CSCI 561
Foundation for Artificial Intelligence

19. Uncertainty and Probability

Professor Wei-Min Shen

University of Southern California

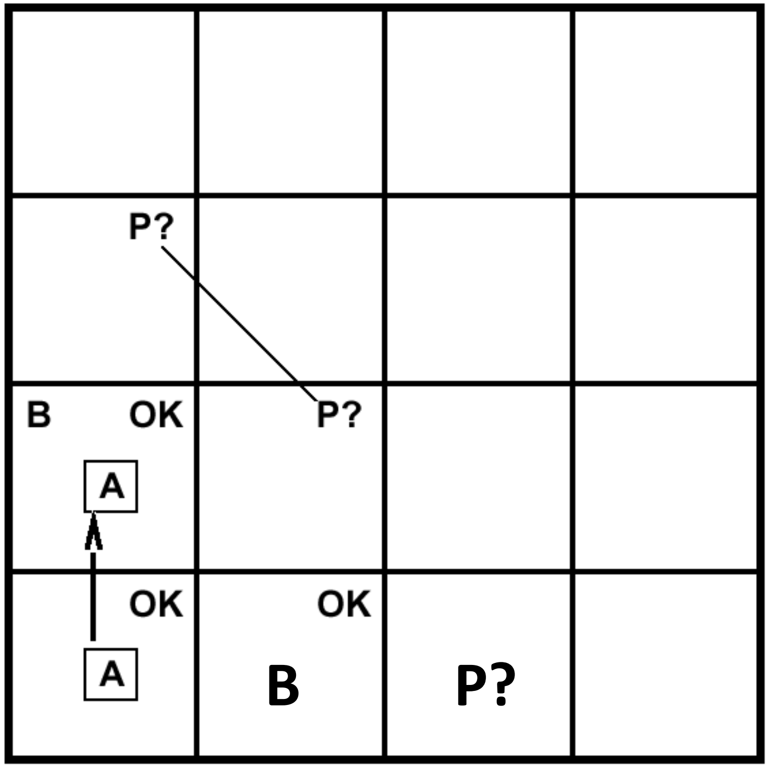
Outline

- Introduction to probability
    - Powerful tool for uncertainty reasoning
    - What and where is probability?
- Models and probability distributions
- Inferences based on probability methods

Example: Seeing vs Believing

- You define two "logic" (random) variables for two propositions:
  - X: "my eyes see the object"
  - Y: "the object is really there"
- Your real experience is as follows:
  - X = [1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1]
  - Y = [1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1]
- Definition of probabilities
  - P(X): the probability of "my eyes see the object"
  - P(Y): the probability of "the object is really there"
  - P(Y^X): the probability that "the object is really there **AND** my eyes see the object"
  - P(Y|X): the probability that "the object is really there **IF** my eyes see the object"
- Computing the probability
  - P(X) = 8/12, P(Y) = 7/12, P(X^Y) = P(XY) = P(X,Y) = 6/12, … (based on counting)
  - In common sense, "Seeing is Believing" ⇔ P(Y|X)=1.0
  - But in this real experience, P(Y|X) = ??? [this and others are what we will learn]

Example: Guessing a Pit



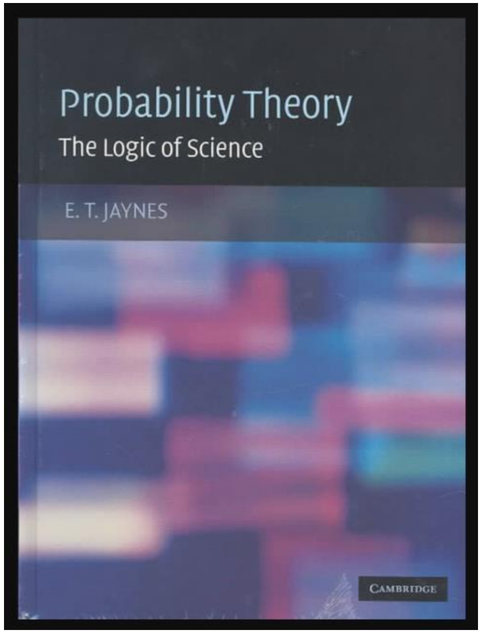
A= Agent
B= Breeze
S= Smell
P= Pit
W= Wumpus
OK = Safe
V = Visited
G = Glitter

Which one has a PIT?

- Logic can only guess randomly which of [1,3],[2,2],[3,1] has a pit
- Using probability you can calculate which one is more likely have a pit than others

Logics and Probabilities

- Not all knowledge are certain
  - Earlier stories of expert systems (medical, legal)
- Two big challenges for logic-like approaches
  - Common sense (vague): “Water flows down”
  - Uncertainty
- Probability: The Logic of Science
  - By E. T. Jaynes
  - >40 years of experience in physics
  - Builds upon principles (Bayes Rule)



Two Key Elements in Probability

- Probability Distribution Model
  - Real World: possible worlds, atomic events, samples
  - Our Mind: Variables, Value assignments (“entailment”)
- Inferences that can be made from the model
  - Sum rule
  - Product rule
  - Conditional
  - Marginalization
  - Normalization

Uncertainty (Random Variables)

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Problems:

1) partial observability (road state, other drivers' plans, etc.)
2) noisy sensors (KCBS traffic reports)
3) uncertainty in action outcomes (flat tire, etc.)
4) immense complexity of modelling and predicting traffic

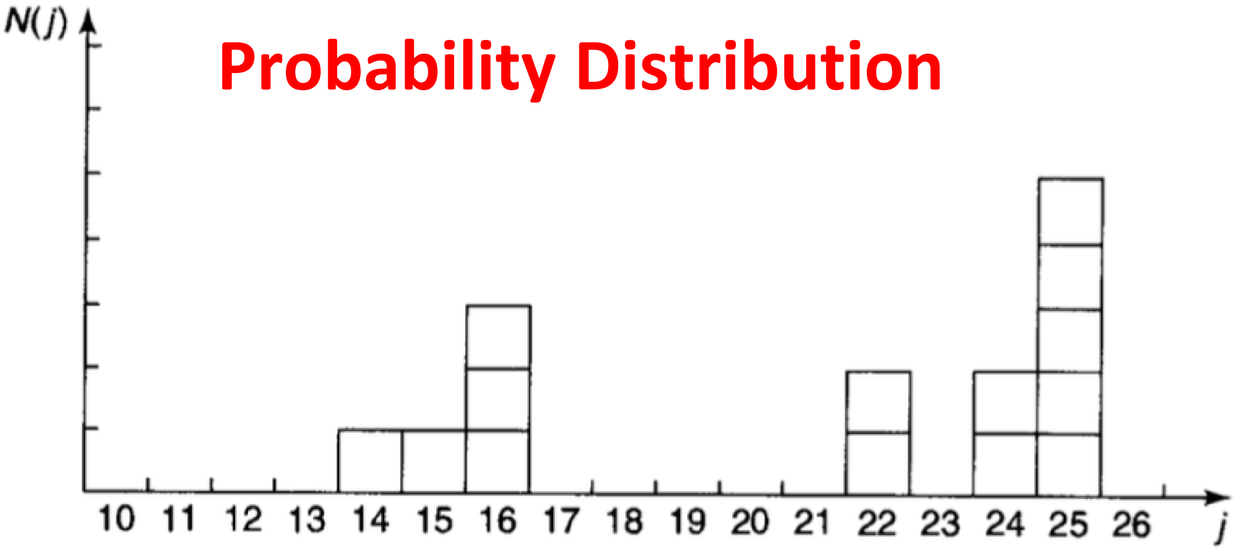
Hence a purely logical approach either
1) risks falsehood: "A_{25} will get me there on time"
or 2) leads to conclusions that are too weak for decision making:
"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time
but I'd have to stay overnight in the airport ...)

Examples of Probability

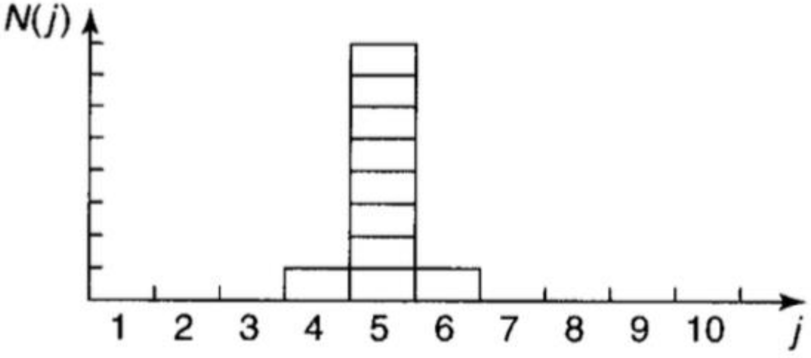
- N people (e.g., N=14) with different age *j*
    - Probability P(j) = N(j)/N
    - Most probable? Media=? Average/Mean <*j*> = ?

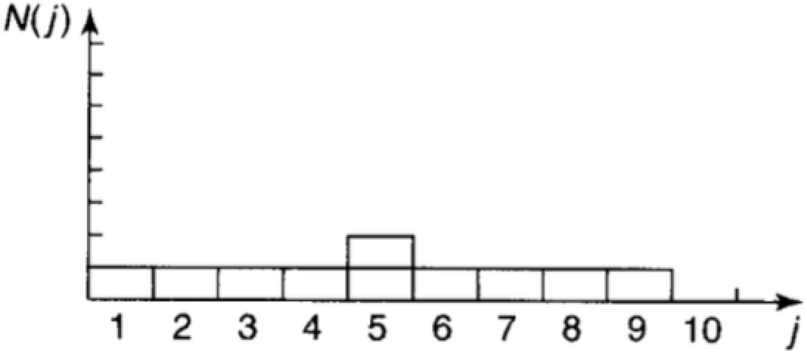
25.0 (14+25)/2 Sum/14 = 21.0


Probability Distribution

Example of Standard Deviation

- Deviation shows how spread the distribution
  - E.g, same median, average, most probable
- Variance $\sigma^2 = < (\Delta j)^2 > = <j^2> - <j>^2$
  - where $\Delta j = j - <j>$





Probability Density ρ(x)

