Probability Spaces

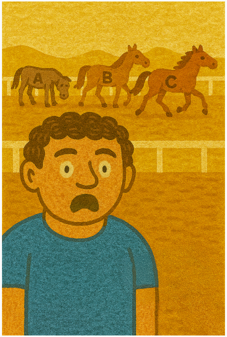
Probabilistic Decision Making — Lecture 2

8th October 2025

Robert Peharz
Institute of Machine Learning and Neural Computation
Graz University of Technology

We offer Tim the following combined bet:

-10 Euro	on Horse A	$(o_A = 50)$
-100 Euro	on Horse B	$(o_B = 5)$
-400 Euro	on Horse C	$(o_C = 1.25)$



outcome	**we win**	
Horse A:	$-10 \times 50 + 510$	$= 10$
Horse B:	$-100 \times 5 + 510$	$= 10$
Horse C:	$-400 \times 1.25 + 510$	$= 10$

Dutch Book—a combination of bets that guarantees loss.

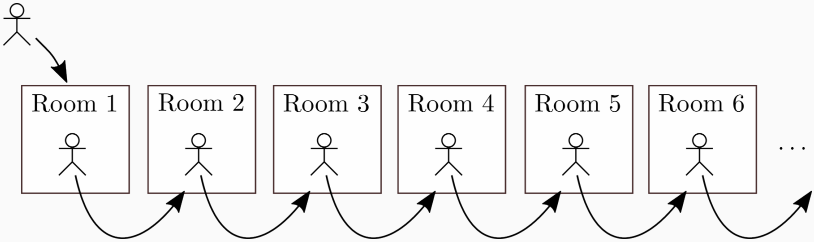
A Dutch Book exists for a set of odds **if and only if** the odds are <u>incoherent</u> with respect to the laws of probability, i.e. $o_A \stackrel{!}{=} \frac{1}{\mathbb{P}(A)}$, where $\mathbb{P}$ is **any** probability distribution.

Preamble

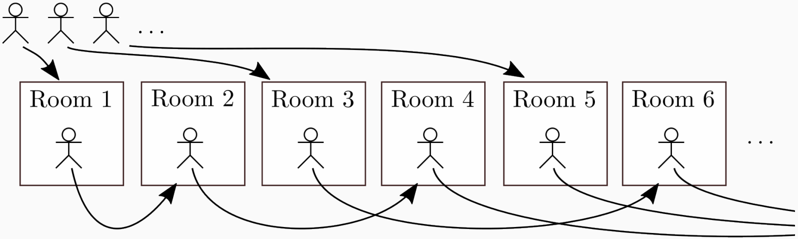
Today we are going to introduce probability formally. First we need to talk about **sets**, **infinities** and **(un)countability**.

Doxiadis & Papadimitriou, Logicomix: An Epic Search for Truth

- **Hilbert's Hotel** has infinitely many rooms, enumerated $1, 2, 3, \ldots$
- all rooms are occupied by guests
- a new person arrives—can they have a room as well?
- **yes!** Hilbert asks each guest, residing in room i, to move to room $i + 1$
- room 1 is now free and can by given to the new person



- now an infinite bus arrives bringing infinitely many new guests
- is there room for them as well?
- **yes!** Hilbert asks each guest, residing in room i to move to room $2i$
- now all rooms with odd numbers are free and can be given to the infinitely many new guests
- **takeaway: infinities behave unintuitively!**



- a set $\mathcal{X}$ is **countable** when it is
  - finite, or
  - infinite and there exists a **one-to-one mapping** between $\mathcal{X}$ and the natural numbers $\mathbb{N}$
- of course, $\mathbb{N}$ itself is countable
- the even numbers $\mathbb{N}_e = \{2, 4, 6, \dots\}$ are countable, even though $\mathbb{N}_e \subset \mathbb{N}$
- same for the odd numbers $\mathbb{N}_o = \{1, 3, 5, \dots\}$
- also the set of integers $\mathbb{Z}$ is countable:

