CS 583: Probabilistic Graphical Models

Topic: Conditional Probability Distributions (CPDs)

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

CPDs

- So far, we assumed tabular representation of Conditional Probability Distributions (CPDs); these are also called Conditional Probability Tables (CPTs)
- A CPD encodes $P(X \mid Pa(X))$ where $\sum P(X \mid Pa(X)) = 1$
- CPTs have significant disadvantages
  - Cannot handle variables with infinite domains
  - The number of independent parameters needed is $|Val(X)\text{-}1| * |Val(Pa(X))|$, which is exponential in the number of parents
  - Ignores the local structure

Example – Medical diagnosis

- A patient can suffer from one or more of many diseases; each disease is represented as a binary variable (Present/NotPresent)
- A symptom, e.g., Fever, can have many causes; this requires many disease nodes to be the parents of a symptom
- If Fever is caused by 10 different diseases, then the CPT for Fever requires $2^{10} = 1{,}024$ independent parameters (assuming a patient can have more than one disease at the same time, and we represent each disease as present/notpresent; and fever is also binary)

$p(fev \mid D_1 \ldots D_n)$

$(2-1) \cdot 2^n$

Key Insight

- A CPD must represent $P(X \mid Pa(X))$ where $\sum P(X \mid Pa(X)) = 1$
- CPD does not have to list all possible combinations
- CPD is a function that maps (x, Pa(x)) to a conditional distribution P(x | Pa(x))
- This representation guarantees that the joint is a well-defined probability distribution

Types of CPDs we will discuss

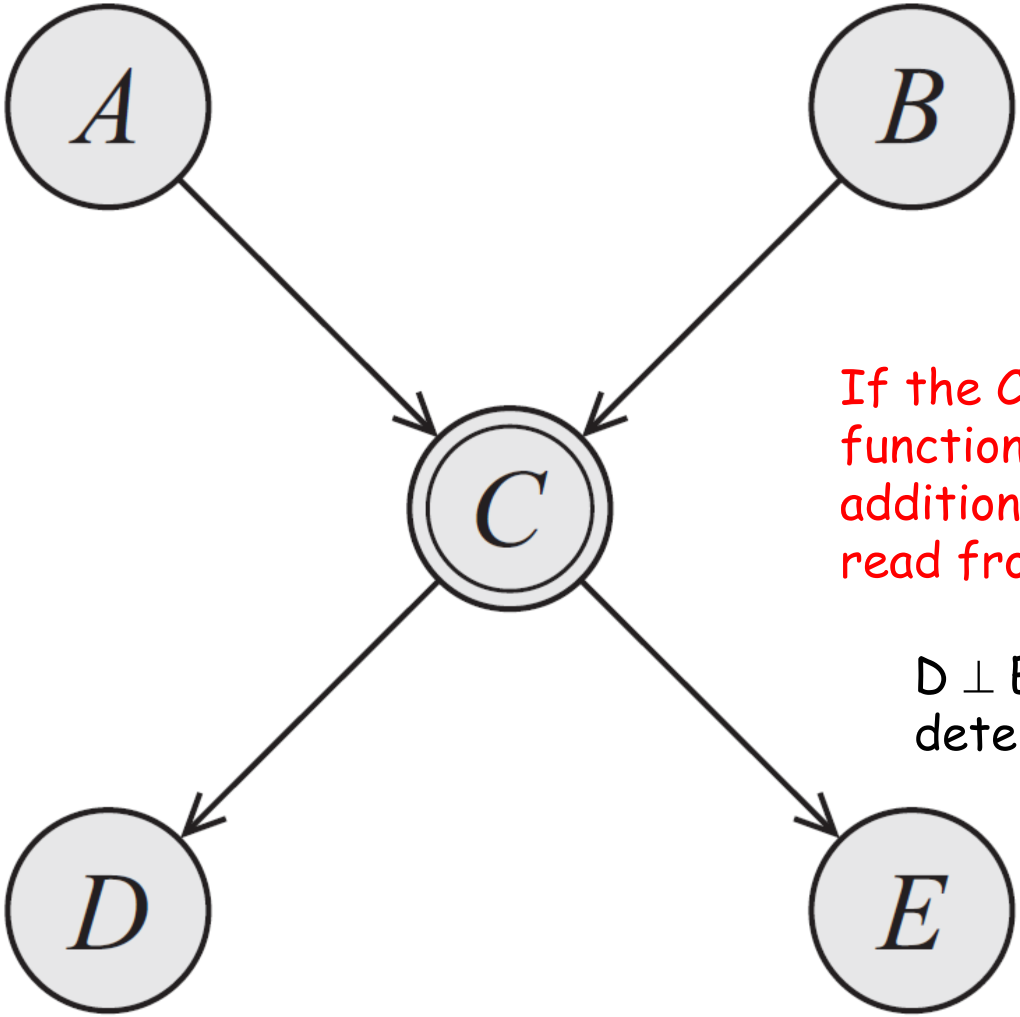
- Deterministic CPDs
- Context-Specific CPDs
  - Tree CPDs
  - Rule CPDs
- Causal independence
  - The noisy-or model
  - Logistic CPD
- Continuous variables
  - Linear Gaussian CPDs

DETERMINISTIC CPDS

- X is a deterministic function of its parents Pa(X)
- There is a deterministic function f : Val(Pa(X)) $\rightarrow$ Val(X) such that
  - $P(x \mid Pa(x)) = 1$ if $x = f(Pa(x))$