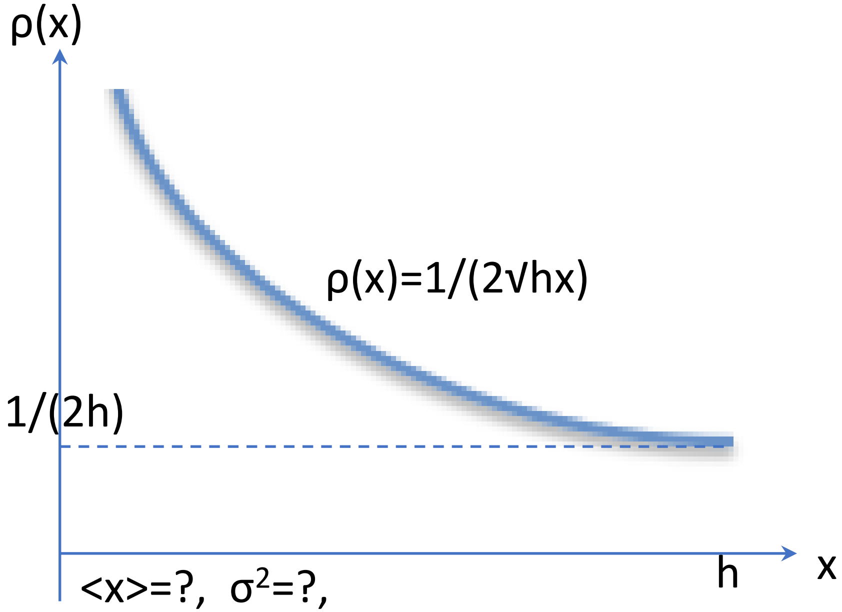
$$P_{ab} = \int_a^b \rho(x)\,dx,$$

$$\int_{-\infty}^{+\infty} \rho(x)\,dx = 1,$$

$$\langle x \rangle = \int_{-\infty}^{+\infty} x\rho(x)\,dx,$$

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x)\rho(x)\,dx,$$

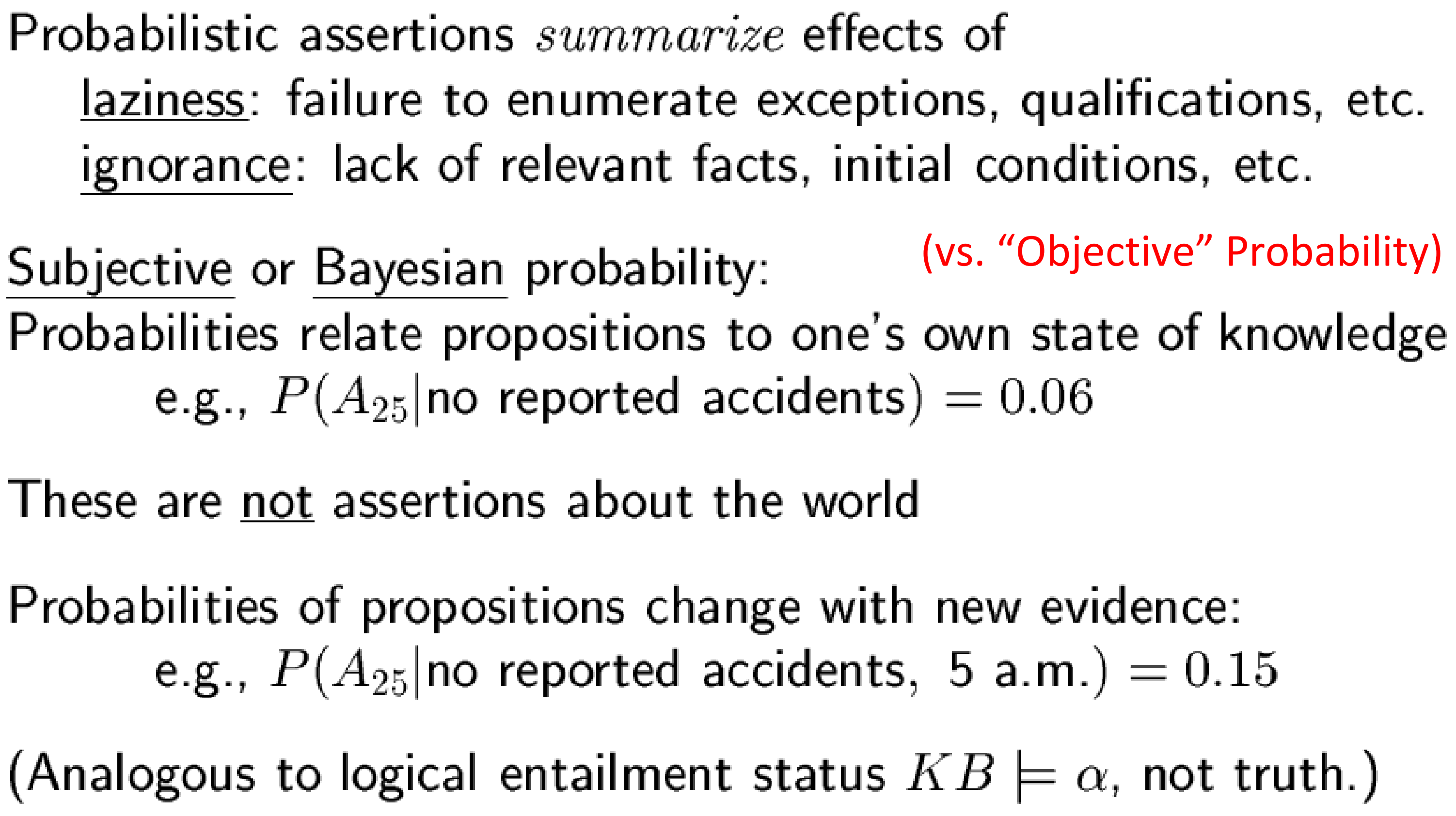
$$\sigma^2 \equiv \langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2.$$



<x>=?, σ^2=?,

Probability

Where is Probability? In the world or in your mind?
E.g., when you are throwing a dice, …



Probabilistic assertions *summarize* effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability: (vs. “Objective” Probability)

Probabilities relate propositions to one's own state of knowledge

e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$

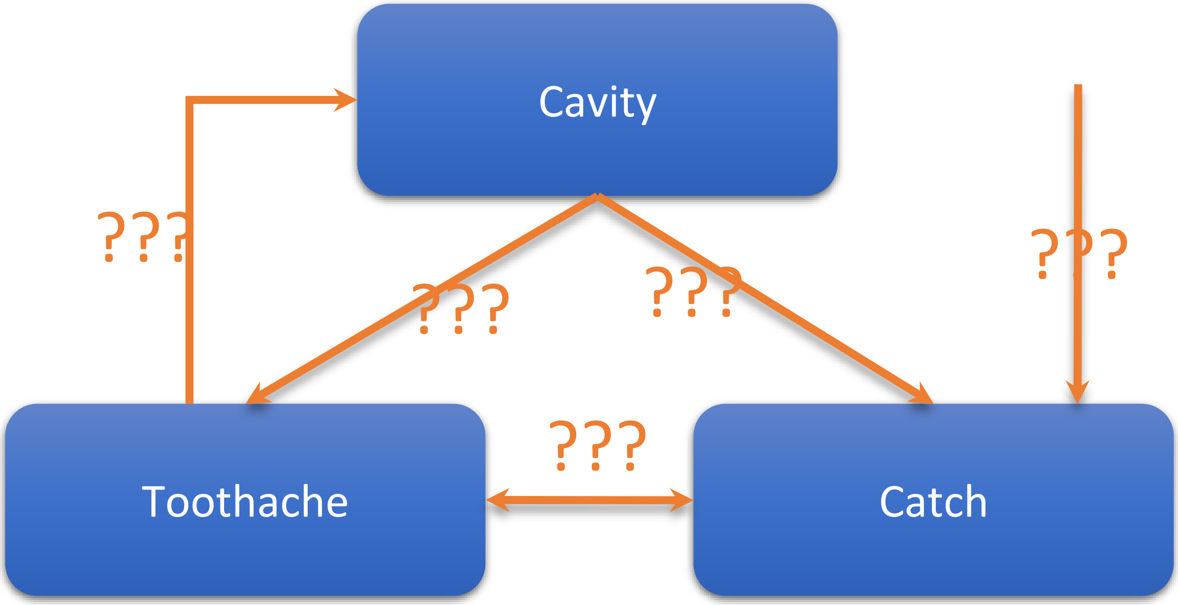
These are not assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$

(Analogous to logical entailment status $KB \models \alpha$, not truth.)

Example: Cavity-Toothache-Catch



Cavity

???

???

???

???

Toothache

???

Catch

Other facts and interactions may exist, but they are either insignificant, unknown, or irrelevant, so we leave them out.

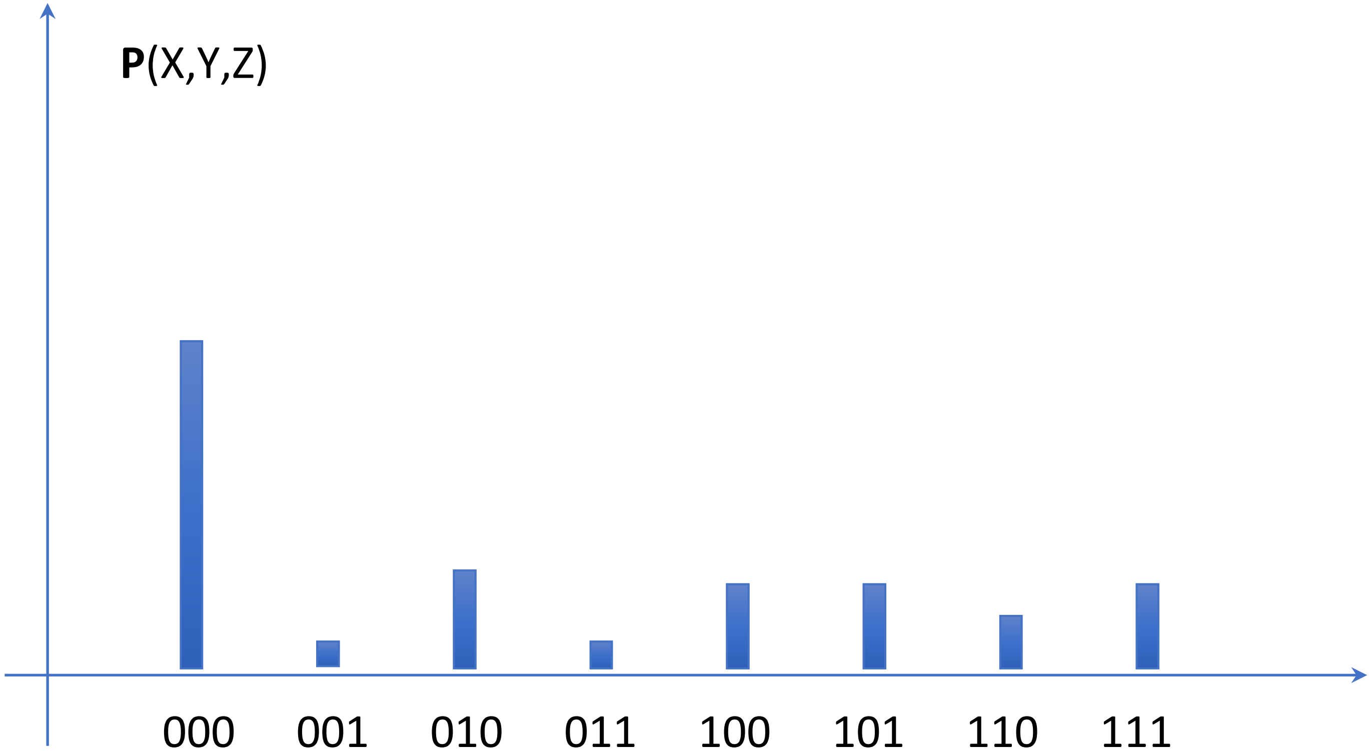
Table of "Uncertainties"

Cavity	Catch	Toothache	Logic Truth	Probability
0	0	0	{0,1}	**0.576**
0	0	1	{0,1}	**0.064**
0	1	0	{0,1}	**0.144**
0	1	1	{0,1}	**0.016**
1	0	0	{0,1}	**0.008**
1	0	1	{0,1}	**0.012**
1	1	0	{0,1}	**0.072**
1	1	1	{0,1}	**0.108**

Sum=1.000

	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

(Fully) Joint Probability Distribution



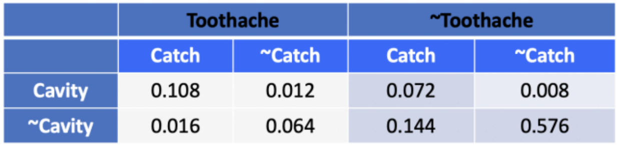
(X,Y,Z)
"Possible worlds"

X: Cavity; Y: Catch; Z: Toothache

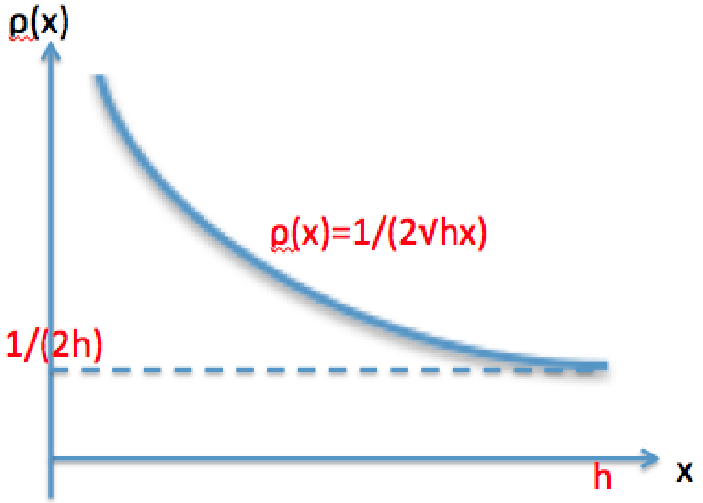
There are SO MUCH in a Distribution!

