Graphical Models & Inference

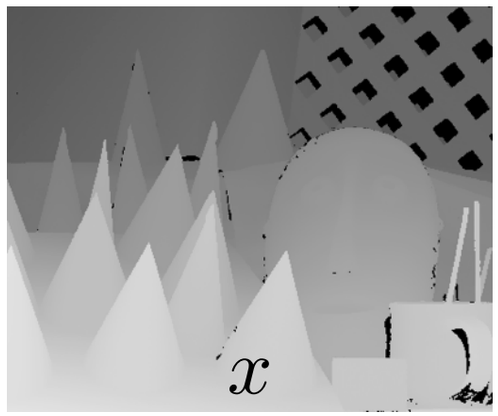
30.04.2014

TECHNISCHE UNIVERSITÄT DARMSTADT

visual inference

Last Lecture: Probabilistic Approach to Vision

Uncertain image measurements



Uncertain state of the world

- Model using posterior distribution: $p(\text{state}|\text{images})$
  - Describe the probability of the state of the world given the image measurements.
- How do we find the “best” state of the world?
  - Using probabilistic inference, e.g. we maximize w.r.t. state x

Last Lecture: Modeling the Posterior

TECHNISCHE UNIVERSITÄT DARMSTADT

- How can we model the posterior?
- We simplified the modeling problem by applying Bayes' rule (generative approach):

$$p(\text{state}|\text{images}) = \frac{p(\text{images}|\text{state}) \cdot p(\text{state})}{p(\text{images})}$$

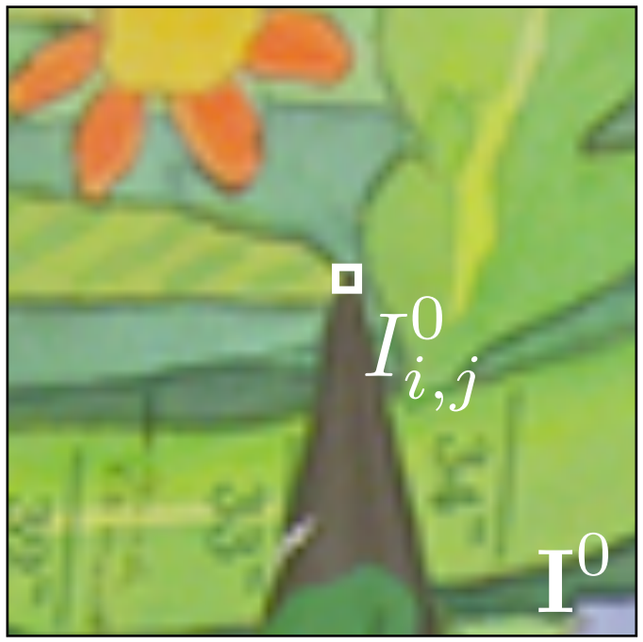
likelihood (observation model)

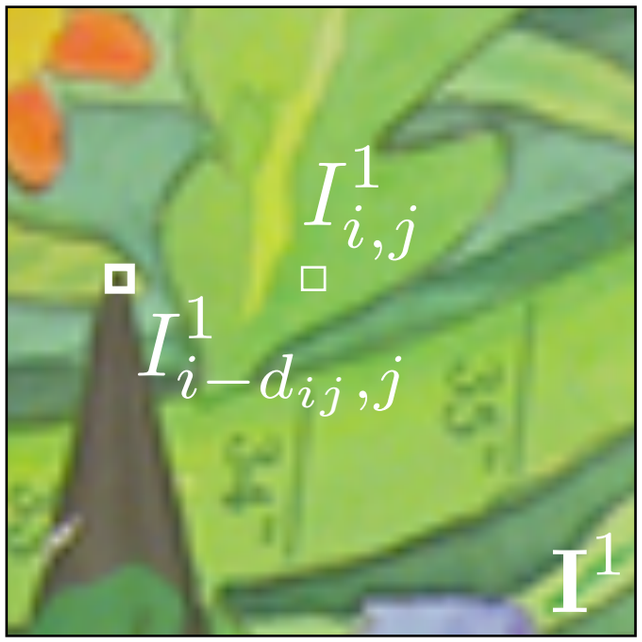
prior

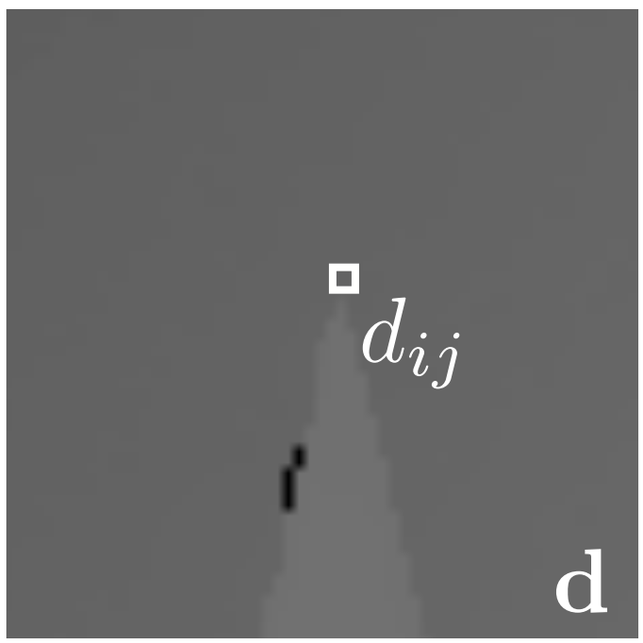
posterior

normalization term (constant)

Last Lecture: Modeling the Likelihood





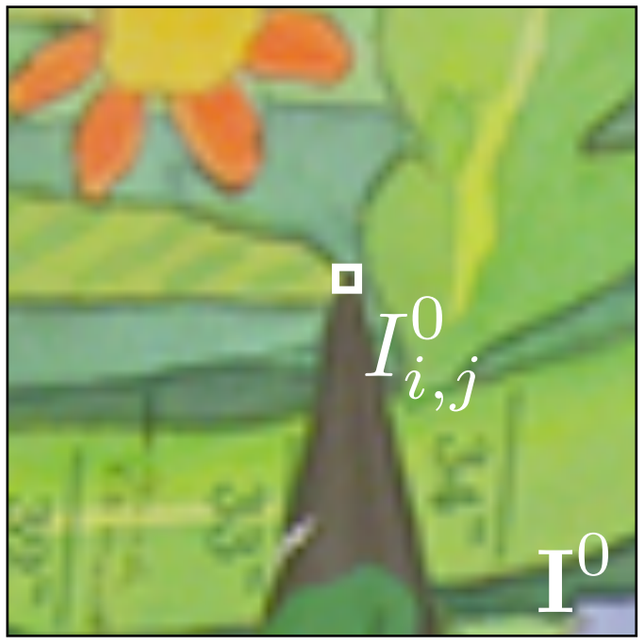


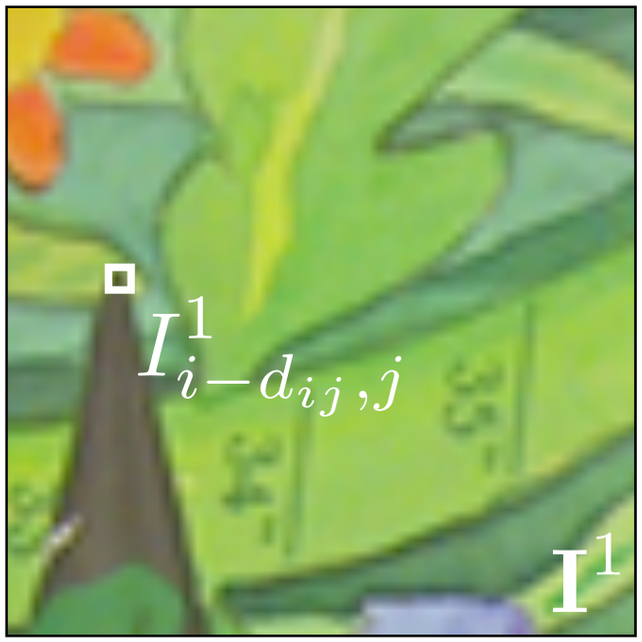
- A simple model:
  - We test how well the corresponding **pixels** match.

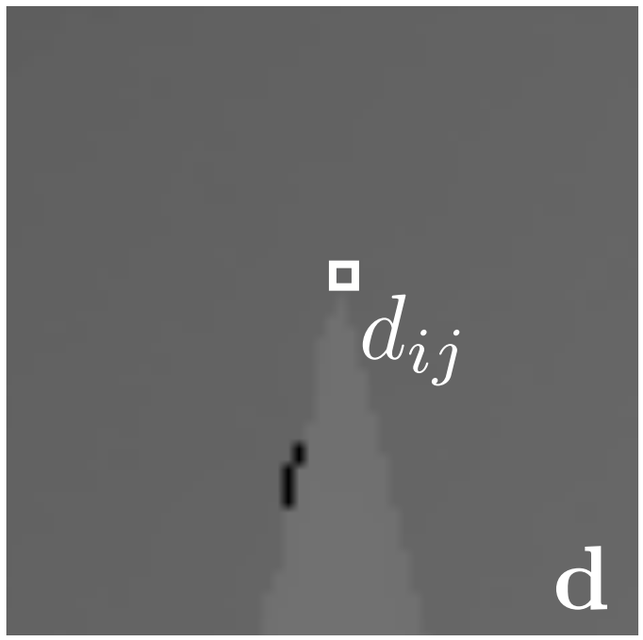
$$p(\mathbf{I}^0, \mathbf{I}^1 | \mathbf{d}) = \prod_{i,j} p(I^0_{i,j}, \mathbf{I}^1 | d_{ij})$$

