CSE 312

Foundations of Computing II

Lecture 7: Bayesian Inference, Chain Rule, Independence

Review Conditional & Total Probabilities

- **Conditional Probability**

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- **Bayes Theorem**

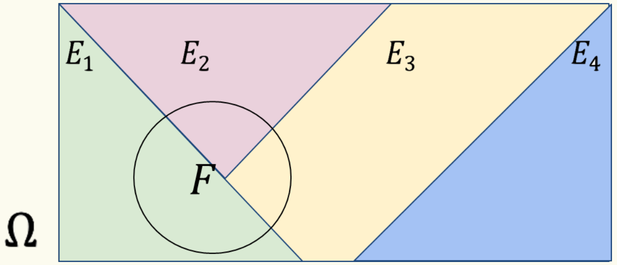
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

if $P(A) \neq 0, P(B) \neq 0$

- **Law of Total Probability**

$E_1, \ldots, E_n$ partition Ω

$$P(F) = \sum_{i=1}^{n} P(F \cap E_i) = \sum_{i=1}^{n} P(F|E_i)P(E_i)$$



Agenda

- Bayes Theorem + Law of Total Probability ◀
- Chain Rule
- Independence
- Infinite process and Von Neumann’s trick
- Conditional independence

Example – Zika Testing

Suppose we know the following Zika stats

- A test is 98% effective at detecting Zika ("true positive") $P(T|Z)$
- However, the test may yield a "false positive" 1% of the time $P(T|Z^c)$
- 0.5% of the US population has Zika. $P(Z)$

What is the probability you test negative (event T^c) if you have Zika (event Z)?

$$P(T^c|Z) = 1 - P(T|Z) = 2\%$$

What is the probability you have Zika (event Z) if you test negative (event T^c)?

By Bayes Rule, $P(Z|T^c) = \dfrac{P(T^c|Z)P(Z)}{P(T^c)}$

By the Law of Total Probability, $P(T^c) = P(T^c|Z)P(Z) + P(T^c|Z^c)P(Z^c)$

$$= \frac{2}{100} \cdot \frac{5}{1000} + \left(1 - \frac{1}{100}\right) \cdot \frac{995}{1000} = \frac{10}{100000} + \frac{98505}{100000}$$

So, $P(Z|T^c) = \frac{10}{10+98505} \approx 0.01\ \%$

Bayes Theorem with Law of Total Probability

Bayes Theorem with LTP: Let $E_1, E_2, \dots, E_n$ be a partition of the sample space, and F and event. Then,

$$P(E_1|F) = \frac{P(F|E_1)P(E_1)}{P(F)} = \frac{P(F|E_1)P(E_1)}{\sum_{i=1}^{n} P(F|E_i)P(E_i)}$$

Simple Partition: In particular, if E is an event with non-zero probability, then

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|E^C)P(E^C)}$$

Bayes Theorem with Law of Total Probability

Bayes Theorem with LTP: Let $E_1, E_2, \ldots, E_n$ be a partition of the sample space, and F and event. Then,

$$P(E_1|F) = \frac{P(F|E_1)P(E_1)}{P(F)} = \frac{P(F|E_1)P(E_1)}{\sum_{i=1}^{n} P(F|E_i)P(E_i)}$$

Simple Partition: In particul

probability, then

We just used this implicity on the negative Zika test example with $E = Z$ and $F = T^c$

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|E^C)P(E^C)}$$

Our First Machine Learning Task: Spam Filtering

Subject: “FREE $$$ CLICK HERE”

What is the probability this email is spam, given the subject contains “FREE”?

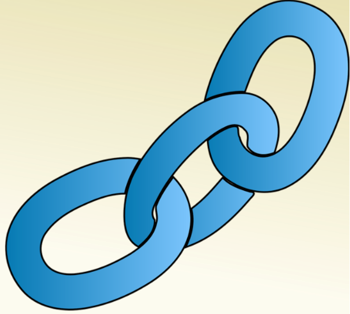
Some useful stats:

- 10% of ham (i.e., not spam) emails contain the word “FREE” in the subject.
- 70% of spam emails contain the word “FREE” in the subject.
- 80% of emails you receive are spam.

Agenda

- Bayes Theorem + Law of Total Probability
- Chain Rule
- Independence
- Infinite process and Von Neumann’s trick
- Conditional independence

