Sylvain Combettes

Master of Science candidate
Ecole des Mines de Nancy
Department of Applied Mathematics

✉ sylvain.combettes [a t] mines-nancy.org

https://github.com/sylvaincom

Supervisors: Alexandre Voisin, Pierre Vallois

Comparison of Empirical Probability Distributions

October 2019 – February 2020

Abstract

This document constitutes the report of my final-year project (one day per week) at Ecole des Mines de Nancy. The end goal of this project is to **compare two empirical probability distributions** from two different methods for computing the Choquet integral.

The first chapter is about the **Choquet integral**, a non-linear aggregation operator. I provide a lot of explanations and examples so that someone new to the Choquet integral can get a good understanding of it.

The second chapter is about **integral probability metrics** (IPMs), a popular estimation of distances on probabilities. In particular, we deal with the **Kantorovich metric** and the Dudley metric. We also study the empirical estimation of the Kantorovich metric and implement it with Python. Note that, under some conditions, the Kantorovich metric is the dual representation of the **Wasserstein distance**.

The third chapter is about f**-divergences**. f-divergences are another method for computing the distance between two probability distributions. In particular, we deal with the **Kullback-Leibler divergence**, the **Hellinger distance** and the **variational distance**. We also study the empirical estimation of these f-divergences and implement them with Python.

The fourth (and last) chapter applies the previous results on IPMs and f-divergences to the data obtained from the two methods for computing the Choquet integral.

All the codes can be found in my GitHub repository: `https://github.com/sylvaincom/comparison-distributions` and are not given in this report.

Contents

Acknowledgments

I would like to thank Alexandre Voisin and Pierre Vallois, my supervisors, for this great experience. I was able to learn a lot from them. I thank them for letting me conduct research in a autonomous way, for their advice and for creating a pleasant work atmosphere.

I would also like to thank Sandie Ferrigno and Antoine Henrot, from Ecole des Mines de Nancy, for making this project possible.

Introduction

Why compare data sets' distributions

The comparison of empirical data distributions from data sets is an important problem in modelling.

For example, in maintenance or **anomaly detection**, we try to detect or predict when a equipment will have a failure. Hence, we can use the comparison of empirical data distributions to detect when a distribution is "shifted" from its "normal" state, a "normal" state corresponding to a state with no failures and a "shifted" state thus corresponding to a failure.

In machine learning, we can try to estimate an unknown probability measure $\mathbb{P}$ by another probability measure $\mathbb{Q}$ with a **generative model**. Let us say that we draw n_q samples from $\mathbb{Q}$. Our goal is to verify that our generative model has drawn n_q samples (from $\mathbb{Q}$) that are similar to the n_p samples (from $\mathbb{P}$). It is a way of measuring the performance of $\mathbb{Q}$ as an estimate of $\mathbb{P}$ (which is unknown).

Moreover, we can try to predict if two data sets come from an actual person or from a **GAN** (generative adversarial network) which has been trained to copy the true person's data. Indeed, GANs are famous for their **deepfakes**: media that take a person in an existing image or video and replace them with someone else's likeness. By comparing empirical data distributions, we can try to determine if a video is fake or not.

Several methods for computing the distance between two empirical distributions exist and they can be split into two categories: **IPMs** (**integral probability metrics**) and f**-divergences**. The goal of this project is to list and compare these methods.

Context: comparing two methods for computing a Choquet integral

In this project, we will not use industrial nor real samples but synthetic ones.

Paper [Petot et al., 2018], submitted by my supervisors, Alexandre Voisin and Pierre Vallois, aims at proving an explicit formula that can give the distribution of the values taken by the **Choquet integral** with stochastic inputs. The Choquet integral [Choquet, 1954] is a non-linear aggregation operator that will be introduced in chapter I. The goal of this project is to show that this formula is correct from an empirical point of view.

For computing the Choquet integral, a direct method (and proven to be correct) is the **Monte-Carlo simulation**. In this project, we try to prove (experimentally) that the formula given in the paper [Petot et al., 2018] is correct by comparing its samples to the direct Monte-Carlo simulations' samples. The program computing the Choquet integral with the formula given in paper [Petot et al., 2018] has been improved in the project [Pajda and Castera, 2019] by Romain Pajda and Guillaume Castera.

Why comparing distributions is different from comparing samples

In the previous section, I explained that our goal is to compare data set distributions. How is comparing data set distributions different from comparing samples?

Let us suppose that we have n_p samples drawn from an unknown probability measure $\mathbb{P}$ and n_q samples drawn from another unknown probability measure $\mathbb{Q}$. Our goal is to check if the n_q samples (from $\mathbb{Q}$) are "similar" to the n_p samples (from $\mathbb{P}$). This notion of similarity will be defined more rigorously further in this report.

If $n_p = n_q$, a naive idea would be comparing the samples. Let X_p be the random variable under $\mathbb{P}$ with n_p samples and X_q the random variable under $\mathbb{Q}$ with n_q samples. Thus, comparing the samples amounts to computing the expectation $\mathbb{E}\left(|X_p - X_q|\right)$ which is the distance between two random variables.

However, we want to compare distributions and not random variables. Indeed, we can have $\mathbb{P} = \mathbb{Q}$, while $\mathbb{E}\left(|X_p - X_q|\right) \neq 0$. For example, that is the case if $X_p, X_q \sim \mathbb{U}[0,1]$ and the random variable have a finite (e.g. "small") number of samples.

Chapter I

The Choquet integral

This chapter is about the Choquet integral, a non-linear aggregation operator. I provide a lot of explanations and examples so that someone new to the Choquet integral can get a good understanding of it. Indeed, understanding deeply how the Choquet integral works can be very useful in order to find the right metric for comparison in the following chapters.

Note that appendix A deals with the numerical computation of the Choquet integral using MATLAB.

This chapter is mainly taken from paper [Petot et al., 2018].

I.1 Definition

I briefly recall the definitions of a capacity and the associated Choquet integral over a finite set $S \coloneqq \{1, 2, \ldots, n\}$. S is a set of criteria.

In order to define the Choquet integral of a vector with respect to a capacity, we need to define what a capacity is.

Definition 1 (Capacity μ).

A **capacity** μ over S is a function defined over the family $\mathscr{P}(S)$ of sets included in S, valued in $[0, 1]$ which is non-decreasing:

$$\mu(A) \leqslant \mu(B) \quad \forall A, B, A \subset B \subset S \tag{I.1}$$

and satisfying:

$$\mu(\varnothing) = 0, \quad \mu(S) = 1 \tag{I.2}$$

Now, I introduce an important notation for permutations that we will use to define the Choquet integral.

Notation 1 (Permutation σ)**.**

1. $\mathscr{S}_n$ stands for the group of permutations of S.
2. Let $x \coloneqq (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. There exists $\sigma \in \mathscr{S}_n$ such that:

$$x_{\sigma(1)} \leqslant x_{\sigma(2)} \leqslant \ldots \leqslant x_{\sigma(n)} \tag{I.3}$$

   σ is unique if $x_i \neq x_j$ for all $i \neq j$. σ depends on x.

3. If $\sigma : S \to S$, we set:

$$\sigma(a:b) \coloneqq \{\sigma(i), a \leqslant i \leqslant b\}, \quad 1 \leqslant a \leqslant b \leqslant n \tag{I.4}$$

   We also set $\sigma(a:b) = \varnothing$ if $a > b$.

Now, we can define the (discrete) Choquet integral.

Definition 2 (Choquet integral C of a vector x)**.**

Let μ be a capacity over S and $x \in \mathbb{R}^n$. The **Choquet integral of x with respect to μ** is the real number:

$$C_\mu(x) \coloneqq \sum_{i=1}^{n} x_{\sigma(i)} \left(\mu\left[\sigma\left(i:n\right)\right] - \mu\left[\sigma\left(i+1:n\right)\right]\right) \tag{I.5}$$

where σ is the permutation defined by (I.3).

Let us note that in (I.5), when $i = n$, we have $\mu[\sigma(n+1:n)] = \mu[\varnothing] = 0$ according to notation 1 and definition 1.

In the context of aggregation, the capacity $\mu(S_1)$ can be seen as the weight or importance of the subset $S_1 \in S$ of criteria in the decision. A graphical representation of a set is given in figure I.1. Let us note that to each node, meaning each set of criteria (such as node $\{1,4\}$), corresponds a value (in $[0,1]$) of the capacity.

An equivalent formula of (I.5) is the following:

$$C_\mu(x) = \sum_{i=1}^{n} \left(x_{\sigma(i)} - x_{\sigma(i-1)}\right) \mu[\sigma(i:n)] \tag{I.6}$$

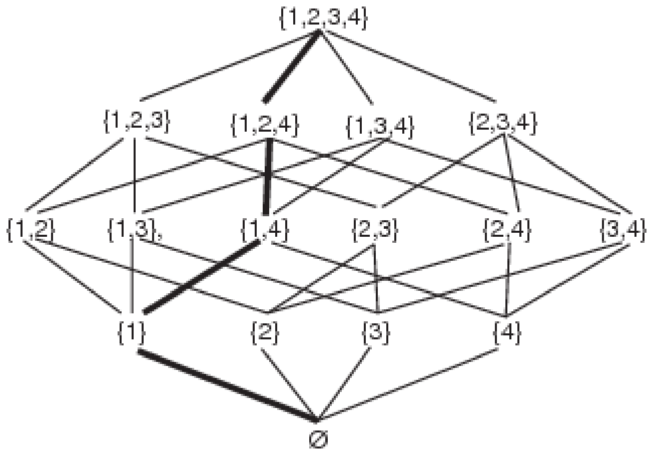


Figure I.1: An example of a graph over the set $S = \{1,2,3,4\}$. Source: [Kojadinovic, 2006].

Indeed, we have:

$$\begin{aligned}
C_\mu(x) &\coloneqq \sum_{i=1}^{n} x_{\sigma(i)} \left(\mu\left[\sigma\left(i:n\right)\right] - \mu\left[\sigma\left(i+1:n\right)\right]\right) \\
&= \sum_{i=1}^{n} x_{\sigma(i)} \mu\left[\sigma\left(i:n\right)\right] - \sum_{i=1}^{n} x_{\sigma(i)} \mu\left[\sigma\left(i+1:n\right)\right] \\
&= \sum_{i=1}^{n} x_{\sigma(i)} \mu\left[\sigma\left(i:n\right)\right] - \sum_{j=2}^{n+1} x_{\sigma(j-1)} \mu\left[\sigma\left(j:n\right)\right] \\
&= \sum_{i=1}^{n} x_{\sigma(i)} \mu\left[\sigma\left(i:n\right)\right] - \sum_{j=1}^{n} x_{\sigma(j-1)} \mu\left[\sigma\left(j:n\right)\right] + \underbrace{x_{\sigma(0)}}_{=0 \text{ (not defined)}} \mu\left[\sigma(1:n)\right] - x_{\sigma(n)} \underbrace{\mu\left[\sigma(n+1:n)\right]}_{=0} \\
&= \sum_{i=1}^{n} \left(x_{\sigma(i)} - x_{\sigma(i-1)}\right) \mu\left[\sigma\left(i:n\right)\right]
\end{aligned}$$

I.2 Examples

I.2.1 Explicit computation

To understand more clearly how formula (I.5) of the Choquet integral works, let us try an example taken from presentation [Kojadinovic, 2006]. Indeed, for someone new to the Choquet integral, it is quite difficult to understand formula (I.5). Moreover, at the end of this subsection, we give an interpretation of the Choquet integral from its graph.

We take the graph given in figure I.1. We have $n = 4$ and we assume $x_3 \leqslant x_2 \leqslant x_4 \leqslant x_1$. The path between the nodes that our Choquet integral takes is in bold in figure I.1. Let us note that the chosen path depends on the values of x and not the values of μ. According to notation 1, we have:

$$\begin{aligned}
\sigma(1) &= 3 \\
\sigma(2) &= 2 \\
\sigma(3) &= 4 \\
\sigma(4) &= 1
\end{aligned}$$

According to formula (I.5), we have:

$$\begin{aligned}
C_\mu(x_1,x_2,x_3,x_4) &= x_{\sigma(1)}\left(\mu[\sigma(1:4)]-\mu[\sigma(2:4)]\right)\\
&\quad + x_{\sigma(2)}\left(\mu[\sigma(2:4)]-\mu[\sigma(3:4)]\right)\\
&\quad + x_{\sigma(3)}\left(\mu[\sigma(3:4)]-\mu[\sigma(4:4)]\right)\\
&\quad + x_{\sigma(4)}\left(\mu[\sigma(4:4)]-\mu[\sigma(5:4)]\right)\\
&= x_3\left(\mu\left[\{\sigma(1),\sigma(2),\sigma(3),\sigma(4)\}\right]-\mu\left[\{\sigma(2),\sigma(3),\sigma(4)\}\right]\right)\\
&\quad + x_2\left(\mu\left[\{\sigma(2),\sigma(3),\sigma(4)\}\right]-\mu\left[\{\sigma(3),\sigma(4)\}\right]\right)\\
&\quad + x_4\left(\mu\left[\{\sigma(3),\sigma(4)\}\right]-\mu\left[\{\sigma(4)\}\right]\right)\\
&\quad + x_1\left(\mu[\{\sigma(4)\}]-\mu[\varnothing]\right)\\
&= x_3\left(\mu[\{3,2,4,1\}]-\mu[\{2,4,1\}]\right)\\
&\quad + x_2\left(\mu[\{2,4,1\}]-\mu[\{4,1\}]\right)\\
&\quad + x_4\left(\mu[\{4,1\}]-\mu[\{1\}]\right)\\
&\quad + x_1\left(\mu[\{1\}]-\mu[\varnothing]\right)
\end{aligned}$$

It is important to visualize the previous formula $C_\mu(x_1,x_2,x_3,x_4) = x_3\left(\mu[\{3,2,4,1\}]-\mu[\{2,4,1\}]\right) + x_2\left(\mu[\{2,4,1\}]-\mu[\{4,1\}]\right) + x_4\left(\mu[\{4,1\}]-\mu[\{1\}]\right) + x_1\left(\mu[\{1\}]-\mu[\varnothing]\right)$ with $x_3 \leqslant x_2 \leqslant x_4 \leqslant x_1$ on the bold path in figure I.1. The nodes in the path are respectively $\varnothing, \{1\}, \{1,4\}, \{1,2,4\}$ and $\{1,2,3,4\}$. Thus, at each step on the graph, we add the index of remaining highest value to our current set.

We start from $\varnothing$. We move to the node corresponding to the highest value of x which is x_1 and multiply the value x_1 at the node $\{1\}$ by the difference between the capacity of the current node and the previous one, thus getting the term $x_1\left(\mu[\{1\}]-\mu[\varnothing]\right)$. Then, we add the highest value to our current set $\{1\}$ which is x_4, thus moving to node $\{1,4\}$. To our previous sum, we add x_4 times the difference between the capacities of the current node and the previous one, thus getting the term $x_4\left(\mu[\{4,1\}]-\mu[\{1\}]\right)$. The process goes on. Let us recall that the chosen path depends on the values of the input x and not the values of the capacity μ.

I.2.2 Why taking the Choquet integral of a vector can be interesting

In this subsection, I give a basic example explaining why the Choquet integral can be highly relevant.

Let us suppose that we have an individual named Lucas. Lucas wishes to go to a restaurant. He can choose between restaurants A, B, C and D. Each restaurant has three dishes: meat, fish and pizza. For each restaurant and each dish, we are given a grade from 0 to 20, grade 20 corresponding to the best possible dish. The grades are given in table I.1.

restaurant	meat	fish	pizza
A	18	15	19
B	15	18	19
C	15	18	11
D	18	15	11

Table I.1: Grades of each dish according to the restaurant

For choosing the restaurant, Lucas' criteria are the following:

- For two restaurants, if the pizzas are equally good, then Lucas prefers the restaurant with the best meat. Example: Lucas prefers A to B.
- For two restaurants, if the pizzas are equally bad, then Lucas prefers the restaurant with the best fish. Example: Lucas prefers C to D.

The above criteria define the values of our capacity μ. However, in this report, I do not compute the exact values of the capacity ourselves, but I trust R's library `kappalab`.

Our problem is the following: between A, B, C and D, which restaurant is Lucas going to choose?

If we choose to solve our problem using the mean of each restaurant, we obtain:

- For restaurant A, we get $\frac{18+15+19}{3} \approx 17.33$, thus mean(A) $\approx$ 17.33
- For restaurant B, we get $\frac{15+18+19}{3} \approx 17.33$, thus mean(B) $\approx$ 17.33
- For restaurant C, we get $\frac{15+18+11}{3} \approx 14.67$, thus mean(C) $\approx$ 14.67
- For restaurant D, we get $\frac{18+15+11}{3} \approx 14.67$, thus mean(D) $\approx$ 14.67

Hence, we have mean(A) = mean(B) and mean(C) = mean(D) so we can not distinguish A from B, nor C from D. Indeed, the mean operator does not take into account Lucas' criteria.

However, if we use the Choquet integral C_μ, we get:

- For restaurant A, we get $C_\mu(A) \approx 17.83$
- For restaurant B, we get $C_\mu(B) \approx 16.83$
- For restaurant C, we get $C_\mu(C) \approx 15.17$
- For restaurant D, we get $C_\mu(D) \approx 14.17$

Hence, we have $C_\mu(A) > C_\mu(B) > C_\mu(C) > C_\mu(D)$ so we can distinguish the restaurants. Lucas is going to prefer restaurant A.

For computing the previous Choquet integrals, I used the following R code given in presentation [Kojadinovic, 2006] (where more details are given):

```r
install.packages('kappalab')
library(kappalab)

a = c(18,15,19)
b = c(15,18,19)
c = c(15,18,11)
d = c(18,15,11)

delta.C = 1
Acp = rbind(c(a,b,delta.C),c(c,d,delta.C))

s = mini.var.capa.ident(3,3,A.Choquet.preorder = Acp)

mu = zeta(s$solution)
mu

Choquet.integral(mu,a)
Choquet.integral(mu,b)
Choquet.integral(mu,c)
Choquet.integral(mu,d)
```

Chapter II

Integral Probability Metrics (IPMs)

This chapter is mainly taken from paper [Sriperumbudur et al., 2012].