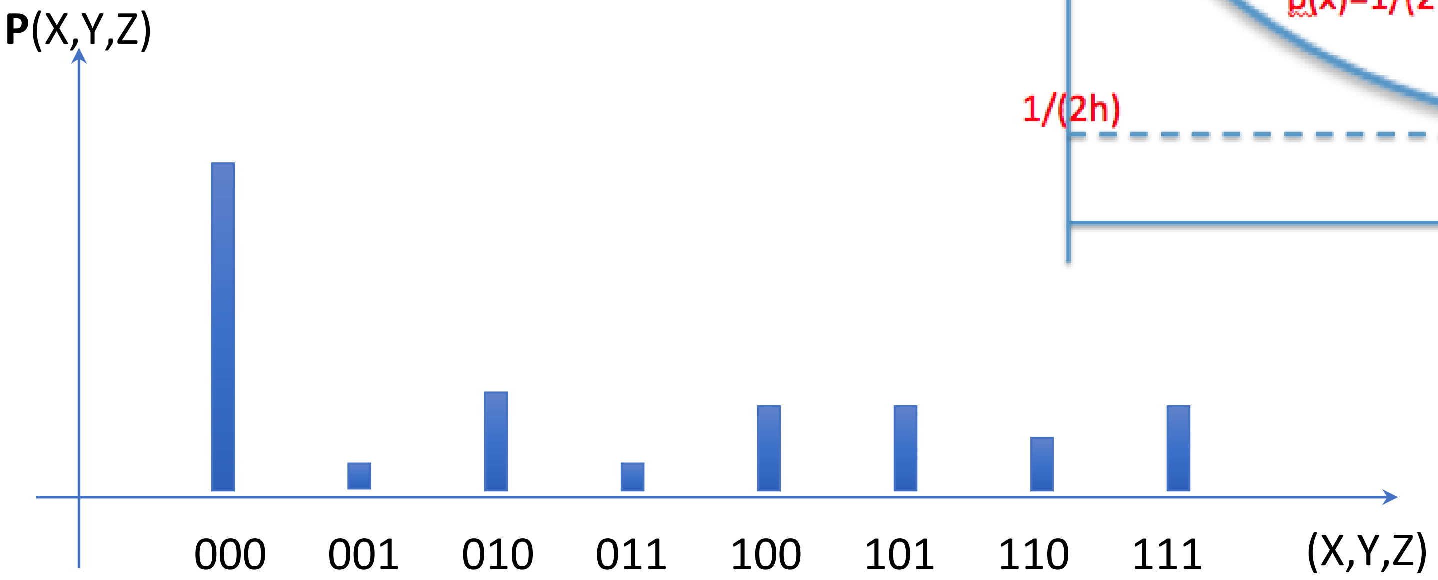
- How likely you have cavity?
- How likely you don’t have cavity?
- How likely you have toothache and cavity?
- How likely you have all three?
- How likely you have none?
- ……
- ……

Probability Distribution Models (discrete or continuous)



	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576





X: Cavity; Y: Catch; Z: Toothache

Full Joint (discrete) Distributions

A complete probability model specifies every entry in the joint distribution for all the variables $\mathbf{X} = X_1, \ldots, X_n$
I.e., a probability for each possible world $X_1 = x_1, \ldots, X_n = x_n$

(Cf. complete theories in logic.)

E.g., suppose $Toothache$ and $Cavity$ are the random variables:

	$Toothache = true$	$Toothache = false$
$Cavity = true$	0.04	0.06
$Cavity = false$	0.01	0.89

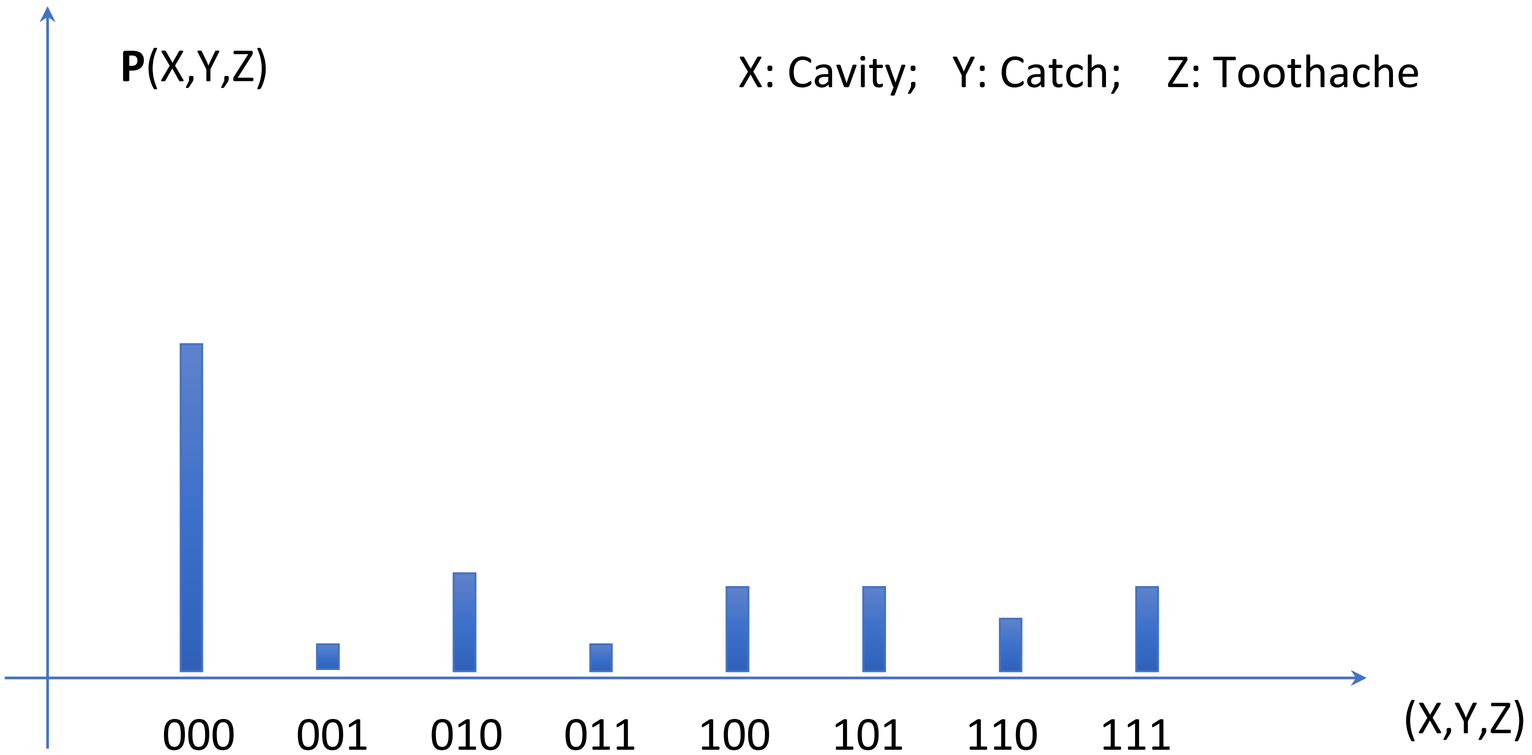
Here, how many possible worlds?

Possible worlds are mutually exclusive $\Rightarrow$ $P(w_1 \wedge w_2) = 0$
Possible worlds are exhaustive $\Rightarrow$ $w_1 \vee \cdots \vee w_n$ is $True$
hence $\Sigma_i P(w_i) = 1$

Possible worlds in Discrete Distribution

- How many possible worlds ?
  - For **P**(X,Y), where X,Y are binary variables, it is 2x2
  - For **P**(X,Y=y), it is 2x1
- In general, the number of possible worlds
  - Is the product of the size of each variable

How many “possible worlds” in a discrete probability distribution?



X: Cavity; Y: Catch; Z: Toothache

“Mutually Exclusive” & “Exhaustive”

“Possible worlds”
“value assignments”

2 x 2 x 2 = 8

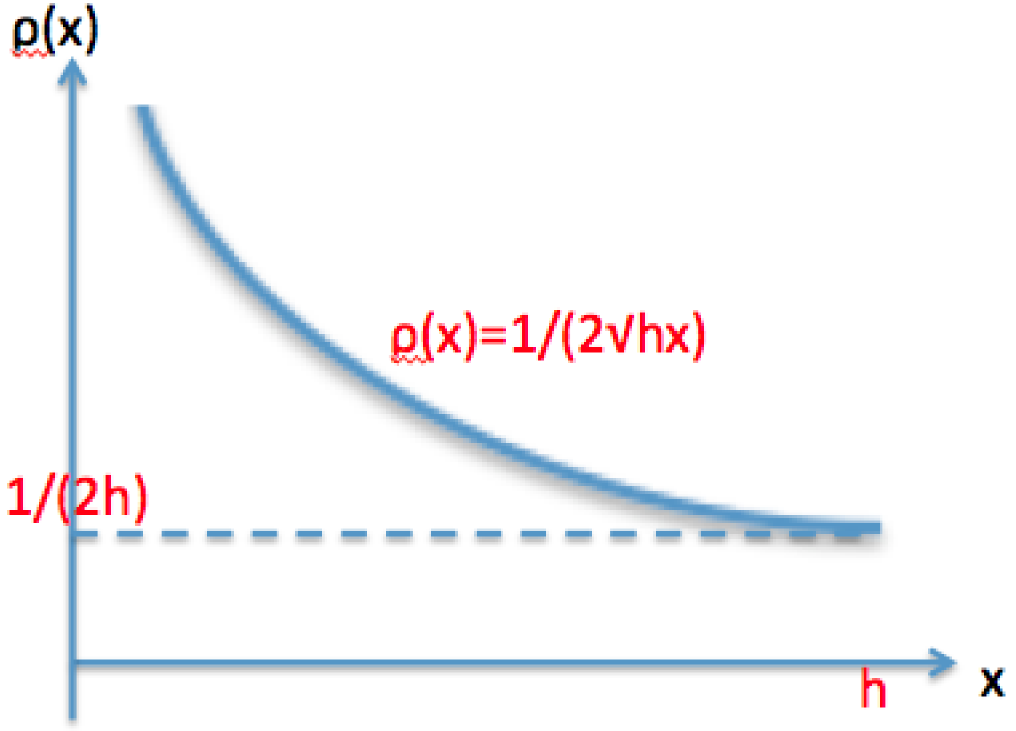
Possible worlds in these distributions?

