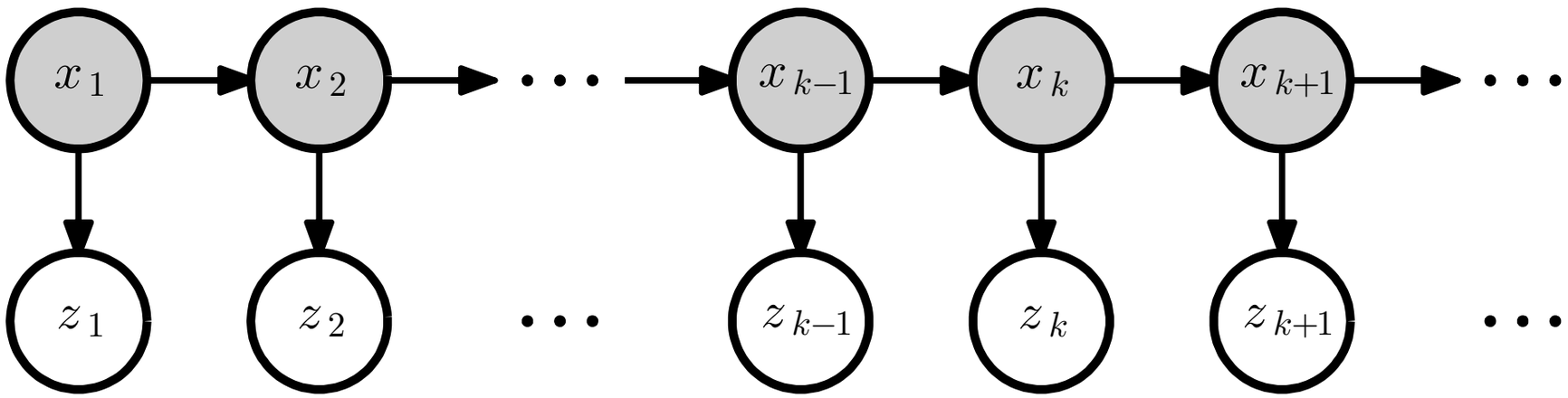
- In an m^{th}-order Markov model the conditional distribution for a particular variable depends on m **previous variables**. The resulting model may be powerful but expensive:
- Suppose observations are discrete random variables that can take K possible values. Then, the conditional distribution $p(x_i|x_{i-1})$ has $K(K-1)$ **parameters** (K–1 parameters for each of the K states of x_{i-1})
- This scales to $K^{m-1}(K-1)$ number of parameters for an m^{th}-order Markov model which is an **exponential growth** – impractical for large values of m



- There is another way to make our model more flexible...

State Space Model

- Let's add **latent or hidden variables** to our model, one for each random variable and let the latent variables form a Markov chain



- Notice the change in notation: we denote **latent** variables by x and **observations** by z (this notation is widely used in particular for LDS)
- It is sometimes common to **shade the nodes** of latent variables in the graphical representation

State Space Model

- In this model, we view the model to describe a system that **evolves on its own**, with **observations of it** occurring in a **separate** process
- This separation makes sense, for example, when observations are obtained from a **noisy sensor**
- This model is called **state space model** or **state observation model**

- **Latent** variables are also known as **hidden** variables. In the context of state space models, they are also called **states**
- They may be of **different type** and **dimensionality** than the observations

State Space Model

- In addition to the independence assumption of the first-order Markov model, we assume that **observations** at time index i are **conditionally independent of the entire state sequence** given the **state variable** at time index i
- The **joint distribution** of this model is derived as follows

$$
\begin{aligned}
p(x_1, \ldots, x_K, z_1, \ldots, z_K) \;=\; & p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots \\
& \cdot p(x_K|x_1, x_2, \ldots, x_{K-1}) \\
& \cdot p(z_1|x_1, x_2, \ldots, x_K)\, p(z_2|x_1, x_2, \ldots, x_K, z_1) \cdots \\
& \cdot p(z_K|x_1, \ldots, x_K, z_1, \ldots, z_{K-1})
\end{aligned}
$$

State Space Model

- In addition to the independence assumption of the first-order Markov model, we assume that **observations** at time index i are **conditionally independent of the entire state sequence** given the **state variable** at time index i
- The **joint distribution** of this model is derived as follows

$$
\begin{aligned}
p(x_1, \ldots, x_K, z_1, \ldots, z_K) \;=\; & p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots \\
& \cdot p(x_K|x_1, x_2, \ldots, x_{K-1}) \\
& \cdot p(z_1|x_1, x_2, \ldots, x_K)\, p(z_2|x_1, x_2, \ldots, x_K, z_1) \cdots \\
& \cdot p(z_K|x_1, \ldots, x_K, z_1, \ldots, z_{K-1})
\end{aligned}
$$