Given samples from two unknown probability measures, it is often of interest to estimate the distance (or divergence) between them. Integral probability metrics (IPMs) is a popular empirical estimation of distances on probabilities. In this report, we will only focus on two of the most popular IPMs: the Kantorovich metric W and the Dudley metric β. For the empirical estimation, we will only deal with the Kantorovich metric W, as it seems to be the most common.

II.1 General definition of IPMs

In this section, we do not take into account the empirical aspect of distributions: we will focus on the empirical aspect starting from section II.2. Thus, we assume that the probability measures are known here.

Definition 3 (Integral probability metric γ)**.**

Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable space S, the **integral probability metric** (IPM) giving the distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\gamma_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathscr{F}} \left| \int_S f \,\mathrm{d}\mathbb{P} - \int_S f \,\mathrm{d}\mathbb{Q} \right| \tag{II.1}$$

where $\mathscr{F}$ is a class of real-valued bounded measurable functions on S.

There are several choices of functions $\mathscr{F}$. The choice of $\mathscr{F}$ is the crucial distinction between different IPMs: each choice of $\mathscr{F}$ leads to a specific IPM.

With formula (II.1), we understand the term "integral" in "integral probability metrics". Indeed, we evaluate a complicated mathematical object – a probability – with an integral, which is easier to manipulate.

Note that for the integrand f, we choose functions that are bounded so that the integrals exist. Moreover, $\mathscr{F}_W$ is restricted so that it is easier to get the sup in formula (II.1).

II.1.1 Kantorovich metric

Definition 4 (Kantorovich metric W)**.**

Setting

$$\mathscr{F} = \{f : \|f\|_L \leqslant 1\} \tag{II.2}$$

in (II.1) yields the **Kantorovich metric** W, where $\|f\|_L$ is the Lipschitz semi-norm of a bounded continuous real-valued function f on a metric space (S, ρ):

$$\|f\|_L \coloneqq \sup\left\{\frac{|f(x) - f(y)|}{\rho(x, y)} : x \neq y \text{ in } S\right\} \tag{II.3}$$

Note that the index L in $\|f\|_L$ stands for Lipschitz.

We introduce the notation $\mathscr{F}_W = \{f : \|f\|_L \leqslant 1\}$.

Note that, under some conditions, the Kantorovich metric is the dual representation of the Wasserstein distance. Indeed, according to the Kantorovich-Rubinstein theorem, when S is separable, the Kantorovich metric W is the dual representation of the Wasserstein distance, more specifically, the $\mathscr{L}_1$-Wasserstein distance W_1. Hence, we understand better the notation W for the Kantorovich metric.

II.1.2 Dudley metric

Definition 5 (Dudley metric β)**.**

Setting

$$\mathscr{F} = \{f : \|f\|_{BL} \leqslant 1\} \tag{II.4}$$

in (II.1) yields the dual-bounded Lipschitz distance – also called the **Dudley metric** β – where:

$$\|f\|_{BL} \coloneqq \|f\|_\infty + \|f\|_L \tag{II.5}$$

with $\|f\|_\infty \coloneqq \sup\{|f(x)| : x \in S\}$.

Note that the index BL in $\|f\|_{BL}$ stands for (dual-)bounded Lipschitz.

We introduce the notation $\mathscr{F}_\beta = \{f : \|f\|_{BL} \leqslant 1\}$.

II.1.3 Example: explicit computation of the Kantorovich metric

In this subsection, I give an example of explicit computation of the Kantorovich metric W to better understand how it works and why it measures a distance between two distributions. For now, we will not deal with the estimator of W.

For ease of computation, let us consider $\mathbb{P}$ and $\mathbb{Q}$ defined on the Borel-algebra of $\mathbb{R}^d$ as product measures:

$$\mathbb{P} = \otimes_{i=1}^{d} \mathbb{P}^{(i)} \quad \text{and} \quad \mathbb{Q} = \otimes_{i=1}^{d} \mathbb{Q}^{(i)} \tag{II.6}$$

where $\mathbb{P}^{(i)}$ and $\mathbb{Q}^{(i)}$ are defined on the Borel σ-algebra of $\mathbb{R}$. In this setting, when $\rho(x,y) = \|x-y\|_1$, it can be shown that:

$$W(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{d} W\left(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}\right) \tag{II.7}$$

where:

$$W\left(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}\right) = \int_{\mathbb{R}} \left|F_{\mathbb{P}^{(i)}}(x) - F_{\mathbb{Q}^{(i)}}(x)\right| \mathrm{d}x \tag{II.8}$$

and:

$$F_{\mathbb{P}^{(i)}}(x) = \mathbb{P}^{(i)}\left((-\infty, x]\right) \tag{II.9}$$

Let $S = \times_{i=1}^{d}[a_i, s_i]$. Suppose $\mathbb{P}^{(i)} = U[a_i, b_i]$ and $\mathbb{Q}^{(i)} = U[r_i, s_i]$, which are uniform distributions on $[a_i, b_i]$ and $[r_i, s_i]$, respectively, where $-\infty < a_i \leqslant r_i \leqslant b_i \leqslant s_i < \infty$. Then, we have:

$$W\left(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}\right) = \frac{s_i + r_i - a_i - b_i}{2} \tag{II.10}$$

and $W(\mathbb{P},\mathbb{Q})$ follows (II.7).

In this example, as can be seen in equation (II.10), the Kantorovich metric W relies on different simulations (the $\mathbb{P}^{(i)}$ and $\mathbb{Q}^{(i)}$) and depends on the parameters a_i, b_i, r_i and s_i of the uniform distributions (and not the values of the realizations of the random variables). Once we chose all the a_i, b_i, r_i and s_i, then $W\left(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}\right)$ is completely determined: it is a real number (and not random variable for example).

How can we interpret the Kantorovich metric W as a distance?

According to (II.10), if we choose the intervals $[a_i, b_i]$ and $[r_i, s_i]$ to be equal, then the distance according to the Kantorovich metric is null, which is intuitive considering the properties of a mathematical distance.

If we choose the intervals to be slightly shifted from each other:

$$\forall i \in [\![1, d]\!] \quad \begin{cases} a_i = 0 \\ b_i = 1 \\ r_i = 1/4 \\ s_i = 5/4 \end{cases}$$

then:

$$\forall i \in [\![1, d]\!] \quad W\left(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}\right) = \frac{1}{4} \quad \text{so that} \quad W(\mathbb{P},\mathbb{Q}) = \frac{d}{4} > 0$$

Moreover, when the dimension d increases, the distance $W(\mathbb{P},\mathbb{Q})$ increases, which is intuitive.

If we choose the intervals to be more shifted from each other:

$$\forall i \in [\![1, d]\!] \quad \begin{cases} a_i = 0 \\ b_i = 1 \\ r_i = 1/2 \\ s_i = 3/2 \end{cases}$$

then:

$$\forall i \in [\![1, d]\!] \quad W\left(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}\right) = \frac{1}{2} \quad \text{so that} \quad W(\mathbb{P}, \mathbb{Q}) = \frac{d}{2} > \frac{d}{4} \tag{II.11}$$

and the distance is greater than previously, which is intuitive.

Hence, the Kantorovich metric W measures the distance between two distributions and increases when the difference between the parameters of the distributions increases. That is why, in section II.4, we are going to carry out experiments to see how the empirical Kantorovich metric of two probability distributions evolves with their parameters.

II.2 Definition: empirical estimation of IPMs

In practice, we are not always able to explicitly compute an IPM, we often need to compute an estimator of the IPM instead. That is the case when $\mathbb{P}$ and $\mathbb{Q}$ are unknown but simulable distributions for example.

In simple words, the empirical estimator of an IPM is the discrete version of formula (II.1).

Definition 6 (Empirical estimator of an IPM)**.**

Given $\left\{X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}\right\}$ and $\left\{X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}\right\}$, which are i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$, respectively, the **empirical estimator of** $\gamma_{\mathscr{F}}(\mathbb{P}, \mathbb{Q})$ is given by

$$\gamma_{\mathscr{F}}(\mathbb{P}_m, \mathbb{Q}_n) := \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{N} \widetilde{Y}_i f(X_i) \right| \tag{II.12}$$

where

$$\mathbb{P}_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i^{(1)}} \quad \text{and} \quad \mathbb{Q}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i^{(2)}}$$

represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$, respectively,

$$N = n + m$$

and

$$\widetilde{Y}_i = \frac{1}{m} \text{ when } X_i = X_i^{(1)} \text{ for } i = 1, \ldots, m$$
$$\widetilde{Y}_{m+i} = -\frac{1}{n} \text{ when } X_{m+i} = X_i^{(2)} \text{ for } i = 1, \ldots, n$$

Let us emphasize that formula (II.12) is a definition and not a result.

Note that the empirical estimator of an IPM inputs the samples directly (and not the empirical probability distributions).

How is formula (II.12) the discrete version of equation (II.1)? We have:

$$\begin{aligned}
\gamma_{\mathscr{F}}\left(\mathbb{P}_m, \mathbb{Q}_n\right) &\coloneqq \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{N} \widetilde{Y}_i f\left(X_i\right) \right| \\
&= \sup_{f \in \mathscr{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f\left(X_i\right) - \frac{1}{n} \sum_{i=m+1}^{N} f\left(X_i\right) \right| \\
&= \sup_{f \in \mathscr{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f\left(X_i^{(1)}\right) - \frac{1}{n} \sum_{j=1}^{n} f\left(X_i^{(2)}\right) \right|
\end{aligned}$$

In this report, for the empirical estimation of IPMs, we are only going to deal with the Kantorovich metric.

II.3 Empirical estimator of the Kantorovich metric

II.3.1 Definition

In order to compute the empirical estimator of the Kantorovich metric W, we need to find the function f that realizes the sup and that solves (II.12) for $\mathscr{F} = \mathscr{F}_W$. Sometimes, we will only be able to approximate f numerically.

Theorem 1 (Empirical estimator of the Kantorovich metric)**.**

For all $\alpha \in [0,1]$, the following definition solves (II.12) for $\mathscr{F} = \mathscr{F}_W$:

$$\forall x \in S \quad f_\alpha(x) \coloneqq \alpha \min_{i=1,\dots,N} \left(a_i^\star + \rho\left(x, X_i\right)\right) + (1-\alpha) \max_{i=1,\dots,N} \left(a_i^\star - \rho\left(x, X_i\right)\right) \tag{II.13}$$

where

$$W\left(\mathbb{P}_m, \mathbb{Q}_n\right) = \sum_{i=1}^{N} \widetilde{Y}_i a_i^\star \tag{II.14}$$

and $\left\{a_i^\star\right\}_{i=1}^{N}$ solve the following linear program,

$$\max_{a_1,\dots,a_N} \left\{ \sum_{i=1}^{N} \widetilde{Y}_i a_i : -\rho\left(X_i, X_j\right) \leqslant a_i - a_j \leqslant \rho\left(X_i, X_j\right), \forall i, j \right\} \tag{II.15}$$

Here, we have a rigorous theorem (based on definition 6) that is proven in paper [Sriperumbudur et al., 2012].

In this project, our goal is to compute the empirical estimator of the Kantotovich metric $W\left(\mathbb{P}_m, \mathbb{Q}_n\right)$ as given in (II.14). Hence, the value of the objective function solving (II.15) will be the "distance" we are looking for. Note that we do not need to compute f_α given by (II.13).

How can we implement theorem 1 into an algorithm? The following subsection will answer our question, as (II.15) is a linear programming problem.

II.3.2 Solving the linear programming problem

How can we solve the linear programming problem (II.15)?

First of all, we must write the linear programming problem (II.15) into its canonical form:

$$\begin{cases} \max\left(c^T a\right) \\ Ma \leqslant b \end{cases} \tag{II.16}$$

where $a = (a_1, \dots, a_N)$ are the variables of the problem, $c = (c_1, \dots, c_N)$ are the coefficients of the objective function, $M \in \mathcal{M}_{p,N}$ and $b = (b_1, \dots, b_p)$ are the non-negative constraints.

Let us transform formula (II.15) into its canonical form (II.16) in a simple case: $N = 4$ (for example).

The objective function is:

$$\sum_{i=1}^{N} \widetilde{Y}_i a_i$$

and amounts to:

$$\widetilde{Y}_1 a_1 + \widetilde{Y}_2 a_2 + \widetilde{Y}_3 a_3 + \widetilde{Y}_4 a_4 \tag{II.17}$$

Hence, we can relate (II.17) to (II.16) with:

$$c = \begin{pmatrix} \widetilde{Y}_1 \\ \widetilde{Y}_2 \\ \widetilde{Y}_3 \\ \widetilde{Y}_4 \end{pmatrix} \tag{II.18}$$

The constraints:

$$\forall (i,j) \in [\![1, N]\!]^2 \quad -\rho\left(X_i, X_j\right) \leqslant a_i - a_j \leqslant \rho\left(X_i, X_j\right)$$

amount to:

$$\begin{aligned}
\begin{cases}
-\rho\left(X_1,X_2\right) \leqslant a_1 - a_2 \leqslant \rho\left(X_1,X_2\right)\\
-\rho\left(X_1,X_3\right) \leqslant a_1 - a_3 \leqslant \rho\left(X_1,X_3\right)\\
-\rho\left(X_1,X_4\right) \leqslant a_1 - a_4 \leqslant \rho\left(X_1,X_4\right)\\
-\rho\left(X_2,X_3\right) \leqslant a_2 - a_3 \leqslant \rho\left(X_2,X_3\right)\\
-\rho\left(X_2,X_4\right) \leqslant a_2 - a_4 \leqslant \rho\left(X_2,X_4\right)\\
-\rho\left(X_3,X_4\right) \leqslant a_3 - a_4 \leqslant \rho\left(X_3,X_4\right)
\end{cases}
&\iff
\begin{cases}
a_1 - a_2 \leqslant \rho\left(X_1,X_2\right)\\
a_1 - a_2 \leqslant \rho\left(X_1,X_2\right)\\
a_1 - a_3 \leqslant \rho\left(X_1,X_3\right)\\
a_3 - a_1 \leqslant \rho\left(X_1,X_3\right)\\
a_1 - a_4 \leqslant \rho\left(X_1,X_4\right)\\
a_4 - a_1 \leqslant \rho\left(X_1,X_4\right)\\
a_2 - a_3 \leqslant \rho\left(X_2,X_3\right)\\
a_3 - a_2 \leqslant \rho\left(X_2,X_3\right)\\
a_2 - a_4 \leqslant \rho\left(X_2,X_4\right)\\
a_4 - a_2 \leqslant \rho\left(X_2,X_4\right)\\
a_3 - a_4 \leqslant \rho\left(X_3,X_4\right)\\
a_4 - a_3 \leqslant \rho\left(X_3,X_4\right)
\end{cases}\\
&\iff
\begin{pmatrix}
1 & -1 & 0 & 0\\
-1 & 1 & 0 & 0\\
1 & 0 & -1 & 0\\
-1 & 0 & 1 & 0\\
1 & 0 & 0 & -1\\
-1 & 0 & 0 & 1\\
0 & 1 & -1 & 0\\
0 & -1 & 1 & 0\\
0 & 1 & 0 & -1\\
0 & -1 & 0 & 1\\
0 & 0 & 1 & -1\\
0 & 0 & -1 & 1
\end{pmatrix}
\begin{pmatrix} a_1\\ a_2\\ a_3\\ a_4 \end{pmatrix}
\leqslant
\begin{pmatrix}
\rho\left(X_1,X_2\right)\\
\rho\left(X_1,X_2\right)\\
\rho\left(X_1,X_3\right)\\
\rho\left(X_1,X_3\right)\\
\rho\left(X_1,X_4\right)\\
\rho\left(X_1,X_4\right)\\
\rho\left(X_2,X_3\right)\\
\rho\left(X_2,X_3\right)\\
\rho\left(X_2,X_4\right)\\
\rho\left(X_2,X_4\right)\\
\rho\left(X_3,X_4\right)\\
\rho\left(X_3,X_4\right)
\end{pmatrix}
\end{aligned} \tag{II.19}$$

Hence, we can relate (II.19) to (II.16) with:

$$M = \begin{pmatrix}
1 & -1 & 0 & 0\\
-1 & 1 & 0 & 0\\
1 & 0 & -1 & 0\\
-1 & 0 & 1 & 0\\
1 & 0 & 0 & -1\\
-1 & 0 & 0 & 1\\
0 & 1 & -1 & 0\\
0 & -1 & 1 & 0\\
0 & 1 & 0 & -1\\
0 & -1 & 0 & 1\\
0 & 0 & 1 & -1\\
0 & 0 & -1 & 1
\end{pmatrix} \in \mathscr{M}_{p,N} \quad \text{and} \quad b = \begin{pmatrix}
\rho\left(X_1,X_2\right)\\
\rho\left(X_1,X_2\right)\\
\rho\left(X_1,X_3\right)\\
\rho\left(X_1,X_3\right)\\
\rho\left(X_1,X_4\right)\\
\rho\left(X_1,X_4\right)\\
\rho\left(X_2,X_3\right)\\
\rho\left(X_2,X_3\right)\\
\rho\left(X_2,X_4\right)\\
\rho\left(X_2,X_4\right)\\
\rho\left(X_3,X_4\right)\\
\rho\left(X_3,X_4\right)
\end{pmatrix} \in \mathbb{R}^p \tag{II.20}$$

From (II.20), we can easily infer that:

$$
\begin{aligned}
p &= 2 \times \left(\sum_{1 \leqslant i < j \leqslant N} 1 \right) && \text{(II.21)} \\
&= 2 \times \left(\sum_{l=1}^{N} l - N \right) && \text{(II.22)} \\
&= 2 \times \left(\sum_{l=1}^{N-1} l \right) && \text{(II.23)} \\
&= 2 \times \left(\frac{N(N-1)}{2} \right) && \text{(II.24)} \\
&= N(N-1) && \text{(II.25)}
\end{aligned}
$$

Hence, we have defined our linear programming problem in the canonical form.

Now, we can try solving it with the Simplex algorithm. Actually, we will use the `PuLP` library from Python [Mitchell et al., 2011] to solve it.

