SANS-MIRI: Review of probability theory

Jorge Garcia Vidal, Jose M. Barcelo Ordinas and Pau Ferrer Cid

July 17, 2023

The course SANS (Statistical Analysis of Networks and Systems) belongs to the Master MIRI (Master of Innovation and Research in Computer Science) of the Faculty of Computer Science of Barcelona. The course is an introduction to some mathematical foundations used in data science. The course content includes an introduction to probability, linear algebra, and estimation.

In this section, we study a number of techniques that are useful to model, analyze and predict the operation of systems and networks. These techniques are mainly based on results from probability theory and statistics.

Probability Theory and Information Theory are classic subjects, and there are many very good books covering the main concepts touched in the course, with different levels of depth. We can recommend some of them: W. Feller, "An Introduction to Probability Theory and its Applications, Vol I", D. Mackay, "Information Theory, Inference, and Learning Algorithms", J. S. Rosenthal, "A First Look at Rigorous Probability Theory", and T. M. Cover and J. A. Thomas, "Elements of Information Theory".

1 Kolmogorov's axioms and Bayesian probability theory

There is not a single way of defining a "probability theory".

In 1933, the outstanding Russian mathematician A. N. Kolmogorov proposed a formalism for probability theory based on measure theory[1]. Surprisingly, in this definition of probability, there is nothing related to "uncertainty", as probability is defined as a *measure*, akin to an area, length, or weight. The link with our intuitive understanding of probability comes from one of the most important results in this theory, the Law of Large Numbers, which states that if the probability of an event E is $p(E)$, and we create a composed experiment in which we perform N *independent* experiments, the relative frequency in which

[1] https://archive.org/details/foundationsofthe00kolm/page/n3/mode/2up

the event E occurs in the sequence of results gets closer to $p(E)$ as N goes to infinity. Some authors refer to this formulation of probability theory based on measure theory as "frequentist probability".

Alternative definitions of probability exist, the most important of them being the Bayesian probability theory (which can be further divided into "objective Bayesian" and "subjective Bayesian"). In this formulation, the probability is closer to logic, as it is defined as a quantification of the degree of belief of a statement being correct, given some previous knowledge or hypothesis[2].

As an example, assume that we are interested in obtaining the probability of obtaining the same results in two fair coins that are tossed independently. In the Kolmogorov approach, we would define a sample space

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\},$$

while the event we are interested in would be $E = \{(H, H), (T, T)\}$. Finally, using the fact that we have fair coins and independent experiments, we would assign the probability as a measure to this event E, $p(E) = 1/2$.

The Bayesian framework would assign a degree of belief on the sentence S: "The result in the two coins is the same", given the hypothesis, H_1: "The two coins are fair", H_2: "The two coin rolls are independent", which in this case, using symmetry and independence arguments, would be $p(S \mid H_1, H_2) = 1/2$.

We will focus on Kolmogorov's approach to probability theory, although when we will study Bayesian estimation, we will discuss with some more detail the Bayesian approach.

1.1 Probability spaces

Kolmogorov's formulation of probability is based on what is known as a *probability space* $(\Omega, \mathbb{A}, p)$, which consists of a *sample set* Ω, a *σ-algebra of events* $\mathbb{A}$, whose elements are subsets of Ω, and a *probability map* $p : \mathbb{A} \mapsto \mathbb{R}$ that assigns a real number to each event included in $\mathbb{A}$.

The sample space Ω can be any arbitrary set, and can be interpreted as the set that contains the possible outcomes of a probabilistic experiment. From the point of view of probability theory, the only relevant characteristic of the set Ω is its cardinality.

Although the theory is in principle the same for any arbitrary Ω, in practice we will usually consider the following three cases:

- Discrete probability spaces: Ω is a countable set. For instance: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ or $\Omega = \mathbb{N}$.

[2] https://bayes.wustl.edu/etj/prob/book.pdf

- Continuous probability spaces: Ω is an uncountable set such as $[0,1]$, $\mathbb{R}^{\ltimes}$ or $\{0,1\}^{\infty}$ (i.e. the set of infinite binary sequences).
- Stochastic Processes: Ω is a set of functions that depend on some parameter $t \in \mathbb{T}$ for some set $\mathbb{T}$. For instance, $\Omega = \mathbb{R}^{[0,1]}$, i.e. the set of real functions defined on the interval $[0,1]$.

As we see, each case includes the previous cases as a particular example. During the course, we will mainly consider discrete and continuous spaces, and only briefly will mention Markov Chains, one of the simplest and more important examples of a stochastic process.

The set of events $\mathbb{A}$ must have a special structure, as it must be a *sigma-algebra* (σ-algebra). This means that:

- $\Omega \in \mathbb{A}$,
- $E \in \mathbb{A} \Rightarrow E^c \in \mathbb{A}$,
- If $\{E_n\}$ is a *countable* collection of sets, with $E_n \in \mathbb{A} \Rightarrow \bigcap E_n \in \mathbb{A}$

In a σ-algebra is also true that if $\{E_n\}$ is a *countable* collection of sets, with $E_n \in \mathbb{A} \Rightarrow \bigcup E_n \in \mathbb{A}$.

The qualifier *countable* deserves special attention, and we will see that this restriction is key throughout the theory.

In general, when we have a set Ω together with a σ-algebra $\mathbb{A}$ of subsets of Ω, we will say that $(\Omega, \mathbb{A})$ is a *measurable* space.

The map p takes as argument one element of $\mathbb{A}$ (i.e. an event) and returns a real number. To qualify as a probability map, it must fulfill three conditions:

- $p(\Omega) = 1$,
- $E \in \mathbb{A} \Rightarrow p(E) = 1 - p(E^c)$,
- If $\{E_n\}$ is a *countable* collection of disjoint events, i.e. $E_n \in \mathbb{A}$, and $E_n \bigcap E_m = \emptyset$, with $n \neq m \Rightarrow p(\bigcup E_n) = \sum p(E_n)$.

These properties correspond to what in measure theory is called a *finite measure.*

We will deal first with discrete probability spaces.

2 Discrete probability spaces

The discrete case avoids many of the mathematical subtleties that appear in continuous probability spaces and in stochastic processes. This does not necessarily means that, in the application examples that we will study, obtaining expressions for the probabilities for specific events of interest in discrete probability spaces will be easier than in the continuous case, as this is very often not the case.