    0 otherwise
- Examples
  - X is (OR, AND, XOR, …) of its parents
  - X is average of its parents

INDEPENDENCIES



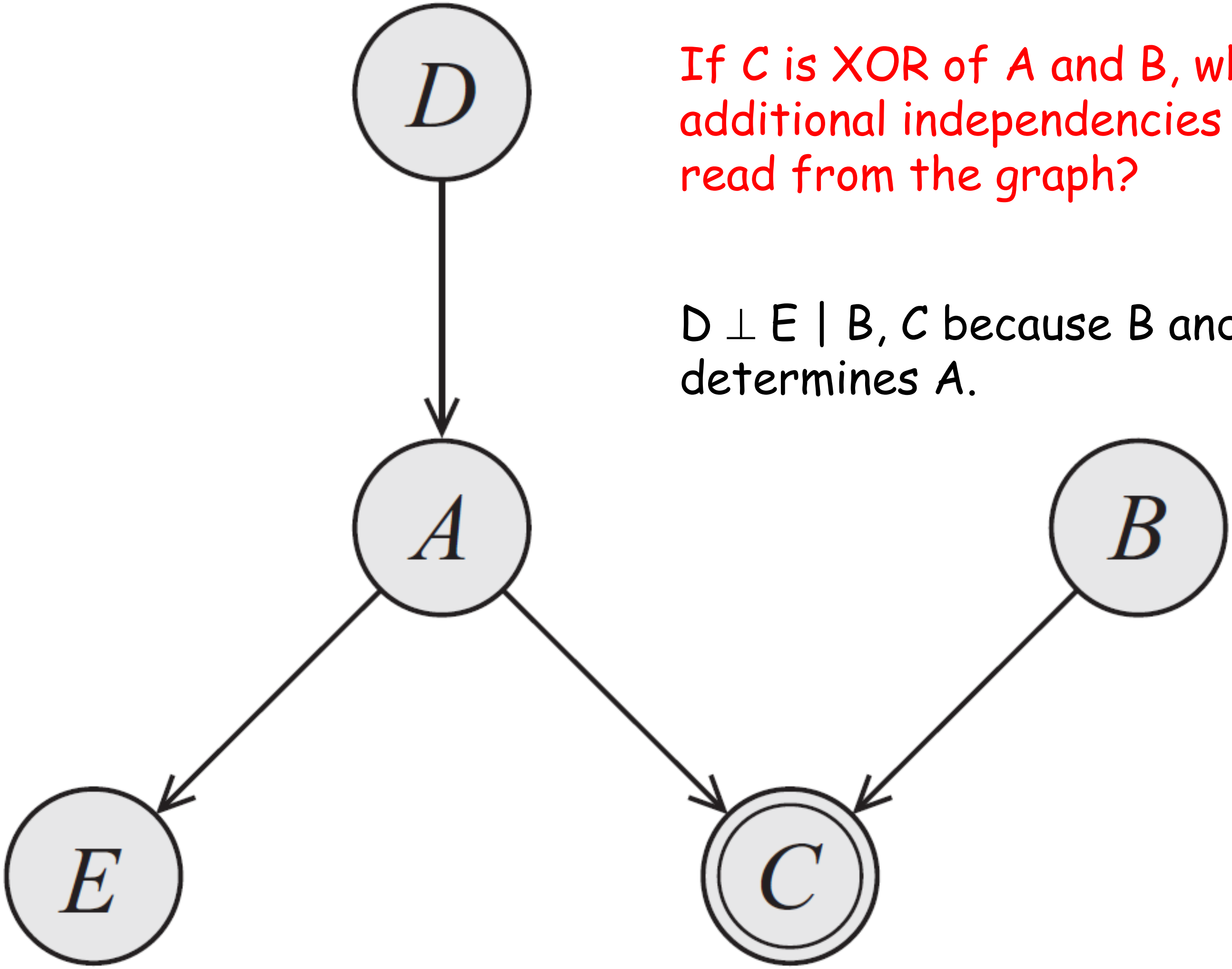
If the CPD of C is a deterministic function of A and B, what additional independencies can we read from the graph?

D $\perp$ E | A, B because A and B determines C.

