Announcements

- HW 5 due **Wednesday, November 16**
- Quiz 10 is due **Thursday, November 17 at 8pm**

Lecture 20: Bayesian Networks

CIS 4190/5190

Fall 2022

Class So Far

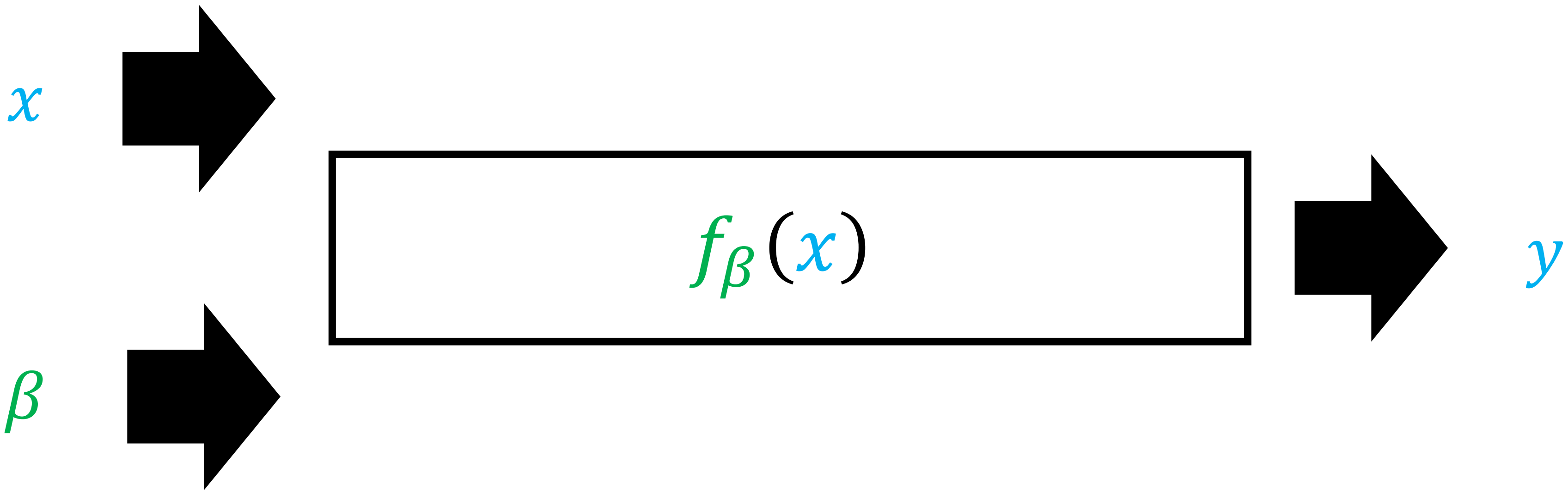
- **Supervised Learning**
  - Linear/logistic regression, MLE, decision trees, ensembles, neural networks
  - Application to computer vision, NLP
- **Unsupervised Learning**
  - PCA, K-Means, neural networks
  - Application to NLP
- **Today:** Bayesian networks
  - Very different viewpoint, but concepts are pervasive in ML research
  - Probability as a unifying framework for machine learning

Design Decisions

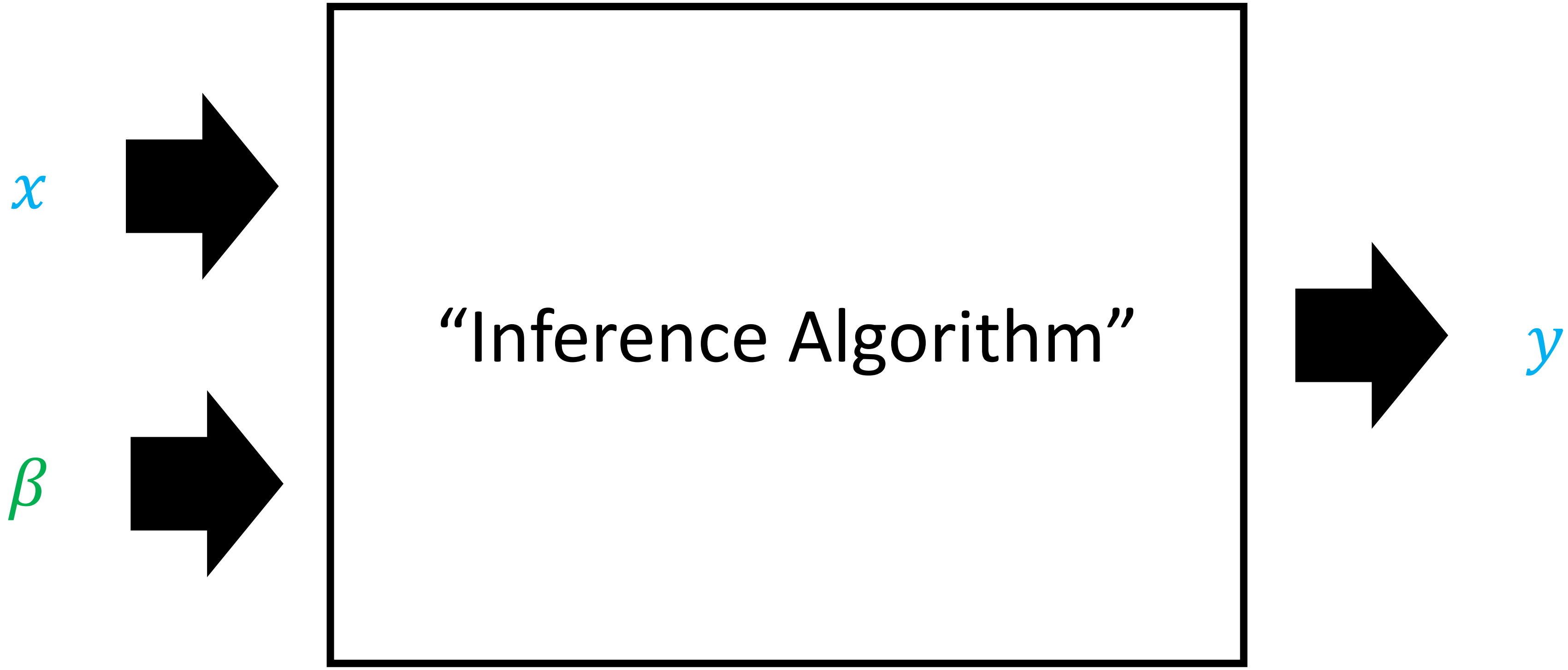
- **Model family**
  - Flexible architectures
  - **Implicit** functions (**inference**)
  - Very different from what we've seen so far!

- **Optimization algorithm**
  - Typically straightforward

Models So Far



Bayesian Network Inference



Logic & AI

- Efficient algorithm for logical reasoning was a major focus of early research on artificial intelligence

- **Logical inference problem**
  - Given a set of “facts”, is a given statement true or false?
  - “Facts” can be formalized as a set of logical formulas

- **Example:**
  - **Facts:** All men are mortal. Socrates is a man.
  - **Question:** Is Socrates mortal?
  - **Answer:** Yes!

Logic & AI

- Pure logic is very limited compared to human reasoning

- **Example (McDermott 1987):**
  - **Facts:** There is an empty can of soda
  - **Question:** Did someone drink soda?
  - **Answer:** Probably!

