

# Artificial Intelligence

CS 165A

Apr 18, 2023

Instructor: Prof. Yu-Xiang Wang

Today

- Counting number of parameters
- Probabilistic modeling
- Factorization and conditional independence

# Notes on Project 1

- Use autograder for testing your code, but do not rely on it too much. It is better to debug your code directly.
- Numerical stability and log-sum-exp.
  - Implementation 1:
  - Implementation 2:

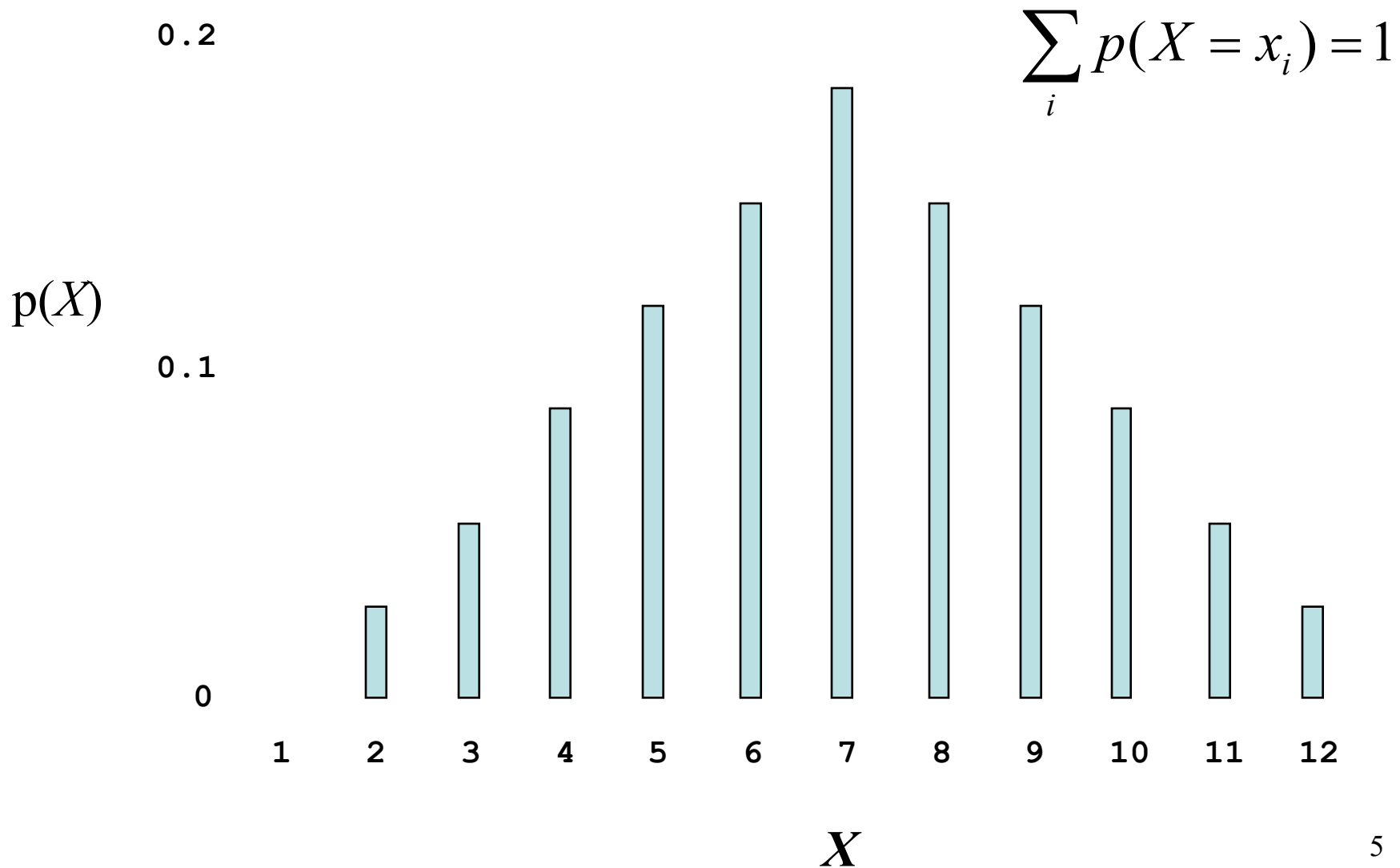
# Recap: Last lecture

- Stochastic gradient descent
- From linear logistic regression to neural networks
- Discriminative vs. Generative modelling
- I asked you to review Slide 26 – 33 of Lecture 4 for a review of probability notations.

# Probability notation and notes

- Probabilities of *propositions*
  - $P(A)$ ,  $P(\text{the sun is shining})$
- Probabilities of *random variables (r.v.)*
  - $P(X = x_1)$ ,  $P(Y = y_1)$ ,  $P(x_1 < X < x_2)$
- $P(A)$  usually means  $P(A = \text{True})$  (**A is a proposition, not a variable**)
  - This is a probability **value**
  - Technically,  $P(A)$  is a probability *function*
- $P(X = x_1)$ 
  - This is a probability **value** ( $P(X)$  is a probability *function*)
- $P(X)$ 
  - This is a probability mass function (**p.m.f.**) for discrete r.v. or a probability density function (**p.d.f.**) for continuous r.v.
- Technically, if  $X$  is an r.v., we should not write  **$P(X) = 0.5$** 
  - But rather  **$P(X = x_1) = 0.5$**

