

# REVIEW OF PROBABILITY THEORY

1

## Outline

- ◇ Uncertainty
- ◇ Probability basics
- ◇ Random variables
- ◇ Probability Distributions
- ◇ Probabilistic Inference
- ◇ Continuous variables
- ◇ Expected Values
- ◇ Independence
- ◇ Bayes' Rule

2

## Representing and Reasoning under Uncertainty

- ◇ Intelligent behavior requires knowledge about the world
- ◇ Propositional logic provides a valuable tool for representing and reasoning with categorical beliefs about the world
- ◇ Any sentence could be true, false, or unknown
- ◇ Often, we are uncertain about the state of the world

3

## Representing and Reasoning under Uncertainty

Example of reasoning under uncertainty

- ◇ Beliefs:
  - If a patient has lung cancer, there is a 60% chance that an X-ray test will come back positive ; and a 40% percent chance negative.
  - If a patient does not have lung cancer, there is 2% percent chance that an X-ray test will come back positive ; and a 98% percent chance negative.
  - Population cancer rate is 1/1000
- ◇ Observation: X-ray test came back positive.
- ◇ Inference task: What is the chance that the patient has lung cancer?

Probability Theory provides a framework for representing and reasoning under uncertainty

4

## Probability basics

Probability deals with chance experiments that have a set of distinct **outcomes**.

For example, we roll a die and the possible outcomes are 1, 2, 3, 4, 5, 6 corresponding to the side that turns up.

We toss a coin with possible outcomes H (heads) and T (tails).

The set of all possible outcomes is called the **sample space**, generally denoted by  $\Omega$ .

Example: A die is rolled once. Then the sample space for this experiment is the 6-element set  $\Omega = \{1, 2, 3, 4, 5, 6\}$ .

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

- a complete specification of the state of the world
- mutually exclusive
- exhaustive

5

## Probability basics

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

$$0 \leq P(\omega) \leq 1$$
$$\sum_{\omega} P(\omega) = 1$$

For example, unless there is reason to believe the die is loaded, the natural assumption is that every outcome is equally likely,

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

An **event**  $A$  is any subset of  $\Omega$ .

Example: The event that the result of the roll is an even number is  $A = \{2, 4, 6\}$ .

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

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

6

## Random variables

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

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

The set of values that a random variable  $X$  can assume is called the **space/domain** of  $X$ .

A probability space  $P$  induces a **probability distribution** for any r.v.  $X$ :

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

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

7

## Random variables

The set of values that a random variable  $X$  can assume is called the **space/domain** of  $X$ ,  $Dm(X)$ . The values are mutually exclusive and exhaustive.

**Propositional** or **Boolean** random variables (have the domain  $\{true, false\}$ )

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

**Discrete** random variables (domain is finite or countably infinite)

e.g.,  $Weather$  is one of  $\{sunny, rain, cloudy, snow\}$

**Continuous** random variables (take on values from the real numbers)

e.g.,  $Temp$

8

## Random variables

Often in AI applications, random variables are basic elements

- the sample points are **defined** by the values of a set of random variables
- A possible world is an assignment of exactly one value to every random variable.
- the sample space is the (Cartesian product of the) domains of the variables

E.g., if the world consists of only two Boolean variables Cavity and Toothache, then there are 4 distinct atomic events or 4 possible worlds:

$$Cavity = false \wedge Toothache = false$$

$$Cavity = false \wedge Toothache = true$$

$$Cavity = true \wedge Toothache = false$$

$$Cavity = true \wedge Toothache = true$$

9

## Probability Distributions

Probability distribution function (probability mass function) gives probability values for all possible values of a random variable

$$P(Weather = sunny) = 0.72, \quad P(Weather = rain) = 0.1$$

$$P(Weather = cloudy) = 0.08, \quad P(Weather = snow) = 0.1$$

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

Probability distribution must satisfy

$$0 \leq P(x) \leq 1, \quad \sum_{x \in Dm(X)} P(x) = 1$$

10

## Joint Probability Distributions

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every combination of the values of the set of r.v.s)