- **Issues**
  - Consider the facts “only people drink soda”, “soda cans start out full”
  - These facts often have many exceptions that can typically be ignored

Probabilistic Inference

- **Solution:** Probabilistic inference
  - **Input:** Facts that hold with some probability, desired query
  - **Output:** Probability of query holding

- Use simplified facts but account for the fact that they may be wrong

Probabilistic Inference

- **Probabilistic models**
  - Probability distribution designed to describe how portion of the world works

$$P(X_1 = x_1, \ldots, X_n = x_n)$$

  - Encode world as set of random variables and their relationships

- Always simplifications (e.g., may not account for every variable, or all dependences between variables)

- **Example: "**Drinking can of soda" and "can being empty/full"

Probabilistic Inference

- **Probabilistic inference:** Compute distribution of unobserved variables
  - **Example:** Explanation (i.e., observe empty soda can, infer someone drank it)
  - **Example:** Prediction (i.e., observe soda can purchase, infer they will drink it)

- **Problem:** Probabilistic inference is computationally challenging!
  - We won't address the question of where the facts come from (huge literature on inducing knowledge graphs that aims to solve this problem)

Bayesian Networks

- **Bayesian networks** (Pearl 1985) are a graph-based data structure for representing probability distributions

- Expose structure in the form of dependences between variables that can make probabilistic inference more tractable

- As with neural networks, you can design the model family!
  - Widely used in computer vision and NLP prior to success of deep learning
  - Incorporated into modern neural network architectures (e.g., VAEs)

Bayesian Networks

- **Logic:** Inference is checking if a fact can be deduced from given facts

- **Bayesian network:** Inference is evaluating the probability of a fact given the probabilities of other facts

Random Variables

- A **random variable** represents a quantity we are uncertain about

- Random variable takes values in a **domain**
  - We will focus on random variables with finite domains

- **Examples:**
  - $R = \text{Is it raining?}$ ($R \in \{\text{true}, \text{false}\}$, which we may write as $\{+r, -r\}$)
  - $T = \text{Is it hot or cold?}$ ($T \in \{\text{hot}, \text{cold}\}$)
  - $D = \text{How long will it take to drive to work?}$ ($D \in [0, \infty)$)

Probability Distributions

- **Probability distribution:** For random variable X, $P(X = x) \in [0,1]$ is probability X has value x

- **Recall:** Probabilities satisfy $P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

- **Notation:** When unambiguous, we drop the random variable

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$

Probability Distributions

- For finite domains, the distribution can be represented as a table

- **Examples:**

T	$P(T)$
hot	0.5
cold	0.5

W	$P(W)$
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Joint Distributions

- Given random variables $X_1, \dots, X_n$, they have a **joint distribution**

$$P(X_1 = x_1, \dots, X_n = x_n)$$

- As before, satisfy
  - $P(X_1 = x_1, \dots, X_n = x_n) \geq 0$
  - $\sum_{(x_1, \dots, x_n)} P(X_1 = x_1, \dots, X_n = x_n) = 1$

Joint Distributions

- For finite domains, the distribution can be represented as a table

- **Example:**

T	R	$P(T, R)$
hot	no rain	0.4
hot	rain	0.1
cold	no rain	0.2
cold	rain	0.3

Designing a Probabilistic Model

- **Naïve idea**
  - Write down the full joint distribution $P(x_1, \ldots, x_n)$
  - Perform inference using this distribution

- **Problem:** For n random variables with domain size $|D| = d$, the table representing the joint distribution has d^n entries!
  - Learning and inference are both intractable!

- Is there structure we can exploit to improve tractability?
  - Yes, **conditional independence**!

Independence

- Two random variables are **independent** (denoted $X \perp\!\!\!\perp Y$) if

$$\forall x, y \,.\; P(x, y) = P(x)P(y)$$

- Here, $P(x) = \sum_y P(x, y)$ is the **marginal distribution**

- That is, the joint distribution factors into two simpler distributions

Independence

- **Example (not independent):**

T	R	$P(T,R)$
hot	no rain	0.4
hot	rain	0.1
cold	no rain	0.2
cold	rain	0.3

Independence

- **Example (independent):**

T	R	$P(T,R)$
hot	no rain	0.3
hot	rain	0.2
cold	no rain	0.3
cold	rain	0.2

$=$

T	$P(T)$
hot	0.5
cold	0.5

$\times$

R	$P(R)$
no rain	0.6
rain	0.4

Independence

- **Example:** Coin flips

X_1	$P(X_1)$
heads	0.5
tails	0.5

$\times \quad \cdots \quad \times$

X_n	$P(X_n)$
heads	0.5
tails	0.5

$=$

X_1	$\cdots$	X_n	$P(X_1, \dots, X_n)$
heads	$\cdots$	heads	2^{-n}
$\cdots$	$\cdots$	$\cdots$	2^{-n}

2^n rows

Independence can lead to much more compact representations!

Conditional Probabilities

- **Conditional probability:**

$$P(\, x \mid y \,) = \frac{P(x, y)}{P(y)}$$

- **Product rule:** $P(x, y) = P(\, x \mid y \,)P(y)$

- **Chain rule:** $P(x_1, \ldots, x_n) = P(x_1)P(\, x_2 \mid x_1 \,) \cdots P(\, x_n \mid x_1, \ldots, x_{n-1} \,)$

Conditional Probabilities

- **Conditional probability:**

$$P(\,x \mid y\,) = \frac{P(x, y)}{P(y)}$$

- **Product rule:** $P(x, y) = P(\,x \mid y\,)P(y)$

- **Chain rule:** $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(\,x_i \mid x_1, \ldots, x_{i-1}\,)$

- **Note:** Independence is equivalently $\forall x, y\,.\; P(y \mid x) = P(y)$

Conditional Independence

- Independence **conditioned** on other random variables

- **Example:** $P(\text{rain}, \text{traffic}, \text{umbrella})$
  - "Having traffic" and "needing an umbrella" are **not independent**!
  - But if we know there is rain, traffic does not depend on umbrella:

$$P(+\text{traffic}|+\text{rain}, +\text{umbrella}) = P(+\text{traffic}|+\text{rain})$$

  - Similarly for not having rain:

$$P(+\text{traffic}|-\text{rain}, +\text{umbrella}) = P(+\text{traffic}|-\text{rain})$$

Conditional Independence

- Traffic is **conditionally independent** of umbrella given rain

$$P(\text{traffic}|\text{rain}, \text{umbrella}) = P(\text{traffic}|\text{rain})$$

- The following statements are equivalent to the one above:
  - $P(\text{umbrella}|\text{rain}, \text{traffic}) = P(\text{umbrella}|\text{rain})$
  - $P(\text{traffic}, \text{umbrella}|\text{rain}) = P(\text{traffic}|\text{rain})P(\text{umbrella}|\text{rain})$

- Traffic and umbrella are **conditionally independent** given rain

Conditional Independence

- X is **conditionally independent** of Y given $Z_1, \dots, Z_n$ if

$$\forall x, y, z_1, \dots, z_n \,.\, P(x, y | z_1, \dots, z_n) = P(x | z_1, \dots, z_n) P(y | z_1, \dots, z_n)$$

- Equivalently:

$$\forall x, y, z_1, \dots, z_n \,.\, P(x | z_1, \dots, z_n, y) = P(x | z_1, \dots, z_n)$$

- Denoted $X \perp\!\!\!\perp Y \mid Z_1, \dots, Z_n$

Designing a Probabilistic Model

- **Idea:** Restrict to joint distributions with given independence relations
  - Posit set of conditional independence relationships $X_i \perp\!\!\!\perp X_j \mid \{X_k\}$
  - Only learn joint distributions $P(x_1, \ldots, x_n)$ that satisfy these relationships
  - **Intuition:** Conditional independences define "local" distributions that are chained together to form "global" distribution

- This is the approach taken by **Bayesian networks**
  - **Note on terminology:** Special kind of **graphical model**

- Rarely have exact independence, but useful modeling assumption

Bayesian Networks

- Represent conditional independences via a directed acyclic graph
- **Nodes/vertices:** Variables $\{(X_k, D_k)\}$ (and their domains)
- **Arcs/edges:** Encode parameter structure
  - **Parameters:** Distribution of each X_i given its parents

Example: Coin Flips

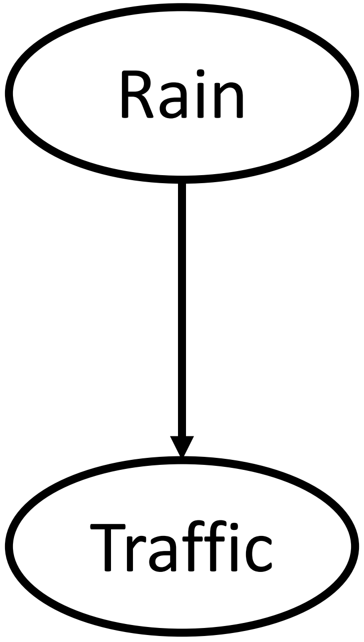
X_1	$P(X_1)$
heads	0.5
tails	0.5

X_n	$P(X_n)$
heads	0.5
tails	0.5

X_1 $\cdots$ X_n

no interactions → all random variables are independent

Example: Weather



Parameters

- Conditional probabilities of node given parents:

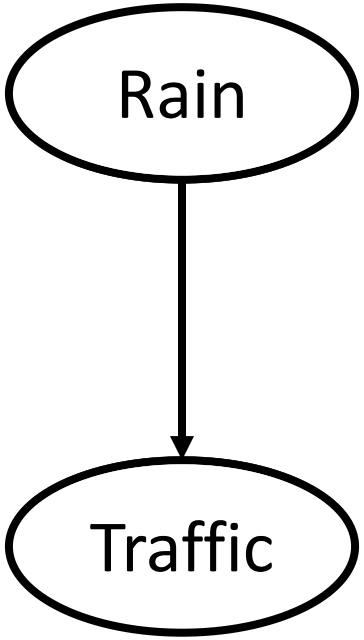
$$\theta_{i,x_1,\ldots,x_{k_i},x} = P\left(X_i = x \mid X_{i_1} = x_1, \ldots, X_{i_k} = x_{k_i} \right)$$

- Here, $x_i \in D_i$ is in the domain of X_i

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4

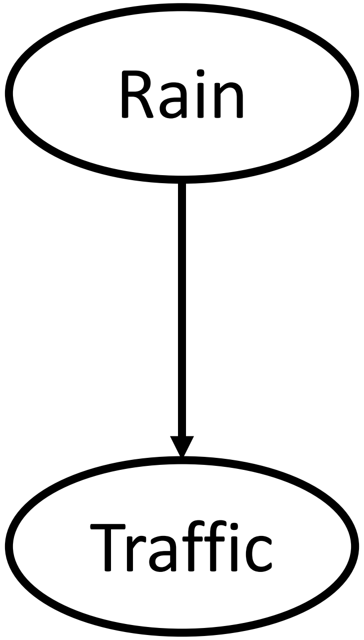
R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75



Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4

R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

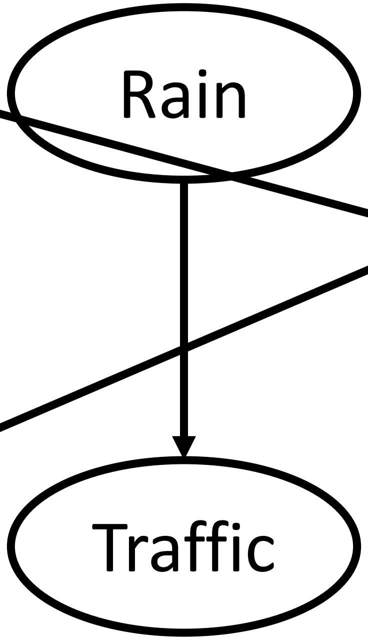


R	T	$P(R,T)$

$$P(R,T) = P(T \mid R)P(R)$$

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4



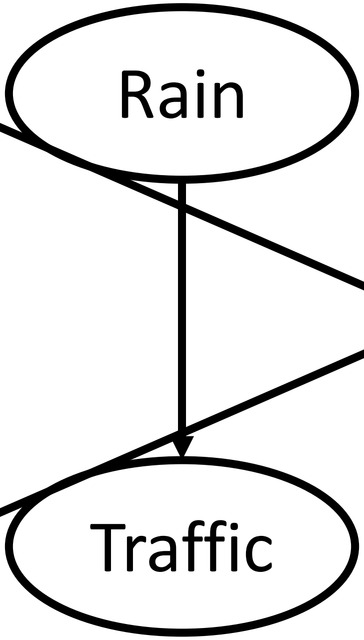
R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R,T)$
no rain	no traffic	0.45

$$P(R,T) = P(T \mid R)P(R)$$

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4



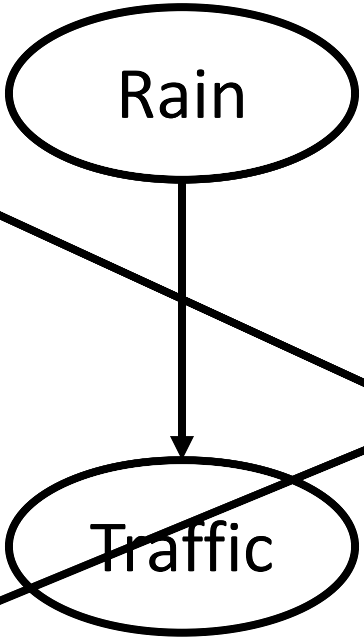
R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R, T)$
no rain	no traffic	0.45
no rain	traffic	0.15

$$P(R, T) = P(T \mid R)P(R)$$

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4



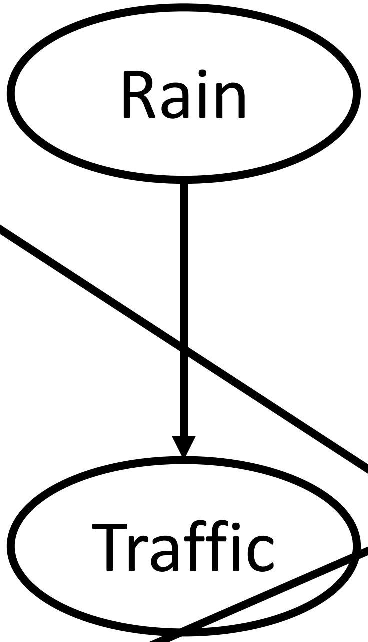
R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R,T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1

$$P(R,T) = P(T \mid R)P(R)$$

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4



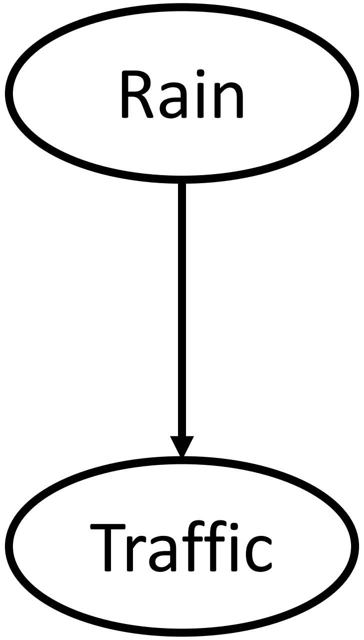
R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R,T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

$$P(R,T) = P(T \mid R)P(R)$$

Example: Weather

R	$P(R)$
no rain	0.6
rain	0.4



R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R, T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

$$P(R, T) = P(T \mid R)P(R)$$

Summary

- **Bayesian network**
  - Nodes represent random variables
  - Edges encode conditional independences
  - For each node, parameters at that node encode probability distribution of node conditioned on its parents

- **Edge directions**
  - Determines parameters
  - Often encode intuitive notion of causality (can be formalized)

Summary

- Any joint distribution satisfying the conditional independencies can be expressed as product of $P\big(\, X_i = x_i \mid \mathrm{parents}(X_i) = \big(x_{i_1}, \dots, x_{i_k}\big)\,\big)$

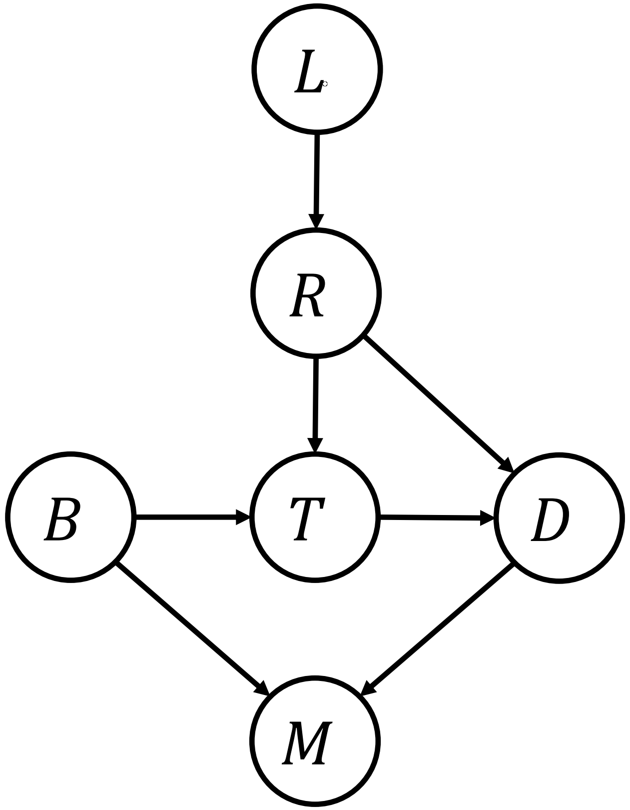
- We can compute the corresponding joint distribution using chain rule:

$$
\begin{aligned}
P(x_1, \dots, x_n) &= \textstyle\prod_{i=1}^{n} P(\, X_i = x_i \mid (X_1, \dots X_{i-1}) = (x_1, \dots, x_{i-1})\,) \\
&= \textstyle\prod_{i=1}^{n} P\big(\, X_i = x_i \mid \mathrm{parents}(X_i) = \big(x_{i_1}, \dots, x_{i_k}\big)\,\big)
\end{aligned}
$$