# Discrete probability distribution



# Continuous probability distribution

0.4

$$\int_{-\infty}^{\infty} p(X) = 1$$

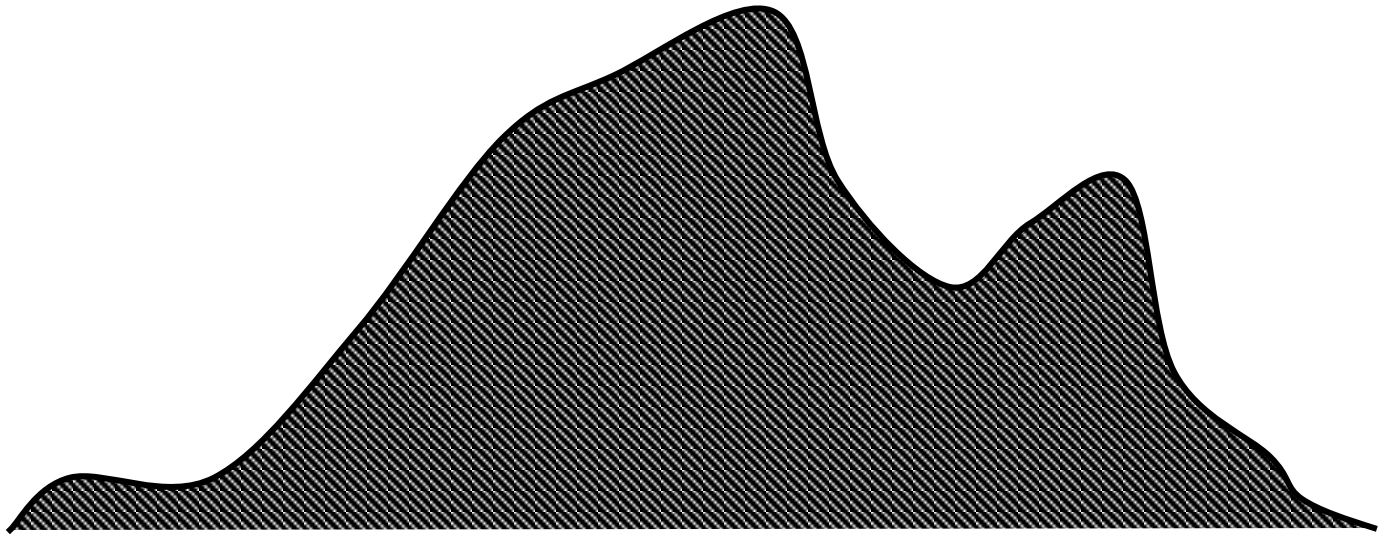
$p(X)$

0.2

0

1 2 3 4 5 6 7 8 9 10 11 12

$X$



# Joint Probabilities

- A **complete probability model** is a single joint probability distribution over all propositions/variables in the domain
  - $P(X_1, X_2, \dots, X_i, \dots)$
- A particular instance of the world has the probability
  - $P(X_1=x_1 \wedge X_2=x_2 \wedge \dots \wedge X_i=x_i \wedge \dots) = p$
- Rather than stating knowledge as
  - $\text{Raining} \Rightarrow \text{WetGrass}$
- We can state it as
  - $P(\text{Raining}, \text{WetGrass}) = 0.15$
  - $P(\text{Raining}, \neg\text{WetGrass}) = 0.01$
  - $P(\neg\text{Raining}, \text{WetGrass}) = 0.04$
  - $P(\neg\text{Raining}, \neg\text{WetGrass}) = 0.8$

	$\neg\text{WetGrass}$	$\text{WetGrass}$
$\neg\text{Raining}$	0.8	0.04
$\text{Raining}$	0.01	0.15

# Marginal and Conditional Probability

- Marginal Probability
  - Marginal probability (distribution) of  $X$ :  $P(X) = \sum_Y P(X, Y)$
  - **Bayesian interpretation:** Probabilities associated with one proposition or variable, **prior** to any evidence
  - E.g.,  $P(\text{WetGrass})$ ,  $P(\neg\text{Raining})$
- Conditional Probability
  - $P(A | B)$  – “The probability of  $A$  given that we know  $B$ ”
  - **Bayesian interpretation:** After (**posterior** to) procuring evidence
  - E.g.,  $P(\text{WetGrass} | \text{Raining})$

$$P(X | Y) = \frac{P(X, Y)}{P(Y)} \quad \text{or} \quad P(X | Y) P(Y) = P(X, Y)$$

Assumes  $P(Y)$  nonzero



# The chain rule: factorizing a joint distribution into marginal and conditionals

$$P(X, Y) = P(X | Y) P(Y)$$

## By the Chain Rule

$$P(X, Y, Z) = P(X | Y, Z) P(Y, Z)$$

$$= P(X | Y, Z) P(Y | Z) P(Z)$$

*or, equivalently*

$$= P(X) P(Y | X) P(Z | X, Y)$$

- Notes:
- Precedence: ‘|’ is lowest
  - E.g.,  $P(X | Y, Z)$  means which?  
 $P((X | Y), Z)$   
 $P(X | (Y, Z)) \leftarrow$

# Chain Rule implies Bayes' Rule

Thomas Bayes: 1701 - 1761



- Since  $P(X, Y) = P(X | Y) P(Y)$   
and  $P(X, Y) = P(Y | X) P(X)$
- Then  $P(X | Y) P(Y) = P(Y | X) P(X)$

$$P(X | Y) = \frac{P(Y | X) P(X)}{P(Y)}$$

Bayes' Rule

**Funny fact:** Thomas Bayes is arguably a frequentist.

Stephen Fienberg. "When did Bayesian inference become 'Bayesian'?" *Bayesian analysis* 1.1 (2006): 1-40.  
<https://projecteuclid.org/euclid.ba/1340371071>

# Representing Probability Distributions using numpy data structures

	<u>Continuous vars</u>	<u>Discrete vars</u>
$P(X)$	Function (of one variable)	m vector
$P(X=x)$	Scalar*	Scalar
$P(X,Y)$	Function of two variables	m×n matrix
$P(X Y)$	Function of two variables	m×n matrix
$P(X Y=y)$	Function of one variable	m vector
$P(X=x Y)$	Function of one variable	n vector
$P(X=x Y=y)$	Scalar*	Scalar

\* - actually zero. Should be  $P(x_1 < X < x_2)$

# Example: Joint probability distribution

From  $P(X, Y)$ , we can always calculate:

$P(X)$        $P(X=x_1)$   
 $P(Y)$        $P(Y=y_2)$   
 $P(X|Y)$      $P(X|Y=y_1)$   
 $P(Y|X)$      $P(Y|X=x_1)$   
 $P(X=x_1|Y)$   
etc.

