Probabilistic reasoning: *representation & inference*

Marco Piastra

Possibility (i.e. from logic to probability)

- Objective knowledge and *plausible* knowledge

  Each (rational) agent is supposed to hold some knowledge which is *objective* (to him/her)

  Call Γ the theory representing such objective knowledge,
  to the agent, the set of *possible worlds* is $W \equiv \{<U, v> : <U, v> \models \Gamma\}$

  Example: the agent knows $\Gamma \equiv \{\varphi \vee \psi\}$
  Therefore, only the worlds $\{<U, v> : <U, v> \models \{\varphi \vee \psi\}\}$ are *possible* (to him/her)

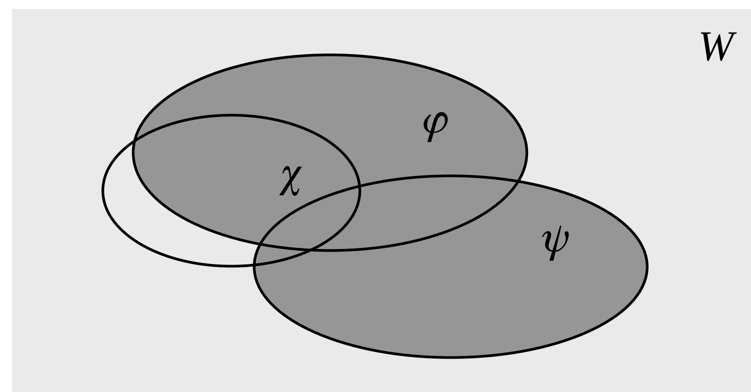
  Does this mean that the *event* $\varphi \vee \psi$ did occur, already?

  On the other hand, the agent might not know χ
  Formally:

  $\varphi \vee \psi \not\models \chi$

  $\varphi \vee \psi \not\models \neg\chi$

  To him/her, both χ and $\neg\chi$ are *plausible*
  (*plausible* = logically *possible*)



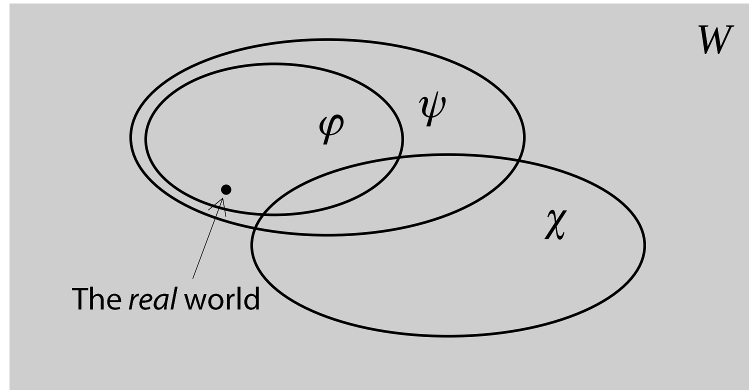
Events as *subsets of possible worlds*

- Subsets of *possible worlds*

We already know that, given an interpreted logical language L (*e.g. of the first order*), every (*closed*) wff is associated to *a subset of possible worlds*

Formally, each (closed) φ is associated to the set $\{<U, v> : <U, v> \models \varphi\}$

(for simplicity, we keep U fixed here)



Intuition:

An **event** can be seen as
a subset of *possible worlds:*
an event is said to <u>*occur*</u>
when the *real* world happens to belong to
the corresponding subset of possible worlds

The agent is not supposed to know which world is the *real* one…

Note:

In 'classical' probability theory, the events need not be defined in a logical fashion

(this fact has also some technical implications, to be clarified later on)