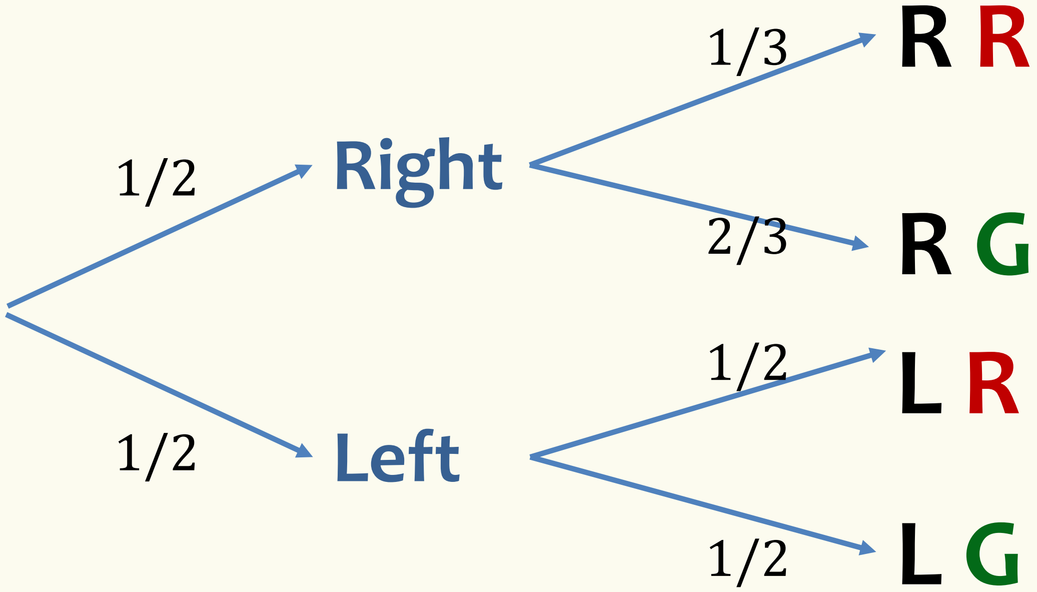
Chain Rule



$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Rightarrow \quad P(A)P(B|A) = P(A \cap B)$$

Often probability space $(\Omega, \mathbb{P})$ is given **implicitly** via sequential process

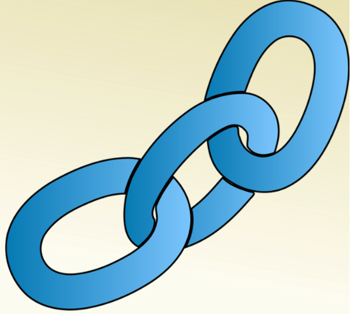
Recall from last time:



$$P(\textbf{R}) = P(\textbf{Left}) \times P(\textbf{R}|\textbf{Left}) + P(\textbf{Right}) \times P(\textbf{R}|\textbf{Right})$$

What if we have more than two (e.g., n) steps?

Chain Rule



$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Rightarrow \quad P(A)P(B|A) = P(A \cap B)$$

Theorem. (Chain Rule) For events $A_1, A_2, \dots, A_n$,

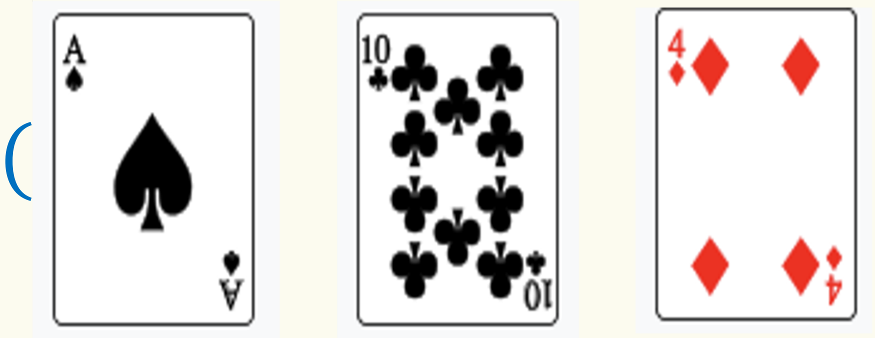
$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2)$$

$$\cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

An easy way to remember: We have n tasks and we can do them sequentially, conditioning on the outcome of previous tasks

Chain Rule Example

Shuffle a standard 52-card deck and draw the top 3 cards.
(uniform probability space)

What is $P($  $) = P(A \cap B \cap C)$?

A: Ace of Spades First
B: 10 of Clubs Second
C: 4 of Diamonds Third

$$P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$

$$\frac{1}{52} \cdot \frac{1}{51} \cdot \frac{1}{50}$$

Agenda

- Bayes Theorem + Law of Total Probability
- Chain Rule
- Independence
- Infinite process and Von Neumann’s trick
- Conditional independence

Independence

Definition. Two events A and B are (statistically) **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

Equivalent formulations:

- If $P(A) \neq 0$, equivalent to $P(B|A) = P(B)$
- If $P(B) \neq 0$, equivalent to $P(A|B) = P(A)$

"The probability that B occurs after observing A" – Posterior
= "The probability that B occurs" – Prior

Independence - Example

Assume we toss two fair coins

"first coin is heads" $A = \{\mathrm{HH}, \mathrm{HT}\}$

"second coin is heads" $B = \{\mathrm{HH}, \mathrm{TH}\}$

$$P(A) = 2 \times \frac{1}{4} = \frac{1}{2}$$

$$P(B) = 2 \times \frac{1}{4} = \frac{1}{2}$$

$$P(A \cap B) = P(\{HH\}) = \frac{1}{4} = P(A) \cdot P(B)$$

Example – Independence

Toss a coin 3 times. Each of 8 outcomes equally likely.