	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

P(Toothache | Cavity) # of possible worlds = 2x2

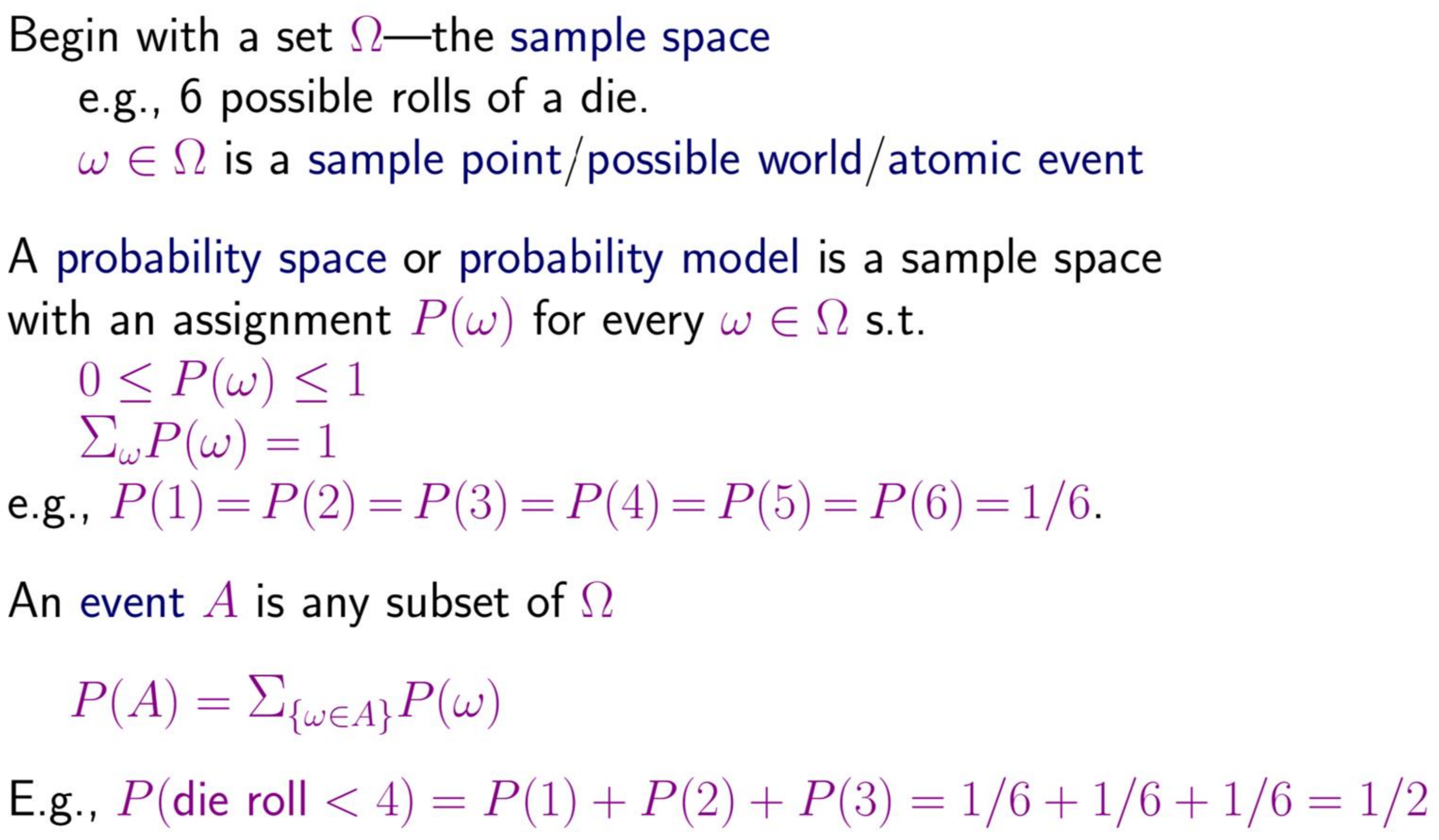
P(Toothache) # of possible worlds = 2

How many possible worlds in a continuous probability distribution?



There are infinite real numbers in [0,h]

Probability Basics (Formal)



Begin with a set Ω—the sample space

e.g., 6 possible rolls of a die.

$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

$0 \leq P(\omega) \leq 1$

$\Sigma_{\omega} P(\omega) = 1$

e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω

$$P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random Variables

A random variable is a function from sample points to some range, e.g., the reals or Booleans

e.g., $Odd(1) = true$.

P induces a probability distribution for any r.v. X:

$$P(X = x_i) = \Sigma_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Propositions

Think of a proposition as the event (set of sample points) where the proposition is true

Given Boolean random variables A and B:

event a = set of sample points where $A(\omega) = true$

event $\neg a$ = set of sample points where $A(\omega) = false$

event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$

Often in AI applications, the sample points are **defined** by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables

With Boolean variables, sample point = propositional logic model

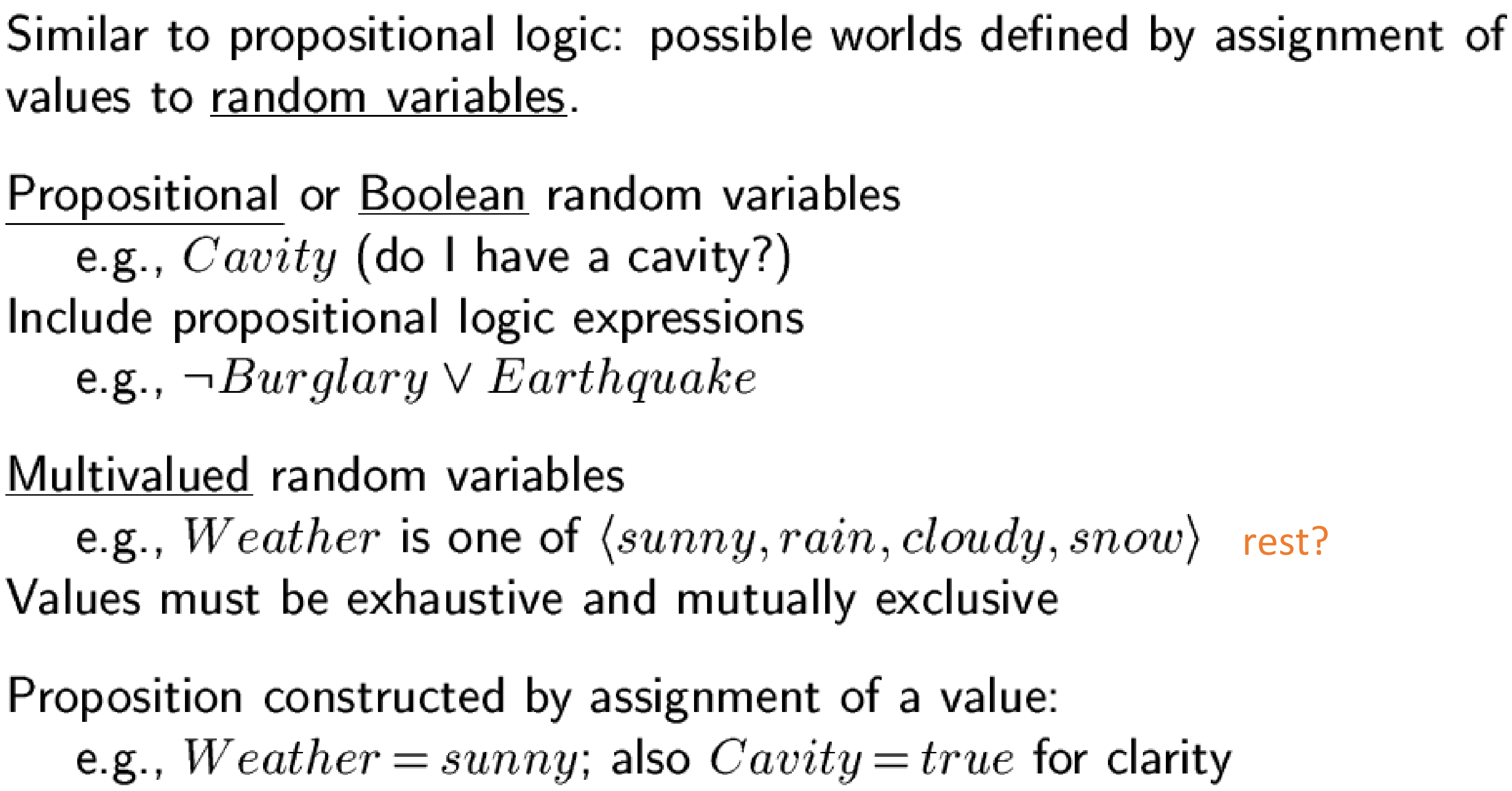
e.g., $A = true$, $B = false$, or $a \wedge \neg b$.

Proposition = disjunction of atomic events in which it is true

e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$

$\Rightarrow \; P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Syntax of Probability



Similar to propositional logic: possible worlds defined by assignment of values to random variables.

Propositional or Boolean random variables

e.g., $Cavity$ (do I have a cavity?)

Include propositional logic expressions

e.g., $\neg Burglary \lor Earthquake$

Multivalued random variables

e.g., $Weather$ is one of $\langle sunny, rain, cloudy, snow \rangle$

Values must be exhaustive and mutually exclusive

Proposition constructed by assignment of a value:

e.g., $Weather = sunny$; also $Cavity = true$ for clarity

rest?

Syntax of Probability

Prior or unconditional probabilities of propositions

e.g., $P(Cavity) = 0.1$ and $P(Weather = sunny) = 0.72$

correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of variables gives values for each possible assignment to all the variables

$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$				
$Cavity = false$				

Syntax

Conditional or posterior probabilities

e.g., $P(Cavity|Toothache) = 0.8$

i.e., given that $Toothache$ is all I know

Notation for conditional distributions:

$\mathbf{P}(Weather|Earthquake) =$ 2-element vector of 4-element vectors

If we know more, e.g., $Cavity$ is also given, then we have

$P(Cavity|Toothache, Cavity) = 1$

Note: the less specific belief *remains valid* after more evidence arrives, but is not always *useful*

New evidence may be irrelevant, allowing simplification, e.g.,

$P(Cavity|Toothache, 49ersWin) = P(Cavity|Toothache) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Syntax Examples

Cavity	Catch	Toothache	Logic Truth	Probability
0	0	0	{0,1}	**0.576**
0	0	1	{0,1}	**0.064**
0	1	0	{0,1}	**0.144**
0	1	1	{0,1}	**0.016**
1	0	0	{0,1}	**0.008**
1	0	1	{0,1}	**0.012**
1	1	0	{0,1}	**0.072**
1	1	1	{0,1}	**0.108**

Exercises:

P(Cavity=1)=?, **P**(Cavity)=?, *P*(Catch=0)=?, **P**(Catch)=?, *P*(Toothache=1)=?, …

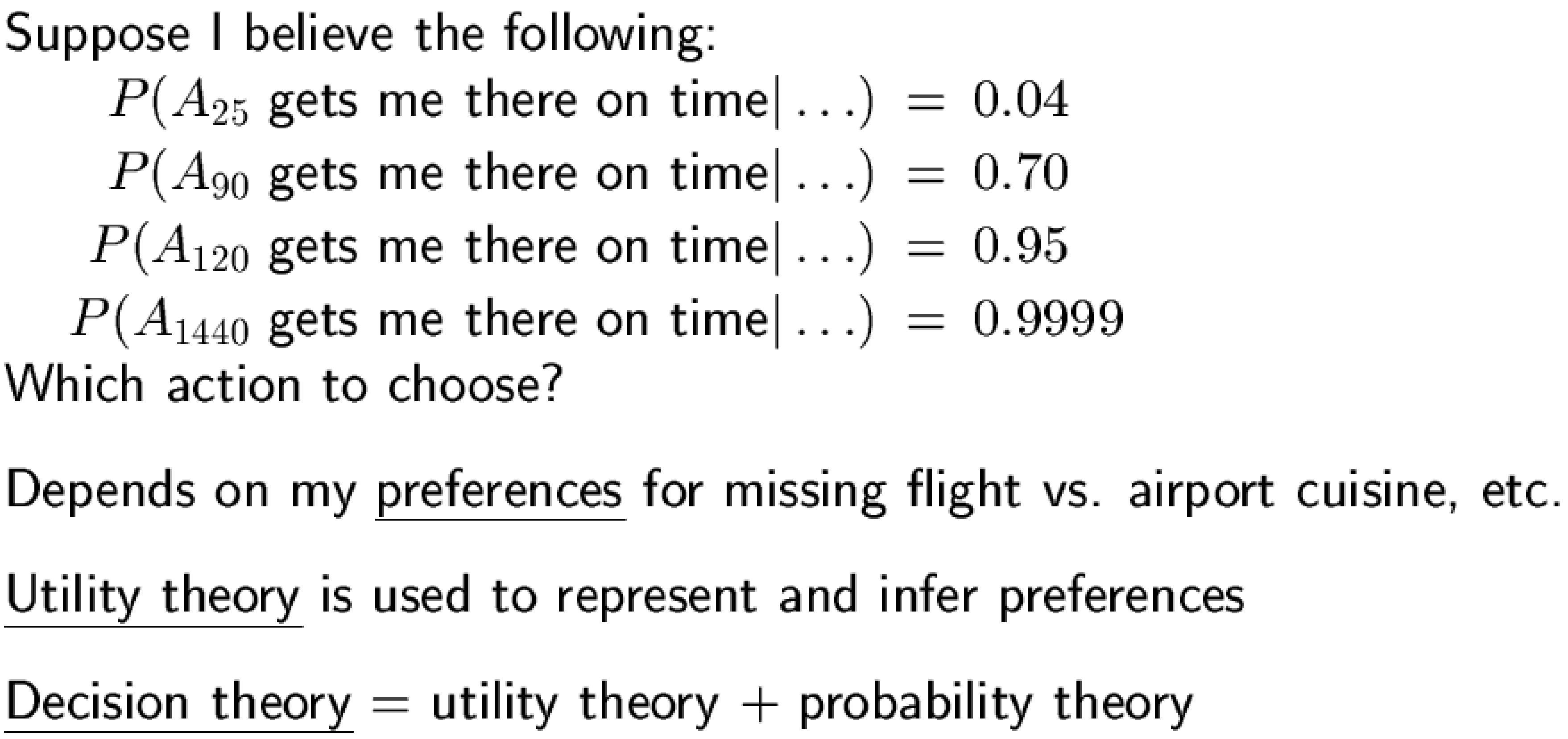
P(Cavity=1, Catch=1, Toothache=1)=?, **P**(Cavity, Catch, Toothache)=?

