

# *Artificial Intelligence*

## Probabilistic reasoning: *representation & inference*

Marco Piastra

# Probability: events as *subsets of possible worlds*

## ■ Boolean algebra

A non-empty collection of subsets  $\Sigma$  of a set  $W$  such that:

1)  $A, B \in \Sigma \implies A \cup B \in \Sigma$

2)  $A \in \Sigma \implies A^c \in \Sigma$

3)  $\emptyset \in \Sigma$

*Corollary:*

The sets  $\emptyset$  e  $W$  belong to any Boolean algebra generated on  $W$

$\Sigma$  is also closed under binary intersection

## ■ $\sigma$ -algebra

A non-empty collection of subsets  $\Sigma$  of a set  $W$  such that:

1)  $A_k \in \Sigma, \forall k \in \mathbb{N}^+ \implies (\bigcup_{k=1}^{\infty} A_k) \in \Sigma$

2)  $A \in \Sigma \implies A^c \in \Sigma$

3)  $\emptyset \in \Sigma$

*Corollary:*

The sets  $\emptyset$  and  $W$  belong to any  $\sigma$ - algebra generated on  $W$

$\Sigma$  is also closed under countable intersection

*This is a stronger requirement:  
closeness under countable union  
Hence a  $\sigma$ -algebra is a boolean algebra  
but not vice-versa*

# Probability: events as *subsets of possible worlds*

## ■ $\sigma$ -algebra (*definition*)

A non-empty collection of subsets  $\Sigma$  of a set  $W$  such that:

- 1)  $A_k \in \Sigma, \forall k \in \mathbb{N}^+ \implies (\bigcup_{k=1}^{\infty} A_k) \in \Sigma$
- 2)  $A \in \Sigma \implies A^c \in \Sigma$
- 3)  $\emptyset \in \Sigma$

## ■ Probability *measure* over a $\sigma$ -algebra

A function  $P : \Sigma \rightarrow [0, 1]$

i.e.  $P$  assigns a measure (i.e. a real number)  
to each elements of a  $\sigma$ -algebra  $\Sigma$  of subsets of  $W$

- 1)  $\forall A \in \Sigma, P(A) \geq 0$
- 2)  $A_k \in \Sigma, \forall k \in \mathbb{N}^+$  are disjoint  $\implies P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$
- 3)  $P(\emptyset) = 0$
- 4)  $P(A^c) = 1 - P(A)$  (which implies  $P(W) = 1$ )

## ■ Probability *space*

A triple  $\langle W, \Sigma, P \rangle$

# Probability: events as *subsets of possible worlds*

- $\sigma$ -algebra
- Probability *measure* over a  $\sigma$ -algebra
- Probability *space*

A triple  $\langle W, \Sigma, P \rangle$

*Why bothering so much with these (very) technical definitions?*

- **Rationale** (*just a few hints*)

Closure w.r.t. *countable unions* of a  $\sigma$ -algebras  
(as well as *countable additivity* of  $P$ )

is required for dealing with *infinite sequences of events* and their properties

However, assuming countable union and additivity is also a *restriction*,  
i.e. to ensure measurability  
(see the so-called *Banach-Tarski paradox* for counterexamples)

# Probability: events as *subsets of possible worlds*

- Probability *measure* over a  $\sigma$ -algebra
- *Disjoint* events

In general

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If  $A \cap B = \emptyset$  then events  $A$  and  $B$  are *disjoint*

$$P(A \cup B) = P(A) + P(B)$$

(\*) Note that  $A \cap B = \emptyset \implies P(A \cap B) = 0$

but not vice-versa: as an event can have zero probability without being empty

(\*\*) Unlike in propositional logic, knowing  $P(A)$  and  $P(B)$  is not sufficient for determining  $P(A \cup B)$

Namely, probability is not *truth-functional* ...

