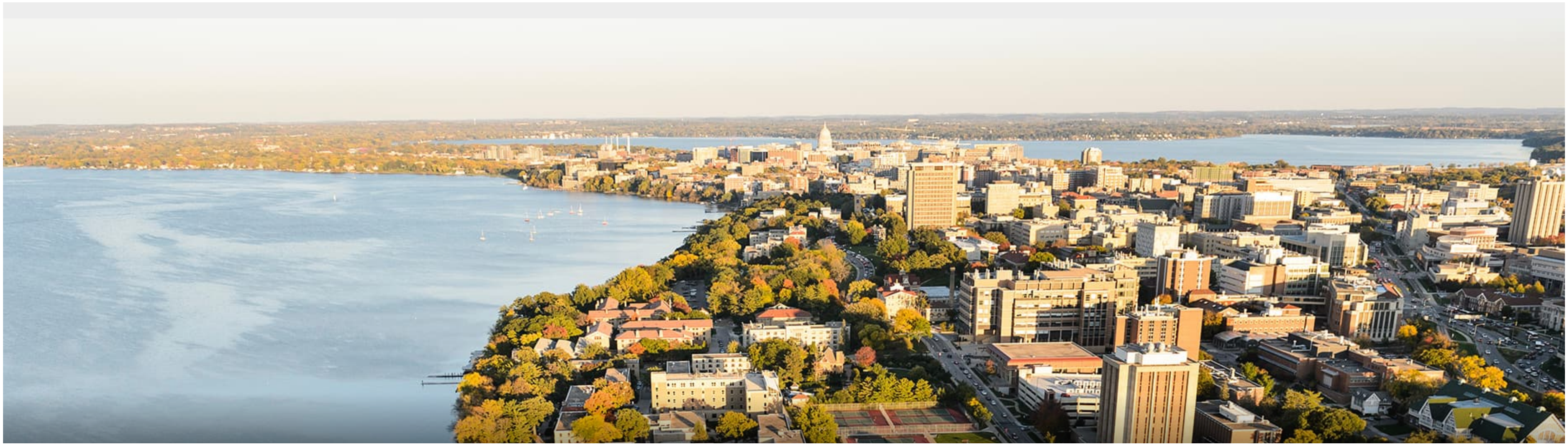


CS 760: Machine Learning

Probability & Graphical Models: Part II

Ilias Diakonikolas

University of Wisconsin-Madison

Nov. 10, 2022

Announcements

- **Logistics**:
  - HW 5 due coming Tuesday.

- Class roadmap:

	Graphical Models
	Less-than-full Supervision
	Unsupervised Learning I
	Unsupervised Learning II

Outline

- **Probability Tutorial**
  - Basics, joint probability, conditional probabilities, etc
- **Bayesian Networks**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families, learning

Outline

- **Probability Tutorial**
  - Basics, joint probability, conditional probabilities, etc
- **Bayesian Networks**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families, learning

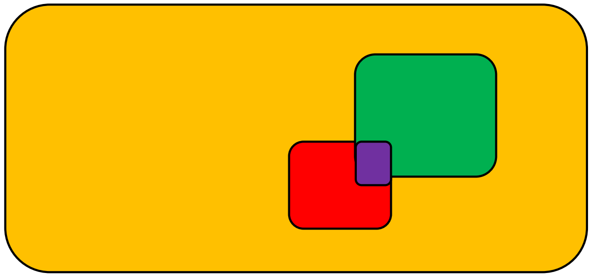
Basics: **Axioms**

- Rules for probability:
  - For all events $E \in \mathcal{F}, P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

- Easy to derive other laws. Ex: non-disjoint events

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

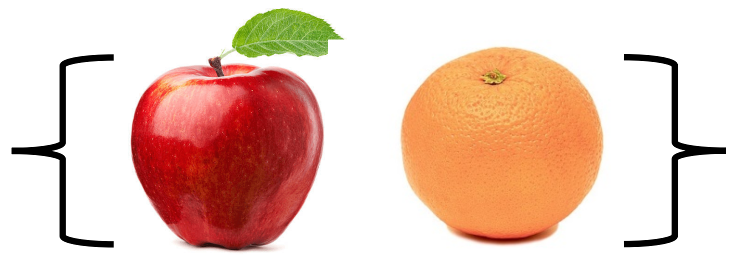


Basics: **Random Variables**

- Really, functions
- Map outcomes to real values

$$X : \Omega \to \mathbb{R}$$

- Why?
  - So far, everything is a set.
  - Hard to work with!
  - Real values are easy to work with



- One requirement, "F measurable". For any *c*,

$$\{\omega : X(\omega) \leq c\} \in \mathcal{F}$$

Basics: **CDF & PDF**

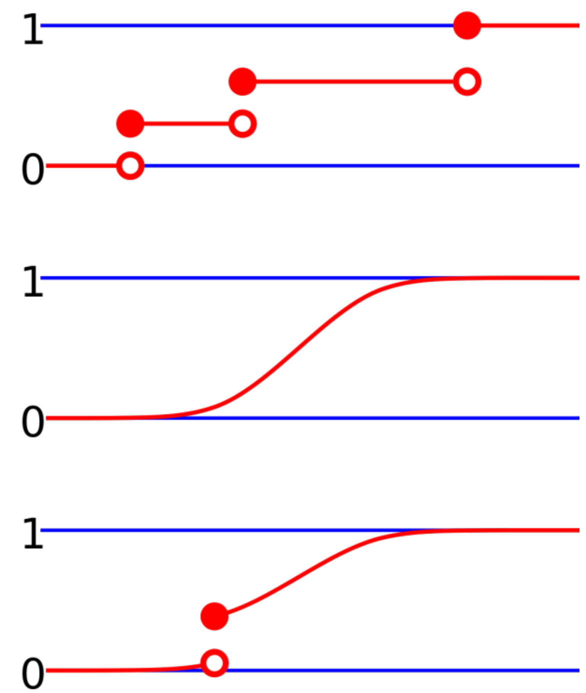
- Can still work with probabilities:

$$P(X = 3) := P(\{\omega : X(\omega) = 3\})$$

- Cumulative Distribution Func. (CDF)

$$F_X(x) := P(X \leq x)$$

- Density / mass function $p_X(x)$
    - Doesn't always exist!



Wiki CDF

Basics: **Expectation** & **Variance**

- Another advantage of RVs are ``summaries''
- Expectation:
  - The "average" $E[X] = \sum_a a \times P(x = a)$
- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of spread
- Raw moments: $E[X], E[X^2], E[X^3], \ldots$
- Note: also don't always exist…
  - **Ex**: Cauchy distribution

Basics: **Expectation** Properties

- Expectation has very useful properties…
  - Linearity: $$E[\sum_i a_i X_i] = \sum_i a_i E[X_i]$$
  - Independence not required!
  - Hat check problem:
    - There is a dinner party where n people check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each person gets their own hat with probability 1/n. What is the expected number of people who get their own hat?

