



Information k-means, fragmentation and syntax analysis. A new approach to unsupervised machine learning

Gautier Appert

► To cite this version:

Gautier Appert. Information k-means, fragmentation and syntax analysis. A new approach to unsupervised machine learning. Machine Learning [stat.ML]. Institut Polytechnique de Paris, 2020. English. NNT : 2020IPPAG011 . tel-03015285

HAL Id: tel-03015285

<https://theses.hal.science/tel-03015285>

Submitted on 19 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS



IP PARIS

Information k -means, fragmentation and syntax analysis: A new approach to unsupervised machine learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 29 Octobre 2020, par

GAUTIER APPERT

Composition du Jury :

Sylvain Arlot Professeur, Université Paris-Saclay	Examineur
Patrice Bertail Professeur, Université Paris Nanterre	Président du jury
Cristina Butucea Professeur, ENSAE Paris	Examineur
Olivier Catoni Directeur de recherche CNRS, ENSAE Paris	Directeur de thèse
Stéphane Gaïffas Professeur, Université Paris Diderot	Rapporteur
Gábor Lugosi Professeur, Pompeu Fabra University	Rapporteur

INFORMATION k -MEANS, FRAGMENTATION AND SYNTAX ANALYSIS:
A NEW APPROACH TO UNSUPERVISED MACHINE LEARNING

Gautier Appert

Why the awe for the Second Law ?
The Second Law of Thermodynamics
defines the ultimate purpose of life,
mind, and human striving: to deploy
energy and information to fight back
the tide of entropy and carve out
refuges of beneficial order.

Steven Pinker

ACKNOWLEDGMENTS/REMERCIEMENTS

Mes remerciements s'adressent en premier lieu à mon directeur de thèse Olivier Catoni sans qui ce projet n'aurait jamais pu voir le jour. Olivier m'a toujours soutenu et orienté dans la démarche à adopter tout au long de la thèse. En effet, lorsque je rencontrais des difficultés Olivier m'a continuellement accompagné et guidé pour les surmonter. C'est pourquoi je suis très reconnaissant du temps qu'Olivier m'a consacré, ainsi que de la patience qu'il m'a accordée. De plus, les rendez-vous qui ont eu lieu durant toute la thèse m'ont permis de découvrir la recherche en mathématiques, une approche pour trouver de nouvelles idées et comment les formaliser. J'ai beaucoup appris à ses côtés, et j'ai pu à la fois admirer et profiter de son esprit créatif. Par ailleurs, je garde aussi un excellent souvenir de toutes les discussions qu'on a pu avoir dans d'autres domaines, notamment la programmation et l'informatique, en particulier toutes les astuces et commandes sous Linux, ainsi que l'administration d'un serveur à distance. Sans compter les discussions autour du sport, et surtout la natation avec ses différentes techniques de nage. Je suis toujours étonné des astuces (justifiées par la mécanique des fluides) sur le crawl. Enfin, je tiens particulièrement à remercier Olivier pour son soutien considérable durant cette période difficile liée au contexte de la crise sanitaire.

I would like also to thank Professors Gabor Lugosi and Stéphane Gaïffas for having accepted to spend time and effort to report my manuscript. I am very grateful for their careful reports along with valuable comments and insightful suggestions. Je remercie aussi sincèrement Sylvain Arlot, Patrice Bertail et Cristina Butucea qui m'ont fait l'honneur d'avoir accepté d'être membres du jury.

Je voudrais également remercier très chaleureusement Pierre Alquier qui m'a suggéré de prendre contact avec Olivier pour démarrer cette thèse. Je te remercie pour toute l'aide administrative liée à la thèse et ta bienveillance générale avec les doctorants. J'ai aussi grandement apprécié les discussions qu'on a pu avoir en mathématique de manière générale durant les nombreuses pauses café. J'ai aussi apprécié tous les restaurants japonais, chinois et vietnamiens qu'on a pu fréquenter à Paris durant cette période. Tu as toujours été un modèle que ce soit sur le plan des maths ou sur le plan humain, et ton charisme mêlé avec ta passion des maths sont hautement contagieux. Je te remercie aussi pour ton aide concernant l'enseignement des travaux dirigés de ton cours de Machine Learning à L'ENSAE. Tu étais toujours disponible quand j'avais des questions concernant l'explication des idées et techniques à employer. Un grand merci aussi pour ta relecture judicieuse de ce manuscrit et le repérage de nombreuses typos.

J'aimerais adresser un remerciement spécial à mon très proche ami Guillaume Salha, pour qui j'ai beaucoup d'affection et de reconnaissance. Durant cette thèse je garde un excellent souvenir des passions qu'on a pu partager, allant des maths au machine learning, en passant par la nourriture, la culture chinoise (l'alcool parfois aussi) et bien d'autres encore. Tu m'as énormément soutenu durant cette thèse, et tes nombreux conseils ont toujours été d'une grande utilité. Tu as toujours été d'un naturel altruiste envers moi et j'en suis très

reconnaissant. Par ailleurs, ce fut un réel plaisir de découvrir ton Pays Basque natal, j'en retiens un très bon souvenir.

De même, je souhaite remercier en particulier mon ami Gabriel Romon, pour la transmission contagieuse de mathématiques pour le moins très (très) techniques. Tu m'as toujours rendu fou avec tes problèmes d'Olympiade de maths tellement sioux ou tes problèmes de théorie de la mesure ou probabilités très subtils. Je ne comprenais jamais rien mais dans l'ensemble ton influence m'a fait progresser dans ce domaine, et ça a été un plaisir de partager cela avec toi. Je tiens aussi à te remercier pour ton aide précieuse quand j'avais des questions pour l'enseignement des travaux dirigés de théorie de la mesure à l'université Paris-Saclay.

Par la même occasion, je tiens à remercier Émilie Le Nhu pour son enthousiasme et sa motivation, que ce soit pour le sport ou les maths. Un grand merci aussi pour les phos préparés par tes parents, ils sont tellement bons.

Je remercie aussi mes amis Félix Pasquier et Zoe Fontier avec qui j'ai passé de très bons moments durant les cours de chinois à l'ENSAE.

Mes remerciements vont aussi à Sisi Yang, avec qui j'ai pu collaborer et appliquer mes connaissances en statistique dans le domaine médical, et plus précisément la radiologie.

Je tiens à remercier aussi la meilleure team Saumon jamais rencontrée: Badr, Mohamed (Simo), Lionel, Goeffrey, Jérémy, Alexis, Jiaying, Avo, Simon, Georgy, Charly, Jean-Baptiste Remy et bien entendu ceux déjà cités auparavant Pierre, Guillaume, Gabriel, Emilie, Zoe, Félix et Sisi.

J'adresse aussi mes remerciements à Lucie, Christophe, Vincent, Jules et Victor pour toute l'aide administrative concernant les enseignements à l'ENSAE et la thèse.

Je souhaite également remercier le laboratoire du CREST avec ses enseignants chercheurs dans son ensemble pour son accueil, son organisation et son ambiance.

J'ai une aussi pensée chaleureuse envers tous mes amis sportifs qui m'ont apporté aussi un soutien considérable: Thierry, Azzedine, Flavio, Christopher, Sonia, Mohamed, Samuel, Nicolas, Sylvia, Maïssoun, Ahmed, Vik, Albert, Ryan, Jérémy, et j'en oublie tellement.

Un énorme merci à mes amis de Toulouse sans qui je ne serais probablement jamais monté à Paris et allé à l'ENSAE: Jérémy Atia, Astrid Beteille, Achraf Elmarhraoui, Dan Hua, Alexandre Crayssac, Maryan Morel et Quentin Villotta.

Je remercie aussi ma famille pour tout son soutien et ses encouragements durant cette thèse, en particulier mon frère Brice pour ses blagues liées au renaaaaaaard.

Mes remerciements vont aussi tout particulièrement à ma femme Man Zhang, mon Oracle, qui a toujours été à mes côtés durant cette thèse. Merci pour tes encouragements, ton soutien permanent et le fait que tu as toujours cru en moi. Enfin mes remerciements s'adressent à mes deux chats Miaomiao et Saumon pour tout l'amour et la bonne humeur qu'ils nous apportent chaque jour.



Figure 1: Chat Saumon et chat Miaomiao.