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

<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	0.576	0.08	0.064	0.08

Joint probability distribution must satisfy

$$0 \leq P(\vec{x}) \leq 1, \quad \sum_{\vec{x} \in Dm(\vec{X})} P(\vec{x}) = 1$$

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

11

## Conditional probability

Prior or unconditional probabilities

e.g.,  $P(\text{Cavity} = \text{true}) = 0.1$  and  $P(\text{Weather} = \text{sunny}) = 0.72$   
correspond to belief prior to arrival of any (new) evidence

Conditional or posterior probabilities

e.g.,  $P(\text{cavity}|\text{toothache}) = 0.8$   
i.e., **given that toothache is all I know**

If we know more, e.g., *cavity* is also given, then we have

$$P(\text{cavity}|\text{toothache}, \text{cavity}) = 1$$

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

$$P(\text{cavity}|\text{toothache}, \text{sunny}) = P(\text{cavity}|\text{toothache}) = 0.8$$

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

Notation for conditional distributions:

$$P(\text{Cavity}|\text{Toothache}) = \text{2-element vector of 2-element vectors}$$

12

## Conditional probability

Definition of conditional probability ( $a$  shorthand for  $A = a$ ):

$$P(a|b) = \frac{P(a,b)}{P(b)} \text{ if } P(b) \neq 0$$

Product rule gives an alternative formulation:

$$P(a,b) = P(a|b)P(b) = P(b|a)P(a)$$

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

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

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

Chain rule is derived by successive application of product rule:

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

13

## Inference by enumeration

Start with the joint distribution:

(Cavity may cause the dentist's probe to catch)

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

For any event, sum the atomic events where it is true

$$P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

14

## Probabilistic Inference

One common task is to extract the distribution over a single variable or some subset of variables, called **marginal distribution**

$$P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

$$P(\neg\textit{toothache}) = \dots = 0.8$$

This process is called **marginalization** or **summing out**. For any sets of variables  $Y$  and  $Z$

$$P(Y) = \sum_z P(Y, z)$$

A distribution over  $Y$  can be obtained by summing out all the other variables from any joint distribution containing  $Y$

A variant of this rule is called **conditioning**

$$P(Y) = \sum_z P(Y|z)P(z)$$

15

## Inference by enumeration

Start with the joint distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

Can also compute conditional probabilities:

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

16

## Probabilistic Inference

Let  $\mathbf{X}$  be all the variables. Typically, we want  
the posterior distribution of the query variables  $\mathbf{Y}$   
given specific values  $\mathbf{e}$  for the evidence variables  $\mathbf{E}$

Let the other variables be  $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

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

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

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

17

## Probabilistic Inference

In principle, joint distributions can be used to answer any probabilistic queries.

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???

18

## Probability for continuous variables

For continuous variables, it is not possible to write out the entire distribution as a table because there are infinitely many values. In fact, it no longer makes sense to talk about the probability that  $X$  has a particular value.

Express distribution as a parameterized function of value, called **probability density function**