Very often, in discrete probability spaces the σ-algebra $\mathbb{A}$ will be the set of all subsets of Ω, also known as the *power set*, denoted as 2^{Ω}. It is easy to prove that the power set is always a σ-algebra. We will see, however, that in continuous probabilistic spaces or in stochastic processes, the power set σ-algebra is usually not used.

If Ω is a countable set, all the events of $\mathbb{A} = 2^{\Omega}$ are also countable sets, and thus any event can be expressed as the countable union of singletons (i.e. of sets formed by a single element): $E \in \mathbb{A}, E = \bigcup\{\omega_n\}$, meaning that $p(E) = \sum p(\{\omega_n\})$. In other words, *if we know the probability for all the singleton events, we can find the probability for any event.*

When one study first this theory it is easy to be confused and think that the probability for singletons is assigned to the elements $\omega \in \Omega$, however, we must remember that probabilities are always assigned to elements of $\mathbb{A}$ (i.e. to specific subsets of Ω), and not to the elements of Ω.

3 Basic combinatorial formulas

For discrete probability spaces, counting the number of elements of a given set is usually an important step in order to obtain the probability of the events of interest. For instance, in the special case of *uniform* discrete probabilities, in which all the singletons sets have the same probability,i.e. $p_n = \frac{1}{|\Omega|}$, we can calculate the probability of any event E as $p(E) = \frac{|E|}{|\Omega|}$. We review here some basic results. Assume that $|E| = n$, then:

- The size of $E \times E \times ... \times E = E^k$ is $|E^k| = n^k$.
- The number of subsets of k elements is $\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n_k}{k!}$.
- The total number of subsets we can create (i.e. the size of the power set 2^E) is 2^n.

Where we have used the convention $n_k = n(n-1)...(n-k+1)$ for $1 \leq k \leq n$, and $n_0 = 1$. Assume that we have n different elements (e.g. $\{1, ..., n\}$), and we create the set of n-tuples: $\{(1, 2, ..., n), (2, 1, ..., n), ..., (n, n-1, ..., 1)\}$ (i.e. the set of permutations of n distinguishable elements). Then the size of this set is $n!$.

Many of these problems are described as finding the number of different ways that n balls (distinguishable or not) can be placed in k bins or n flags can be placed on k poles, etc. See volume 1 of the book by W. Feller for a rich number of examples. One must understand some conventions: when we place balls in a bin, balls are not ordered, while when we place flags on a pole, the order becomes important.

Some basic results are:

- There are n^k ways of placing n distinguishable balls in k bins,
- There are $(n+k-1)_k$ ways of placing n distinguishable flags on k poles.
- There are $\binom{n+k-1}{k}$ ways of placing n indistinguishable balls in k bins (or equivalently n indistinguishable flags on n poles. In this case, if $n \leq k$, there are $\binom{k-1}{n-1}$ configurations with non-empty bins or poles.

3.1 Optional: Inclusion/exclusion formula

If $E_1, E_2 \in \mathbb{A}$ we have $p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1, E_2)$. The *inclusion/exclusion* formula generalizes this for the union of more than two events:

$$E_1, ..., E_n \in \mathbb{A} \Rightarrow p(\cup_n E_n) = S_1 - S_2 + ... + (-1)^{n+1} S_n$$

with $S_k = \sum p(E_i, ..., E_l)$ where the sum includes the $\binom{n}{k}$ different combinations of k events.

For instance:

$$\begin{aligned} p(E_1 \cup E_2 \cup E_3) &= S_1 - S_2 + S_3 \\ &= p(E_1) + p(E_2) + p(E_3) - \\ &\quad [p(E_1, E_2) + p(E_1, E_3) + p(E_2, E_3)] + p(E_1, E_2, E_3) \end{aligned}$$

4 Conditional probability and conditional independence

Let E and F be two events with $p(F) \neq 0$. We define the *conditional probability of E given F* as:

$$p(E \mid F) = \frac{p(E \bigcap F)}{p(F)}.$$

4.1 Independence and conditional independence of events

- Two events E and F are *independent* when $p(E \bigcap F) = p(E)p(F)$. An equivalent definition for independent events is $p(E|F) = p(E)$. A common notation used to indicate that two events E and F are independent is $E \perp\!\!\!\perp F$.

- Two events E and F are *conditional independent* given a third event G with $p(G) \neq 0$, when $p(E \bigcap F \mid G) = p(E \mid G)p(F \mid G)$. The notation used in this case is $E \perp\!\!\!\perp F|G$

Independence of events (or random variables) is probably one of the most important concepts in probability theory and the key aspect that differentiates probability theory from general measure theory (see for instance page 8 of Kolmogorov's text).

4.2 Total probability and Bayes formulas

Two useful identities are very often used to find expressions for the probability of events:

- Total probability formula:

  $$P(E) = \sum p(E \mid F_n),$$

  with $\{F_n\}$ a collection of disjoint events with $\bigcup F_n = \Omega$.

- Bayes formula:

  $$p(E \mid F) = \frac{p(F \mid E)p(E)}{p(F)}$$

  with $p(F) \neq 0$.

4.3 Identities involving conditioning of events

It is useful to have some familiarity with manipulations involving conditional probabilities. For instance:

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

etc, where we assume that the terms in the denominators are non-zero.

5 Random variables

Usually, our intuitive idea of a random variable corresponds to a "random", "unpredictable" number. However, the mathematical definition of a random variable corresponds to a *measurable function*: Let $(\Omega, \mathbb{A})$ and $(\Psi, \mathbb{B})$ be two measurable spaces, a function $X : \Omega \to \Psi$ is a *random variable* (or *measurable* *function*) when it also fulfills a technical condition: $F \in \mathbb{B} \Rightarrow X^{-1}(F) \in \mathbb{A}$. Otherwise stated, in this course $\Psi = \mathbb{N}$ or $\mathbb{R}$. If Ω is a countable set, we will say that X is a *discrete random variable*.

If we have a probability space $(\Omega, \mathbb{A}, p)$, a random variable X between $(\Omega, \mathbb{A})$ and $(\Psi, \mathbb{B})$ induces a probability space $(\Psi, \mathbb{B}, p_X)$, with p_X defined as:

$$p_X(F \in \mathbb{B}) = p(X^{-1}(F) \in \mathbb{A})$$

In the case in which $\Psi = \mathbb{Z}$, we only need to know the probability for the singletons $\{n\}$:

$$p_X(n) = p(X^{-1}(\{n\})).$$

From now on, we will always assume that random variables take values in $\Psi = \mathbb{R}$ or $\Psi = \mathbb{Z}$ or $\mathbb{N}$.