  - First equality holds for any distribution by chain rule
  - Second equality holds by assumption (assumes topological order)

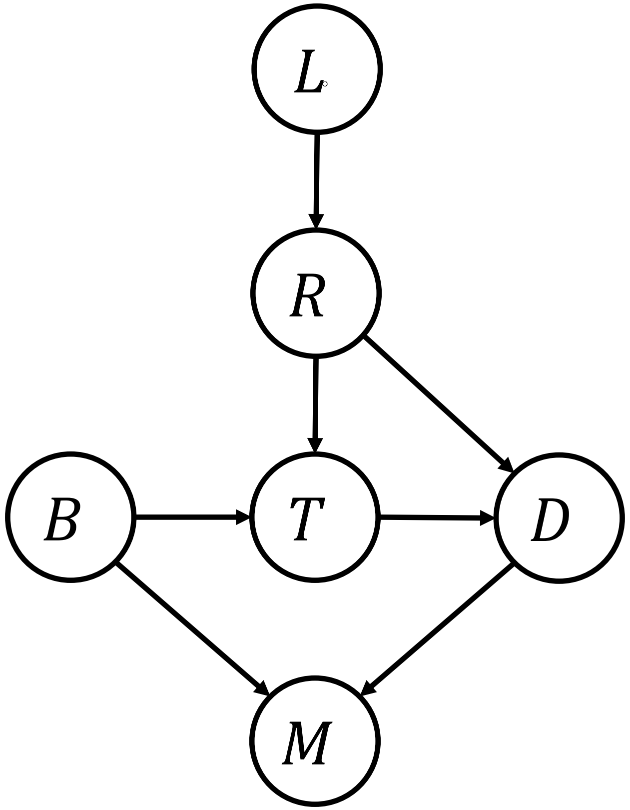
Example: More Complex Traffic Model

- **Variables:**
  - Low pressure (L)
  - Rain (R)
  - Traffic (T)
  - Roof damage (D)
  - Ballgame (B)
  - Mood (M)

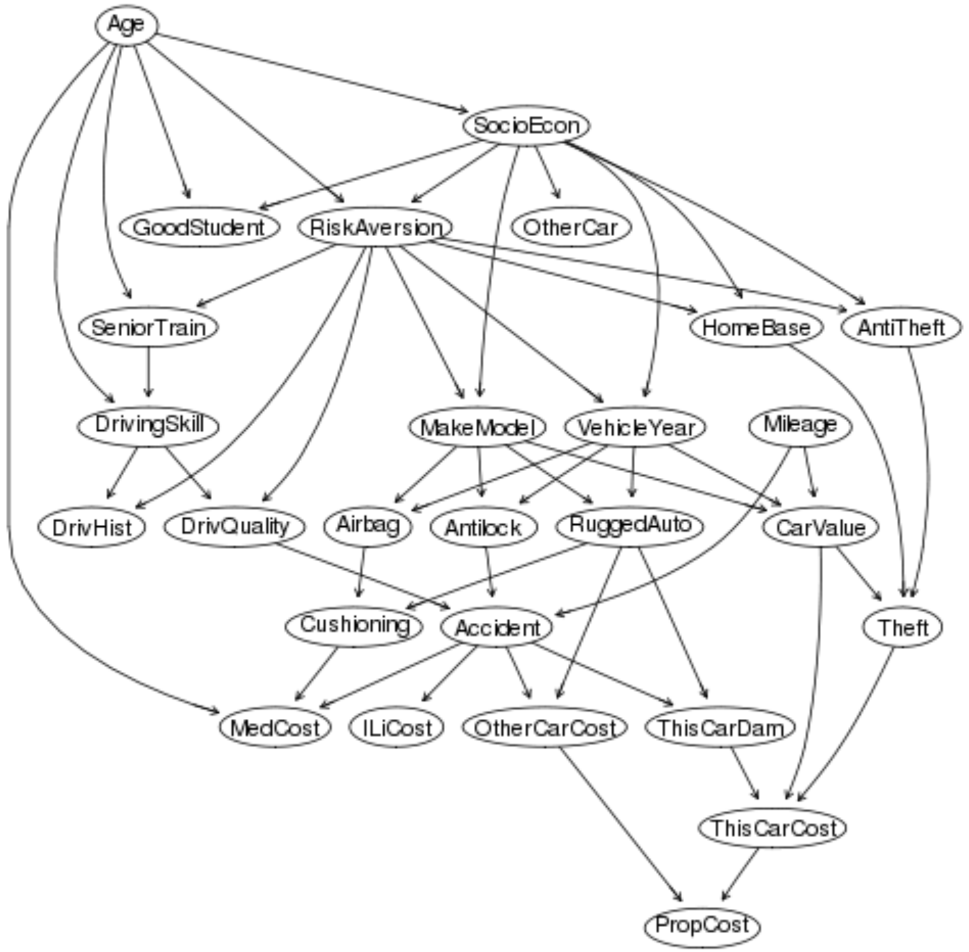


Example

$P(L, B, R, T, D, M) =$
$P(L)$
$P(B)$
$P(\ R \mid L\ )$
$P(\ T \mid R, B\ )$
$P(\ D \mid R, T\ )$
$P(\ M \mid B, D\ )$



Example: Insurance



https://www.bnlearn.com/bnrepository/discrete-medium.html#insurance

Queries on Bayesian Networks

- Which variables are conditionally independent?
  - For **any** values of the parameters
  - Called **d-separation**

- What is the most likely assignment, i.e., $\max_{x_1,\ldots,x_n} P(x_1, \ldots, x_n)$?
  - Called **maximum a posteriori (MAP) inference**

- What is the conditional distribution $P\left(X_i \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}\right)$?
  - For any X_i and any $X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}$
  - Called **marginal inference**

Queries on Bayesian Networks

- Which variables are conditionally independent?
  - For **any** values of the parameters
  - Called **d-separation**
- What is the most likely assignment, i.e., $\max_{x_1,\ldots,x_n} P(x_1, \ldots, x_n)$?
  - Called **maximum a posteriori (MAP) inference**
- What is the conditional distribution $P\left(X_i \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}\right)$?
  - For any X_i and any $X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}$
  - Called **marginal inference**

D-Separation Strategy

- **Step 1:** Look at three special cases
  - Causal chain
  - Common cause
  - Common effect

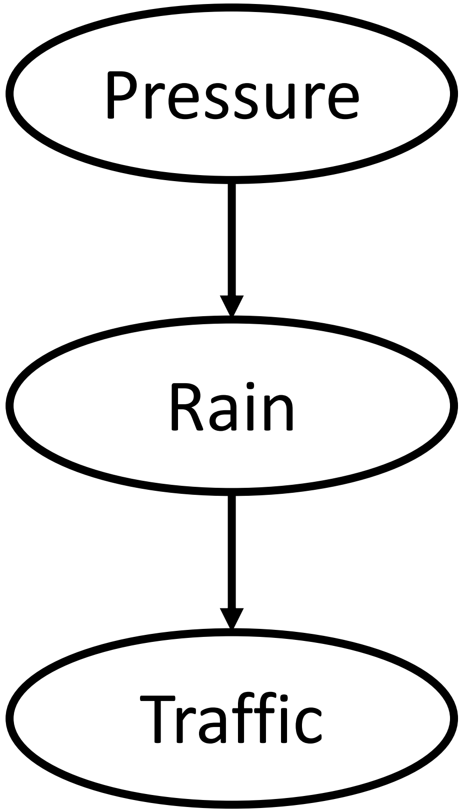
- **Step 2:** Piece them together

Causal Chain

- $X \rightarrow Y \rightarrow Z$

- Is $X \perp\!\!\!\perp Z$? Not necessarily
  - E.g., Rain $=$ Pressure and Traffic $=$ Rain