		<b>X</b>		
		$x_1$	$x_2$	$x_3$
<b>Y</b>	$y_1$	0.2	0.1	0.1
	$y_2$	0.1	0.2	?

**P(X,Y)**

	$x_1$	$x_2$	$x_3$
$y_1$	0.2	0.1	0.1
$y_2$	0.1	0.2	0.3

**P(X)**

$x_1$	$x_2$	$x_3$
0.3	0.3	0.4

**P(Y)**

$y_1$	0.4
$y_2$	0.6

**P(X|Y)**

	$x_1$	$x_2$	$x_3$
$y_1$	0.5	0.25	0.25
$y_2$	0.167	0.333	0.5

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

$P(X|Y=y_1) = ?$

$P(X=x_1|Y) = ?$

**P(Y|X)**

	$x_1$	$x_2$	$x_3$
$y_1$	0.667	0.333	0.25
$y_2$	0.333	0.667	0.75

# Quick checkpoint

- Probability notations
  - $P(A)$  is a number when  $A$  is an event / predicate.
  - $P(X)$  is a vector/function when  $X$  is a random variable.
- Joint probability distribution
  - Enumerating all combinations of events.
  - All values the random variables can take.
  - Assign a non-negative value to each.
- Marginals, conditionals
  - How they are related: Chain rule, Bayes rule

# You should know HOW TO do the following:

- For discrete probability distributions for multiple random variables
  - Know **the number of possible values** these RVs can take
  - Know the **shape of the numpy arrays** that you need to represent Joint-distribution, conditional distribution
  - Know the **number of independent parameters** you need to specify these distributions. (we often need to subtract 1 here or there. Why is that?)
- More generally: Know the distinctions between
  - p.m.f -- probability mass function (for discrete distribution)
  - p.d.f. -- probability density function (for continuous distribution)
  - CDF -- cumulative distribution function (for both)

# Plan for today

- Modeling the world with probabilities:
  - Case study of “author classification”
- Conditional independences and factorization
- Introduction to BayesNet
- Reading materials:
  - Today’s lecture: AIMA Ch. 12.1-12.6
  - Thursday’s lecture: AIMA Ch. 13.1 – 13.3 / Jordan PGM Ch. 2.1



# How can a joint-distribution help us?

- A principled way to model the world
  - Handles missing data/variables
  - Easy to incorporate prior knowledge
- With the joint distribution, we can do anything we want
  - Design classifiers
$$\hat{y} = \arg \max_y P(Y = y|X)$$
  - We can make Bayesian inference (probabilistic reasoning)
    - Sherlock Holmes: P(Murderer | Observed Evidence)
    - Doctor: P( Disease | Symptoms ), **P(Effect | Treatment)**
    - Parenting:
      - P( Dirty Diaper, Hungry, Lonely | 5 a.m., Baby crying)
      - **P(Baby crying at 5 a.m. | feeding at 2 a.m.)**
      - **P(Baby crying at 5 a.m. | feeding at 1 a.m.)**

# A probabilistic reasoning agent

- PEAS for a probabilistic reasoning agent
  - Environment: A joint distribution of variables
  - Actuators: Compute conditional probabilities
  - Sensors: Observe evidence
  - Performance measure: How accurate is the model / inference?
- Modeling / Inference / Learning
  - Modeling: Specifying a family of joint distributions
  - Learning: Finding one joint distribution that fits / explains the data
  - Inference: Making probabilistic inferences

# (3 min discussion) Modeling the world with probability distribution

- Example: Author attribution
  - Variables: *Word 1, Word 2, Word 3, ..., Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Jane Austen, Tolkien, George RR. Martin, ... , Xueqin Cao, Douglas Adams}
  - A vocabulary of size 3000
- Questions:
  - What is the dimension(s) of the joint probability table?
  - How many **independent** parameters are needed to represent this distribution?
- Answers:

# Statistical Independences

## (Marginal / absolute) Independence

- X and Y are independent iff
  - $P(X, Y) = P(X) P(Y)$  [by definition]
  - $P(X | Y) = P(X)$     Since  $P(X | Y) = P(X, Y)/P(Y) = P(X) P(Y)/P(Y)$

## Conditional Independence

- If X and Y are (conditionally) independent given Z, then
  - $P(X | Y, Z) = P(X | Z)$
  - Example:
    - $P(\text{WetGrass} | \text{Season}, \text{Rain}) = P(\text{WetGrass} | \text{Rain})$

# Example of Conditional Independence

- In practice, conditional independence is more common than marginal independence.
  - $P(\text{Final exam grade} \mid \text{Weather}) \neq P(\text{Final exam grade})$ 
    - i.e., they are not independent
  - $P(\text{Final exam grade} \mid \text{Weather, Effort}) = P(\text{Final exam grade} \mid \text{Effort})$ 
    - But they are conditionally independent given Effort

# (Example continued) Modeling the world with probability distribution

- Example: Author attribution
  - Variables: *Word 1, Word 2, Word 3, ..., Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Jane Austen, Tolkien, George RR. Martin, ... , Xueqin Cao, Douglas Adams}
  - A vocabulary of size 3000
- In addition, assume that: *Word 1, ..., Word N* are **mutually independent** given *Author*
  - $P(\text{Word } 1, \dots, \text{Word } N \mid \text{Author}) = P(\text{Word } 1 \mid \text{Author}) \times \dots \times P(\text{Word } N \mid \text{Author})$
- Questions:
  - What is the dimension(s) of the joint probability table?
  - How many **independent** parameters?