# Studying basic properties: *a finitary setting*

*It can be useful to adopt, at least for a while, a simpler setting that allows a simpler definition of fundamental properties*

- Finite algebras of events

$\Sigma$  is a finite collection of subsets

*In this setting, boolean algebra =  $\sigma$ -algebra*

*Events could also be defined via propositional logic  
(à la de Finetti, 1937)*

- Finitely additive probability measures

*Just summations, no integrals*

*Computability is always guaranteed*

# Partitions, random variables\*

(\*) In a finitary setting

## ■ Partition

A *finite* collection  $A_i$  of *disjoint* subsets (i.e. events) such that

$$\bigcup_i A_i = W$$

A  $\sigma$ -algebra can be generated from a *partition* by taking its closure under union and complement

## ■ Random Variable

Is a convenient way to define a  $\sigma$ -algebra over  $W$

# Partitions, random variables\*

## ■ Partition

A finite collection  $A_i$  of disjoint subsets (i.e. events) such that

$$\bigcup_i A_i = W$$

## ■ Random Variable

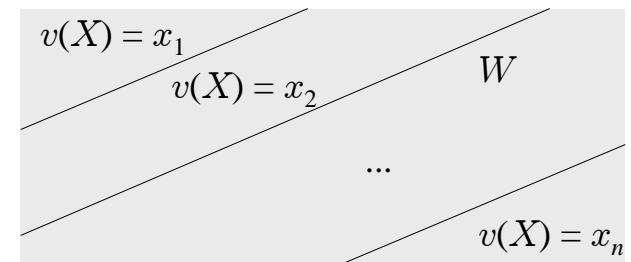
Let  $X$  be a variable having a finite set of values  $\{x_1, x_2, \dots, x_n\}$

- In each possible world, the variable  $X$  has a specific value  $x_i$
- The set of values  $X = x_1, X = x_2, \dots, X = x_n$  defines a partition of  $W$
- Each constraint  $X = x_i$  defines an event (i.e. a subset of  $W$ )
- When  $X = x_i$  and  $X = x_j$  are disjoint events (i.e.  $i \neq j$ )

$$P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$$

Random variables having binary values are also said to be *binomial* (also *Bernoullian*)

Random variables with multiple values are also said to be *multinomial*





# Random variables, joint distribution\*

## ■ Multiple random variables

*In practice, in a probabilistic representation, there will be multiple random variables*

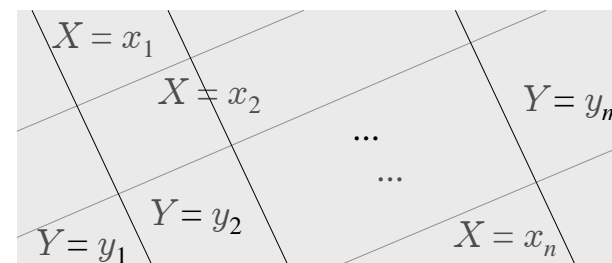
Example:

$X_i$  occurrence of a word  $i$  in the body of an email (binomial)

$Y$  classification of that email as spam (binomial)

Together, a collection of random variable defines  
a partition of  $W$

The intersection of two or more  $\sigma$ -algebras is a  $\sigma$ -algebra



## ■ Joint probability distribution

for a given set of random variables, e.g.  $X, Y, Z$

It is a function that associates a value in  $[0, 1]$  to each individual combination of values

$$P(X = x, Y = y, Z = z)$$

Given that  $X, Y$  e  $Z$  define each a *partition* of  $W$ :

$$\sum_x \sum_y \sum_z P(X = x, Y = y, Z = z) = 1$$

# Random variables: notation

- Random variables, events and  $\sigma$ -algebras

*(sometimes the notation can be ambiguous)*

Examples:

$$P(X)$$

*This is the probability measure over the  $\sigma$ -algebra generated by the random variable  $X$*

$$P(X = x)$$

*This the probability (i.e. a value in  $[0,1]$ ) associated to the event  $X = x$*

$$P(X, Y = y)$$

*This is the probability measure over the  $\sigma$ -algebra generated by the random variable  $X$  in the subspace of  $W$  corresponding to the event  $Y = y$*

# Marginalization

*Removing a random variable from a joint distribution*

Given a joint probability distribution

$$P(X = x, Y = y)$$

The marginal probability  $P(X = x)$  is obtained via summation:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

A marginal probability can be a joint probability too ...

