

Undirected Graphical Models

Dr. Shuang LIANG

School of Software Engineering
Tongji University
Fall, 2012

Today's Topics

- Introduction
- Parameterization
- Gibbs Distributions
- Reduced Markov Networks
- Markov Network Independencies
- Learning Undirected Models

Today's Topics

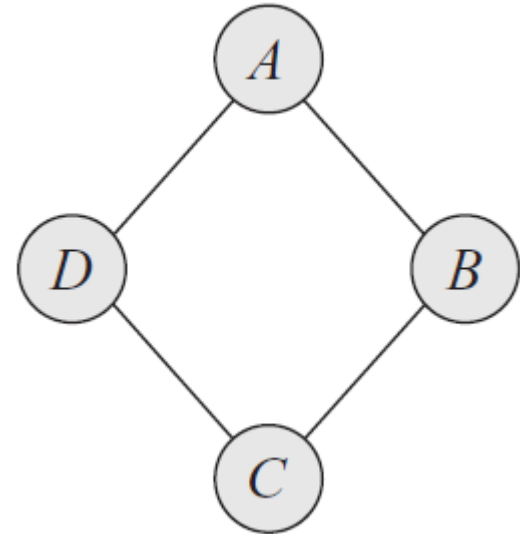
- *Introduction*
- Parameterization
- Gibbs Distributions
- Reduced Markov Networks
- Markov Network Independencies
- Learning Undirected Models

Introduction

- We looked at directed graphical models whose structure and parameterization provide a natural representation for many real-world problems.
- Undirected graphical models are useful where one cannot naturally ascribe a directionality to the interaction between the variables.

Introduction

- An example model that satisfies
 - $(A \perp C | \{B, D\})$
 - $(B \perp D | \{A, C\})$
 - No other independencies
- These independencies cannot be naturally captured in a Bayesian network.



An example undirected graphical model

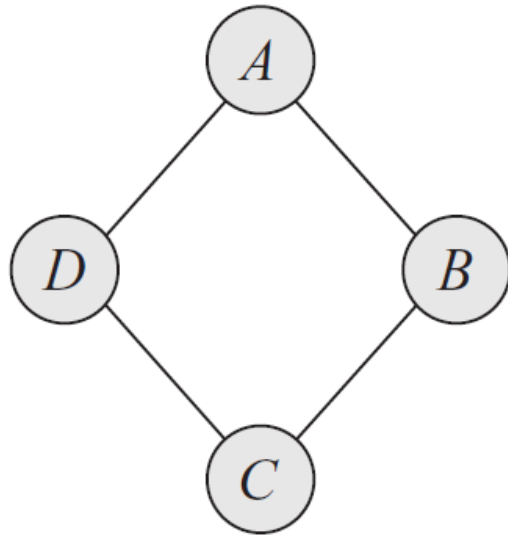
An Example

- Four students are working together in pairs on a homework.
- Alice and Charles cannot stand each other, and Bob and Debbie had a relationship that ended badly.
- Only the following pairs meet: Alice and Bob; Bob and Charles; Charles and Debbie; and Debbie and Alice.
- The professor accidentally misspoke in the class, giving rise to a possible misconception.
- In study pairs, each student transmits her/his understanding of the problem.

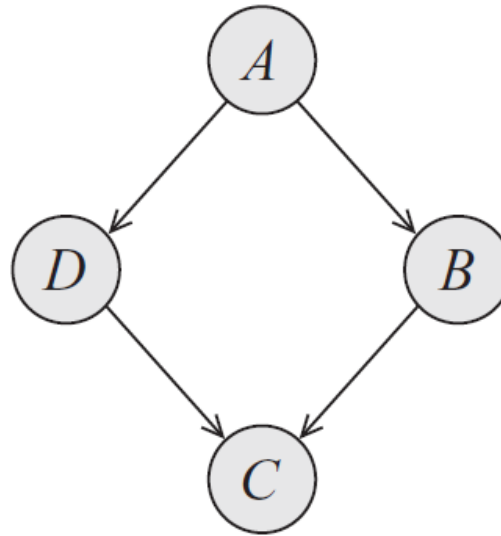
An Example

- Four binary random variables are defined, representing whether the student has a misconception or not.
- Assume that for each $X \in \{A, B, C, D\}$, x^1 denotes the case where the student has the misconception, and x^0 denotes the case where she/he does not.
- Alice and Charles never speak to each other directly, so A and C are conditionally independent given B and D.
- Similarly, B and D are conditionally independent given A and C.