- $A = \{\text{at most one } T\} = \{HHH, HHT, HTH, THH\}$
- $B = \{\text{at most } 2\ H'\text{s}\} = \{HHH\}^c$

Independent?

$$P(A \cap B) \overset{?}{=} P(A) \cdot P(B)$$

$$\frac{3}{8} \neq \frac{1}{2} \cdot \frac{7}{8}$$

Poll:

A. Yes, independent

B. No

pollev/stefanotessaro617

Multiple Events – Mutual Independence

Definition. Events $A_1, \ldots, A_n$ are **mutually independent** if for every non-empty subset $I \subseteq \{1, \ldots, n\}$, we have

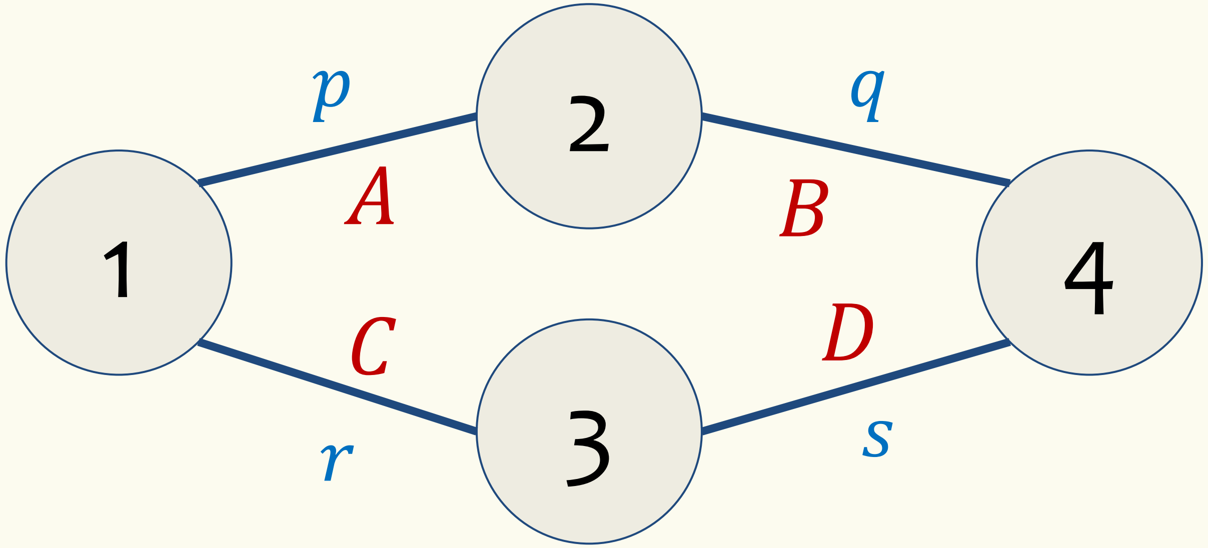
$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i).$$

Example – Network Communication

Each link works with the probability given, **independently**

i.e., mutually independent events A, B, C, D with

$$P(A) = p$$
$$P(B) = q$$
$$P(C) = r$$
$$P(D) = s$$

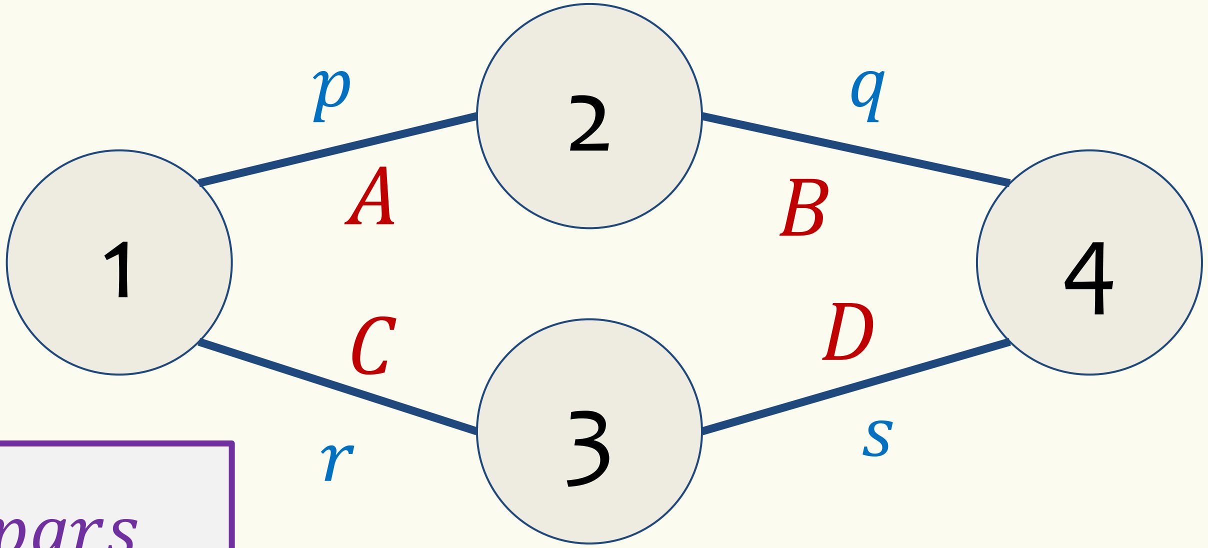


Example – Network Communication

If each link works with the probability given, **independently**:
What's the probability that nodes 1 and 4 can communicate?