The complexity of our linear programming problem grows fast because $M \in \mathscr{M}_{p,N}$ and $b \in \mathbb{R}^p$ with $p = N(N-1)$. For example, if we take $N = 200$ (meaning only 200 samples total for both distributions), we have $p = 200 \times 199 = 39\ 800$, meaning $p = 39\ 800$ constraints, which is already quite huge to solve! Moreover, the number of values in M is $p \times N = 39\ 800 \times 200 = 7\ 960\ 000$, thus approximately 8 million! Thus, in this report, from a numerical view point, we will only be able to compute empirical Kantorovich metrics for short numbers of samples (less than 1 000 total samples). Note that we are confronted with a memory error while the processing time is quite reasonable (less than 30 seconds for $N = 300$).

For further work, as M contains a lot of zeros, we could try coding it as a sparse matrix. However, I am not sure if the `PuLP` library can work with sparse matrices. Moreover, dealing with M as a sparse matrix does not change the fact that we have a lot of constraints.

Thus, in this report, computing the Kantorovich metric can only be done when we have less than 1 000 total samples (which is sadly not much!).

II.4 How does the Kantorovich metric of two probability distributions evolve with their parameters?

In this section, we are going to interpret the empirical Kantorovich metric W by running several simulations using Python. We will always choose $\rho(x,y) = |x-y|$ for $(x,y) \in \mathbb{R}^2$ as the space metric, because the samples will always be one-dimensional.

We are going to compare two (empirical) normal distributions by modifying their parameters (means and standard deviations) and interpret how their Kantorovich metric evolves. We are going to do the same with two exponential distributions and two uniform distributions.

The aim is to see how the distance evolves according to the parameters but also to set some reference values. For example, reference values will allow to say that a value of an IPM of value 1 is "low", meaning that the distributions are "close".

Let us recall that, contrary to f-divergences, IPMs input the empirical samples, and not the empirical probability distributions. Moreover, due to the memory issue, we can only compute the Kantorovich metric for a total number of samples that is inferior to 1 000.

In this chapter, to fit a model, we use the the linear regression function from `scikit-learn` [Pedregosa et al., 2011]. The `score` method returns the coefficient of determination R^2 of the prediction.

II.4.1 Comparison of two normal distributions

In this subsection, we consider two normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$. X_p are the samples drawn from $\mathbb{P}$ and X_q are the samples drawn from $\mathbb{Q}$. Let n_p be the number of samples generated from $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and n_q the number of samples generated from $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$.

II.4.1.1 Influence of the difference between the means

First of all, let us plot the histograms from the samples to visualize how the Kantorovich metric evolves with the difference $\mu_q - \mu_p$. Note that the data is samples, from which we have drawn histogram, then approximated their density. See figure II.1. All the parameters are the same, only the means are different. On the graphic on the left, we have $\mu_q - \mu_p = 1$ and $W = 3.721$, while on the right we have $\mu_q - \mu_p = 5$ and $W = 7.459$. Hence, when the difference between the means increases (thus the distributions are "more different"), the Kantorovich metric W increases which justifies why W is a "distance".

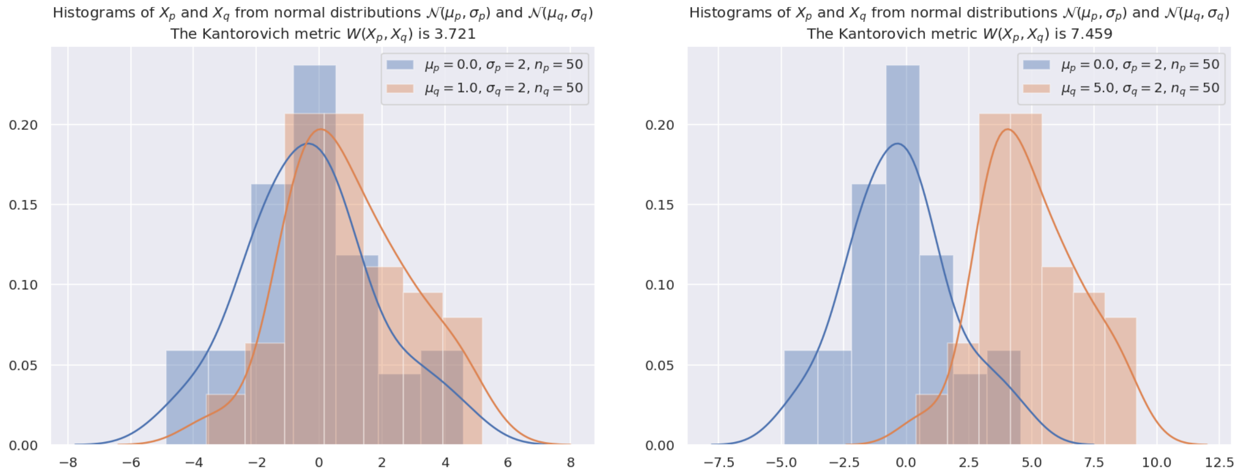


Figure II.1: Visualizing the influence of the difference between the means of two normal distributions with histograms

Now, we are going to plot the Kantorovich metric itself. We choose $n_p = n_q = 30$, which is a sufficient number of samples according to the results of our simulations. Thus, we have $N = 60$. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters: $\mu_p = -5$ and $\sigma_p = 2$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 2$, while μ_q varies from 0 to 20. See figure II.2.

Observing figure II.2, can we say that the dependency of $W(X_p, X_q)$ to $\mu_q - \mu_p$ is linear? Yes, from an empirical point of view. Indeed, the linear regression score R^2 is 0.997.

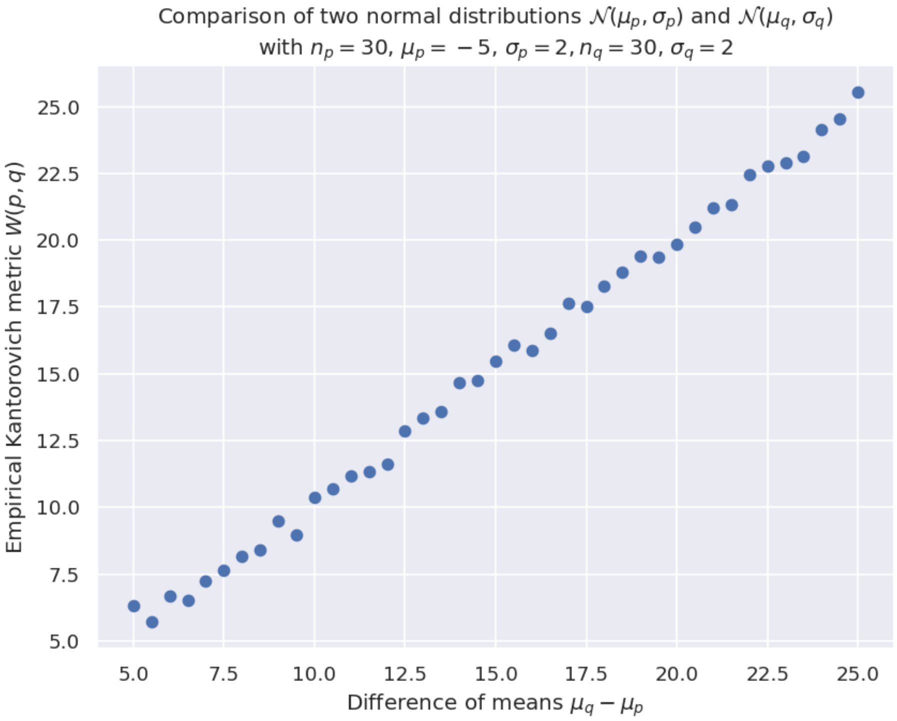


Figure II.2: Evolution of the Kantorovich metric with the difference between the means of two normal distributions

Now we do the same as previously, but by changing the fixed value of μ_p and the fixed values of the $\sigma_p = \sigma_q$. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters: $\mu_p = 3$ and $\sigma_p = 4$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 4$, while μ_q varies from 0 to 20. See figure II.3. It seems to be symmetric and there are two linear regressions (one decreasing on the left and the other one increasing on the right).

II.4.1.2 Influence of the difference between the standard deviations

Now, we do the same as previously, but by modifying the standard deviations (and keeping the means fixed). See figure II.4. When the difference between the standard deviations increases, the Kantorovich metric W increases.

Observing figure II.4, can we say that the dependency of $W(X_p, X_q)$ to $\sigma_q - \sigma_p$ is linear? Not really. Indeed, the linear regression score R^2 is 0.856. Note that only the part at the beginning seems to cause problem for a linear model, which explains why the regression score is not closer to 1.

II.4.1.3 Influence of the number of samples

Now, we do the same as subsubsection II.4.1.1 with the difference between means, but three times, each time changing the number of samples: $n_p = n_q = 10$ then $n_p = n_q = 30$ then $n_p = n_q = 50$. See figure II.5 for $n_p = n_q = 10$ and $n_p = n_q = 50$. See the previous figure II.2 for $n_p = n_q = 30$.

The number of samples does not seem to have an influence, only that the linear regression slightly gets closer to 1 when the number of samples increases. Indeed, for $n_p = n_q = 10$ we have $R^2 = 0.991$, for $n_p = n_q = 30$ we have $R^2 = 0.997$ and for $n_p = n_q = 50$ we have $R^2 = 0.998$. Hence, in this report,

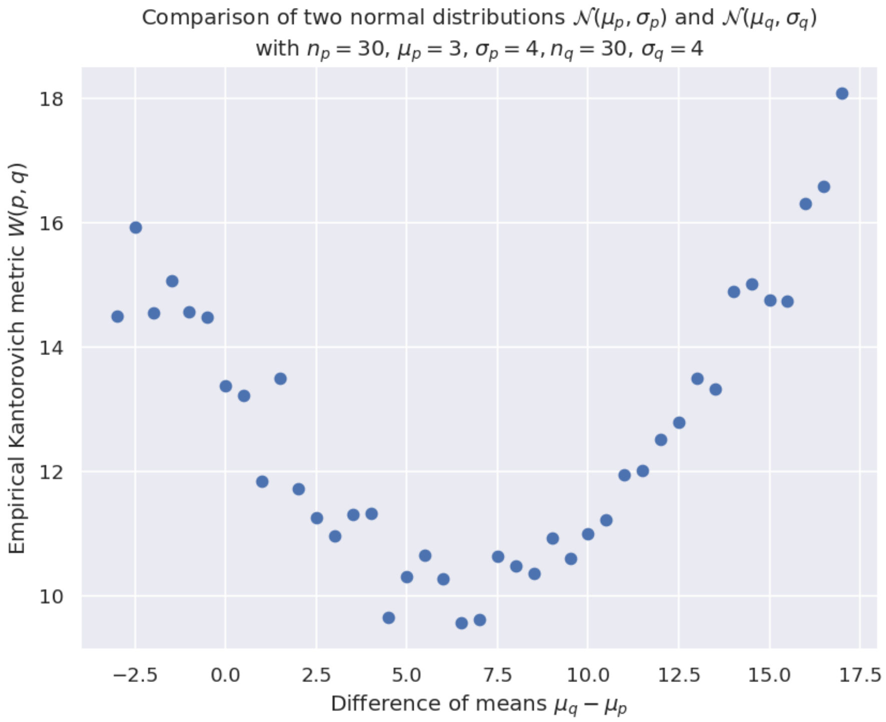


Figure II.3: Evolution of the Kantorovich metric with the difference between the means of two normal distributions

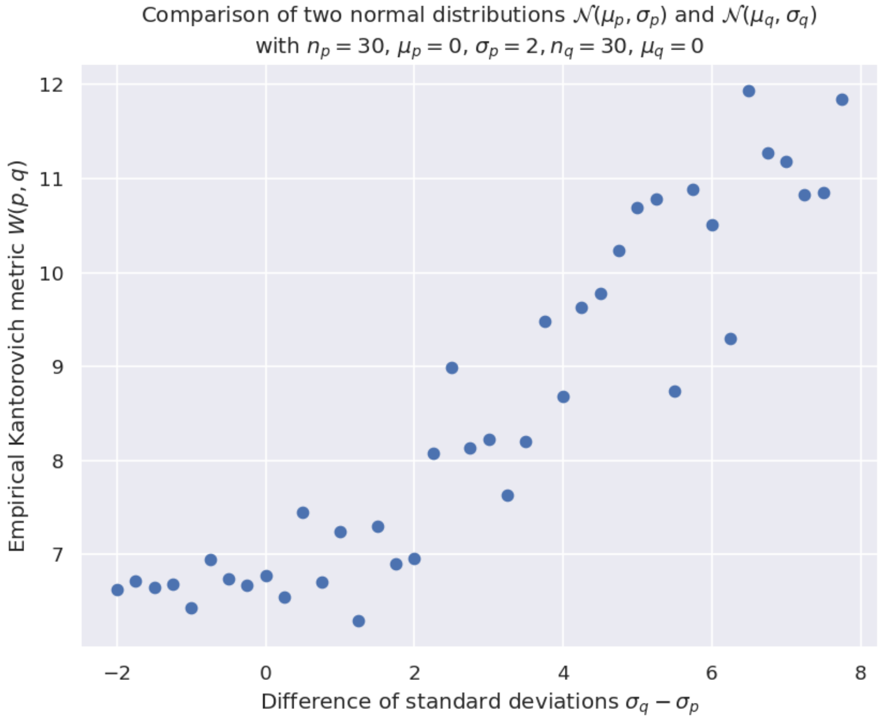


Figure II.4: Evolution of the Kantorovich metric with the difference between the standard deviations of two normal distributions

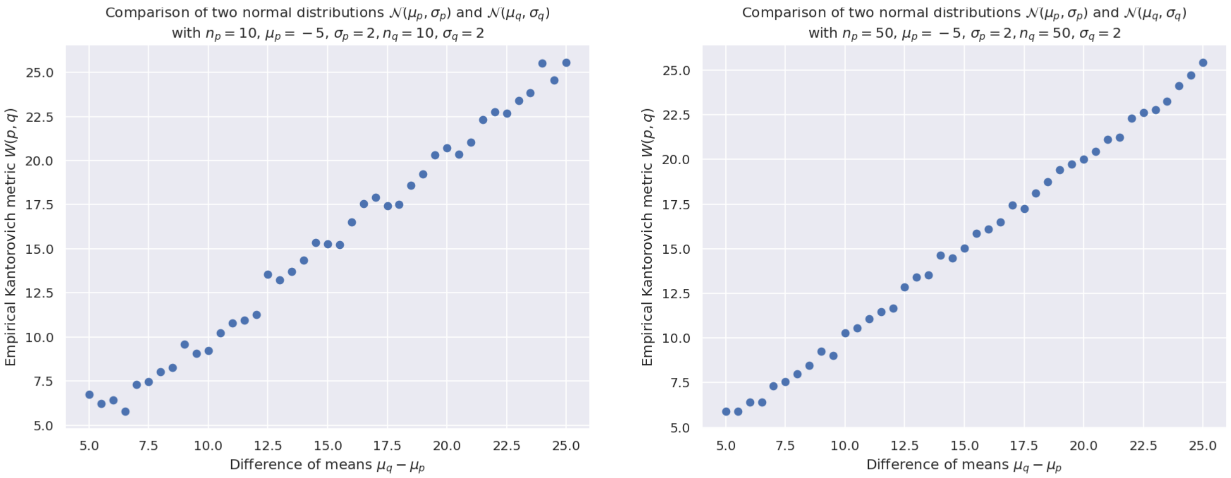


Figure II.5: Evolution of the Kantorovich metric with the difference between the means of two normal distributions, for several numbers of samples

we will not consider the number of samples as a relevant parameter.

II.4.2 Comparison of two exponential distributions

We choose $n_p = n_q = 30$, which is a sufficient number of samples according to the results of our simulations. Thus, we have $N = 60$. For $\mathbb{P} = \mathscr{E}(\lambda_p)$, we fix $\lambda_p = 1$. For $\mathbb{Q} = \mathscr{E}(\lambda_q)$, λ_q varies from 1 to 20. See figure II.6. When the difference between the parameters increases, the Kantorovich metric W increases.

Observing figure II.6, can we say that the dependency of $W(X_p, X_q)$ to $\lambda_q - \lambda_p$ is linear? Not really. Indeed, the linear regression score R^2 is 0.865.

II.4.3 Comparison of two uniform distributions

We choose $n_p = n_q = 30$, which is a sufficient number of samples according to the results of our simulations. Thus, we have $N = 60$. We consider $\mathbb{P} = \mathbb{U}([a, a+h])$ and $\mathbb{Q} = \mathbb{U}([r, r+h])$ where h is the length of the intervals. We fix $h = 2$. For $\mathbb{P}$, we fix $a = 0$. For $\mathbb{Q}$, r varies from 1 to 20. See figure II.7. When the difference between the parameters increases, the Kantorovich metric W increases.

Observing figure II.7, can we say that the dependency of $W(X_p, X_q)$ to $r - a$ is linear? Yes, from an empirical point of view. Indeed, the linear regression score R^2 is 1. The linear dependency obtained numerically is coherent with the explicit formula (II.10).