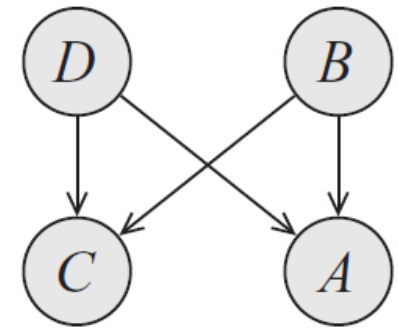
An Example



(a)



(b)



(c)

Example models for the misconception example. (a) An undirected graph modeling study pairs over four students. (b) An unsuccessful attempt to model the problem using a Bayesian network. (c) Another unsuccessful attempt.

Today's Topics

- Introduction
- *Parameterization*
- Gibbs Distributions
- Reduced Markov Networks
- Markov Network Independencies
- Learning Undirected Models

Parameterization

- How to parameterize this undirected graph?
- We want to capture the affinities between related variables.
- Conditional probability distributions cannot be used because they are not symmetric, and the chain rule need not apply.
- Marginals cannot be used because a product of marginals does not define a consistent joint.
- A general purpose function: *factor* (also called *potential*).

Parameterization

- Let D is a set of random variables.
 - A factor Φ is a function from $\text{Val}(D)$ to \mathbb{R} .
 - A factor is nonnegative if all its entries are nonnegative.
 - The set of variables D is called the scope of the factor.
- In the example, an example factor is

$$\phi_1(A, B) : \text{Val}(A, B) \mapsto \mathbb{R}^+.$$

Parameterization

- Factors for the misconception example.

$\phi_1(A, B)$			$\phi_2(B, C)$			$\phi_3(C, D)$			$\phi_4(D, A)$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Parameterization

- The value associated with a particular assignment a , b denotes the affinity between these two variables: the higher the value $\Phi_1(a, b)$, the more compatible these two values are.
- For Φ_1 , if A and B disagree, there is less weight.
- For Φ_3 , if C and D disagree, there is more weight.
- A factor is not normalized, i.e., the entries are not necessarily in $[0, 1]$.

Parameterization

- The Markov network defines the local interactions between directly related variables.
- To define a global model, we need to combine these interactions.
- We combine the local models by multiplying them as

$$P(a, b, c, d) = \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(d, a).$$

Parameterization

- However, there is no guarantee that the result of this process is a normalized joint distribution.
- Thus, it is normalized as

$$P(a, b, c, d) = \frac{1}{Z} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a)$$

where

$$Z = \sum_{a,b,c,d} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a).$$

- Z is known as the partition function.

Parameterization

- Joint distribution for the misconception example

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300,000	0.04
a^0	b^0	c^0	d^1	300,000	0.04
a^0	b^0	c^1	d^0	300,000	0.04
a^0	b^0	c^1	d^1	30	4.110^{-6}
a^0	b^1	c^0	d^0	500	6.910^{-5}
a^0	b^1	c^0	d^1	500	6.910^{-5}
a^0	b^1	c^1	d^0	5,000,000	0.69
a^0	b^1	c^1	d^1	500	6.910^{-5}
a^1	b^0	c^0	d^0	100	1.410^{-5}
a^1	b^0	c^0	d^1	1,000,000	0.14
a^1	b^0	c^1	d^0	100	1.410^{-5}
a^1	b^0	c^1	d^1	100	1.410^{-5}
a^1	b^1	c^0	d^0	10	1.410^{-6}
a^1	b^1	c^0	d^1	100,000	0.014
a^1	b^1	c^1	d^0	100,000	0.014
a^1	b^1	c^1	d^1	100,000	0.014