# (Example continued) Modeling the world with probability distribution

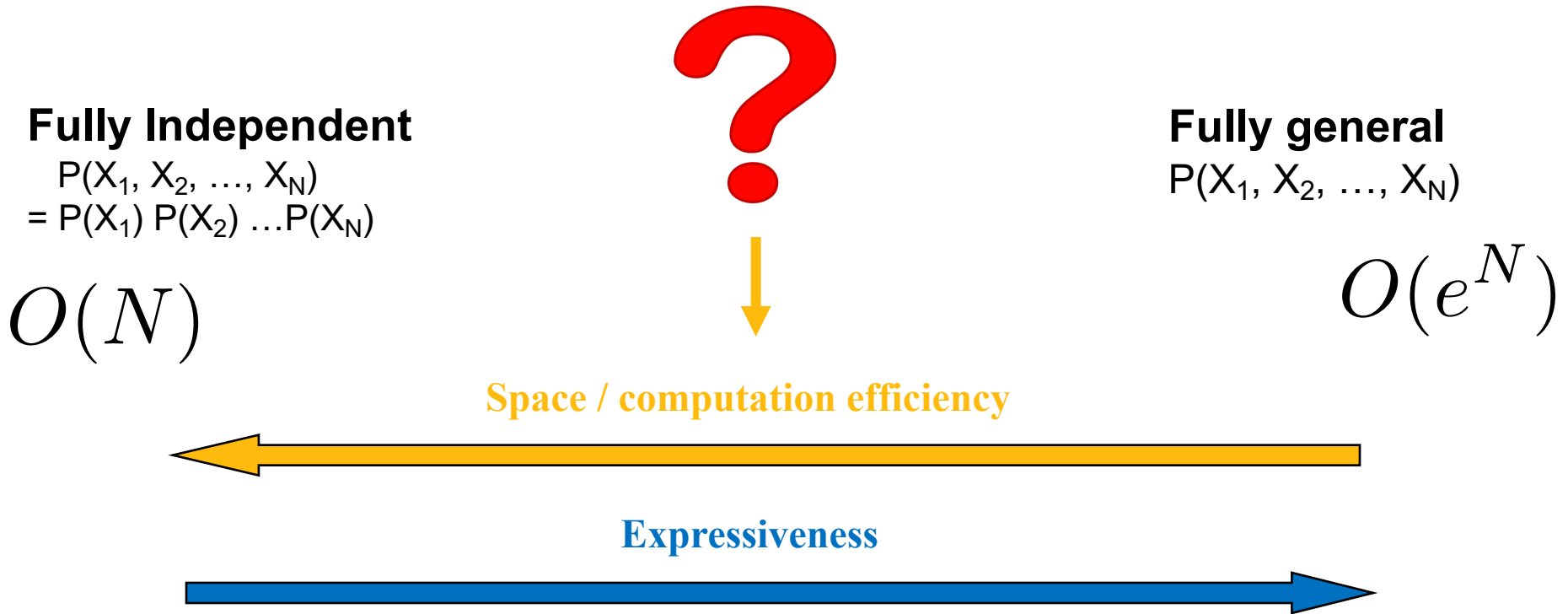
- Example: Author attribution
  - Variables: *Word 1, Word 2, Word 3, ..., Word N, Author*
  - 15 authors in total: {Dickens, Shakespeare, Jane Austen, Tolkien, George RR. Martin, ... , Xueqin Cao, Douglas Adams}
  - A vocabulary of size 3000
- In addition, assume that: *Word 1, ..., Word N* are **mutually independent** and **Identically Distributed** given *Author*?
  - $P(\text{Word } 1, \dots, \text{Word } N \mid \text{Author}) = P(\text{Word } 1 \mid \text{Author}) \times \dots \times P(\text{Word } N \mid \text{Author})$
- Questions:
  - What is the dimension(s) of the joint probability table?
  - How many **independent** parameters?

# Quiz time: Representing a joint Probability

- Joint probability:  $P(X_1, X_2, \dots, X_N)$ 
  - Defines the probability for any possible state of the world
  - Let the variables **be binary**. How many numbers (“free parameters”) does it take to define the joint distribution?
  
- If the variables are independent, then
$$P(X_1, X_2, \dots, X_N) = P(X_1) P(X_2) \dots P(X_N)$$
  - How many numbers does it take to define the joint distribution?



# Tradeoffs in our model choices



## Idea:

1. Independent groups of variables?
2. Conditional independences?

# Benefit of conditional independence

- If some variables are conditionally independent, the joint probability can be specified with many fewer than  $2^N-1$  numbers (or  $3^N-1$ , or  $10^N-1$ , or...)
- For example: (for binary variables  $W, X, Y, Z$ )
  - $P(W,X,Y,Z) = P(W) P(X|W) P(Y|W,X) P(Z|W,X,Y)$ 
    - # of params?
  - But if  $Y$  and  $W$  are independent given  $X$ , and  $Z$  is independent of  $W$  and  $X$  given  $Y$ , then
    - $P(W,X,Y,Z) = P(W) P(X|W) P(Y|X) P(Z|Y)$
    - # of params?
- This is often the case in real problems.

# What happens to ChatGPT?

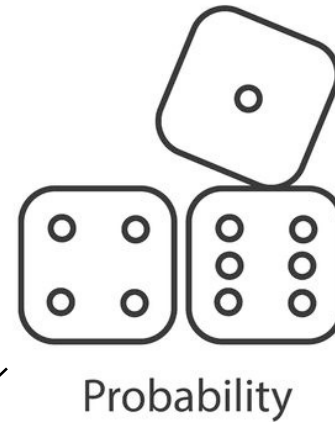
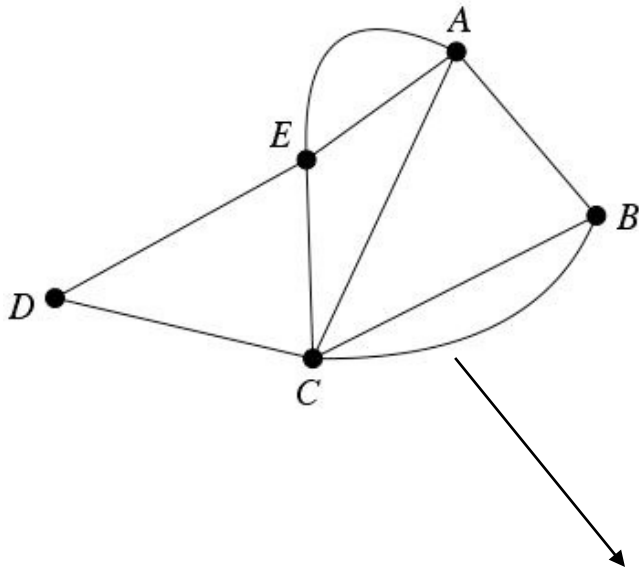
- How does ChatGPT models the joint distribution of words?
- “It predicts the next word!” How does it make inference (in math)?