$$P(\text{1-4 connected}) = P\big((A \cap B) \cup (C \cap D)\big)$$
$$= P(A \cap B) + P(C \cap D) - P(A \cap B \cap C \cap D)$$

$$P(A \cap B) = P(A) \cdot P(B) = pq$$
$$P(C \cap D) = P(C) \cdot P(D) = rs$$
$$P(A \cap B \cap C \cap D)$$
$$= P(A) \cdot P(B) \cdot P(C) \cdot P(D) = pqrs$$



$$\boxed{P(\text{1-4 connected}) = pq + rs - pqrs}$$

Independence as an assumption

- People often assume it **without justification**

- Example: A skydiver has two chutes

  A: event that the main chute doesn't open $\quad P(A) = 0.02$

  B: event that the back-up doesn't open $\quad P(B) = 0.1$

- What is the chance that at least one opens assuming independence?

Assuming independence doesn't justify the assumption!

Both chutes could fail because of the same rare event e.g., freezing rain.

Independence – Another Look

Definition. Two events A and B are (statistically) **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

“Equivalently.” $P(A|B) = P(A)$.

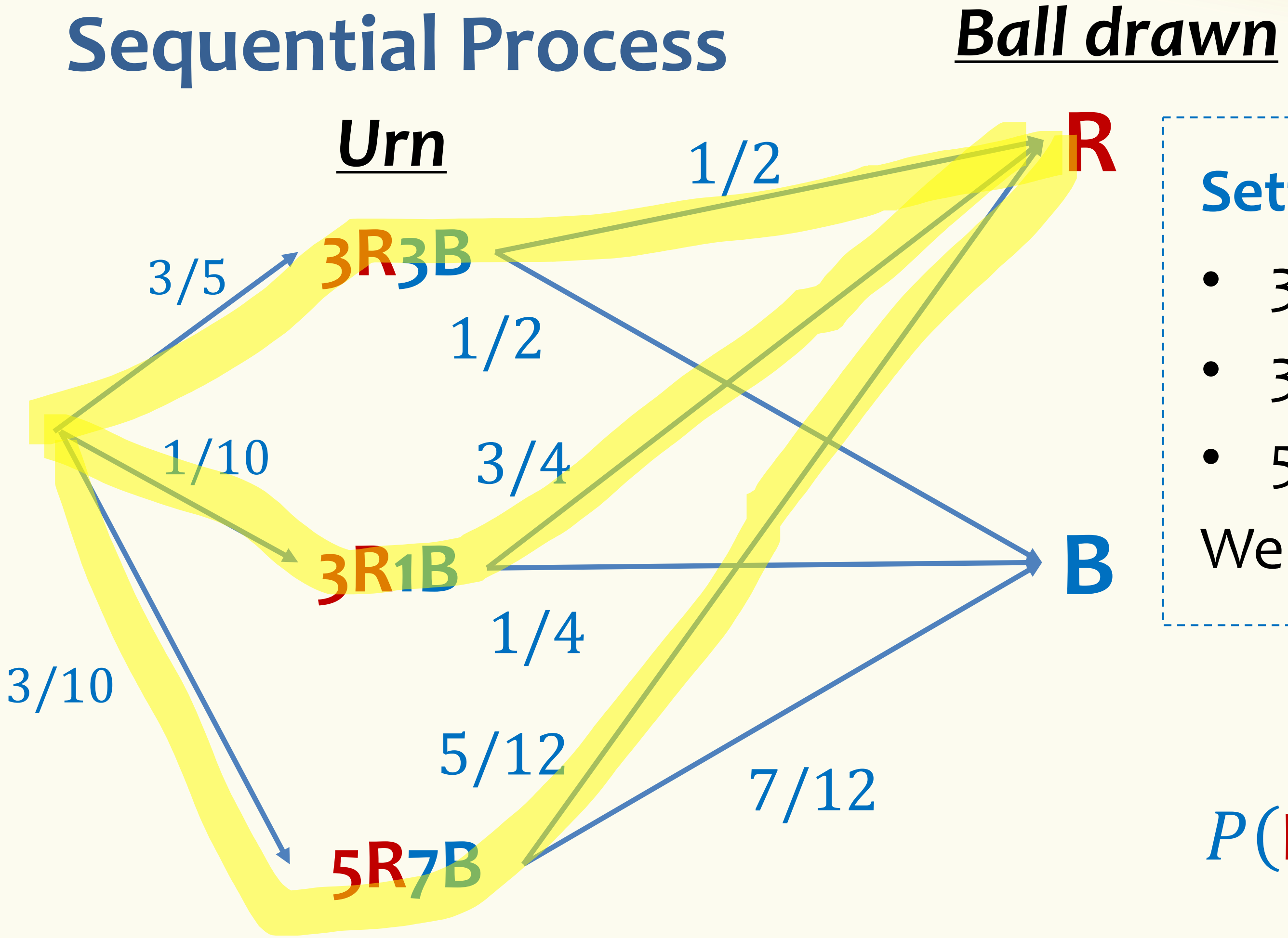
It is important to understand that independence is a property of probabilities of outcomes, not of the root cause generating these events.