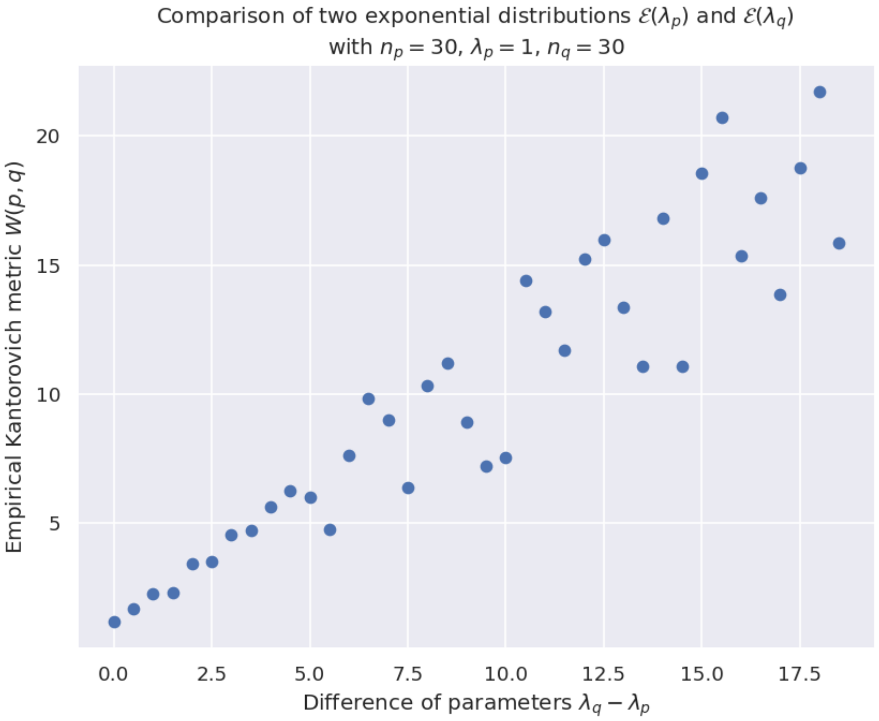


Figure II.6: Evolution of the Kantorovich metric with the difference between the parameters of two exponential distributions

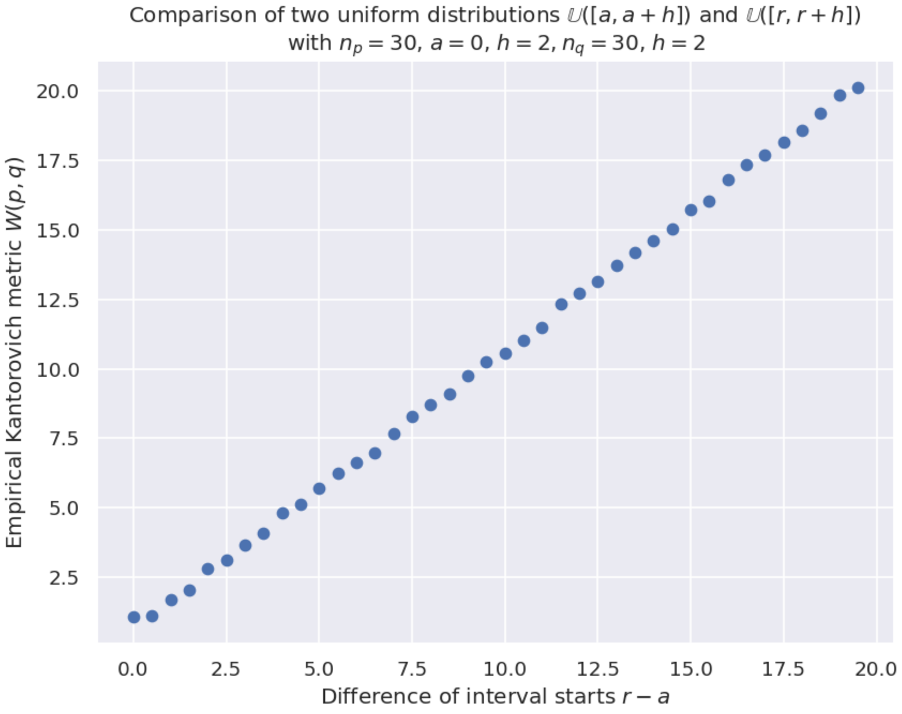


Figure II.7: Evolution of the Kantorovich metric with the difference between the parameters of two uniform distributions

II.5 Conclusion on IPMs (Kantorovich metric)

In this chapter, we studied integral probability metrics $\gamma_{\mathscr{F}}$. IPMs are a popular estimation of distance on two probability distributions, that input the samples drawn from these distributions.

The choice of $\mathscr{F}$ is the crucial distinction between different IPMs: each choice of $\mathscr{F}$ leads to a specific IPM. We dealt with the Kantorovich metric W and the Dudley metric β, which are the most popular.

We obtained an explicit formula for the Kantorovich metric W of two uniform distributions and we realized that W intuitively corresponds to the common notion of distance: when the difference between the parameters (of the distributions) increases, W increases.

Hence, we carried out several experiments in order to observe how IPMs evolve with the parameters of the distributions. For the empirical estimation, we focused on the Kantorovich metric W of normal, exponential and uniform distributions. Note that the number of samples does not have a relevant influence on the value of W and that we took $N = 60$ total samples in our simulations.

For the numerical computation of W, we can not have too many samples (less than 1 000 total samples) because of the memory issue met when solving the linear programming problem (we used the PuLP library). Let us not forget that we must choose the space metric ρ. In this report we always chose $\rho(x, y) = |x - y|$ for $(x, y) \in \mathbb{R}^2$ as the samples were always one-dimensional.

The empirical results are grouped in table II.1. When the difference between the parameters increases, the distance increases.

distribution $\mathbb{P}$	distribution $\mathbb{Q}$	difference between the parameters $\varDelta$	R^2 of the linear regression of $W \propto \varDelta$
$\mathscr{N}(\mu_p, \sigma_p)$	$\mathscr{N}(\mu_q, \sigma_q)$	$\mu_q - \mu_p$	0.997
$\mathscr{N}(\mu_p, \sigma_p)$	$\mathscr{N}(\mu_q, \sigma_q)$	$\sigma_q - \sigma_p$	0.856
$\mathscr{E}(\lambda_p)$	$\mathscr{E}(\lambda_q)$	$\lambda_q - \lambda_p$	0.865
$\mathbb{U}([a, a+h])$	$\mathbb{U}([r, r+h])$	$r - a$	1

Table II.1: Synthesis of the simulations on the Kantorovich metric W of two distributions $\mathbb{P}$ and $\mathbb{Q}$: evolution of W according to the difference between the parameters of $\mathbb{P}$ and $\mathbb{Q}$

Chapter III

f-divergences

In probability theory, a f-divergence is a function $D_f(\mathbb{P}\|\mathbb{Q})$ that measures the difference between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$.

III.1 General definition of f-divergences

According to [Basseville, 1988], the general notion of f-divergence has been introduced by Csiszar and indepentently by Ali & Silvey. It is based upon the fact that it is intuitively "natural" to measure the remoteness of two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ with the aid of the "dispersion" – with respect to $\mathbb{P}$ – of the likelihood ratio $\Phi(x) = \frac{\mathbb{Q}(x)}{\mathbb{P}(x)}$: if $\mathbb{P}$ and $\mathbb{Q}$ are two densities on $\mathbb{R}$, when they "move" away from each other, Φ increases on a set of decreasing $\mathbb{P}$-probability and decreases on a set of increasing $\mathbb{P}$-probability.

According to [Basseville, 1988], we have the continuous version:

Definition 7 (f-divergence D_f (continuous version))**.**

Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions. Let f be a convex real function on $\mathbb{R}_+$. Then, the **f-divergence of $\mathbb{P}$ from $\mathbb{Q}$** is defined as:

$$D_f(\mathbb{P},\mathbb{Q}) \coloneqq \mathbb{E}_{\mathbb{P}}\left[f\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)\right] \tag{III.1}$$

According to [Csiszár and Shields, 2004], we have the discrete version:

Definition 8 (f-divergence D_f (discrete version))**.**

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a space S. Let f be a continuous convex real function on $\mathbb{R}_+$, with $f(1) = 0$. Then, the **f-divergence of $\mathbb{P}$ from $\mathbb{Q}$** is defined as:

$$D_f(\mathbb{P},\mathbb{Q}) \coloneqq \sum_{x\in S} \mathbb{Q}(x)\; f\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\right) \tag{III.2}$$

Many common divergences, such as Kullback-Leibler divergence, Hellinger distance, and total variation distance, are special cases of f-divergence, coinciding with a particular choice of f in (III.2). In the following subsections, we are going to define the three previous divergences (in the discrete case), using [Jordan, xxxx]. We choose only these three because they are the most popular.

III.1.1 Kullback-Leibler divergence

The definition is taken from [Jordan, xxxx] and the Deep Learning textbook pages 71 and 72 [Goodfellow et al., 2016].

For the Kullback-Leibler divergence (III.3), we choose $f(u) = u\log(u)$ in (III.2).

Definition 9 (Kullback-Leibler divergence D_{KL})**.**

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a space S. The **Kullback-Leibler divergence** (or KL divergence) of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q}) \coloneqq \sum_{x\in S} \mathbb{P}(x)\log\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\right) \tag{III.3}$$

if $\forall x \in S, \mathbb{P}(x)\mathbb{Q}(x) > 0$.

In the case of discrete variables, the KL divergence $D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q})$ is the extra amount of information (measured in bits if we use the base-2 logarithm, but in machine learning we usually use nats and the natural logarithm) needed to send a message containing symbols drawn from probability distribution $\mathbb{P}$, when we use a code that was designed to minimize the length of messages drawn from probability distribution $\mathbb{Q}$.

The KL divergence has many useful properties, most notably being non-negative. The KL divergence is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution in the case of discrete variables, or equal “almost everywhere” in the case of continuous variables. Because the KL divergence is non-negative and measures the difference between two distributions, it is often conceptualized as measuring some sort of distance between these distributions. It is not a true distance because it is **not symmetric**: $D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q}) \neq D_{\mathrm{KL}}(\mathbb{Q},\mathbb{P})$ for some $\mathbb{P}$ and $\mathbb{Q}$. This asymmetry means that there are important consequences to the choice of whether to use $D_{\mathrm{KL}}(\mathbb{P},\mathbb{Q})$ or $D_{\mathrm{KL}}(\mathbb{Q},\mathbb{P})$.

As can be seen in formula (III.3), D_{KL} is a sum of positive terms. Thus, the more terms we have, the higher the KL divergence is. Hence, for the numerical computation, D_{KL} will increase when our empirical distributions $\mathbb{P}$ and $\mathbb{Q}$ have more samples. See figure III.1. This is problematic because we want to measure the "distance" between two distributions and the KL divergence obtained should not depend on the number of samples, or we could not compare KL divergences with each other. For example, if we have $\mathbb{P} = \mathbb{Q}$, thus we should have an empirical KL divergence close to 0, but if we have a lot of samples, the estimated KL divergence would be high.

We should normalize formula (III.3) by dividing the KL divergence by the number of samples of a distribution $n_p (= n_q)$. See figure III.2. In the rest of this report, we will only use the normalized KL divergence. We refer to the **normalized Kullback-Leibler divergence** as D_{nKL}.

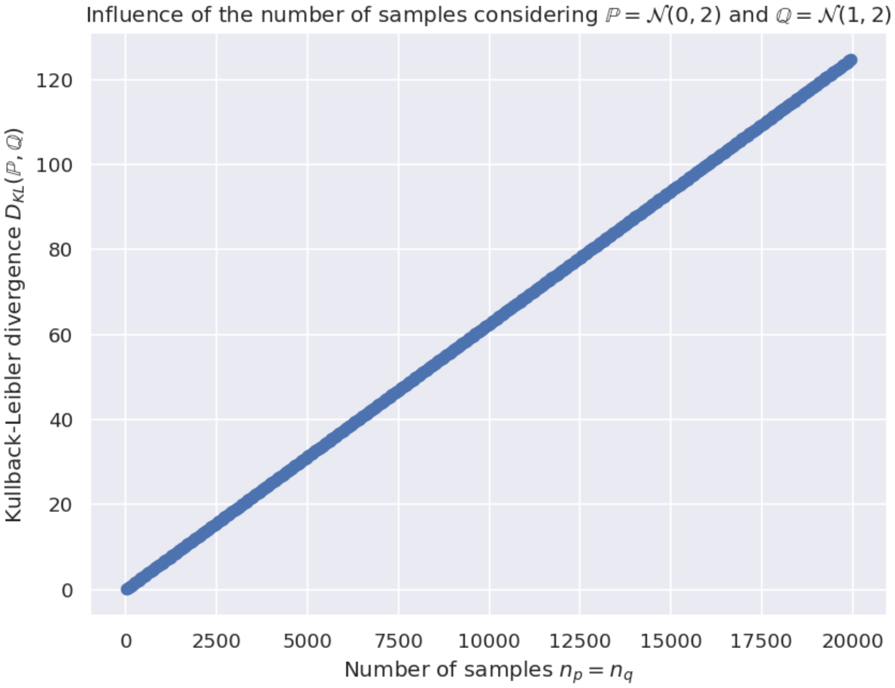


Figure III.1: When our empirical distributions have more samples, the KL divergences increases.

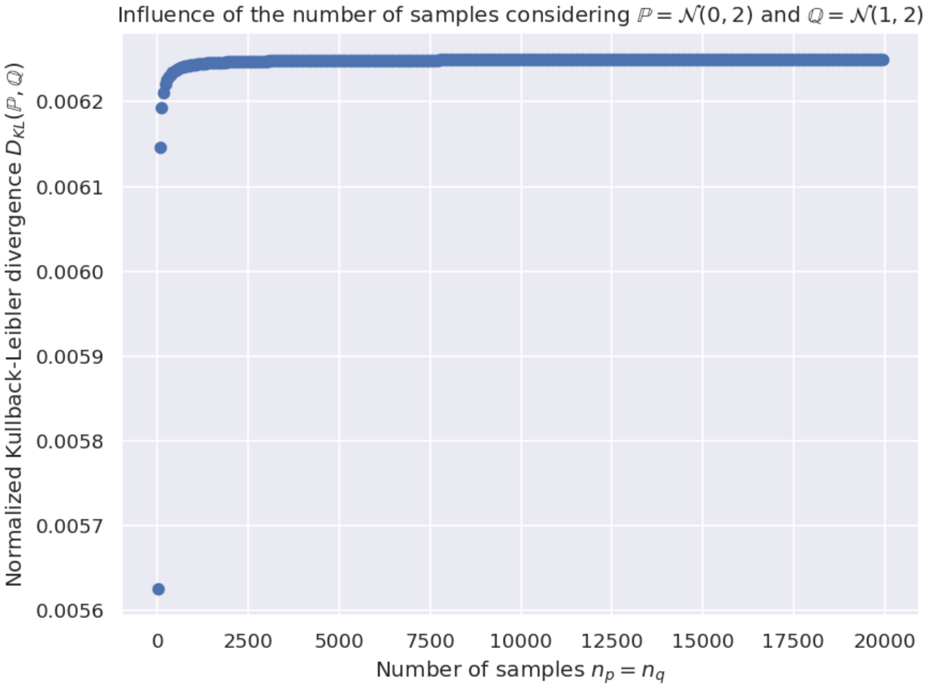


Figure III.2: When our empirical distributions have more samples, the normalized KL divergences stay constant.

III.1.2 Hellinger distance

For the Hellinger distance (III.4), we choose $f(u) = \left(\sqrt{u} - 1\right)^2$ in (III.2).

Definition 10 (Hellinger distance)**.**

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a space S. The **Hellinger distance** of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \coloneqq \sum_{x \in S} \left(\sqrt{\mathbb{P}(x)} - \sqrt{\mathbb{Q}(x)}\right)^2 \tag{III.4}$$

The Hellinger distance is non-negative, is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution (in the case of discrete variables) and symmetric. As its name suggests, contrary to the KL divergence, the Hellinger distance is a true distance.

As can be seen in formula (III.4), D_{H} is a sum of positive terms. Similar to the KL divergence, we should normalize formula (III.4) by dividing the Hellinger distance by the number of samples of a distribution $n_p (= n_q)$. In the rest of this report, we will only use the normalized Hellinger distance. We refer to the **normalized Hellinger distance** as D_{nH}.

III.1.3 Variational distance

For the variational distance (III.5), we choose $f(u) = |u - 1|$ in (III.2).

Definition 11 (Variational distance)**.**

Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability distributions over a space S. The **variational distance** of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:

$$D_{\mathrm{V}}(\mathbb{P}, \mathbb{Q}) \coloneqq \sum_{x \in S} |\mathbb{P}(x) - \mathbb{Q}(x)| \tag{III.5}$$

The variational distance is non-negative, is 0 if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution (in the case of discrete variables) and symmetric. As its name suggests, contrary to the KL divergence, the variational distance is a true distance.

As can be seen in formula (III.5), D_{V} is a sum of positive terms. Similar to the KL divergence, we should normalize formula (III.5) by dividing the Variational distance by the number of samples of a distribution $n_p (= n_q)$. In the rest of this report, we will only use the normalized variational distance. We refer to the **normalized variational distance** as D_{nV}.

III.2 How does the Kullback-Leibler divergence of two probability distributions evolve with their parameters?

In this section and the following ones III.3 and III.4, we are going to interpret the empirical f-divergences by running several simulations using Python. The samples will always be one-dimensional.

We are going to compare two (empirical) normal distributions by modifying their parameters (means and standard deviations) and interpret how their f-divergence evolves. We are going to do the same with two exponential distributions and two uniform distributions.

Let us recall that, contrary to IPMs, f-divergences input the empirical probability distributions, and not the empirical samples. Moreover, contrary to the Kantorovich metric, there is no memory issue due to the number of samples.

In this chapter, to fit a model, we use the the linear regression function from `scikit-learn`. The `score` method returns the coefficient of determination R^2 of the prediction.

Here, we start with the normalized Kullback-Leibler divergence D_{nKL}. We recall that the normalized KL divergence does not depend upon the number of samples: it is parameter that we will not study.

III.2.1 Comparison of two normal distributions

In this subsection, we consider two normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$.

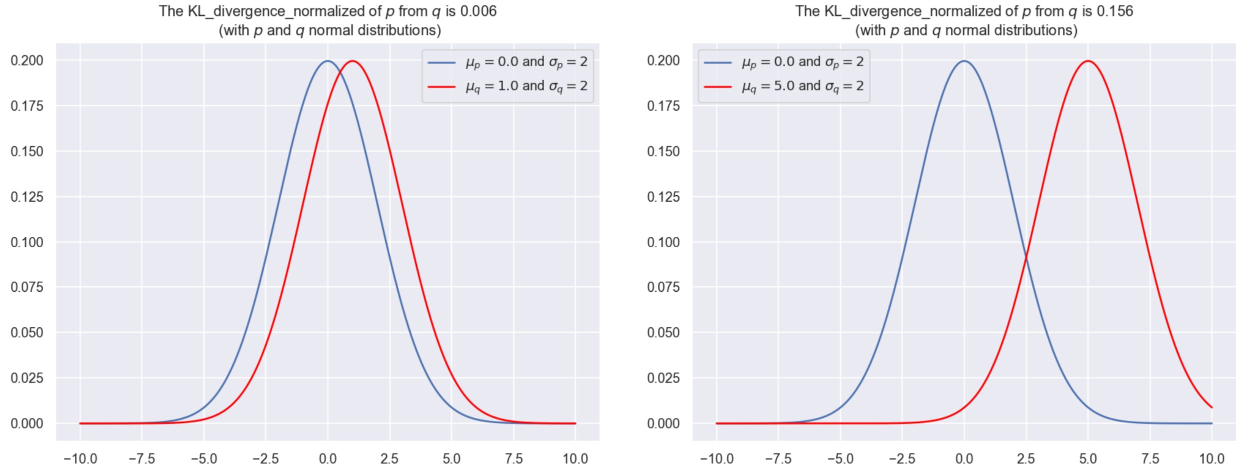
III.2.1.1 Influence of the difference between means

First of all, let us plot the probability distributions to visualize how the normalized KL divergence evolves with the difference $\mu_q - \mu_p$. See figure III.3. All the parameters are the same, only the means are different. On the graphic on the left, we have $\mu_q - \mu_p = 1$ and $D_{\text{nKL}} = 0.06$, while on the right we have $\mu_q - \mu_p = 5$ and $D_{\text{nKL}} = 0.156$. Hence, when the difference between the means increases (thus the distribution are "more different"), D_{nKL} increases which justifies why D_{nKL} is a "distance".