$$P[x \in (a, b)] = \int_a^b p(x)dx$$

$$p(x) \geq 0, \quad \int p(x)dx = 1$$

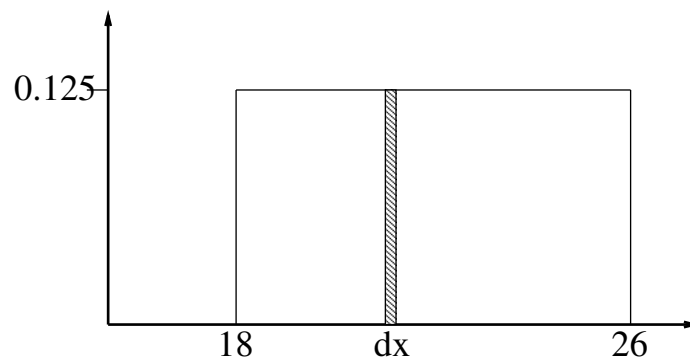
$$p(x, y) = p(y|x)p(x)$$

$$p(x) = \int p(x, y)dy$$

19

## Probability for continuous variables

$p(x) = U[18, 26](x)$  = uniform density between 18 and 26



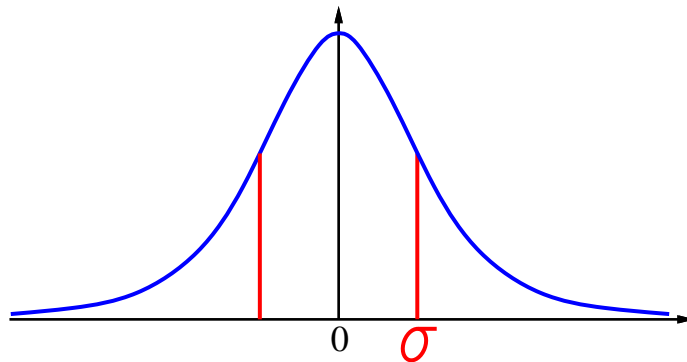
Here  $p(x)$  is a density; integrates to 1.

$$\lim_{dx \rightarrow 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

20

## Gaussian/Normal density

$$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



$$P(|x - \mu| \leq \sigma) = 0.68$$

21

## Expected Values

The **expectation** of the r.v.  $X$  (expected value, mean, average)

$$E[X] = \sum_x xP(x)$$

The expectation of function  $f(x)$

$$E[f(x)] = \sum_x f(x)P(x)$$

$$E[f(x)] = \int_x f(x)p(x)dx$$

The **variance** (and the **standard deviation**) provides a measure of variability around mean value

$$var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

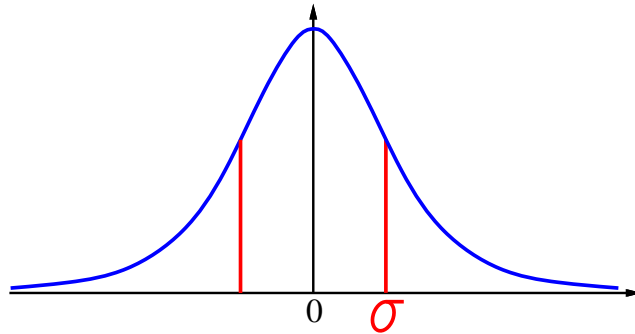
The **covariance**: a measure of statistical dependence between  $X$  and  $Y$

$$cov[X, Y] = E[(X - E[X])(Y - E[Y])] = \sum_{x,y} (x - E[X])(y - E[Y])P(x, y)$$

22

## Gaussian/Normal density

$$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



$$E[X] = \mu, \quad \text{var}[X] = \sigma^2$$

$\sigma$  is called the **standard deviation**

$$P(|x - \mu| \leq \sigma) = 0.68$$

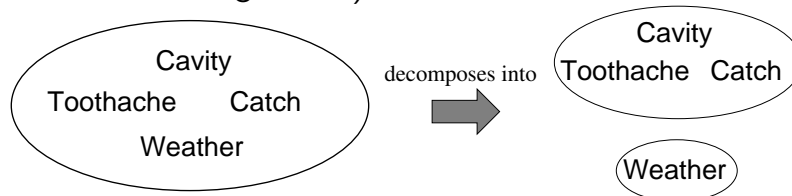
23

## Independence

Two sets of variables  $A$  and  $B$  are **independent** iff

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{or} \quad \mathbf{P}(B|A) = \mathbf{P}(B) \quad \text{or} \quad \mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$$

(for all possible value assignments)



$$\begin{aligned} &\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather}) \\ &= \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity})\mathbf{P}(\textit{Weather}) \end{aligned}$$

32 entries reduced to 12; for  $n$  independent biased coins,  $2^n \rightarrow n$