$$= \prod_{i,j} f(I^0_{i,j} - I^1_{(i-d_{ij}),j})$$

Last Lecture: Modeling the Likelihood







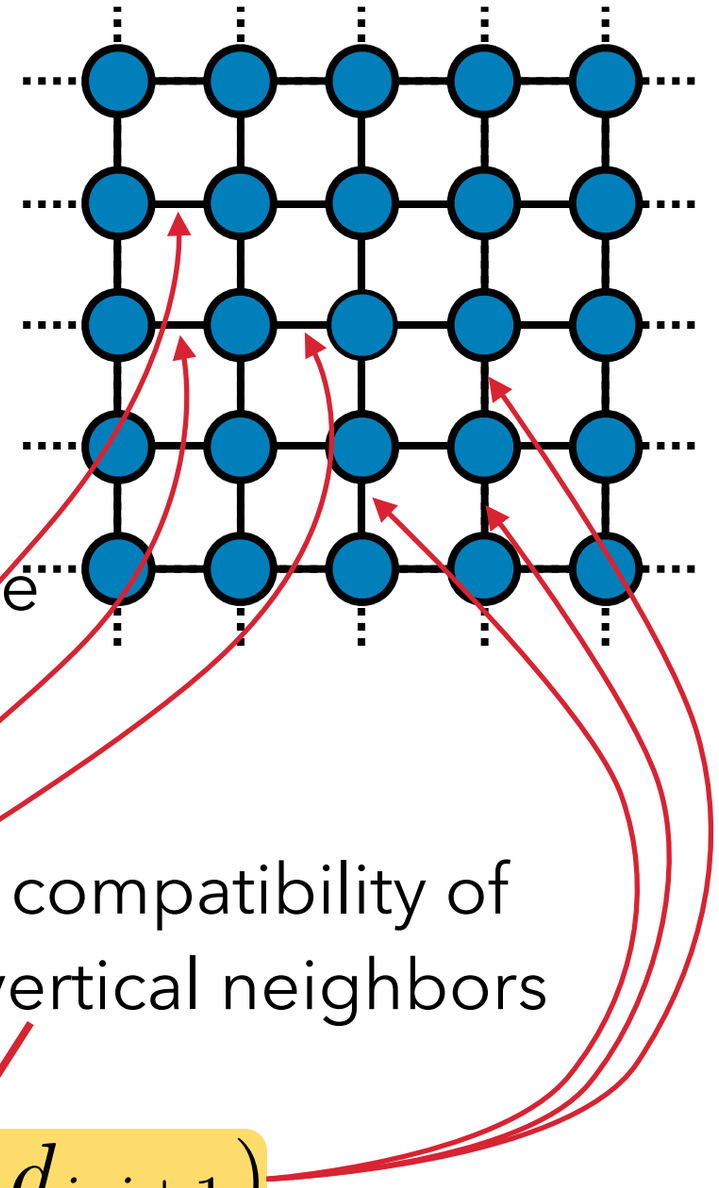
- $f(\cdot)$ is a probabilistic model of how well two pixels match that are related by the local disparity.
  - We assumed for simplicity that it is Gaussian. Other models are easily derived from this.

$$p(\mathbf{I}^0, \mathbf{I}^1 | \mathbf{d}) = \prod_{i,j} f(I^0_{i,j} - I^1_{(i-d_{ij}),j}) = \prod_{i,j} \mathcal{N}(I^0_{i,j} - I^1_{(i-d_{ij}),j};\ 0, \sigma^2)$$

Last Lecture: Markov Random Fields

TECHNISCHE UNIVERSITÄT DARMSTADT

- We formulated the prior as a so-called **graphical model**, or more specifically a so called **Markov random field**.
  - Each edge (in this particular graph) corresponds to a term in the prior that models how compatible two neighboring pixels are in terms of their disparity:



compatibility of horizontal neighbors

compatibility of vertical neighbors

$$p(\mathbf{d}) = \frac{1}{Z} \prod_{i,j} f_H(d_{i,j}, d_{i+1,j}) \cdot f_V(d_{i,j}, d_{i,j+1})$$

product over all the pixels

Potts Model

- Define very simple compatibility functions:

$$f_H(d_{i,j} - d_{i+1,j}) = \frac{1}{Z(T)} \exp\left\{ \frac{1}{T} \delta(d_{i,j} - d_{i+1,j}) \right\}$$

- Kronecker delta:

$$\delta(a-b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

- This prior:
  - Prefers to have the same disparities at neighboring pixels.
  - But allows for disparity discontinuities with no penalty for large discontinuities.
  - Is called a **Potts model**.
    - Originally from statistical physics (magnetism)

Graphical Models - What and Why?

TECHNISCHE UNIVERSITÄT DARMSTADT

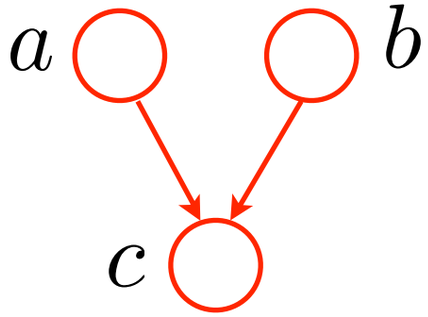
- Probabilistic graphical models:
  - Marriage between probability theory and graph theory.
  - Provide a natural tool for dealing with uncertainty and complexity.
  - Formalize and visualize the structure of a probabilistic models through a graph.
  - Give insight into the structure of a probabilistic model.
  - Become increasingly important for computer vision.
  - It would take a whole semester to get into all the details, so we will only review the very basics.

Graphical Models

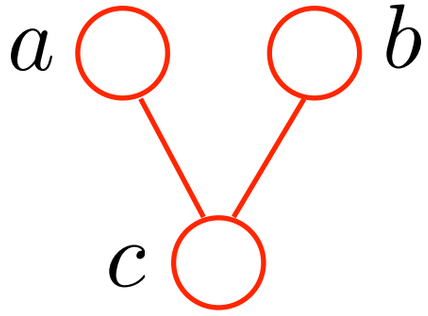
TECHNISCHE UNIVERSITÄT DARMSTADT

- There are two basic kinds of graphical models:
  - Directed graphical models, or Bayesian networks.
  - Undirected graphical models, or Markov random fields.

- Two key components:
  - Nodes
  - Edges (directed or undirected)



directed graphical model



undirected graphical model

- We will start with directed models and an example...

Example 1: Icy Roads

TECHNISCHE UNIVERSITÄT DARMSTADT

- Inspector Smith
  - is waiting for „Dr. Watson“ and „Mr. Holmes“.
  - He looks out the window and wonders whether the roads are icy.
  - Smith knows that the danger of have an accident is much larger, when the roads are icy.
- Smith receives a call
  - Dr. Watson had an accident.
  - Smith concludes that the roads are probably icy and that
  - the probability that Holmes will have an accident, too, is greater.

Example 1: Icy Roads

TECHNISCHE UNIVERSITÄT DARMSTADT

- Smith's secretary
  - says though that the temperature is above 0 degrees Celsius, so the streets cannot possibly be icy.

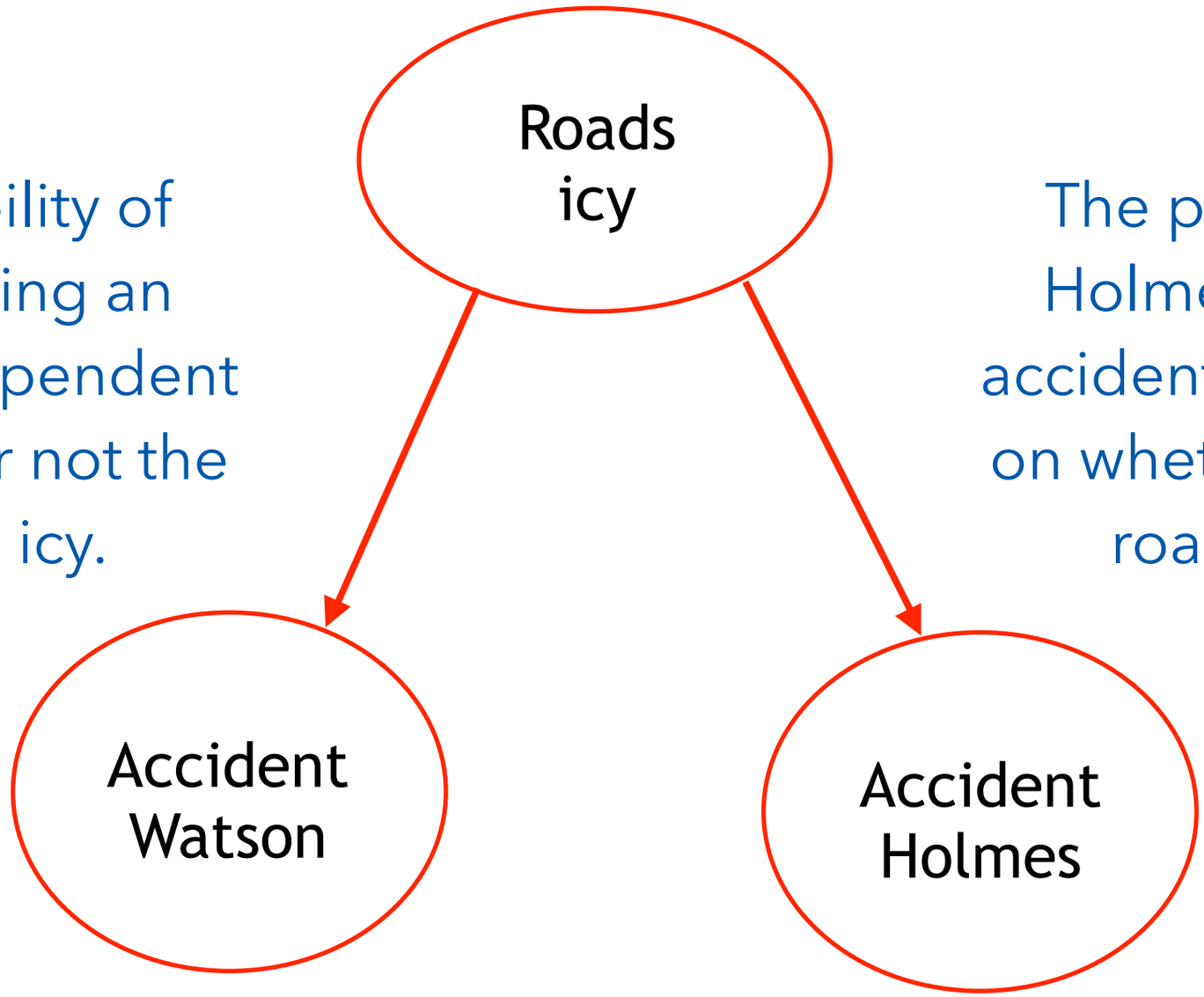
- Smith concludes
  - that the reason of Watson's accident cannot have been icy roads, and that there must have been another reason.
  - And so he also concludes that an accident of Holmes is less likely again.

Example 1: Icy Roads

TECHNISCHE UNIVERSITÄT DARMSTADT

- Directed graphical model / Bayesian network:

The probability of Watson having an accident is dependent on whether or not the roads are icy.



The probability of Holmes having an accident is dependent on whether or not the roads are icy.