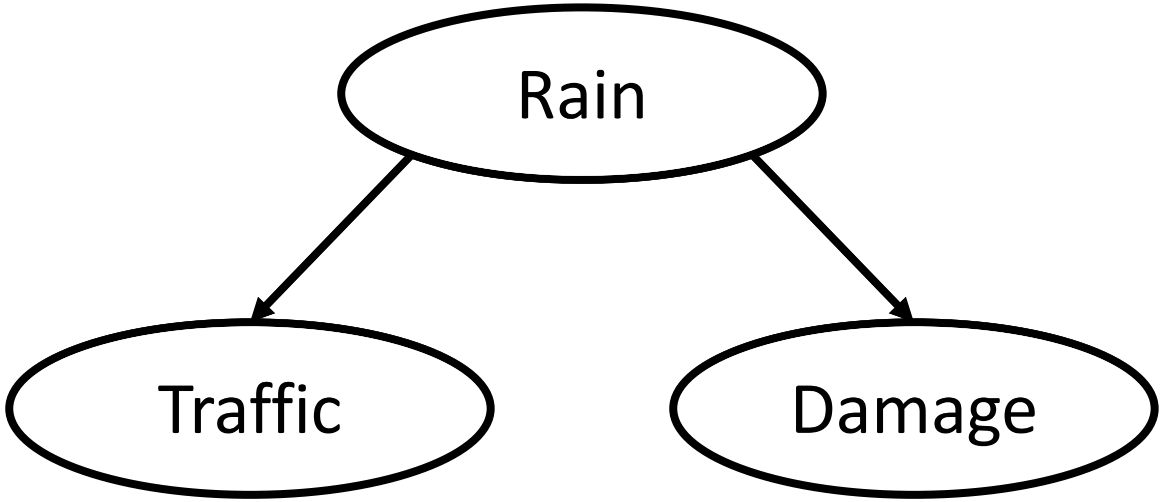
- Is $X \perp\!\!\!\perp Z \mid Y$? Yes
  - $P(z \mid x, y) = \frac{P(x,y,z)}{P(x,y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} = P(z \mid y)$



Common Cause

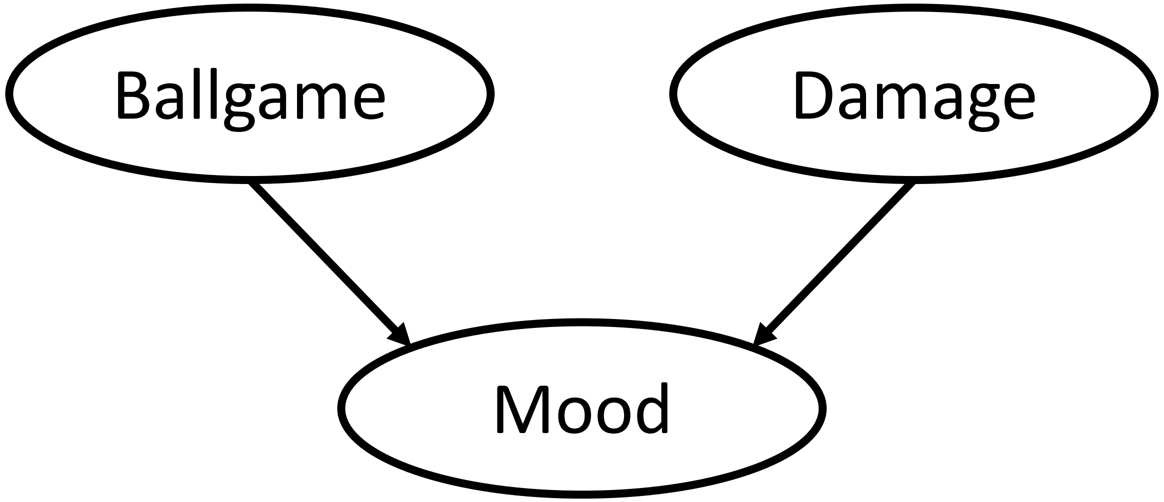
- $X \leftarrow Y \rightarrow Z$
- Is $X \perp\!\!\!\perp Z$? Not necessarily
  - E.g., $\text{Traffic} = \text{Rain and}$ $\text{Damage} = \text{Rain}$
- Is $X \perp\!\!\!\perp Z \mid Y$? Yes
  - $P(z \mid x, y) = \frac{P(x,y,z)}{P(x,y)}$

$$= \frac{P(x)P(x|y)P(z|y)}{P(x)P(x|y)} = P(z \mid y)$$



Common Effect

- $X \rightarrow Y \leftarrow Z$
- Is $X \perp\!\!\!\perp Z$? Yes
  - Proof left as exercise
- Is $X \perp\!\!\!\perp Z \mid Y$? Not necessarily
  - E.g., for $Y = X \oplus Z$ (XOR), then if $Y = \text{False}$, then $X = \neg Z$
  - **Example:** Medical diagnosis
- **Observation “activates” path**



General Case

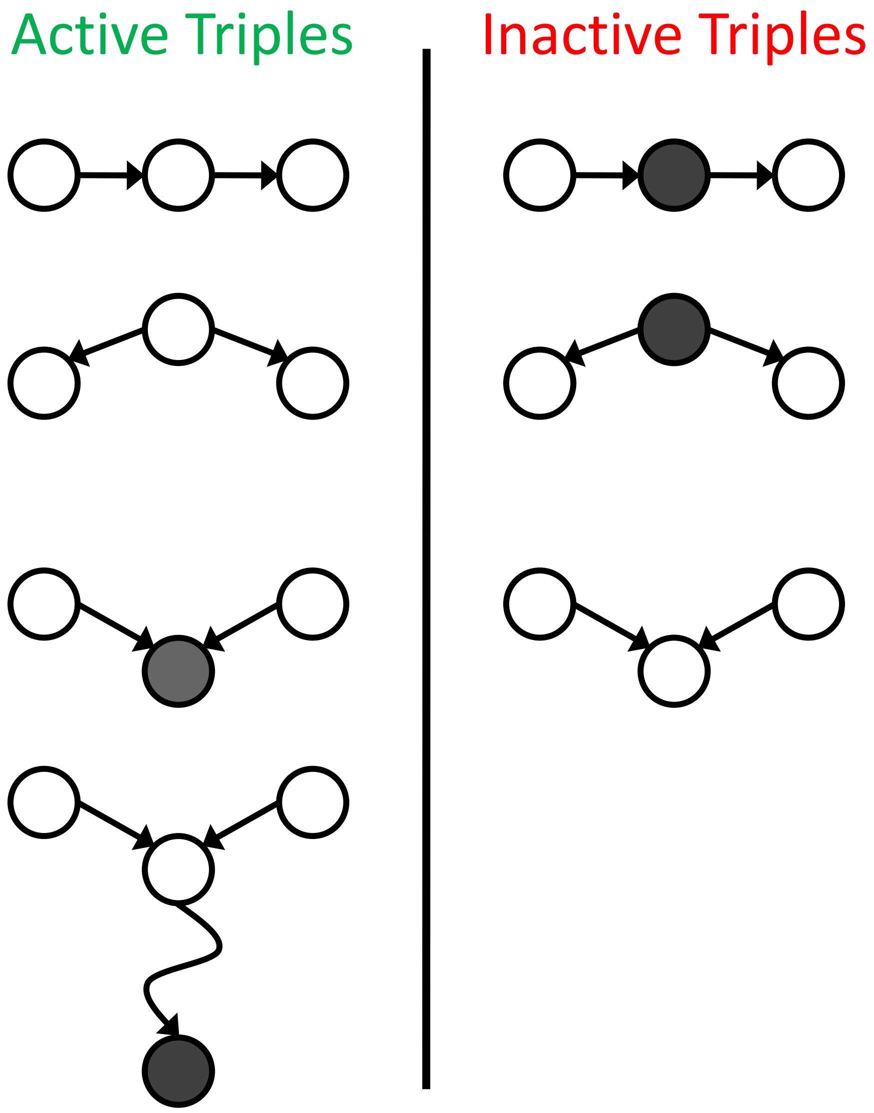
- **Query:** For a general Bayesian network, is $X \perp\!\!\!\perp Y \mid Z_1, \dots, Z_k$?