Basics: **Joint Distributions**

- Move from one variable to several
- Joint distribution

$$P(X = a, Y = b)$$

- Or more variables.

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

Basics: **Marginal** Probability

- Given a joint distribution

$$P(X = a, Y = b)$$

  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the "marginal" distribution.

Basics: **Marginal** Probability

$$P(X=a)=\sum_b P(X=a,Y=b)$$

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

$$[P(\text{hot}),P(\text{cold})]=[\tfrac{195}{365},\tfrac{170}{365}]$$

Independence

- Independence for a set of events $A_1, \ldots, A_k$

$$P(A_{i_1} A_{i_2} \cdots A_{i_j}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_j})$$

for all the $i_1, \ldots, i_j$ combinations

- Why useful? Dramatically reduces the complexity
- Collapses joint into **product** of marginals
  - Note sometimes we have only pair-wise, etc independence

Uncorrelatedness

- For random variables, uncorrelated means

$$E[XY] = E[X]E[Y]$$

Note: weaker than independence.

- Independence implies uncorrelated (easy to see)
- Other way around: usually false (but not always).
- If X,Y independent, functions are not correlated:

$$E[f(X)f(Y)] = E[f(X)]E[f(Y)]$$

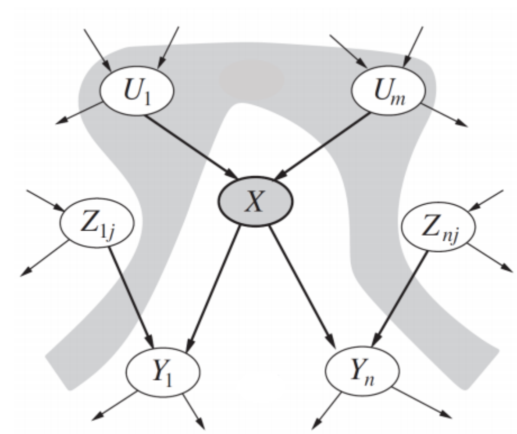
Conditional Probability

- For when we know something,

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

- Leads to **conditional independence**

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



Credit: Devin Soni

Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \dots, A_n)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1)\dots P(A_n|A_{n-1}, \dots, A_1)$$

- Note: still big!
  - If some **conditional independence**, can factor!
  - Leads to **probabilistic graphical models (this lecture)**

Law of Total Probability

- Partition the sample space into disjoint $B_1, \ldots, B_k$
- Then,

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

- Useful way to control A via conditional probabilities.
  - **Example**: there are 5 red and 2 green balls in an urn. A random ball is selected and replaced by a ball of the other color; then a second ball is drawn. What is the probability the second ball is red?

Bayesian Inference

- Conditional Prob. & Bayes:

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1, \ldots, E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- Has more evidence.
  - Likelihood is hard---but **conditional independence assumption**

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H)P(E_2|H)\cdots, P(E_n|H)P(H)}{P(E_1, E_2, \ldots, E_n)}$$

Random Vectors & Covariance

- Recall variance: $\mathbb{E}[(X - E[X])^2]$
- Now, for a **random vector** (same as joint of *d* RVs)
  - Note: size *d x d.* All variables are centered

$$\Sigma = \begin{bmatrix} \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] & \dots & [(X_1 - \mathbb{E}[X_1])((X_n - \mathbb{E}[X_n])] \\ \vdots & \vdots & \vdots \\ [(X_n - \mathbb{E}[X_n])((X_1 - \mathbb{E}[X_1])] & \dots & \mathbb{E}[(X_n - \mathbb{E}[X_n])^2] \end{bmatrix}$$

Cross-variance

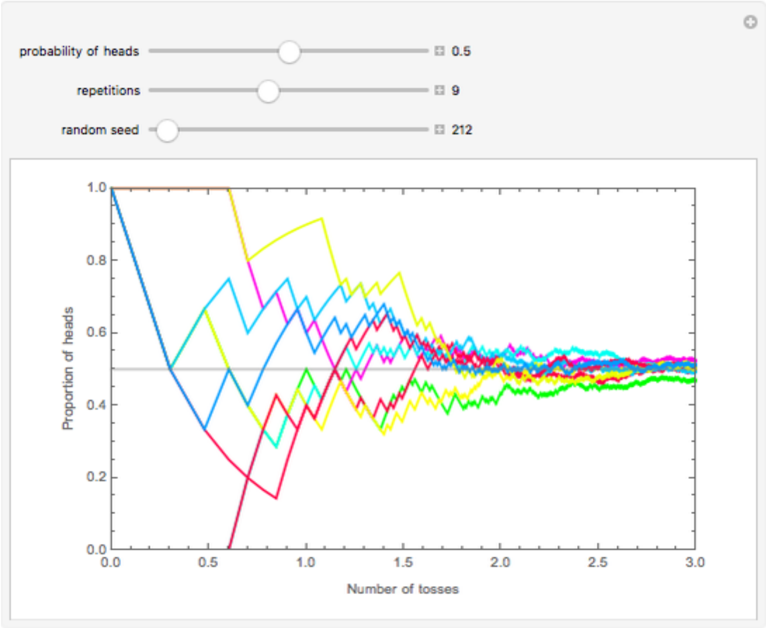
Diagonals: Scalar Variance

Estimation Theory

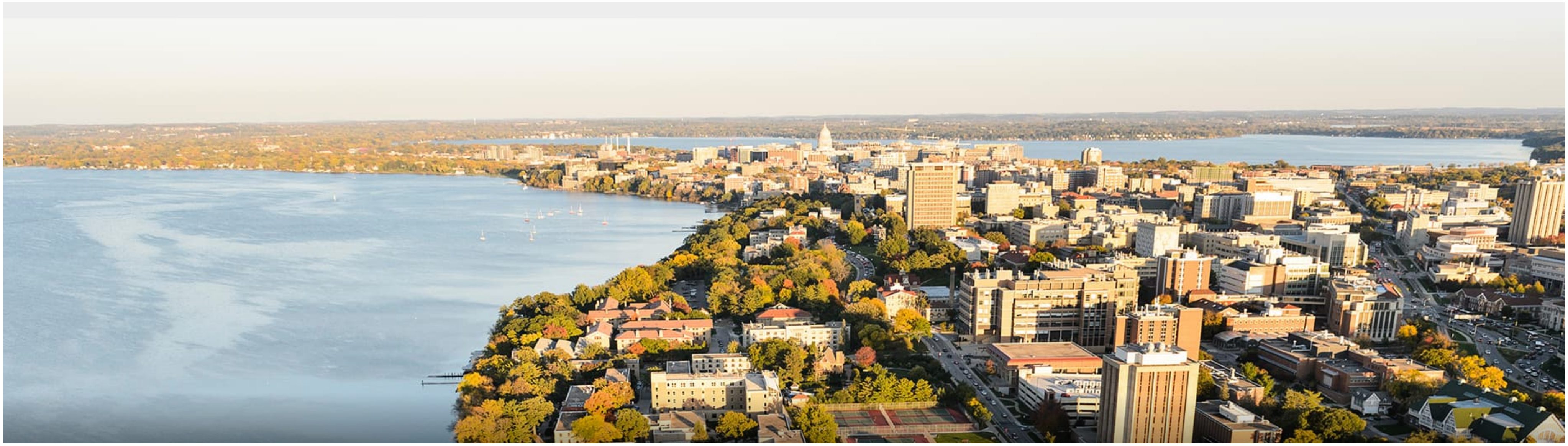
- How do we know that the sample mean is a good estimate of the true mean?
  - Concentration inequalities