It is usual to use the following convention when dealing with random variables: $p(X \leq n) = p(E)$ where $E \in \mathbb{A}$ with $E = X^{-1}((-\infty, n])$, etc.

5.1 Joint (discrete) random variables and functions of random variables

Assume that X and Y are two discrete random variables defined on the *same* probability space $(\Omega, \mathbb{A}, p)$. We can defined then a *joint* random variable (X, Y) on the same probability space with $(X, Y)(\omega) = (X(\omega), Y(\omega))$.

The distribution of this joint random variable (we can also say *the joint distribution* of X and Y) is given by:

$$p_{X,Y}(n, m) = p(X^{-1}(\{n\}) \cap Y^{-1}(\{m\})).$$

The *marginal distribution* of X (i.e. $p_X(n)$) is given by:

$$p_X(n) = \sum_m p_{X,Y}(n, m).$$

The *conditional distribution* of X given Y would be:

$$p_{X|Y}(n, m) = p(X^{-1}(\{n\})|Y^{-1}(\{m\})) = \frac{p_{X,Y}(n, m)}{p_Y(m)}.$$

These definitions for joint and marginal distributions can be extended to the continuous case in a straightforward manner. The extension of the conditional distribution definition requires more attention as it may happen that we are conditioning on events of zero probability.

5.2 Independence and conditional independence of random variables

Two joint random variables X and Y are independent when $p_{X,Y}(n,m) = p_X(n)p_Y(m)$.

Analogously, X and Y are independent given Z when $p_{X,Y|Z}(n,m|k) = p_{X|Z}(n|k)$ $p_{Y|Z}(m|k)$

5.3 Function of random variables

If we have a random variable X and f is a function (e.g. a real function) we can define a new random variable $Z = f(X)$ with $Z(\omega) = f(X(\omega))$. E.g. $Z = X^2$, $Z = e^X$, etc.

If we have joint random variables X and Y and f is a function of two variables (e.g. a real function of two variables) we can define a new random variable $Z = f(X,Y)$ with $Z(\omega) = f(X(\omega), Y(\omega))$. E.g. $Z = X + Y$, $Z = e^{XY}$, etc.

6 Expected value, variance, and moments

We define the *expected value* of a discrete random variable X as

$$\mathbb{E}(X) = \mu_X = \sum_{\omega\in\Omega} X(\omega)p(\{\omega\})$$

if this sum converges. In the mathematical theory of probability the expected values are usually defined in terms of the *Lebesgue integral*, i.e.

$$\mathbb{E}(X) = \int_\Omega X(\omega)dp(\omega),$$

which coincides with the previous summation formula for discrete random variables. We do not study Lebesgue integrals, though.

We can also use the distribution function of X:

$$\mathbb{E}(X) = \mu_X = \sum np(X^{-1}(\{n\})) = \sum np_X(n),$$

if this sum converges.

We define the *variance* of a discrete random variable X as

$$\sigma_X^2 = \mathbb{E}((X - \mathbb{E}(X))^2),$$

if this expected value exists.

We define the *k-th moment* of a discrete random variable X is $\mathbb{E}(X^k)$ if this expected value exists.

A useful identity is $\sigma_X^2 = \mathbb{E}(X^2) - \mathbb{E}^2(X)$.

6.1 Expected value and variance of the sum of random variables

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Here *it is not required that X and Y are independent.*
- If X and Y *are independent*: $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$.

For the case of product:

- If $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$ we say that X and Y are *uncorrelated.*
- If X and Y *are independent then they are also uncorrelated.*

6.2 Optional: Random variables as vectors

Let us take the example of a sample space Ω = {a,b,c}. We define a random variable X on Ω. We can represent this random variable as a three-component vector $(X(a), X(b), X(c)) \in \mathbb{R}^3$. A constant random variable 1 would be the vector $(1,1,1)$. The set of all random variables that we can define on Ω is a vector space.

We can define in this vector space a scalar product between two random variables as $\langle X, Y\rangle = \mathbb{E}(XY)$. The expected value of X corresponds to $\mathbb{E}(X) = \langle X, 1\rangle$, and give us the projection of the vector X on the axis (i.e. the subspace) created by the vector $(1,1,1)$. The variance is the square of the length of the projection of X on the plane that includes the origin and is perpendicular to the axis $(1,1,1)$.

The identity $\sigma^2_X + \mathbb{E}^2(X) = \mathbb{E}(X^2)$ is thus the Pythagoras theorem applied to these random variables.

We can extend the same ideas to a general sample space. When Ω has an infinite number of points (e.g. $\Omega = \mathbb{N}$) one have infinite dimensional vector spaces, which requires the use of the techniques of functional analysis. If we only consider the random variables X for which $\mathbb{E}(X^2) < \infty$ we say that the corresponding vector space of random variables is a *Hilbert Space* called l^2.

7 Common examples of discrete probability distributions

If we have a pair $(\Omega, \mathbb{A} = 2^\Omega)$ and a sequence of real numbers $\{p_n\}$ with $p_n \geq 0$ and $\sum p_n = 1$ $\{p_n\}$, we can define a probability map p with $p(\{\omega_n\}) = p_n$, and $p(E) = \sum_{\omega_n \in E} p_n$ for $E \in \mathbb{A}$. If we define a discrete random variable $X(\omega_n) = n$

we say that $P_X(n) = p_n$ is a *discrete probability distribution.* Common examples are:

7.1 Bernouilli

- $p_X(1) = p$ and $p_X(0) = 1 - p$ for $0 \leq p \leq 1$. $p_X(k) = 0$ for $k = 2, 3, ...$
- $\mathbb{E}(X) = p$; $Var(X) = p(1-p)$

7.2 Binomial

- $p_X(k) = \binom{N}{k}\ p^k(1-p)^{N-k}$ for $k = 0, ..., N$ and $0 \leq p \leq 1$.
- $\mathbb{E}(X) = pN$; $Var(X) = p(1-p)N$

7.3 Multinomial

- $p_{X_1,...,X_m}(k_1, ..., k_m) = \frac{N!}{k_1!...k_m!} p_1^{k_1} ... p_m^{k_m}$,
  for $0 \leq k_i \leq N$, $\sum k_i = N$, $0 \leq p_i \leq 1$, and $\sum p_i = 1$.
- $\mathbb{E}(X_i) = p_i N$; $Var(X_i) = p_i(1-p_i)N$; $Cov(X_i, X_j) = -Np_ip_j, i \neq j$.