INDEPENDENCIES

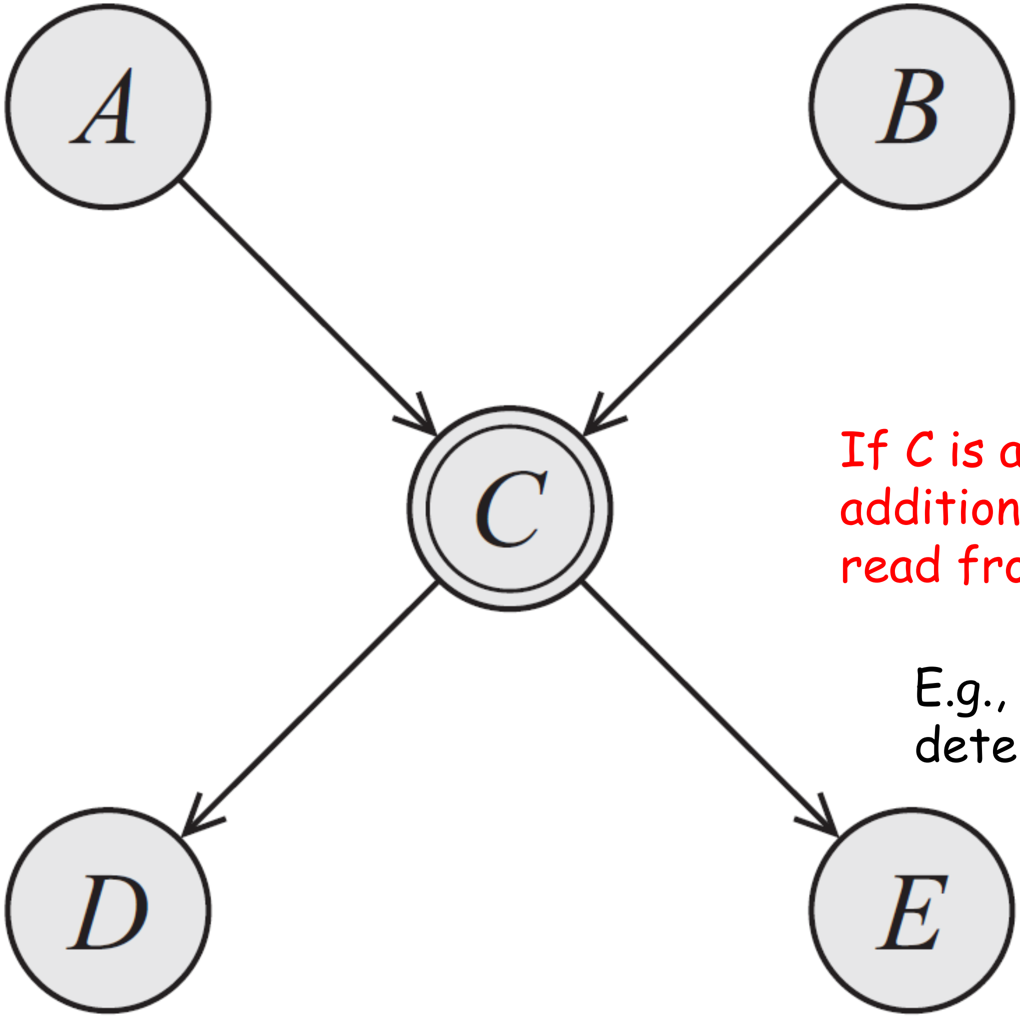
$D \perp E \mid A$

If C is XOR of A and B, what additional independencies can we read from the graph?



$D \perp E \mid B, C$ because B and C determines A.

INDEPENDENCIES



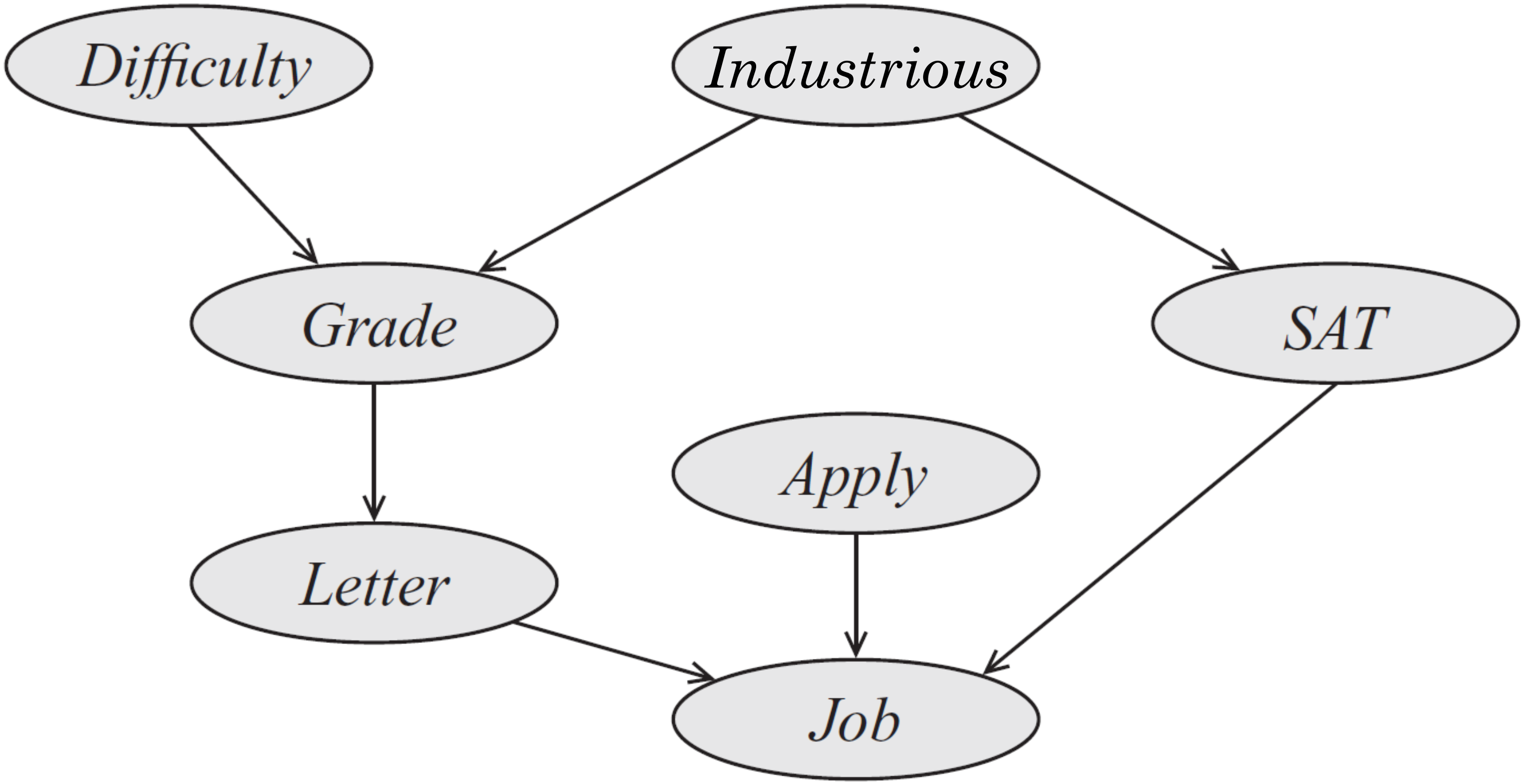
If C is an OR of A and B, what additional independencies can we read from the graph?