$$P(|\mathbb{E}[X] - \hat{\mathbb{E}}[X]| \geq t) \leq \exp(-2nt^2)$$

  - Law of large numbers
  - Central limit theorems, etc.



Wolfram Demo

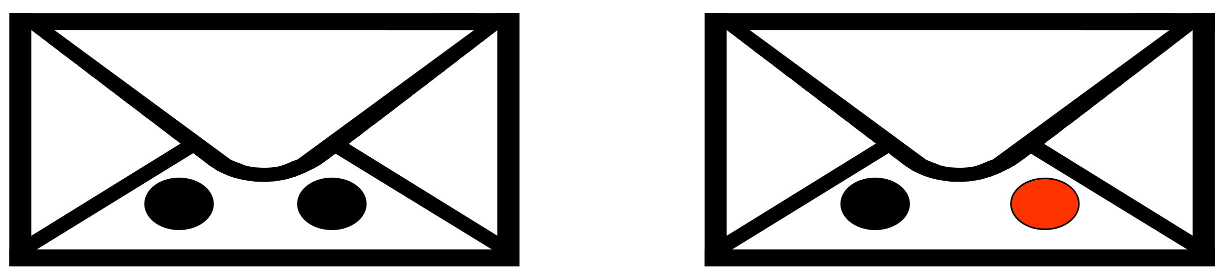


Break & Quiz

Break & Quiz

Q 1-1: We have two envelopes:

- E_1 has two black balls, E_2 has one black, one red
- The **red** one is worth \$100. Others, zero
- Open an envelope, see one ball. Then, can switch (or not).
- You see a black ball. **Switch?**



Break & Quiz

Q 1-1: Solution

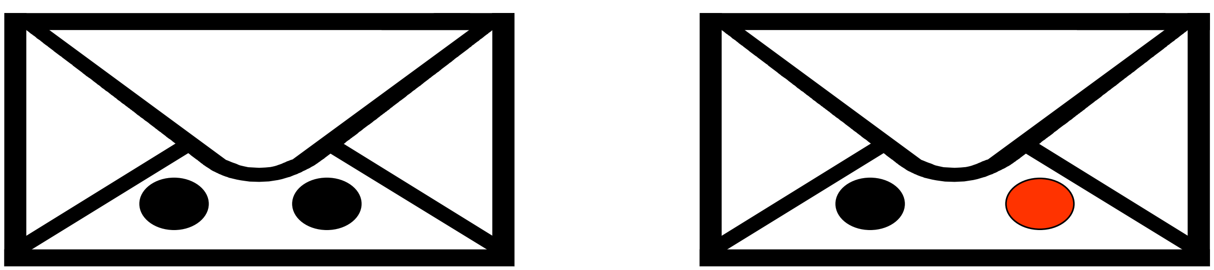
$$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- Now plug in:

$$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!



Break & Quiz

Q 1-2: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. 5/104
B. 95/100
C. 1/100
D. 1/2

Break & Quiz

Q 1-2**:** 50% of emails are spam. Software has been applied to filter spam. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. **5/104**
B. 95/100
C. 1/100
D. 1/2

Outline

- **Probability Tutorial**
  - Basics, joint probability, conditional probabilities, etc
- **Bayesian Networks**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families, learning

Bayesian Networks Example

- Consider the following 5 binary random variables:

  B = a burglary occurs at the house

  E = an earthquake occurs at the house

  A = the alarm goes off

  J = John calls to report the alarm

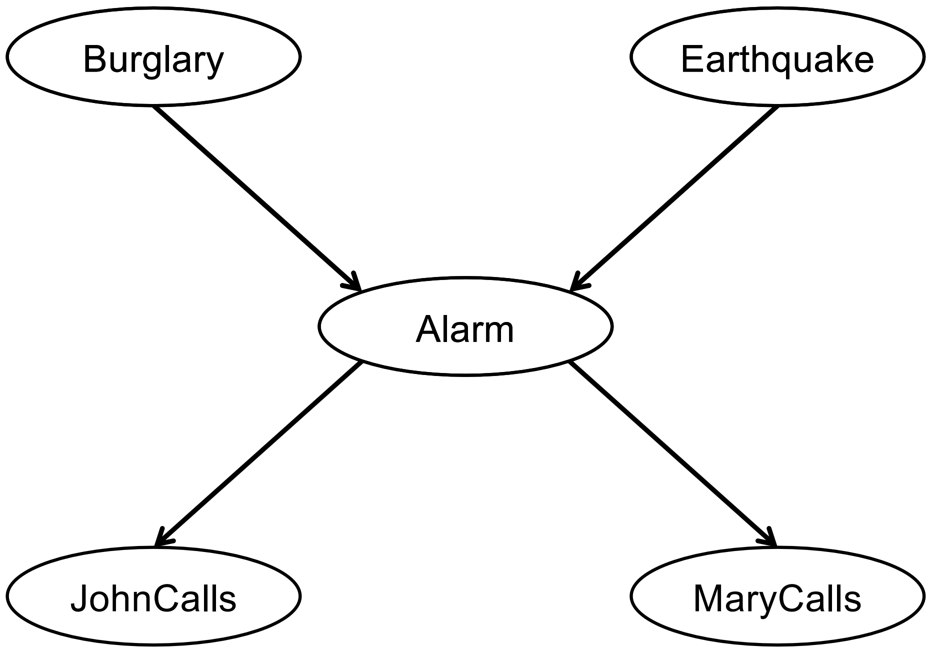
  M = Mary calls to report the alarm

- Suppose Burglary or Earthquake can trigger Alarm, and Alarm can trigger John's call or Mary's call

- Now we want to answer queries like what is $P(B \mid M, J)$?