7.4 Poisson

- $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, ...$ and $\lambda > 0$
- $\mathbb{E}(X) = \lambda$; $Var(X) = \lambda$.

7.5 Geometric

- $p_X(k) = p(1-p)^k$ for $k = 0, 1, ...$ and $0 \leq p \leq 1$
- $\mathbb{E}(X) = \frac{1-p}{p}$; $Var(X) = \frac{1-p}{p^2}$.

8 Inequalities

8.1 Markov and Chebychev inequalities

Assume that X only takes non-negative values (i.e. is a *non-negative random variable*), then:

$$\epsilon p(X \geq \epsilon) = \sum_{\omega : X(\omega) \geq \epsilon} \epsilon p(\omega) \leq \sum_{\omega : X(\omega) \geq \epsilon} X(\omega)p(\omega) \leq \sum_{\omega \in \Omega} X(\omega)p(\omega) = \mathbb{E}(X),$$

or alternatively,

$$\epsilon p(X \geq \epsilon) = \sum_{k \geq \epsilon} \epsilon p_X(k) \leq \sum_{k \geq \epsilon} k p_X(k) \leq \sum_{k \geq 0} k p_X(k) = \mathbb{E}(X),$$

leading to the *Markov inequality*:

$$p(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}.$$

Assume that X is a discrete random variable (which now can take positive and negative values). We can define a new *positive* random variable $Y = (X - \mathbb{E}(X))^2$ defined on $(\Omega, \mathbb{A})$. We have that the event $E \in \mathbb{A}$ which maps to $Y(E) \geq \epsilon^2$ is the same as the event which maps to $\mid \sqrt{Y(E)} \mid \geq \epsilon$. Applying the Markov inequality to this positive random variable we derive the *Chebychev inequality*:

$$p(\mid X - \mathbb{E}(X) \mid \geq \epsilon) = p((X - \mathbb{E}(X))^2 \geq \epsilon^2) \leq \frac{\sigma_X^2}{\epsilon^2}.$$

8.2 Jensen inequality

Another important result is the Jensen inequality. Assume that f(x) is a *convex real function*, defined in some interval I, i.e. $f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$ for $0 \leq \theta \leq 1$ for all $x_1, x_2 \in I$. Then

$$\mathbb{E}(f(X)) \leq f(\mathbb{E}(X)).$$

9 The (Weak) Law of Large Numbers

Assume $\{X_n\}$ is an infinite collection of independent joint random variables with the same expected value (i.e. $\mathbb{E}(X_n) = \mu, \forall n$ and variance (i.e. $\sigma_{X_n}^2 = \sigma^2, \forall n$).

Define a collection of random variables $\{X_n^*\}$ defined as $X_n^* = \frac{\sum_{k \leq n} X_k}{n}$

It is easy to see that: $\mathbb{E}(X_n^*) = \mu$ and $\sigma_{X_n^*}^2 = \frac{\sigma^2}{n}$.

The (Weak) Law of Large Numbers (LLN) says that the distribution of X_n^* concentrates around its average for large n:

$$p(\mid X_n^* - \mu \mid \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

9.1 Chernoff inequality

In our proof of the weak LLN we have used the Chebychev bound. In our derivation we had not required that the n random variables $\{X_n\}$ are jointly independent, but only that every pair of random variables are independent.

A stronger result can be obtained when the n random variables $\{X_k\}$ are in fact jointly independent, by using a general trick known as "Chernoff inequality".

Defining as before $X_n^* = \frac{\sum_{k\leq n} X_k}{n}$, the main idea is to obtain the Markov bound of the random variable $e^{sX_n^*}$ for a positive value s and minimize the obtained bound for $s > 0$, exploiting the joint independence of X_k.

As a simple but important example, let's apply this trick to the sum of n jointly independent Bernoulli r.v. and positive s:

$$p(X_n^* \leq a) = p(e^{sX_n^*} \leq e^{sa}) \leq \frac{\mathbb{E}(e^{sX_n^*})}{e^{sa}} \leq min_{s>0}\ \ e^{-sa}\prod_k \mathbb{E}(e^{sX_k}).$$

For Bernoulli r.v. we have: $\mathbb{E}(e^{sX_k}) = 1 + p(e^s - 1) \leq e^{p(e^s-1)}$, obtaining: $\prod_k \mathbb{E}(e^{sX_k}) \leq e^{np(e^s-1)}$.

If now we express a in terms of a positive factor δ as $a = (1+\delta)np$, and finding the value s that minimizes the bound we arrive to:

$$p(X_n^* \leq (1+\delta)np) \leq (\frac{e^\delta}{(1+\delta)^{(1+\delta)}})^{np}.$$

This expression gives a bound that decreases *exponentially* fast with n, which is a much tighter bound than the one given by using the Chebychev bound.

10 Optional: convergence of sequences of random variables

Assume that we have a succession of random variables $\{X_n\}_{n\in\mathbb{N}}$ defined on the same probability space $(\Omega, \mathbb{A}, p)$.

We say that the succession *converges weakly* to a random variable X defined on $(\Omega, \mathbb{A}, p)$, when:

$$\forall\epsilon > 0, \lim p(|Xn - X| \geq \epsilon) = 0.$$

We say that the succession *converges strongly* to a random variable X defined on $(\Omega, \mathbb{A}, p)$, when:

$$p(\{\omega \in \Omega; \lim |X_n(\omega) - X(\omega)| = 0\}) = 1.$$

Strong convergence implies weak convergence, but the reverse is not true.

Assume that we have a succession of discrete random variables $\{X_n\}_{n\in\mathbb{N}}$, which may be defined in different probability spaces. Assume that $F_{X_n}(x)= p(\{\omega; X_n(\omega) \in (-\infty, x]\})$ are the cumulative distribution functions for the random variables X_n. We say that we have *convergence in distribution* to another cumulative distribution function $F_X(x)$ when $\lim F_{X_n}(x) = F_X(x)$, for all the continuity points of $F_X(x)$.