When given a problem with many variables.

[CS165A Lecture attendance,  
Project 1, Project 2, Project 3,  
Readings, Ed Discussion, Final Grade,  
Weather, Job Offer]

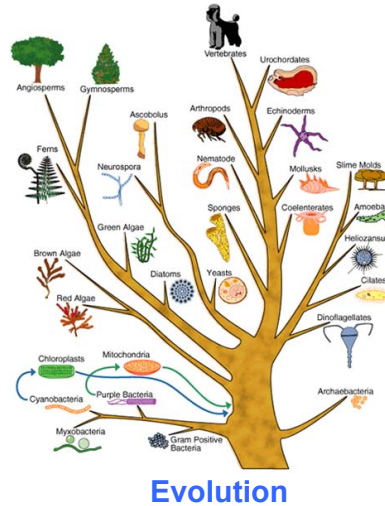
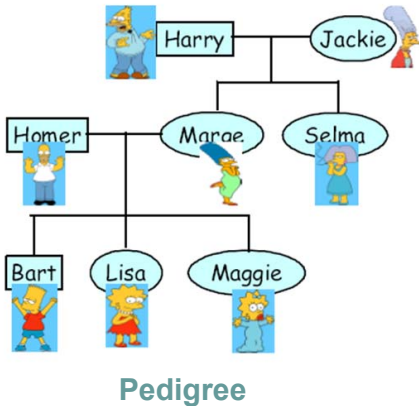
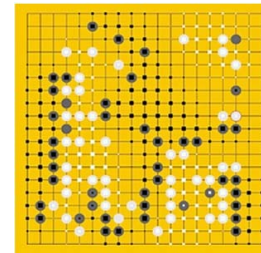
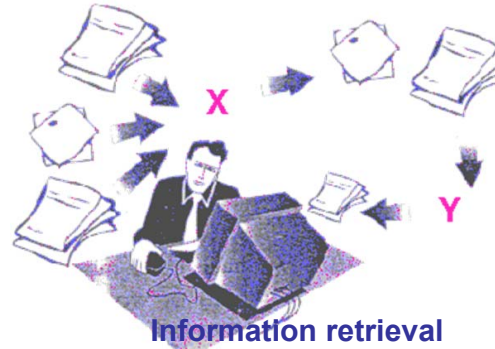
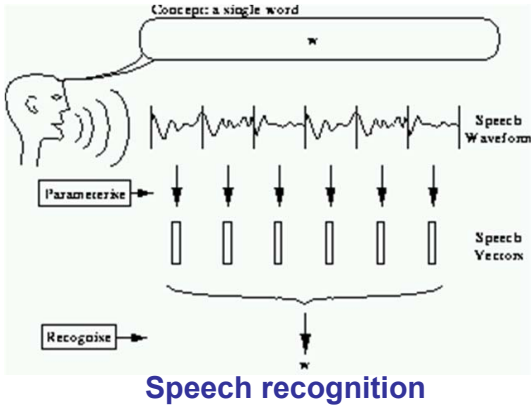
How do we know **which conditional independence(s)**  
to include in the joint distribution?

**Graphical models** come out of the marriage of graph theory and probability theory



**Directed Graph => Bayesian Networks / Belief Networks**  
**Undirected Graph => Markov Random Fields**

# Used as a modeling tool. Many applications!

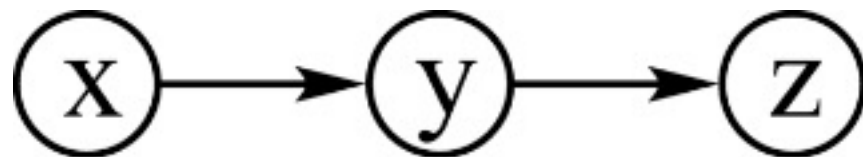


© Eric Xing @ CMU, 2005-2014

(Slides from Prof. Eric Xing)

# Two ways to think about Graphical Models

- A particular factorization of a joint distribution
  - $P(X,Y,Z) = P(X) P(Y|X) P(Z|Y)$
- A collection of conditional independences
  - $\{ X \perp Z \mid Y, \dots \}$



**Represented using a graph!**

# Bayesian Networks (BayesNets)

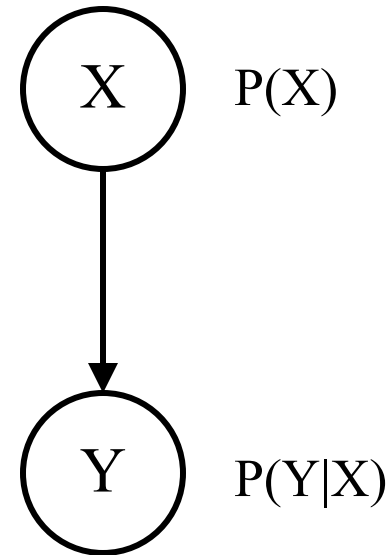
a.k.a. directed graphical models, Belief nets etc.

