# Artificial Intelligence

## A Course About Foundations

# Probabilistic Reasoning: Representation & Inference

Marco Piastra

# Probability Space

# Probability Space *(preliminary definition)*

- ***Probability space***

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

*Possible worlds (a.k.a. Sample Space)*

*Event Space (a collection of subsets over $W$)*

*Probability Measure*

$$P : \Sigma \to [0, 1]$$

*The intuitive definition is simple enough,*
*its mathematical translation ... not so much*

# Event Space: *a collection of 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) $\varnothing \in \Sigma$

  *Corollary*:
  The sets $\varnothing$ e $W$ belong to any Boolean algebra generated on $W$
  $\Sigma$ is also closed under *<u>binary</u> 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) $\varnothing \in \Sigma$

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

  *Corollary*:
  The sets $\varnothing$ and $W$ belong to any $\sigma$- algebra generated on $W$
  $\Sigma$ is also closed under *<u>countable</u> intersection*

# Probability Measure

- $\sigma$-algebra (*Event Space*)

  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) $\varnothing \in \Sigma$

- Probability *measure* over a $\sigma$-algebra (i.e., over the events)

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

  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) $A_1, A_2 \in \Sigma$ are *disjoint* $\implies P(A_1 \cup A_2) = P(A_1) + P(A_2)$ *Finite additivity*

     $A_k \in \Sigma,\ \forall k \in \mathbb{N}^+$ are all *disjoint* $\implies P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$ *Countably infinite additivity*
  3) $P(\varnothing) = 0$
  4) $P(A^c) = 1 - P(A)$ (which implies $P(W) = 1$)

# Probability Space

- ***Probability space***

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

  *Possible worlds*

  *Event Space*
  *(a $\sigma$-algebra over $W$)*

  *Probability Measure*

  $$P : \Sigma \to [0, 1]$$

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

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

  Closure w.r.t. *countable unions* of a $\sigma$-algebra (as well as *countable additivity* of $P$) is required for dealing with *infinite sequences of events*

  A $\sigma$-algebra is included in the *power set* of $W$ (i.e., the collection of all its possible subsets): requiring closure on *countable* union is a *restriction*, to ensure *measurability*

  *(see the so-called Banach-Tarski paradox for counterexamples)*

# *An Aside:* Probability is Systemic

In general

*It follows from the additivity property*

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

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

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

(*) Note that $A \cap B = \varnothing \implies P(A \cap B) = 0$
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 *compositional* …

# Discrete Probability

# Studying basic properties*: (*) in a <u>*finitary*</u> setting

*A simpler setting that allows a more intuitive definition of fundamental properties*

*Basic assumption*: the set of possible worlds $W$ is <u>finite</u>

- <u>Finite</u> *event space*

  $\Sigma$ is a <u>*finite*</u> collection of subsets

  *In such setting*
  *boolean algebra* $\equiv$ *$\sigma$-algebra*

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

- <u>Finitely additive</u> *probability measure*

  *Just summations, no integrals*

  *Computability will be always guaranteed*

# Random Variables*

(*) In a finitary setting

# Partitions, random variables*

- 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*

# Partitions, random variables*

(*) In a finitary setting

- **Random Variable** (i.e. a convenient way to define a $\sigma$-*algebra*)

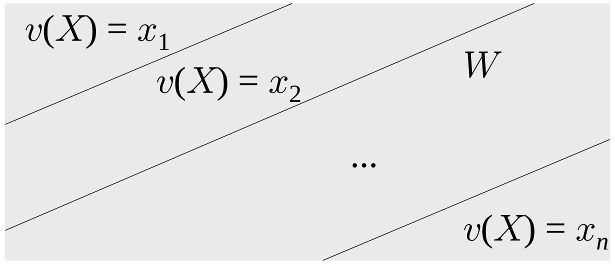
  Let $X$ be a variable having a *finite* set of possible values $\{x_1, x_2, \ldots, x_n\}$
  In each possible world, the variable $X$ is assigned a specific value $x_i$

  - The set of possible assignments $\{X = x_1, X = x_2, \ldots X = x_n\}$ defines a *partition* over $W$
  - A $\sigma$-*algebra* can be obtained by taking the closure of the partition under union and complement

  - $X = x_i$ defines an *event* (i.e. a subset of $W$)
  - $X = x_i$ and $X = x_j$ are *disjoint events* , whenever $i \neq j$

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

*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*

(*) In a finitary setting

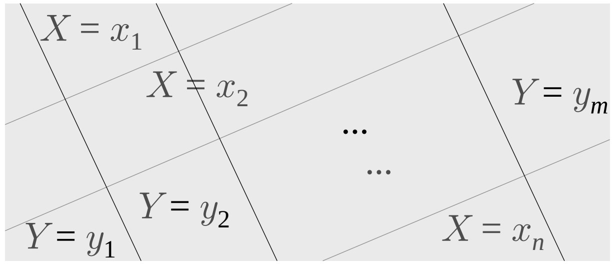
Multiple random variables

*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)