Bayesian Networks Example

- Set up a network that shows how random variables influence others:

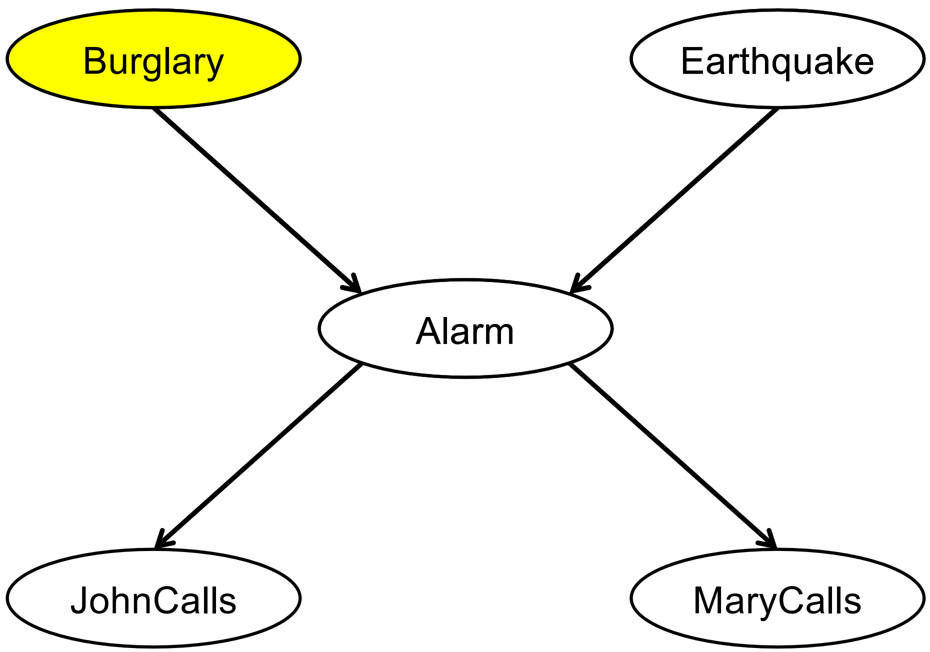


Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (*B*)

t	f
0.001	0.999

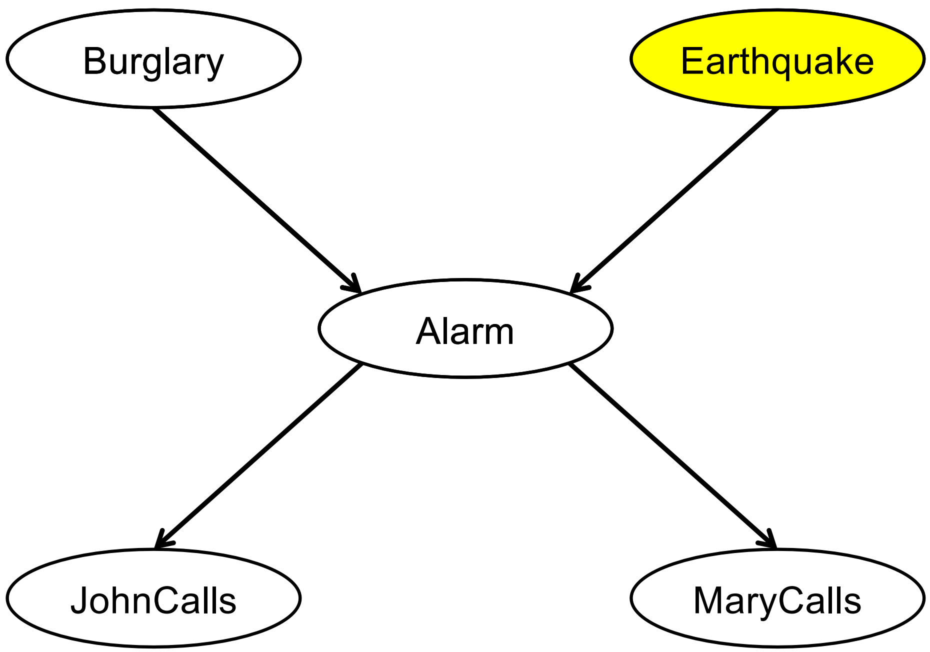


Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (*B*)

t	f
0.001	0.999



P (*E*)

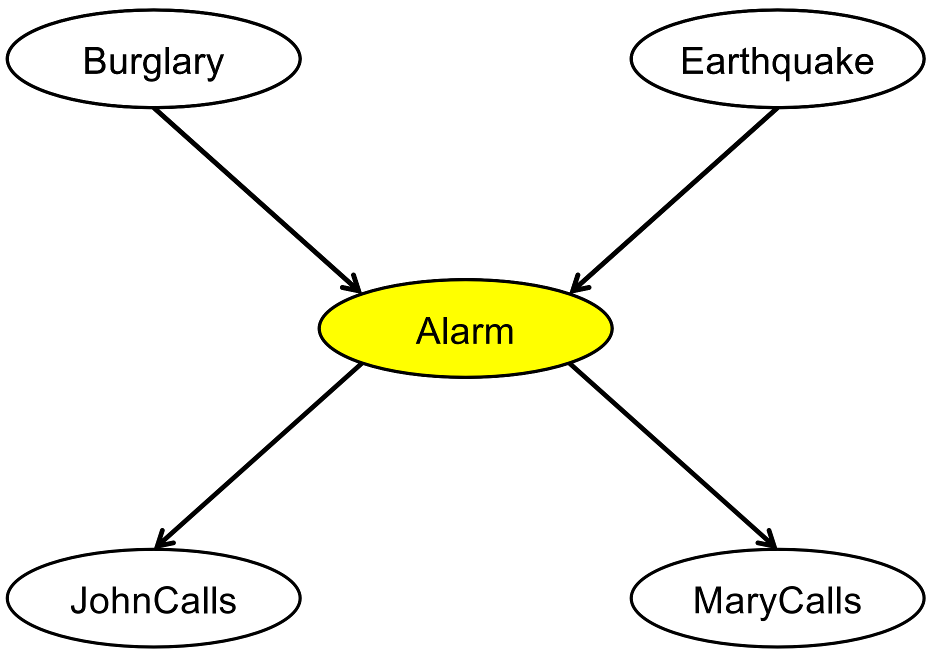
t	f
0.001	0.999

Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (B)

t	f
0.001	0.999



P (E)

t	f
0.001	0.999

P (A | B, E)