11 Continuous probability spaces

In many applications, we define probability spaces with $\Omega = \mathbb{R}^n$ or $\Omega = \{0,1\}^\infty$, which are examples of non-countable sets. We will call these spaces *continuous* probability spaces. There are two main complications in the theory compared with the discrete case:

- Even for simple probabilistic experiments (e.g. picking a random number in the interval $[0,1]$ following a uniform distribution or in an infinite number of independent experiments of tossing a fair coin), one can find subsets of Ω for which probability cannot be assigned while fulfilling the properties of a proper probability map (see appendix). It is important to understand that this does not mean that the probability cannot be computed, or that the probability is zero, but that we cannot assign a probability for these subsets. We will thus define our probability space using a σ-algebra of events that do not include these *non-measurable* sets. It is also important to point out that the fact that a given subset is or is not measurable is fully dependent on the specific probability map that we are considering. In practice, non-measurable sets in continuous probability spaces are exotic sets, difficult to define. However, they must be excluded from the σ-algebra of events in order to have a correct theory. In the case of stochastic processes, though, many of the subsets that we consider in practical applications may be non-measurable.
- Many events will have a non numerable number of elements, meaning that knowing the probability for singletons will be not enough in order to find the probability of these events. We need thus an alternative method for assigning probability maps to events. As we will see the *extensions theorems* provide the adequate machinery for this.

11.1 σ-algebra generated by a collection of sets

Assume that we have a (possibly uncountable) collection of σ-algebras defined on the same set Ω. Then it is easy to proof that the intersection of all these σ-algebras is itself a σ-algebra. Assume also that C is a collection of subsets of Ω. We create the intersection of all σ-algebras on Ω that contain the subsets of C. The results of this intersection $\sigma(C)$ is again a σ-algebra, and it is the coarsest σ-algebra containing C. We say $\sigma(C)$ is the *σ-algebra generated by C*.

11.2 Semi-algebra of intervals and Borel σ-algebra

As we have already mentioned (see in the appendix of these lecture notes) when $\Omega = \mathbb{R}$ even in the simplest case of picking a random number uniformly in an interval, there are subsets of the sample space for which we cannot assign a probability. In these models, we often restrict the σ-algebra of events in our probability space to be the *Borel σ-algebra*, that we define now. The name comes from Emile Borel, a French mathematician who played a prominent role in the development of measure and probability theory.

As an example, assume $\Omega = \mathbb{R}$. Let J be the set of all intervals contained in Ω. This set J is a *semi-algebra* as: (i) the complementary of an interval can be expressed as the union of a *finite* number of disjoint intervals, (ii) a *finite* number of the intersection of intervals is another interval, and (iii) Ω and $\emptyset$ are intervals.

The *Borel σ-algebra* is the σ-algebra generated by J, $\sigma(J)$. We usually use the notation $\mathbb{B}(\Omega)$. In more general terms, the Borel σ-algebra is the sigma-algebra generated by the set of all open sets (according to some topology that defines the open sets).

When $\Omega = \{0,1\}^{\mathbb{N}}$ we use the *cylinder* σ-algebra instead.

11.3 Defining continuous probability spaces

For discrete probability spaces, we can assign a probability for every event $E \in \mathbb{A}$ by simply defining the probability for every singleton $p(\{\omega_n\})$, in such a way that fulfills some conditions ($p(\{\omega_n\}) \geq 0$, $\sum_{\omega_n \in \Omega} p(\{\omega_n\}) = 1$), and then use $p(E) = \sum_{\omega_n \in E} p(\{\omega_n\})$.

For a continuous probability space $(\Omega, \mathbb{A}, p)$ with $\Omega = \mathbb{R}$ or $\Omega = \{0,1\}^{\mathbb{N}}$, this procedure cannot be used as events, in general, are uncountable. We need then a procedure that assigns a probability measure to every event of the sigma-algebra $\mathbb{A}$. The σ-algebras we use in these cases (usually, the Borel or the cylinder σ-algebra) are an extremely rich collection of sets for which we cannot assign probabilities following a direct procedure.

The *Caratheodory extension theorem* provides a procedure to assign a probability to an arbitrary event of $\mathbb{A}$, and it is the standard method for defining continuous probability spaces.

11.4 Optional: The Caratheodory Extension theorem

Let's see how this is done for the case of the Borel σ-algebra. A similar procedure is used for the case of cylinder σ-algebra.

Assume $\Omega = \mathbb{R}$, and that we define a function set p that assigns a non-negative value $p(I_n)$ to each interval $I_n \in J$, in such a way that: (i) $p(\Omega) = 1$, and (ii) $p(\bigcup I_n) = \sum p(I_n)$ whenever $\{I_n\}$ is a countable collection of disjoint intervals of J with $I = \bigcup I_n \in J$.

The theorem says that under these conditions we can *extend* this function set

p, defined on J, to a probability map defined on the Borel σ-algebra generated by J, using the following formula:

$p(E) = \mathit{inf}\,\{\sum_{E\subseteq \cup_n I_n} p(I_n)\}$, with $E \in \sigma(J) = \mathbb{B}(\mathbb{R})$, and $I_n \in J$.

It can be shown that this defines a legitimate probability function on the probability triple $(\mathbb{R}, \mathbb{B}(\mathbb{R}), p)$. *This is the usual procedure used for defining probability maps on real numbers.*

12 Cumulative Distribution Function (CDF)

As we have seen, once we find a function set p with some specific properties on intervals, we can extend this to a probability map defined on the corresponding Borel σ-algebra.

The usual way of defining these probabilities for intervals is to use a function $F : \mathbb{R} \rightarrow [0, 1]$, which is:

- non decreasing,
- rigth continuous i.e. $\lim_{y\rightarrow x+} F(y) = F(x^+) = F(x), \forall x \in \mathbb{R}$.
- and with $F(-\infty) = 0$, and $F(+\infty) = 1$.

We define $p((-\infty, x]) = F(x)$, $p((x, -\infty]) = 1 - F(x)$, $p((-\infty, x)) = F(x^-)$, $p((y, x]) = F(x) - F(y^-)$, etc. In other words, we can determine a function set for any interval in J, and it can be proved that it fulfills the properties we require for the extension theorem. The function F is called a *cumulative distribution function, (CDF)*.

In many cases, we define *continuous random variables* as a mapping between an arbitrary set Ω equipped with a σ-algebra $\mathbb{A}$ to the set $\mathbb{R}$ (or an interval of $\mathbb{R}$) equipped with the Borel sigma-algebra. As we see, once we know the probability $p(\{X(\omega) \in (-\infty, x]\}) = p(\{X(\omega) \leq x\})$, we can assign a probability to any event in the borel-sigma algebra. Again we can define a CDF $F_X(x)$ as $p(\{X(\omega) \leq x\}) = F_X(x)$, where the subindex X is used to refer to the corresponding random variable.

12.1 Continuous and absolutely continuous random variables

A random variable X is *continuous* when every singleton in $\mathbb{R}$ has zero probability, in other words, $p(X = x) = 0$.