- Bayesian Network
  - A data structure (depicted as a graph) that represents the dependence among variables and allows us to concisely specify the joint probability distribution
  - The graph itself is known as an “influence diagram”
- A BayesNet is a **directed acyclic graph** where:
  - The nodes represent the set of random variables (one node per random variable)
  - Arcs between nodes represent *influence*, or *causality*
    - A link from node X to node Y means that X “directly influences” Y
  - Each node has a *conditional probability table* (CPT) that defines **P(node | parents)**



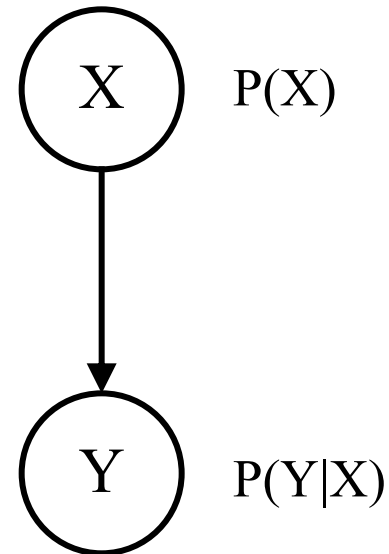
# Example

- Random variables  $X$  and  $Y$ 
  - $X$  – It is raining
  - $Y$  – The grass is wet
- $X$  has an *effect* on  $Y$   
Or,  $Y$  is a *symptom* of  $X$
- Draw two nodes and link them
- Define the CPT for each node
  - $P(X)$  and  $P(Y | X)$
- Typical use: we observe  $Y$  and we want to query  $P(X | Y)$ 
  - $Y$  is an *evidence variable*
  - $X$  is a *query variable*



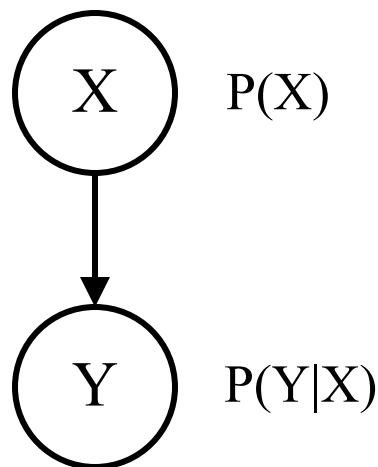
# We can write everything we want as a function of the CPTs. Try it!

- What is  $P(X | Y)$ ?
  - Given that we know the CPTs of each node in the graph

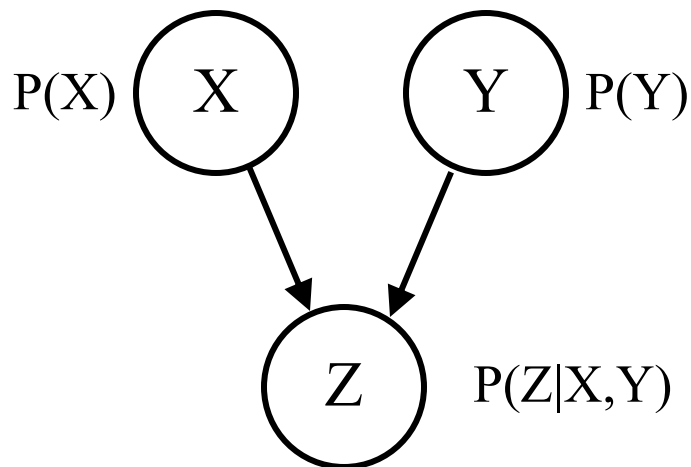


# BayesNets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(\text{var}_1, \dots, \text{var}_N) = \prod_i P(\text{var}_i | \text{Parents}(\text{var}_i))$



$$P(X, Y) = P(X) P(Y|X)$$



$$P(X, Y, Z) = P(X) P(Y) P(Z|X, Y)$$

# Three steps in modelling with BayesNet

1. Choose variables in the environments, represent them as nodes.
2. Connect the variables by inspecting the “direct influence”: cause-effect
3. Fill in the probabilities in the CPTs.

# Example: Modelling with BayesNet

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links
- This defines the joint probability
  - $P(\text{JohnCalls}, \text{MaryCalls}, \text{Earthquake}, \text{Burglar}, \text{Alarm})$
- What do we want to know?