Events generated independently ➔ *their probabilities satisfy independence*

~~⬅~~ *Not necessarily*

This can be counterintuitive!

Sequential Process



Are **R** and **3R3B** independent?

Setting: An urn contains:

- 3 **red** and 3 **blue** balls w/ probability 3/5
- 3 **red** and 1 **blue** balls w/ probability 1/10
- 5 **red** and 7 **blue** balls w/ probability 3/10

We draw a ball at random from the urn.

$$P(\mathbf{R}) = \frac{3}{5}\times\frac{1}{2}+\frac{1}{10}\times\frac{3}{4}+\frac{3}{10}\times\frac{5}{12}=\frac{1}{2}$$

$P(\mathbf{3R3B})\times P(\mathbf{R} \mid \mathbf{3R3B})$

Independent! $P(\mathbf{R}) = P(\mathbf{R} \mid \mathbf{3R3B})$

Agenda

- Bayes Theorem + Law of Total Probability
- Chain Rule
- Independence
- Infinite process and Von Neumann’s trick
- Conditional independence

Often probability space (Ω, P) is given **implicitly** via sequential process

- *Experiment proceeds in n sequential steps, each step follows some local rules defined by the chain rule and independence*
- *Natural extension:* Allows for easy definition of experiments where $|\Omega| = \infty$

Fun: Von Neumann's Trick with a biased coin

- How to use a biased coin to get a fair coin flip:
  - Suppose that you have a biased coin:
    - $P(H) = p \qquad P(T) = 1 - p$

1. Flip coin twice: If you get HH or TT go to step 1
2. If you got HT output H; if you got TH output T.

Why is it fair? $P(H) = P(HT) = p(1-p) = (1-p)p = P(TH) = P(T)$

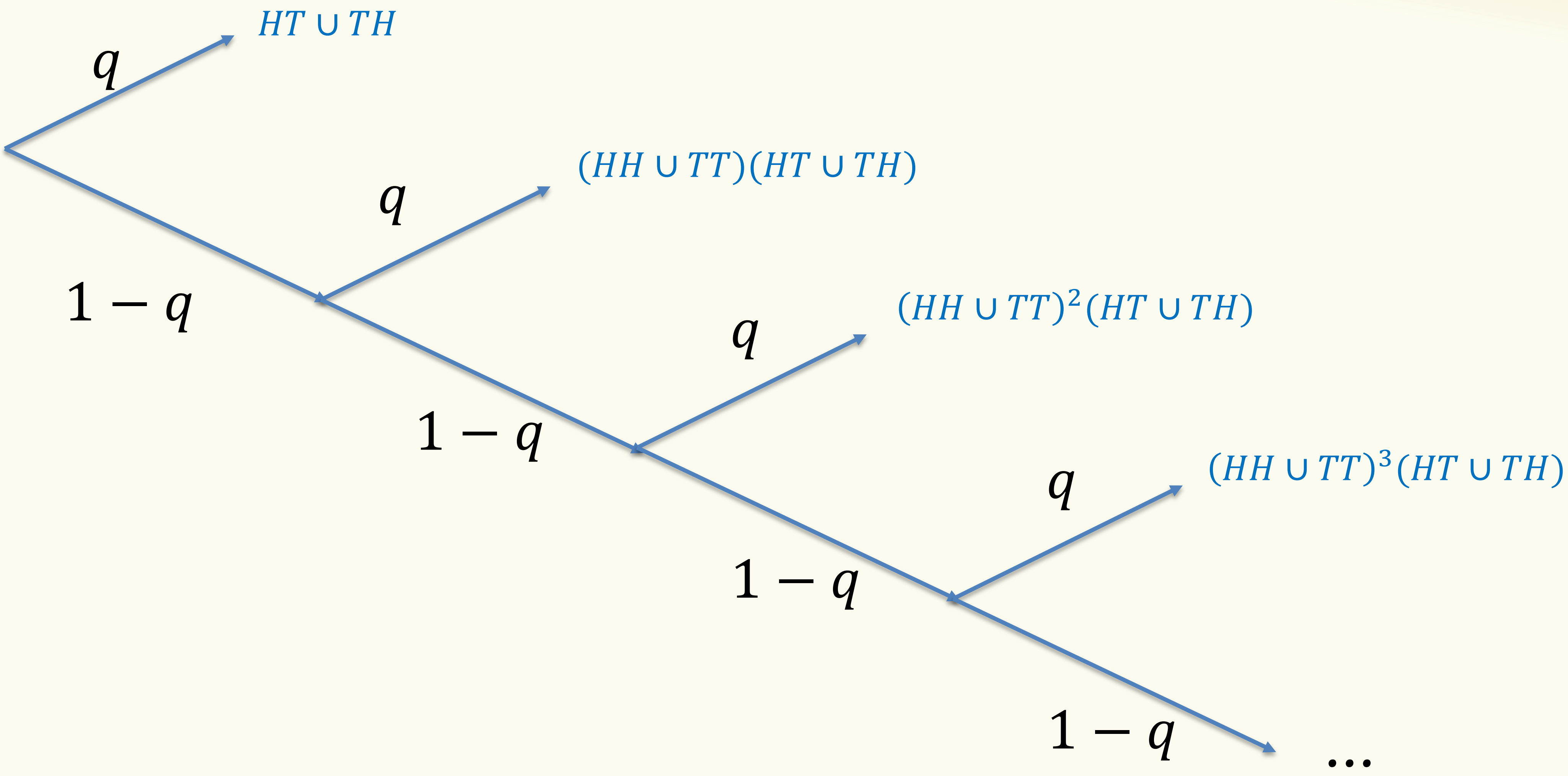
Drawback: You may never get to step 2.