Example 2: Wet Lawn

TECHNISCHE UNIVERSITÄT DARMSTADT

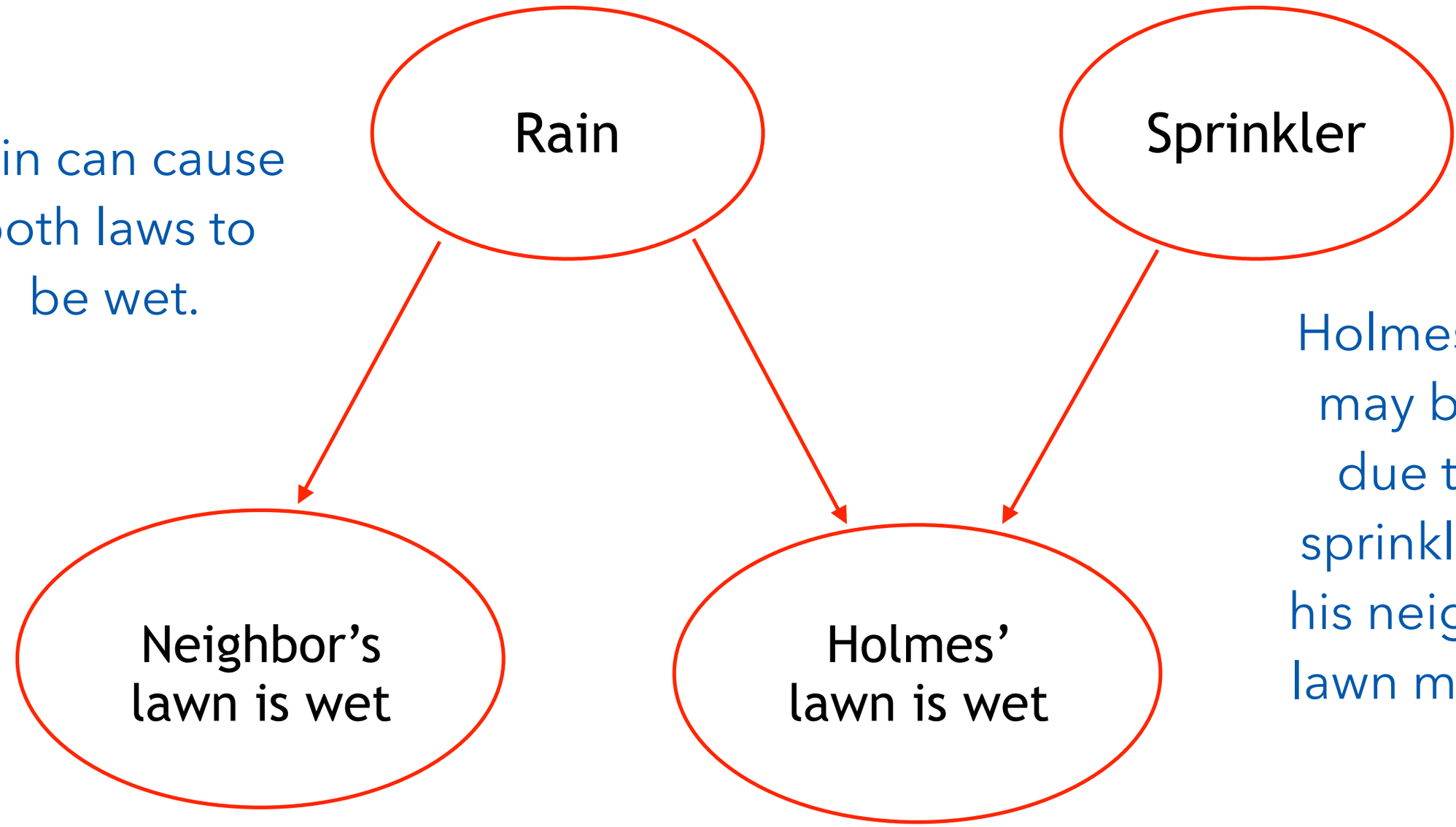
- Mr. Holmes leaves his house:
  - He sees that the lawn in front of his house is wet.
  - This can have several reasons: Either it rained, or Holmes forgot to shut the sprinkler off.
  - Without any further information, the probability of both events (rain, sprinkler) increases (knowing that the lawn is wet).
- Now Holmes looks at his neighbor's lawn:
  - The neighbor's lawn is also wet.
  - This information increases the probability that it rained. And it lowers the probability for the sprinkler.

Example 2: Wet Lawn

TECHNISCHE UNIVERSITÄT DARMSTADT

- Directed graphical model / Bayesian network:

Rain can cause both laws to be wet.



Holmes' lawn may be wet due to his sprinkler, but his neighbor's lawn may not.

Example 3: Earthquake or Burglary?

TECHNISCHE UNIVERSITÄT DARMSTADT

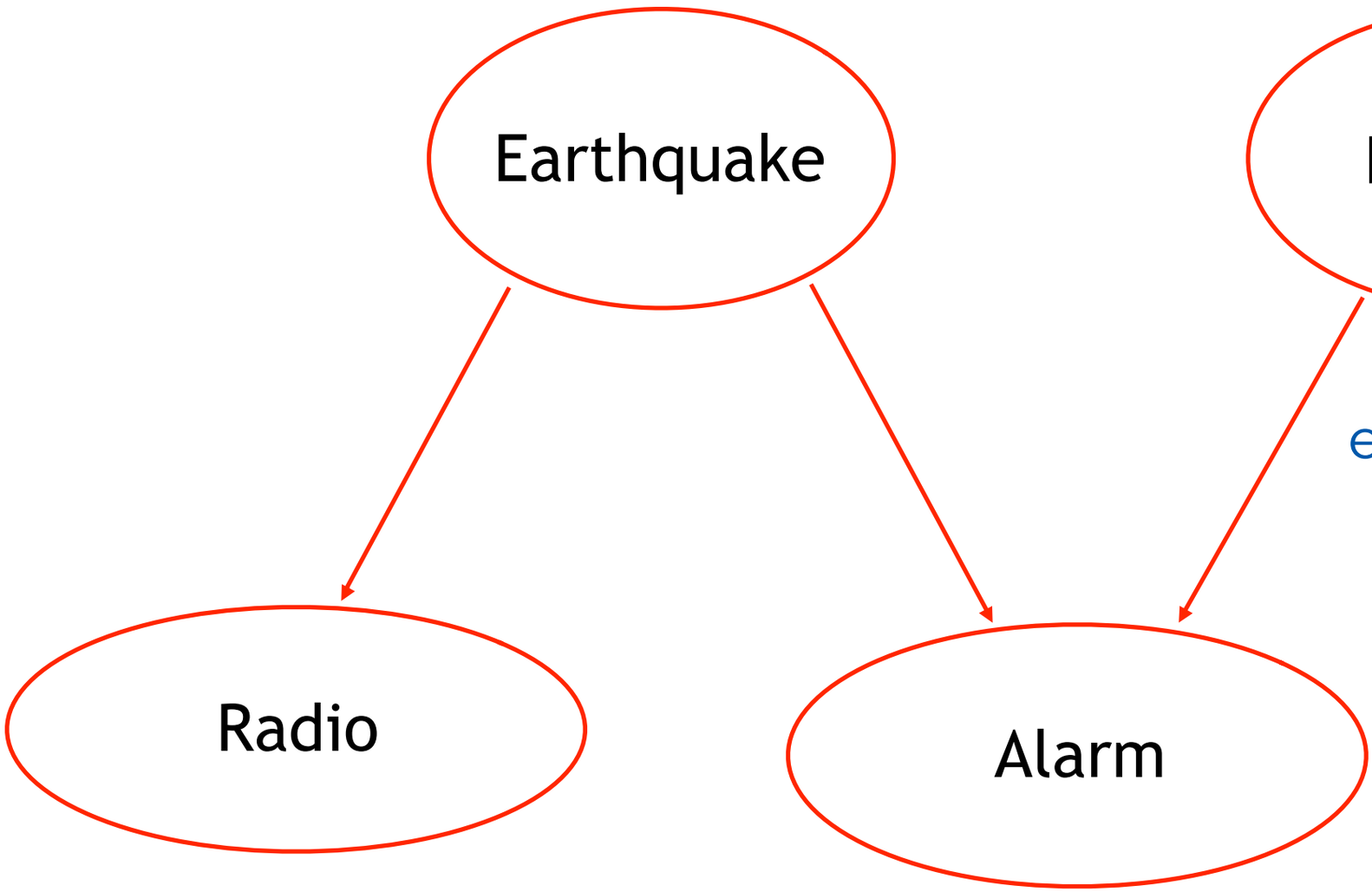
- Mr. Holmes is in his office.
  - He receives a call from a neighbor saying that the alarm in his house went off.
  - He concludes that there must have been a burglary at his home.

- Shortly after, he hears on the radio that there was a light earthquake:
  - Since his alarm has frequently gone off when there was an earthquake, he no longer assumes that there was a burglary.

Example 3: Earthquake or Burglary?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Directed graphical model / Bayesian network:



If we know that an earthquake was reported on the radio, this reduces our belief that a burglary must have happened.

Difference of Examples 2 and 3

TECHNISCHE UNIVERSITÄT DARMSTADT

- The structures of the directed graphical models in both examples are the same:
  - Qualitatively, the interpretation is the same.
  - But the dependencies of the various events can be more or less strong, and thus change the quantitative interpretation.
    - Different quantitative interpretations can also arise from different a-priori probabilities of the events.
- (Example) Assumption:
  - The prior probability of rain and sprinkler are similar.
  - But the prior probability of a burglary is much higher than that of an earthquake.

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- or Bayesian networks
  - Are based on a directed graph.
  - The nodes correspond to the random variables.
  - The directed edges correspond to the (causal) dependencies among the variables.
    - The notion of a causal nature of the dependencies is somewhat hard to grasp.
    - We will typically ignore the notion of causality here.
  - The structure of the network qualitatively describes the dependencies of the random variables.

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Nodes or random variables:
  - We usually know the range of the random variables.
  - The value of a variable may be known or unknown.
  - If they are known, we usually shade the node:

- Examples of variable nodes:
  - Binary events: rain (yes / no), sprinkler (yes / no)
  - Discrete variables: Ball is red, green, blue, ...
  - Continuous variables: Age of a person, ...

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most often, we are interested in
  - quantitative statements
  - i.e. the probabilities (or densities) of the variables.
  - Example: What is the probability of an earthquake or a burglary if the alarm went off, ...
  - These probabilities change if we have more knowledge, less knowledge, or different knowledge about the other variables in the network.

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Simplest case:

  - This model encodes:
    The value of b depends on the value of a.