A random variable X is *absolutely continuous* when every set of length zero has zero probability. Not all the continuous random variables are also absolutely continuous (although the reverse is obviously true). However, most continuous random variables that we find in applications are also absolutely continuous.

12.2 Probability density functions

Let $F_X(x)$ be a Cumulative Distribution Function of a random variable X which is *absolutely continuous*. The function $f_X(x)$ defined by $f_X(x) = dF_X(x)dx$, when $F_X(x)$ is differentiable at x, is called the *probability density function* (PDF) of X.

This definition can be extended to multivariate CDFs. For instance, for absolutely continuous joint random variables X and Y, with CDF $F_{X,Y}(x, y)$, we can define the probability density function as $f_{X,Y}(x, y) = \partial^2 F_X(x)/\partial x \partial y$, at the points where the double partial derivative exists. The order of derivation is irrelevant.

We can give the following interpretation for the probability density function: $p(X \in (x, x + \Delta x]) = f_X(x)\Delta x + o(\Delta x)$. For the bivariate case we would have: $p(X \in (x, x + \Delta x], Y \in (y, y + \Delta y]) = f_{X,Y}(x, y)\Delta x\Delta y + o(\Delta x\Delta y)$.

Let D be a measurable set included in $\mathbb{R}$, then we have $p(X \in D) = \int_D f_X(x)\, dx$. For the bivariate case, we have that if D is a measurable set included in $\mathbb{R}^2$, then we have $p((X, Y) \in D) = \iint_D f_{XY}(x, y)\, dx\, dy$.

Regarding expected values, we will have $\mathbb{E}(X) = \int_D x\, f_X(x)\, dx$, and more in general $\mathbb{E}(g(X)) = \int_D g(x)\, f_X(x)\, dx$.

13 Common examples of continuous probability distributions

13.1 (Continuous) Uniform

- $f_X(x) = \frac{1}{b-a}$ if $x \in [a, b]$, and $f_X(x) = 0$ otherwise.
- $\mathbb{E}(X) = \frac{b+a}{2}$; $Var(X) = \frac{(b-a)^2}{12}$.

13.2 Exponential

- $f_X(x) = \lambda e^{-\lambda x}$ if $x \geq 0$, and $f_X(x) = 0$ otherwise, with $\lambda > 0$.
- $\mathbb{E}(X) = \frac{1}{\lambda}$; $Var(X) = \frac{1}{\lambda^2}$.

13.3 Univariate Gaussian

- $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ with $\sigma \neq 0$.
- $\mathbb{E}(X) = \mu$; $Var(X) = \sigma^2$.

13.4 Beta

- $f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ if $x \in [0,1]$, with $f_X(x) = 0$ otherwise. $\alpha > 0$, $\beta > 0$. $B(\alpha, \beta)$ is the *beta function*.
- $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$; $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

14 Multivariate Gaussian

14.1 Linear combination of independent Gaussians is a Gaussian

Let X and Y be two joint independent Gaussian distributions, with $X \sim$ Gaussian$(\mu_x, \sigma_x^2; x)$ and $Y \sim$ Gaussian$(\mu_y, \sigma_y^2; y)$. Then $Z = \alpha X + \beta Y$, where α and β are arbitrary constants, is a Gaussian distributions with $\mu_z = \alpha\mu_x + \beta\mu_y$ and $\sigma_z^2 = \alpha^2\sigma_x^2 + \beta^2\sigma_y^2$.

14.2 Multivariate Gaussian distribution

Let $\boldsymbol{X} = (X_1, ..., X_n)$ a *random vector* with components n joint random variables. We say that $\boldsymbol{X}$ follows a multivariate Gaussian distribution of parameters $\boldsymbol{\mu}$ and Σ, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and Σ is a *positive definite matrix*, if the joint probability density function of $\boldsymbol{X}$ is of the form:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

We use the notation $Gaussian(\boldsymbol{\mu}, \Sigma; \boldsymbol{x})$ for multivariate Gaussian distributions. The expected value of $\boldsymbol{X}$ is: $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, and Σ is the variance-covariance matrix.

14.3 The variance-covariance matrix

The variance-covariance matrix of a general (not necessarily multivariate Gaussian) random vector X is defined as

$$Cov(\boldsymbol{X}) = \mathbb{E}((\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))^\top)$$

In general $Cov(\boldsymbol{X})$ is a $n \times n$ positive semi-definite matrix, with components: $Cov(\boldsymbol{X})_{i,i} = \sigma^2_{X_i}$ and $Cov(\boldsymbol{X})_{i,j} = \sigma_{X_i X_j}$. If X_i and X_j are uncorrelated (for instance, if they are independent), $\sigma_{X_i X_j} = 0$.

14.4 The variance-covariance matrix of a multivariate Gaussian

For the specific case of $\boldsymbol{X}$ being a multivariate Gaussian, we have $Cov(\boldsymbol{X}) = \Sigma$. In this case, $\Sigma_{i,j} = 0$ implies that X_i and X_j are independent (and not simply uncorrelated). As an example, if all the components of the ransom vector $\boldsymbol{X}$ are independent random variables, the matrix Σ would be a diagonal matrix.

Note that in our definition of multivariate Gaussian rv we forced Σ to be positive definite (and not only positive semidefinite), meaning that it has always an inverse Σ^{-1} called the *precision matrix*.

Determinants of positive definite matrices are positive, meaning that $|\Sigma|^{1/2} \geq 0$ always.

14.5 Isocontour lines

For multivariate Gaussian, the points that fulfill the equation

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2$$

correspond to points for which the probability density function takes a constant value, $f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{c^2}{2}}$.

These points form an n-dimensional ellipsoid centered at $\boldsymbol{\mu}$.

The axis of this ellipsoid are aligned with $\boldsymbol{q_i}$, the eigenvectors of Σ (which are orthogonal vectors according to the spectral theorem, as Σ is an asymmetric definite positive matrix).

The semi-axis of this ellipsoid aligned with vector $\boldsymbol{q_i}$ has a length of $c\sqrt{\lambda_i}$, where λ_i is the eigenvalue associated with $\boldsymbol{q_i}$, which will be a positive real number as Σ is a symmetric definite positive matrix.

15 Some concepts of Information Theory

15.1 Entropy of a distribution

Assume that $(X_1, ..., X_N)$ is a vector of N joint iid random variables. Assume first that X_k follow a Bernoulli distribution with $p(X_k = 1) = p$ and $p(X_k = 0) = 1 - p$.