Contents

General notation	9
1 Introduction	11
2 Overview	17
2.1. GENERAL IDEAS	17
2.2. GENERALIZATION BOUNDS FOR FRAGMENTATION	27
2.3. DESCRIPTION OF THE SIGNAL FRAGMENTATION ALGORITHM	29
2.4. SYNTAX ANALYSIS	31
2.5. RELATION WITH STATISTICAL ESTIMATION	37
3 Information k-means and information fragmentation algorithms	39
3.1. INFORMATION k -MEANS ALGORITHMS.	39
3.2. INFORMATION FRAGMENTATION	49
3.2.1. RECALL OF THE INFORMATION k -MEANS SETTING	49
3.2.2. FIRST GENERALIZATION: ESTIMATING A JOINT DISTRIBUTION	50
3.2.3. INFORMATION FRAGMENTATION	51
3.3. SIGNAL FRAGMENTATION	56
4 PAC-Bayesian bounds for information k-means and information fragmentation	63
4.1. A PAC-BAYESIAN BOUND FOR INFORMATION k -MEANS.	63
4.1.1. EXPLICIT BOUND IN THE INFORMATION k -MEANS SETTING	71
4.1.2. CLASSICAL k -MEANS QUANTIZATION IN A SEPARABLE HILBERT SPACE	73

4.1.3. DISCUSSION ABOUT THE BOUNDS	74
4.2. A BOUNDED CRITERION FOR INFORMATION k -MEANS.	75
4.3. A BOUNDED CRITERION FOR THE EUCLIDEAN k -MEANS	80
4.4. PAC-BAYESIAN BOUNDS FOR INFORMATION FRAGMENTATION.	81
4.5. FASTER BOUNDS	94
4.5.1. FASTER BOUNDS FOR INFORMATION k -MEANS AND CLASSICAL k -MEANS	94
4.5.2. FASTER BOUNDS FOR BOTH THE BOUNDED INFORMATION k -MEANS AND THE BOUNDED k -MEANS CRITERION	103
4.5.3. FASTER BOUNDS FOR INFORMATION FRAGMENTATION	105
5 Experiment on digital images	115
6 Conclusion	127
A Code highlights	131
B Présentation générale	139
References	151

General notation

We will use the following notation throughout this document.

On some measurable probability space Ω , we will consider various random variables $X : \Omega \rightarrow \mathfrak{X}$, $Y : \Omega \rightarrow \mathfrak{Y}$, etc. that are nothing but measurable functions. We will also consider several probability measures on Ω , and typically two measures \mathbb{P} and $Q \in \mathfrak{M}_+^1(\Omega)$, where \mathbb{P} describes the usually unknown data distribution and Q describes an estimation of \mathbb{P} . Then we will use the short notation \mathbb{P}_X for the push forward measure $\mathbb{P} \circ X^{-1}$. Similarly we will let $Q_X = Q \circ X^{-1}$. In the same way $\mathbb{P}_{X,Y} \in \mathfrak{M}_+^1(\mathfrak{X} \times \mathfrak{Y})$ will be the joint distribution of the couple (X, Y) under \mathbb{P} and $\mathbb{P}_{Y|X}$ the corresponding regular conditional probability measure of Y knowing X when it exists. We will always work under sufficient hypotheses to ensure that the decomposition

$$\mathbb{P}_{X,Y} = \mathbb{P}_X \mathbb{P}_{Y|X} \tag{1}$$

is valid, meaning that for any bounded measurable function $f(X, Y)$

$$\int f \, d\mathbb{P}_{X,Y} = \int \left(\int f \, d\mathbb{P}_{Y|X} \right) d\mathbb{P}_X.$$

Moreover, we will use the short notation

$$\int f \, d\mathbb{P}_{X,Y} = \mathbb{P}_{X,Y}(f),$$

so that the previous formula becomes

$$\mathbb{P}_{X,Y}(f) = \mathbb{P}_X[\mathbb{P}_{Y|X}(f)].$$

We will often use the Kullback Leibler divergence

$$\mathfrak{K}(Q, \mathbb{P}) = \begin{cases} Q \left(\log \left(\frac{dQ}{d\mathbb{P}} \right) \right) & \text{when } Q \ll \mathbb{P}, \\ +\infty & \text{otherwise.} \end{cases}$$

We will always be in this memoir in a situation where the decomposition

LEMMA 1

$$\begin{aligned} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) &= \mathfrak{K}(Q_X, \mathbb{P}_X) + Q_X[\mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X})] \\ &= \mathfrak{K}(Q_Y, \mathbb{P}_Y) + Q_Y[\mathfrak{K}(Q_{X|Y}, \mathbb{P}_{X|Y})] \end{aligned}$$

is valid.

PROOF. It follows from the decomposition (1). A precise statement and a rigorous proof dealing with measurability issues can be found in [Cat04, Appendix section 1.7 page 50]. \square

CHAPTER 1

Introduction

Over the last decades, machine learning and statistical models have been used extensively to treat large amounts of digital data of any type, such as images, speech, and texts ... One of the main problems faced by data scientists is to predict the value of one specific variable as a function of others, based on past history, consigned in a training set. The situation is usually formalized in the following way: we consider a training set

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

consisting in n pairs of independent and identically distributed random variables with unknown joint distribution $\mathbb{P}_{X,Y}$. The goal of the machine learning algorithm is to compute some function $f(X)$ that can predict the outcome Y associated to X . This framework is called supervised learning and can be divided into two categories depending on whether Y is a quantitative or qualitative variable. In the first case, when $Y \in \mathbb{R}^d$, we are dealing with a regression problem and in the second case when Y belongs to a finite set, we are dealing with a classification problem. The computation of the function f is usually performed using an algorithm that minimizes an empirical risk criterion

$$\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i)$$

where \mathcal{F} represents a certain class of functions and $L(\cdot)$ some loss function adapted to the problem. For instance, in the classification setting, one can take $L(Y, f(X)) = \mathbb{1}(Y \neq f(X))$ and in the regression framework $L(Y, f(X)) = \|Y - f(X)\|_2^2$. The set of functions \mathcal{F} depends also on the setting, typically in the regression framework, one can consider the class of linear predictors $\mathcal{F} = \{f_\beta(X) = X^\top \beta, \beta \in \mathbb{R}^d\}$, whereas in the classification case one can take $\mathcal{F} = \{f_\beta(X) = \mathbb{1}(\mathbb{P}_\beta(Y = 1 | X) \geq 1/2)\}$, where $\mathbb{P}_\beta(Y = 1 | X) = 1/(1 + e^{-X^\top \beta})$, which corresponds to the logistic model.

However, beforehand, data scientists have usually to extract features from the variables (X_1, \dots, X_n) on the occasion of an exploratory analysis. It consists in extracting the essential information carried by the explanatory variables and in some cases in reducing the dimension of the feature space, typically using principal component analysis. This data

preprocessing step is necessary to learn afterwards a classification function f and obtain a better classification performance.

This preprocessing step is comparable to the case when the available data set

$$(X_1, \dots, X_n) \sim \mathbb{P}_X^{\otimes n}$$

does not come with attached labels (Y_1, \dots, Y_n) . The question is then simply to identify patterns or characteristic structures. Typically, one would like to cluster the data into meaningful groups, or to extract informative patterns. This task can be viewed as finding a new representation of the data that highlights its content in a more usable way.

We are touching here on the daunting matter of unsupervised learning, where we do not even know how to measure the quality of the result. Nonetheless, this is a key issue, as most of the available data are unlabelled, and as asking human experts to provide labels in order to train supervised learning algorithms is costly. This need is at the origin of this project, where we propose new ideas with a focus on digital images. We have a double interest in clustering and in changing the representation and will follow a simple guideline : use data compression as a tool for data understanding.

To be more precise, we will describe a series of transformations of the representation of an i.i.d. sample of digital images $(X_1, \dots, X_n) \sim \mathbb{P}_X^{\otimes n}$, where $X_i \in \mathbb{R}^d$. We will start with a lossy coding scheme based on labelling image regions, that can be seen as a multiple labels extension of the k -means algorithm.

This procedure will learn a classification function

$$\ell : \llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \mapsto \llbracket 1, k \rrbracket$$

defined by the family of product sets

$$\ell^{-1}(j) = A_j \times B_j \subset \llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket$$

and a set of image fragments $C_j \in \mathbb{R}^d, 1 \leq j \leq k$ where $\text{supp}(C_j) \subset B_j$. They define an approximation Y_i of each image X_i given by the formula

$$Y_i = \sum_{j=1}^k \mathbf{1}(i \in A_j) C_j, \quad i \in \llbracket 1, n \rrbracket,$$

or equivalently by the formula

$$Y_{i,s} = C_{\ell(i,s),s}, \quad i \in \llbracket 1, n \rrbracket, s \in \llbracket 1, d \rrbracket.$$

We will consider a coding distribution $q(\theta)$ on the parameter $\theta = (A, B, C)$, and a measure $D(\theta)$ of the distortion of the representation of (X_1, \dots, X_n) by θ , based on an entropy criterion. We will learn a representation $\hat{\theta}(X_1, \dots, X_n)$ whose distortion is under a given level and whose ideal code length $-\log_2(q(\hat{\theta}))$ is as small as possible. The procedure produces image fragments C_j entering into the description of a set of sample images A_j as large as

possible. From this perspective, it can be viewed as a block indexing scheme reminiscent of the Lempel Ziv algorithm. Another interesting feature of our algorithm is that it does not use the image geometry. Although it has been tested on images, it could have been used on any vectors of d measurements. It is indeed invariant by any permutation of the pixel indices and could be without modification applied to other types of digital signals, like rgb images, stereoscopic images, 3D images, video samples, speech signals etc. Moreover the image fragmentation algorithm produces a discrete representation that can be augmented with other discrete descriptors, like text annotations.