- **Algorithm**
  - Look for paths from X to Y
  - Segment $A - B - C$ only "active" (from previous three cases, see next slide)

- If there are **no** paths from X to Y such that **all** segments are active, then $X \perp\!\!\!\perp Y \mid Z_1, \dots, Z_k$
  - Otherwise, conditional independence is not guaranteed

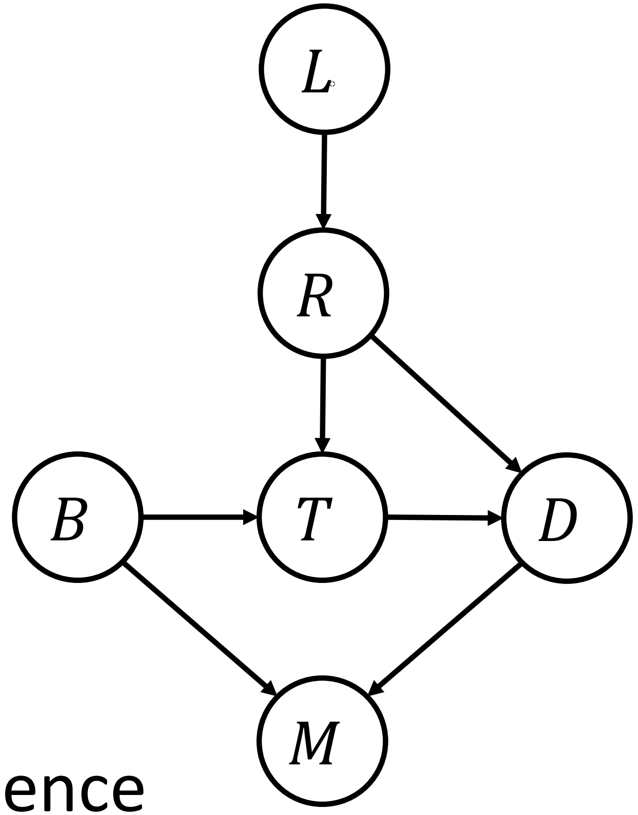
General Case

- **Causal chain**
  - $A \rightarrow B \rightarrow C$
  - Active iff $B \notin \{Z_i\}$
- **Common cause**
  - $A \leftarrow B \rightarrow C$
  - Active iff $B \notin \{Z_i\}$
- **Common effect**
  - $A \rightarrow B \leftarrow C$
  - Active iff $B \in \{Z_i\}$ (or descendant $\in \{Z_i\}$)



Example

- **Query:** Is $L \perp\!\!\!\perp M$?
  - No, $L \rightarrow R \rightarrow D \rightarrow M$
- **Query:** Is $L \perp\!\!\!\perp B$?
  - Yes!
  - $L \rightarrow R \rightarrow T \leftarrow B$
  - $L \rightarrow R \rightarrow D \leftarrow T \leftarrow B$
  - $L \rightarrow R \rightarrow D \rightarrow M \leftarrow B$
- **Note:** If we observe T, D, or M, breaks independence
  - None of $L \perp\!\!\!\perp B \mid T$, $L \perp\!\!\!\perp B \mid D$, and $L \perp\!\!\!\perp B \mid M$ hold



Queries on Bayesian Networks

- Which variables are conditionally independent?
  - For **any** values of the parameters
  - Called **d-separation**

- What is the most likely assignment, i.e., $\max_{x_1,\ldots,x_n} P(x_1, \ldots, x_n)$?
  - Called **maximum a posteriori (MAP) inference**

- What is the conditional distribution $P\left(X_i \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}\right)$?
  - For any X_i and any $X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}$
  - Called **marginal inference**

Marginal Inference

- **Input:**
  - **Evidentiary variables:** $E_1 = e_1, \ldots, E_k = e_k$ (features)
  - **Query variable:** Q (label)
  - **Hidden variables:** $H_1, \ldots, H_m$ (all remaining, “latent” variables)

- **Goal:** For each q, compute

$$P(Q = q \mid E_1 = e_1, \ldots, E_k = e_k)$$

- **Equivalently:** Likelihood $p(y \mid x)$

Enumerative Algorithm

- **Step 1:** Construct table for joint distribution $P(q, h_1, \dots, h_m, e_1, \dots, e_k)$

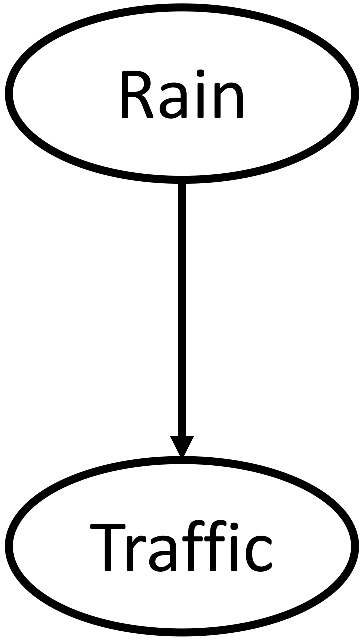
- **Step 2:** Select rows consistent with evidence
  - I.e., $P(q, h_1, \dots, h_m, e_1, \dots, e_k)$ for some $h_1, \dots, h_m$

- **Step 3:** Sum out hidden variables and normalize:

$$P(Q = q \mid e_1, \dots, e_k) = \frac{1}{Z} \sum_{h_1, \dots, h_h} P(q, h_1, \dots, h_m, e_1, \dots, e_k)$$

  - Normalizing constant $Z = \sum_{q, h_1, \dots, h_h} P(q, h_1, \dots, h_m, e_1, \dots, e_k)$

Step 1: Construct Joint Distribution



R	$P(R)$
no rain	0.6
rain	0.4

R	T	$P(T \mid R)$
no rain	no traffic	0.75
no rain	traffic	0.25
rain	no traffic	0.25
rain	traffic	0.75

R	T	$P(R,T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

$$P(R,T) = P(T \mid R)P(R)$$

Query: $P(R \mid \text{traffic})$

Step 2: Select Rows

R	T	$P(R,T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

Query: $P(R \mid \text{traffic})$

Step 2: Select Rows

R	T	$P(R, T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

Query: $P(R \mid \text{traffic})$

Step 3: Sum and Normalize

R	T	$P(R, T)$
no rain	no traffic	0.45
no rain	traffic	0.15
rain	no traffic	0.1
rain	traffic	0.3

$$P(\text{rain} \mid \text{traffic}) = \frac{0.15}{0.3+0.15} = \frac{1}{3}$$

$$P(\text{no rain} \mid \text{traffic}) = \frac{0.3}{0.3+0.15} = \frac{2}{3}$$

Query: $P(R \mid \text{traffic})$

Enumerative Algorithm

- Constructing the joint distribution is very computationally expensive!
- NP hard in general, but we can do better in practice
- **Idea:** Marginalize hidden variables before the end

Factors and Operations

- **Factor:** A table encoding a distribution $P(\, x_1, \dots, x_k \mid y_1, \dots, y_h \,)$
  - In general, we denote factors by $\phi(z_1, \dots, z_m)$

- **Join:** Given $\phi(x_1, \dots, x_k, y_1, \dots, y_m)$ and $\phi(x_1, \dots, x_k, z_1, \dots, z_n)$ output

$$\phi(x_1, \dots, x_k, y_1, \dots, y_m, z_1, \dots, z_n) = \phi(x_1, \dots, x_k, y_1, \dots, y_m)\phi(x_1, \dots, x_k, z_1, \dots, z_n)$$

- **Eliminate:** Given $\phi(x, y_1, \dots, y_k)$ output

$$\phi(y_1, \dots, y_k) = \sum_x \phi(x, y_1, \dots, y_k)$$

Enumerative Algorithm

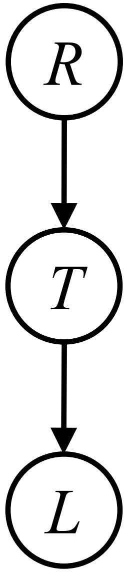
- **Step 0:** Initial factors are $P(\,X_i \mid \mathrm{parents}(X_i)\,)$ for each node X_i
  - Immediately drop rows conditioned on evidentiary variables