Note that we can visualize that D_{nKL} is asymmetric on figure III.4. Indeed, we switched the values of the parameters of $\mathbb{P}$ and $\mathbb{Q}$, but the values of the normalized KL divergence are different.

Now, we are going to plot the normalized KL divergence itself. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters with $\mu_p = 0$ and $\sigma_p = 2$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 2$, while μ_q varies from 0 to 20. See figure III.5. Observing figure III.5, can we say that the dependency of $D_{\text{nKL}}(\mathbb{P}, \mathbb{Q})$ to $(\mu_q - \mu_p)^2$ is linear? Yes, from an empirical point of view. Indeed, the linear regression score R^2 is 1.

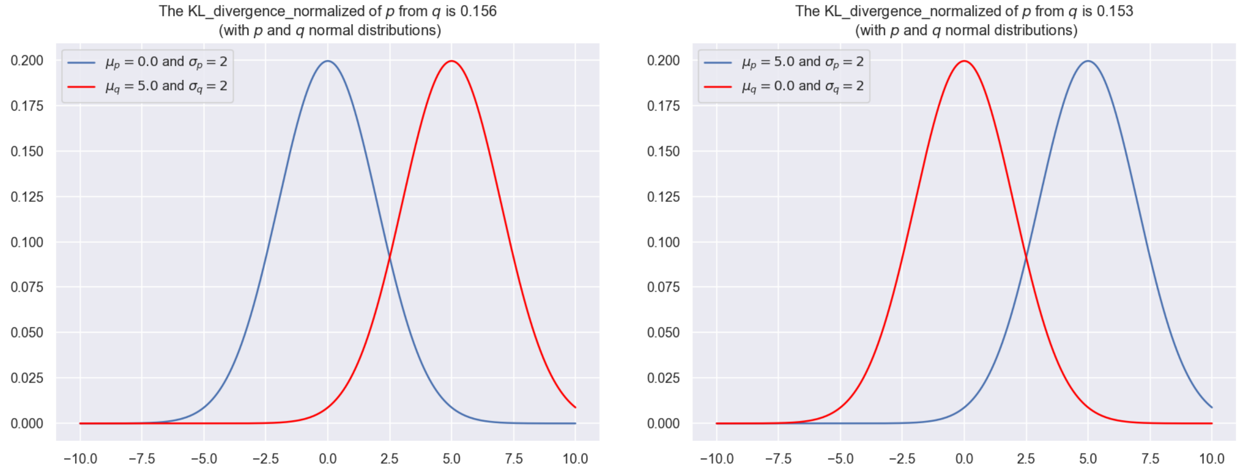
Now we do the same as previously, but by changing the fixed value of μ_p and the fixed values of the $\sigma_p = \sigma_q$. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters: $\mu_p = 3$ and $\sigma_p = 4$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 4$, while μ_q varies from 0 to 20. See figure III.6. Observing figure III.6, can we say that the dependency of $D_{\text{nKL}}(\mathbb{P}, \mathbb{Q})$ to $(\mu_q - \mu_p)^2$ is linear? Yes, from an empirical point of view. Indeed, the linear regression score R^2 is once again 1. We obtain the same result as previously.



The KL_divergence_normalized of p from q is 0.006 (with p and q normal distributions)

The KL_divergence_normalized of p from q is 0.156 (with p and q normal distributions)

Figure III.3: Visualizing the influence of the difference means of two normal distributions plotting the (empirical) distributions. We have $n_p = n_q = 20\,000$.



The KL_divergence_normalized of p from q is 0.156 (with p and q normal distributions)

The KL_divergence_normalized of p from q is 0.153 (with p and q normal distributions)

Figure III.4: Visualizing the asymmetry of the normalized KL divergence plotting the (empirical) distributions. We have $n_p = n_q = 20\,000$ for the two normal distributions.

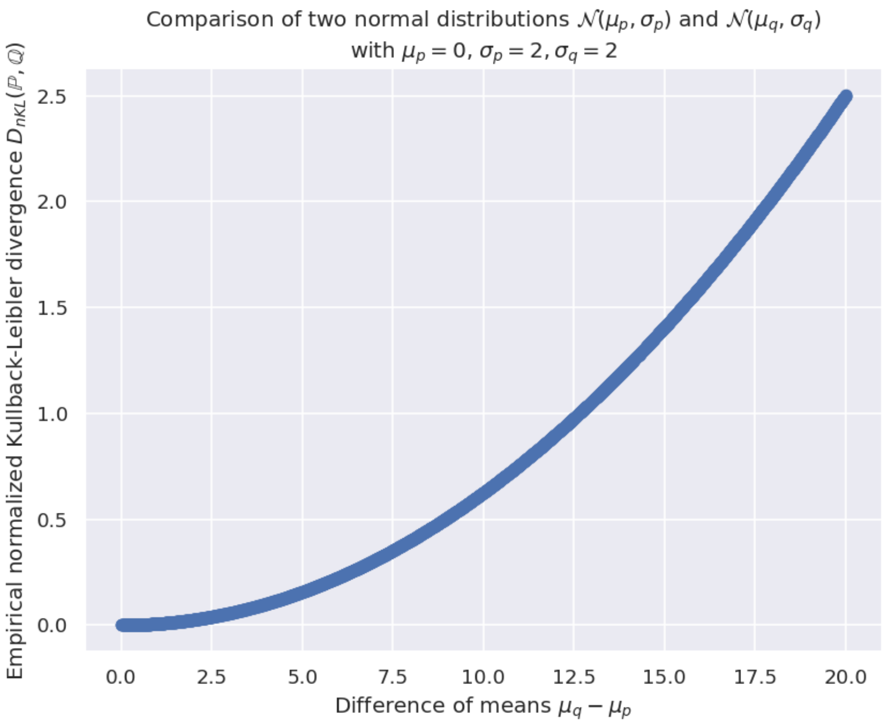


Figure III.5: Evolution of the normalized KL divergence with the difference between means for normal distributions

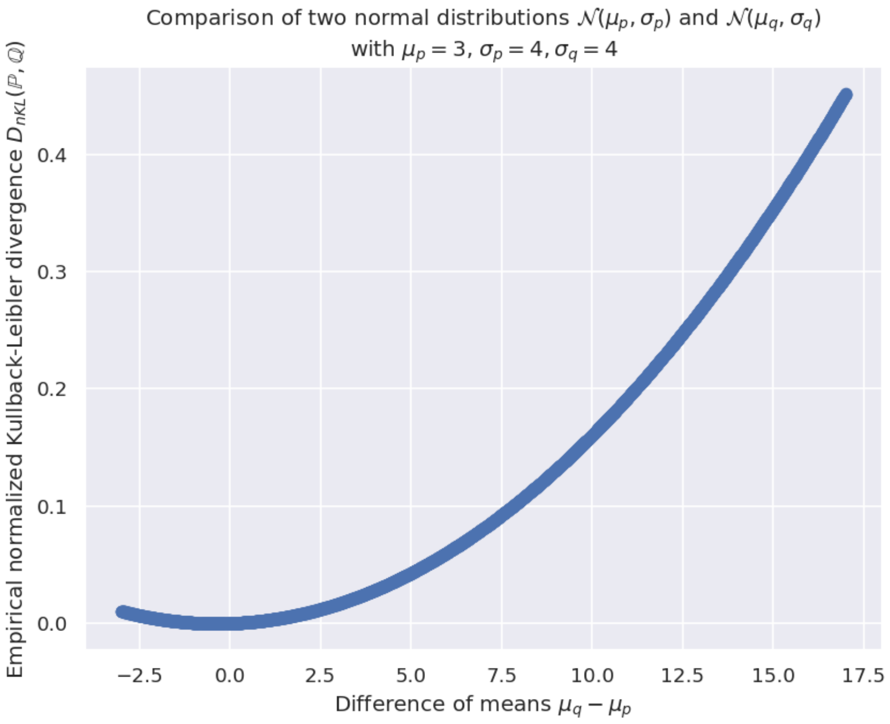


Figure III.6: Evolution of the normalized KL divergence with the difference between means of two normal distributions

III.2.1.2 Influence of the difference between the standard deviations

Now, we do the same as previously, but by modifying the standard deviations (and keeping the means fixed). See figure III.7. When the difference between the standard deviations increases, D_{nKL} increases. Observing figure III.7, can we say that the dependency of $D_{\text{nKL}}(\mathbb{P},\mathbb{Q})$ to $\sqrt{(\sigma_q - \sigma_p)}$ is linear? Yes, from an empirical point of view. Indeed, the linear regression score R^2 is 0.991.

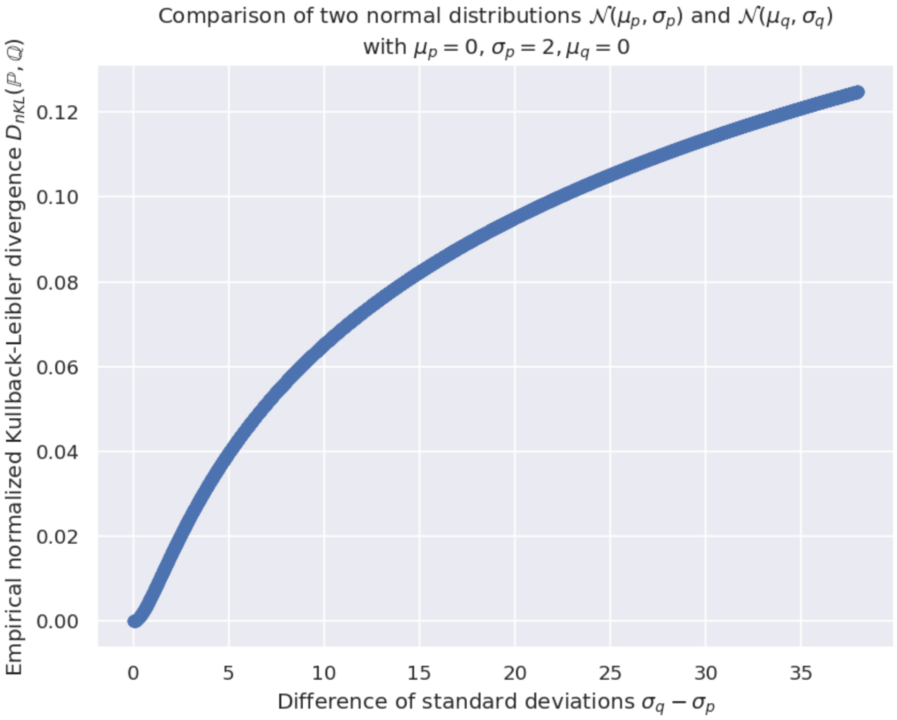


Figure III.7: Evolution of the normalized KL divergence with the difference between standard deviations of two normal distributions

III.2.2 Comparison of two exponential distributions

For $\mathbb{P} = \mathscr{E}(\lambda_p)$, we fix $\lambda_p = 1$. For $\mathbb{Q} = \mathscr{E}(\lambda_q)$, λ_q varies from 2 to 20. See figure III.8. When the difference between the parameters increases, D_{nKL} increases. Note that we received a compilation error because we had to divide by 0 a few times. Observing figure III.8, it is not relevant to try to model D_{nKL} according to $\lambda_q - \lambda_p$.

III.2.3 Comparison of two uniform distributions

We consider $\mathbb{P} = \mathbb{U}([a, a+h])$ and $\mathbb{Q} = \mathbb{U}([r, r+h])$ where h is the length of the intervals. We fix $h = 2$. For $\mathbb{P}$, we fix $a = 0$. For $\mathbb{Q}$, r varies from 0 to 20. See figure III.9. When the difference between the parameters increases, D_{nKL} increases. Note that we received a compilation error because we had to divide by 0 (a uniform distribution often takes value 0), thus $r - a$ only goes to 1.0. Observing figure III.9, can we say that the dependency of $D_{\text{nKL}}(\mathbb{P},\mathbb{Q})$ to $r - a$ is linear? Yes, from an empirical point of view. Note that we were not able to compute the linear regression score R^2, because there are some infinite values.

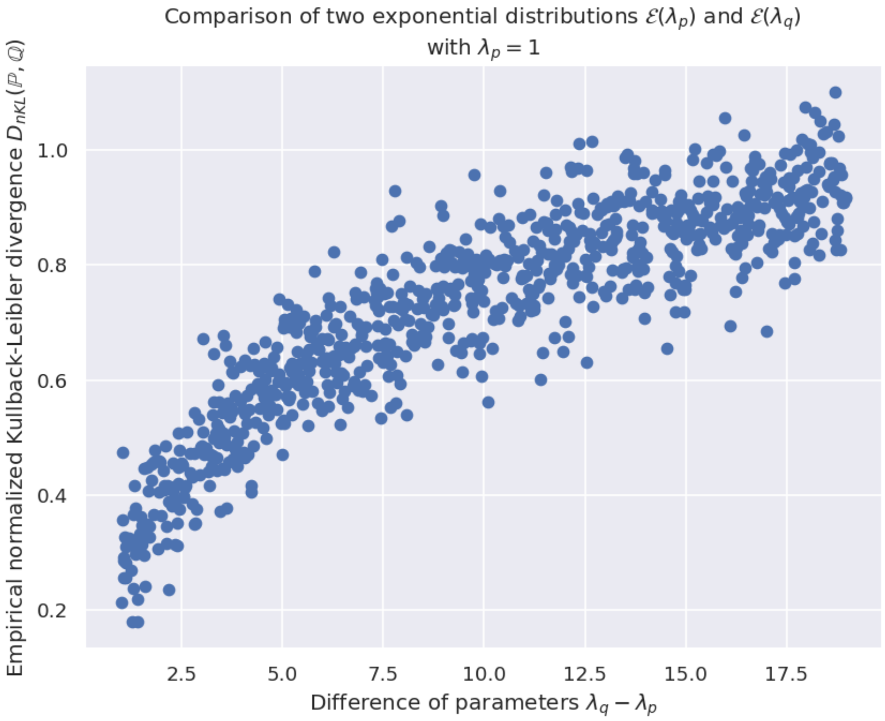


Figure III.8: Evolution of the normalized KL divergence with the difference between the parameters of two exponential distributions

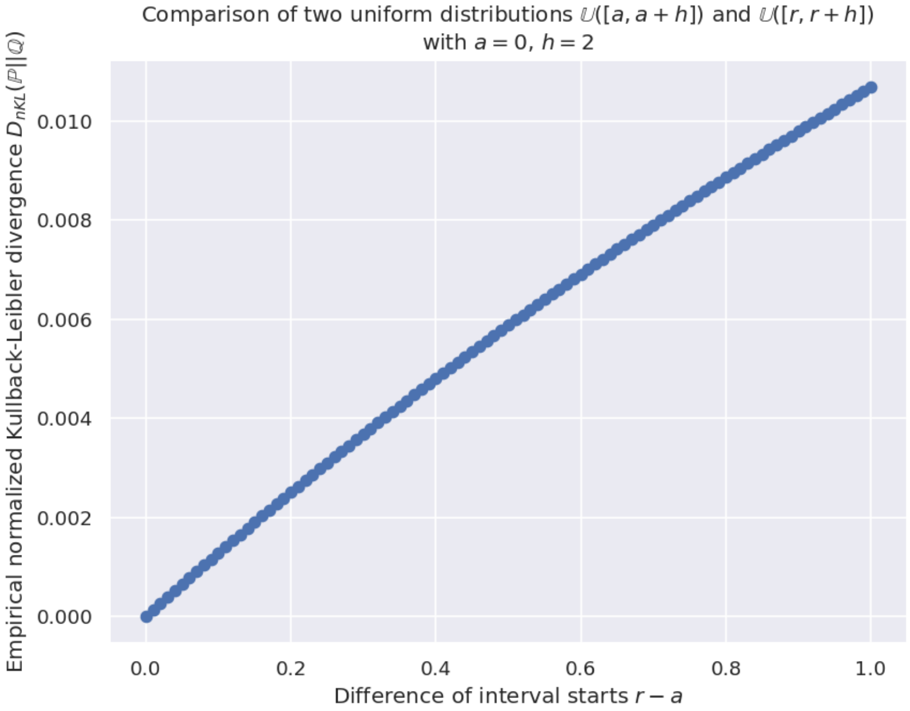


Figure III.9: Evolution of the normalized KL divergence with the difference between the parameters of two uniform distributions. Computation error for $r - a > 1$ because we divide by 0.

III.3 How does the Hellinger distance of two probability distributions evolve with their parameters?

See the first paragraphs of section III.2 for more information about the goal and structure of this current section.

III.3.1 Comparison of two normal distributions

In this subsection, we consider two normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$.

III.3.1.1 Influence of the difference between means

Now, we are going to plot the normalized KL divergence itself. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters with $\mu_p = 0$ and $\sigma_p = 2$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 2$, while μ_q varies from 0 to 20. See figure III.10. Observing figure III.10, it is not relevant to try to model D_{nH} according to $\mu_q - \mu_p$. Note that for $\mu_q - \mu_p > 15$, D_{nH} seems to be constant. What is most surprising is the decreasing part around the middle.