E.g., $D \perp E \mid A=T$ because $A=T$ determines C.

Context-specific independence

- **Definition**: Let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ be pairwise disjoint sets of variables, $\boldsymbol{C}$ be a set of variables (that might overlap with $\boldsymbol{X}\cup\boldsymbol{Y}\cup\boldsymbol{Z}$), and let $\boldsymbol{c} \in \mathrm{val}(\boldsymbol{C})$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *contextually independent* given $\boldsymbol{Z}$ <u>and</u> the context $\boldsymbol{c}$ if
  - $P(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{c}) = P(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{c})$ whenever $P(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{c}) > 0$.

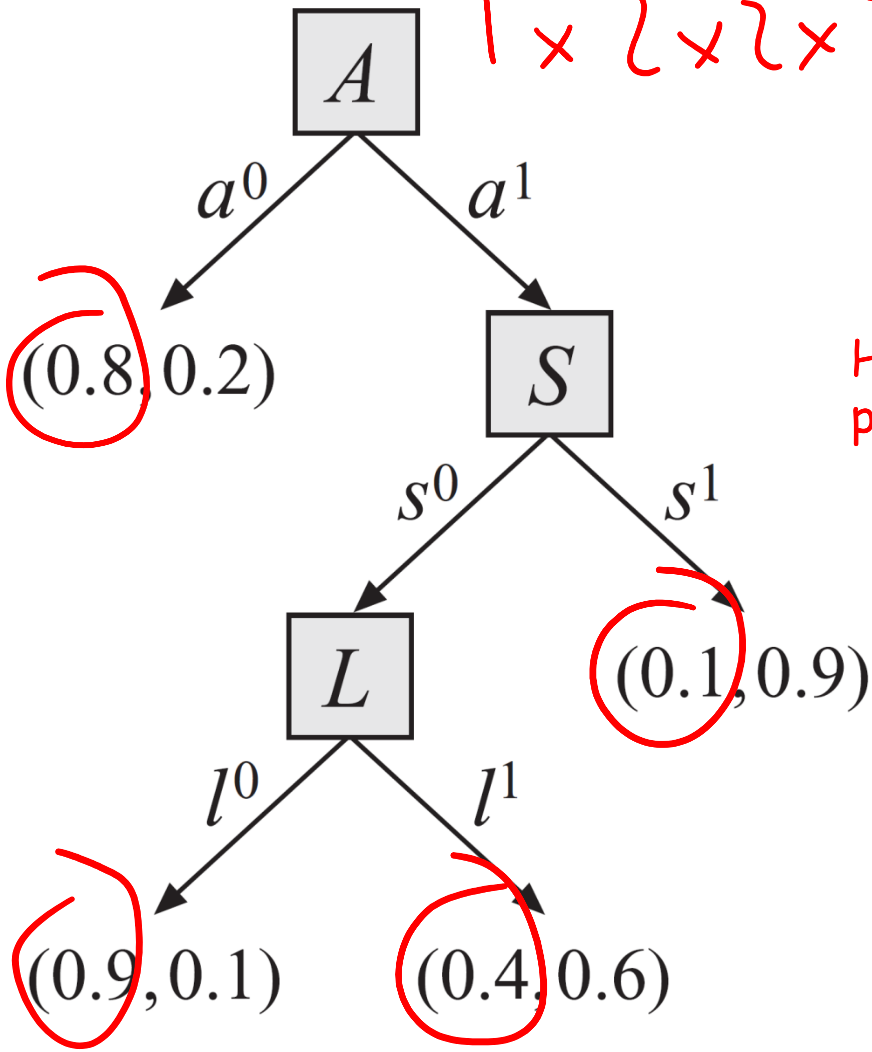
Example: job application



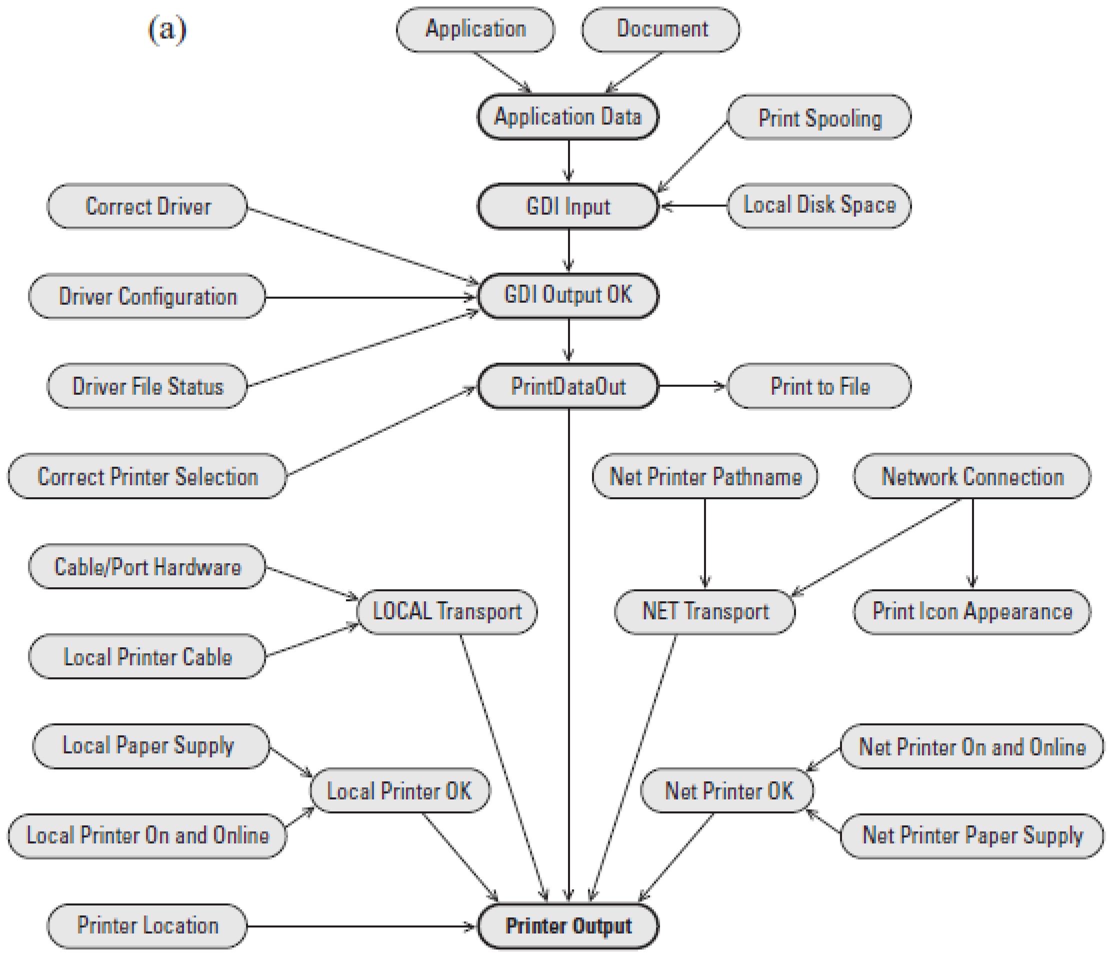
Tree-CPDs