Once the fragmentation step is performed, the signal is described as a random set of labels indexing image fragments. Borrowing ideas from linguistic theory, we then learn rewriting rules to compress even more the representation. The algorithm searches repeatedly for the most frequent pair of labels to be reindexed by a single new label. Decoding is then performed by rewriting the new label into the original pair of labels. Again, this compression scheme is reminiscent of the Lempel Ziv algorithm. While in the Lempel Ziv algorithm though, a new indexed block is created by adding one bit to an old block, here a new block is made from a pair of non overlapping old blocks. This type of procedure learns automatically a set of rewriting rules, forming a context free grammar. Further compression steps are then performed by factorizing and compressing again the grammar obtained at the previous step. We obtain a grammar of the grammar that performs a kind of syntax analysis and defines syntax labels. The process can then be repeated to obtain a hierarchy of syntax labels forming a syntax tree (or rather a syntax forest in the general case).

Thus, instead of estimating a statistical model for the data, we learn a grammar through a compression algorithm. The two approaches are different, although a link can be made, considering that a lossless compression code defines a sub-probability measure, due to the Kraft inequality. A good binary code is hard to compress further, so that its distribution is close to a sequence of i.i.d. Bernoulli random variables with parameter $1/2$. For this reason, this compression approach can be seen as an alternative to the selection and estimation of a statistical model based on conditional independence properties.

It avoids the difficulty of performing multiple tests of independence that was present in [Mai14]. It allows to bypass the estimation of conditional probability distributions in high dimension and overcomes the problem of model selection. To a certain extent, it is able to circumvent the statistical issues due to the curse of dimensionality. From a practical point of view, since it computes a simpler (compressed) representation of images, there is hope it is scalable and can cope with large datasets (although we did not test this yet).

Nevertheless, we will establish a link with the estimation of the joint probability distribution of random sets of labels. In particular, we will provide a sub-probability measure estimator thanks to the Kraft inequality. Indeed, according to the Kraft inequality, for any prefix binary code $(c_i)_{i \geq 1}$

$$\sum_{i \geq 1} 2^{-l(c_i)} \leq 1,$$

where $l(c_i)$ is the length in bits of c_i .

Then, we will provide an estimation of the law of one single random set using some kind of Bayesian Shtarkov type estimator [Tri16]. We will provide also an oracle inequality to measure the quality of this estimation.

Let us conclude this introduction with a presentation of k -means clustering, the algorithm at the heart of vector quantization and lossy compression that will serve us as a starting point.

The k -means algorithm was suggested by [Llo06] and it is sometimes referred to as Lloyd's algorithm. It provides a partition of a data set $\mathcal{D}_n = \{X_1, \dots, X_n \mid X_i \in \mathbb{R}^p\} \sim \mathbb{P}_X^{\otimes n}$, into k distinct non-overlapping clusters by minimizing the k -means criterion or within-cluster inertia criterion

$$\inf_{\ell: \mathbb{R}^d \rightarrow \{1, \dots, k\}} \inf_{(\mu_1, \dots, \mu_k) \in \mathbb{R}^{d \times k}} \frac{1}{n} \sum_{i=1}^n d(X_i, \mu_{\ell(i)})^2, \quad (1.1)$$

where $\ell: \mathbb{R}^p \rightarrow \{1, \dots, k\}$ denotes the labeling function or cluster assignment function, whereas $(\mu_j)_{j \in [1, k]}$ are the cluster centers (also called centroids) and $d(., .)$ indicates some distance or dissimilarity measure. It should be pointed out that criterion (1.1) is the empirical counterpart of the theoretical k -means objective function

$$\inf_{\ell: \mathbb{R}^d \rightarrow \{1, \dots, k\}} \inf_{(\mu_1, \dots, \mu_k) \in \mathbb{R}^{d \times k}} \mathbb{P}_X [d(X, \mu_{\ell(X)})^2],$$

where $\mathbb{P}_X[.]$ denotes the expectation with respect to X . Minimizing the previous criterion (1.1) requires to explore all partitions of $\{1, \dots, n\}$ into k groups. However this task represents a combinatorial optimization problem which is known to be NP-hard and hence infeasible at a practical level.

As an alternative, the k -means algorithm tries to decrease the value of the objective function in an iterative fashion by allocating each data point to the cluster with the nearest centroid and recomputing the center from this partition, see a complete description in algorithm 1 described below. In this way, the k -means algorithm decreases the value of the criterion at each step but converges generally to a local minimum.

Besides, the results obtained depend on the initialization of the k -means algorithm, that is for this reason often randomized. One way to avoid bad clustering due to bad initialization, is to repeat the operation several times and select the one for which the within-cluster inertia criterion is the smallest.

By far, the Euclidean distance $d(x, y) = \|x - y\|_2$ is the most common choice, which leads to compute centroids as empirical means inside each cluster. However, the Euclidean distance can give rise to bad clustering since the empirical mean is not robust. One way to overcome this point is to employ robust estimators for the mean of a random vector. For instance, one may use the estimator suggested in [CG18], since the resulting estimator satisfies tight concentration bounds and is straightforward to compute. An other way to

tackle this issue is to consider other distances more robust to noise and outliers, for instance the L^1 norm.

Algorithm 1 k -means

Initialization :

Pick at random k individuals in the data set that represent the initial centroids of the k clusters.

Iterate the following steps until the within-cluster inertia criterion converges :

- Put each individual in the cluster indexed by the closest centroid, that is for all $1 \leq i \leq n$

$$\ell^*(i) = \arg \min_{j \in \llbracket 1, k \rrbracket} d(x_i, \mu_j)^2.$$

- Compute the centers of gravity of each cluster, that is for all $1 \leq j \leq k$

$$\mu_j^* = \arg \min_{\mu_j} \sum_{i \in \ell^{-1}(j)} d(x_i, \mu_j)^2.$$

These centers of gravity become the new centroids.

return $((\ell^{-1}(j))_{j \in \llbracket 1, k \rrbracket}, (\mu_j^*)_{j \in \llbracket 1, k \rrbracket})$

In what follows we will introduce a fragmentation algorithm that uses a generalization of the k -means criterion. However, we will propose an algorithm that is very different from Lloyd's algorithm. We will indeed keep the criterion under a given level while optimizing what stands for the number k of centroids. Doing this, we will avoid the difficult question of the choice of k . We will take this opportunity to prove new dimension free generalization bounds for the k -means criterion and for the fragmentation criterion. These non asymptotic bounds decrease like $(k \log(k)/n)^{1/4}$ or like $\log(n/k) \sqrt{k \log(k)/n}$ and improve on previously known ones. They are based on PAC-Bayesian lemmas and some new kind of PAC-Bayesian chaining. Bounds for fragmentation show what quantity takes the place of the number k of centroids in this generalized framework: it is the sum of the sizes of the fragments (normalized by the image size, so that we get back k when we consider in the original k -means setting each centroid as a big fragment covering the whole pixel grid).

CHAPTER 2

Overview

2.1. GENERAL IDEAS

In this study, we will make new proposals for the unsupervised classification of signals, taking the example of digital images.

The aim of unsupervised classification, as we conceive it, is to propose various classification functions, with the hope that at least some of them may be useful to solve interesting tasks.

We will take as a starting point the k -means algorithm with Euclidean distance. Given $(X_1, \dots, X_n) \in \mathbb{R}^{d \times n}$ a training sample, and k centers $c_i, 1 \leq i \leq k$, the empirical loss of the k -means algorithm is

$$L(c_1, \dots, c_k) = \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \|X_i - c_j\|^2 = \inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \bar{\mathbb{P}}_X(\|X - c_{\ell(X)}\|^2),$$

where $\bar{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure. This empirical loss is related to the expected loss

$$\inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \mathbb{P}_X(\|X - c_{\ell(X)}\|^2).$$

The first thing we will do, starting from there, is to see what we get if we consider the square of the Euclidean norm as a Gaussian Kullback divergence. To make this interpretation, we add to $X \in \mathbb{R}^d$ a second random variable Y such that $\mathbb{P}_{Y|X} = \mathcal{N}(X, \sigma^2 I_d)$, where I_d is the identity matrix of size $d \times d$. We get that

$$\|X - c_{\ell(X)}\|^2 = 2\sigma^2 \mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X}),$$

where $Q_{Y|X} = \mathcal{N}(c_{\ell(X)}, \sigma^2 I_d) = Q_{Y|\ell(X)}$. It is interesting to consider $\mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X})$ rather than $\mathfrak{K}(\mathbb{P}_{Y|X}, Q_{Y|X})$ that are equal in this case, because of the following property.