$$P(\mathbf{B} \mid \mathbf{J}, \neg \mathbf{M})$$

# How should we connect the nodes? (3 min discussion)

Burglary

Earthquake

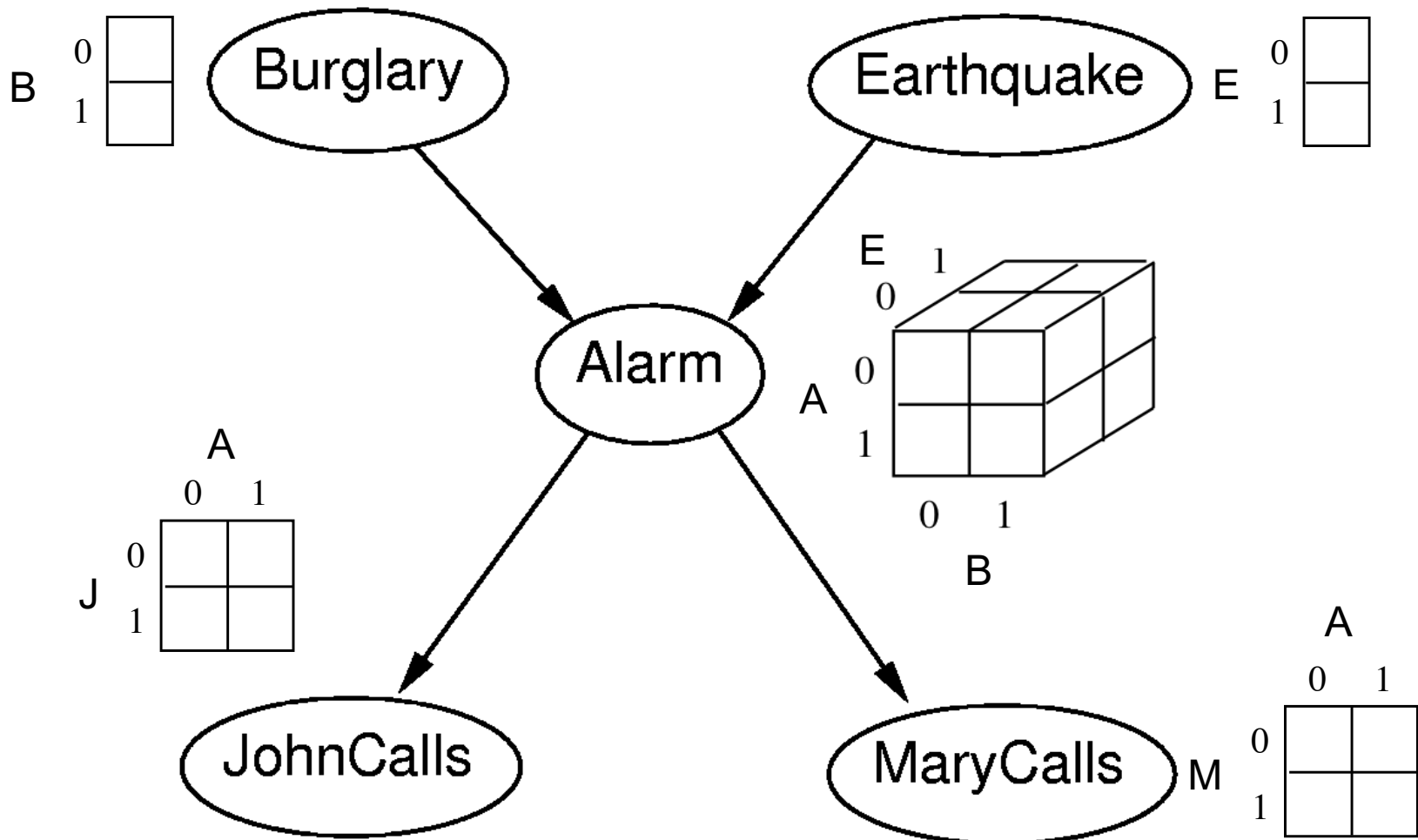
Alarm

JohnCalls

MaryCalls

Links and CPTs?

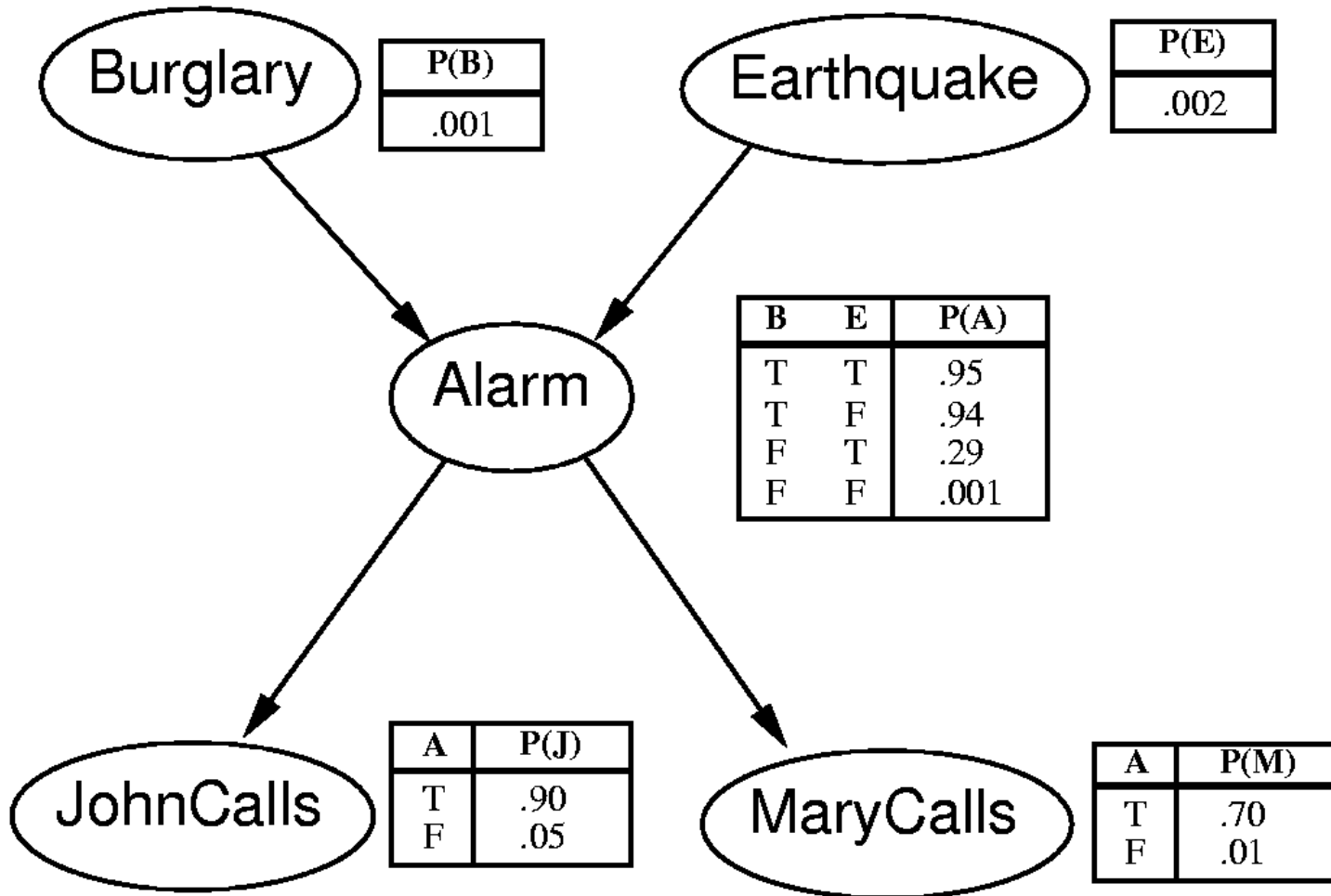
# What are the CPTs? What are their dimensions?



**Question: How to fill values into these CPTs?**

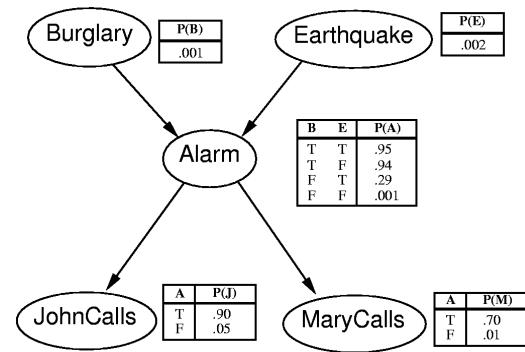
**Ans: Specify by hands. Learn from data (e.g., MLE).**

# Example



Joint probability?  $P(J, \neg M, A, B, \neg E)$ ?