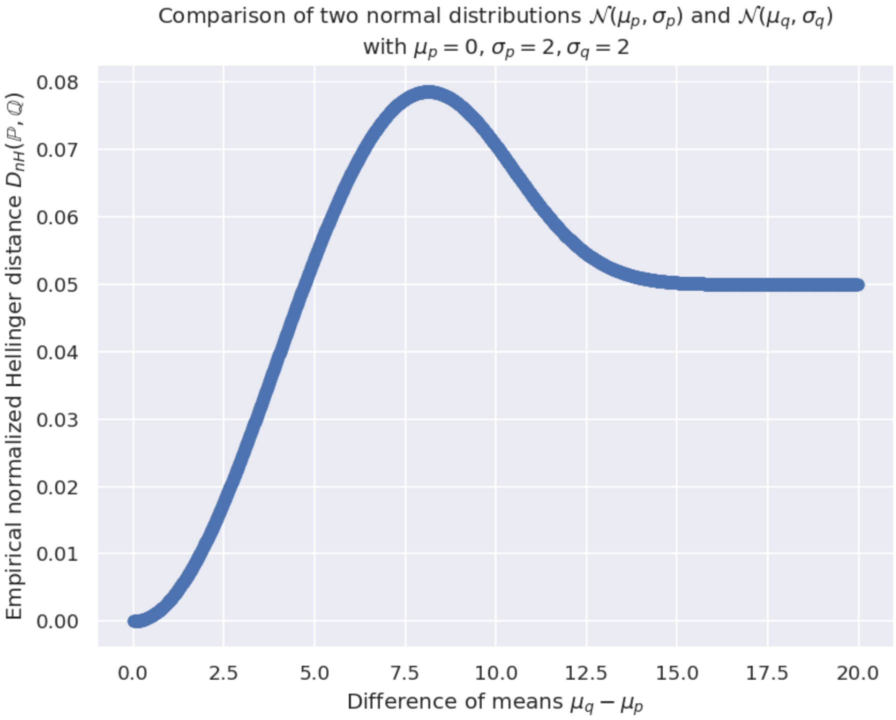


Figure III.10: Evolution of the normalized Hellinger distance with the difference between means for normal distributions

III.3.1.2 Influence of the difference between the standard deviations

Now, we do the same as previously, but by modifying the standard deviations (and keeping the means fixed). See figure III.11. When the difference between the standard deviations increases,

D_{nH} increases. Observing figure III.11, it is not relevant to try to model D_{nH} according to $\sigma_q - \sigma_p$.

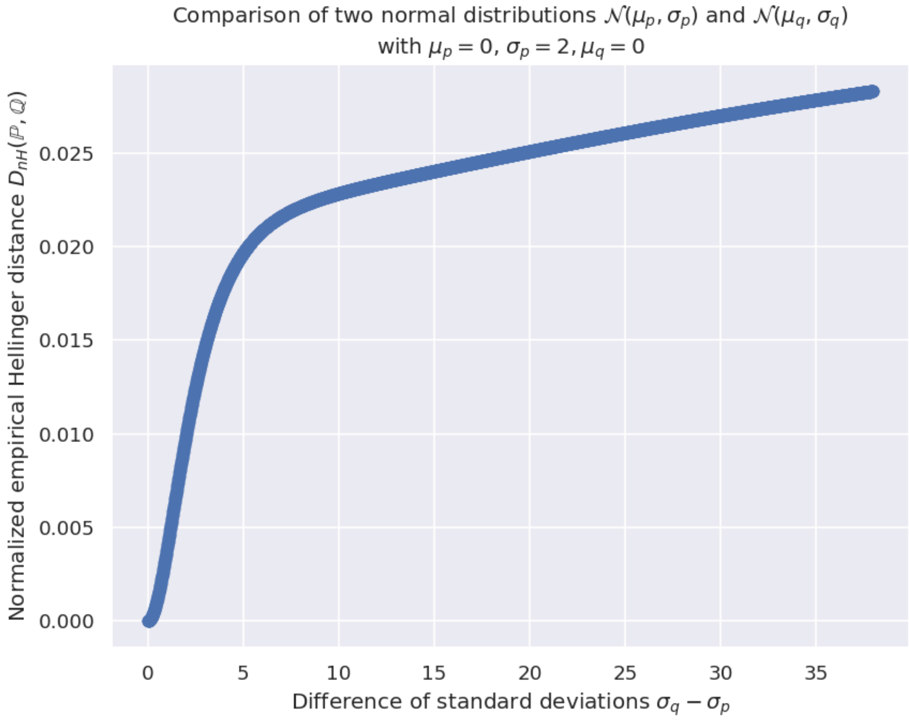


Figure III.11: Evolution of the normalized Hellinger distance with the difference between standard deviations of two normal distributions

III.3.2 Comparison of two exponential distributions

For $\mathbb{P} = \mathcal{E}(\lambda_p)$, we fix $\lambda_p = 1$. For $\mathbb{Q} = \mathcal{E}(\lambda_q)$, λ_q varies from 2 to 20. See figure III.12. When the difference between the parameters increases, D_{nH} increases. Note that we received a compilation error because we had to divide by 0 a few times. Observing figure III.12, it is not relevant to try to model D_{nH} according to $\lambda_q - \lambda_p$.

III.3.3 Comparison of two uniform distributions

We consider $\mathbb{P} = \mathbb{U}([a, a+h])$ and $\mathbb{Q} = \mathbb{U}([r, r+h])$ where h is the length of the intervals. We fix $h = 2$. For $\mathbb{P}$, we fix $a = 0$. For $\mathbb{Q}$, r varies from 0 to 20. See figure III.13. Observing figure III.13, it is not relevant to try to model D_{nH} according to $r - a$. The curve even decreases on the right side.

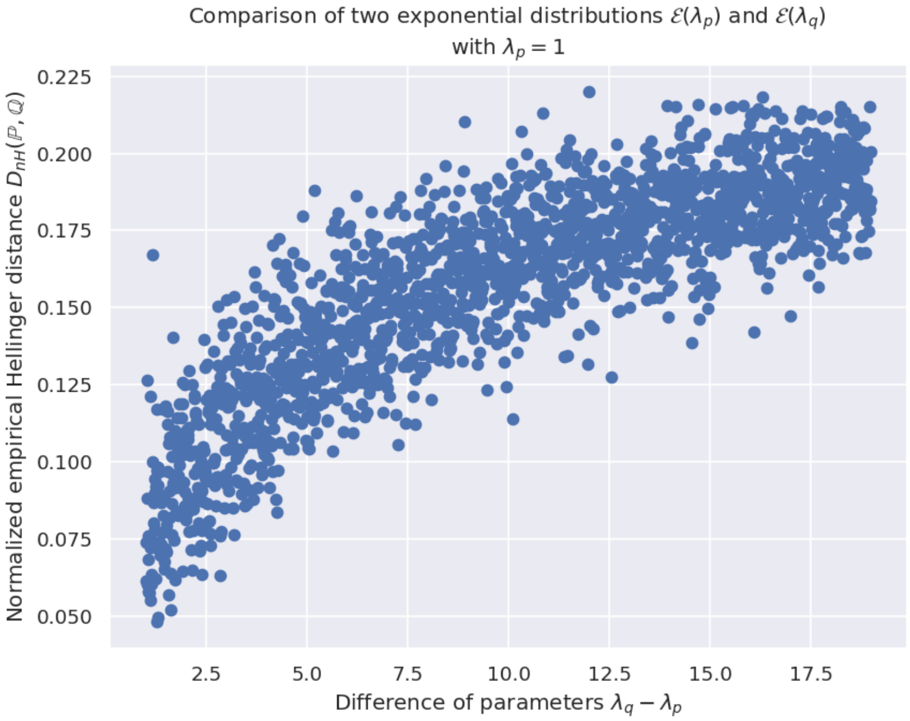


Figure III.12: Evolution of the normalized Hellinger distance with the difference between the parameters of two exponential distributions

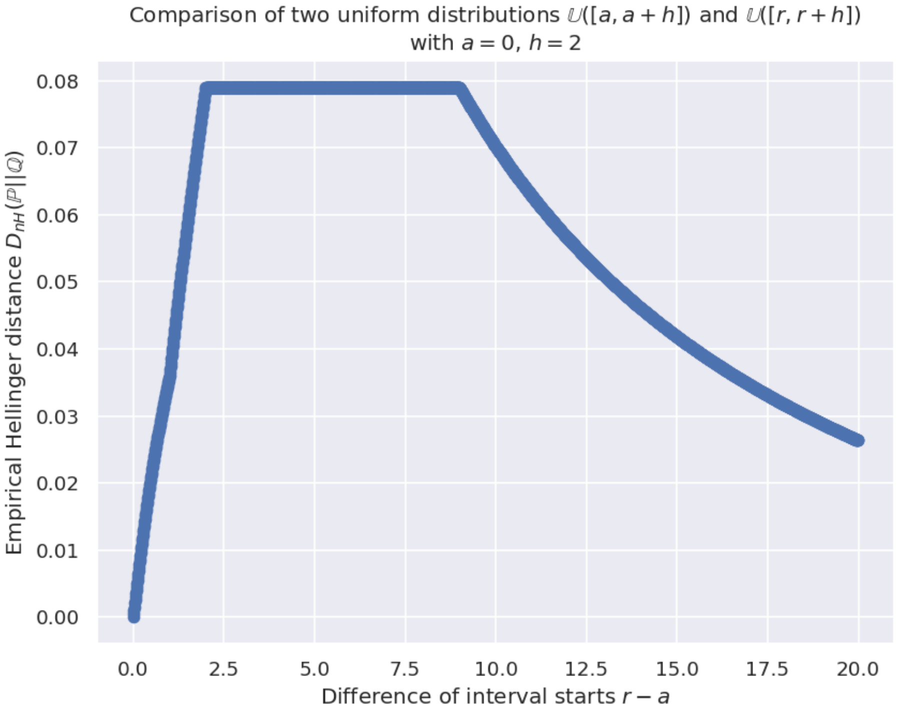


Figure III.13: Evolution of the Hellinger distance with the difference between the parameters of two uniform distributions

III.4 How does the variational distance of two probability distributions evolve with their parameters?

See the first paragraphs of section III.2 for more information about the goal and structure of this current section.

III.4.1 Comparison of two normal distributions

In this subsection, we consider two normal distributions $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$ and $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$.

III.4.1.1 Influence of the difference between means

Now, we are going to plot the normalized variational distance itself. For $\mathbb{P} = \mathcal{N}(\mu_p, \sigma_p)$, we fix all the parameters with $\mu_p = 0$ and $\sigma_p = 2$. For $\mathbb{Q} = \mathcal{N}(\mu_q, \sigma_q)$, we only fix $\sigma_q = 2$, while μ_q varies from 0 to 20. See figure III.14. Observing figure III.14, it is not relevant to try to model D_{nV} according to $\mu_q - \mu_p$. Note that for $\mu_q - \mu_p > 15$, D_{nV} seems to be constant. What is most surprising is the decreasing part around the middle.

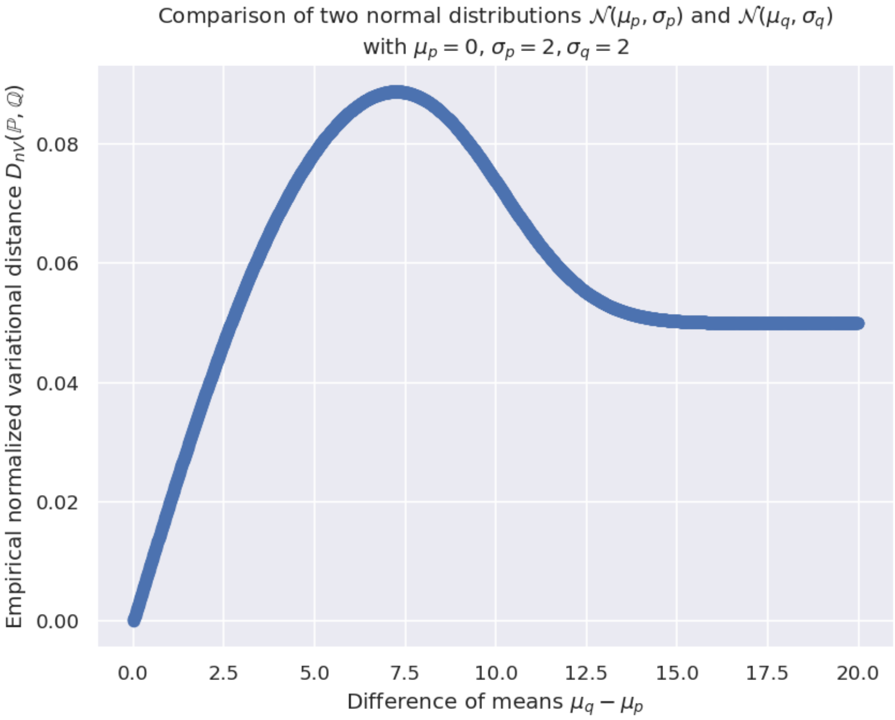


Figure III.14: Evolution of the normalized variational distance with the difference between means for normal distributions

III.4.1.2 Influence of the difference between the standard deviations

Now, we do the same as previously, but by modifying the standard deviations (and keeping the means fixed). See figure III.15. When the difference between the standard deviations increases, D_{nV}

increases (up to a point of saturation). Observing figure III.15, it is not relevant to try to model D_{nV} according to $\sigma_q - \sigma_p$.

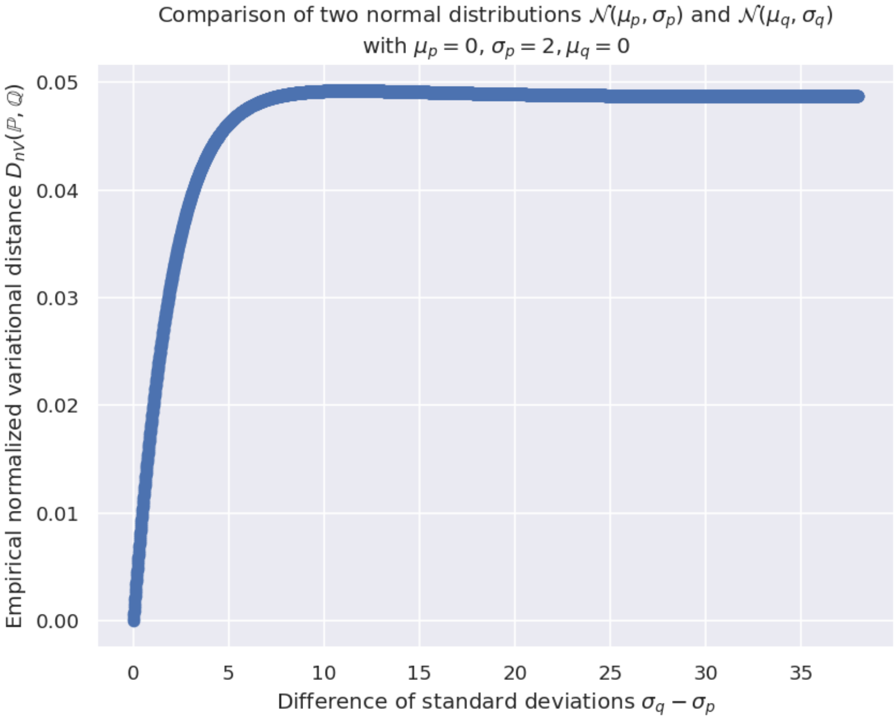


Figure III.15: Evolution of the variational distance with the difference between standard deviations of two normal distributions

III.4.2 Comparison of two exponential distributions

For $\mathbb{P} = \mathscr{E}(\lambda_p)$, we fix $\lambda_p = 1$. For $\mathbb{Q} = \mathscr{E}(\lambda_q)$, λ_q varies from 2 to 20. See figure III.16. When the difference between the parameters increases, D_{nV} increases. Note that we received a compilation error because we had to divide by 0 a few times. Observing figure III.16, it is not relevant to try to model D_{nV} according to $\lambda_q - \lambda_p$.

III.4.3 Comparison of two uniform distributions

We consider $\mathbb{P} = \mathbb{U}([a, a+h])$ and $\mathbb{Q} = \mathbb{U}([r, r+h])$ where h is the length of the intervals. We fix $h = 2$. For $\mathbb{P}$, we fix $a = 0$. For $\mathbb{Q}$, r varies from 0 to 20. See figure III.17. Observing figure III.17, it is not relevant to try to model D_{nV} according to $r - a$. The curve even decreases on the right side.