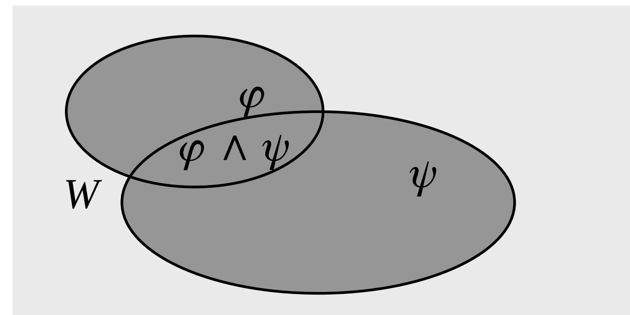
Probability

- Probability is a measure over the subsets of W

  In particular over $W \equiv \{<\mathrm{U}, v> : <\mathrm{U}, v> \models \Gamma\}$

  i.e. the set of worlds that are deemed possible by an agent who knows Γ

  Technically, $P(.)$ is a *function* that assigns a measure (i.e. a real number) to each elements of a σ*-algebra* Σ of subsets of W



σ-algebra (definition)

A collection of subsets Σ of a set W such that:

1) Σ is not empty
2) If $\varphi \in \Sigma$ then $\neg\varphi \in \Sigma$
   ($\neg\varphi$ is intended as the *complement* of φ in W)
3) For any *countable* collection of subsets $\{\varphi_i\}$, $\varphi_i \in \Sigma$, we have $\bigcup_i \varphi_i \in \Sigma$

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

Each element of a σ-algebra is an ***event***

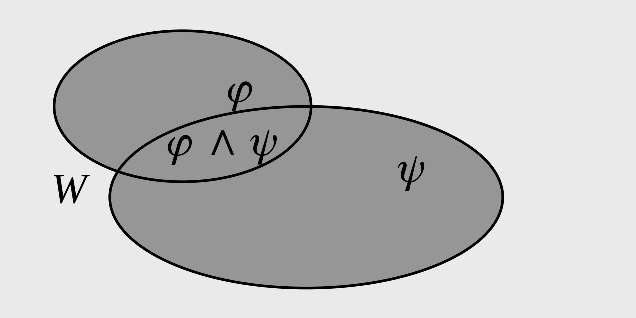
Probability

- Probability is a measure over the subsets of W

  In particular over $W \equiv \{<U, v> : <U, v> \models \Gamma\}$

  i.e. the set of worlds that are deemed possible by an agent who knows Γ

  Technically, $P(.)$ is a *function* that assigns a measure (i.e. a real number) to each elements of a *σ-algebra* Σ of subsets of W



$P(.)$ is a *measure* defined over the σ-algebra Σ

1) For each event $\varphi \in \Sigma$, $P(\varphi) \geq 0$
2) $P(W) = 1$
3) For every <u>*countable*</u> sequence φ_i of *disjoint* events *in* Σ (*disjoint* $\Leftrightarrow$ $\varphi_i \cap \varphi_j \equiv \emptyset$ se $i \neq j$):
   $P(\varphi_1 \vee \varphi_2 \vee \ldots \vee \varphi_n) = \sum_i P(\varphi_i)$

Corollary:
For any *event* $\varphi \in \Sigma$, $0 \leq P(\varphi) \leq 1$

Partitions, random variables*

(*) When W is discrete

- Partition

  A collection φ_i of *disjoint* events such that

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

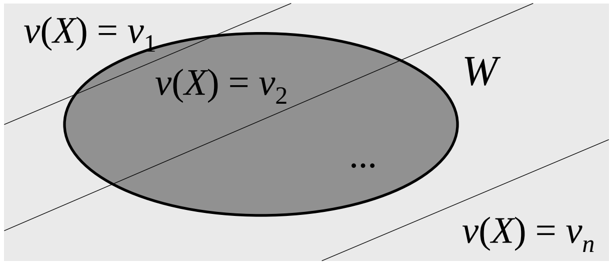
- Random Variable

  Let X be a variable having $\{v_1, v_2, \dots, v_n\}$ as its domain.