PROPOSITION 2 *The Euclidean k -means criterion can be viewed as an information k -means criterion due to the identity*

$$\inf_{c_1, \dots, c_k} \mathbb{P}_X(\|X - c_{\ell(X)}\|^2) = 2\sigma^2 \inf_Q \mathbb{P}_X[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X})].$$

The important thing here is that we do not have to restrict the infimum to

$$\left\{ Q : Q_{Y|X} = \mathcal{N}(c_{\ell(X)}, \sigma^2 I_d) \right\}.$$

PROOF. This is a consequence of Proposition 11 on page 45. \square

This proposition shows that the k -means quadratic criterion is a special case of an information k -means criterion

$$\inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \inf_Q \mathbb{P} \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right],$$

and of its empirical counterpart

$$\inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \inf_Q \overline{\mathbb{P}} \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right].$$

This information k -means criterion broadens the scope of the k -means algorithm from classifying vectors to classifying distributions. In this case, the data set $X_1, \dots, X_n \sim \mathbb{P}_X^{\otimes n}$ is replaced with a set of conditional probability measures p_{X_1}, \dots, p_{X_n} . For instance in text-mining, histograms made of word counts, called *bags of words* are often used to represent documents. Similarly in computer vision, images may be represented by histograms of visual features. It is worth noticing that bag of visual features are often the result of some clustering algorithm, that may be the k -means algorithm, run on a specific set of image patches to produce a dictionary of features. For more details, one can refer to [Tsa12].

In the information k -means setting, we try to approximate $\mathbb{P}_{Y|X}$ by $Q_{Y|\ell(X)}$. This is an invitation to consider as a variant of the same idea the approximation of the joint distribution $\mathbb{P}_{X,Y}$ by $Q_{X,Y}$ such that $Q_{Y|X} = Q_{Y|\ell(X)}$.

From the disintegration theorem notice that

$$\mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = \mathcal{K}(Q_X, \mathbb{P}_X) + Q_X[\mathcal{K}(Q_{Y|X}, \mathbb{P}_{Y|X})].$$

Thus, considering the specific model

$$\mathcal{Q} = \left\{ Q_{X,Y} : Q_X = \mathbb{P}_X, Q_{Y|X} = Q_{Y|\ell(X)}, \ell(X) \in \{1, \dots, k\} \right\},$$

the information k -means can be expressed as an information projection

$$\inf_{Q_{X,Y} \in \mathcal{Q}} \mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \inf_{Q_{Y|\ell(X)} \in \mathcal{M}_+^1(\mathcal{Y})} \mathbb{P}_X \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right].$$

The information projection, also called I-projection [Csi75], consists in projecting a probability measure P onto a set \mathcal{Q} of probability distributions, solving

$$\inf_{Q \in \mathcal{Q}} \mathcal{K}(Q, P).$$

This concept appears also in Sanov's theorem [Csi84] which provides a bound on the probability of the empirical measure $\bar{\mathbb{P}}_n$ to belong to a set of probability distributions \mathfrak{Q} , informally

$$-\log\left(\mathbb{P}_X^{\otimes n}(\bar{\mathbb{P}}_n \in \mathfrak{Q})\right) \sim n \inf_{Q \in \mathfrak{Q}} \mathcal{K}(Q, \mathbb{P}_X).$$

The difference between maximum likelihood estimation, that can be written as

$$\hat{\theta}_{\text{MLE}} \in \arg \min_{\theta \in \Theta} \mathcal{K}(\bar{\mathbb{P}}_n, Q_\theta), \quad (2.1)$$

at least when the state space is finite, and I-projection is due to the fact that the Kullback Leibler divergence is not symmetric. In particular it is finite only if its first argument is absolutely continuous with respect to its second argument. In other words, maximum likelihood estimation tends to over-estimate the support of the data distribution, whereas I-projection under-estimate it. The difference in terms the support's estimation is very well illustrated in the Gaussian case in [Bis06], figure 10.2 and 10.3, chap 10. Besides, one can see that (2.1) is equivalent to the maximization of the expectation of a loss function

$$\hat{\theta}_{\text{MLE}} \in \arg \max_{\theta \in \Theta} \bar{\mathbb{P}}_n \left(\log \left(\frac{dQ_\theta}{d\nu} \right) \right),$$

where ν is some dominating measure ($Q_\theta \ll \nu$, for each $\theta \in \Theta$), whereas its theoretical counterpart is written as

$$\theta_{\text{MLE}}^* \in \arg \max_{\theta \in \Theta} \mathbb{P} \left(\log \left(\frac{dQ_\theta}{d\nu} \right) \right).$$

In the same manner, we will propose a loss function for the estimation of the classification parameter $\ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket$ equivalent to the minimization of the information k -means criterion. Indeed, from Lemma 1 on page 9 and Lemma 6 on page 41, one can remark that

$$\inf_{Q_{X,Y} : Q_{Y|X} = \mu_{\ell(X)}} \mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = -\log \left\{ \mathbb{P}_X \left[\exp \left(-\mathcal{K}(\mu_{\ell(X)}, \mathbb{P}_{Y|X}) \right) \right] \right\}, \quad \mu \in \mathfrak{M}_+^1(\mathcal{Y})^k.$$

It shows that the minimization of the information k -means criterion is related to the minimization of the expectation of a loss function $\gamma_{\mu,\ell}(X)$

$$(\mu^*, \ell^*) \in \arg \min_{\mu \in \mathfrak{M}_+^1(\mathcal{Y})^k, \ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X(\gamma_{\mu,\ell}(X)), \quad (2.2)$$

where $\gamma_{\mu,\ell}(X) = 1 - \exp\left(-\mathcal{K}(\mu_{\ell(X)}, \mathbb{P}_{Y|X})\right)$. This loss function is completely observed (we assume that $\mathbb{P}_{Y|X}$ is known) and plays the role of $\log\left(\frac{dQ_\theta}{d\nu}\right)$ in the maximum likelihood framework. Note that the loss function $\gamma_{\mu,\ell}(X)$ belongs to the unit interval, since the Kullback divergence is non negative. We will study the excess risk

$$\mathbb{P}_X(\gamma_{\hat{\mu}, \hat{\ell}}(X)) - \mathbb{P}_X(\gamma_{\mu^*, \ell^*}(X)),$$

where

$$(\hat{\mu}, \hat{\ell}) \in \arg \min_{\mu \in \mathfrak{M}_+^1(\mathcal{Y})^k, \ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket} \bar{\mathbb{P}}_X(\gamma_{\mu,\ell}(X)).$$

Furthermore, information projection appears in many machine learning algorithms, like for instance in variational Bayes methods (VB) for Bayesian estimation. Indeed, VB methods try to approximate a posterior distribution by I-projection onto a family of tractable distributions. These methods represent an alternative to slower MCMC approximations of the posterior distribution. We refer the reader to [Bis06], [BKM17], [AR20] and [ARC16] for more details on this topic. It turns out that VB methods appear also as an appealing tool for unsupervised clustering, especially to compute variational autoencoder (VAE), see [Doe16] for a complete review of the subject. It also emerged in graph variational autoencoders to perform clustering of nodes in a graph, see [Sal+19b] and [Sal+19a].

We should mention that clustering (conditional) probability distributions based on the Kullback divergence or other information criteria is not a new subject. It has been extensively used in text categorization, especially in word clustering to extract features or reduce the original space dimension. For instance, [PTL02] introduce what they call distributional clustering which consists in clustering nouns with respect to the conditional distribution of the associated verb given the noun. The grouping is performed by measuring the Kullback divergence between the conditional distribution knowing each noun and its associated centroid distribution. The centroid distribution is set to an intra-cluster average of conditional distributions that minimizes the average of the Kullback divergence.

However, in the information k -means framework, we follow a different route. We perform the grouping step by minimizing the Kullback divergence with respect to its first argument, which leads to very different centroids, computed as geometric means of conditional distributions. This is to the best of our knowledge a new addition to the literature.

The clustering of conditional distributions in [PTL02] is a particular version of a more general problem called *information bottleneck*, see [TPB01]. In the sequel, we will see that information k -means is a particular version of a more general clustering problem called information fragmentation.

Besides, in [Dhi+03], the authors propose a type of k -means algorithm that decreases a loss function based on the Jensen-Shannon entropy, written also as a loss of mutual information, leading to centers equal to weighted means of conditional distributions. In particular, they show that their entropy criterion can be expressed as a k -means type criterion using the Kullback divergence as the distortion. More formally, their criterion reduces to the following objective problem

$$\inf_{\ell: \mathbf{Q} \rightarrow \llbracket 1, k \rrbracket} \inf_{q_1, \dots, q_k} \sum_{j=1}^k \sum_{i \in \ell^{-1}(j)} \pi_i \mathfrak{K}(p_i, q_j),$$

where \mathbf{Q} is a set of discrete probability distributions and $\pi_i > 0$ represents some weights associated with distribution p_i . We can notice here that centroids are computed by minimizing the Kullback divergence with respect to the second argument, so that they are leading to compute centroids as

$$q_j^* = \sum_{i \in \ell^{-1}(j)} \frac{\pi_i p_i}{\sum_{i \in \ell^{-1}(j)} \pi_i},$$

with ℓ fixed and compute the best classification function as

$$\ell^*(i) = \arg \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(p_i, q_j^*).$$

In the same way, [Cao+13] and [Wu12] go further in this direction, studying what they call *Info K -means*. In particular, [Cao+13] proposes a new algorithm to deal with the practical issues of Info- K means, that arise from computing the Kullback divergence in high dimension. They apply this algorithm to cluster a sample of digital images presenting 11 different landmarks. They first preprocess images by extracting visual features and by quantizing those features, in order to consider each image as a bag of visual features. Then, they cluster images using Info- K means with $K = 11$ and obtain promising results by recovering a large portion of the original partition given by the types of landmarks. We should point out that when we will conduct our experiments, we will not use any kind of preprocessing step such as feature selection with quantization. We will perform directly the information fragmentation algorithm on the original digital images and this represents a strong point of our approach. Besides, we refer also to [Jia+11], who propose a k -medoid algorithm to decrease a k -means loss based on the Kullback divergence in both the discrete and continuous cases, and provide in addition an estimator of the Kullback divergence in a continuous setting.

