Probabilistic reasoning: *representation & inference*

Marco Piastra

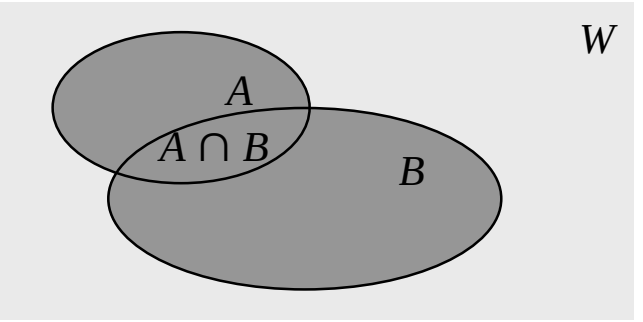
Probability: events as *subsets of possible worlds*

- Boolean algebra

  A non-empty collection of subsets Σ 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
  Σ is also closed under *intersection*



- σ-algebra

  A non-empty collection of subsets Σ of a set W such that:

  1) $A_k \in \Sigma,\ \forall k \in \mathbb{N}^+ \implies (\bigcup_{k=1}^{\infty} A_k) \in \Sigma$ ⟵ *This is a stronger requirement: closeness under <u>countable</u> union*
  2) $A \in \Sigma \implies A^c \in \Sigma$
  3) $\varnothing \in \Sigma$

  Hence a σ-algebra is a boolean algebra but not vice-versa

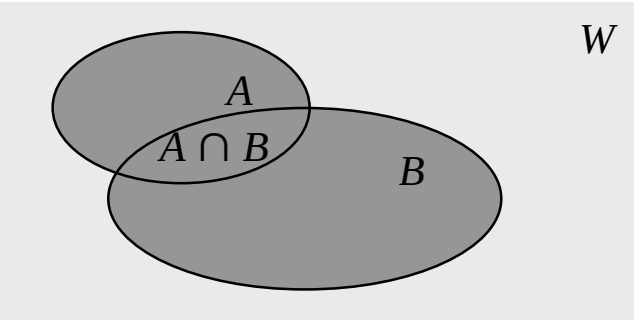
  Corollary:
  The sets $\varnothing$ e W belong to any σ- algebra generated on W
  Σ is also closed under *<u>countable</u> intersection*

Probability: events as *subsets of possible worlds*

- σ-algebra (*definition*)

  A non-empty collection of subsets Σ 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 σ-algebra

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

  i.e. P assigns a measure (i.e. a real number)
  to each elements of a σ-*algebra* Σ 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*** $\implies P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$
  3) $P(\varnothing) = 0$
  4) $P(A^c) = 1 - P(A)$ (which implies $P(W) = 1$)

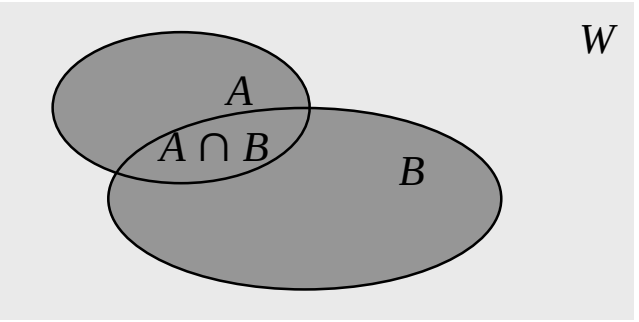
- Probability *space*

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

Probability: events as *subsets of possible worlds*

- σ-algebra
- Probability *measure* over a σ-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 σ-algebras
  (as well as *countable additivity* of P)
  is required for dealing with *infinite sequences of events* and their properties

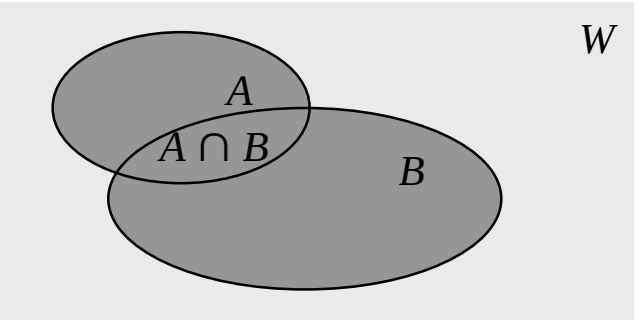
  However, the properties about *countable* union and additivity are also restrictions, to ensure *measurability*
  (see the so-called Banach-Tarski paradox for counterexamples)