$\mathbb{N}$	1	2	3	4	5	6	7	8	9	....
$\mathbb{Z}$	0	1	-1	2	-2	3	-3	4	-4	....

even though $\mathbb{N} \subset \mathbb{Z}$

Countable Sets cont'd

- using **Cantor's diagonal argument** it turns out that also **Cartesian products** of countable sets are countable, e.g.

$$\mathbb{Z} \times \mathbb{Z} = \{(a, b) \mid a, b \in \mathbb{Z}\}$$
$$\mathbb{N}^D = \{(a_1, \ldots, a_D) \mid a_1, \ldots, a_D \in \mathbb{N})\}$$

- hence, also the set of all rational numbers $\mathbb{Q}$ is countable (as they can be represented as tuples of two integers)
- there are **infinitely many rational numbers in every interval** $(a, a + \epsilon)$, yet they are **countable!**

Uncountable Sets

- however, the set all real numbers $\mathbb{R}$ is **uncountable**
- **Cantor's argument (roughly):**
  - $\mathbb{R}$ was countable, we could write all real numbers between 0 and 1 in an infinite list, each number written as a (unique) infinite binary expansion
  - construct a new real number that differs in one digit from any other real number $\Rightarrow$ not in the list, contradiction
- **there are strictly more real numbers than natural numbers!**
- this leads to different **sizes of infinity** (**cardinalities**), where countable is the "smallest" infinity:

$$\overbrace{|\mathbb{R}|}^{\text{uncountable}} > \overbrace{|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}|}^{\text{countable}}$$

- let S be any set
- the **power set** 2^S is the sets of all subsets of S

$$2^S = \{A \mid A \subseteq S\}$$

- for example, if $S = \{a, b, c\}$

$$2^S = \{\{\}, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

- $2^{\mathbb{N}} = \{A \mid A \subseteq \mathbb{N}\}$ all possible subsets of natural numbers
- $2^{\mathbb{R}} = \{A \mid A \subseteq \mathbb{R}\}$ all possible subsets of real numbers

Cantor's Theorem

Let S be any set. The cardinality of the power set 2^S is **strictly larger** then the cardinality of S.

- for any finite set S we have $|2^S| = 2^{|S|}$
- one can show that $|\mathbb{R}| = |2^{\mathbb{N}}| > |\mathbb{N}|$, i.e. there are as many real numbers as there are subsets of natural numbers
- but also
$$\cdots > |2^{(2^{\mathbb{R}})}| > |2^{\mathbb{R}}| > |\mathbb{R}| > |\mathbb{N}|$$
- hence, there are **infinitely many infinities!**

Probability Spaces

Why Probability Spaces?

Probability spaces, the central object in probability, are perceived as a bit "abstract", and one rarely works with them directly.

Why do we need to discuss them then?

- they rigorously describe "what probability really is"
- albeit abstract, they are actually quite simple
- hence, to understand the "true nature of probability", one should grasp the basic notion of probability space
- essentially, they are the **"Linux kernel" of probability:** you rarely need to work with them directly, but knowing about them will make you a better user!

A **probability space** is a triple

$$(\Omega, \mathcal{F}, \mathbb{P})$$

where

- Ω is a **sample space** (**any** nonempty set)
- $\mathcal{F}$ is a **sigma-algebra** over Ω
- $\mathbb{P}$ is a **probability measure** defined on $\mathcal{F}$

We are going to discuss these three components in detail.