Marginal probability of an event (shorthand notation, values of  $Y$  omitted):

$$P(X = x) = \sum_Y P(X = x, Y)$$

Marginal probability of a  $\sigma$ -algebra (shorthand notation, values of  $Y$  omitted):

$$P(X) = \sum_Y P(X, Y)$$

# Conditional probability

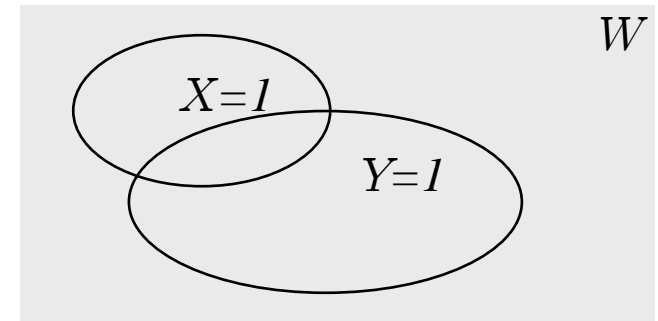
(\*) In a finitary setting

## ■ Definition

$$P(X|Y = y) := \frac{P(X, Y = y)}{P(Y = y)}$$

It is a form of *inference*: from a set  $W$  to a set  $W'$

Therefore, from a probability space to another probability space



Example:  $W$  is the set of possible worlds,  $X, Y$  are binary random variables and  $P(X, Y)$  is the joint probability distribution

Suppose the agent learns that event  $Y = 1$  has occurred: the event  $Y = 0$  is now *impossible* (to him/her)

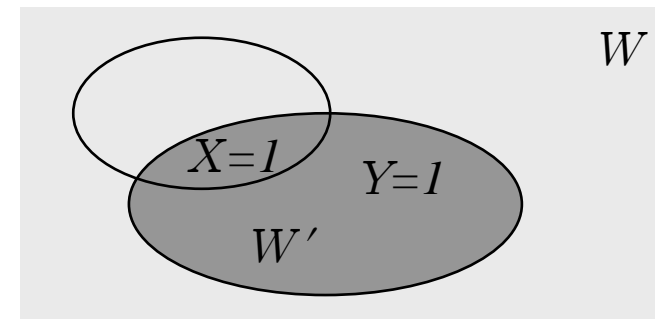
$W' := \{w \in W | Y = 1\}$  is the new set of possible worlds

$P(X|Y = 1)$  is the new probability of  $X$

More in general

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

Denotes the conditional probabilities for the whole  $\sigma$ -algebra of events generated by  $Y$



# Bayes' Theorem (T. Bayes, 1764)



## ■ Definition

A relation between conditional and marginal probabilities

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

$P(Y|X)$  is also called *likelihood*  $L(X|Y)$

The theorem follows from the definition of conditional probability (*chain rule*)

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

Furthermore, given the definition of marginalization:

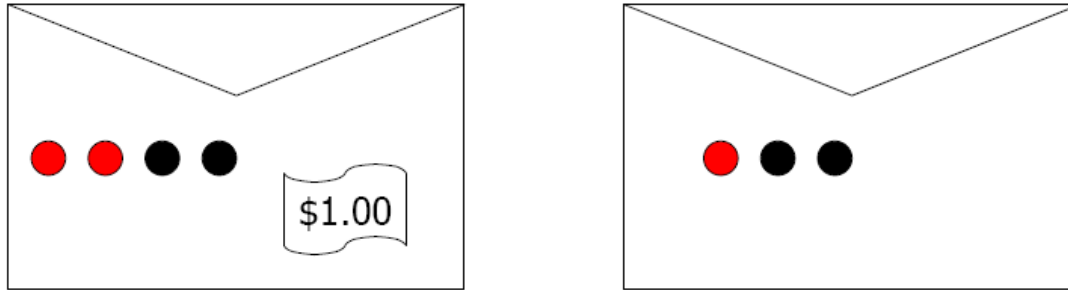
$$P(Y) = \sum_X P(X, Y) = \sum_X P(Y|X)P(X)$$