The sample space for Von Neumann's trick

- For each round of Von Neumann's trick we flipped the biased coin twice.
  - If HT or TH appears, the experiment ends:
    - Total probability each round: $2p(1-p)$ call this q
  - If HH or TT appears, the experiment continues:
    - Total probability each round: $p^2 + (1-p)^2$ this is $1-q$

- Probability that flipping ends in round t is $(1-q)^{t-1} \cdot q$
  - Conditioned on ending in round t, $P(H) = P(T) = 1/2$

Sequential Process – Example


$HT \cup TH$
q
$(HH \cup TT)(HT \cup TH)$
q
$1 - q$
$(HH \cup TT)^2(HT \cup TH)$
q
$1 - q$
$(HH \cup TT)^3(HT \cup TH)$
q
$1 - q$
$1 - q$
…

The sample space for Von Neumann's trick

More precisely, the sample space contains the successful outcomes:

$$\cup_{t=1}^{\infty}(HH \cup TT)^{t-1}(HT \cup TH)$$

which together have probability $\sum_{t=1}^{\infty}(1-q)^{t-1}q$ for $q = 2p(1-p)$

as well as all of the failing outcomes in $(HH \cup TT)^{\infty}$.

Observe that $q \neq 0$ iff $0 < p < 1$. We have two cases:

- If $q \neq 0$ then $\sum_{t=1}^{\infty}(1-q)^{t-1} = 1/q$ so successful outcomes account for total probability 1.
- If $q = 0$ then either:
  - $p = 1$ and $(HH)^{\infty}$ has probability 1.
  - $p = 0$ and $(TT)^{\infty}$ has probability 1.

Agenda

- Bayes Theorem + Law of Total Probability
- Chain Rule
- Independence
- Infinite process and Von Neumann’s trick
- Conditional independence ◀

Conditional Independence

Definition. Two events A and B are **independent** conditioned on C if

$$P(C) \neq 0 \text{ and } P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

- If $P(A \cap C) \neq 0$, equivalent to $P(B|A \cap C) = P(B \mid C)$
- If $P(B \cap C) \neq 0$, equivalent to $P(A|B \cap C) = P(A \mid C)$

Plain Independence. Two events A and B are **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

- If $P(A) \neq 0$, equivalent to $P(B|A) = P(B)$
- If $P(B) \neq 0$, equivalent to $P(A|B) = P(A)$

Example – Throwing Dice

Suppose that Coin 1 has probability of heads 0.3
and Coin 2 has probability of head 0.9.

We choose one coin randomly with equal probability and flip that coin 3 times independently. What is the probability we get all heads?

$$P(HHH) = P(HHH \mid C_1) \cdot P(C_1) + P(HHH \mid C_2) \cdot P(C_2)$$

Law of Total Probability (LTP)

$$= P(H \mid C_1)^3 \, P(C_1) + P(H \mid C_2)^3 \, P(C_2)$$

Conditional Independence

$$= 0.3^3 \cdot 0.5 + 0.9^3 \cdot 0.5 = 0.378$$

C_i = coin i was selected

Conditional independence and Bayesian inference in practice: Graphical models

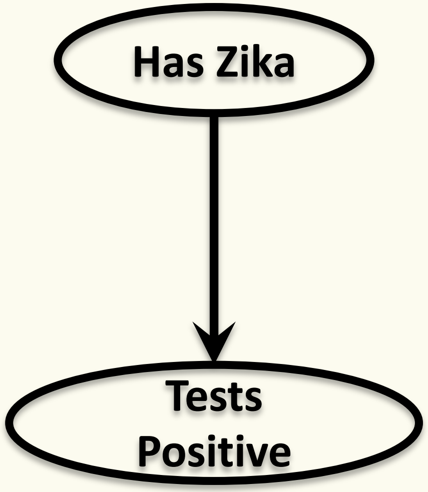
- The sample space Ω is often the Cartesian product of possibilities of many different variables

- We often can understand the probability distribution P on Ω based on ***local properties*** that involve a few of these variables at a time

- We can represent this via a directed acyclic graph augmented with probability tables (called a Bayes net) in which each node represents one or more variables…

Graphical Models/Bayes Nets

- Bayes net for the Zika testing probability space (Ω, P)