  In each possible world, X has a specific value v_i
  The set of values $v(X) = v_1, v(X) = v_2, \dots, v(X) = v_n$ define a *partition* of W

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

(*) When W is discrete

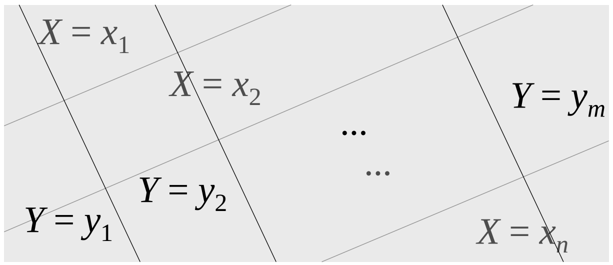
- Many random variables

  In practice, in a probabilistic representation, multiple random variables have to 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 r.v.s define a partition of W

  Any combination of values of the above r.v.s defines an *event*



- Joint probability distribution

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

  It is a <u>*function*</u> $P(X=x_i \wedge Y=y_j \wedge Z=z_k)$ that associates a real value to each indivudual combination of values $<x_i, y_j, z_k>$

  Alternative notiation: $P(X=x_i, Y=y_j, Z=z_k)$ more frequently, just: $P(X, Y, Z)$

  Given that X, Y e Z define a partition of W: $$\sum_i \sum_j \sum_k P(X=x_i, Y=y_j, Z=z_k) = 1$$

Marginalization

Removing a random variable from a joint distribution

Given a joint probability distribution

$P(X=x_i, Y=y_j, Z=z_k)$

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

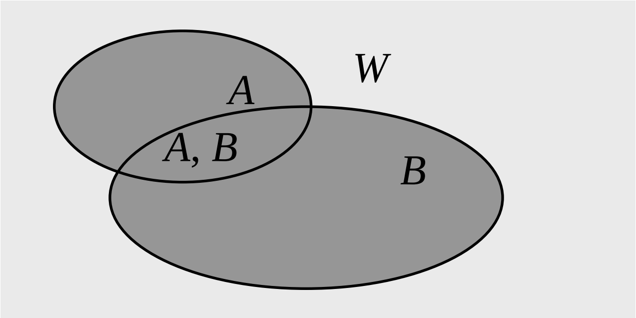
$$P(X=x_i, Y=y_j) = \sum_k P(X=x_i, Y=y_j, Z=z_k)$$

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

Conditional probability

- Definition

$$P(A|B) = \frac{P(A,B)}{P(B)}$$



- Meaning

It is a form of *inference*: switching from a set W to a set W' (i.e. a subset of the former)

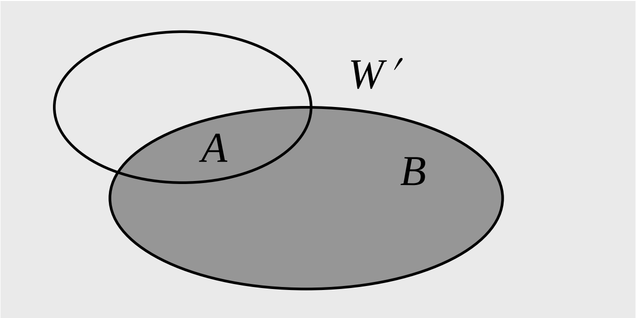
Therefore, from a probability measure to another one

Consider an agent who thinks that W is the set of possible worlds

$P(A)$ is the probability (to him/her) that event A occurs

Suppose that the agent then learns that event B occurred

The event $\neg B$ is now *impossible* (to him/her)



$W' \equiv B$ is the new set of possible worlds

$P(A \mid B)$ is the new probability of A

Bayes' Theorem (T. Bayes, 1764)

- Definition

A relation between conditional and marginal probabilities

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$P(B \mid A)$ is also called *likelihood* $L(A \mid B)$