- the **sample space** Ω can be **any** non-empty set
- it establishes the "model universe" of "domain of discourse"
- the elements $\omega \in \Omega$ are called **atomic events** or **elementary events**

Sample Space **Example**

a suitable sample space for a 6-sided die is:

$$\Omega_{\text{die}} = \{1, 2, 3, 4, 5, 6\}$$

the set of TU Graz students can be a sample space:

$$\Omega_{\text{students}} = \{\text{student}_1, \text{student}_2, \dots\}$$

$\mathbb{R}$ the set of real number can be a sample space:

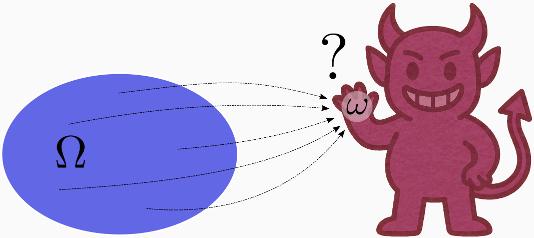
$$\Omega_{\text{real}} = \mathbb{R}$$

$\{f\}$ the set of all real-valued functions can be a sample space:

$$\Omega_{\text{function}} = \{f \colon \mathbb{R} \mapsto \mathbb{R}\}$$

Random Mechanism

A **random selection mechanism** picks an atomic event ω from Ω—**but we don't know how nor which one is picked.**



Idea of Probability

Capture the **randomness**/**uncertainty**/**ignorance**/**lack of information** about ω and do useful computation with it.

With our "universe" Ω fixed, any possible (composite) event can be described as whether the selected ω is contained in some subset $A \subset \Omega$. Hence, such a subset A is called event:

Event

Consider a sample space Ω. An **event** is some subset $A \subseteq \Omega$. The event "happens" if the selected $\omega \in A$. Note the nomenclature:

- **atomic events** ω are the **elements** of $\omega \in \Omega$
- **events** A are **subsets** of $A \subseteq \Omega$.

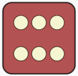
Mutually Exclusive Events

Two events $A \subseteq \Omega$ and $B \subseteq \Omega$ with $A \cap B = \emptyset$ are called **disjoint** or **mutually exclusive**.

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Prime number
$A = \{2, 3, 5\}$

≥ 5
$A = \{5, 6\}$



Fibonacci number
$A = \{1, 2, 3, 5\}$

< 5
$A = \{1, 2, 3, 4\}$

Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

Events

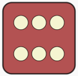
Example

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

disjoint

Prime number
$A = \{2, 3, 5\}$

Fibonacci number
$A = \{1, 2, 3, 5\}$



≥ 5
$A = \{5, 6\}$

< 5
$A = \{1, 2, 3, 4\}$

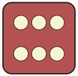
Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Prime number
$A = \{2, 3, 5\}$

Fibonacci number
$A = \{1, 2, 3, 5\}$



≥ 5
$A = \{5, 6\}$

< 5
$A = \{1, 2, 3, 4\}$

disjoint

Odd
$A = \{1, 3, 5\}$

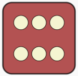
Even
$A = \{2, 4, 6\}$

Events

Example

$\Omega = \{1, 2, 3, 4, 5, 6\}$

roll...



Prime number
$A = \{2, 3, 5\}$

Fibonacci number
$A = \{1, 2, 3, 5\}$

≥ 5
$A = \{5, 6\}$

< 5
$A = \{1, 2, 3, 4\}$

Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

Events

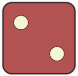
Example

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$\omega = 2$

Prime number
$A = \{2, 3, 5\}$

≥ 5
$A = \{5, 6\}$



Fibonacci number
$A = \{1, 2, 3, 5\}$

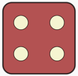
< 5
$A = \{1, 2, 3, 4\}$

Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\omega = 4$$



Prime number
$A = \{2, 3, 5\}$

Fibonacci number
$A = \{1, 2, 3, 5\}$

≥ 5
$A = \{5, 6\}$

< 5
$A = \{1, 2, 3, 4\}$

Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

Events

Example

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$\omega = 3$

Prime number
$A = \{2, 3, 5\}$

≥ 5
$A = \{5, 6\}$

Fibonacci number
$A = \{1, 2, 3, 5\}$

< 5
$A = \{1, 2, 3, 4\}$

Odd
$A = \{1, 3, 5\}$

Even
$A = \{2, 4, 6\}$

Sigma-Algebra

Let Ω be a sample space and $\mathcal{F}$ a set of events (i.e., a set of subsets of Ω). $\mathcal{F}$ is called a **sigma-algebra** over Ω if