P(Cavity=0, Catch=0)=?, **P**(Cavity, Catch=0)=?

Two Key Elements in Probability

- Probability Distribution Model
  - Variables, Value assignments (possible worlds)
  - Represented as a table or a graph
- Inferences that can be made from the model
  1. Sum rule
  2. Product rule
  3. Conditional
  4. Normalization
  5. Marginalization

Making Decisions under Uncertainty



Suppose I believe the following:

$$\begin{aligned} P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\ P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\ P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\ P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999 \end{aligned}$$

Which action to choose?

Depends on my preferences for missing flight vs. airport cuisine, etc.

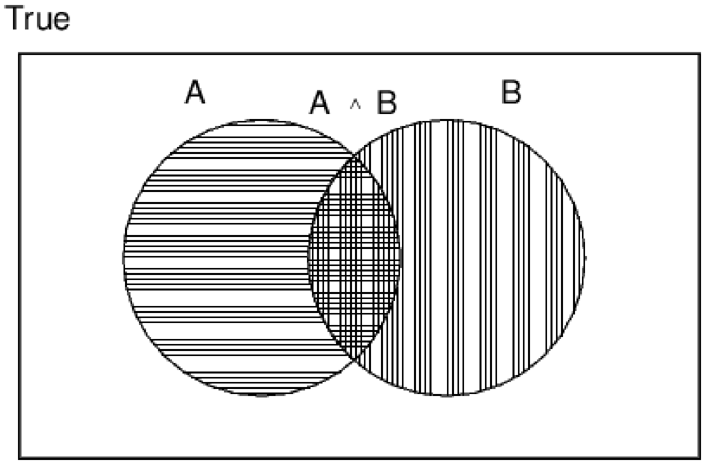
Utility theory is used to represent and infer preferences

Decision theory = utility theory + probability theory

Axioms of Probability

For any propositions A, B

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$ and $P(False) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

Two Inference Rules (1)(2)

- Rules: You only need to remember two ☺
  - Rule (1) Sum Rule
    - **$P(A|B) + P(\sim A|B) = 1$**
  - Rule (2) Product Rule for P(A^B)
    - $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$
    - Or more general: $P(AB|C) = P(A|C)\, P(B|AC) = P(B|C)\, P(A|BC)$

Remember these two rules!

- Notations
  - Variable X, value x_i, $P(X=x_i)$,
  - **P**(X) denotes for all values of X as shown in the table, aka “joint probability distribution”
  - Mixed variables and values: **P**(X,Y), **P**(X, y)

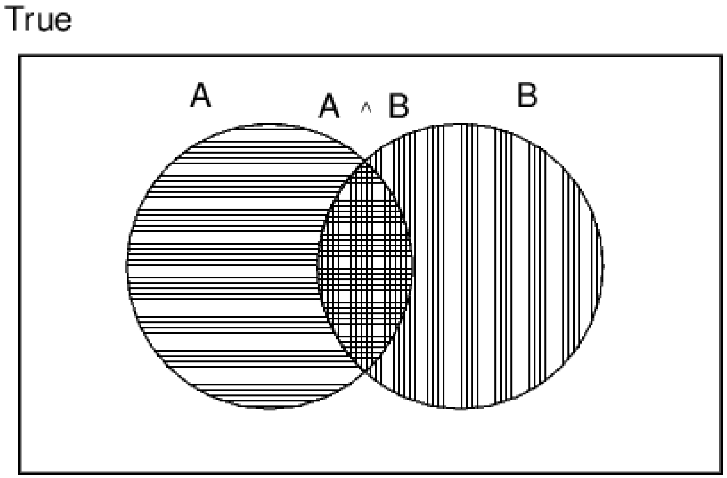
Inference for P(A v B)

For any propositions A, B

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$ and $P(False) = 0$
3. $\boxed{P(A \vee B) = P(A) + P(B) - P(A \wedge B)}$

Can you prove it by Sum Rule?

$P(AVB) = P(\neg\neg(AVB)) =$
$P(\neg(\neg A \neg B)) = 1 - P(\neg A \neg B)$



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

(3) Conditional Probability P(A|B)

Definition of conditional probability:

$$P(A|B) = \frac{P(A \land B)}{P(B)} \text{ if } P(B) \neq 0$$

Derivable from Product Rule

Product rule gives an alternative formulation:

$$P(A \land B) = P(A|B)P(B) = P(B|A)P(A)$$

Product Rule

A general version holds for whole distributions, e.g.,

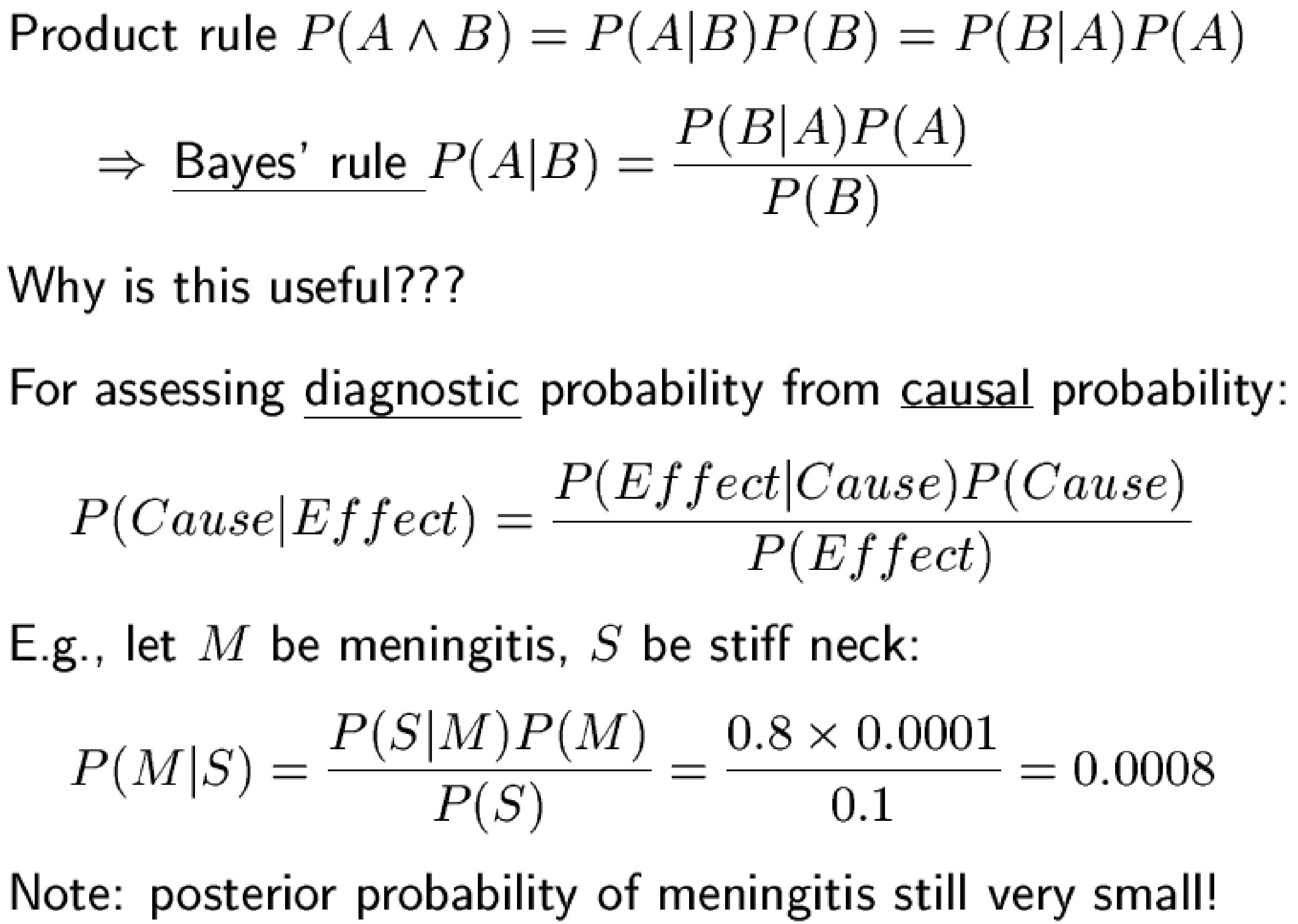
$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, *not* matrix mult.)

Chain rule is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n_1}|X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \Pi_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})\end{aligned}$$

(3) Bayes' Rule: derivable from product



Product rule $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$

$$\Rightarrow \text{ Bayes' rule } P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Why is this useful???

For assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

(4) Normalization

("make the distribution sum to 1")

Suppose we wish to compute a posterior distribution over A given $B = b$, and suppose A has possible values $a_1 \ldots a_m$

We can apply Bayes' rule for each value of A:

$$P(A = a_1|B = b) = P(B = b|A = a_1)P(A = a_1)/P(B = b)$$

$$\ldots$$

$$P(A = a_m|B = b) = P(B = b|A = a_m)P(A = a_m)/P(B = b)$$

Adding these up, and noting that $\Sigma_i P(A = a_i|B = b) = 1$:

$$1/P(B = b) = 1/\Sigma_i P(B = b|A = a_i)P(A = a_i) \quad = \alpha$$

This is the normalization factor, constant w.r.t. i, denoted α:

$$\boxed{\mathbf{P}(A|B = b) = \alpha \mathbf{P}(B = b|A)\mathbf{P}(A) \; = \alpha \, \mathbf{P}(B{=}b \wedge A)}$$

Typically compute an unnormalized distribution, normalize at end

e.g., suppose $\mathbf{P}(B = b|A)\mathbf{P}(A) = \langle 0.4, 0.2, 0.2 \rangle$

then $\mathbf{P}(A|B = b) = \alpha \langle 0.4, 0.2, 0.2 \rangle = \frac{\langle 0.4, 0.2, 0.2 \rangle}{0.4 + 0.2 + 0.2} = \langle 0.5, 0.25, 0.25 \rangle$

Normalization Example

Cavity	Catch	Toothache	Logic Truth	Probability
0	0	0	{0,1}	**0.576**
0	0	1	{0,1}	**0.064**
0	1	0	{0,1}	**0.144**
0	1	1	{0,1}	**0.016**
1	0	0	{0,1}	**0.008**
1	0	1	{0,1}	**0.012**
1	1	0	{0,1}	**0.072**
1	1	1	{0,1}	**0.108**

P(Toothache | Cavity=1) =
α**P(**Toothache ^ Cavity=1) =
α <0.108+0.012, 0.072+0.008> =
α <0.12, 0.08> =
<0.12/0.2, 0.08/0.2> =
<0.6, 0.4>

	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

Normalization once for all?