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

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




- 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 $X$, $Y$ and $Z$ define each a *partition* over $W$:

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

# Random variables: notation*

(*) In a finitary setting

- 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 $Y = y$

# Fundamental Operations*

(*) In a finitary setting

# Marginalization*

(*) In a finitary setting

*Removing a random variable from a joint distribution*

Given a joint probability distribution

$$P(X, Y)$$

The *marginal probability* $P(X)$ is obtained via a summation:

$$P(X) \ := \ \sum_{y \in \mathcal{Y}} P(X, Y = y)$$

A marginal probability can be a joint probability as well …

$$P(X, Y) \ := \ \sum_{z \in \mathcal{Z}} P(X, Y, Z = z)$$

Marginal probability, shorthand notation with values of $Y$ omitted:

$$P(X) \ = \ \sum_{Y} P(X, Y)$$

*Shorthand notation*

# Conditionalization*

(*) In a finitary setting

*Exploring 'what if' something becomes known*

Given a joint probability distribution

$$P(X, Y)$$

The ***conditional probability*** $P(X \mid Y = y)$ is defined as:

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

A conditional probability can be a joint probability as well …

$$P(X, Y \mid Z = z) \;:=\; \frac{P(X, Y, Z = z)}{P(Z = z)}$$

Conditional probability, more general notation:

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

Conditional probabilities for the whole $\sigma$-algebra generated by $Y$
*(it represents a family of probability measures)*

# Conditionalization*

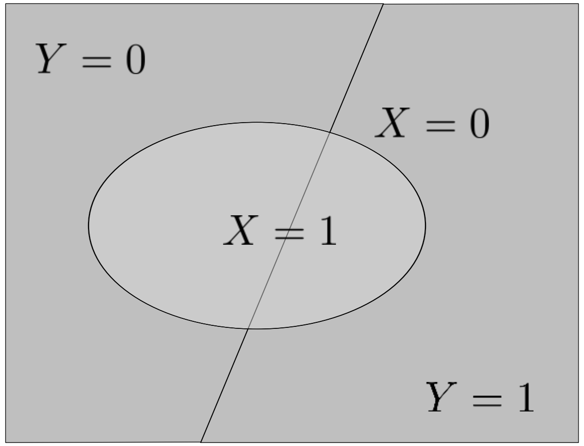
(*) In a finitary setting

*Exploring 'what if' something becomes known*

Conditional probability, more general notation:

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

Assume both variables are *binary* $X, Y \in \{0,1\}$




$$P(X \mid Y=0) \ := \ \frac{P(X,Y=0)}{P(Y=0)}$$

$$P(X \mid Y=1) \ := \ \frac{P(X,Y=1)}{P(Y=1)}$$

*Each value of the conditioning variable defines a distinct event (sub)space*

# Chain Rule

Starting point: a joint distribution

$$P(X, Y, Z)$$

conditioning on $X$ :

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

which implies:

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

then, conditioning on $Y$ :

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

which implies:

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

hence and finally:

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

*Univariate factorization* of the joint distribution

*Any other sequence of conditionalization would do*

# Inference
# (without *learning*)

# Probabilistic Inference* *(general structure)*

(*) In a finitary setting

- General structure of probabilistic inference problems

  The starting point is a fully-specified joint probability distribution

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

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

  1) Observed variables $\{X_o\}$ *having a definite (and supposedly known) value*
  2) Irrelevant variables $\{X_i\}$ *which are neither observed nor directly part of the answer*
  3) Relevant variables $\{X_r\}$ *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" ("computability") is guaranteed *(* in a finitary setting)*

    Given that the joint probability distribution is completely specified
  - Computational efficiency can be a problem

    The number of value combinations to be considered in the summation grows exponentially with the number of random variables in $\{X_r\} \cup \{X_i\}$

# Bayes' Theorem* (T. Bayes, 1764)

(*) In a finitary setting

- 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 the *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:

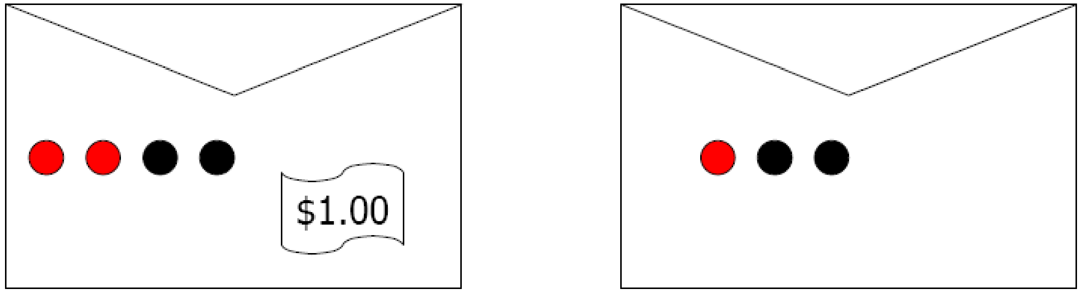
$$P(Y) = \sum_X P(X,Y) = \sum_X P(Y|X)P(X)$$