- $\Omega \in \mathcal{F}$
- $\mathcal{F}$ is **closed under complement**
  if $A \in \mathcal{F}$ then also $A^c = \Omega \setminus A \in \mathcal{F}$
- $\mathcal{F}$ is **closed under countable union**
  if $A_1, A_2, \dots \in \mathcal{F}$ then also $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

It follows from the definition that also

- $\emptyset \in \mathcal{F}$ and
- $\mathcal{F}$ is **closed under countable intersection**: if $A_1, A_2, \dots \in \mathcal{F}$ then also $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$

Sigma-Algebras: A Logically Closed System of Events

The properties of the sigma-algebra just means that it is closed under logical operations, i.e. that any **logical combinations** of events are also in $\mathcal{F}$:

- **complement** is **logical not**:

$$\omega \in A^c \quad \text{means} \quad \omega \notin A$$

- **union** is **logical or**:

$$\omega \in A \cup B \quad \text{means} \quad \omega \in A \;\text{ or }\; \omega \in B$$

- **intersection** is **logical and:**

$$\omega \in A \cap B \quad \text{means} \quad \omega \in A \;\text{ and }\; \omega \in B$$

Further logical operations follow, e.g. **xor** (symmetric set difference).

- For any Ω, the set $\mathcal{F} = \{\emptyset, \Omega\}$ is a sigma-algebra (**trivial sigma-algebra**, smallest one).
- For any Ω, the power set 2^Ω is a sigma-algebra (**discrete sigma-algebra**, largest one).
- For $\Omega = \{1, 2, 3, 4, 5, 6\}$, the following are all sigma-algebras:
  - $\mathcal{F}_1 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\}$
  - $\mathcal{F}_2 = \{\emptyset, \{1, 2, 3, 4\}, \{5, 6\}, \{1, 2, 3, 4, 5, 6\}\}$
  - $\mathcal{F}_3 = \{\emptyset, \{1, 2\}, \{3, 4\}, \{5, 6\},$
    $\{3, 4, 5, 6\}, \{1, 2, 5, 6\}, \{1, 2, 3, 4\},$
    $\{1, 2, 3, 4, 5, 6\}\}$

Probability Measures

Probability Measure

Let Ω be a sample space and $\mathcal{F}$ be a sigma-algebra over Ω. Let

$$\mathbb{P}: \mathcal{F} \mapsto [0, 1]$$

be a function assigning real numbers to each event in $A \in \mathcal{F}$. Function $\mathbb{P}$ is called a **probability measure** if

- $\mathbb{P}(\Omega) = 1$
- if $A_1, A_2, \dots \in \mathcal{F}$ are **disjoint**, then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$

Note that $\mathbb{P}$ is a function that takes **sets** as arguments!

Note that it follows that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(\emptyset) = 0$.

Probability Space

- Let Ω be a non-empty set (**sample space**),
- $\mathcal{F}$ be a **sigma-algebra** over Ω
- $\mathbb{P}\colon \mathcal{F} \mapsto [0, 1]$ be a function (**probability measure**) with
  - $\mathbb{P}(\Omega) = 1$
  - For any **disjoint** $A_1, A_2, A_3, \ldots$ from $\mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad (\textbf{sigma-additivity})$$

(Note that this includes finite unions as well, by setting all but finitely many $A_i = \emptyset$)

Any such triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**. The elements of Ω are called **atomic events**. The elements of $\mathcal{F}$ are called **(composite) events**.

This is the universally accepted definition of probability!

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = 2^{\Omega}$
- $\mathbb{P}(A) = \frac{|A|}{6}$ for all $A \in \mathcal{F}$

$(\Omega, \Sigma, \mathbb{P})$ is a **probability space**.