Following the ideas of [Dhi+03], [BDG04] presents a general k -means framework based on the Bregman divergence. The authors show that such criteria can be minimized iteratively. The Bregman distance encompasses many traditional similarity measures such as the Euclidean distance, the Kullback divergence, the logistic loss and many others. However, in the Kullback case, the minimization is performed with respect to the second argument, and not the first as in our proposal.

Coming back to information k -means, we have seen that if we choose freely the distribution Q_X instead of setting it to \mathbb{P}_X , we get the bounded loss function of equation (2.2) on page 19. In the quadratic case, we get the criterion

$$\inf_{\ell} \mathbb{P}_X \left[1 - \exp \left(-\frac{1}{2\sigma^2} \|X - c_{\ell(X)}\|^2 \right) \right] = \mathbb{P}_X \left[1 - \exp \left(-\frac{1}{2\sigma^2} \min_{j \in \llbracket 1, k \rrbracket} \|X - c_j\|^2 \right) \right].$$

We will see that we can state a generalization bound for this kind of robust criterion under weaker hypotheses than for the original criterion.

So far, we have described the extension of k -means to information k -means. The next extension we would like to discuss is from k -means to fragmentation. In the k -means setting, we use a center in \mathbb{R}^d to represent nearby points. This is rather crude from a classification point of view. We do not really expect whole images (or more generally whole signals) to correspond to a relevant class. We would rather like to label *parts* of images. This may be done by labeling pixels in images with different labels, producing a *fragmentation* of each image into various areas. In the usual quadratic k -means setting, we propose to approximate X by

$$Y = \sum_{j \in A_X} c_j, \text{ where } \llbracket 1, d \rrbracket = \bigsqcup_{j \in A_X} \text{supp}(c_j).$$

Here A_X replaces $\ell(X)$ and contains the labels of the components of the signal $X \in \mathbb{R}^d$. This framework can be seen as a generalization of the k -means setting that corresponds to $A_X = \{\ell(X)\}$. The quadratic criterion becomes

$$\mathbb{P}_X \left(\left\| X - \sum_{j \in A_X} c_j \right\|^2 \right).$$

Introducing π_j , the orthogonal projection on the vector space spanned by the support of c_j , it can also be written as

$$\mathbb{P}_X \left(\sum_{j \in A_X} \left\| \pi_j(X) - c_j \right\|^2 \right) = \sum_{j=1}^k \mathbb{P}_X \left(\mathbb{1}(X \in A_j) \left\| \pi_j(X) - c_j \right\|^2 \right),$$

where $A_j = \{x \in \mathbb{R}^d : j \in A_x\}$.

To give an information projection description of fragmentation, we have to introduce the pixel location S as a random variable. More precisely, we replace the representation Y of X with two random variables $S \in \llbracket 1, d \rrbracket$ and $V \in \mathbb{R}$, defined by

$$\mathbb{P}_{S|X} = \frac{1}{d} \sum_{j=1}^d \delta_j \text{ and } \mathbb{P}_{V|X, S=j} = \mathcal{N}(X_j, \sigma^2).$$

This being put, we can describe the quadratic k -means criterion as

$$\inf_Q \mathbb{P}_{X,S} \left[\mathcal{K}(Q_{V|S, \ell(X)}, \mathbb{P}_{V|S, X}) \right],$$

while the global entropy criterion is

$$\begin{aligned} & \inf_{Q: Q_{S,V|X} = \mathbb{P}_S Q_{V|S, \ell(X)}} \mathcal{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}) \\ &= -\log \sup_{Q: Q_{S,V|X} = \mathbb{P}_S Q_{V|S, \ell(X)}} \left\{ \mathbb{P}_X \left[\exp \left(-\mathcal{K}(Q_{S,V|\ell(X)}, \mathbb{P}_{S,V|X}) \right) \right] \right\}. \end{aligned}$$

The modification to be made to obtain the quadratic fragmentation criterion is just to make the classification function ℓ depend on S , the pixel location. We get

$$\inf_Q \mathbb{P}_{X,S} \left[\mathcal{K}(Q_{V|S, \ell(X,S)}, \mathbb{P}_{V|X,S}) \right].$$

The corresponding global entropy criterion is

$$\inf_{Q: Q_{S,V|X, \ell(X,S)} = Q_{S,V|\ell(X,S)}} \mathcal{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}).$$

It has the following interesting properties.

PROPOSITION 3 (GLOBAL FRAGMENTATION CRITERION) *Consider k centers $\rho_j \in \mathfrak{M}_+^1(\mathbb{R}^d)$, $1 \leq j \leq k$. Define*

$$\mathfrak{T}_2 = \left\{ B \subset \llbracket 1, k \rrbracket : \rho_i \perp \rho_j, i \neq j \in B \right\},$$

the set of (possibly partial) tilings by mutually singular probability measures ρ_j . The partial minimum

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V})$$

taken on all probability measures Q , such that for some measurable function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$

$$Q \left[Q_{X|S,V,\ell(X,S)} = \rho_{\ell(X,S)} \right] = 1. \quad (2.3)$$

is equal to

$$\begin{aligned} & - \sup_{\ell} \log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1} \left[\rho_j(\ell_S^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right) \\ & = - \log \mathbb{P}_S \left(\sup_{B \in \mathfrak{T}_2} \sum_{j \in B} \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right), \end{aligned}$$

where

$$\begin{aligned} \ell_s : \mathbb{R}^d & \rightarrow \llbracket 1, k \rrbracket \\ x & \mapsto \ell(x, s). \end{aligned}$$

For any choice of ℓ , and in particular for the optimal one, putting $W = \ell(X, S)$, the optimum in $Q_{W,S,V}$ is reached when

$$\frac{dQ_{W,S,V}}{d\mathbb{P}_{W,S} \otimes \lambda_V} = Z^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right] g_{\sigma, \rho_W(X_S)}(V), \quad (2.4)$$

where λ_V is the Lebesgue measure on \mathbb{R} and

$$g_{\sigma, m}(v) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{(v - m)^2}{2\sigma^2} \right).$$

In particular, for the optimal choice of $Q_{W,S,V}$, $Q_{V|S,W} = \mathcal{N}(\rho_W(X_S), \sigma^2)$ is a Gaussian probability measure and

$$\frac{dQ_{S|W}}{d\mathbb{P}_{S|W}} = Z_W^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right].$$

On the other hand, consider k centers $\mu_{S,V}^{(j)} \in \mathfrak{M}_+^1(\llbracket 1, d \rrbracket \times \mathbb{R})$, $1 \leq j \leq k$ such that

$$\mu_{V|S}^{(j)} = \mathcal{N}(\mu_{V|S}^{(j)}(V), \sigma^2), \quad 1 \leq j \leq k.$$

Define

$$\mathfrak{T}_1 = \left\{ A \subset \llbracket 1, k \rrbracket : \mu_S^{(i)} \perp \mu_S^{(j)}, i \neq j \in A \right\},$$

the set of tilings by mutually singular probability measures $\mu_S^{(j)}$ (or equivalently by mutually singular probability measures $\mu_{S,V}^{(j)}$).

The partial minimum

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V})$$

taken on all probability measures $Q \in \mathfrak{M}_+^1(\Omega)$ such that, for some measurable function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$

$$Q \left[Q_{S,V|X, \ell(X,S)} = \mu_{S,V}^{(\ell(X,S))} \right] = 1, \quad (2.5)$$

is equal to

$$\begin{aligned} & - \sup_{\ell} \log \mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_S^{(j)} \left(\ell_X^{-1}(j) \right) = 1 \right] \right. \\ & \quad \times \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \Bigg) \\ & = -\log \mathbb{P}_X \left(\sup_{A \in \mathcal{I}_1} \sum_{j \in A} \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \right), \end{aligned}$$

where

$$\begin{aligned} \ell_x & : \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket \\ s & \mapsto \ell(x, s). \end{aligned}$$

For any value of ℓ , and in particular for the optimal one, considering $W = \ell(X, S)$, the minimum in $Q_{X,W}$ is reached when

$$\frac{dQ_{X,W}}{d\mathbb{P}_{X,W}} = Z^{-1} \exp \left\{ -\mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) - \mu_S^{(W)} \left[\left(\mu_{V|S}^{(W)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\}. \quad (2.6)$$

Alternating these two partial optimization steps, we can converge to a local minimum for the optimization problem

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}),$$

where the infimum is taken over probability measures $Q \in \mathfrak{M}_+^1(\Omega)$ satisfying, for some measurable classification function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$,

$$Q \left[Q_{X,S,V|\ell(X,S)} = Q_{X|\ell(X,S)} \otimes Q_{S,V|\ell(X,S)} \right] = 1. \quad (2.7)$$

The proof will be given later, see Proposition 16 on page 56. The second part of the proposition shows that the entropy criterion can be viewed as an expectation with respect to \mathbb{P}_X . This expectation can be estimated by an expectation with respect to the empirical measure $\bar{\mathbb{P}}_X$. The last part of the proposition describes the pendent of Lloyd's algorithm (in the case where we replace the unknown \mathbb{P}_X by the empirical measure $\bar{\mathbb{P}}_X$).