Independence assertions are usually based on knowledge of the domain. They can help in reducing the size of the domain representation and the complexity of the inference. Unfortunately absolute independences are rare. E.g., Dentistry involves hundreds of diseases/symptoms, all of which interrelated

24

## Conditional independence

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

$$(1) P(\text{catch}|\text{toothache}, \text{cavity}) = P(\text{catch}|\text{cavity})$$

The same independence holds if I haven't got a cavity:

$$(2) P(\text{catch}|\text{toothache}, \neg\text{cavity}) = P(\text{catch}|\neg\text{cavity})$$

*Catch* is conditionally independent of *Toothache* given *Cavity*:

$$\mathbf{P}(\text{Catch}|\text{Toothache}, \text{Cavity}) = \mathbf{P}(\text{Catch}|\text{Cavity})$$

25

## Conditional independence

$X$  is independent of  $Y$  given  $Z$ , denoted  $I(X, Z, Y)$ , iff

$$P(x|y, z) = P(x|z), \forall x \in Dm(X), y \in Dm(Y), z \in Dm(Z)$$

Equivalent statements:

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

$X_1, \dots, X_n$  are mutually independent given  $Z$  if

$$P(X_1, \dots, X_n|Z) = \prod_i P(X_i|Z)$$

26

## Conditional independence contd.

$\mathbf{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$  has  $2^3 - 1 = 7$  independent entries

Write out full joint distribution using chain rule:

$$\begin{aligned}\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) &= \mathbf{P}(\textit{Toothache}|\textit{Catch}, \textit{Cavity})\mathbf{P}(\textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache}|\textit{Catch}, \textit{Cavity})\mathbf{P}(\textit{Catch}|\textit{Cavity})\mathbf{P}(\textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache}|\textit{Cavity})\mathbf{P}(\textit{Catch}|\textit{Cavity})\mathbf{P}(\textit{Cavity})\end{aligned}$$

I.e.,  $2 + 2 + 1 = 5$  independent numbers (equations 1 and 2 remove 2)

In some cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in  $n$  to linear in  $n$ .

Conditional independence assertions are much more commonly available than absolute independence assertions

**Conditional independence is one of the most basic and robust form of knowledge about uncertain environments.**

27

## Bayes' Rule

Product rule  $P(a, b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Rightarrow \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

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

( $1/P(X)$  is often viewed as a normalization constant  $\alpha$ )

Useful for assessing **diagnostic** probability from **causal** probability:

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

Plays a central role in machine learning

28

## Example use of Bayes' Rule

A patient takes a X-ray test, comes back positive for lung cancer

Test accuracy: false negative rate of .4

$$P(\text{test} = + | \text{cancer} = \text{yes}) = .6$$

false positive rate of .02:

$$P(\text{test} = + | \text{cancer} = \text{no}) = .02$$

Population cancer rate .001

$$P(\text{cancer} = \text{yes}) = .001.$$

Answer

$$P(\text{yes} | +) = .6 * .001 / (.6 * .001 + .02 * .999) = .029$$

29

## Bayes' Rule and conditional independence

$$\begin{aligned} P(\text{Cavity} | \text{toothache}, \text{catch}) \\ &= \alpha P(\text{toothache}, \text{catch} | \text{Cavity}) P(\text{Cavity}) \\ &= \alpha P(\text{toothache} | \text{Cavity}) P(\text{catch} | \text{Cavity}) P(\text{Cavity}) \end{aligned}$$

This is an example of a **naive Bayes** model:

$$P(\text{Cause} | \text{Effect}_1, \dots, \text{Effect}_n) = \alpha P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})$$



Total number of parameters is **linear** in  $n$

30

## Summary

Probability is a rigorous formalism for uncertain knowledge

Joint probability distribution specifies probability of every atomic event

Queries can be answered by summing over atomic events

For nontrivial domains, we must find a way to reduce the joint size

Independence and conditional independence provide the tools