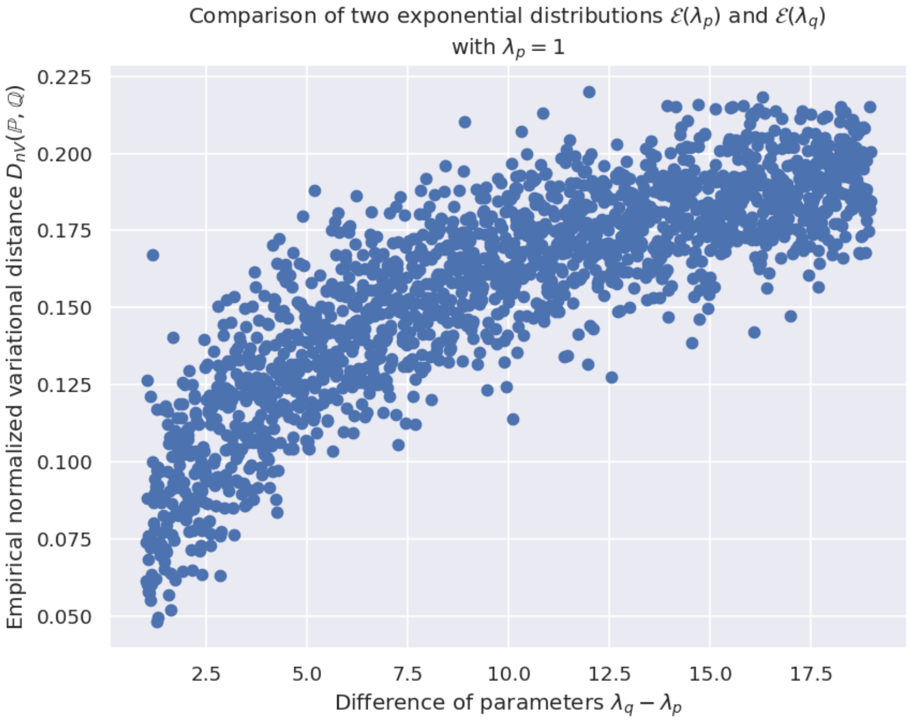


Figure III.16: Evolution of the normalized variational distance with the difference between the parameters of two exponential distributions

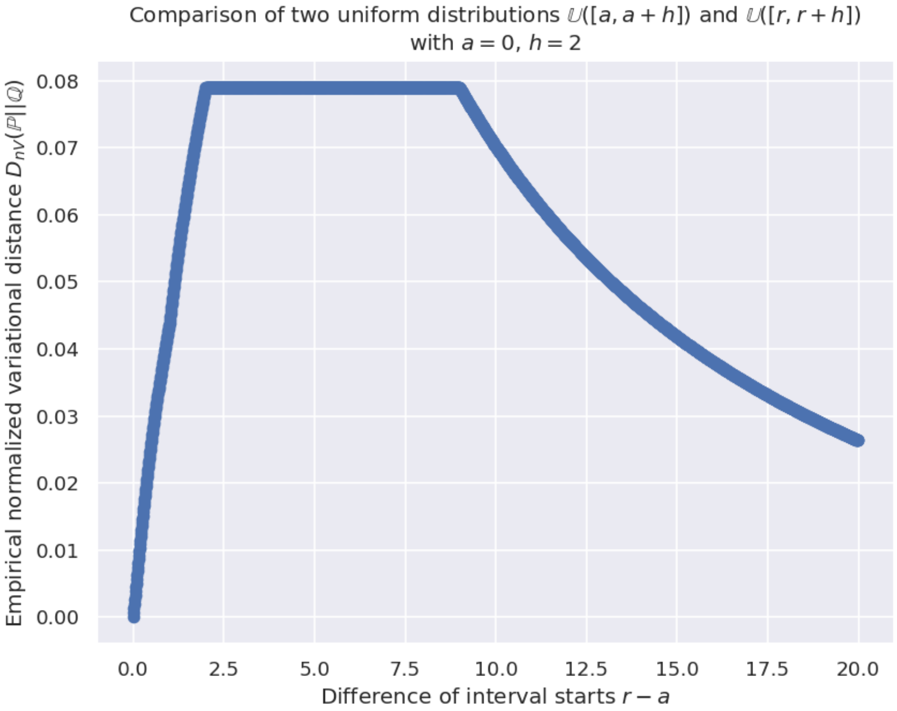


Figure III.17: Evolution of the variational distance with the difference between the parameters of two uniform distributions

III.5 Conclusion on f-divergences

In this chapter, we studied f-divergences D_f. f-divergences are a popular estimation of distance on two probability distributions, that input the empirical probability distributions.

The choice of f is the crucial distinction between different f-divergences: each choice of f leads to a specific f-divergence. We dealt with the Kullback-Leibler divergence D_{KL}, the Hellinger distance D_{H} and the variational distance D_{V}, which are the most popular.

D_{KL}, D_{H} and D_{V} are all-non negative, equal to zero if and only if $\mathbb{P}$ and $\mathbb{Q}$ are the same distribution. The numerical computation of these f-divergences is direct and very easy: contrary to IPMs, we do not need to solve any linear programming problem and we have no memory issue.

Note that D_{KL} is less convenient than D_{H} and D_{V} because it is not symmetric, thus it is not a true distance. Moreover, in the computation of the empirical $D_{\text{KL}}(p,q)$ if q takes the value 0, as q is in the denominator, we get Inf.

As D_{KL}, D_{H} and D_{V} are all sum of positive terms, we should normalize them so that they do not depend on the number of samples of the empirical distributions.

We carried out several experiments in order to observe how f-divergences evolve with the parameters of the distributions. For the empirical estimation, we focused on the f-divergences of normal, exponential and uniform distributions. Note that after normalization, the number of samples does not have a relevant influence. The samples were always one-dimensional.

The empirical results for the normalized Kullback-Leibler divergence D_{nKL} are grouped in table III.1. When the difference between the parameters increases, D_{nKL} increases.

distribution $\mathbb{P}$	distribution $\mathbb{Q}$	difference between the parameters Δ	R^2 of the linear regression of $D_{\text{nKL}} \propto \Delta$
$\mathcal{N}(\mu_p, \sigma_p)$	$\mathcal{N}(\mu_q, \sigma_q)$	$(\mu_q - \mu_p)^2$	1
$\mathcal{N}(\mu_p, \sigma_p)$	$\mathcal{N}(\mu_q, \sigma_q)$	$\sqrt{\sigma_q - \sigma_p}$	0.991
$\mathcal{E}(\lambda_p)$	$\mathcal{E}(\lambda_q)$	$\lambda_q - \lambda_p$	no model
$\mathbb{U}([a, a+h])$	$\mathbb{U}([r, r+h])$	$r - a$	1

Table III.1: Synthesis of the simulations on the normalized Kullback-Leibler divergence D_{nKL} of two distributions $\mathbb{P}$ and $\mathbb{Q}$: evolution of D_{nKL} according to the difference between the parameters of $\mathbb{P}$ and $\mathbb{Q}$.

For the normalized Hellinger distance D_{nH} and the normalized variational distance D_{nV}, there is no trivial model that seems to work. We can not say that, when the difference between the parameters increases, the distance increases. Hence, from the point of view of the modelling, D_{nKL} seems more convenient.

Chapter IV

Application of IPMs and f-divergences to the Choquet integral

In this chapter, we address the end goal of this project: to compare two empirical probability distributions obtained from two different methods for computing the Choquet integral. There are two different methods for computing the Choquet integral that are detailed in paper [Petot et al., 2018]: we will refer to them as the direct method and the method with the new formula introduced in paper [Petot et al., 2018].

Actually, we try to prove (experimentally) that the new formula given in the paper [Petot et al., 2018] is correct by comparing its output samples to the direct method's output samples, as the direct method is proven to be correct. If the "distance" between these samples is "small", we can say that the new formula gives "similar" results than the direct method, thus that the new formula is "correct".

Hence, we are going to use two main different measures for comparing the samples from the two methods: integral probability metrics introduced in chapter II and f-divergences introduced in chapter III. The Choquet integral operator was introduced in chapter I.

We will not focus on how these methods for computing the Choquet integral work: we will only see them as input data for our IPMs and f-divergences.

Let us take the example of comparing the two methods for computing the Choquet integral of normal distributions. The Choquet integral takes as input some normal distributions $\mathcal{N}_{in}$ and outputs a distribution. For the Choquet integral computed with the direct method, this output is noted $\mathbb{P}$. For the Choquet integral computed with the new formula, this output is noted $\mathbb{Q}$. Note that, as the Choquet integral is a non-linear aggregation operator, $\mathbb{P}$ and $\mathbb{Q}$ are (empirical) distributions that have no reason to be normal distributions as their corresponding input $\mathcal{N}_{in}$. Now, we compute the (empirical) Kantorovich metric $W(\mathbb{P},\mathbb{Q})$: if it is "small", we claim that $\mathbb{P}$ is "close" to $\mathbb{Q}$, thus that the new formula is correct. We also compute the f-divergences.

Contrary to the previous ones, this chapter will not deal with uniform distributions, as we have no available data about the Choquet integral of uniform distributions. We will only address normal and exponential distributions.

IV.1 For the Choquet integral of normal distributions

As introduced in the beginning of this chapter, we are going to compare two (empirical) distributions $\mathbb{P}$ and $\mathbb{Q}$, both outputs of the Choquet integral of normal distributions.

IV.1.1 Presenting the data

For computing IPMs, we need the samples X_p and X_q drawn from $\mathbb{P}$ and $\mathbb{Q}$. These samples are given figure IV.1, but in the form of histograms (and their estimated empirical probability distribution).

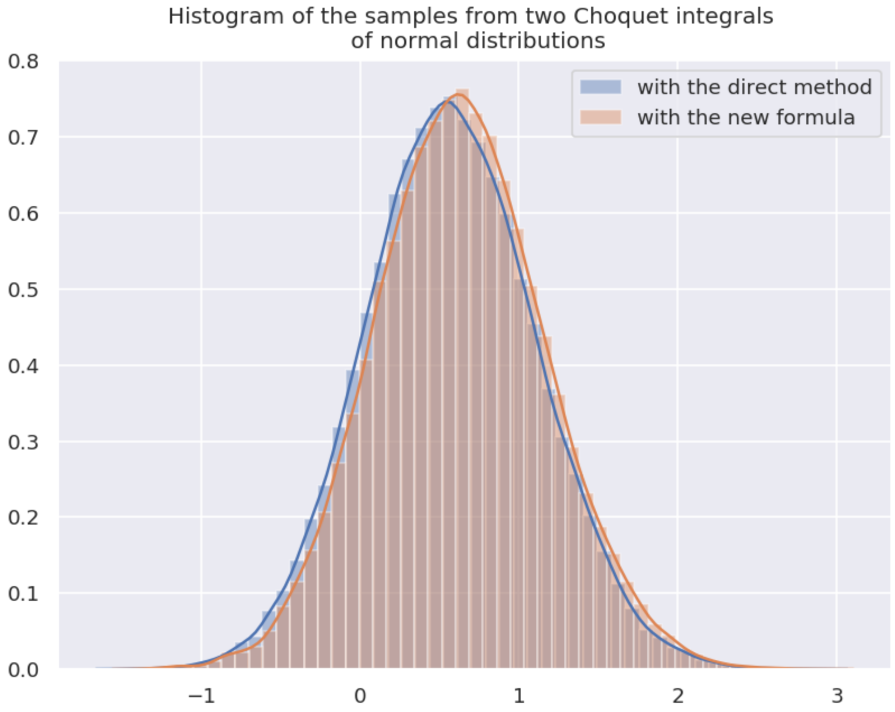


Figure IV.1: Data for IPMs: Histograms of the samples from two methods for computing the Choquet integral of normal distributions

$\mathbb{P}$ and $\mathbb{Q}$ look like normal distributions but are not. Indeed, we performed D'Agostino and Pearson's test using `SciPy`, the null hypothesis being that the samples come from a normal distribution. For X_p, we obtained a pvalue of 10^{-26}. For X_q, we obtained a pvalue of 10^{-28}. Thus we can (strongly) reject the null hypothesis in both cases.

For computing f-divergences, we need the probability distributions, given figure IV.2. Each distribution has 61 samples.

IV.1.2 IPMs: Kantorovich metric

Because of the memory issue, we can only select (randomly) 100 samples per distribution: X'_p and X'_q then compute $W(X'_p, X'_q)$.

Hence, we perform ten random samplings of 100 samples per distribution, compute the Kantorovich metric of each pair of distributions. The mean of the ten Kantorovich metrics is $W = 0.894$ and the

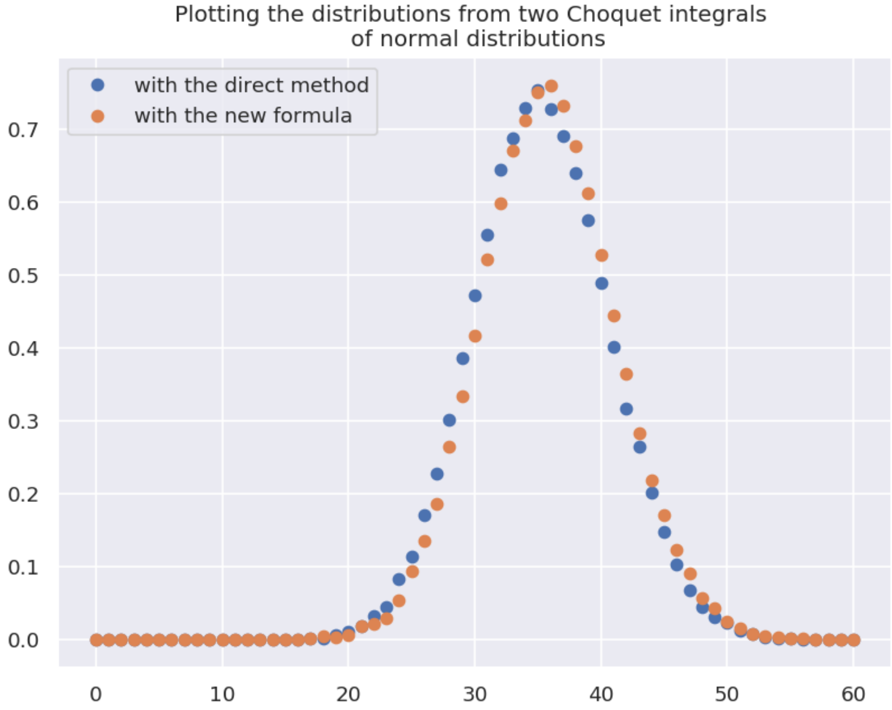


Figure IV.2: Data for f-divergences: empirical distributions from two methods for computing the Choquet integral of normal distributions

standard deviation is 0.033.

By comparing $W = 0.894$ to the values from the simulations of subsection II.4.1, we claim that $W = 0.894$ is a "very small" value, thus that the two distributions are "very close". Hence, from an empirical point of view, the new formula of paper [Petot et al., 2018] seems to be correct.

IV.1.3 f-divergences

The results are displayed in table IV.1. Note that we have a problem with the normalized KL divergence because we must divide by 0.

Normalized Kullback-Leibler divergence D_{nKL}	Inf
Normalized Hellinger distance D_{nH}	5.179×10^{-4}
Normalized variational distance D_{nV}	1.381×10^{-2}

Table IV.1: f-divergences

By comparing the results of table IV.1 (except D_{nKL}) to the simulations of subsections III.2.1, III.3.1 and III.4.1, we claim that these f-divergences are "very small" values, thus that the two distributions are "very close". Hence, from an empirical point of view, the new formula of paper [Petot et al., 2018] seems to be correct.

IV.2 For the Choquet integral of exponential distributions

We are going to compare two (empirical) distributions $\mathbb{P}$ and $\mathbb{Q}$, both outputs of the Choquet integral of exponential distributions.

IV.2.1 Presenting the data

For computing IPMs, we need the samples X_p and X_q drawn from $\mathbb{P}$ and $\mathbb{Q}$. These samples are given figure IV.3, but in the form of histograms (and their estimated empirical probability distribution).

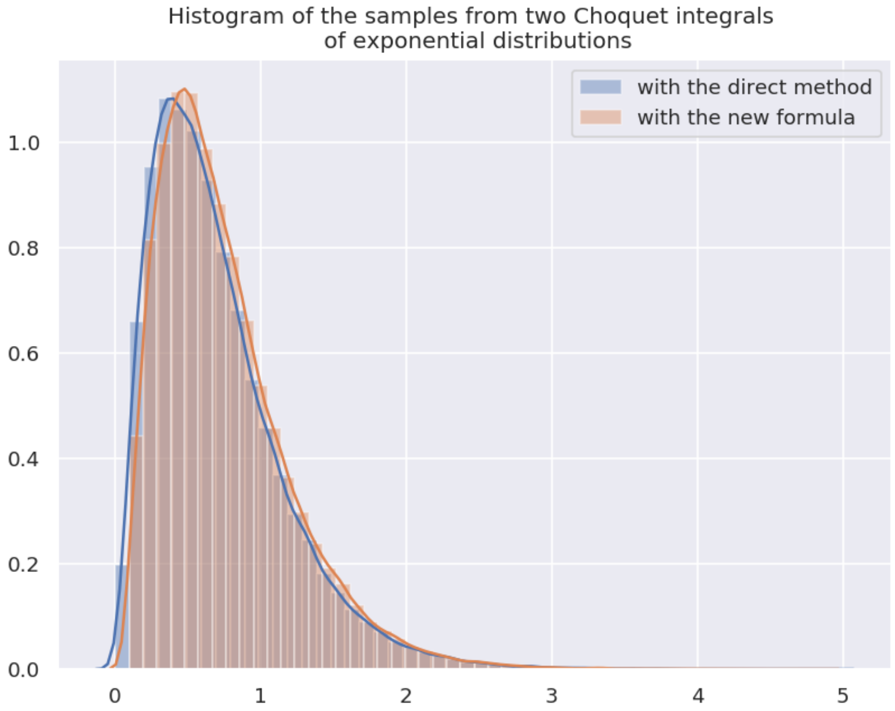


Figure IV.3: Data for IPMs: Histograms of the samples from two methods for computing the Choquet integral of exponential distributions

For computing f-divergences, we need the probability distributions, given figure IV.4. Each distribution has 51 samples.

IV.2.2 IPMs: Kantorovich metric

Because of the memory issue, we can only select (randomly) 100 samples per distribution: X'_p and X'_q then compute $W(X'_p, X'_q)$.

Hence, we perform ten random samplings of 100 samples per distribution, compute the Kantorovich metric of each pair of distributions. The mean of the ten Kantorovich metrics is $W = 0.776$ and the standard deviation is 0.051.

By comparing $W = 0.776$ to the values from the simulations of subsection II.4.2, we claim that $W = 0.776$ is a "very small" value, thus that the two distributions are "very close". Hence, from an empirical point of view, the new formula of paper [Petot et al., 2018] seems to be correct.

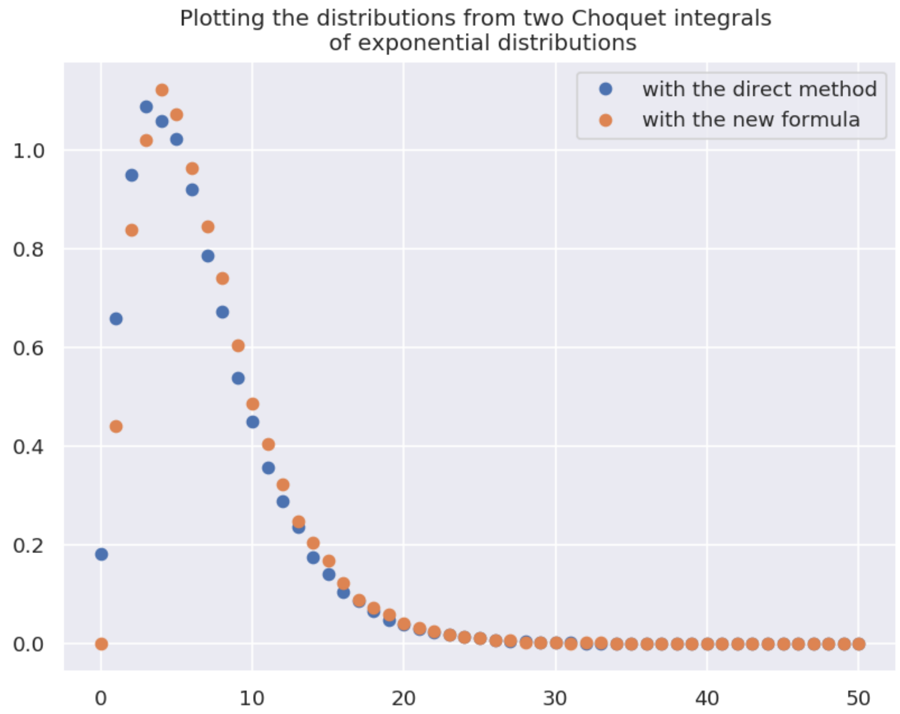


Figure IV.4: Data for f-divergences: empirical distributions from two methods for computing the Choquet integral of exponential distributions

IV.2.3 f-divergences

The results are displayed in table IV.2. Note that we have a problem with the normalized KL divergence because we must divide by 0.