- **Step 1:** Join all factors

- **Step 2:** Eliminate all hidden variables

- **Output:** Remaining factor is $P(Q, e_1, \dots, e_k)$, which can be normalized

Example Query



$P(R)$

+r	0.1
-r	0.9

$P(T \mid R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Query: $P(L)$

Step 0: Initial Factors

$P(R)$

+r	0.1
-r	0.9

$P(T \mid R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Step 1: Join All Factors

$P(R)$

+r	0.1
-r	0.9

$P(T \mid R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R, T) = P(T \mid R)P(R)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R, T, L) = P(L \mid T)P(R, T)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Step 2: Eliminate Hidden Variables

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

$P(T, L) = \sum_r P(R, T, L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

$P(L) = \sum_t P(T, L)$

+l	0.134
-l	0.886

Variable Elimination Strategy

$P(R)$

+r	0.1
-r	0.9

$P(T \mid R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R,T) = P(\,T \mid R\,)P(R)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(T) = \sum_r P(R,T)$

+t	0.17
-t	0.83

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

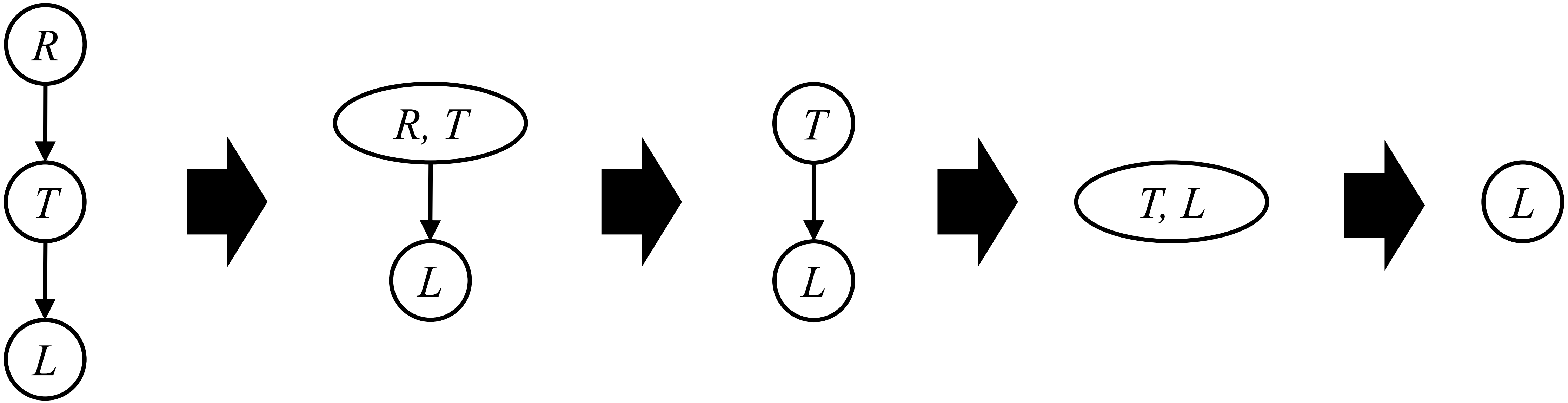
$P(T,L) = P(\,L \mid T\,)P(T)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

$P(L) = \sum_t P(L,T)$

+l	0.134
-l	0.866

Variable Elimination Strategy



What about evidence?

- When there are evidentiary variables, select those rows first

$P(R)$

+r	0.1
-r	0.9

$P(T \mid R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(+r)$

+r	0.1

$P(T \mid +r)$

+r	+t	0.8
+r	-t	0.2

$P(L \mid T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Query: $P(L \mid +r)$

What about evidence?

- At the end, obtain an unnormalized distribution, which we normalize

$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

$P(L \mid +r)$

+l	0.26
-l	0.74

Query: $P(L \mid +r)$

Alternative View

$$P(\ell) = \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(\ell \mid t)\underbrace{P(r)P(t \mid r)}_{\text{join on } r}}_{\text{join on } t}}_{\text{eliminate } r}}_{\text{eliminate } t}$$

Enumeration

$$P(\ell) = \underbrace{\sum_t \underbrace{P(\ell \mid t) \underbrace{\sum_r \underbrace{P(r)P(t \mid r)}_{\text{join on } r}}_{\text{eliminate } r}}_{\text{join on } t}}_{\text{eliminate } t}$$

Variable Elimination

General Variable Elimination Strategy

- **Step 0:** Initial factors are $P(\, X_i \mid \text{parents}(X_i)\,)$ for each node X_i
  - Immediately drop rows conditioned on evidentiary variables

- **Step 1:** For each H_i:
  - **Step 1a:** Join all factors containing H_i
  - **Step 1b:** Eliminate H_i

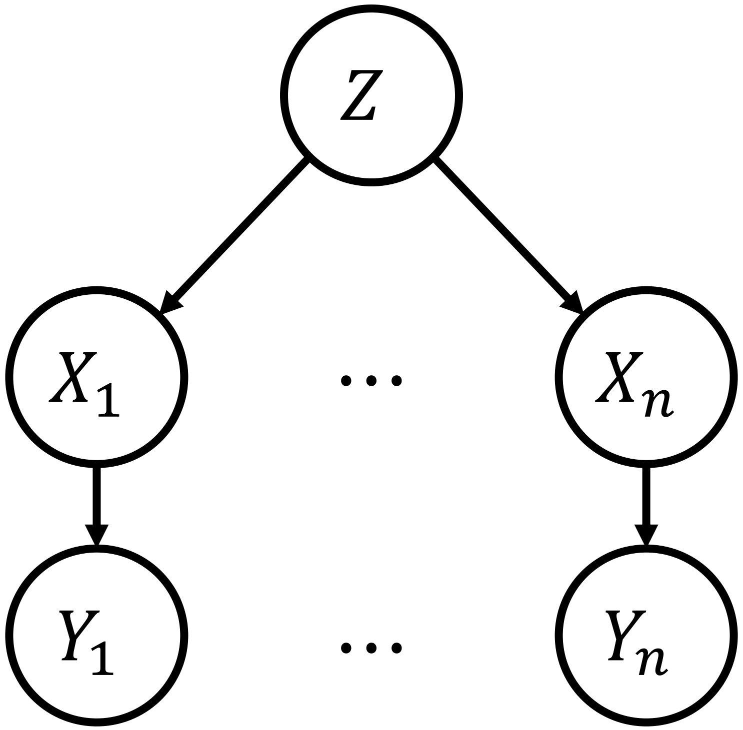
- **Output:** Join all remaining factors and normalize

Variable Elimination Order

- **Query:** $P(\, X_n \mid y_1, \ldots, y_n \,)$

- Eliminating Z first results in factor of size 2^{n+1}

- Eliminating $X_1, \ldots, X_{n-1}$ first results in factors of size 2



Variable Elimination Order

- Order in which hidden variables are eliminated can greatly affect performance (e.g., exponential vs. constant)

- May not exist an efficient ordering (problem is NP hard in general)

- Computing optimal ordering is also NP hard

Learning Bayesian Networks

- **Supervised learning**
  - Features x are evidentiary variables
  - Label y is query variable
  - Parameters are the conditional probabilities
  - Marginal inference evaluates likelihood $p(y \mid x)$

- How to learn the parameters?