Probability Space – Single Die II

Example

We can also construct a different probability space as follows:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4\}, \{5, 6\}, \{1, 2, 3, 4\}, \{3, 4, 5, 6\}, \{1, 2, 5, 6\}, \Omega\}$
- Let $\mathbb{P}$ be given as

A	$\emptyset$	$\{1, 2\}$	$\{3, 4\}$	$\{5, 6\}$	$\{1, 2, 3, 4\}$	$\{1, 2, 5, 6\}$	$\{3, 4, 5, 6\}$	Ω_{die}
$\mathbb{P}(A)$	0	0.5	0.4	0.1	0.9	0.6	0.5	1

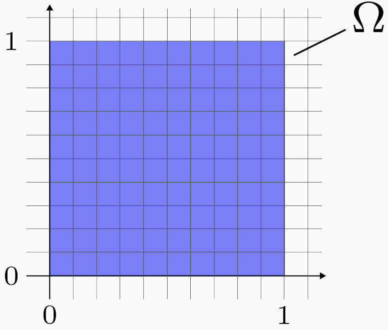
$(\Omega, \mathcal{F}, \mathbb{P})$ is a **probability space**.

Note. We don't have probability for atomic events here!

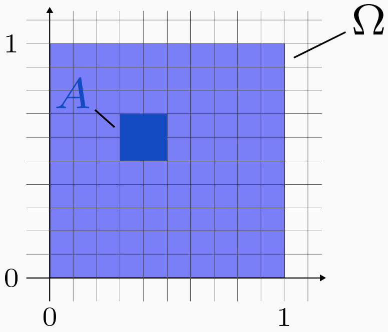
- a **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ containing
    - **sample space** Ω (**any** non-empty set)
    - **sigma-algebra** $\mathcal{F}$ (logically closed system of subsets of Ω)
    - **probability measure** $\mathbb{P}$ (mapping $\mathcal{F}$ to real numbers)
- **rules of probability:**
    - $\mathbb{P}(\Omega) = 1$
    - probabilities of disjoint events just add up
- **this concisely describes what probability really is!**
- in the following, let's discuss:
    - why this abstractness, and why sigma-algebras?
    - what does probability "mean"?

(Non-)Measurable Sets

- consider the unit square $\Omega = [0,1]^2 = \{(x,y) \mid 0 \le x, y \le 1\}$
- it's **area** is obviously 1

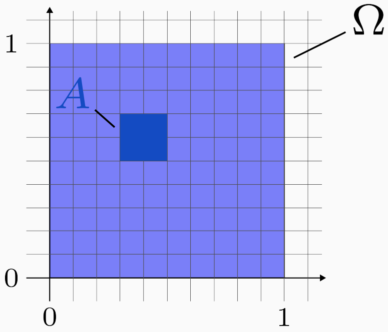


What is the area of the subset $A \subset \Omega$ shown below?

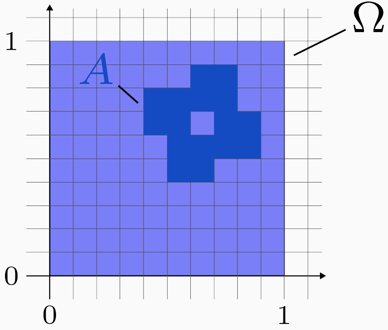


What is the area of the subset $A \subset \Omega$ shown below?

You will probably conclude 0.04, is it is a square of length 0.2.

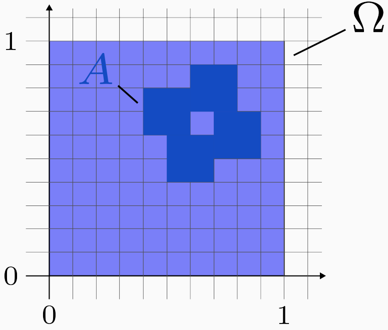


What is the area of this subset $A \subset \Omega$?

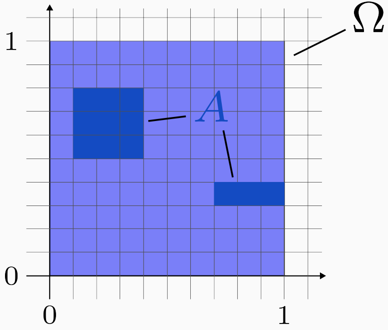


What is the area of this subset $A \subset \Omega$?