Probability: events as *subsets of possible worlds*

- Probability *measure* over a σ-algebra
- *Disjoint* events

  In general

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



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 *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 algebra of events

  Σ is a finite collection of subsets

  In this setting, boolean algebra = σ-algebra

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

- Finitely additive probability measure

  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

  $$\bigcup_i P(A_i) = W$$

  A σ-algebra can be generated from a *partition* by adding the empty set $\varnothing$ and all the unions of two or more subsets (i.e. events) in the partition

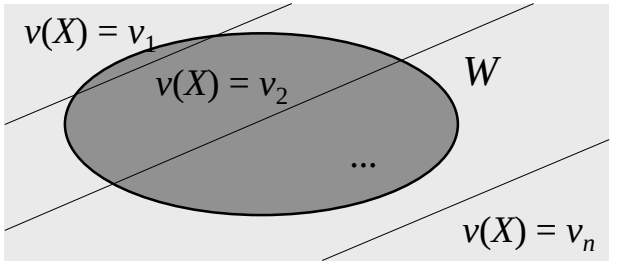
- Random Variable

  Let X be a variable having a *finite* set $\{v_1, v_2, \ldots, v_n\}$ as its domain.

  In each possible world, X has a specific value v_i
  The set of values $X = v_1$, $X = v_2$, ..., $X = v_n$ defines a *partition* of W

  - X is a *random variable*
  - Each constraint $X = v_i$ defines an event (i.e. a subset of W)
  - *Given that* $X=v_i$ e $X=v_j$ are disjoint, $P(X=v_i \vee X=v_j) = P(X=v_i) + P(X=v_j)$ whenever $i \neq j$

Random variable having binary values are also said to be *Bernoullian*
Random variables with vectorial values are also said to be *multinomial*



Random variables, joint distribution*

(*) In a finitary setting

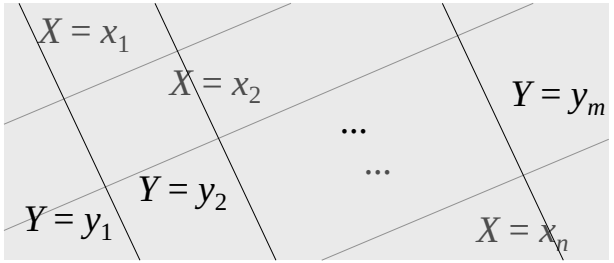
- Multiple random variables

  In practice, in a probabilistic representation, multiple random variables can coexist

  Example:
  X_i occurrence of a word I in the body of an email (0/1)
  Y classification of that email as spam (0/1)

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

  The intersection of two or more σ-algebras *is a* σ-algebra



- Joint probability distribution

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

  It is a *function* $P(X{=}x\ ,\ Y{=}y\ ,\ Z{=}z)$ that associates a value in $[0,\ 1]$ to each individual combination of values $<x\ ,\ y\ ,\ z>$

  Given that X, Y e Z define a partition of 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 σ-algebras

  (*sometimes the notation can be ambiguous*)

  Examples:

  $P(X)$

  This is the probability measure over the σ-algebra *generated by 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 σ-algebra *generated by 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

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

The *marginal probability* $P(X=x, Y=y)$ is obtained via summation:

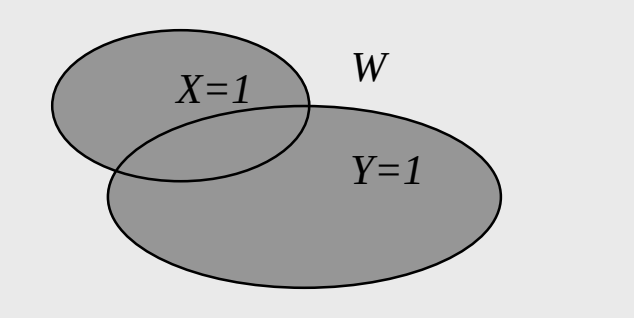
$$P(X = x, Y = y) = \sum_z P(X = x, Y = y, Z = z)$$

A marginal probability, in general, is still a joint probability

Conditional probability

- Definition

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



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 bernoullian 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)

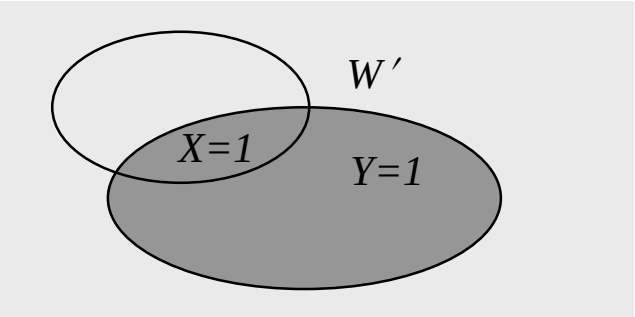
$W' \equiv Y=1$ is the new set of possible worlds