Normalized Kullback-Leibler divergence D_{nKL}	Inf
Normalized Hellinger distance D_{nH}	4.391×10^{-3}
Normalized variational distance D_{nV}	2.295×10^{-2}

Table IV.2: f-divergences

By comparing the results of table IV.2 (except D_{nKL}) to the simulations of subsections III.2.2, III.3.2 and III.4.2, we claim that these f-divergences are "very small" values, thus that the two distributions are "very close". Hence, from an empirical point of view, the new formula of paper [Petot et al., 2018] seems to be correct.

Conclusion

Recapitulation

In this report, we studied several measures to compare two empirical probability distributions. Two categories of measures were addressed: integral probability metrics (IPMs) and f-divergences.

Let us recall that we wrote two intermediate conclusions on integral probability metrics (IPMs) at section II.5 and on f-divergences at section III.5.

For IPMs $\gamma_{\mathscr{F}}$, each choice of $\mathscr{F}$ leads to a specific IPM. For the numerical computation, we only dealt with the most popular IPM: the Kantorovich metric W.

For f-divergences, each choice of f leads to a specific f-divergence. We focused on the most popular f-divergences: the Kullback-Leibler divergence D_{KL}, the Hellinger distance D_{H} and the Varational distance D_{H}.

For both IPMs and f-divergences, we have carried out several experiments in order to observe how these measures evolve with the parameters of the distributions. For each measure, we analyzed the results on three common distributions: normal, exponential and uniform.

For the numerical computation of W, we need to solve a linear programming problem, which causes memory issues, thus we need the number of samples of the inputs to be "small" (inferior to 1 000). We used the `PuLP` library to solve the linear programming problem. The empirical results showed that W is a distance: when the difference between the parameters of the distributions increases, W increases. For each distribution, we can model W according to the difference between the parameters (of the distribution) and the model is trivial, thus relevant.

On the other hand, the numerical computation of f-divergences is very direct and effective: there are no memory issues. Their explicit formula can be "translated" into an algorithm with one line of code. We need to normalize these measures. The empirical results showed that D_{KL} is a "distance" (though it is not symmetric): when the difference of the parameters of the distributions increases, W increases; and the model is trivial. For D_{H} and D_{V}, no trivial model can be applied. However, D_{KL} does not deal well with empirical distributions that take the value 0.

Each "distance" being a compromise, there is no "distance" that appears to be better than the others: each "distance" has its advantages and drawbacks.

We applied IPMs and f-divergences to compare the output samples from the direct method for computing the Choquet integral – to the ones obtained with the new formula of paper [Petot et al., 2018]. By comparing to the "benchmark" of simulations in chapters II and III, we claimed that the distances were "very small", thus that the new formula of paper [Petot et al., 2018] seems to be correct (from an empirical point of view).

Further work

One could try to do more explicit (theoretical) computations of IPMs and f-divergences: we only did an explicit computation of the Kantorovich metric W for uniform distributions at subsection II.1.3. Hence, we could try to relate the distances to the parameters of the distributions in a more formal way.

One could also look into the empirical estimation of the Dudley metric β, an IPM introduced in chapter II, that requires solving a linear programming problem.

From a numerical point of view, one could try to address the issue of dividing by `0` in the computation of D_{KL} (without forgetting the normalization issue).

Furthermore, one could try to implement the solving of the linear programming problem in the computation of W in a more efficient way, so that the inputs could have more samples. Sparse matrices seem promising for example but they could not be implemented in the `PuLP` library.

Bibliography

M. Basseville. Distance measures for signal processing and pattern recognition. [Research Report]RR-0899, INRIA. 1988. inria-00075657, 1988. https://hal.inria.fr/inria-00075657.

G. Choquet. Theory of capacities. Annales de l'Institut Fourier, 5:131–295, 1954. doi: 10.5802/aif.53. URL https://aif.centre-mersenne.org/item/AIF_1954__5__131_0/.

I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. Foundations and Trends in Communications and Information Theory, 1, 2004.

I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

M. I. Jordan. f-divergences and surrogate loss functions, xxxx. https://people.eecs.berkeley.edu/~jordan/jordan-ssg.pdf.

I. Kojadinovic. Représentation de préférences à l'aide de l'intégrale de choquet, 2006.

S. Mitchell, S. M. Consulting, and I. Dunning. Pulp: A linear programming toolkit for python, 2011.

R. Pajda and G. Castera. Accélération du calcul de la densité de l'intégrale de choquet pour des entrées aléatoires, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

Y. Petot, P. Vallois, and A. Voisin. Choquet integral with stochastic entries. HAL, 2018.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. Electronic Journal of Statistics, 2012. http://php.scripts.psu.edu/users/b/k/bks18/euclid.ejs.1347974672.pdf.

Appendix A

Numerical computation of the Choquet integral of a vector

In this appendix, we are going to explain how to define a MATLAB function that returns the Choquet integral given a capacity μ and an input vector x.

In our example, the capacity μ is stored into `capa3.mat` with the variable name `Capa`. Section A.1 explains how we stored a capacity into a vector. For example, from an algorithmic point of view, how do we encode $\mu([\{1,2\}]) = 0.3651$?

A.1 How can we numerically store the values of a capacity?

We assume that the values of our capacity are already stored in the MATLAB file `capa3.mat` and we are going to explain how they were stored. `capa3.mat` is a vector of length 8, as is shown in the following MATLAB script:

```
>> load('capa3.mat')

>> size(Capa)
ans =
     1     8

>> Capa
Capa =
          0    0.0999    0.2538    0.3651    0.0118    0.6651
    0.4383    1.0000
```

`capa3.mat` represents a capacity μ with $\log_2(8) = 3$ criteria. (Indeed, $2^3 = 8$.) The values of μ are given in table A.1. The capacity index is just the MATLAB index of the vector (that starts with 1 in MATLAB). We introduce what we call the decimal index, that is the capacity index shifted so the index start at 0. We will understand in the next paragraphs why we prefer using the decimal index rather than the capacity index.

To which set do the capacity index correspond to? For example, as shown in table A.1, for a capacity index of 4, we have a capacity value of 0.3651, but to which set does the capacity value 0.3651 correspond to?

We encode a set into a decimal index is which then encoded into a capacity index (with a simple

capacity index	1	2	3	4	5	6	7	8
decimal index	0	1	2	3	4	5	6	7
capacity value	0	0.0999	0.2538	0.3651	0.0118	0.6651	0.4383	1

Table A.1: Values taken by the capacity `capa3.mat`

shift as shown in table A.1). Table A.2 explains how we can relate the (capacity) index of `capa3.mat` to their corresponding set (or node).

We obtain the binary encoding from the set: we put a 1 if the number (corresponding to the column) appears in the set and 0 otherwise. For example, the set $\{1,2\}$ contains the numbers 1 and 2, so that for the binary encoding, we put a value of one in the column corresponding to 1 and 2: $\underline{011}_2$.

We obtain the decimal encoding from the binary encoding in the usual way. For example, $\underline{011}_2 = \underline{\left(0\times 2^2+1\times 2^1+1\times 2^0\right)}_{10} = \underline{3}_{10}$ so we get a 3 for the decimal encoding of $\underline{011}_2$.

set	binary encoding			decimal encoding
(or node)	3	2	1	(or decimal index)
$\{\varnothing\}$	0	0	0	$0\times 2^2+0\times 2^1+0\times 2^0=0$
$\{1\}$	0	0	1	$0\times 2^2+0\times 2^1+1\times 2^0=1$
$\{2\}$	0	1	0	$0\times 2^2+1\times 2^1+0\times 2^0=2$
$\{1,2\}$	0	1	1	$0\times 2^2+1\times 2^1+1\times 2^0=3$
$\{3\}$	1	0	0	$1\times 2^2+0\times 2^1+0\times 2^0=4$
$\{1,3\}$	1	0	1	$1\times 2^2+0\times 2^1+1\times 2^0=5$
$\{2,3\}$	1	1	0	$1\times 2^2+1\times 2^1+0\times 2^0=6$
$\{1,2,3\}$	1	1	1	$1\times 2^2+1\times 2^1+1\times 2^0=7$

Table A.2: Correspondence between sets, their binary encoding and their decimal encoding

Now, we can now join table A.1 with table A.2 on the decimal index to obtain table A.3. Table A.3 gives the value of the capacity for a given set (or node). For example, set $\{1,2\}$ has a decimal index of 3 according to table A.2 and the decimal index of 3 corresponds to a capacity value of 0.3651 according to table A.1, so set $\{1,2\}$ has a capacity value of 0.3651.

set	capacity value
$\{\varnothing\}$	0
$\{1\}$	0.0999
$\{2\}$	0.2538
$\{1,2\}$	0.3651
$\{3\}$	0.0118
$\{1,3\}$	0.6651
$\{2,3\}$	0.4383
$\{1,2,3\}$	1

Table A.3: Capacity values of each set

Finally, we represent A.3 as a graph in table A.4. Each node in the graph corresponds to a set and to each set corresponds a capacity value. For example, $\{1,2\}_{0.3651}$ means that set $\{1,2\}$ has a capacity value of 0.3651.

	$\mathbf{\{1,2,3\}_1}$	
$\{1,2\}_{0.3651}$	$\mathbf{\{1,3\}_{0.6651}}$	$\{2,3\}_{0.4383}$
$\{1\}_{0.0999}$	$\{2\}_{0.2538}$	$\mathbf{\{3\}_{0.0118}}$
	$\boldsymbol{\varnothing_0}$	

Table A.4: Graph of sets with their capacity values

A.2 Computing by hand the Choquet integral with an encoded capacity

In this section, in order to understand how a MATLAB function for computing the Choquet integral should work, we compute a Choquet integral for a given x and capacity `capa3.mat` from section A.1 by hand.

Hence, we can check if the MATLAB function written in section A.4 returns the same and exact value.

Suppose that we have:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.12 \\ 0.7 \end{pmatrix}$$

thus having $x_2 \leqslant x_1 \leqslant x_3$ and $\sigma(1) = 2, \sigma(2) = 1, \sigma(3) = 3$. The path in the graph taken by our Choquet integral is in bold in table A.4.

For the capacity, we take the capacity `capa3.mat` given in table A.4:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.0999 \\ 0.2538 \\ 0.3651 \\ 0.0118 \\ 0.6651 \\ 0.4383 \\ 1 \end{pmatrix}$$

According to formula (I.5) and table A.4, let us compute the Choquet integral of x with respect to μ:

$$\begin{aligned} C_\mu(x_1, x_2, x_3) &= x_2 \left(\mu[\{2,1,3\}] - \mu[\{1,3\}]\right) \\ &\quad + x_1 \left(\mu[\{1,3\}] - \mu[\{3\}]\right) \\ &\quad + x_3 \left(\mu[\{3\}] - \mu[\varnothing]\right) \\ &= 0.12\,(1 - 0.6651]) \\ &\quad + 0.25\,(0.6651 - 0.0118) \\ &\quad + 0.7\,(0.0118 - 0) \\ &= 0.2118 \end{aligned}$$

Note that when computing the Choquet integral manually, we can directly move to the penultimate line (the second to last line) in the equations above by following the path on the graph on table A.4.

A.3 Getting the capacity index from the iteration on a specific path

In this section, we explain a trick we will use in our MATLAB function for computing the Choquet integral. We are going to give a way to get the capacity index of the nodes that belong to the path of our Choquet integral. Indeed, in practice, we do not need to get the capacity values of all our nodes, but only those that belong to the specific path.

We introduce the permutation $\tau \in \mathscr{S}_n$ such that:

$$x_{\tau(1)} \geqslant x_{\tau(2)} \geqslant \ldots \geqslant x_{\tau(n)} \tag{A.1}$$

We take the same values for x and μ as in subsection A.2. We have $\tau(1) = 3, \tau(2) = 1, \tau(3) = 2$.

Note that τ is different from the permutation σ introduced in formula (I.3).

Let c be the vector containing the capacity indexes of the nodes in our specific path. Note that c is the vector whose values we want to compute with a trick. We have:

$$c = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

where c_1 is capacity index corresponding to the set $\{\tau(1)\}$, c_2 to $\{\tau(1), \tau(2)\}$ and c_3 to $\{\tau(1), \tau(2), \tau(3)\}$.

Actually, c_1 is the capacity index of the first node of our path (after node $\{\varnothing\}$), c_2 the capacity index of second node in our path and c_3 the capacity index of the third node in our path. Actually, $i \in [\![1,3]\!]$ corresponds to a step in the path.

Here, we have $\{\tau(1)\} = \{3\}, \{\tau(1), \tau(2)\} = \{2,1\}, \{\tau(1), \tau(2), \tau(3)\} = \{2,1,3\}$. Thus, according to table A.2, we have:

$$c = \begin{pmatrix} 5 \\ 6 \\ 8 \end{pmatrix}$$

Can we get a formula for computing the values of c? Yes. We must take a closer look at table A.2. We note that the capacity index of the set $\{2\}$ is $2^{2-1} + 1 = 3$, the capacity index of the set $\{1,2\}$ is $2^{1-1} + 2^{2-1} + 1 = 4$, the capacity index of the set $\{3,2\}$ is $2^{3-1} + 2^{2-1} + 1 = 7$, the capacity index of the set $\{1,2,3\}$ is $2^{1-1} + 2^{2-1} + 2^{3-1} + 1 = 8$, and so on. Thus, we have:

$$\forall i \in [\![1,3]\!] \quad c_i = 1 + \sum_{j \in \tau(1:i)} 2^{\tau(j)-1} \tag{A.2}$$

For verification, according to (A.2), we have:

$$\begin{aligned}
c_1 &= 1 + 2^{\tau(1)-1} = 1 + 2^{3-1} = 5 \\
c_2 &= 1 + 2^{\tau(1)-1} + 2^{\tau(2)-1} = 1 + 2^{3-1} + 2^{1-1} = 6 \\
c_3 &= 1 + 2^{\tau(1)-1} + 2^{\tau(2)-1} + 2^{\tau(3)-1} = 1 + 2^{3-1} + 2^{1-1} + 2^{2-1} = 8
\end{aligned}$$

which is correct.

A.4 MATLAB function returning the Choquet integral

Taking into account the previous sections, we can now define a MATLAB function returning the Choquet integral for a given input vector x and a given capacity μ.

For the Choquet integral, we use formula (I.6) but instead of σ, we take τ as defined in (A.1). Thus, we have:

$$C_\mu(x) = \sum_{i=1}^{n} \left(x_{\tau(i)} - x_{\tau(i+1)}\right) \mu[\tau(i:n)] \tag{A.3}$$

The MATLAB code for computing the Choquet integral I received from Alexandre Voisin is the following:

```matlab
function OUT=CI(mu,IN)

n=size(IN,1);
NCrit=size(IN,2);
[Xv Xi]=sort(IN,2,'descend');
OUT=zeros(n,1);

%calcul des indices des mus associ? aux sous ensemble de crit?rea i.e.
%c(1) --> x(1)
%c(2) --> {x(1),x(2)}
%c(3) --> {x(1),x(2),x(3)}...
if size(IN,1)==1 %cas o? l'on a une seule valeur ? calculer
    c=cumsum(2.^(Xi-1))+1;

    for j=1:NCrit-1
        OUT=OUT+(Xv(:,j)-Xv(:,j+1)).*mu(c(j));
    end
    OUT=OUT+Xv(NCrit).*mu(c(NCrit));

else
    c=cumsum(2.^(Xi-1),2)+1; %cas o? l'ion a un ensemble de valeurs ? calculer

    for j=1:NCrit-1
        OUT=OUT+(Xv(:,j)-Xv(:,j+1)).*mu(c(:,j))';
    end
    OUT=OUT+Xv(:,NCrit).*mu(c(:,NCrit))';
end
end
```

I changed it into:

```matlab
function [CI] = choquet_integral(mu, X)
% Computes the (discrete) Choquet integral of a vector.
% Parameters
%   mu : capacity
%   X : input values
% Returns
%   CI : Choquet integral of each input value

n = size(X,1) ; % number of input values, number of vectors
p = size(X,2) ; % number of criteria
CI = zeros(n,1) ; % Choquet integral

```

```
% We do the permutation:
[X_tau, tau] = sort(X, 2, 'descend') ; % and not ascent??

% We compute the vector containing the capacity index of our path:
c = cumsum(2.^(tau-1),2)+1 ;

% We compute the Choquet integral:
for j=1:p-1
    CI = CI + (X_tau(:,j)-X_tau(:,j+1)).*mu(c(:,j))' ;
end
CI = CI + X_tau(:,p).*mu(c(:,p))' ;

end
```

Running the following script:

```
% Loading the capacity
load('capa3.mat') % has the variable name Capa

X = [0.25, 0.12, 0.7] ; % input values
mu = Capa ; % capacity
choquet_integral(mu, X)

% Computation by hand:
res = 0.12*(1-0.6651) + 0.25*(0.6651-0.0118) + 0.7*(0.0118-0)
```

gives us:

```
ans =
    0.2118

res =
    0.2118
```

Thus, our MATLAB function `choquet_integral` returns the same result as the computation by hand of section A.2. We have defined a function that can compute the Choquet integral.