Parameterization

- There is a tight connection between the factorization of the distribution and its independence properties.
- For example, $P \models (X \perp Y | Z)$ if and only if we can write P in the form $P(X) = \Phi_1(X, Z) \Phi_2(Y, Z)$.
- From the example

$$P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A),$$

we can infer that

$$P \models A \perp C | \{B, D\},$$

$$P \models B \perp D | \{A, C\}.$$

Parameterization

- Factors do not correspond to either probabilities or to conditional probabilities.
- It is harder to estimate them from data.
- One idea for parameterization could be to associate parameters directly with the edges in the graph.
 - This is not sufficient to parameterize a full distribution.
- A more general representation can be obtained by allowing factors over arbitrary subsets of variables.

Parameterization

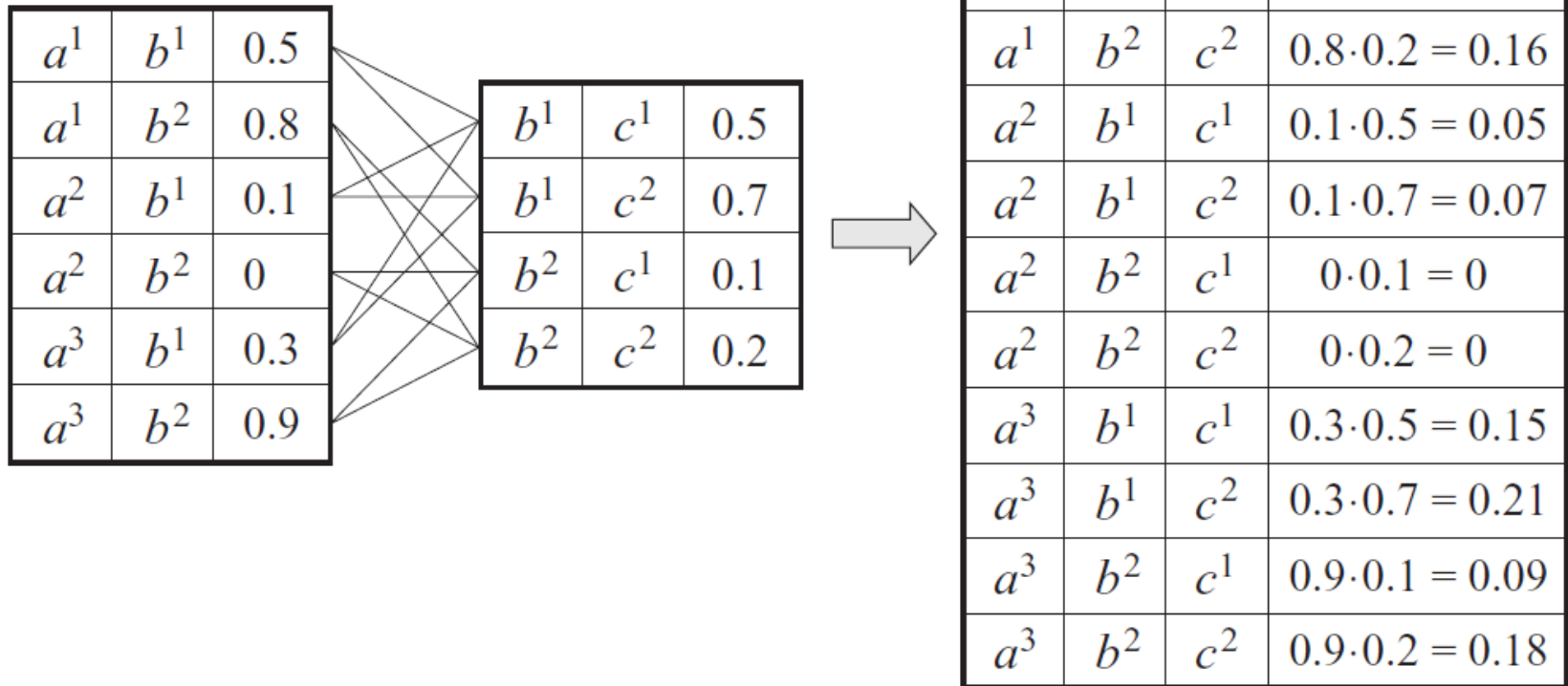
- Let X , Y , and Z be three disjoint sets of variables, and let $\Phi_1(X,Y)$ and $\Phi_2(Y,Z)$ be two factors.
- We define the factor product $\phi_1 \times \phi_2$ to be a factor $\psi : \text{Val}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \mapsto \mathbb{R}$ as follows:

$$\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \phi_1(\mathbf{X}, \mathbf{Y})\phi_2(\mathbf{Y}, \mathbf{Z}).$$

- The key aspect is the fact that the two factors Φ_1 and Φ_2 are multiplied in way that matches up the common part Y .

Parameterization

- An example of factor product.



Parameterization

- Note that the factors are not marginals.
- In the misconception model, the marginal over A,B is

$$\begin{array}{l|l} a^0 & b^0 \\ a^0 & b^1 \\ a^1 & b^0 \\ a^1 & b^1 \end{array} \left| \begin{array}{l} 0.13 \\ 0.69 \\ 0.14 \\ 0.04 \end{array} \right.$$

but the factor is

$$\begin{array}{l|l} a^0 & b^0 \\ a^0 & b^1 \\ a^1 & b^0 \\ a^1 & b^1 \end{array} \left| \begin{array}{l} 30 \\ 5 \\ 1 \\ 10 \end{array} \right.$$

- A factor is only one contribution to the overall joint distribution.
- The distribution as a whole has to take into consideration the contributions from all of the factors involved.

Today's Topics

- Introduction
- Parameterization
- *Gibbs Distributions*
- Reduced Markov Networks
- Markov Network Independencies
- Learning Undirected Models

Gibbs Distributions

- We can use the more general notion of factor product to define an undirected parametrization of a distribution.
- A distribution P_{Φ} is a *Gibbs distribution* parametrized by a set of factors $\Phi = \{\phi_1(\mathbf{D}_1), \dots, \phi_K(\mathbf{D}_K)\}$ if it is defined as follows:

$$P_{\Phi}(X_1, \dots, X_n) = \frac{1}{Z} \phi_1(\mathbf{D}_1) \times \dots \times \phi_K(\mathbf{D}_K)$$

where

$$Z = \sum_{X_1, \dots, X_n} \phi_1(\mathbf{D}_1) \times \dots \times \phi_K(\mathbf{D}_K)$$

is the partition function.

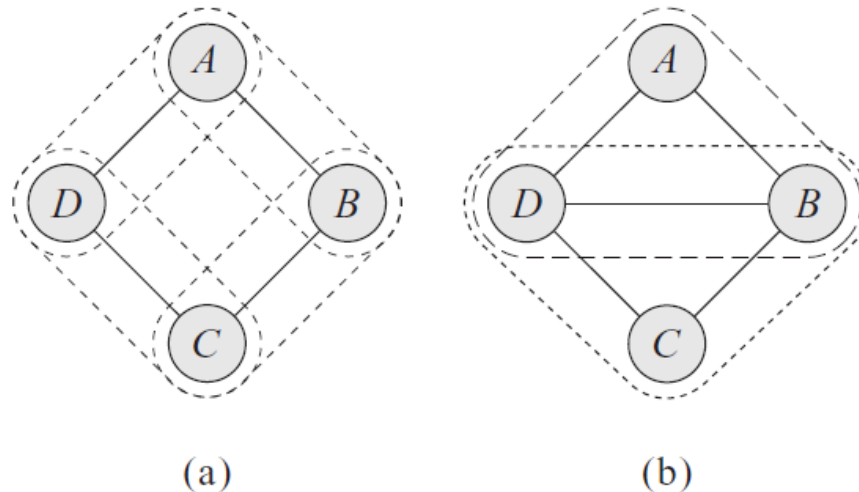
- The D_i are the scopes of the factors.