  - This dependency is expressed through the conditional probability:
    $$p(b|a)$$
  - Knowledge about a is expressed through the prior probability:
    $$p(a)$$
  - The whole graphical model describes the joint probability of a and b:
    $$p(a,b) = p(b|a)p(a)$$

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- If we have such a representation, we can derive all other interesting probabilities from the joint:
  - E.g. marginalization:

$$p(a) = \sum_b p(a,b) = \sum_b p(b|a)p(a)$$

$$p(b) = \sum_a p(a,b) = \sum_a p(b|a)p(a)$$

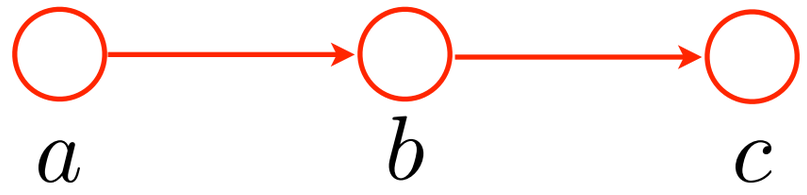
- With the marginals, we can also compute conditional probabilities:

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Chains of nodes:



  - As before, we can compute

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

  - But we can also compute the joint distribution of all three variables:

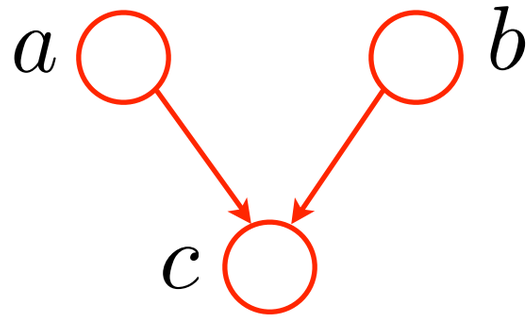
$$\begin{aligned} p(a, b, c) &= p(c|a, b)p(a, b) \\ &= p(c|b)p(b|a)p(a) \end{aligned}$$

  - We can read off from the graphical representation that variable c does not depend on a, if b is known. How?

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Convergent connections:



  - Here the value of c depends on both variables a and b. This is modeled with the conditional probability:

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

- Hence, the joint probability of all three variables is given as:

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

Example: Wet lawn, revisited

TECHNISCHE UNIVERSITÄT DARMSTADT

$p(C)$

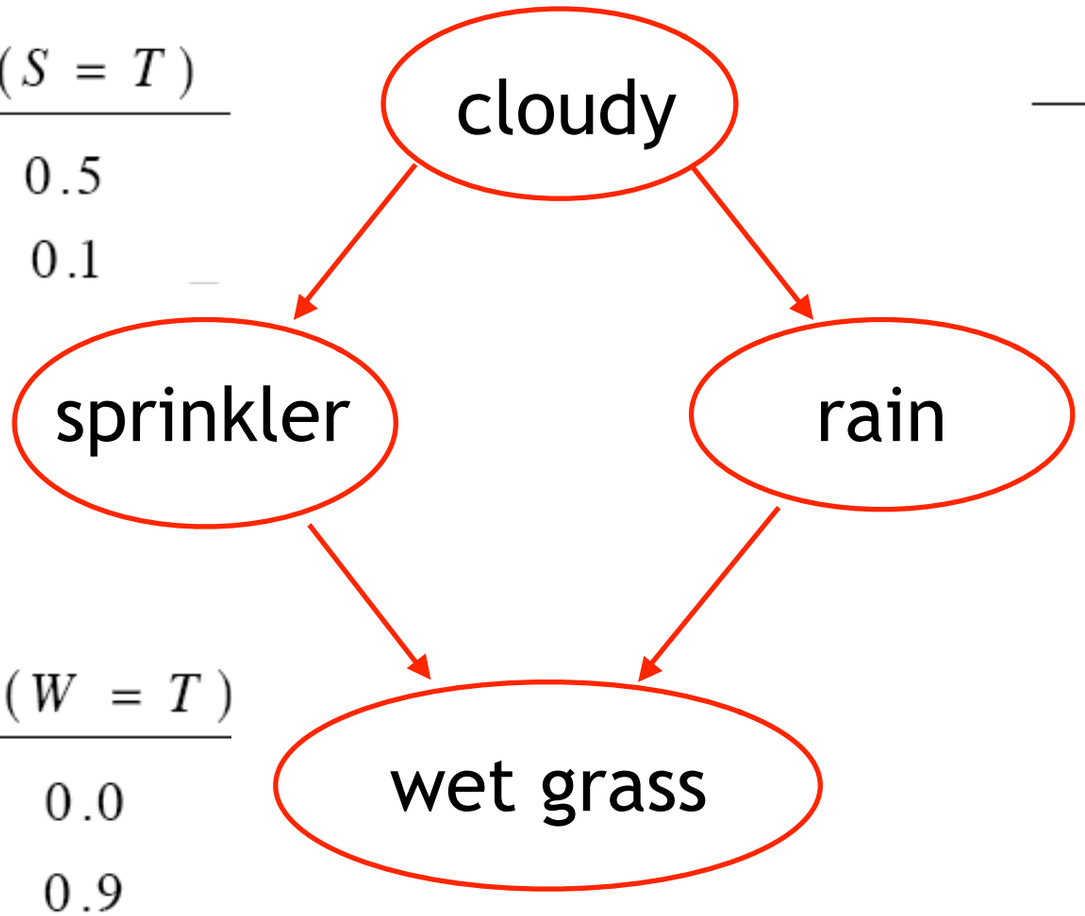
$p(C = F)$	$p(C = T)$
0.5	0.5

$p(S|C)$

C	$p(S = F)$	$p(S = T)$
F	0.5	0.5
T	0.9	0.1

$p(R|C)$

C	$p(R = F)$	$p(R = T)$
F	0.8	0.2
T	0.2	0.8



$p(W|R, S)$

SR	$p(W = F)$	$p(W = T)$
FF	1.0	0.0
TF	0.1	0.9
FT	0.1	0.9
TT	0.01	0.99

Example: Wet lawn, revisited

TECHNISCHE UNIVERSITÄT DARMSTADT

- We start with the simple product rule:

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

- This means that we can rewrite the joint probability of the 4 variables as:

$$p(C, S, R, W) = p(C)p(S|C)p(R|C, S)p(W|C, S, R)$$

- But the Bayesian network says that

$$p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)$$

  - i.e. rain is independent of sprinkler (given the cloudiness).
  - Wet grass is independent of the cloudiness (given the state of the sprinkler and the rain).
  - This is a factorized representation of the joint probability.

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- A general directed graphical model / Bayesian network consists of
  - A set of variables: $V = \{x_1, \ldots, x_n\}$
  - A set of directed edges between the variable nodes.

- The variables and the directed edges give an acyclic graph:
  - Acyclic means that there is no directed cycle in the graph.

- For each variable x_i with parent nodes $\mathrm{Parents}(i)$ in the graph, we require knowledge of a conditional probability:

$$p(x_i | \{x_j | j \in \mathrm{Parents}(i)\})$$

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Given
  - Variables: $V = \{x_1, \ldots, x_n\}$
  - Directed acyclic graph: $G = (V, E)$
    - V: nodes = variables, E: directed edges

- We can express / compute the joint probability as:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p\left(x_i \,\middle|\, \{x_j | j \in \text{Parents}(i)\}\right)$$

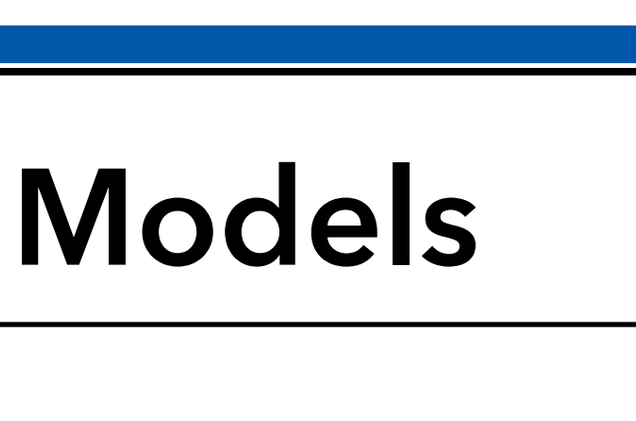
  - We can express the joint as a product of all the conditional distributions from the parent-child relations in the graph.
  - We obtain a factorized representation of the joint.

Directed Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p\Big(x_i \,\Big|\, \{x_j | j \in \text{Parents}(i)\}\Big)$$

- Example:



$$p(x_0, x_1, x_2, x_3) = p(x_0)p(x_1|x_0)p(x_2)p(x_3|x_1, x_2)$$

Factorized Representation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Reduction of complexity:
  - Joint probability of n binary variables requires us to represent values brute force:

$$\mathcal{O}(2^n) \quad \text{terms}$$

- The factorized form obtained from the graphical model only requires

$$\mathcal{O}(n \cdot 2^k) \quad \text{terms}$$

  - k: maximum number of parents of a node.

Conditional Independence

TECHNISCHE UNIVERSITÄT DARMSTADT

- Suppose with have a joint density with 4 variables:

$$p(x_0, x_1, x_2, x_3)$$

  - For example, 4 subsequent words in a sentence:

$$x_0 = \text{Machine}, \quad x_1 = \text{learning}, \quad x_2 = \text{is}, \quad x_3 = \text{fun}$$

- The product rule tells us that we can rewrite the joint density:

$$\begin{aligned} p(x_0, x_1, x_2, x_3) &= p(x_3|x_0, x_1, x_2)p(x_0, x_1, x_2) \\ &= p(x_3|x_0, x_1, x_2)p(x_2|x_0, x_1)p(x_0, x_1) \\ &= p(x_3|x_0, x_1, x_2)p(x_2|x_0, x_1)p(x_1|x_0)p(x_0) \end{aligned}$$

Conditional Independence

$$p(x_0, x_1, x_2, x_3) = p(x_3|x_0, x_1, x_2)p(x_2|x_0, x_1)p(x_1|x_0)p(x_0)$$

- Now we can make a simplifying assumption:
  - Only the previous word is what matters, that is given the previous word we can forget about every word *before* the previous one.
  - E.g.: $p(x_3|x_0, x_1, x_2) = p(x_3|x_2) \quad \text{or} \quad p(x_2|x_0, x_1) = p(x_2|x_1)$
  - That seems reasonable, for example, because the probability that "fun" follows "is" doesn't change (that much) whether the first word is "machine" or not.
  - Such assumptions are called conditional independence assumptions.