0.16, as it is a union of 4 squares of area 0.04

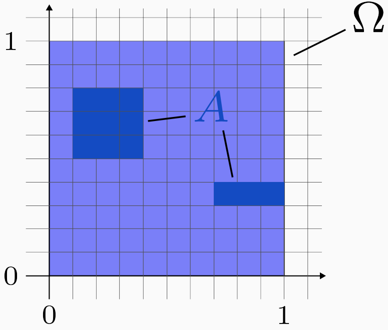


What is the area of this subset $A \subset \Omega$?

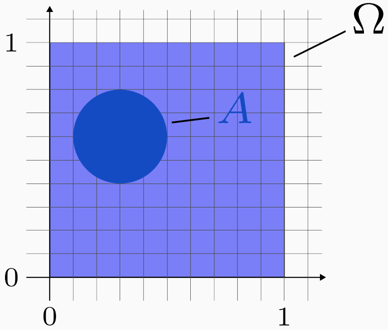


What is the area of this subset $A \subset \Omega$?

0.12, as it is a union of 2 rectangles of areas 0.09 and 0.03, respectively. It doesn't hurt that the squares are not connected.

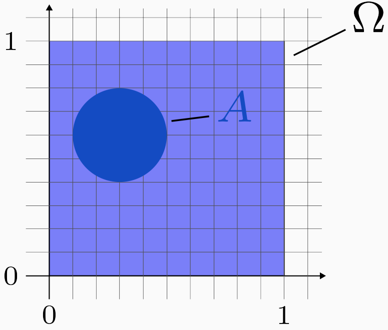


What is the area of this subset $A \subset \Omega$?



What is the area of this subset $A \subset \Omega$?

$0.126\ldots$ Even though a circle is **not a finite union** of rectangles, it can be represented as a **countable union of rectangles**, leading to the formula $r^2\pi$.



Measures

Note what we did in the previous example:

- we were considering some space $\Omega = [0,1]^2$ (the unit square)
- we assigned an area to all kind of subsets $A \subseteq \Omega$
- this can be understood as a function λ, that assigns real numbers (their area) to sets:

$$\lambda(A) = \text{“area of } A\text{”}$$

- **note how similar this is to the probability measure $\mathbb{P}$:**
  - both assign real numbers to subsets of a some space Ω
  - the numbers of disjoint sets just add up to yield the number assigned to the union (sigma-additivity)
- such functions are called **measure** (areas, volume, length, masses, charges, probabilities, etc.)

Measures cont'd

- **a probability measure is just a measure assigning 1 to Ω!**
- in the previous example Ω was the unit square with area 1, hence λ is actually a probability measure in this case (the **uniform distribution** on the unit square)
- generally, λ is called **Lebesgue measure**, that assigns **length**, **area**, **volume**, etc. to Euclidean space $\Omega = \mathbb{R}^D$
- **classical question in measure theory:** can λ be defined on <u>**all**</u> subsets of Ω?
- in other words, can λ be defined on the power set of $\mathbb{R}^D$?

$$\lambda : 2^{\Omega} \mapsto [0, \infty] \quad (?)$$

Non-Measurable Sets

- the answer is **no**
- one can show that there exists subsets of the Euclidean space that are in conflict with the desiderata of the Lebesgue measure—hence λ does not exist on 2^{Ω}
- these "bad" sets are called **non-measurable sets**
- they are **very** abstract and are constructed indirectly via the **axiom of choice** (one of the foundational axioms of mathematics)
- it turns out that the axiom of choice is equivalent with the existence of non-measurable sets!