Gibbs Distributions

- If our parameterization contains a factor whose scope contains both X and Y , we would like the associated Markov network structure \mathcal{H} to contain an edge between X and Y .
- We say that a distribution P_{Φ} with $\Phi = \{\phi_1(\mathbf{D}_1), \dots, \phi_K(\mathbf{D}_K)\}$ factorizes over a Markov network \mathcal{H} if each $\mathbf{D}_k, k = 1, \dots, K$, is a *complete subgraph* of \mathcal{H} .
- The factors that parameterize a Markov network are often called *clique potentials*.

Gibbs Distributions

- We can reduce the number of factors by allowing factors only for *maximal cliques*.
- However, the parameterization using maximal clique potentials generally obscures structure that is present in the original set of factors.



The cliques in two simple Markov networks. (a) $\{A,B\}$, $\{B,C\}$, $\{C,D\}$, and $\{D,A\}$. (b) $\{A,B,D\}$ and $\{B,C,D\}$.

Today's Topics

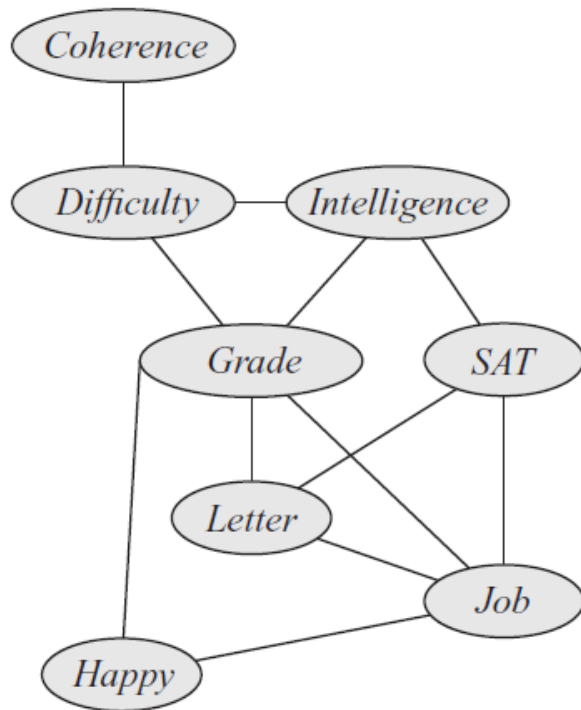
- Introduction
- Parameterization
- Gibbs Distributions
- *Reduced Markov Networks*
- Markov Network Independencies
- Learning Undirected Models

Reduced Markov Networks

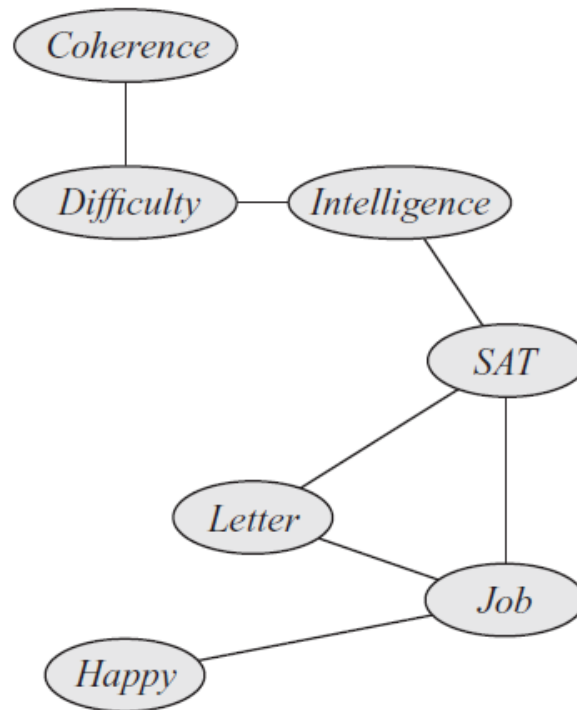
- If we observe some values, $U = u$, in the factor value table, we can eliminate the entries which are inconsistent with $U = u$.
- Let H be a Markov network over X and $U = u$ a context. The reduced Markov network $H[u]$ is a Markov network over the nodes $W = X - U$, where we have an edge $X - Y$ if there is an edge $X - Y$ in H .