Conditional Independence

TECHNISCHE UNIVERSITÄT DARMSTADT

- The notion of conditional independence means that:
  - Given a certain variables, other variables become independent.
  - More concretely here:

$$p(x_3|x_0, x_1, x_2) = p(x_3|x_2)$$

    - This means that x_3 is conditionally independent from x_0 and x_1 given x_2.

$$p(x_2|x_0, x_1) = p(x_2|x_1)$$

    - This means that x_2 is conditionally independent from x_0 given x_1.
  - Why is this?

$$p(x_0, x_2|x_1) = p(x_2|x_0, x_1)p(x_0|x_1)$$

$$= p(x_2|x_1)p(x_0|x_1)$$

independent given x_1

Conditional Independence

TECHNISCHE UNIVERSITÄT DARMSTADT

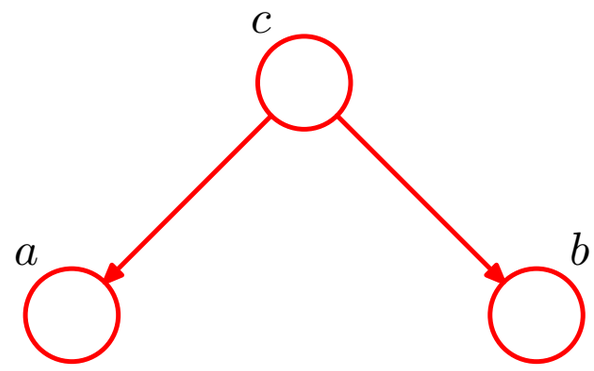
- Directed graphical models are not only useful...
  - because the joint factorized into a product of simpler conditional distributions.
  - but also, because we can read off conditional independence of variables.

- Let us study this.

First Case: Divergent

TECHNISCHE UNIVERSITÄT DARMSTADT

- Divergent model:



  - Are a and b independent?

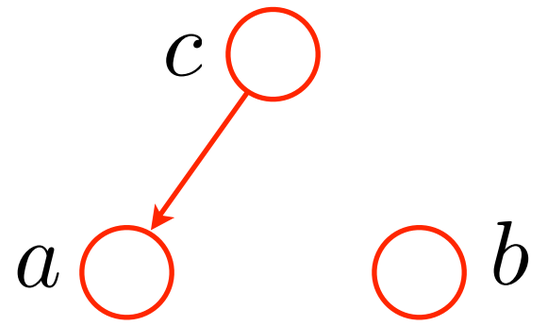
  - Marginalize out c:

$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a|c)p(b|c)p(c)$$

  - In general, this is not equal to $p(a)p(b)$, hence the variables are not independent.

First Case: Divergent

- What about now?



  - Are a and b independent?
    - Marginalize out c:

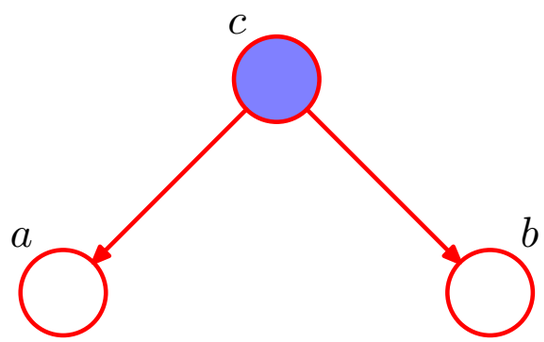
$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a|c)p(b)p(c) = p(a)p(b)$$

- If there is no undirected connection between two variables, then they are independent.

First Case: Divergent

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let us return to the original graph, but now we assume that we observe the value of c:



- The conditional probability is given as:

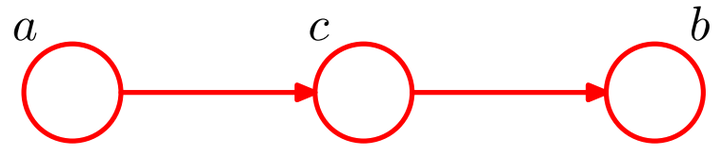
$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$$

- If c becomes known, the variables a and b become conditionally independent.

Second Case: Chain

TECHNISCHE UNIVERSITÄT DARMSTADT

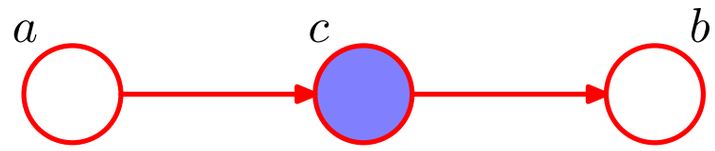
- Let us consider a slightly different graphical model:



chain model

  - Are a and b independent? No.

$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(b|c)p(c|a)p(a) = p(b|a)p(a)$$



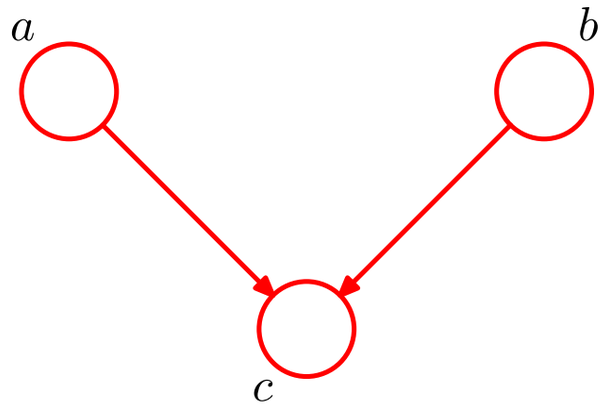
  - If c becomes known, are a and b conditionally independent? Yes.

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

Third Case: Convergent

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let us look at a final case: Convergent graph



- Are a and b independent? **YES!**

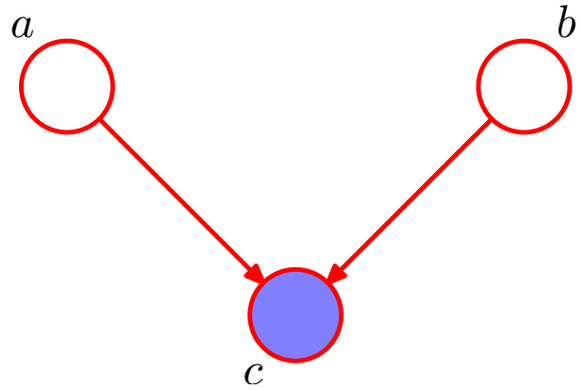
$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(c|a,b)p(a)p(b) = p(a)p(b)$$

- This is very different from the previous cases.
- Even though a and b are connected, they are independent.

Third Case: Convergent

TECHNISCHE UNIVERSITÄT DARMSTADT

- Now we assume that c is observed:



- Are a and b conditionally independent? **NO!**

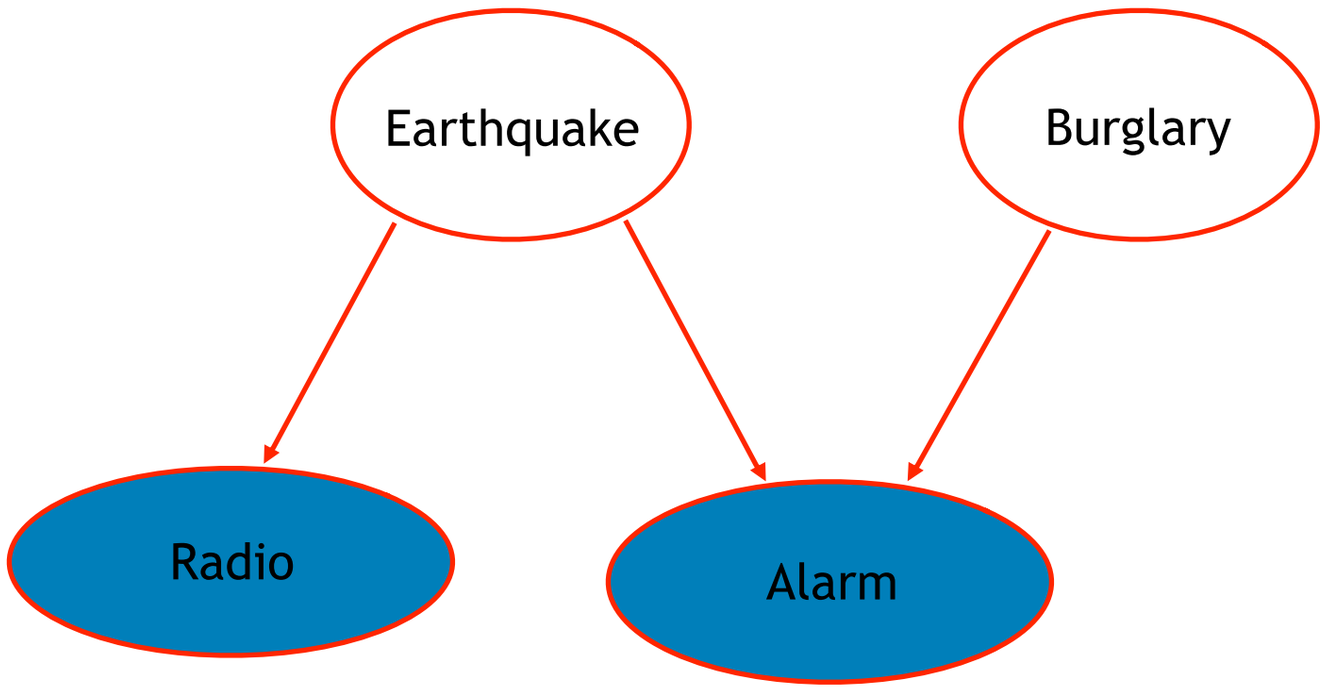
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

- In general, they are not conditionally independent.
- This case is the opposite of the previous cases!

Explaining Away

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let us look at this previous example again:



- Both earthquakes and burglaries increase the probability for an alarm and a radio alert.
- But knowing that there is an earthquake decreases the probability for a burglary.
- The burglary is **explained away**.

Undirected Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Directed graphical models are quite versatile, but they are not always appropriate:
  - It may not always be convenient to provide conditional distributions.
  - There are certain conditional independence structures that a directed graph cannot represent.

- There is a second class of graphical models called **undirected graphical models** or Markov random fields.

Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Given:
  - The nodes again correspond to the random variables:

$$V = \{x_1, \ldots, x_n\}$$

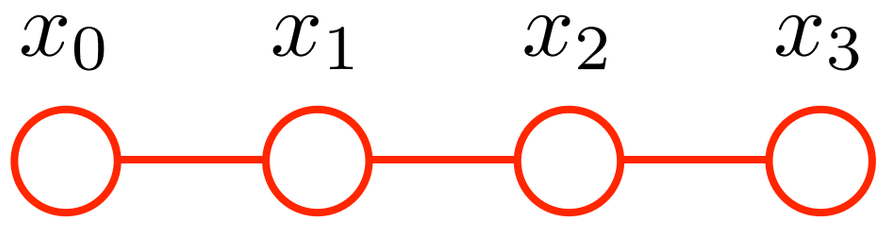
  - The edges of the graph $G = (V, E)$ are now undirected.
  - The graph may have cycles.

- The edges tell us how the joint distribution can be written as a product of simpler factors.
- But the factors are arbitrary non-negative functions - so-called potential functions (and not conditional distributions).

Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- First example:
  - Chain graph:



  - The joint probability is a product of pairwise factors in this case:

$$p(x_0, x_1, x_2, x_3) = \frac{1}{Z} f(x_0, x_1) f(x_1, x_2) f(x_2, x_3)$$

  - The term $\frac{1}{Z}$ is a normalization factor, that ensures that the probability (density) sums (integrates) to 1.
    - This is necessary, because the factors are in general not related to any (conditional) probability distribution.

Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- More generally:

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{c \in \mathcal{C}} f_c\Big(\{x_j | j \in c\}\Big)$$

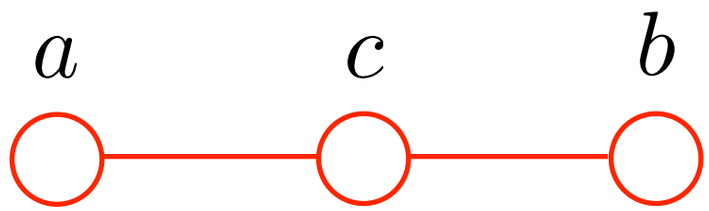
cliques of the graph

  - Note: The **factors** f_c are non-negative functions, but unnormalized.
    They are not probabilities.

Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Look at a simple example:



  - Are a and b independent?

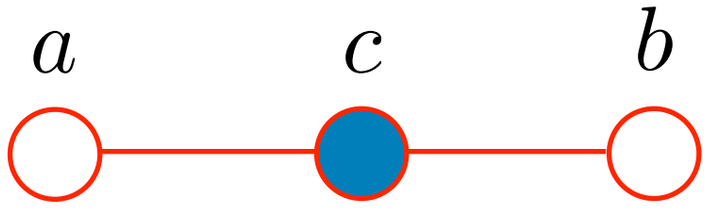
$$p(a,b) = \sum_c p(a,b,c) = \frac{1}{Z} \sum_c f_1(a,c) f_2(b,c)$$

  - In general, this does not simplify to $p(a)p(b)$, hence no independence.

Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- What if c is observed?



  - Are a and b conditionally independent given c?

$$p(a,b|c) = \frac{1}{p(c)}\frac{1}{Z} f_1(a,c) f_2(b,c) = \frac{1}{Z'} \hat{f}_a(a,c) \cdot \frac{1}{Z''} \hat{f}_b(b,c)$$

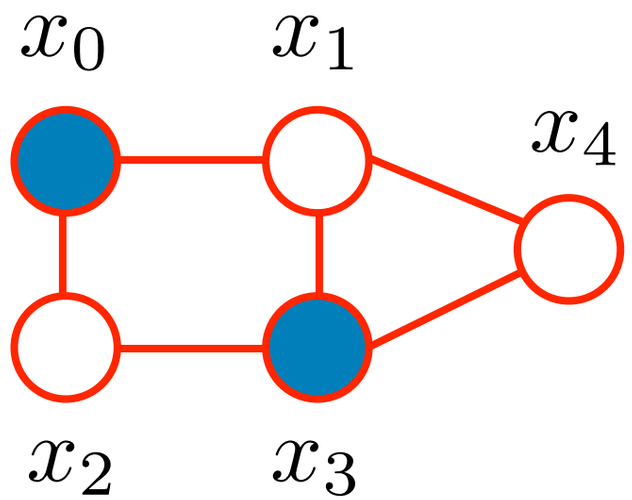
  - The factors are proportional to $p(a|c)$ resp. $p(b|c)$.
  - Hence conditional independence of the variables.

Conditional Independence

TECHNISCHE UNIVERSITÄT DARMSTADT

- More generally
  - In particular, two variables, say a and b, are conditionally independent given a set other variables S, if you cannot reach a from b in the graph without passing through S.
  - The conditional independence statements are somewhat easier to read off than in directed models.

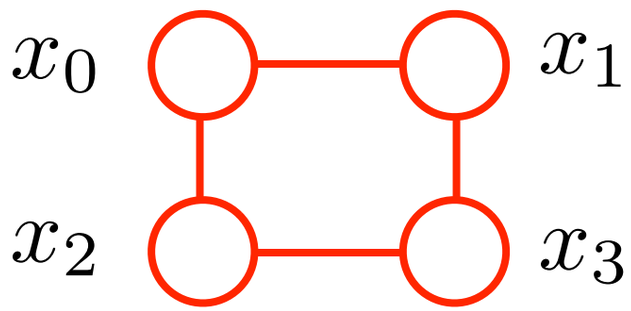
- Example:
  - x_2 and x_4 are conditionally independent given x_0 and x_3.



Undirected Graphical Models

TECHNISCHE UNIVERSITÄT DARMSTADT

- Loopy graph:



$$p(x_0, x_1, x_2, x_3) = \frac{1}{Z} f_0(x_0, x_1) f_1(x_0, x_2) f_2(x_2, x_3) f_3(x_1, x_3)$$

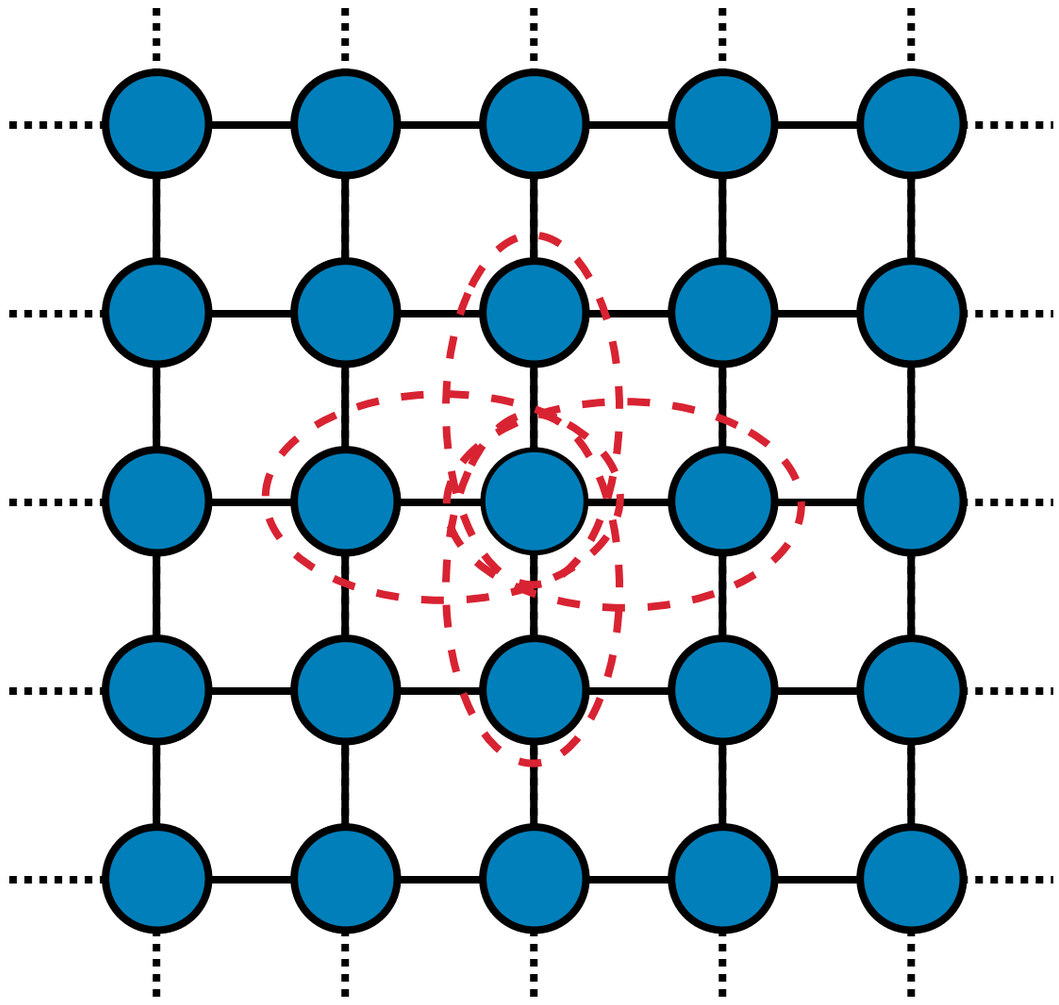
- In this case, it is not possible to express the same conditional independence statements with a directed model.

- Are any variables independent?
- If two variables are reachable in the graph, they are dependent.

Markov Random Fields

TECHNISCHE UNIVERSITÄT DARMSTADT

- Back to stereo...
- In this simple graph all the cliques are pairs of pixels.
  - Pairwise MRF.
- Now we can fully understand the factor structure of the joint density:

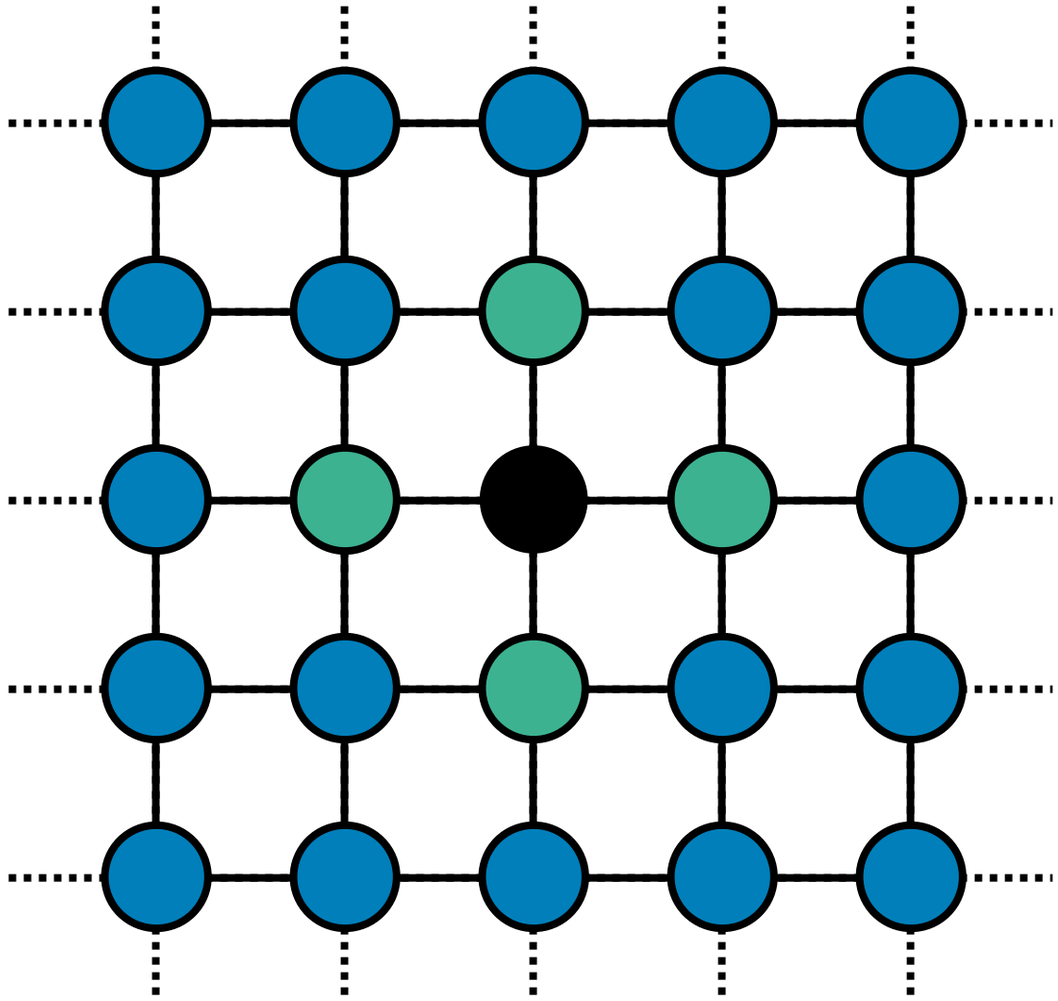


$$p(\mathbf{d}) = \frac{1}{Z} \prod_{i,j} f_H(d_{i,j}, d_{i+1,j}) \cdot f_V(d_{i,j}, d_{i,j+1})$$

Markov Random Fields

TECHNISCHE UNIVERSITÄT DARMSTADT

- Back to stereo...
- Markov property:
  - Given its 4 neighboring pixels, a pixel is independent of all other pixels!