Also called  
'law of total probability'

it follows an alternative formulation of the Bayes' theorem:

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

# Example: information and bets



- Two envelopes, only one is extracted

One envelope contains two red tokens and two black tokens, it is worth \$1.00

One envelope contains one red token and two black tokens, it is valueless

The envelope has been extracted.

Before posing you bet, you are allowed to extract one token from it

a) The token is black. How much do you bet ?

b) The token is red. How much do you bet ?

Purpose: showing that Bayes' Theorem makes the representation easier

# Independence, conditional independence

## ■ Independence (also *marginal independence*)

Two events are independent

iff their joint probability is equal to the product of the marginals

$$\langle X \perp Y \rangle \Rightarrow P(X, Y) = P(X)P(Y)$$

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

## ■ Conditional independence

Two events are conditional independent, given a third event,

iff their joint conditional probability is equal to the product of the *conditional marginals*

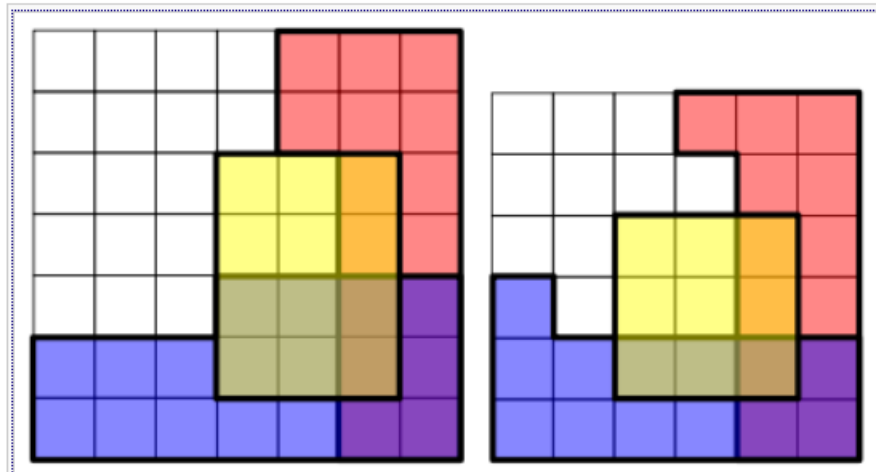
$$\langle X \perp Y | Z \rangle \Rightarrow P(X, Y | Z) = P(X | Z)P(Y | Z)$$

$$\Rightarrow P(X | Y, Z) = \frac{P(X, Y | Z)}{P(Y | Z)} = \frac{P(X | Z)P(Y | Z)}{P(Y | Z)} = P(X | Z)$$

**CAUTION:** *the two forms of independence are distinct!*

$$\langle X \perp Y \rangle \not\Rightarrow \langle X \perp Y | Z \rangle, \quad \langle X \perp Y | Z \rangle \not\Rightarrow \langle X \perp Y \rangle$$

# Independence, conditional independence



[from Wikipedia, "Conditional Independence"]

These are two examples illustrating **conditional independence**. Each cell represents a possible outcome. The events  $R$ ,  $B$  and  $Y$  are represented by the areas shaded red, blue and yellow respectively. And the probabilities of these events are shaded areas with respect to the total area. In both examples  $R$  and  $B$  are conditionally independent given  $Y$  because:

$$\Pr(R \cap B \mid Y) = \Pr(R \mid Y) \Pr(B \mid Y)^{[1]}$$

but not conditionally independent given not  $Y$  because:

$$\Pr(R \cap B \mid \text{not } Y) \neq \Pr(R \mid \text{not } Y) \Pr(B \mid \text{not } Y).$$