- If you use the inference rules properly, then normalization would be preserved
- For the last example, we may use the Product Rule
  - **P**(Toothache|Cavity=1) = α **P**(Cavity=1 ^ Toothache)
  - **P**(Toothache|Cavity=1) = **P**(Cavity=1^Toothache) / **P**(Cavity=1)
  - The results will be the same
  - Try it

(5) Marginalization

Introducing a variable as an extra condition:

$$P(X|Y) = \Sigma_z P(X|Y, Z = z) P(Z = z|Y)$$

Why? To plug in the numbers we know

Intuition: often easier to assess each specific circumstance, e.g.,

$$\begin{aligned} &P(RunOver|Cross) \\ &\quad = P(RunOver|Cross, Light = green) P(Light = green|Cross) \\ &\quad + P(RunOver|Cross, Light = yellow) P(Light = yellow|Cross) \\ &\quad + P(RunOver|Cross, Light = red) P(Light = red|Cross) \end{aligned}$$

When Y is absent, we have summing out or marginalization:

$$P(X) = \Sigma_z P(X|Z = z) P(Z = z) = \Sigma_z P(X, Z = z)$$

Why? To plug in the known numbers

In general, given a joint distribution over a set of variables, the distribution over any subset (called a marginal distribution for historical reasons) can be calculated by summing out the other variables.

Marginalization Example

	Toothache		~Toothache	
	Catch	~Catch	Catch	~Catch
Cavity	0.108	0.012	0.072	0.008
~Cavity	0.016	0.064	0.144	0.576

P(Toothache|Cavity) = Σ **P**(Toothache ^ Catch | Cavity)

// Insert Catch

= **P**(Toothache, Catch=1 | Cavity)
+ **P**(Toothache, Catch=0 | Cavity)

// Insert Catch & Cavity

P(Toothache) = Σ **P**(Toothache ^ Cavity ^ Catch)

= **P**(Toothache, Cavity=1, Catch=0)
+ **P**(Toothache, Cavity=0, Catch=0)
+ **P**(Toothache, Cavity=1, Catch=1)
+ **P**(Toothache, Cavity=0, Catch=1)

Margin from Joint Distributions

Typically, we are interested in **P(Y|E)**

the posterior joint distribution of the query variables $\mathbf{Y}$

given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$ // where **X** denotes all variables

Then the required summation of joint entries is done by summing out the hidden variables:

These are known from the joint distribution

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E}=\mathbf{e}) = \alpha \Sigma_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E}=\mathbf{e}, \mathbf{H}=\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables **X**

Obvious problems:

1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Example

- Assume the full joint distribution (like a truth table!)

	toothache		~toothache	
	catch	~catch	catch	~catch
cavity	0.108	0.012	0.072	0.008
~cavity	0.016	0.064	0.144	0.576

Y = Cavity: we want to know whether we have a cavity

E = Toothache: we know we have a toothache

H = Catch: we don’t know whether probe would catch or not

$$P(Y|e) = \alpha P(Y, e) = \alpha \sum_{h} P(Y, e, h)$$

Two Key Elements in Probability

✓ - Probability Distribution Model
  - Variables, Value assignments (possible worlds)
  - Represented as a table or a graph

✓ - Inferences that can be made from the model
  1. Sum rule
  2. Product rule
  3. Conditional
  4. Normalization
  5. Marginalization

- Now we can do any general inference we like!

General Inference for any Proposition

- 0 <= P(A) <= 1
- P(False) = 0;   P(True) = 1
- P(A ^ B) = P(AB) = P(A) P(B|A) = P(B) P(A|B)
- P(A v B) = P(A) + P(B) – P(A ^ B)

General Inference for Sentences

1) For any proposition ϕ defined on the random variables w_i

$\phi(w_i)$ is true or false

2) ϕ is equivalent to the disjunction of w_is where $\phi(w_i)$ is true

Hence $\boxed{P(\phi)} = \underline{\Sigma_{\{w_i:\ \phi(w_i)\}} P(w_i)}$

Sum Rule is for disjunction

I.e., the unconditional probability of any proposition is computable as the sum of entries from the full joint distribution

Conditional probabilities can be computed in the same way as a ratio:

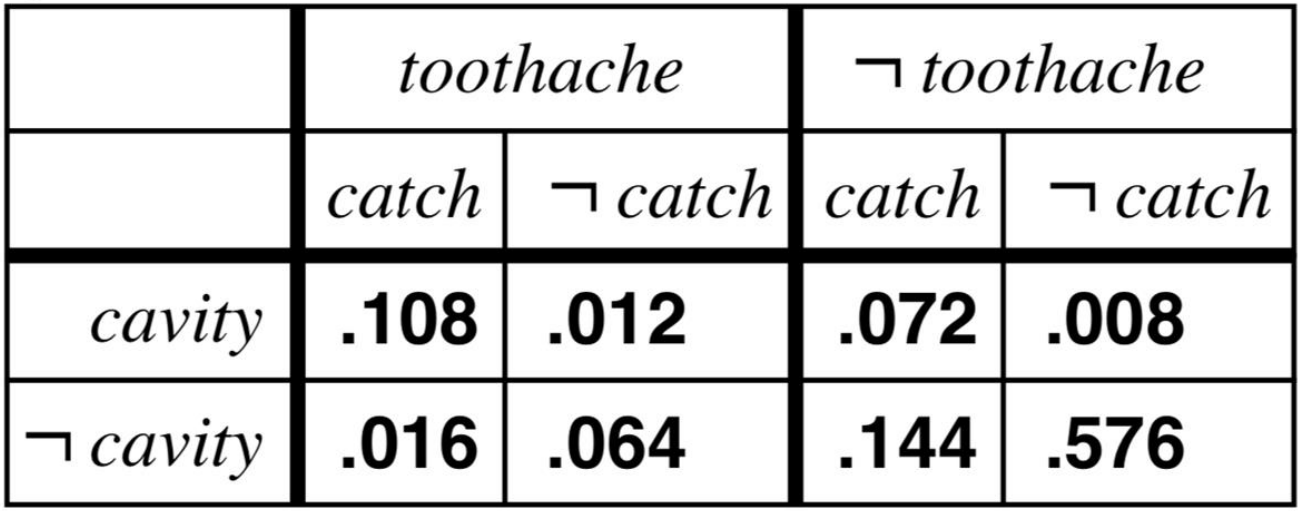
$$\boxed{P(\phi|\xi)} = \frac{P(\phi \wedge \xi)}{P(\xi)}$$

E.g.,

$$P(Cavity|Toothache) = \frac{P(Cavity \wedge Toothache)}{P(Toothache)} = \frac{0.04}{0.04 + 0.01} = 0.8$$

Inference by Summation (enumeration)

Start with the joint distribution:



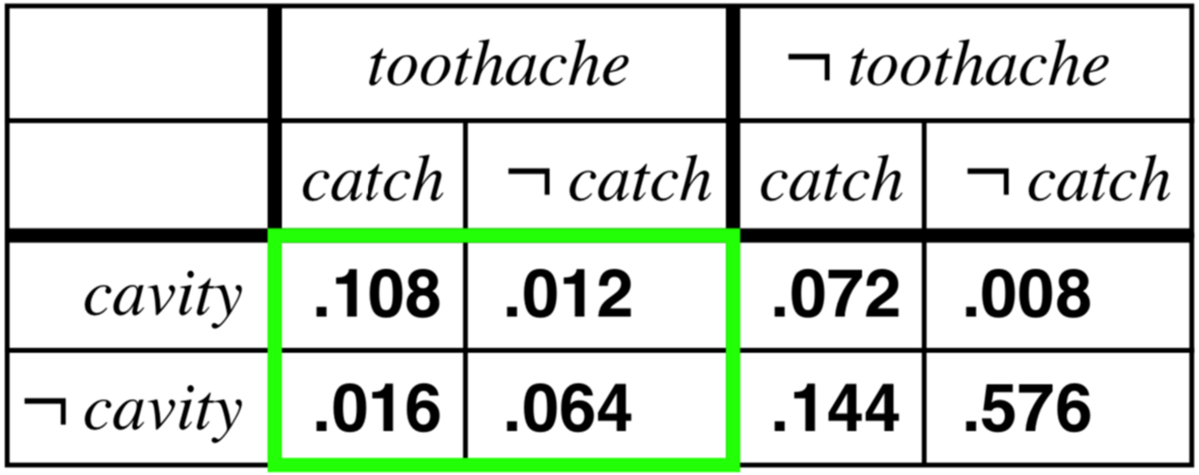
	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$$

Inference by summation/enumeration

Start with the joint distribution:



	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

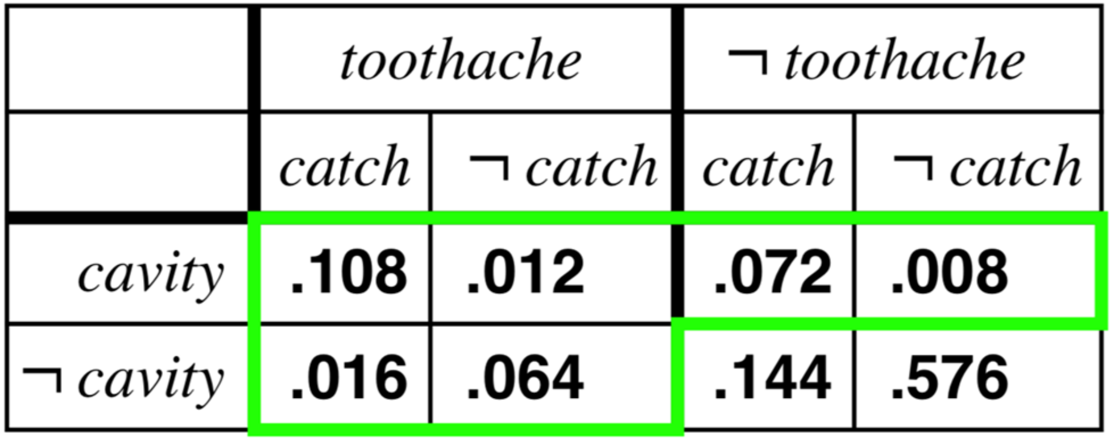
For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by summation/enumeration

Start with the joint distribution:



	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

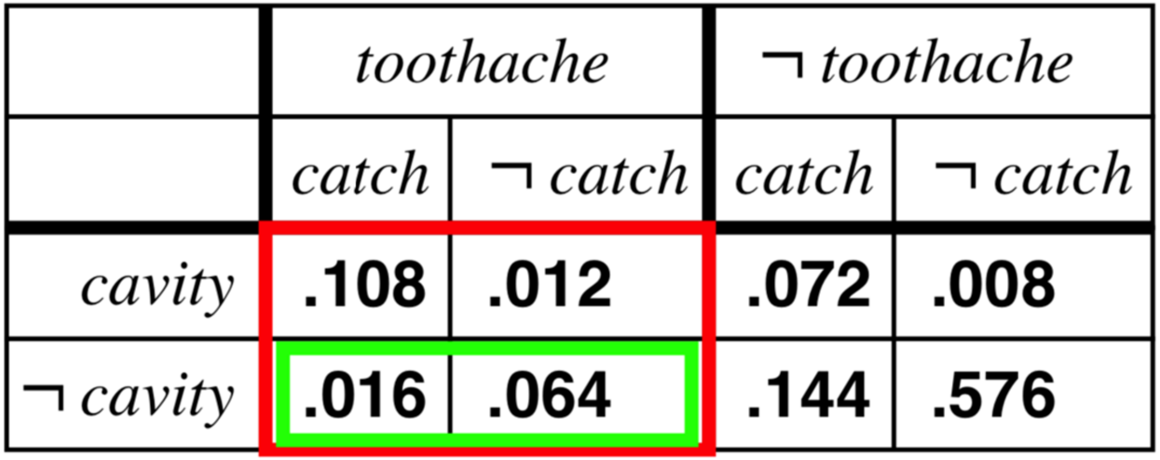
For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega\models\phi}P(\omega)$$