Now that we have a criterion for fragmentation, we need an algorithm to compute a fragmentation based on this criterion.

We will use the criterion

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}),$$

to define a distortion function. Let X_1, \dots, X_n be an i.i.d. training set. We can represent its content by the distribution

$$\begin{aligned}\bar{\mathbb{P}}_{I,S,V} &= \left(\frac{1}{n} \sum_{i=1}^n \delta_i \right) \mathbb{P}_{S,V|X=X_I} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \delta_i \right) \left(\frac{1}{d} \sum_{j=1}^d \delta_j \right) \mathcal{N}(X_{I,S}, \sigma^2) \\ &= \frac{1}{nd} \sum_{i=1}^n \sum_{j=1}^d \delta_{i,j} \mathcal{N}(X_{i,j}, \sigma^2).\end{aligned}$$

Here I is a random index ranging in $\llbracket 1, n \rrbracket$. We see immediately that (X_1, \dots, X_n) is a function of $\bar{\mathbb{P}}$, since

$$X_{i,s} = \bar{\mathbb{P}}_{V|S=s, I=i}(V), \quad i \in \llbracket 1, n \rrbracket, \quad s \in \llbracket 1, d \rrbracket.$$

Consider a finite codebook $\mathfrak{C} \subset \mathbb{R}$, for instance $\mathfrak{C} = \{m2^{-8} : m \in \llbracket 0, 255 \rrbracket\}$ if we are to code light intensities ranging in the unit interval $[0, 1]$ on eight bits as is usually the case. For any classification function

$$\ell : \llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \longrightarrow \llbracket 1, k \rrbracket$$

defined by

$$\ell^{-1}(j) = A_j \times B_j, \quad 1 \leq j \leq k,$$

where $(A_j \times B_j, 1 \leq j \leq k)$ is a partition of $\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket$ and any family of centers $(C_j, 1 \leq j \leq k) \in \mathfrak{C}^{d \times k}$, where $\text{supp}(C_j) \subset B_j$, define a parameter

$$\theta = (A_j, B_j, C_j)_{j=1}^k$$

and the corresponding model

$$\begin{aligned}\mathfrak{Q}_\theta &= \left\{ Q_{I,S,V} \in \mathfrak{M}_+^1(\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \times \mathbb{R}) \right. \\ &\quad \left. : Q_{S,V|I, (I,S) \in A_j \times B_j} = \mathbb{P}_{S|S \in B_j} \mathcal{N}(C_{j,S}, \sigma^2), j \in \llbracket 1, k \rrbracket \right\},\end{aligned}$$

where we recall that $\mathbb{P}_S = \frac{1}{d} \sum_{s=1}^d \delta_s$ is known and is the uniform measure on the pixel locations. To make a connection between \mathfrak{Q}_θ and the model (2.5) on page 24 defined in Proposition 3 on page 22, one can see that \mathfrak{Q}_θ is equivalent to impose

$$\mu_S^{(j)} = \mathbb{P}_{S|S \in B_j} \text{ and } \mu_{V|S}^{(j)} = \mathcal{N}(C_{j,S}, \sigma^2).$$

In other words, $\mathfrak{Q}_\theta \subset \mathfrak{Q}_\ell$, where

$$\mathfrak{Q}_\ell = \left\{ Q_{I,S,V} \in \mathfrak{M}_+^1(\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \times \mathbb{R}) : Q_{I,S,V|\ell(I,S)} = Q_{I|\ell(I,S)} \otimes Q_{S,V|\ell(I,S)} \right\}.$$

We define the distortion $D(\theta)$ of the representation of the sample (X_1, \dots, X_n) by the parameter θ as

$$\begin{aligned}
D(\theta) &= \inf_{Q \in \mathfrak{Q}_\theta} \left\{ \mathfrak{K}(Q_{I,S,V}, \bar{P}_{I,S,V}) \right\} \\
&= -\log \bar{P}_I \left(\sum_{j=1}^k \mathbb{1}(I \in A_j) \mathbb{P}_S(B_j) \exp \left\{ -\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} \left[(X_{I,S} - Y_{I,S})^2 \right] \right\} \right),
\end{aligned}$$

according to Proposition 3 on page 22.

Note that it makes sense to optimize in $Q \in \mathfrak{Q}_\theta$, since the quantization of X_i , given by

$$Y_{i,s} = Q_{V|S=s, I=i}(V) = \sum_{j=1}^k \mathbb{1}(i \in A_j) C_{j,s}, \quad 1 \leq i \leq n, \quad 1 \leq s \leq d,$$

does not depend on $Q \in \mathfrak{Q}_\theta$, but only on θ . In fact, it does not depend on $Q_{I,S}$ so that, as a variant, we could have optimized even more in the definition of $D(\theta)$.

Note that this notion of distortion satisfies

$$\inf_{Q \in \mathfrak{Q}_\ell} \mathfrak{K}(Q_{I,S,V}, \bar{P}_{I,S,V}) \leq D(\theta) \leq \bar{P}_{I,S} \left[(X_{I,S} - Y_{I,S})^2 \right] = \frac{1}{nd} \|X_1^n - Y_1^n\|^2.$$

Given a coding distribution $q(\theta)$ and an acceptable distortion level $\eta \geq 0$, the fragmentation algorithm will compute a lossy representation $\hat{\theta}(X_1, \dots, X_n)$ with distortion $D(\hat{\theta}) \leq \eta$ and with an ideal code length $-\log(q(\hat{\theta}))$ as small as possible.

We will use a coding distribution of the form

$$q(\theta) = q(A, B, C) = q(A) q(B, C).$$

After this fragmentation step leading to the computation of $\hat{\theta}$, we will perform a syntax analysis step where we will replace the ideal code $q(A)$ by a more efficient code $\tilde{q}(A)$. This improvement will be obtained using the Bayesian Shtarkov approach. More precisely we will consider a family $q_\alpha(A)$ of coding distributions depending on a new parameter α and a prior coding distribution $\mu(\alpha)$, and we will improve on $q(A)$ by considering

$$\tilde{q}(A) = \max_{\alpha} \mu(\alpha) q_\alpha(A).$$

Since

$$\bar{q}(A) = \sum_{\alpha} \mu(\alpha) q_\alpha(A)$$

is a probability measure, \tilde{q} is a subprobability measure and thus a valid coding distribution. From a Bayesian point of view, $\tilde{q}(A)$ can also be seen as the maximum a posteriori probability estimate (MAP). Introducing

$$\hat{\alpha} \in \arg \max_{\alpha} \mu(\alpha) q_\alpha(\hat{A}),$$

we see that $\tilde{q}(\hat{A}) = \mu(\hat{\alpha}) q_{\hat{\alpha}}(\hat{A})$ and that $\hat{\alpha}$ is a function of the sample (X_1, \dots, X_n) , since this is the case for \hat{A} . When we will detail the construction of $\hat{\alpha}$ we will see that it performs some kind of syntax analysis and in particular leads to the computation of a syntax tree for each image X_i , $1 \leq i \leq n$ of the sample.

2.2. GENERALIZATION BOUNDS FOR FRAGMENTATION

The fragmentation algorithm computes a classification function

$$\ell_k : \llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \longrightarrow \llbracket 1, k \rrbracket.$$

We can deduce from it

$$\begin{aligned} \bar{\ell}_k : \{X_1, \dots, X_n\} \times \llbracket 1, d \rrbracket &\longrightarrow \llbracket 1, k \rrbracket \\ (X_i, s) &\longmapsto \bar{\ell}(X_i, s) = \ell_k(i, s). \end{aligned}$$

A natural question is to extend $\bar{\ell}(x, s)$ to $x \notin \{X_1, \dots, X_n\}$ in some meaningful way. To this purpose, introduce the set of fragments used to represent X_i

$$\bar{A}_i = \{j \in \llbracket 1, k \rrbracket : i \in A_j\}.$$

Remarking that $\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) = -\log(\mathbb{P}_S(B_j))$ and noticing that $\mu_{V|S}^{(j)}(V) = C_{j,S}$, one gets from Proposition 3 on page 22

$$\begin{aligned} D(A, B, C) &= -\log \bar{\mathbb{P}}_I \left(\sum_{j \in \bar{A}_I} \mathbb{P}_S(B_j) \exp \left[-\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} [(C_{j,S} - X_{I,S})^2] \right] \right) \\ &\geq -\log \bar{\mathbb{P}}_X \left(\exp [-D(X, B, C)] \right), \end{aligned}$$

where $(A, B, C) = ((A_j)_{j=1}^k, (B_j)_{j=1}^k, (C_j)_{j=1}^k)$ for short and where

$$D(X, B, C) = -\log \max_{\bar{A} \in \mathfrak{T}} \left(\sum_{j \in \bar{A}} \mathbb{P}_S(B_j) \exp \left\{ -\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} [(X_S - C_{j,S})^2] \right\} \right)$$

and

$$\mathfrak{T} = \{\bar{A} \subset \llbracket 1, k \rrbracket : B_i \cap B_j = \emptyset, i \neq j \in \bar{A}\}.$$

We can see $D(X, B, C)$ as the optimal distortion for a single image when it is represented by its best approximation in the codebook (B, C) of image fragments. An optimal set of fragments \bar{A}_X for X is given by the formula

$$\bar{A}_X \in \arg \max_{\bar{A} \in \mathfrak{T}} \left(\sum_{j \in \bar{A}} \mathbb{P}_S(B_j) \exp \left\{ -\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} [(X_S - C_{j,S})^2] \right\} \right). \quad (2.8)$$

It is well defined even when $X \notin \{X_1, \dots, X_n\}$.