*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 on 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

# Independence, conditional independence

- **Independence** (also *marginal independence*)

  Two variables are independent
  iff their joint probability can be factorized into the product of *marginals*

$$< X \perp Y > \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \quad \Leftrightarrow \quad < Y \perp X >$$

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

- **Conditional independence**

  Two variables are conditionally independent given a third variable,
  iff their joint conditional probability can be factorized into the product of *conditional marginals*

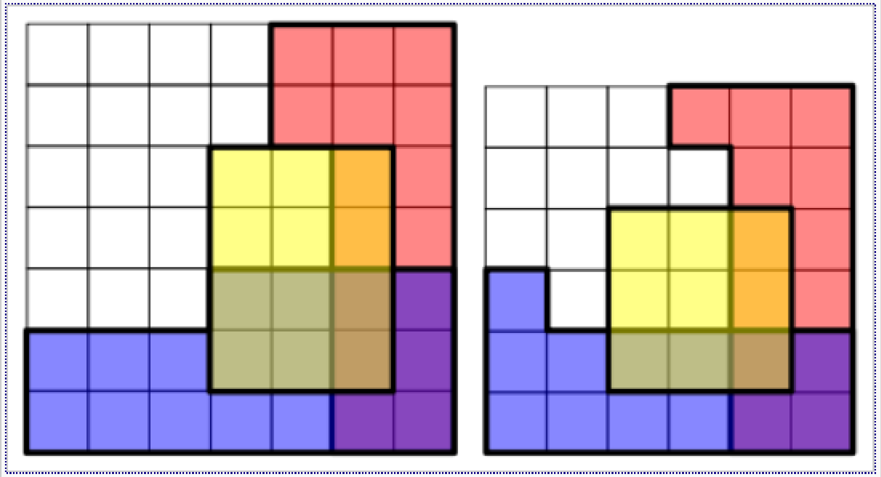
$$< X \perp Y \,|\, Z > \quad \Leftrightarrow \quad P(X,Y|Z) = P(X|Z)P(Y|Z) \quad \Leftrightarrow \quad < Y \perp X \,|\, Z >$$

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

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

$$< X \perp Y > \; \not\Rightarrow \; < X \perp Y \,|\, Z > \qquad < X \perp Y \,|\, Z > \; \not\Rightarrow \; < X \perp Y >$$

# Independence, 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:

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

[from Wikipedia, "Conditional Independence"]

*R, B and Y here are subsets, i.e. <u>events</u>, not random variables*

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

# Continuous Random Variables

# Continuous random variables (hints)

*Although intuitively similar, dealing with continuous random variables is technically difficult*

Consider a **continuous** random variable $X \in \mathcal{X}$

*A continuous value space:*
*e.g. the real interval* $[0, 1]$
*or even* $[-\infty, +\infty]$

$X = x$ does not describe an *event* in a continuous setting

For technical reasons (of *measurability*), a point must have probability zero

Events need to be *subsets*, or better, intervals:

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

*Assuming* $a < b$

Probability measures these *subsets*

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

*These two events are disjoint*

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

$$P_X(x) = P(X \leq x)$$

*Shorthand notation*

# Density and Cumulative Distribution

- **Probability Density Function (pdf)**

  *It is defined as the* derivative

$$p_X(x) := \frac{dP_X(x)}{dx}$$

  provided that it <u>*exists*</u> *everywhere in* $\mathcal{X}$ *and is non-negative:* $p_X(x) \geq 0$

- **Probability Measure as Cumulative Distribution Function (CDF)**

*Cumulative Distribution Function (CDF)*

$$P(a < X \leq b) := \int_a^b p_X(x)\, dx$$

*probability density function (pdf)*

  *As a probability measure, it must integrate to unity*

$$\int_{x \in \mathcal{X}} p_X(x)\, dx = 1$$

*Note that* $p(x)$ *may well be above* $1$ *(<u>its integral</u> over the value space* $\mathcal{X}$ *will be equal to one)*

# 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**

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

*Continuous case*

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

*Also denoted as:*

$\mu_X$

Expectation is a linear operator

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

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

Conditional expectation*

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

(*) In a finitary setting

# Variance of a random variable

Basic definition

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

*Also denoted as:* $\sigma_X^2$

where*:

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

$$\sigma_X := \sqrt{\mathrm{Var}(X)}$$

*Standard Deviation*

*Variance is <u>not</u> a linear operator*

Conditional Variance

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

Variance lemma

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

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

(*) In a finitary setting