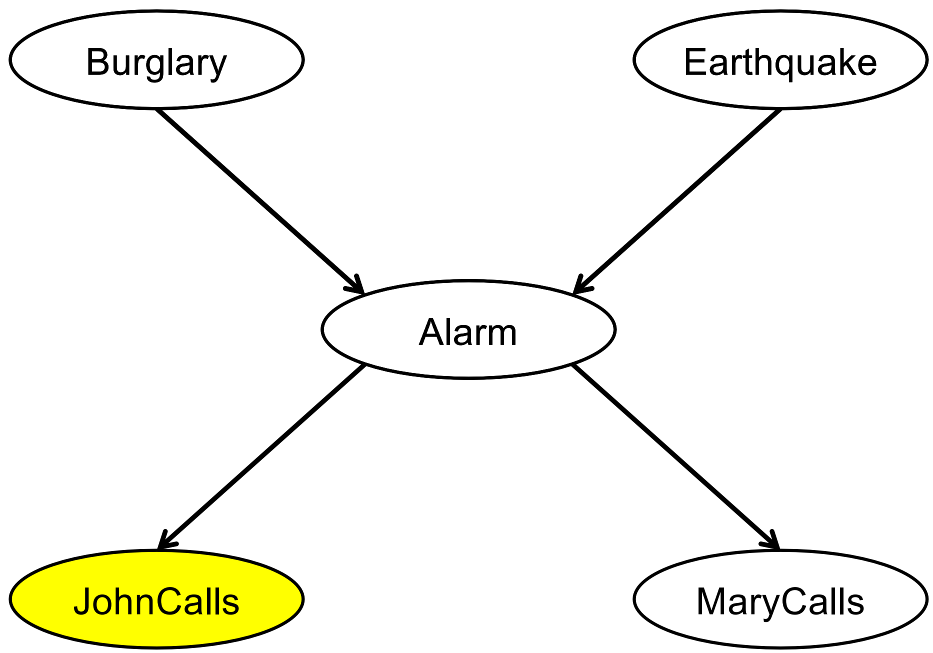
B	*E*	t	f
t	t	0.95	0.05
t	f	0.94	0.06
f	t	0.29	0.71
f	f	0.001	0.999

Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (*B*)

t	f
0.001	0.999



P (*E*)

t	f
0.001	0.999

P (*A* | *B*, *E*)

B	*E*	t	f
t	t	0.95	0.05
t	f	0.94	0.06
f	t	0.29	0.71
f	f	0.001	0.999

P (*J* | *A*)

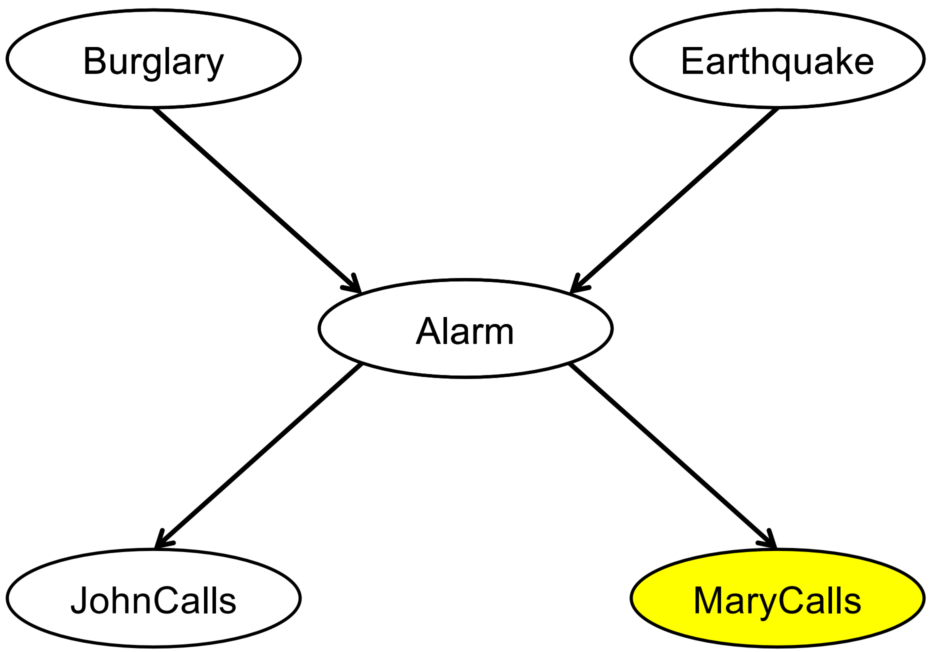
A	t	f
t	0.9	0.1
f	0.05	0.95

Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (B)

t	f
0.001	0.999



P (E)

t	f
0.001	0.999

P (A | B, E)

B	*E*	t	f
t	t	0.95	0.05
t	f	0.94	0.06
f	t	0.29	0.71
f	f	0.001	0.999

P (J | A)

A	t	f
t	0.9	0.1
f	0.05	0.95

P (M | A)

A	t	f
t	0.7	0.3
f	0.01	0.99

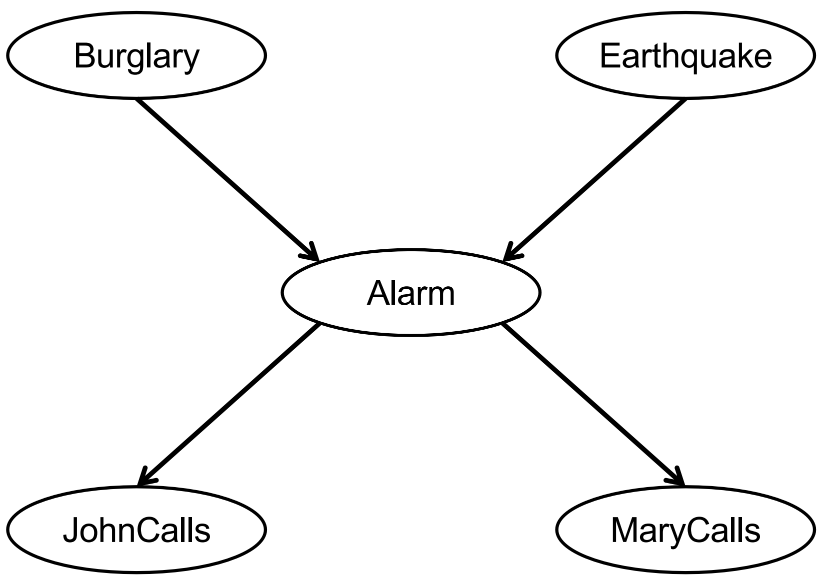
Bayesian Networks: Definition

- A BN consists of a **Directed Acyclic Graph (DAG**) and a set of **conditional probability distribution**s
- The DAG:
  - each node denotes a random variable
  - each edge from X to Y represents that X *directly influences* Y
  - (formally: each variable X is independent of its non-descendants given its parents)
  - **Each CPD: represents** $P(X \mid Parents(X)\ )$

$$p(x_1, \ldots, x_d) = \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)})$$