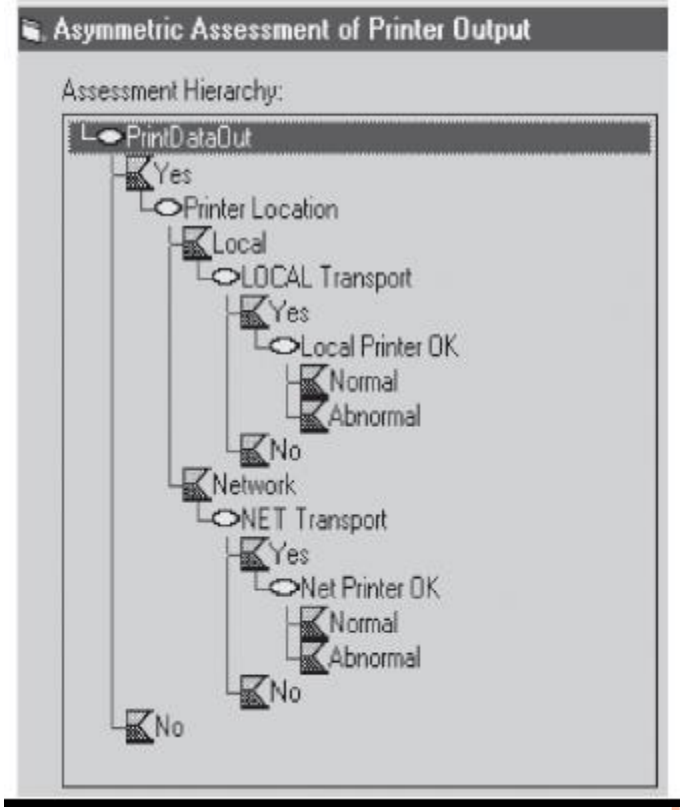
$P(J \mid A, S, L)$

$1 \times 2 \times 2 \times 2 = 8$



How many independent parameters are needed?


(a)
Application
Document
Application Data
Print Spooling
Correct Driver
GDI Input
Local Disk Space
Driver Configuration
GDI Output OK
Driver File Status
PrintDataOut
Print to File
Correct Printer Selection
Net Printer Pathname
Network Connection
Cable/Port Hardware
LOCAL Transport
NET Transport
Print Icon Appearance
Local Printer Cable
Local Paper Supply
Net Printer On and Online
Local Printer OK
Net Printer OK
Local Printer On and Online
Net Printer Paper Supply
Printer Location
Printer Output


Asymmetric Assessment of Printer Output
Assessment Hierarchy:
PrintDataOut
Yes
Printer Location
Local
LOCAL Transport
Yes
Local Printer OK
Normal
Abnormal
No
Network
NET Transport
Yes
Net Printer OK
Normal
Abnormal
No
No