For each rv X_k we define a new random variable through the function $f(X_k)$ given by: $f(0) = -log_2(1-p)$ and $f(1) = -log_2(p)$. At first, this seems a *strange* self-referencing rv, as its values are dependent on the parameters of the distribution of X_k, but this is still a valid definition for a function.

With this definition for f, $(f(X_1), ..., f(X_N))$ is now a vector of N joint iid random variables following a Bernoulli distribution $p(f(X_k) = -log_2(p)) = p$, and $p(f(X_k) = -log_2(1-p)) = 1-p$.

We can apply the LLN to the components of this vector:

$$lim \;\; p(|\frac{\sum_{k=1}^{N} f(X_k)}{N} - \mathbb{E}(f(X_k))| \leq \epsilon) = 1.$$

or in other words, for every $\epsilon, \delta > 0$, there is a value N_0 so that $N > N_0$ implies:

$$p(|\frac{\sum_{k=1}^{N} f(X_k)}{N} - \mathbb{E}(f(X_k))| \leq \epsilon) \geq 1 - \delta.$$

This inequality tells us that the aggregated probability of the sequences $(X_1, ..., X_N)$ that fulfill the condition $|\frac{\sum_{k=1}^{N} f(X_k)}{N} - \mathbb{E}(f(X_k))| \leq \epsilon$ is almost 1. We call this set the *ϵ-typical set*, or simply the *typical set* (TS).

Note that the aggregated probability of the typical set is almost one, but this does not mean that the typical set contains the most probable sequences. For instance, if $p > 1/2$, the most probable sequence is $(1, ..., 1)$, which does not belong to the TS.

The expected value $H(X_k) = \mathbb{E}(f(X_k)) = -p \; log_2(p) - (1-p) \; log_2(1-p)$ is called the *entropy* of the random variable X_k.

Using a similar reasoning, for a discrete rv taking values in the set $\{0, ..., m-1\}$ with probabilities $p(i) = p_i$ we define the entropy $H(X)$ as:

$$H(X) = -\sum_{k=0}^{m-1} p_i \; log_2(p_i),$$

and the typical set as the set of sequences $(X_1, ..., X_N)$ for which

$$2^{-N(H+\epsilon)} \leq p(X_1, ..., X_N) \leq 2^{-N(H-\epsilon)}. \tag{15.1.1}$$

we know that the accumulated probability of these sequences in the TS is at least $1 - \delta$.

15.2 The Asymptotic Equipartition Property (AEP)

If we add the previous inequalities for all the sequences in the TS we obtain:

$$|TS|2^{-N(H+\epsilon)} \leq \sum_{TS} p(X_1, ..., X_N) \leq |TS|2^{-N(H-\epsilon)}$$

and:

$$(1-\delta) \leq \sum_{TS} p(X_1, ..., X_N) \leq 1,$$

which leads to the following bounds to the size of the TS:

$$(1-\delta)2^{N(H-\epsilon)} \leq |TS| \leq 2^{N(H+\epsilon)}.$$

As δ and ϵ can be chosen as close to zero as we want, this means that for large enough N the size of the TS is approximately 2^{NH} and that the sequences of the TS have probability approximately equal to 2^{-NH}. This is the Asymptotic Equipartition Property (AEP) of the TS.

15.3 Shannon source coding theorem

Assume that we want to compress the sequences $(X_1, ..., X_N)$ for very large N. A possible (not very practical) method would be the following:

- We number the sequences in the TS, using $log_2(|TS|) \leq NH$ bits (in fact, would be the ceiling function of NH).
- If a sequence belongs to the TS, we code the sequence as $(0, n)$, where n is the order number of the sequence in the TS (remember that this number can be represented using NH bits).
- If a sequence does not belong to the TS, we code it as $(1, X_1, ..., X_N)$, (in other words, we do not compress the sequence).

The expected number of bits required to represent a sequence would be:

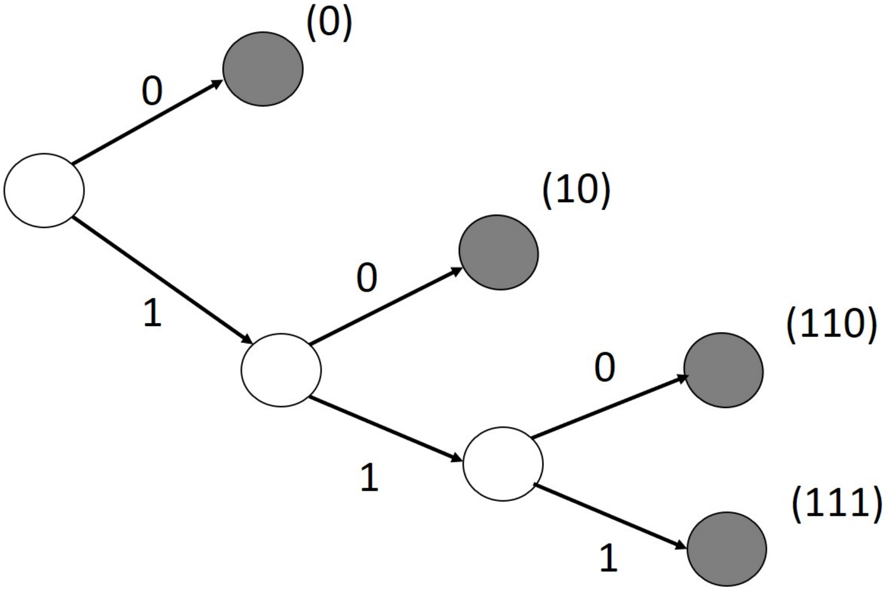
$$p((x_1, ..., X_N) \in TS)(NH+1) + (1 - p((x_1, ..., X_N) \in TS))(N+1).$$

As for large N $p((x_1, ..., X_N) \in TS)$ is near to 1, we obtain that the average number of bits to represent a sequence f length N would be $NH+1$, or in other words, *the factor of compression for large N would be the entropy H.*

15.4 Prefix-free codes and Kraft inequality

It is obvious that the coding method that we use in the proof of the Shannon source coding theorem is not very practical. An alternative way of coding sequences from a source that produces m symbols (i.e. symbols from the alphabet $\{1, ..., m\}$) would be the following:

Create a binary tree (in general a D-tree) of depth l_{max}. Pick m nodes in the trees to be used as codeword, in such as way that in the path between the root of the tree and a chosen node, there are no other chosen nodes. Codeword i is represented as a sequence of l_i bits, each bit indicating how we traverse the tree from the root until we reach the node. Assign the m nodes of the tree to the symbols in the alphabet.