Bayesian Networks: Parameter Counting

- Parameter reduction: a standard representation of the joint distribution for the Alarm example has $2^5 = 32$ parameters
- the BN representation of this distribution has 20 parameters



$$
\begin{aligned}
&P(B, E, A, J, M) \\
&= P(B) \\
&\times P(E) \\
&\times P(A \mid B, E) \\
&\times P(J \mid A) \\
&\times P(M \mid A)
\end{aligned}
$$

Inference in Bayesian Networks

Given: values for some variables in the network (*evidence*), and a set of ***query*** variables

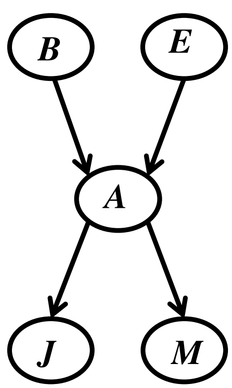
Do: compute the posterior distribution over the query variables

- Variables that are neither evidence variables nor query variables are ***hidden*** variables
- The BN representation is flexible enough that any set can be the evidence variables and any set can be the query variables

Inference by Enumeration

- Let a denote $\boldsymbol{A}$=true, and $\neg a$ denote $\boldsymbol{A}$=false
- Suppose we're given the query: $P(b \mid j, m)$

  "probability the house is being burglarized given that John and Mary both called"
- From the graph structure we can first compute:



$$P(b, j, m) = \sum_{e,\neg e} \sum_{a,\neg a} P(b)P(E)P(A \mid b, E)P(j \mid A)P(m \mid A)$$

sum over possible values for E and A variables (e, $\neg e$, a, $\neg a$)

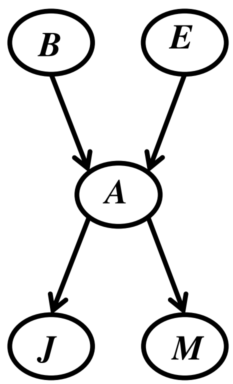
Inference by Enumeration

$$P(b, j, m) = \sum_{e,\neg e} \sum_{a,\neg a} P(b)P(E)P(A \mid b, E)P(j \mid A)P(m \mid A)$$

$$= P(b) \sum_{e,\neg e} \sum_{a,\neg a} P(E)P(A \mid b, E)P(j \mid A)P(m \mid A)$$

P(B)
0.001

P(E)
0.001



B	E	P(A)
t	t	0.95
t	f	0.94
f	t	0.29
f	f	0.001

A	P(J)
t	0.9
f	0.05

A	P(M)
t	0.7
f	0.01

$B \quad E \quad A \quad J \quad M$

$$= 0.001 \times (0.001 \times 0.95 \times 0.9 \times 0.7 + \qquad e, a$$

$$0.001 \times 0.05 \times 0.05 \times 0.01 + \qquad e, \neg a$$

$$0.999 \times 0.94 \times 0.9 \times 0.7 + \qquad \neg e, a$$

$$0.999 \times 0.06 \times 0.05 \times 0.01) \qquad \neg e, \neg a$$

Inference by Enumeration

- Next do equivalent calculation for $P(\neg b,\ j, m)$ and determine $P(b \mid j, m)$

$$P(b \mid j,m) = \frac{P(b, j,m)}{P(j,m)} = \frac{P(b, j,m)}{P(b, j,m) + P(\neg b, j,m)}$$

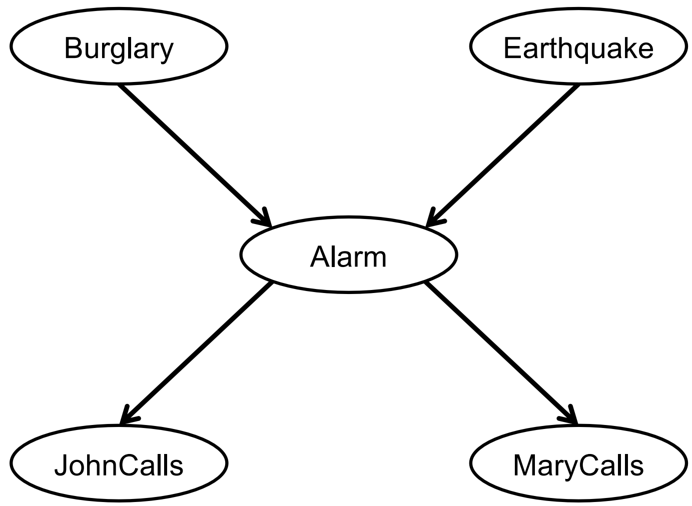
So: exact method, but can be intractably hard.

- Some cases: efficient
- Approximate inference sometimes available

Learning Bayes Nets

- **Problem 1 (parameter learning)**: given a set of training instances, the graph structure of a BN

B	E	A	J	M
f	f	f	t	f
f	t	f	f	f
f	f	t	f	t
		…		



- **Goal**: infer the parameters of the CPDs

Learning Bayes Nets

- **Problem 2 (structure learning)**: given a set of training instances

B	E	A	J	M
f	f	f	t	f
f	t	f	f	f
f	f	t	f	t
		…		

- **Goal**: infer the graph structure (and then possibly also the parameters of the CPDs)

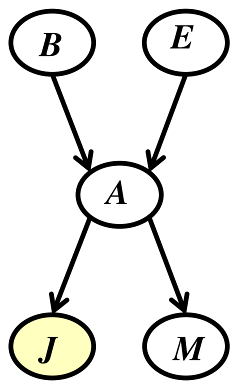
Parameter Learning: MLE

- **Goal**: infer the parameters of the CPDs
- As usual, can use MLE

$$
\begin{aligned}
L(\theta : D, G) = P(D \mid G, \theta) &= \prod_{d \in D} P(x_1^{(d)}, x_2^{(d)}, \ldots, x_n^{(d)}) \\
&= \prod_{d \in D} \prod_i P(x_i^{(d)} \mid Parents(x_i^{(d)})) \\
&= \prod_i \underbrace{\left(\prod_{d \in D} P(x_i^{(d)} \mid Parents(x_i^{(d)})) \right)}_{\text{independent parameter learning problem for each CPD}}
\end{aligned}
$$

Parameter Learning: MLE Example

- **Goal**: infer the parameters of the CPDs
- Consider estimating the CPD parameters for B and J in the alarm network given the following data set