RULE CPDS

- Definition: A rule-based CPD $P(X \mid Pa_X)$ is a set of rules $\mathcal{R}$ such that
  - For each rule $r \in \mathcal{R}$, Scope[r] $\subseteq \{X\} \cup Pa_X$
  - For each assignment to $(x, \boldsymbol{u})$, we have precisely one rule $<\boldsymbol{c};p> \in \mathcal{R}$ such that $\boldsymbol{c}$ is compatible with $(x, \boldsymbol{u})$. In this case, $P(X=x \mid Pa_X = \boldsymbol{u}) = p$.
  - The resulting CPD $P(X \mid Pa_X)$ is a legal CPD,
    - $\Sigma_x P(x \mid \boldsymbol{u}) = 1$
- More general than tree CPDs
  - Every tree CPD can be represented as a rule-based CPD but the converse is not true

Example: rule CPDs

$$\begin{tikzcd} A \arrow[dr] & B \arrow[d] & C \arrow[dl] \\ & X & \end{tikzcd}$$

$$\rho_1 : \langle a^1, b^1, x^0; 0.1 \rangle \qquad \rho_2 : \langle a^1, b^1, x^1; 0.9 \rangle$$
$$\rho_3 : \langle a^0, c^1, x^0; 0.2 \rangle \qquad \rho_4 : \langle a^0, c^1, x^1; 0.8 \rangle$$
$$\rho_5 : \langle b^0, c^0, x^0; 0.3 \rangle \qquad \rho_6 : \langle b^0, c^0, x^1; 0.7 \rangle$$
$$\rho_7 : \langle a^1, b^0, c^1, x^0; 0.4 \rangle \qquad \rho_8 : \langle a^1, b^0, c^1, x^1; 0.6 \rangle$$
$$\rho_9 : \langle a^0, b^1, c^0; 0.5 \rangle$$

X	$a^0b^0c^0$	$a^0b^0c^1$	$a^0b^1c^0$	$a^0b^1c^1$	$a^1b^0c^0$	$a^1b^0c^1$	$a^1b^1c^0$	$a^1b^1c^1$
x^0	0.3	0.2	0.5	0.2	0.3	0.4	0.1	0.1
x^1	0.7	0.8	0.5	0.8	0.7	0.6	0.9	0.9

Independence of causal models

- Variable of interest Y depends on several causes $X_1, \ldots, X_k$
- Even though the interaction between X_i and Y can be arbitrary, it is often reasonable to assume that the combined influence of X_i on Y is a simple combination of the individual influences of X_i on Y in isolation.
  - Noisy-or model
  - Logistic CPD

Noisy-OR Model

- A professor writes a good letter if
  - The student asked good questions in class, or
  - The student wrote a good final paper
- However, the professor
  - Might forget that student asked good questions, with 0.2 probability
  - Might not be able to read student's handwriting, with 0.1 probability
- What is the probability that the professor will write a good letter if the student
  - Did not ask good questions, and did not write a good final paper? 0
  - Asked good questions but did not write a good final paper? $1-0.2=0.8$
  - Did not ask good questions but wrote a good final paper? $1-0.1=0.9$
  - Asked good questions, and wrote a good final paper?

$1-0.1\times0.2=0.98$

NOISY-OR MODEL

$P(L \mid Q, F)$

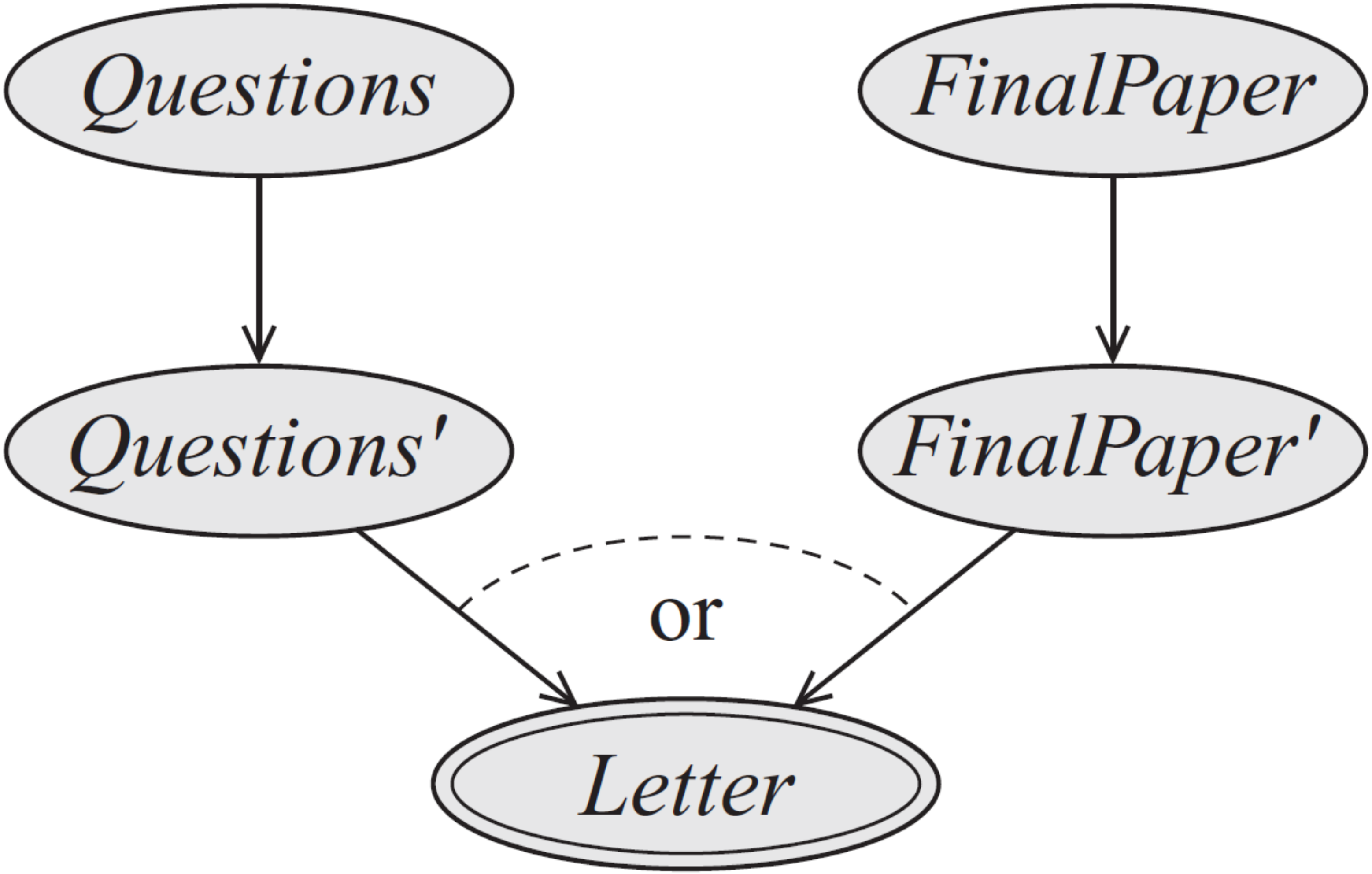
Q, F	l^0	l^1
q^0, f^0	1	0
q^0, f^1	0.1	0.9
q^1, f^0	0.2	0.8
q^1, f^1	?	?