$R$ ,  $B$  and  $Y$  here are subsets, i.e. events, not random variables

*The example above shows that (conditional) independence of two specific events does NOT imply (conditional) independence of the whole  $\sigma$ -algebras*



# Probabilistic Inference (no *learning*)

## ■ General structure of probabilistic inference problems

The starting point is a fully-specified joint probability distribution

$$P(X_1, X_2, \dots, X_n)$$

In an *inference* problem, the set of random variables  $\{X_1, X_2, \dots, X_n\}$  is divided into three categories:

- 1) *Observed variables*  $\{X_o\}$  , i.e. having a definite (and certain) value
- 2) *Irrelevant variables*  $\{X_i\}$  , i.e. which are not directly part of the answer
- 3) *Relevant variables*  $\{X_r\}$  , i.e. which are part of the answer we seek

In general, the problem is finding:

$$P(\{X_r\}|\{X_o\}) = \sum_{\{X_i\}} P(\{X_r\}, \{X_i\}|\{X_o\})$$

- “Decidability” (actually “computability”) is not an issue (\*in a finitary setting)  
Given that the joint probability distribution is completely specified
- Computational efficiency can be a problem  
The number of value combinations grows exponentially with the number of random variables

# Continuous random variables (hint)

Although conceptually the same, dealing with continuous random variables is technically difficult

Consider a **continuous** random variable  $X \in \mathcal{X}$  A continuous domain  
e.g. the real interval  $[0, 1]$

$X = x$  does not describe a proper event

For technical reasons (i.e. *measurability*), a point must have probability zero

$$X \leq a \quad X \leq b \quad a < X \leq b$$

(where  $a < b$ ) these are *subsets* are proper events (i.e. they may have non-zero probability)

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

These two events are disjoint

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

Assume that the derivative  $p(X) := \frac{dP(X)}{dX}$  exists

cumulative distribution function (cdf)

$$P(a < X \leq b) = \int_a^b p(X) dX$$

probability density function (pdf)

# Expected value of a random variable

(also *expectation*)

## Basic definition

$$\mathbb{E}_X[X] := \sum_{x \in \mathcal{X}} x P(X = x)$$

*More concise notation*

$$\mathbb{E}[X] := \sum_{x \in \mathcal{X}} x P(x)$$

## Expectation is a linear operator

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[cX] = c\mathbb{E}[X]$$

*Continuous case*

$$\mathbb{E}[X] := \int_{x \in \mathcal{X}} x p(x) dx$$

## Conditional expectation

$$\mathbb{E}_X[X|Y = y] = \mathbb{E}[X|Y = y] := \sum_{x \in \mathcal{X}} x P(X = x|Y = y)$$

## Iterated expectation (*see Wikipedia*)

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$$

# Variance of a random variable

## Basic definition

$$\text{Var}(X) := \mathbb{E}_X[(X - \mathbb{E}_X[X])^2] = \mathbb{E}_X[(X - \mu_X)^2]$$

where  $\mu_X := \mathbb{E}_X[X]$

$$\text{Var}(X) := \sum_{x \in \mathcal{X}} P(X = x) (x - \mu)^2$$

variance is not a linear operator

## Conditional variance

$$\text{Var}(X|Y = y) := \mathbb{E}_X[(X - \mathbb{E}_X[X|Y = y])^2 | Y = y]$$

## Variance lemma

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - 2\mu_X \mathbb{E}[X] + \mu_X^2 \\ &= \mathbb{E}[X^2] - 2\mu_X^2 + \mu_X^2 = \mathbb{E}[X^2] - \mu_X^2 \end{aligned}$$

$$\mathbb{E}[X^2] = \mu_X^2 + \sigma_X^2$$

where  $\sigma_X := \sqrt{\text{Var}(X)}$  *standard deviation*