Maximum Likelihood Learning

- Minimize the NLL:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{d} \log P_{\theta}\left(X_j = x_{i,j} \mid \text{parents}(X_j) = \left(x_{i,k_1}, \dots, x_{i,k_j}\right)\right)$$

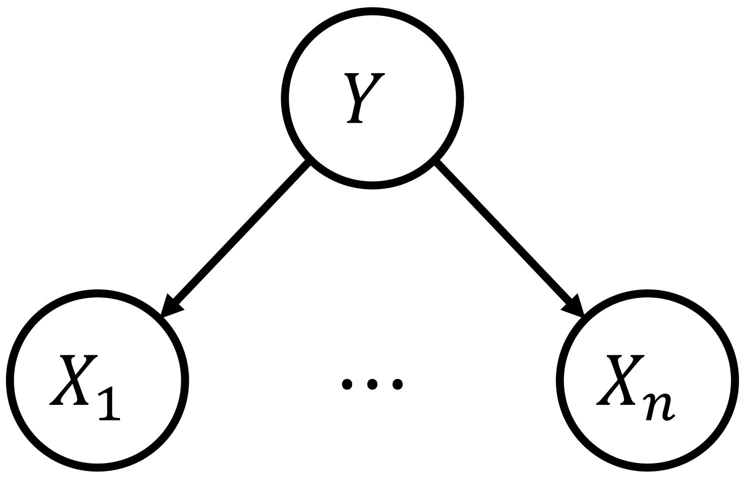
- Can use gradient descent to optimize
  - There is a nice formula for the gradient

Simplest Example: Naïve Bayes

- Model:

$$P(Y, X_1 \dots, X_n) = P(Y) \prod_{i=1}^{n} P(\,X_i \mid Y\,)$$

- If Y has domain D_Y and X_i has domain D_X, then $n \cdot |D_X| \cdot |D_Y|$ parameters



Inference in Naïve Bayes

- **Step 1:** For each $y \in D_Y$, compute joint probability distribution

$$P(y, x_1, \dots, x_n) = P(y) \prod_{i=1}^{n} P(\, x_i \mid y \,)$$

- **Step 2:** Normalize distribution:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y, x_1, \dots, x_n)}{Z}$$

  - Here, $Z = \sum_{y' \in D_Y} P(y', x_1, \dots, x_n)$

Naïve Bayes for Spam Detection

- Bag of words model

- Parameter sharing via "tied" distribution: For all i, j, constrain

$$P(\, X_i = x \mid Y\,) = P(X_j = x \mid Y$$

- Encodes invariant structure in bag of words models

Naïve Bayes for Spam Detection

$P(y)$

```
not spam: 0.66
spam:     0.33
```

$P(x \mid \text{spam})$

```
the :  0.0156
to  :  0.0153
and :  0.0115
of  :  0.0095
you :  0.0093
a   :  0.0086
with:  0.0080
from:  0.0075
...
```

$P(x \mid \text{not spam})$

```
the :  0.0210
to  :  0.0133
of  :  0.0119
2002:  0.0110
with:  0.0108
from:  0.0107
and :  0.0105
a   :  0.0100
...
```

Maximum Likelihood Learning

- Minimize the NLL for Naïve Bayes for text:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \left\{ \log P_{\theta}(y_i) + \log \sum_{j=1}^{d} P_{\theta}\left(x_{i,j} \mid y_i \right) \right\}$$

- Can show that parameters are counts:

$$P_{\theta}(x \mid y) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{d} 1\left(y_i = y \wedge x_{i,j} = x\right)}{\sum_{i=1}^{n} \sum_{j=1}^{d} 1(y_i = y)}$$

Maximum Likelihood Learning

- Can overfit
  - If a word never occurs in the training dataset, probabilities are all undefined

- Regularization via **Laplace smoothing**
  - Assume each word occurs k extra times in the dataset (increase counts by k)

$$P_\theta(\, x \mid y \,) = \frac{k + \sum_{i=1}^{n} \sum_{j=1}^{d} 1\big(y_i = y \wedge x_{i,j} = x\big)}{k \cdot d + \sum_{i=1}^{n} \sum_{j=1}^{d} 1(y_i = y)}$$

  - Can be interpreted as a prior on θ (in particular, the Dirichlet prior)

Example: Works Well

Recipients

Send me the money right away

??????,

???????? ! ???? ?????, wire the money! ???????

????? ??????

Prince ?????

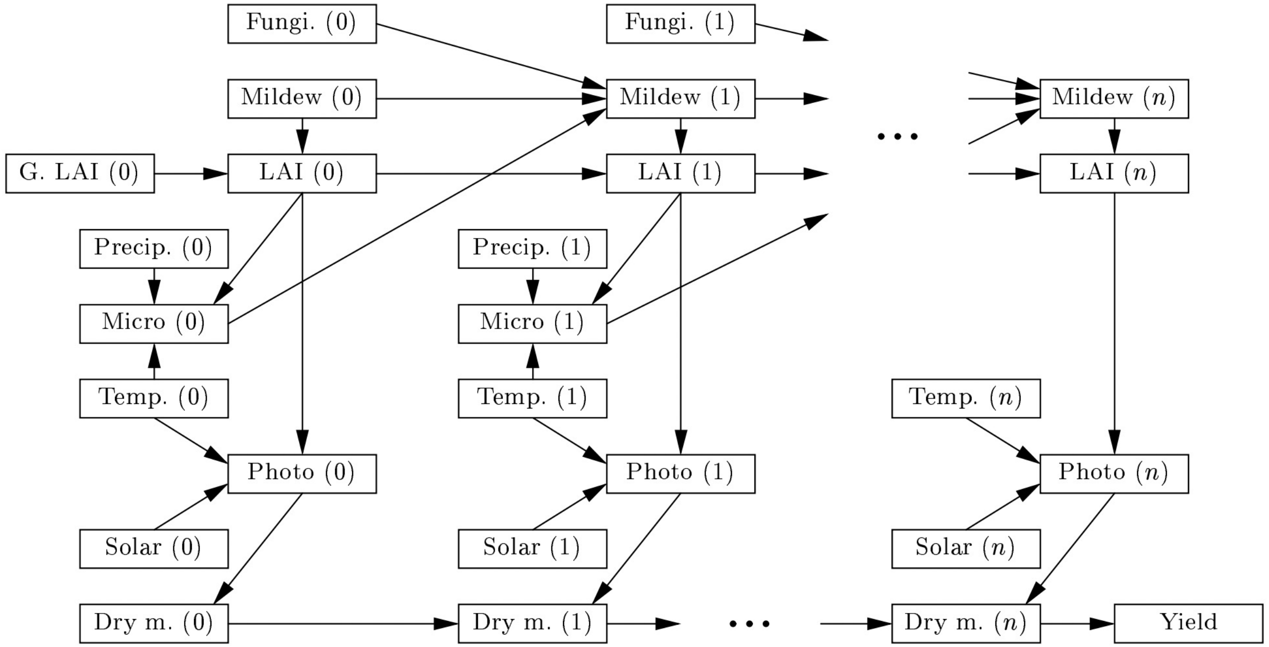
Example: Works Poorly

I wanted to love XXXXX, but I couldn't.

I wanted to love XXXXX, and I did!

Reasoning Through Time

- One strength of the framework is for modeling time varying processes
  - E.g., use (partial) measurements of factors to estimate future crop yield

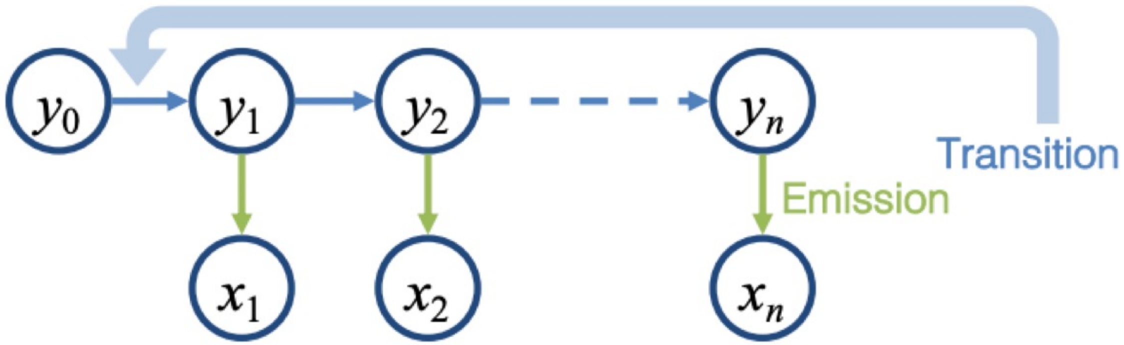


http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=01495B0A3249E426AD4C55FA02C14200?doi=10.1.1.30.9931&rep=rep1&type=pdf

Hidden Markov Model

- Speech recognition, machine translation, object tracking

We want a model of sequences y and observations x



$$p(x_1 \ldots x_n, y_1 \ldots y_n) = q(STOP|y_n) \prod_{i=1}^{n} q(y_i|y_{i-1}) e(x_i|y_i)$$

where $y_0 = START$ and we call $q(y_i \mid y_{i-1})$ the transition distribution and $e(x_i \mid y_i)$ the emission (or observation) distribution.