For the professor to write a bad letter when student asked good questions and a good final paper, the professor

- Forgets the student's participation AND
- Cannot read the student's handwriting

$$P(l^0 \mid q^1, f^1) = 0.1 * 0.2 = 0.02$$
$$P(l^1 \mid q^1, f^1) = 1 - 0.02 = 0.98$$

Noisy-or model



Letter example $\lambda_0 = 0$ $\lambda_Q = 0.8$

NOISY-OR MODEL

$\lambda_F = 0.9$

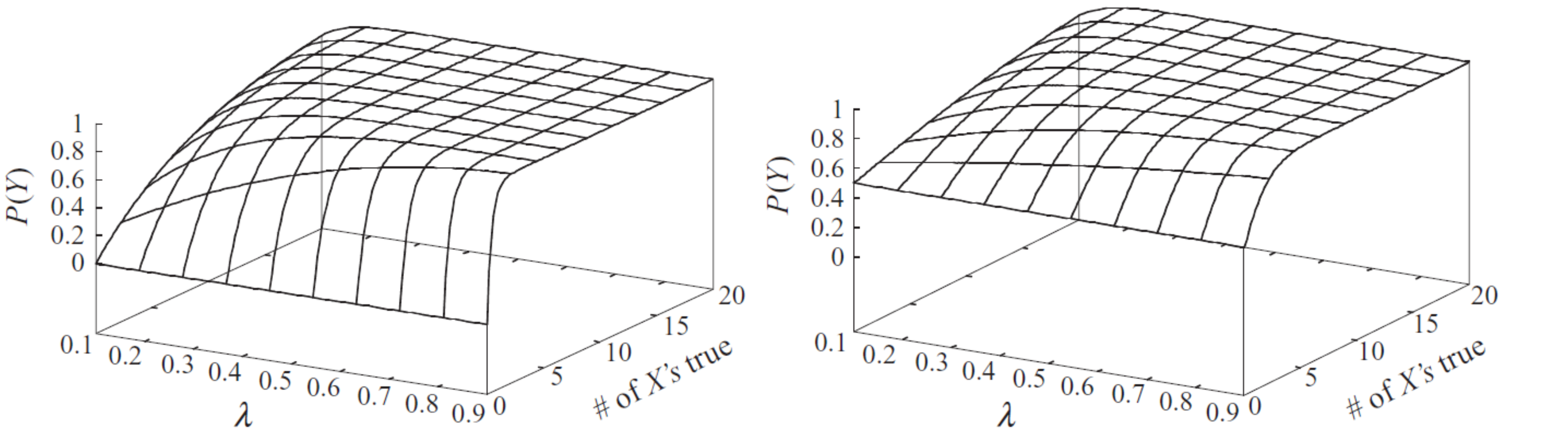
- Let Y be a binary random variable with k binary parents $X_1, \ldots, X_k$. The CPD $P(Y \mid X_1, \ldots, X_k)$ is a noisy-or if there are $k+1$ parameters $\lambda_0, \lambda_1, \ldots, \lambda_k$ such that

$Q = T$, $F_P = T$ $(1-0)\cdot(1-0.8)(1-0.9)$

$$P(y^0 \mid X_1, \ldots, X_k) = (1-\lambda_0) \prod_{i:X_i = x_i^1} (1-\lambda_i)$$

$$P(y^1 \mid X_1, \ldots, X_k) = 1-(1-\lambda_0) \prod_{i:X_i = x_i^1} (1-\lambda_i)$$

NOISY-OR MODEL



(a) (b)