- This is a pretty reasonable assumption, but it is oversimplifying the problem somewhat.
  - Circumventing this is a current research problem!

Today

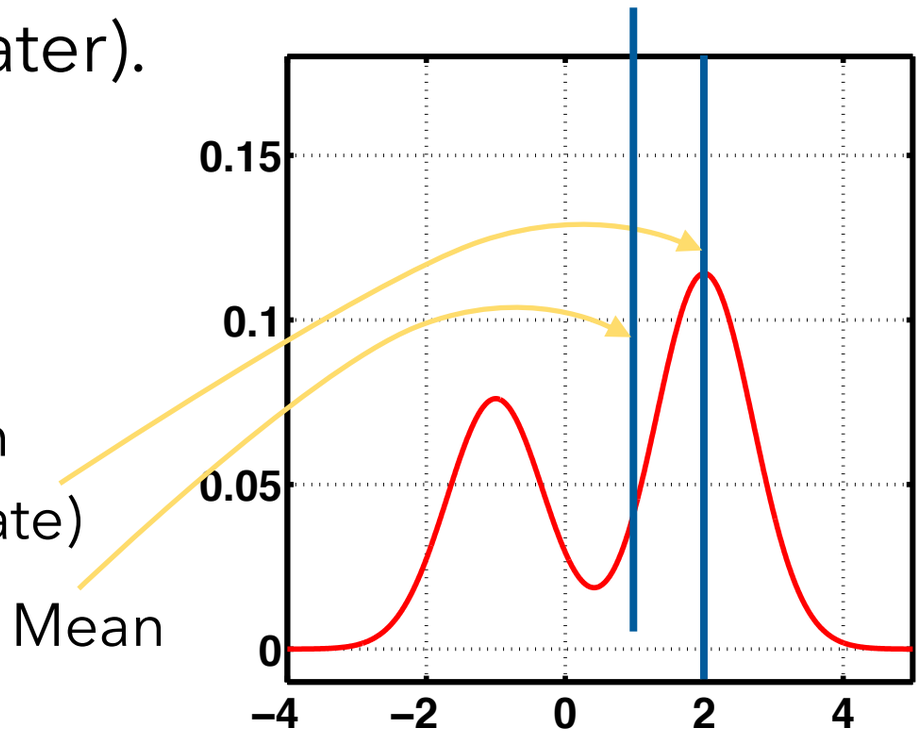
TECHNISCHE UNIVERSITÄT DARMSTADT

- Once we have given such a probabilistic model of a vision problem, such as stereo, how do we figure out (i.e. infer) the solution?
- We do that by means of **probabilistic inference**.
- This is a very general set of tools that is:
  - Well understood.
  - Used very widely, not only in computer vision, but in robotics, computational biology, natural language processing, etc.
- Unfortunately, it is not all that easy to understand and is a key topic of machine learning classes.
- Hence we will only cover the very basics.

Probabilistic Inference

TECHNISCHE UNIVERSITÄT DARMSTADT

- ... generally means one of three things:
  - Computing the maximum of the posterior distribution $p(x|y)$, that is computing the state that is the most probable given our observations (maximum a-posteriori (MAP) estimation).
  - Computing expectations over the posterior distribution, such as the mean of the posterior $p(x|y)$.
  - Computing marginal distributions (later).
- How do these differ?
  - Assume we have the following posterior distribution:
  - The posterior may be multi-modal!



Probabilistic Inference

TECHNISCHE UNIVERSITÄT DARMSTADT

- We will talk about MAP estimation first:
  - First "attempt": Continuous optimization methods.
  - Second "attempt": Graph-based methods (graph cuts).
- Later we will briefly get into computing expectations and marginal distributions:
  - Belief propagation or the sum-product algorithm.
    - This can be extended to MAP estimation: The max-product algorithm.

Continuous Optimization

TECHNISCHE UNIVERSITÄT DARMSTADT

- The most straightforward idea for maximizing the posterior is to apply well-known continuous optimization techniques.
- Especially gradient techniques have found widespread use, e.g.:
  - Simple **gradient ascent**, also called hill-climbing.
  - Conjugate gradient methods.
  - And many more.
- Since the posterior may be multi-modal (more on that later), this will typically only give us a **local optimum** and not the global optimum.

Gradient Ascent

TECHNISCHE UNIVERSITÄT DARMSTADT

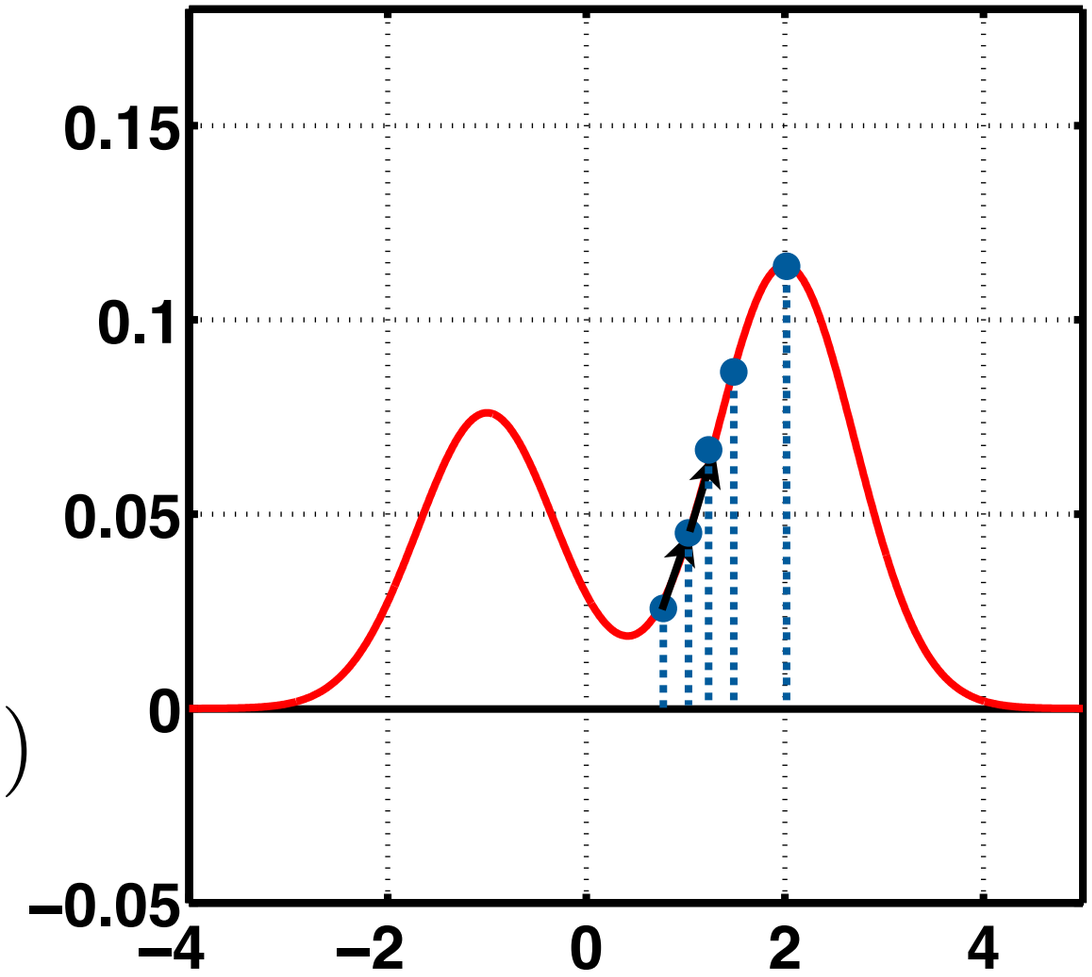
- Iteratively maximize a function $f(x)$:
  - Initialize somewhere: $x^{(0)}$
  - Compute the derivative: $\frac{\mathrm{d}}{\mathrm{d}x} f(x) = f'(x)$
  - Take a step in the direction of the derivative:

$$x^{(1)} \leftarrow x^{(0)} + \eta \cdot f'(x^{(0)})$$

step size

  - Repeat...