For instance, for an alphabet of $m = 4$ symbols, a possible election for nodes in a binary tree would be $\{0, 10, 110, 111\}$. Let us assume the following symbols to codewords: $1 \longrightarrow$ ”0”, $2 \longrightarrow$ ”10”, $3 \longrightarrow$ ”110”, and $4 \longrightarrow$ ”0”. In this case, the sequence $(1, 3, 4, 4, 2)$ would be coded as: (01101110111010). Note that the original symbols can be recovered uniquely from the coded sequence. Moreover, once we have identified a code word in the compressed sequence, there is no question about whether this is the prefix or not of a valid code word (these are called ”prefix-free codes”).

The *Kraft inequality* states that there is a prefix-free code of lengths $l_1, ..., l_m$

iif:

$$\sum_{k=1..m} \frac{1}{2^{l_k}} \leq 1.$$

For instance, in our previous example we have $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = 1$.

Regarding the optimal assignment of code words to symbols, we can try to minimize the average compressed message length subject to the Kraft inequality constraint: Minimize $\sum_{k=1,\ldots,m} p(k) l_k$, subject to $\sum_{k=1..m} \frac{1}{2^{l_k}} \leq 1$.

We can use the Lagrange multiplication method, taking derivatives respect to l_k and to λ, obtaining $l_k^{opt} \geq -log_2(p_k)$, which leads to the result that the optimal average compression rate is upper bounded by the entropy of the source H.

15.5 Source coding using the wrong distribution. Cross entropy

Assume that our source of m symbols follows a distribution given by p. Assume also that we do not know the true distribution p and we estimate it instead by another distribution q. If we now source-code the symbols of the source, we would obtain compressed messages of compression rates:

$$H(p,q) = -\sum_{k=1,\ldots,m} p_k log_2(q_k).$$

This quantity is called *cross entropy*, and from our previous discussion, we expect to be a value larger than the entropy of the distribution p, i.e. $H(p,q) \leq H(p)$. This can be shown to be true using Jensen's inequality.

15.6 Kullback-Leibler (KL) divergence

The difference between the cross entropy of p and q with respect to the "true" entropy of p is a measure of how different are both distributions and it is called *Kullback-Leibler (KL) divergence*:

$$D(p||q) = H(p,q) - H(p) = -\sum_{k=1,\ldots,m} p_k \; log_2(\frac{q_k}{p_k}).$$

The KL divergence can be interpreted as a distance between two distributions p and q, although it is not a true distance as it is not symmetric, i.e. in general $D(p||q) \neq D(q||p)$.

15.7 Optional: Cross entropy as a cost function in supervised classification

Assume that we have a collection of pictures coded as a string of bits X. We want to design a classification function f, that takes a picture and assigns a probability of whether the picture belongs to class 1, e.g. "cat". Obviously, $1 - f$ would be the probability that the picture belongs o class -1, e.g. "cat".

In supervised machine learning, we have a set of n pictures for which we know the true class of each picture (i.e. a training set). Our function f belongs to a set of possible functions f_θ, parametrized by a value θ. Our goal is to find the optimal value of θ according to a given criterion, usually the value of θ that minimizes a loss function.

There are many possible cost functions that can be used. A popular choice is to use an estimation of the cross entropy between the results given by our classification function and the true values that we know in the training set (TS). Assume that $1(t_k)$ is an indicator function that takes the value 1 if $t_k == 1$, and the value 0 otherwise. We define a cross-entropy loss function as:

$$J(\theta) = \frac{1}{|TS|} \sum_{(x_k, t_k) \in \ TS} -\frac{1(t_k)}{n} log f_\theta(x_k) - \frac{(1 - 1(t_k))}{n} log(1 - f_\theta(x_k)).$$

and find the parameter θ that minimizes this loss function. Take into account that this function would also minimize the KL divergence between the true distribution and the distributions estimated by our classification function.

A Non-measurable sets

A.1 Example of a non-measurable result in an infinity fair coin tossing experiment

Assume that $\Omega = \{0, 1\}^{\mathbb{N}}$, i.e. the set of binary sequences indexed by $\mathbb{N}$. Assume that $\omega = (a_0, a_1.a_2, ...)$. We define the flipping operation $F^S(\omega)$ applied to ω, where S is a *finite* subset of $\mathbb{N}$ as follows: We flip the value of a_k from 0 to 1 or from 0 to 1 for the indexes that are contained in the set S. We can also apply the flipping operation to a set E by applying the flipping operation to every element of the set E.

For the fair coin-tossing experiment we would have $p(E) = p(F^S(E))$, provided that the probability $p(E)$ is well defined.

We are now establishing a partition of the set of binary sequences, by grouping in the same equivalence class sequences that are related by a flipping operation regarding a finite set S: $\omega, \gamma \in \Omega$, $\omega \sim \gamma$ when for some finite subset of integers S we have $\omega = F^S(\gamma)$. Here the fact that S is finite has a special importance.

Each equivalence class, V_x has a countable number of elements (as the number of finite subsets S is countable), and as Ω is noncountable, we will have a noncountable number of such equivalence classes.

The axiom of choice allows us to pick one element from each equivalence class V_x. Let us call U the set of all such elements. If we apply the F^S operation to the set U for all finite subsets S, we will generate the whole set on sequences: $\bigcup_S F^S(U) = \Omega$. These sets are disjoint (think about it), and by countable additivity of probability, we should have $\sum_S p(F^S(U)) = p(\Omega) = 1$. But all the terms in the sum have the same value (due to the flip invariant property of the fair coin toss experiment), meaning that we arrive at a contradiction: If $p(U) = 0$ the sum would be 0, whereas if $p(U) > 0$ the sum would be ∞.

This means that we cannot apply a probability to the set U in the case of a fair coin toss experiment, or in other words: *U is a non-measurable set.*

A.2 Measurability is not an intrinsic property of a set

It is important to note that a set is measurable or not depending on the specific probability map that we are considering. For instance, assume that the probability of "0" is 1, meaning that $p(\{(..0, 0, 0, ...)\}) = 1$, and $p(\{\omega\}) = 0$ otherwise. In this case if the the set U contains $(..0, 0, 0, ...)$ we would have $p(U) = 1$, while $p(U) = 0$ otherwise.

A.3 Example of a nonmeasurable result in a uniform probability map on an interval

See the *Vitali set.*