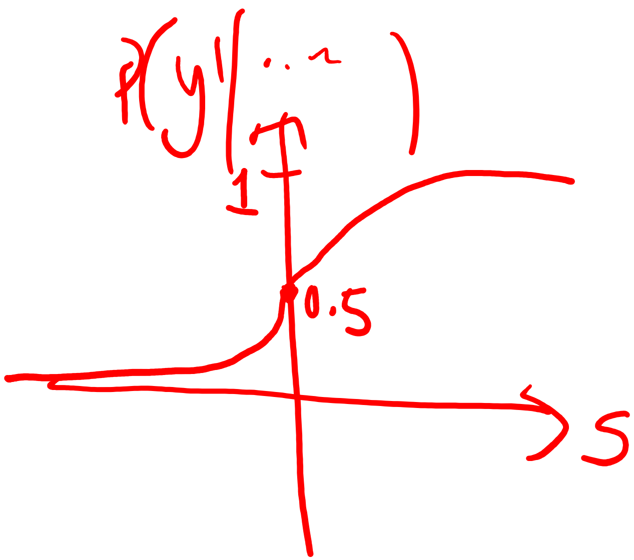
Logistic CPD

- Let Y be a binary variable with k parents: $X_1, \ldots, X_k$ that take on numerical values. The CPD P(Y| $X_1, \ldots, X_k$) is a *logistic CPD* if there are k+1 weights $w_0, w_1, \ldots, w_k$ such that

$$P(y^1 \mid X_1, \ldots, X_k) = sigmoid\left(w_0 + \sum_{i=1}^{k} w_i X_i\right)$$

S

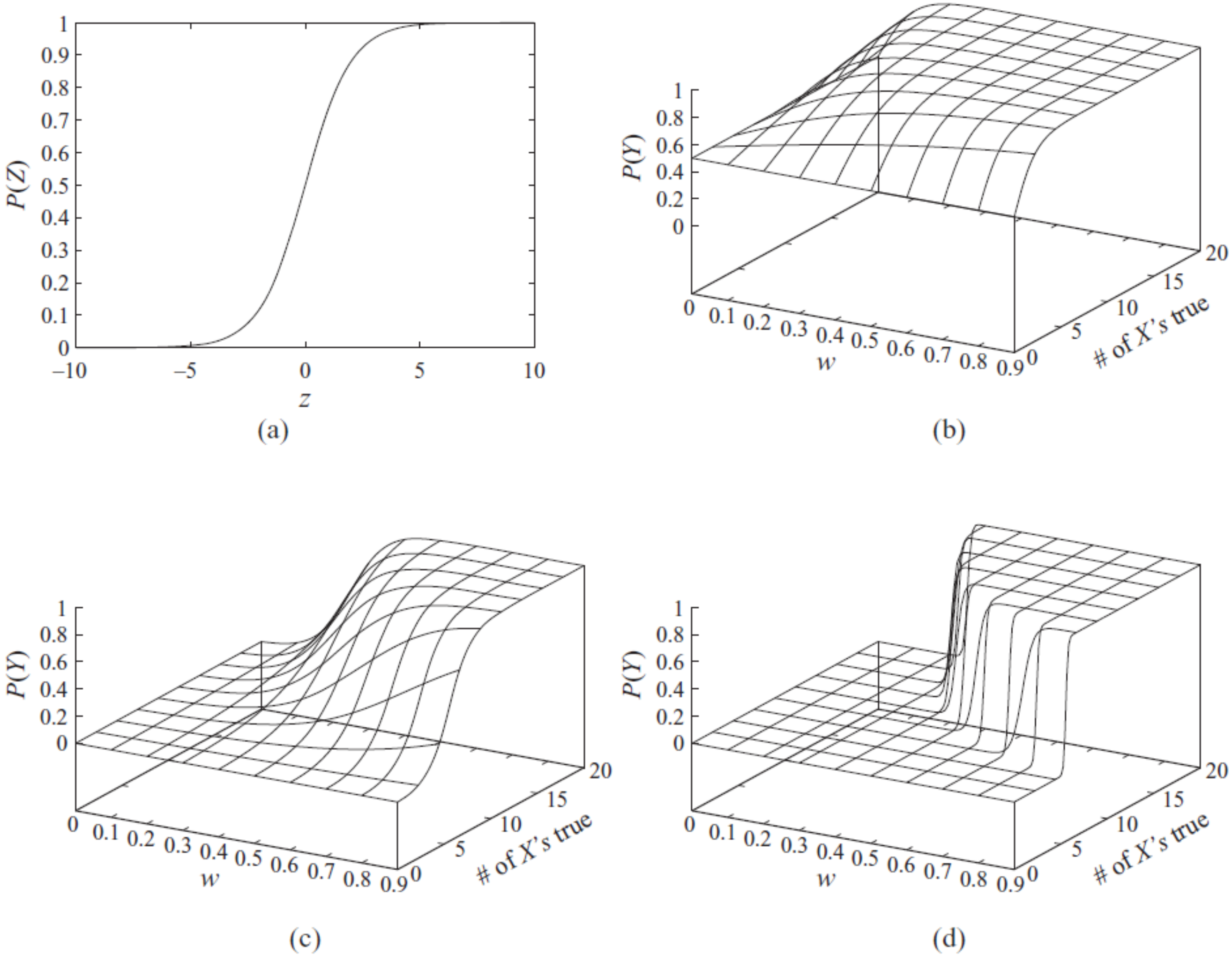
$$= \frac{e^{w_0 + \sum_{i=1}^{k} w_i X_i}}{1 + e^{w_0 + \sum_{i=1}^{k} w_i X_i}}$$



$s \to -\infty$

$s \to \infty$

Logistic CPD



Linear Gaussian CPD

- Let Y be a continuous variable with continuous parents X_1, … , X_k. We say that Y has a linear Gaussian model if there are parameters β_0, ..., β_k and σ^2 such that

$$p(Y \mid x_1, \ldots x_k) = \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k ; \sigma^2\right)$$