B	E	A	J	M
f	f	f	t	f
f	t	f	f	f
f	f	f	t	t
t	f	f	f	t
f	f	t	t	f
f	f	t	f	t
f	f	t	t	t
f	f	t	t	t

$$P(b) = \frac{1}{8} = 0.125$$

$$P(\neg b) = \frac{7}{8} = 0.875$$

$$P(j \mid a) = \frac{3}{4} = 0.75$$

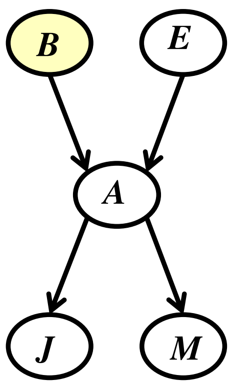
$$P(\neg j \mid a) = \frac{1}{4} = 0.25$$

$$P(j \mid \neg a) = \frac{2}{4} = 0.5$$

$$P(\neg j \mid \neg a) = \frac{2}{4} = 0.5$$

Parameter Learning: MLE Example

- **Goal**: infer the parameters of the CPDs
- Consider estimating the CPD parameters for B and J in the alarm network given the following data set



B	E	A	J	M
f	f	f	t	f
f	t	f	f	f
f	f	f	t	t
t	f	f	f	t
f	f	t	t	f
f	f	t	f	t
f	f	t	t	t
f	f	t	t	t

$$P(b) = \frac{0}{8} = 0$$

$$P(\neg b) = \frac{8}{8} = 1$$

do we really want to set this to 0?

Parameter Learning: Laplace Smoothing

- Instead of estimating parameters strictly from the data, we could start with some prior belief for each
- For example, we could use *Laplace estimates*

$$P(X = x) = \frac{\overbrace{n_x + 1}^{}}{\sum_{v \in \text{Values}(X)} (n_v + 1)}$$

pseudocounts

where n_v represents the number of occurrences of value v

- Recall: we did this for Naïve Bayes

Structure Learning

- Generally a hard problem, many approaches.
  - Exponentially (or worse) many structures in # variables
  - Can either use heuristics or restrict to some tractable subset of networks. Ex: **trees**

- Chow-Liu Algorithm
  - Learns a BN with a tree structure that maximizes the likelihood of the training data

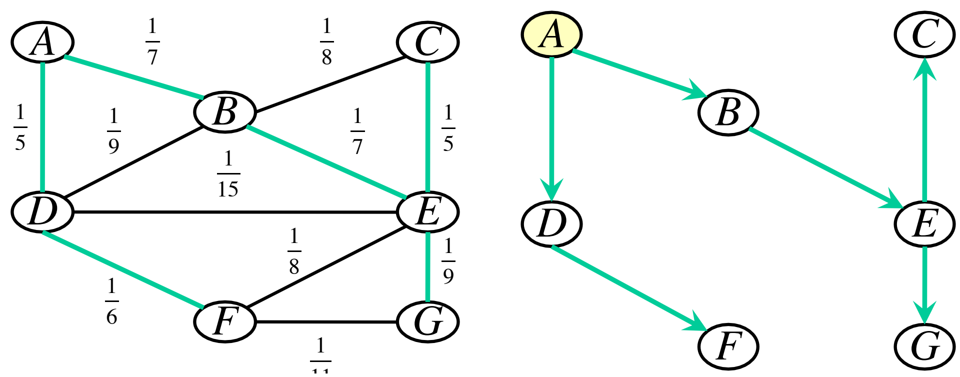
1. Compute weight $I(X_i, X_j)$ of each possible edge (X_i, X_j)
2. Find maximum weight spanning tree (MST)
3. Assign edge directions in MST

Structure Learning: Chow-Liu Algorithm

Chow-Liu Algorithm

1. Compute weight $I(X_i, X_j)$ of each possible edge (X_i, X_j)
2. Find maximum weight spanning tree (MST)
3. Assign edge directions in MST

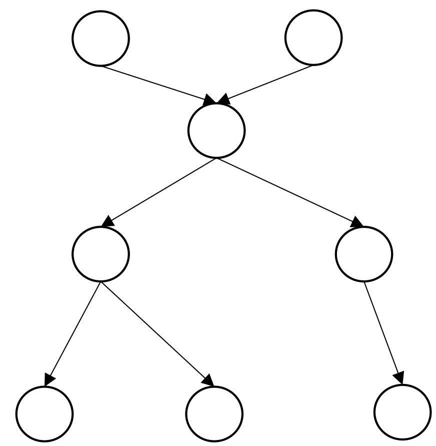
- 1. Empirical mutual information: $O(n^2)$ computations
- 2. Compute MST. (Ex: Kruskal's algorithm)
- 3. Assign directions by picking a root and making everything directed from root



Break & Quiz

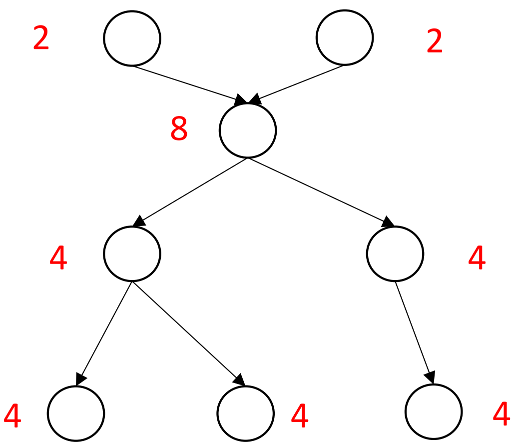
Q2-1: Consider a case with 8 binary random variables, how many parameters does a BN with the following graph structure have?

1. 24
2. 28
3. 32
4. 52



Q2-1: Consider a case with 8 binary random variables, how many parameters does a BN with the following graph structure have?

1. 24
2. 28
3. 32
4. 52



So we have 32 parameters in total.

Outline

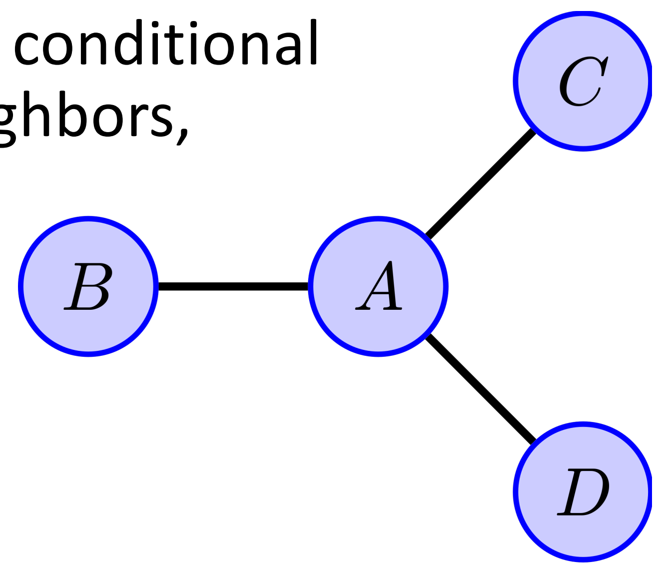
- **Probability Tutorial**
  - Basics, joint probability, conditional probabilities, etc
- **Bayesian Networks**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families, learning