Z	$\neg Z$
0.005	0.995

	T	$\neg T$
Z	0.98	0.02
$\neg Z$	0.01	0.99

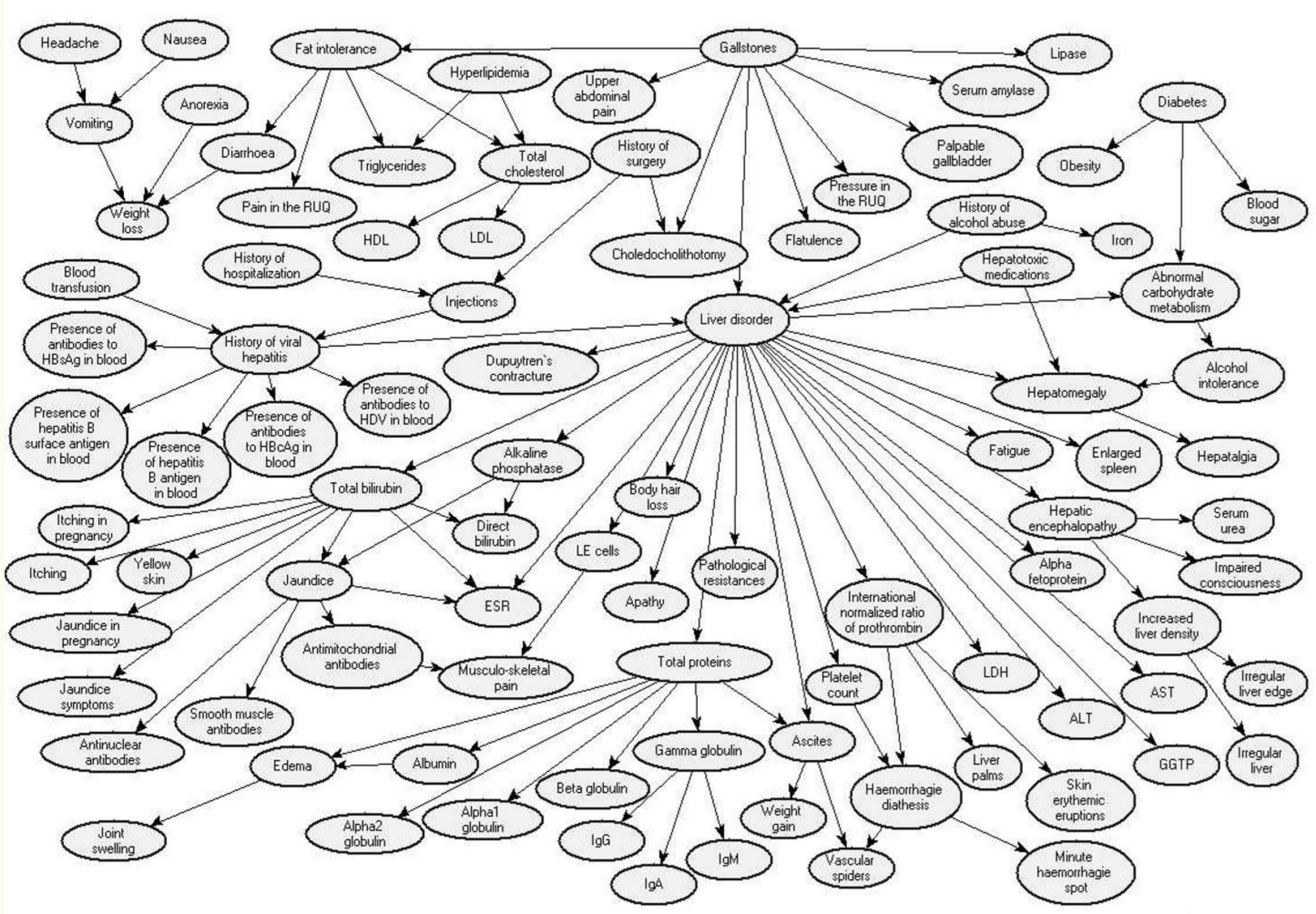
$P(T|\neg Z)$

Conditional Probability Table:

- One column for each value of the variables at the node
- One row for each combination of values of immediate predecessors

Ω = Cartesian product of possible value assignments at all nodes.

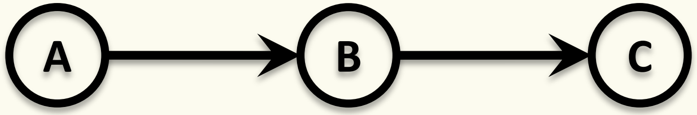
Graphical Models/Bayes Nets



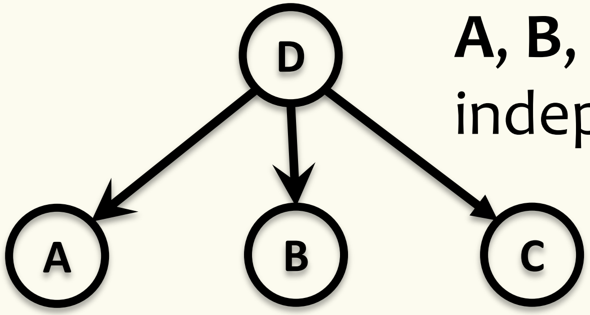
“A Bayesian Network Model for Diagnosis of Liver Disorders” – Agnieszka Onisko, M.S., Marek J. Druzdzel, Ph.D., and Hanna Wasyluk, M.D.,Ph.D.- September 1999.