We can then define the optimal empirical distortion of the codebook (B, C) as

$$D(B, C) = -\log \bar{\mathbb{P}}_X \left(\exp [-D(X, B, C)] \right) \leq D(A, B, C)$$

and ask for its relationship with its expected counterpart

$$\mathbf{D}(B, C) = -\log \mathbb{P}_X \left(\exp [-D(X, B, C)] \right).$$

For this we need deviation inequalities for

$$\overline{\mathbb{P}}_X \left(\exp[-D(X, B, C)] \right)$$

that are uniform with respect to the parameter (B, C) (here the parameter is the fragments codebook).

Note that the optimal expected distortion $\mathbf{D}(B, C)$ is related to the estimation of the distribution $\mathbb{P}_{X, S, V}$. Indeed, consider the model

$$\begin{aligned} \mathfrak{Q}_{B, C} &= \left\{ Q_{X, S, V} \in \mathfrak{M}_+^1(\mathbb{R}^d \times \llbracket 1, d \rrbracket \times \mathbb{R}) : \right. \\ &\quad \left. Q_{S, V | X} = \sum_{j \in \overline{A}_X} Q_{S | X}(B_j) \mathbb{P}_{S | S \in B_j} \mathcal{N}(C_{j, S}, \sigma^2) \right\} \\ &= \left\{ Q_{X, S, V} \in \mathfrak{M}_+^1(\mathbb{R}^d \times \llbracket 1, d \rrbracket \times \mathbb{R}) : \right. \\ &\quad \left. Q_{S, V | X, \bar{\ell}(X, S)=j} = \mathbb{P}_{S | S \in B_j} \mathcal{N}(C_{j, S}, \sigma^2) \right\}, \end{aligned}$$

where \overline{A}_X , the optimal set of fragments, is defined by equation (2.8) on page 27 and where, putting $A_j = \{x \in \mathbb{R}^d : j \in \overline{A}_x\}$, $\bar{\ell}$ is defined by the formula

$$\bar{\ell}^{-1}(j) = A_j \times B_j, \quad 1 \leq j \leq k.$$

Remark that

$$\mathbf{D}(B, C) = \inf_{Q \in \mathfrak{Q}_{B, C}} \mathfrak{K}(Q_{X, S, V}, \mathbb{P}_{X, S, V}).$$

Bounding $\mathbf{D}(B, C)$ in terms of $D(B, C)$ thus appears as a generalization bound for some information k -means algorithm.

We will also derive similar bounds for the more classical k -means algorithms described above.

Our proofs will be based on PAC-Bayesian lemmas. We will first rewrite the risk (of information k -means or fragmentation) using a mapping to a reproducing kernel Hilbert space. This will allow to see the risk as the expectation of the minimum of linear functions of the parameter, in a separable Hilbert space of possibly infinite dimension. We will then establish dimension free PAC-Bayesian bounds suitable to this situation. Borrowing ideas from the construction of the isonormal Gaussian process [MPS07, section 3.5], we will use the distribution of an infinite sequence of shifted Gaussian random variables both for the prior and the posterior parameter distribution. We will also use arguments from the proofs of [CG18] and [CG17], concerning the estimation of the mean of a random vector. We will prove generalization bounds going to zero as $k \log(k)/n$ goes to zero, which is better than other bounds already published. Concerning the speed of decrease of the bounds, we first prove bounds in $(k \log(k)/n)^{1/4}$ and then, introducing a more sophisticated chaining argument, bounds in $\log(n/k) \sqrt{k \log(k)/n}$. We will work with weak hypotheses and will in particular not consider the kind of margin assumptions that are necessary to get bounds

decreasing faster than $\sqrt{1/n}$ for a given value of k . See [BDL08], [Fis10], [Lev13],[Lev15] and [BFL20].

To get a $\sqrt{1/n}$ speed as in [BDL08], we take inspiration from the classical chaining procedure for bounding the expected suprema of sub-Gaussian processes (see section 13.1 in [BLM13]). We create a PAC-Bayesian version of chaining in which the concept of δ -net and δ -covering is replaced by the use of a sequence of Gaussian perturbations parametrized by a variance ranging on a logarithmic grid. We combine this PAC-Bayesian chaining with the use of the influence function ψ described in [Cat12] to decompose the excess risk into a sub-Gaussian part and an other part representing extreme values. Doing so, we will recover the speed $1/\sqrt{n}$ proved in [BDL08], but with a better dependence in k , since we get a non asymptotic bound of order $\log(n/k)\sqrt{k \log(k)/n}$ instead of k/\sqrt{n} .

2.3. DESCRIPTION OF THE SIGNAL FRAGMENTATION ALGORITHM

To start with, we will devote to each image $X_i \in \mathbb{R}^d$ of an i.i.d. training set X_i , $1 \leq i \leq n$ a single fragment equal to the whole X_i .

This means that we will start with $k = n$ and a classification function ℓ_n defined by

$$\ell_n^{-1}(j) = A_{n,j} \times B_{n,j} = \{j\} \times \llbracket 1, d \rrbracket.$$

From there, we will iteratively define ℓ_k for larger values of k setting for some pair $J_k \subset \llbracket 1, k \rrbracket$ of labels

$$\begin{aligned} A_{k+1,k+1} &= \bigsqcup_{j \in J_k} A_{k,j}, & A_{k+1,j} &= A_{k,j}, \quad j \in \llbracket 1, k \rrbracket, \\ B_{k+1,k+1} &\subset \bigcap_{j \in J_k} B_{k,j}, & B_{k+1,j} &= \begin{cases} B_{k,j} \setminus B_{k+1,k+1}, & j \in J_k, \\ B_{k,j}, & j \in \llbracket 1, k \rrbracket \setminus J_k. \end{cases} \end{aligned}$$

These equations define ℓ_{k+1} from ℓ_k . We will explain later on how to choose J_k and $B_{k+1,k+1}$. In all cases, however, we can readily remark that for any $k \geq n$,

$$\{A_{k,j} \times B_{k,j} : 1 \leq j \leq k, B_{k,j} \neq \emptyset\}$$

is a partition of $\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket$. Note that this partition may contain less than k components (and even less than n components) due to the fact that $B_{k,j}$ is decreasing with $k \geq j$ for a fixed value of j and may become empty. Consider

$$Q_{I,S,V}^{(k)} = \overline{\mathbb{P}}_{I,S} \mathcal{N}(C_{k,\ell_k(I,S),S}, \sigma^2).$$

where

$$C_{k,j,s} = \arg \min \left\{ |c - \overline{\mathbb{P}}_{I|I \in A_{k,j}}(X_{I,s})| : c \in \mathfrak{C} \right\}, \quad s \in B_j,$$

and choose

$$B_{k+1,k+1} = \left\{ s \in \bigcap_{j \in J_k} B_{k,j} : \mathbf{Var}(X_{I,s} | \bar{\mathbb{P}}_{I|I \in A_{k+1,k+1}}) + \min_{c \in \mathfrak{C}} \left(\bar{\mathbb{P}}_{I|I \in A_{k+1,k+1}}(X_{I,s}) - c \right)^2 \leq \alpha \right\}. \quad (2.9)$$

Doing so, we are sure at each step that

$$D(A_k, B_k, C_k) \leq \mathfrak{K}(Q_{I,S,V}^{(k)}, \bar{\mathbb{P}}_{I,S,V}) \leq \frac{\alpha}{2\sigma^2}.$$