Reduced Markov Networks

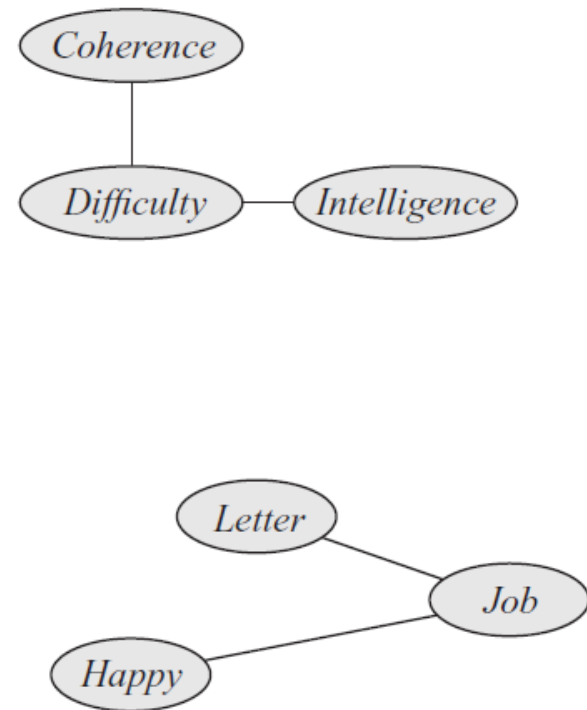
- A reduced Markov network example. (a) Original set of factors. (b) Reduced to the context $G = g$. (c) Reduced to the context $G = g, S = s$.



(a)



(b)



(c)

Reduced Markov Networks

- Conditioning on a context U in Markov networks eliminates edges from the graph.
- In a Bayesian network, conditioning on evidence can create new dependencies.

Reduced Markov Networks

- Markov Random Fields:
 - Pairwise Markov network.
 - They are simple.
 - Interactions on edges are an important special case that often arises in practice.

Today's Topics

- Introduction
- Parameterization
- Gibbs Distributions
- Reduced Markov Networks
- *Markov Network Independencies*
- Learning Undirected Models

Markov Network Independencies

- Let H be a Markov network and let $X_1 - \dots - X_k$ be a path in H .
- Let $Z \subseteq \mathcal{X}$ be a set of observed variables.
- The path $X_1 - \dots - X_k$ is active given Z if none of the X_i 's, $i = 1, \dots, k$, is in Z .
- A set of nodes Z separates X and Y in H , denoted $\text{sep}_H(X; Y | Z)$, if there is no active path between any node $X \in X$ and $Y \in Y$ given Z .
- We define the global independencies associated with H to be

$$\mathcal{I}(H) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : \text{sep}_H(\mathbf{X}; \mathbf{Y} | \mathbf{Z})\}$$

Today's Topics

- Introduction
- Parameterization
- Gibbs Distributions
- Reduced Markov Networks
- Markov Network Independencies
- *Learning Undirected Models*

Learning Undirected Models

- Like in Bayesian networks, once the joint distribution is generated, any kind of question can be answered using conditional probabilities and marginalization.
- However, a key distinction between Markov networks and Bayesian networks is normalization.
- Markov networks use a global normalization constant called the partition function.
- Bayesian networks involve local normalization within each conditional probability distribution.

Learning Undirected Models

- The global factor couples all of the parameters across the network, preventing us from decomposing the problem and estimating local groups of parameters separately.
- The global parameter coupling has significant computational ramifications.
- Even the simple maximum likelihood parameter estimation with complete data cannot be solved in closed form.

Learning Undirected Models

- We generally have to resort to iterative methods such as gradient ascent.
- The good news is that the likelihood objective is concave, so the methods are guaranteed to converge to the global optimum.
- The bad news is that each of the steps in the iterative algorithm requires that we run inference on the network, making even simple parameter estimation a fairly expensive process.