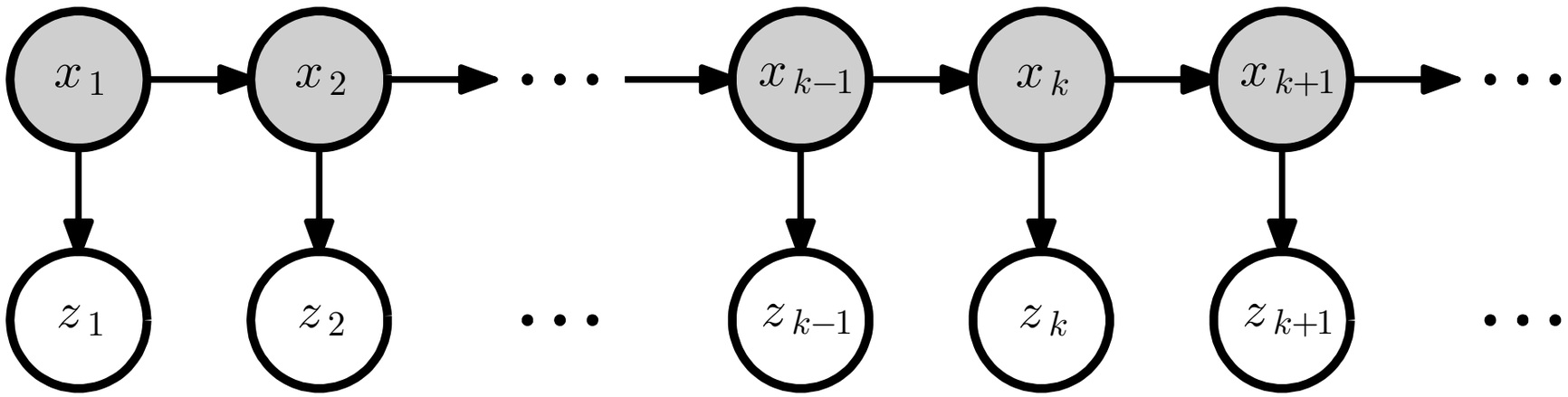
State Space Model

- In addition to the independence assumption of the first-order Markov model, we assume that **observations** at time index i are **conditionally independent of the entire state sequence** given the **state variable** at time index i
- The **joint distribution** of this model is derived as follows

$$
\begin{aligned}
p(x_1, \ldots, x_K, z_1, \ldots, z_K) &= p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots \\
&\quad \cdot p(x_K|x_1, x_2, \ldots, x_{K-1}) \\
&\quad \cdot p(z_1|x_1, x_2, \ldots, x_K)\, p(z_2|x_1, x_2, \ldots, x_K, z_1) \cdots \\
&\quad \cdot p(z_K|x_1, \ldots, x_K, z_1, \ldots, z_{K-1}) \\
&= p(x_1) \left[\prod_{i=2}^{K} p(x_i|x_{i-1}) \right] \prod_{i=1}^{K} p(z_i|x_i)
\end{aligned}
$$

State Space Model

- There are two important models for sequential data that are described by this graph



- If the latent variables are **discrete**, then we obtain a **hidden Markov Model** (HMM). Observed variables can either be discrete or continuous in HMMs
- If both the latent and the observed variables are **continuous**, then we obtain the **linear dynamical system** (LDS)

Summary

- Probabilistic graphical models represent a joint distribution over a domain of random variables **using a graph**
- The graph encodes a set of **conditional independence assumptions** that encode and leverage structure in the joint distribution
- There are two components to a Bayesian network
  - The **graph structure** (conditional independence assumptions)
  - The **numerical probabilities** (for each variable given its parents)
- Answering queries in a Bayesian network, called **inference** or **reasoning**, amounts to the computation of conditional probabilities
- **Markov models** are temporal models able to describe **sequential data**
- The **Markov property** denotes the assumption that variables in a Markov chain depend only on **the most recent** observation
- The **state space model** describes systems that **evolve on their own**, with **observations of it** occurring in a **separate** process

Sources Used for These Slides and Further Reading

The slides mainly follow the books by Bischop [1] (chapters 8 and 13) and Koller and Friedman [2] (chapters 1, 3 and 6). Small bits are taken from [3] and [4]

Bischop [1] and also Prince [4] have well written compact introductions to probabilistic graphical models. A comprehensive treatment of this topic is the book by Koller and Friedman [2].

[1] C.M. Bischop, “Pattern Recognition and Machine Learning”, Springer, 2nd ed., 2007. See http://research.microsoft.com/en-us/um/people/cmbishop/prml

[2] D. Koller, N. Friedman, “Probabilistic graphical models: principles and techniques”, MIT Press, 2009. See http://pgm.stanford.edu

[3] K. Murphy, “An introduction to Bayesian Networks and the Bayes Net Toolbox for Matlab”, MIT AI Lab, May 2003

[4] S.J.D. Prince, “Computer vision: models, learning and inference”, Cambridge University Press, 2012. See www.computervisionmodels.com