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by Conditional Probability

Start with the joint distribution:



	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

$$
\begin{aligned}
P(\neg cavity|toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\
&= \frac{0.016+0.064}{0.108+0.012+0.016+0.064} = 0.4
\end{aligned}
$$

Example for General Inference

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a normalization constant α

$$
\begin{aligned}
&\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache) \\
&\quad = \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&\quad = \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&\quad = \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}
$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

Inference of Probability vs Logic

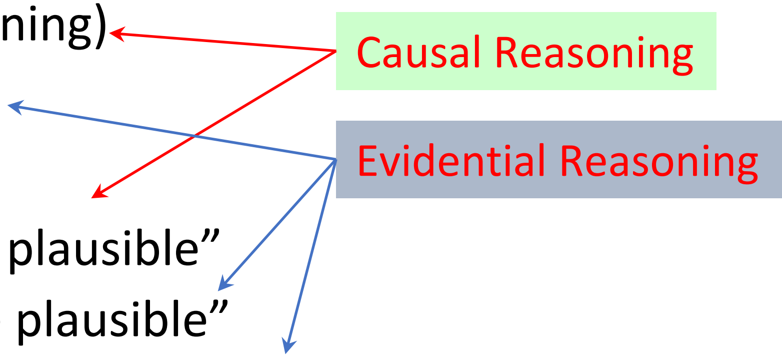
- $P(AB|C) = P(A|C)P(B|AC) = P(B|C)P(A|BC)$

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

- Probability can do much more than logics
  - Deductive reasoning:
    - If A->B and A, then B
    - If A->B and ~B, then ~A
    - If A->B and B, then “A become more plausible”
  - Inductive reasoning:
    - If A->B and ~A, then “B become less plausible”
    - If A->”B becomes more plausible” and B, then “A become more plausible”

Probability vs. Logic

- Probability can do much more than logics
  - Deductive reasoning (logic and probability):
    - If A->B and A, then B (forward reasoning)
    - If A->B and ~B, then ~A (resolution)
  - Inductive reasoning (probability):
    - If A->B and ~A, then “B become less plausible”
    - If A->B and B, then “A become more plausible”
    - If A->”B becomes more plausible” and B, then “A become more plausible”

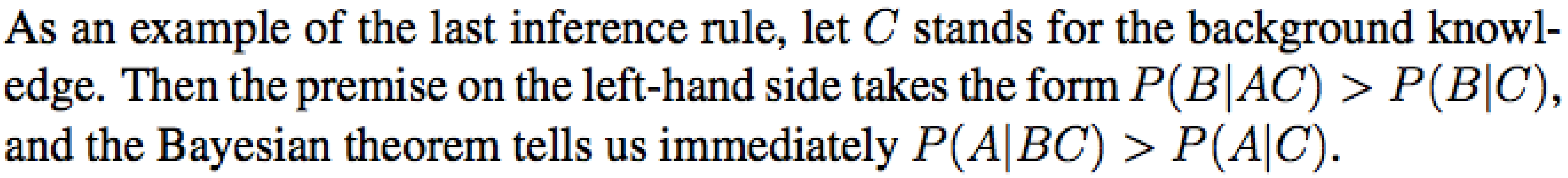


Causal Reasoning

Evidential Reasoning

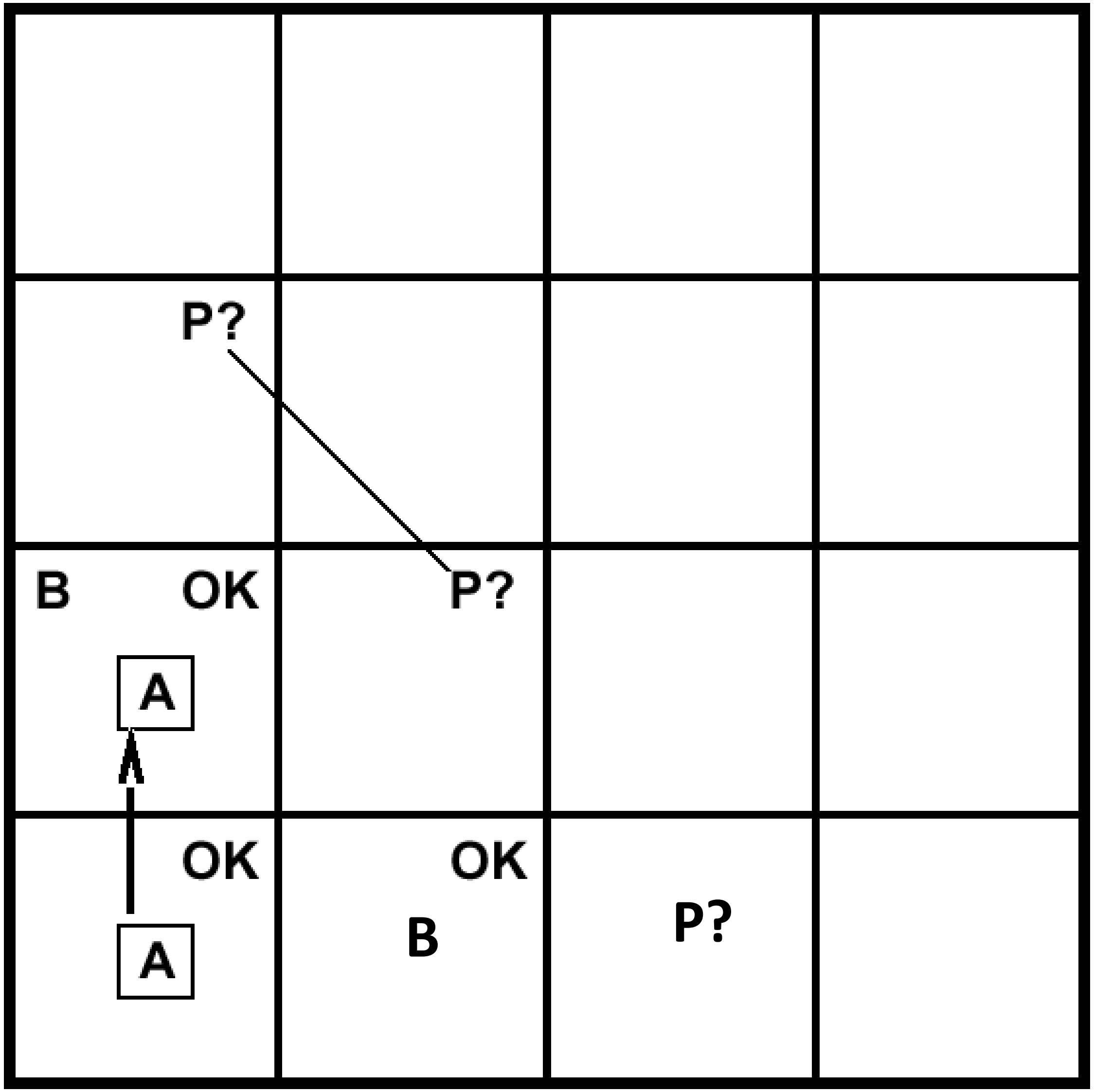
Why Can Probability Do More?

- Can you prove the following using probability?
  1. If A->B and A, then B
  2. If A->B and ~B, then ~A
  3. If A->B and B, then “A become more plausible”
  4. If A->B and ~A, then “B become less plausible”
  5. If A->”B becomes more plausible” and B,
     then “A become more plausible”
- These proofs are for your exercises at home



As an example of the last inference rule, let C stands for the background knowledge. Then the premise on the left-hand side takes the form $P(B|AC) > P(B|C)$, and the Bayesian theorem tells us immediately $P(A|BC) > P(A|C)$.

Inferences in the Wumpus world



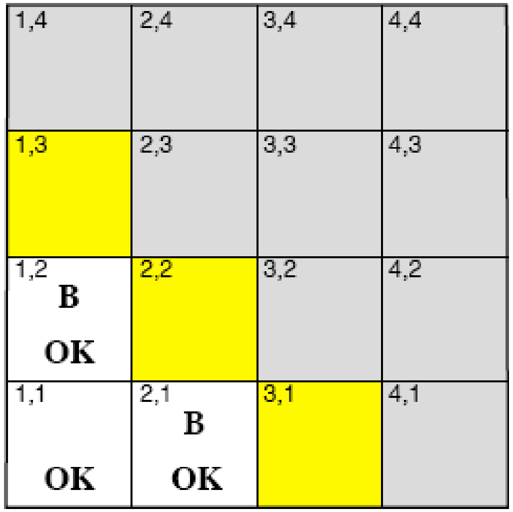
A= Agent
B= Breeze
S= Smell
P= Pit
W= Wumpus
OK = Safe
V = Visited
G = Glitter

Which one has a PIT?

Inferences in the Wumpus world

- Logic can only guess randomly which of [1,3],[2,2],[3,1] has a pit
- Using probability you can calculate which one is more likely have a pit than others

Details in Wumpus World



1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

- Any of [1,3], [2,2] or [3,1] may have a pit, but which one is riskier/safer to try, assuming pits are relatively rare?
  - Either [1,3] or [3,1] should be less risky than [2,2] because most probable pattern given the evidence is one pit at [2,2]
- Need a probabilistic rather than a logical model

P_{ij} = *true* iff [i,j] contains a pit

B_{ij} = *true* iff [i,j] is breezy

} *Boolean random variables*

For simplicity, will include only $B_{12}, B_{21}, P_{11}, \ldots P_{44}$ in probability model

Construct a Probability Model (for the Wumpus world)

- Construct the probability model
  - Select your random (binary) variables: $B_{12},B_{21},P_{11},\ldots,P_{44}$ (see the last slide)
  - Construct the full joint probability distribution (exclusive, exhaustive, sum=1)

$$\mathbf{P}(B_{12},B_{21},P_{11},\ldots,P_{44})$$

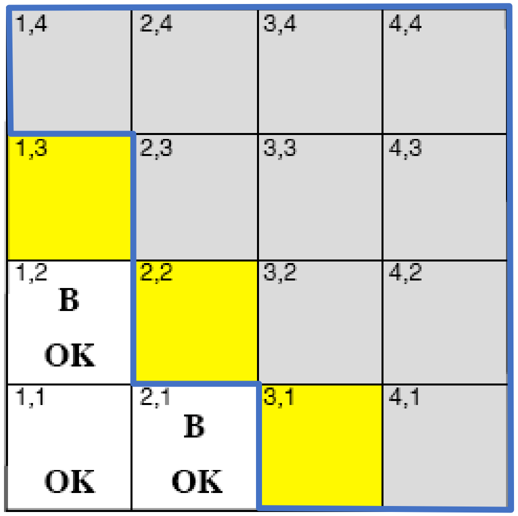
- Use the model to compute the probabilities in interest
  - $\mathbf{P}(P_{11},\ldots,P_{44})$ is the probability of pit distribution (uniformly)
    - Assume that pits are placed randomly with P(p_{ij})=.2 for all *i,j*
    - Assume that pits are placed independently
    - Then we have

$$\mathbf{P}(P_{11},\ldots,P_{44}) = \mathbf{P}(P_{11})\bullet\mathbf{P}(P_{12})\bullet\bullet\bullet\bullet\mathbf{P}(P_{44})$$

  - $\mathbf{P}(B_{12},B_{21} \mid P_{11},\ldots,P_{44})$ is the probability of breeze at [1,2] [21].
  - We can compute this using the product rule

$$\mathbf{P}(B_{12},B_{21} \mid P_{11},\ldots,P_{44}) = \mathbf{P}(B_{12},B_{21},P_{11},\ldots,P_{44}) / \mathbf{P}(P_{11},\ldots,P_{44})$$