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

More in general

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

Denotes the conditional probabilities for the whole σ-algebra of events generated by Y



Bayes' Theorem (T. Bayes, 1764)

- Definition

  A relation between conditional and marginal probabilities

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

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

$$P(X \mid Y) \propto L(X \mid Y)\ P(X)$$

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

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

  Given the definition of marginalization:

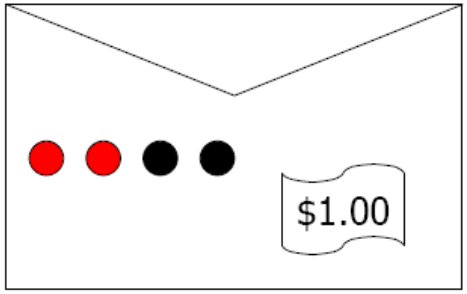
$$P(Y) = \sum_X P(X, Y) = \sum_X P(Y \mid X)\ P(X)$$

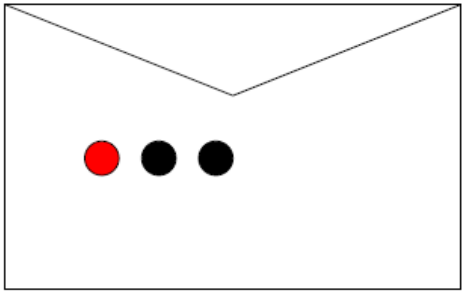
  ← *Also called 'law of total probability'*

  follows an alternative formulation of the Bayes' theorem:

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

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, conditional independence

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

  Two events are independent iff their joint probability is equal to the product of the marginals

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

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

- **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*

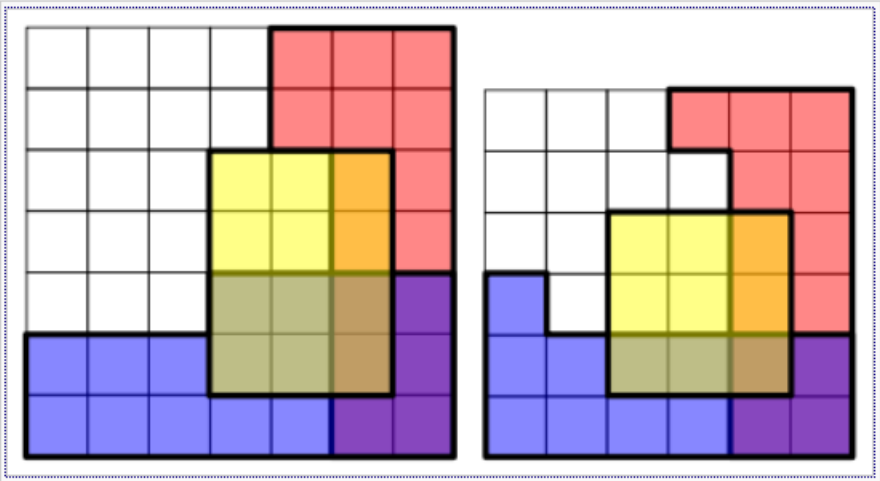
  $$< X \perp Y \mid Z > \Rightarrow P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

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

  CAUTION: the two forms of independence are distinct!

  $$< X \perp Y > \not\Rightarrow < X \perp Y \mid Z >, \quad < X \perp Y \mid Z > \not\Rightarrow < X \perp Y >$$

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"]

CAUTION: in the example above: independence on one event generated by Y does NOT imply independence on the whole σ-algebra of events generated by Y

Unless stated otherwise, independence is assumed to be on the whole σ-algebra generated by the conditioning random variables

Probabilistic Inference

- General setting

  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* $\{X_o\}$, 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:

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

  - "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 variable 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 an *event*

It is not a subset, it always has (probability) measure zero

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

describe *subsets* hence *events* ($a < b$ is assumed here)

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

since the two events on the left hand side are *disjoint*

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

Assuming 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

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

Continuous case

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

A linear operator

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

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

Conditional expectation

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

Iterated expectation (*see Wikipedia*)

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

Variance of a random variable

Basic definition

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

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

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

variance is *not* 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

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

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

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