$$x^{(n+1)} \leftarrow x^{(n)} + \eta \cdot f'(x^{(n)})$$



Gradient Ascent

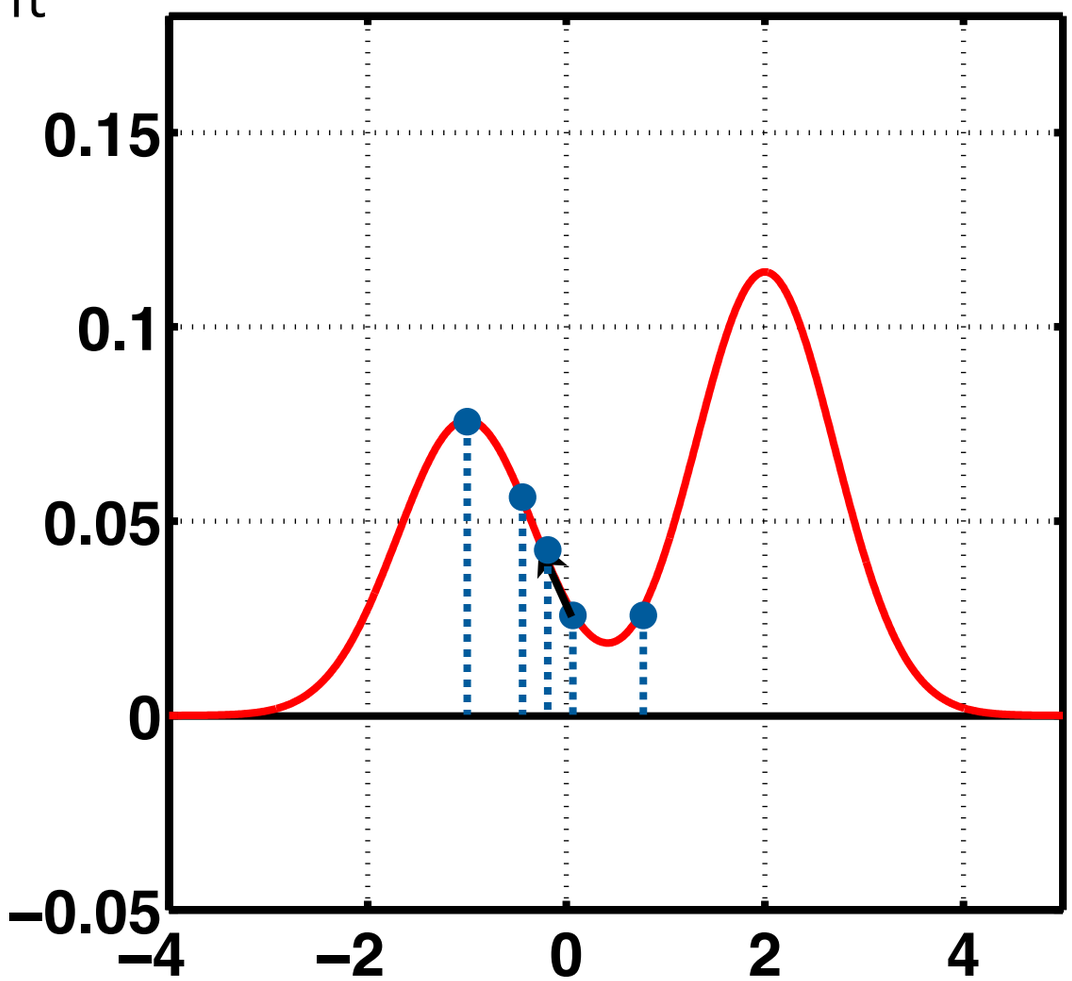
TECHNISCHE UNIVERSITÄT DARMSTADT

- We can do the same in multiple dimensions:

$$\mathbf{x}^{(n+1)} \leftarrow \mathbf{x}^{(n)} + \eta \cdot \nabla f(\mathbf{x}^{(n)})$$

gradient

- Problems:
  - What if we initialize wrong?
    - We end up in the wrong local optimum!
  - How do we choose the step size η?
    - The wrong one can lead to instabilities or slow convergence.



Stereo with Continuous Optimization

TECHNISCHE UNIVERSITÄT DARMSTADT

- Concrete example:
  How can we do stereo using these techniques?
- We usually work with the log of the posterior:
  - Numerically much more stable.
  - What is the gradient of the log-posterior?
    - Let us forget about the likelihood for now, only look at the log-prior:

$$\log p(\mathbf{d}) = \log \left[\frac{1}{Z} \prod_{i,j} f_H(d_{i,j}, d_{i+1,j}) \cdot f_V(d_{i,j}, d_{i,j+1}) \right]$$

$$= \sum_{i,j} \log f_H(d_{i,j}, d_{i+1,j}) + \log f_V(d_{i,j}, d_{i,j+1}) + \text{const}$$

Gradient of the Log-Prior

TECHNISCHE UNIVERSITÄT DARMSTADT

- Calculate the partial derivative w.r.t. a particular pixel $d_{k,l}$

$$\begin{aligned}\frac{\partial}{\partial d_{k,l}} \log p(\mathbf{d}) &= \frac{\partial}{\partial d_{k,l}} \sum_{i,j} \log f_H(d_{i,j}, d_{i+1,j}) + \log f_V(d_{i,j}, d_{i,j+1}) + \text{const} \\ &= \sum_{i,j} \frac{\partial}{\partial d_{k,l}} \log f_H(d_{i,j}, d_{i+1,j}) + \frac{\partial}{\partial d_{k,l}} \log f_V(d_{i,j}, d_{i,j+1})\end{aligned}$$

- Only the 4 terms from the 4 neighbors remain:

$$\begin{aligned}\frac{\partial}{\partial d_{k,l}} \log p(\mathbf{d}) &= \frac{\partial}{\partial d_{k,l}} \log f_H(d_{k,l}, d_{k+1,l}) + \frac{\partial}{\partial d_{k,l}} \log f_H(d_{k-1,l}, d_{k,l}) + \\ &+ \frac{\partial}{\partial d_{k,l}} \log f_V(d_{k,l}, d_{k,l+1}) + \frac{\partial}{\partial d_{k,l}} \log f_V(d_{k,l-1}, d_{k,l})\end{aligned}$$

Gradient of the Log-Prior

TECHNISCHE UNIVERSITÄT DARMSTADT

- Almost there... Simply apply the chain rule:

$$\frac{\partial}{\partial d_{k,l}} \log f_H(d_{k,l}, d_{k+1,l}) = \frac{\frac{\partial}{\partial d_{k,l}} f_H(d_{k,l}, d_{k+1,l})}{f_H(d_{k,l}, d_{k+1,l})}$$

- But what is the derivative of the compatibility function (or potential function)?

$$\frac{\partial}{\partial d_{k,l}} f_H(d_{k,l}, d_{k+1,l})$$

- Problem: The Potts model is not differentiable!

$$f_H(d_{k,l}, d_{k+1,l}) = \frac{1}{Z(T)} \exp\left\{ \frac{1}{T} \delta(d_{k,l}, d_{k+1,l}) \right\}$$

Stereo with Continuous Optimization

TECHNISCHE UNIVERSITÄT DARMSTADT

- But that is not a big problem, because we said that the Potts model is not very good anyway.
  - We could use some differentiable compatibility (potential) function
    - e.g. a robust function such as the Lorentzian / Student-t
  - Once we can differentiate that, we're done with the prior part.

- What about the log-likelihood?

$$\begin{aligned}
\log p(\mathbf{I}^0, \mathbf{I}^1 | \mathbf{d}) &= \log \prod_{i,j} \mathcal{N}(I^0_{i,j} - I^1_{(i-d_{ij}),j};\ 0, \sigma^2) \\
&= \sum_{i,j} \log \mathcal{N}(I^0_{i,j} - I^1_{(i-d_{ij}),j};\ 0, \sigma^2) \\
&= -\sum_{i,j} \frac{1}{2\sigma^2} (I^0_{i,j} - I^1_{(i-d_{ij}),j})^2 + \text{const}
\end{aligned}$$

Gradient of the Log-Likelihood

- To take the derivative, let us rewrite the log-likelihood assuming that the image is a continuous function:

$$\begin{aligned}\log p(\mathbf{I}^0, \mathbf{I}^1 | \mathbf{d}) &= -\sum_{i,j} \frac{1}{2\sigma^2} (I^0_{i,j} - I^1_{(i-d_{ij}),j})^2 + \text{const} \\ &= -\sum \frac{1}{2\sigma^2} \left(I^0(i,j) - I^1(i - d_{i,j}, j)\right)^2 + \text{const}\end{aligned}$$

- Now we can compute the partial as follows:

$$\frac{\partial}{\partial d_{k,l}} \log p(\mathbf{I}^0, \mathbf{I}^1 | \mathbf{d}) = \frac{1}{\sigma^2} \left(I^0(k,l) - I^1(k - d_{k,l}, l)\right) \cdot \frac{\partial}{\partial d_{k,l}} I^1(k - d_{k,l}, l)$$

Horizontal image derivative at $k - d_{k,l}$

Stereo with Continuous Optimization

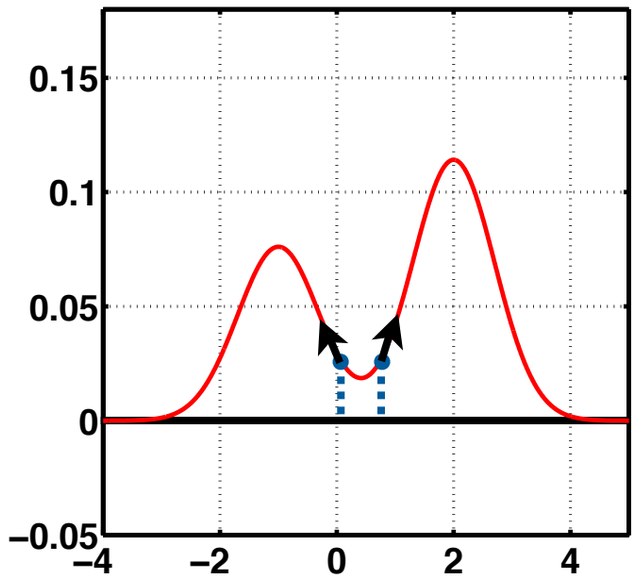
TECHNISCHE UNIVERSITÄT DARMSTADT

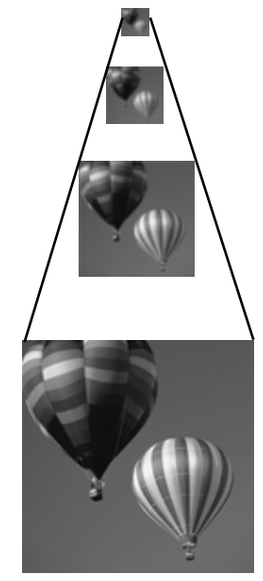
- In reality, our image is not a continuous function, requiring us to approximate the horizontal derivative.
  - A typical approach is to use bicubic interpolation.
  - Yields interpolated disparities and derivatives.
- Now we can put the derivative expressions of log-prior and log-likelihood together, and we can do MAP estimation for stereo.
- But will this actually work well?
  - No, this does not work particularly well. Why?
  - The log-posterior has many **local optima**, because the data term can produce many incorrect matches that are still locally optimal.

Avoiding the Wrong Local Optima

TECHNISCHE UNIVERSITÄT DARMSTADT

- We have seen that initialization matters with gradient techniques.
- We could initialize in a smarter way:
  - For example with the output of a different stereo algorithm, such as window-based correlation.
  - Even better, we can estimate disparity in a coarse-to-fine way:
    - Scale the input images down and estimate disparity there.
    - Use this estimate to initialize at a finer resolution.
    - Use a Gaussian pyramid to do this.
    - More when we talk about optical flow.





Why did we go through all this?

- We just went through a lot of math only to find out that gradient methods may not be such a great idea for stereo.
- Why did we do this here? Several good reasons:
  - Gradient methods are widely used in vision.
  - We will see them over and over again in image processing, optical flow, and tracking.
  - You understand some of the challenges that we face.