$$P(A \mid B) \propto L(A \mid B)\, P(A)$$

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

$$P(A, B) = P(B \mid A)\, P(A)$$

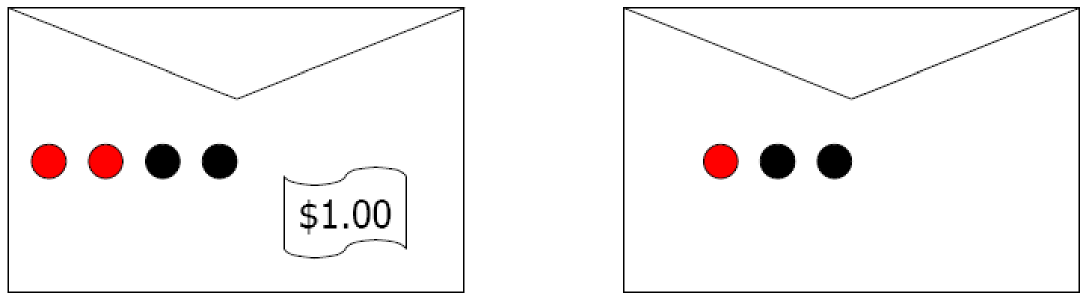
Given the definition of marginalization:

$$P(B) = \sum_A P(A, B) = \sum_A P(B \mid A)\, P(A)$$

also follows (Bayes' theorem alternative formulation):

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{\sum_A P(B \mid A)\, P(A)}$$

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

  $$\langle A \perp B \rangle \quad \Rightarrow \; P(A, B) = P(A)\, P(B)$$

- **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*

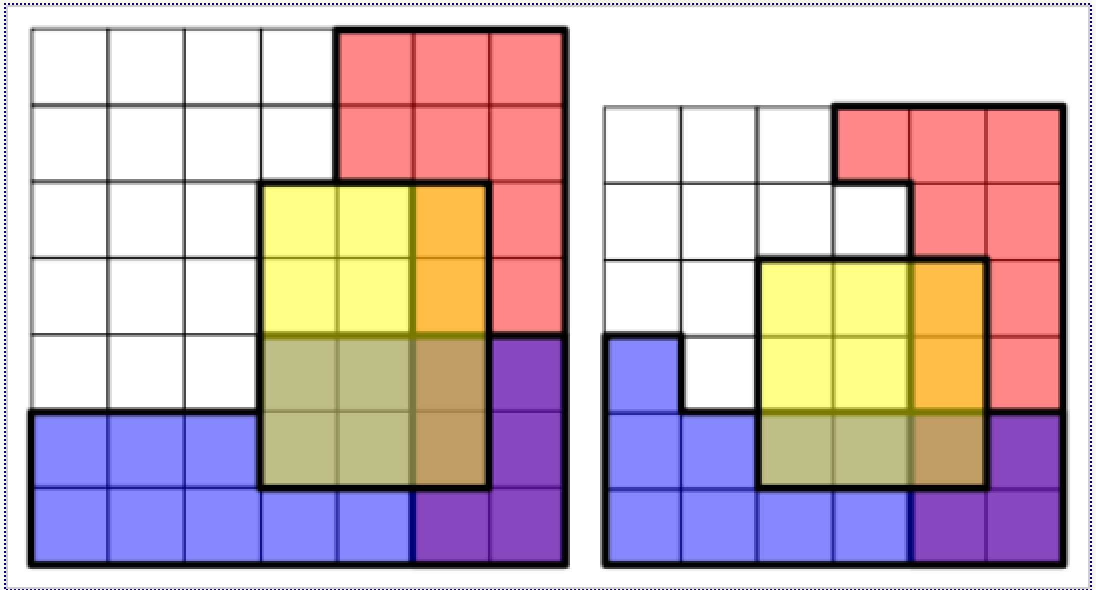
  $$\langle A \perp B \mid C \rangle \; \Rightarrow \; P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

  $$\Rightarrow \; P(A \mid B, C) = \frac{P(A, B \mid C)}{P(B \mid C)} = \frac{P(A \mid C)\, P(B \mid C)}{P(B \mid C)} = P(A \mid C)$$

  CAUTION: the two forms of independence are distinct!

  $\langle A \perp B \rangle \not\Rightarrow \langle A \perp B \mid C \rangle, \; \langle A \perp B \mid C \rangle \not\Rightarrow \langle A \perp B \rangle$

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"]

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

In general, the problem is finding:

$$P(\{X_f\} | \{X_e\}) = \sum_{\{X_r\}} P(\{X_f\}, \{X_r\} | \{X_e\})$$

- "Decidability" (actually "computability") is not an issue (*in a discrete 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