The Reason for Sigma-Algebras

- the existence of non-measurable sets forbids that we use the power set 2^Ω as domain for (probability) measures
- hence, we need to specify a sigma-algebra $\mathcal{F}$ "that works" and excludes non-measurable sets
- there are also other good reasons why sigma-algebras are useful (not discussed here)
- **Note:** the troubles with non-measurabilty only occurs when Ω is uncountable—**when Ω is countable, it is save to assume $\mathcal{F} = 2^\Omega$!**

- let E be **any** collection of subsets of Ω
- for example think of $\Omega = \mathbb{R}$ and $E = \{(a, b) \mid a, b \in \mathbb{R}\}$ the collection of all open intervals (E is no sigma-algebra)
- the **sigma-algebra $\sigma(E)$ generated by E** is defined as the smallest sigma-algebra that contains E
- formally, it is the intersection of **all** sigma-algebras that contain E

$$\sigma(E) = \bigcap_{\mathcal{F}:\ E \subseteq \mathcal{F}} \mathcal{F}$$

- they are a very powerful tool in measure/probability theory

- when $\Omega = \mathbb{R}$, there is a standard choice for the sigma-algebra: the **Borel sigma-algebra** $\mathcal{B}(\mathbb{R})$
- it is defined as

$$\mathcal{B}(\mathbb{R}) \coloneqq \sigma(E)$$

  where E can be (equivalently)
  - the set of all open intervals
  - the set of all closed intervals
  - the set of all open sets
  - the set of all closed sets
  - ...
- it contains all sets "one can imagine" and excludes non-measurable sets
- the Lebesgue measure (length) and all probability measures of interest can be defined without problems on it

The Borel sigma-algebra generalizes to $\mathbb{R}^D$, where it can equivalently be defined as the sigma-algebra generated by

- the open hypercubes
- the closed hypercubes
- the open sets
- the closed sets
- ...

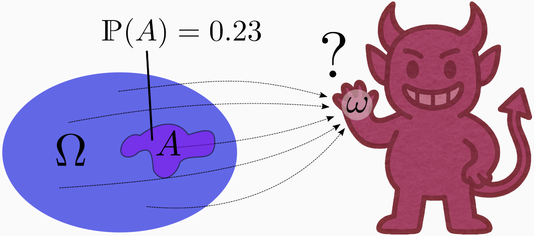
- **classical problem in measure theory:** measures such as probability, length, area volume, etc, cannot be defined on all **subsets**, i.e. the power set 2^{Ω}
- **reason:** the foundational axioms of mathematics
- to this end, one needs to work with smaller sigma-algebras
- standard choice for real-valued spaces: the Borel-sets
- if Ω is "just" countable (i.e. discrete), the power set works fine

Interpretations of Probability

Probability: The Big Picture

Probability is about assigning real numbers (probabilities) to events (subsets of Ω) such that

- $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\Omega) = 1$
- for **disjoint** $A_1, A_2, A_3 \ldots$ $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$



Ok. But what do these numbers mean?

Interpretations of Probability

There are at least 4 good interpretations of probability:

- **fair prices**, **avoidance of Dutch books**, see last lecture (works for finitely many events)
- the **frequentist interpretation**
- the **Bayesian interpretation**
- the interpretation as **Shannon information**

I personally don't think that one of the interpretations is superior to the others—rather I switch between them and treat them as different mental pictures of a rigorous **calculus of reasoning**.

Frequentist Interpretation

- the **frequentist interpretation** requires **repetition**
- rather than drawing a **single** $\omega \in \Omega$, we draw a **sequence** $(\omega_1, \ldots, \omega_N)$, **independently and identically distributed (i.i.d.)**, which we write as

$$\omega_i \overset{i.i.d.}{\sim} \mathbb{P} \qquad i \in [N]$$

- $\sim$ means "distributed according to"
- $[N]$ is a shorthand for $\{1, \ldots, N\}$
- we essentially introduce N **independent copies** $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ **of the original probability space** $(\Omega, \mathcal{F}, \mathbb{P})$