Calculate  $P(J, \neg M, A, B, \neg E)$

Read the joint pf from the graph:

$$P(J, M, A, B, E) = P(B) P(E) P(A|B,E) P(J|A) P(M|A)$$

Plug in the desired values:

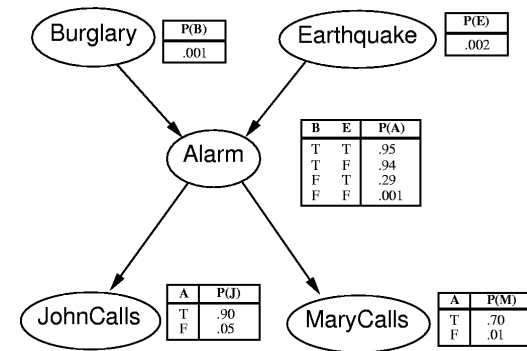
$$\begin{aligned} P(J, \neg M, A, B, \neg E) &= P(B) P(\neg E) P(A|B, \neg E) P(J|A) P(\neg M|A) \\ &= 0.001 * 0.998 * 0.94 * 0.9 * 0.3 \\ &= 0.0002532924 \end{aligned}$$

**How about  $P(B | J, \neg M)$  ?**

Remember, this means  $P(B=\text{true} | J=\text{true}, M=\text{false})$

Calculate  $P(B | J, \neg M)$

$$P(B | J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M)}$$



By marginalization:

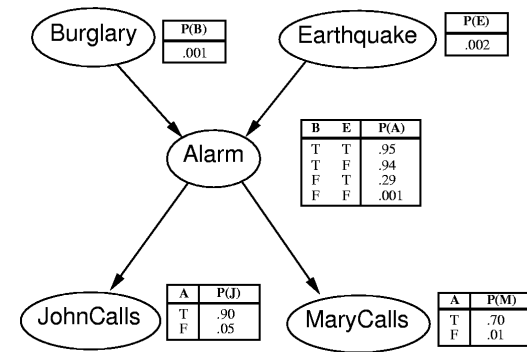
$$\begin{aligned}
 & \sum_i \sum_j P(J, \neg M, A_i, B, E_j) \\
 = & \frac{\sum_i \sum_j \sum_k P(J, \neg M, A_i, B_j, E_k)}{\sum_i \sum_j \sum_k P(B_j)P(E_k)P(A_i | B_j, E_k)P(J | A_i)P(\neg M | A_i)}
 \end{aligned}$$

# Quick checkpoint

- BayesNet as a modelling tool
- By inspecting the cause-effect relationships, we can draw directed edges based on our domain knowledge
- The product of the CPTs give the joint distribution
  - We can calculate  $P(A | B)$  for any A and B
  - The factorization makes it computationally more tractable

**What else can we get?**

# Example: Conditional Independence



- Conditional independence is seen here
  - $P(\text{JohnCalls} \mid \text{MaryCalls}, \text{Alarm}, \text{Earthquake}, \text{Burglary}) = P(\text{JohnCalls} \mid \text{Alarm})$
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
- Does this mean that an earthquake or a burglary do not influence whether or not John calls?
  - No, but the influence is already accounted for in the Alarm variable
  - JohnCalls is conditionally independent of Earthquake, but not marginally independent of it

**\*This conclusion is independent to values of CPTs!**

# Question

If  $X$  and  $Y$  are independent, are they therefore independent given any variable(s)?

I.e., if  $P(X, Y) = P(X) P(Y)$  [ i.e., if  $P(X|Y) = P(X)$  ], can we conclude that

$$P(X | Y, Z) = P(X | Z)?$$

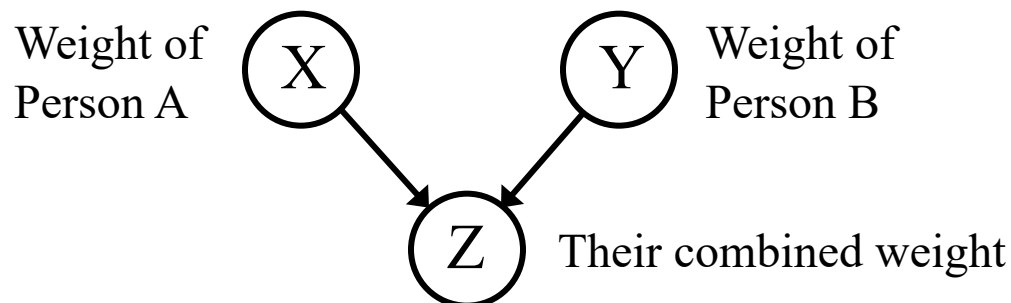
# Question

If  $X$  and  $Y$  are independent, are they therefore independent given any variable(s)?

I.e., if  $P(X, Y) = P(X) P(Y)$  [ i.e., if  $P(X|Y) = P(X)$  ], can we conclude that

$$P(X | Y, Z) = P(X | Z)?$$

The answer is **no**, and here's a counter example:



$$P(X | Y) = P(X)$$
$$P(X | Y, Z) \neq P(X | Z)$$

Note: Even though  $Z$  is a deterministic function of  $X$  and  $Y$ , it is still a random variable with a probability distribution

**\*Again: This conclusion is independent to values of CPTs!**

# Key points of today's lecture

- Representing a joint-distribution
  - number of parameters exponential in the number of variables
  - Calculating marginals and conditionals from the joint-distribution.
- Conditional independences and factorization of joint-distributions
  - Saves parameters, often exponential improvements
- Intro to Bayesian networks / directed graphical models.

## Next lecture

- More Bayesian networks, directed graphical models.
- Read off conditional independences from the graph!
- More examples
- Reading materials:
  - Today's lecture: AIMA Ch. 12.1-12.6
  - Thursday's lecture: AIMA Ch. 13.1 – 13.3 / Jordan PGM Ch. 2.1