Undirected Graphical Models

- Still want to encode conditional independence, but not in an "ordered" way (ie, no parents, direction)
  - **Why**? Allows for modeling other distributions that Bayes nets can't, allows for other algorithms

- Idea: graph directly encodes a type of conditional independence. If nodes i,j are not neighbors,

$$X_i \perp X_j | X_{V \setminus \{i,j\}}$$

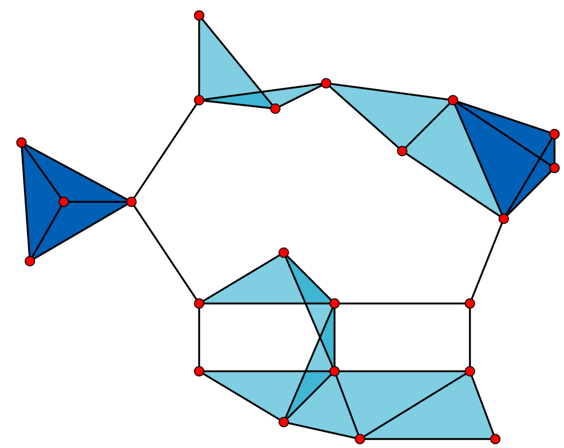


Markov Random Fields

- A particularly popular kind of undirected model. As above, can describe in terms of:
  - 1. Conditional independence: $X_i \perp X_j | X_{V \setminus \{i,j\}}$
  - 2. Factorization. (Clique: maximal fully-connected subgraphs)
    - Bayes nets: factorize over CPTs with **parents**; MRFs: factorize over **cliques**

$$P(X) = \prod_{C \in \text{cliques}(G)} \phi_C(x_C)$$

"**Potential**" functions



Exponential Families

- MRFs (under some conditions) can be written as exponential families. General form:

$$P(x_1, \ldots, x_d) = \frac{1}{Z} \exp(\sum_i \theta_i^T f_i(x_{\{i\}}))$$

Partition function
(ensures that probabilities integrate to 1)

Sufficient statistics

- Lots (but not all) distributions have this form.

Exponential Families: Multivariate Gaussian

- MRFs (under some conditions) can be written as exponential families. General form:

$$P(x_1, \ldots, x_d) = \frac{1}{Z} \exp(\sum_i \theta_i^T f_i(x_{\{i\}}))$$

- Multivariate Gaussian:

$$\frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$\frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} \sum_{i,j} K_{i,j} (x_i - \mu_i)(x_j - \mu_j)\right)$$

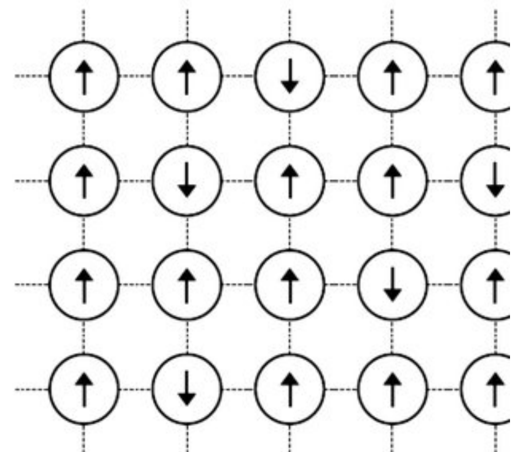
Partition function

Inverse Covariance Matrix

Ising Models

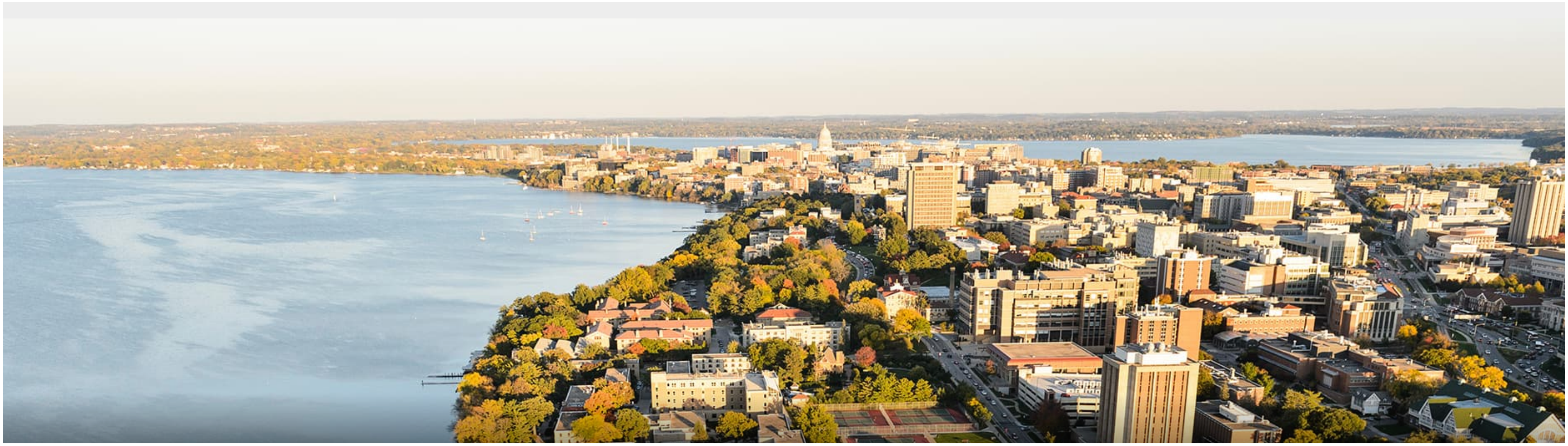
- Ising models: a particular kind of MRF usually written in exponential form
  - Popular in statistical physics
  - **Idea**: pairwise interactions (biggest cliques of size 2)

$$P(x_1, \ldots, x_d) = \frac{1}{Z} \exp\Big(\sum_{(i,j) \in E} \theta_{ij} x_i x_j \Big)$$



Khudier and Fawaz

- Challenges:
  - Compute partition function
  - Perform inference/marginalization



Thanks Everyone!

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, Pedro Domingos, Jerry Zhu, Yingyu Liang, Volodymyr Kuleshov, Fei-Fei Li, Justin Johnson, Serena Yeung, Pieter Abbeel, Peter Chen, Jonathan Ho, Aravind Srinivas, Fred Sala