Frequentist Interpretation cont'd

- for any event $A \in \mathcal{F}$ we might count how many ω_i are in A, and normalize the result:

$$\hat{\mathbb{P}}_N(A) := \frac{N_A}{N} = \frac{\sum_{i=1}^{N} \mathbb{1}[\omega_i \in A]}{N}$$

  where $\mathbb{1}$ is the **indicator function** which evaluates to 1 if its argument it true and 0 otherwise
- one can be easily show that $\hat{\mathbb{P}}_N(A)$ is a probability measure for any N
- moreover, when N goes to infinity, it recovers $\mathbb{P}$:

$$\lim_{N \to \infty} \hat{\mathbb{P}}_N(A) = \mathbb{P}(A)$$

Frequentist Interpretation cont'd

Hence, in the **frequentist interpretation** the number $\mathbb{P}(A)$ is seen as the **relative frequency** of trials with the outcome $\omega_i \in A$, when N grows very large.

Think of A as a bucket at which we throw a large number of peas in an identical manner. The percentage of peas ending up in the bucket is its probability.

Bayesian Interpretation

Classical logic can be "emulated" in probability theory, by restricting probabilities to Boolean values 0 and 1. For example:

- **Logical Not** and **Double Negation:**

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - (1 - \mathbb{P}(A))$$

- **Logical Or:** The probability of $A \cup B$ can be written as

$$\mathbb{P}(A \cup B) = \overbrace{\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)}^{1 \text{ iff } \mathbb{P}(A)=1 \text{ or } \mathbb{P}(B)=1}$$

- **Implication (Modus Ponens)** can be written as

$$\mathbb{P}(A \cap B) = \overbrace{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}^{1 \text{ iff } \mathbb{P}(A)=1 \text{ and } \mathbb{P}(B \mid A)=1}$$

(we will learn about conditional probability $\mathbb{P}(B \mid A)$ later.)

Bayesian Interpretation

- classical logic considers only `True` (1) and `False` (0)
- the Bayesian interpretation sees probability as a **relaxation of Boolean truth values** $\{0, 1\}$ to the **interval** $[0, 1]$
- hence, probabilities are **degrees of plausibility** or **degrees of (un)certainty** (unfortunately, also called **subjective beliefs**)
- **Cox's Theorem: under relatively mild assumptions, probability is the <u>only</u> generalization of logic!**

Shannon Information

- in his development of **information theory**, Shannon readily accepted probability theory
- he devised the **information content** (**surprise**) $I(A)$ of an event by requiring
  - the more likely an event, the less information it contains

$$\mathbb{P}(A) > \mathbb{P}(B) \quad \Rightarrow \quad I(A) < I(B)$$

  - $I(A)$ is continuous in $\mathbb{P}(A)$
  - for independent events A, B, the information content of the event that both happen is the sum of the individual information contents:

$$I(A \cap B) = I(A) + I(B)$$

Shannon Information cont'd

- there is a single function (up to scaling) satisfying these desiderata: the **negative logarithm**

$$I(A) = -\log \mathbb{P}(A)$$

- with log to the base 2 the unit of information is called `bits`
- with log to the base e (Euler number) the unit of information is called `nats` (conversion: $1\,\texttt{bit} = 0.693\,\texttt{nat}$)
- for example, the outcome of a coin flip contains

$$-\log_2(0.5) = 1\,\texttt{bit}$$

- **information content measures the message length or memory required for the event**

Shannon Information cont'd

- information content and probability are related **one-to-one** via

$$I(A) = -\log \mathbb{P}(A) \qquad \mathbb{P}(A) = \exp(-I(A))$$

- they are equivalent quantities and **hence, we might just interpret probability as an information value!**
- information theory has been wildly successful and can be seen as an extension of probability theory

- **what is probability?**
  formally, described via **probability spaces** $(\Omega, \mathcal{F}, \mathbb{P})$
- **why is it defined that way (why $\mathcal{F}$)?**
  problems with measurability ("bug" in foundations of mathematics)
- **how can we interpret $\mathbb{P}(A)$?**
  - **fair prices**, non-existence of **Dutch books**
  - **frequentist interpretation**: normalized frequency
  - **Bayesian interpretation**: relaxed logical value
  - **Shannon information**: information content, surprise