Graphical Models/Bayes Nets

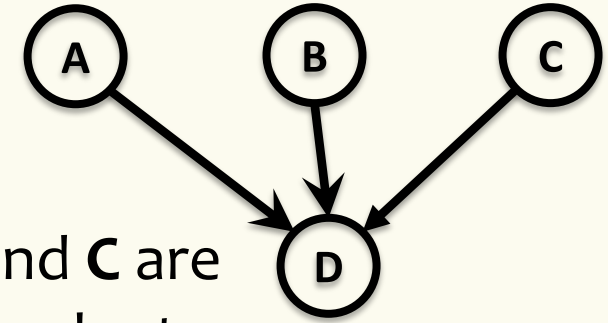
Bayes Net assumption/requirement

- The only dependence between variables is given by paths in the Bayes Net graph:
  - if only edges are 

    then **A** and **C** are *conditionally independent* given the value of **B**



A, B, C conditionally independent given **D**



A, B, and **C** are independent

Defines a unique global probability space (Ω, P)

Inference in Bayes Nets

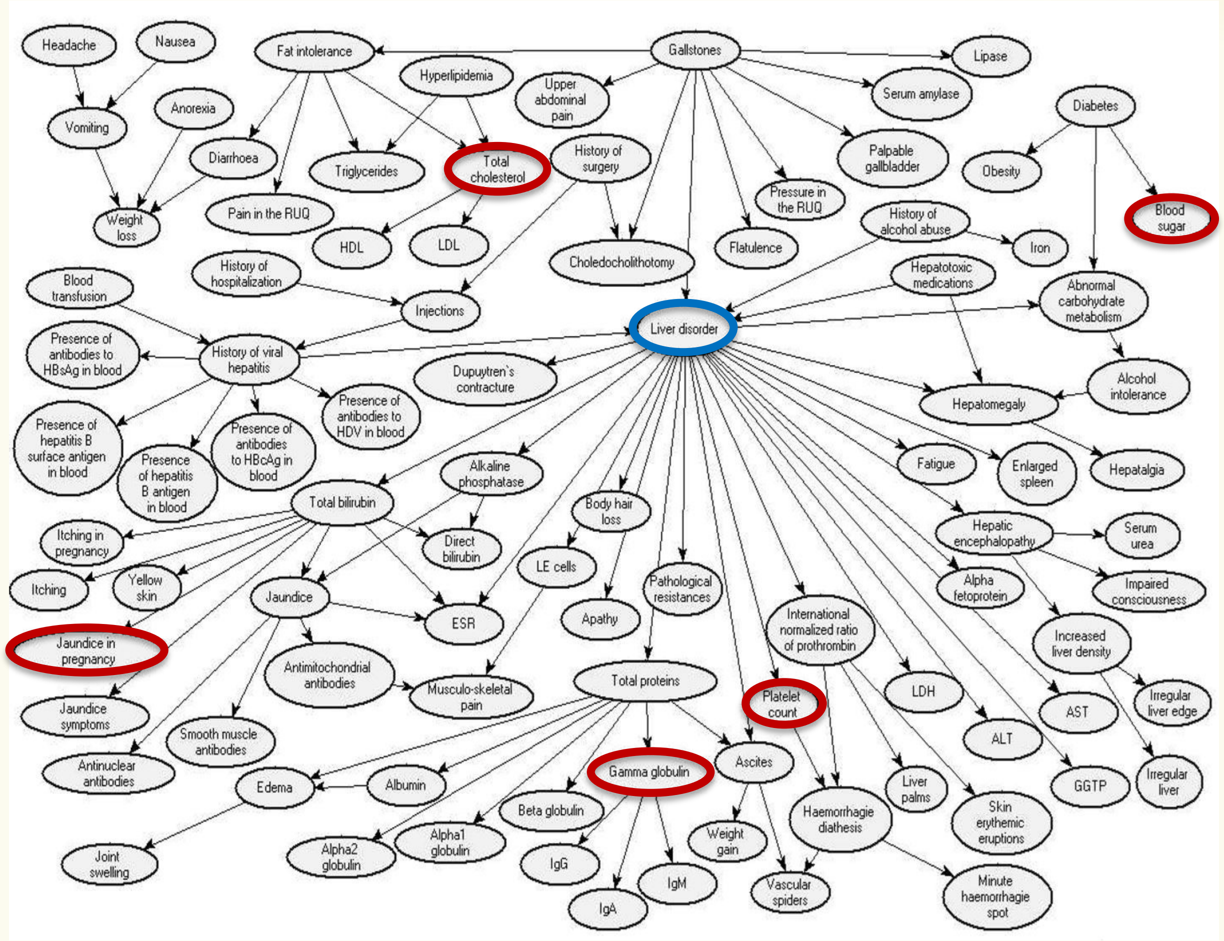
For much more see CSE 473

Given

- Bayes Net
  - graph
  - conditional probability tables for all nodes
- Observed values of variables at some nodes
  - e.g., clinical test results

Compute

- Probabilities of variables at other nodes
  - e.g., diagnoses



"A Bayesian Network Model for Diagnosis of Liver Disorders" – Agnieszka Onisko, M.S., Marek J. Druzdzel, Ph.D., and Hanna Wasyluk, M.D.,Ph.D.- September 1999.