Observations and Query



- Have observed these facts:

  $b = b_{12} \wedge b_{21}$

  $known = {\sim}p_{11} \wedge {\sim}p_{12} \wedge {\sim}p_{21}$

- Our query is: $\mathbf{P}(P_{13} \mid known, b)$
  - Also want to query P_{22} and P_{31}
- Define *unknown* = the set of P_{ij}s other than P_{13} & *known*
- For inference by enumeration, we have

  $\mathbf{P}(P_{13} \mid known, b) = \alpha \, \Sigma_{unknown} \mathbf{P}(P_{13}, unknown, known, b)$

  Grows exponentially with number of squares

Conditional Independence to the Rescue

- **Basic Insight**
  - If partition *unknown* variables into *fringe* and *other*, then observed breezes are conditionally independent of *other* variables, given *known*, *query* and *fringe* variables
- Reformulate query into usable form
  - If *Unknown* = *Fringe* or *Other,* then

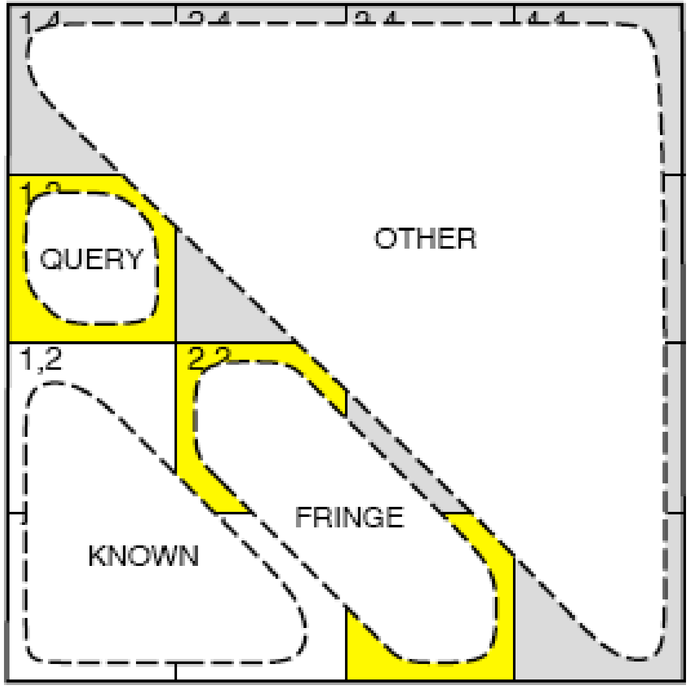
    $\mathbf{P}(b \mid P_{13}, Known, Unknown) = \mathbf{P}(b \mid P_{13}, Known, Fringe)$
  - Use this to convert

    $\mathbf{P}(P_{13} \mid known, b) = \alpha \Sigma_{unknown} \mathbf{P}(P_{13}, unknown, known, b)$

  Into (via sequence of transformations)

  $\alpha' \, \mathbf{P}(P_{13}) \, \Sigma_{fringe} \mathbf{P}(b \mid known, P_{13}, fringe) \, \mathbf{P}(fringe)$

  [where $\alpha' = \alpha P(known)$]
  - *Greatly reduces computation*



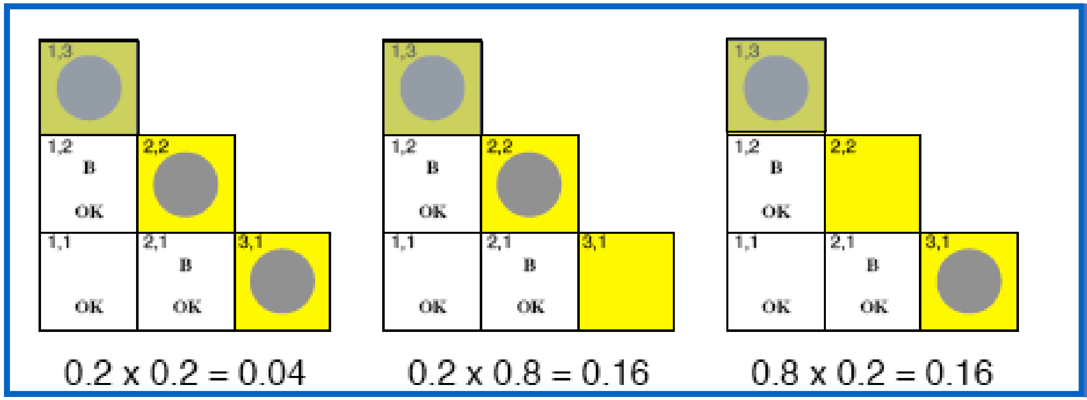
Results

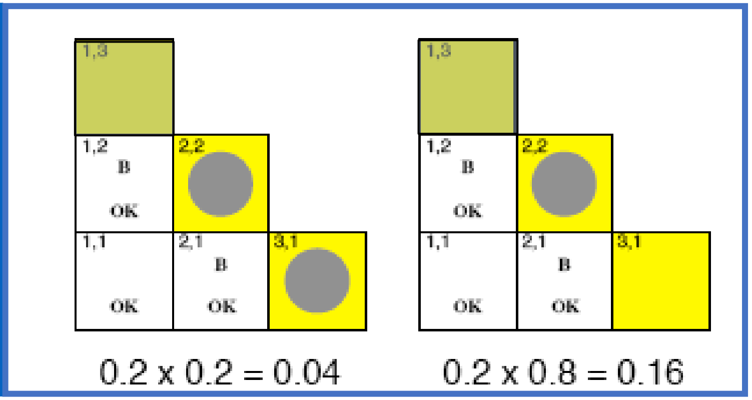
$\mathbf{P}(P_{13} \mid known, b) = \alpha' \, \mathbf{P}(P_{13}) \, \Sigma_{fringe} \mathbf{P}(b \mid known, P_{13}, fringe) \, \mathbf{P}(fringe)$

- $\mathbf{P}(b \mid known, P_{13}, fringe)$ is 0 or 1 depending on if fringe is consistent with b
- P(*fringe*) is shown in figure for situations in which fringe is consistent with b
- Now need to partition action events according to whether P_{13} is *true* or *false*

$$= \alpha'[.2(.04 + .16 + .16), 0.8(.04 + .16)] = \alpha'[.072, .16] \; [.31, .69]$$

$\mathbf{P}(P_{22} \mid known, b) = [.86, .14]$





P_{13} = *true*

Alternative: Using Bayesian Rule!

- Can compute it nicely by comparing:
  - $\mathbf{P}(P_{13} \mid known,b)$
  - $\mathbf{P}(P_{31} \mid known,b)$
  - $\mathbf{P}(P_{22} \mid known,b)$
- **Let's use Bayesian Rule**
  - $\mathbf{P}(P_{13}|known,b)=\mathbf{P}(P_{13})\mathrm{P}(known,b|P_{13})/\mathrm{P}(known,b)$
  - $\mathbf{P}(P_{22}|known,b)=\mathbf{P}(P_{22})\mathrm{P}(known,b|P_{22})/\mathrm{P}(known,b)$
  - $\mathbf{P}(P_{31}|known,b)=\mathbf{P}(P_{31})\mathrm{P}(known,b|P_{31})/\mathrm{P}(known,b)$
- Since $\mathbf{P}(P_{13})=\mathbf{P}(P_{22})=\mathbf{P}(P_{31})$, only need compare
  - $\mathrm{P}(known,b|P_{13})$
  - $\mathrm{P}(known,b|P_{22})$ // this is larger than the other two, why?
  - $\mathrm{P}(known,b|P_{31})$

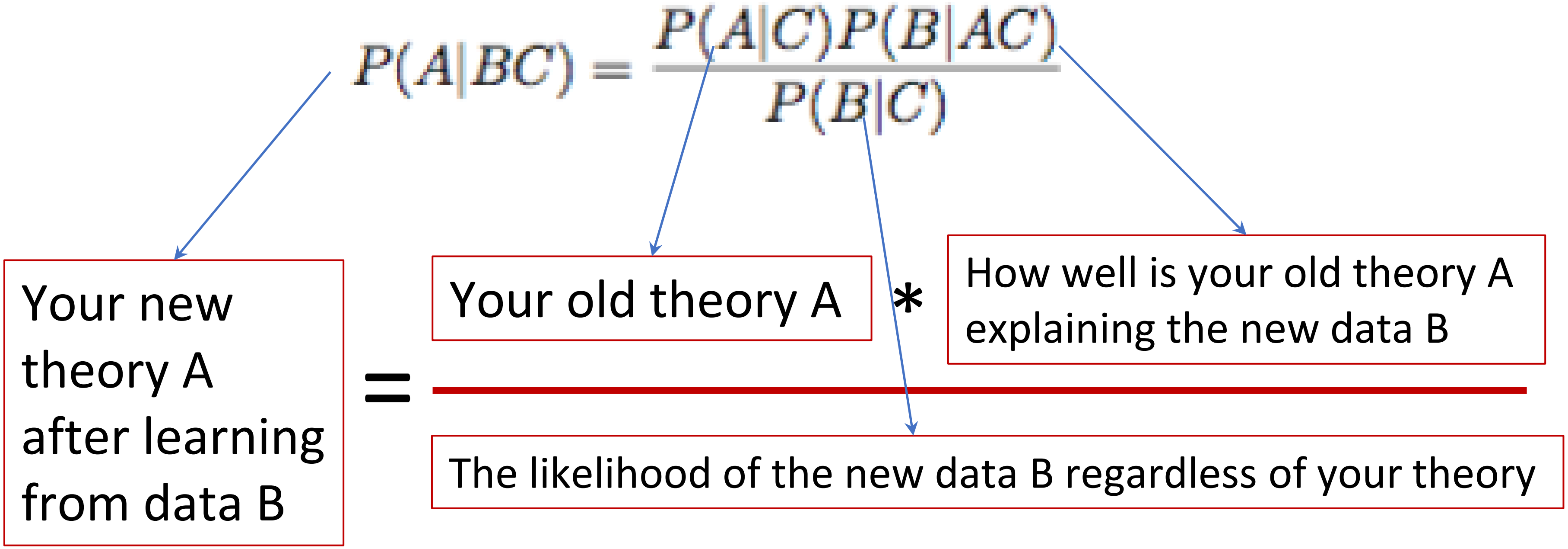
Compare the three choices

- Using chain rules
  - $P(known,b|p_{13})=P(known|b,p_{13})P(b|p_{13})$
  - $P(known,b|p_{22})=P(known|b,p_{22})P(b|p_{22})$
  - $P(known,b|p_{31})=P(known|b,p_{31})P(b|p_{31})$
- The first terms are equal
  - because pits are independent from each other
- The second term:
  - $P(b_{12}\wedge b_{21}|p_{13}) < 1$
  - $P(b_{12}\wedge b_{21}|p_{22}) = 1$
  - $P(b_{12}\wedge b_{21}|p_{13}) < 1$

  - *Where known=*$\sim p_{11}\wedge\sim p_{12}\wedge\sim p_{21}$, and $b=b_{12}\wedge b_{21}$

Bayesian Rule for Learning

- Bayesian Rule is a powerful tool for learning
  - Let A be “your theory”; B be the “new data”; and C be the “background information”



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

Your new theory A after learning from data B

=

Your old theory A

*

How well is your old theory A explaining the new data B

The likelihood of the new data B regardless of your theory

Summary

- Probability is a rigorous formalism for uncertain knowledge
    - Conditional probabilities enable reasoning with uncertain evidence
- A *full joint probability distribution* specifies the probability of every *atomic event*
    - Queries answerable by summing over probabilities of atomic events
- Bayesian theorem/rule provides the basis for the most modern diagnostic reasoning in AI
    - Converts uncertain causal information into diagnostic conclusions
- For nontrivial domains, we must find a way to reduce the size of the joint distribution
    - *(Conditional) independence* provides the tools
      (next lecture)