Indeed, from the decomposition of the Kullback divergence (see Lemma 1 on page 9) and the law of iterated expectations, we see that

$$\begin{aligned} \mathfrak{K}(Q_{I,S,V}^{(k)}, \bar{\mathbb{P}}_{I,S,V}) &= \frac{1}{2\sigma^2} \bar{\mathbb{P}}_{I,S} [(C_{k,\ell_k(I,S),S} - X_S)^2] \\ &= \frac{1}{2\sigma^2} \sum_{j=1}^k \bar{\mathbb{P}}_{I,S}(A_{k,j} \times B_{k,j}) \bar{\mathbb{P}}_{S|S \in B_{k,j}} \left\{ \bar{\mathbb{P}}_{I|I \in A_{k,j}} [(C_{k,j,S} - X_{I,S})^2] \right\} \\ &= \frac{1}{2\sigma^2} \sum_{j=1}^k \bar{\mathbb{P}}_{I,S}(A_{k,j} \times B_{k,j}) \bar{\mathbb{P}}_{S|S \in B_{k,j}} \left[\mathbf{Var}(X_{I,S} | \bar{\mathbb{P}}_{I|I \in A_{k,j}}) + (\bar{\mathbb{P}}_{I|I \in A_{k,j}}(X_{I,S}) - C_{k,j,S})^2 \right]. \end{aligned}$$

We have now to discuss the choice of $J_k = \{i_k, j_k\}$. Assume that $\max\{n, d, |\mathfrak{C}|\} < 2^L$, so that all coordinates in (A_k, B_k, C_k) can be coded with L bits. In this case (which is an obvious case of the Kraft inequality),

$$q(A_k, B_k, C_k) = 2^{-L(|A_k| + 2|B_k| + 3|C_k|)},$$

where $|A_k| = \sum_{j=1}^k |A_{k,j}|$, and $|B_k| = \sum_{j=1}^k |B_{k,j}|$, is a sub-probability measure (the factor $3k$ comes from the use of a separator, for instance the index 0 that is not used otherwise, at the end of the enumerations of the sets $A_{k,j}$, $B_{k,j}$ and $C_{k,j}$). Remark now that the code length decrease is

$$\log_2(q(A_{k+1}, B_{k+1}, C_{k+1})) - \log_2(q(A_k, B_k, C_k)) = 2L|B_{k+1,k+1}| - L|A_{k+1,k+1}| - 3L.$$

Maximizing the code length decrease would lead to choose

$$J_k = \{j_{k,1}, j_{k,2}\} \in \arg \max_{J_k} (2|B_{k+1,k+1}| - |A_{k+1,k+1}|).$$

This requires to compute $A_{k+1,k+1}$ and $B_{k+1,k+1}$ for all possible choices of the pair J_k . We have tested a faster approximation to this minimization, consisting in choosing

$$j_{k,1} \in \arg \max_{j \in \llbracket 1, k \rrbracket} (2|B_{k,j}| - |A_{k,j}|)$$

and then

$$j_{k,2} \in \arg \max_{j_{k,2} \in \llbracket 1, k \rrbracket \setminus \{j_{k,1}\}} (2|B_{k+1,k+1}| - |A_{k+1,k+1}|).$$

This requires to compute $A_{k+1,k+1}$ and $B_{k+1,k+1}$ only for $k-1$ possible values of J_k , namely for

$$J_k \in \left\{ \{j_{k,1}, j\} : j \in \llbracket 1, k \rrbracket \setminus \{j_{k,1}\} \right\},$$

instead of $k(k-1)/2$ possible values if we opt for a full optimization. This heuristic simplification is based on the fact that

$$2|B_{k+1,k+1}| - |A_{k+1,k+1}| \leq 2|B_{k,j_{k,1}}| - |A_{k,j_{k,1}}|,$$

so that the left-hand side cannot be big if the right-hand side is already small, justifying the idea of maximizing the right-hand side to choose $j_{k,1}$ and then the left-hand side to choose $j_{k,2}$.

A natural stopping rule in this framework is to continue as long as we can decrease the code length, that is as long as

$$2|B_{k+1,k+1}| - |A_{k+1,k+1}| > 3.$$

2.4. SYNTAX ANALYSIS

Now that we have a mean to represent an i.i.d. sample of images (X_1, \dots, X_n) by sets of fragments $(\bar{A}_1, \dots, \bar{A}_n)$, drawn from a fragment codebook (B, C) , we will carry further the change of representation by defining a fragment grammar.

Our approach will still be based on compression theory as we have mentioned in the introduction.

The algorithm we will now describe consists in a sequence of lossless codes for the sample $(\bar{A}_1, \dots, \bar{A}_n)$ made of n random sets of labels ranging in $\llbracket 1, k \rrbracket$. As in the previous section, our guide will be to decrease the code length at each step. This can be coded to start with by listing the content of each set, with separators, resulting in the sequence

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge,$$

where \wedge is a supplementary label used as a separator and where the labels $w_{i,j} \in \llbracket 1, k \rrbracket$ and \wedge are coded with a prefix binary code (so that they can for instance either be coded with a fixed length code or with a prefix binary code for the integers). The syntax of this initial representation is

$$\{\{w\}\wedge\}$$

where we have used $\{\}$ to denote repetition, as in Extended Backus Naur specifications and where $w \leq k$ are initial labels. From there we will put $\bar{A}_i = A_{0,i}$ and choosing a pair of labels $J_1 \subset \llbracket 1, k \rrbracket$ and a new label numbered $k+1$, we will define the new sequence of sets

$$A_{1,i} = \begin{cases} (A_{0,i} \setminus J_1) \cup \{k+1\}, & \text{when } J_1 \subset A_{0,i}, \\ A_{0,i}, & \text{otherwise.} \end{cases}$$

To recover $A_{0,i}$, $1 \leq i \leq n$ from $A_{1,i}$, $1 \leq i \leq n$, we need to code also the value of $J_1 = \{a, b\}$. We obtain in this way a new code. The syntax for this new code is of the form

$$\{\{w|p\}\wedge\}pab$$

where $p = k + 1 > k$ is a pair label and where $|$ means *or* as in Extended Backus Naur specifications. In order to shrink as much as possible the length of the representation, we have to choose J_1 to maximize

$$\sum_{i=1}^n \mathbb{1}(J_1 \subset A_{0,i}),$$

the number of times it appears in the random sets $A_{0,i}$, $1 \leq i \leq n$ (at least if we code labels with a fixed length code, as we will assume for simplicity in all this discussion). We can repeat this process choosing at each step a most frequent pair, to form new sequences of random sets

$$A_{0,i}, \dots, A_{m,i}, 1 \leq i \leq n.$$

We obtain a new shorter code at each time by concatenating the code for $A_{m,i}$, $1 \leq i \leq n$ and the list of rewriting rules defining the pairs. The syntax of such a code is thus of the form

$$\{\{w|p\} \wedge\} \{pab\}$$

(We do not need separators in the description of the pairs, since we know they are always triplets.) Note that $w \in \llbracket 1, k \rrbracket$, $p \in \llbracket k + 1, k + m \rrbracket$ and that $a, b \in \llbracket 1, k + m \rrbracket$ may be so to speak either terminal or non terminal indices (or symbols). Note also that we can forget to represent p explicitly in the $\{pab\}$ section of the representation, since we can assume that

$$\{pab\} = p_1 a_1 b_1, \dots, p_m a_m b_m,$$

where $p_i = k + i$. Doing this, we obtain a code of the type

$$\{\{w|p\} \wedge\} \{ab\}$$

Remark that, as long as the chosen pair appears at least three times, we decrease the length of the representation at each step. This can be taken as a stopping rule. We can also use a higher threshold to stop earlier (given that the frequency of the most frequent pair is decreasing).

From an image perspective, this compression step consists in merging two fragments a and b that are frequently seen together in the image sample $(X_1 \dots, X_n)$. Inside a fixed pair, elements can be viewed as a context of each other. It is rather a "main" context as it is in a way the most frequent one. Therefore, this merging operation is twofold : it simplifies the representation of the set $A_{m,i}$ and determines in the meantime the context of each fragment. Notice that this procedure is very similar to the fragmentation step and can be seen as its dual version in the sense that we exchange the role played by the sets A and B . Indeed, in contrast with fragmentation, here we make $A_{m,i}$ smaller and create a bigger fragment B_p .

Prepare now to push the compression of the representation further by performing permutations on the list of pairs

$$a_1 b_1 \dots a_m b_m$$

Put $a_i = c_{i,1}$ and $b_i = c_{i,2}$. We represent the pairs by the permutation

$$f(c_{\sigma(1),\psi(1,1)})f(c_{\sigma(1),\psi(1,2)}) \cdots f(c_{\sigma(m),\psi(m,1)})f(c_{\sigma(m),\psi(m,2)}), \quad (2.10)$$

where the transformation

$$f(c) = \begin{cases} c, & \text{when } c \in \llbracket 1, k \rrbracket, \\ k + \sigma^{-1}(c - k), & \text{when } c \in \llbracket k + 1, k + m \rrbracket, \end{cases}$$

of the pair indices should also be applied to the first part of the representation, that is to $\{\{w|p\}\wedge\}$, where p becomes $f(p)$. Introduce the counter

$$\xi(c) = \sum_{(i,j) \in \llbracket 1, m \rrbracket \times \{1,2\}} \mathbb{1}(c_{i,j} = c)$$

choose

$$c_* \in \arg \max_{c \in \llbracket 1, k+m \rrbracket} \xi(c),$$

and choose

$$\sigma(1), \psi(1,1), \psi(1,2), \dots, \sigma(\xi(c_*)), \psi(\xi(c_*),1), \psi(\xi(c_*),2)$$

in such a way that

$$c_{\sigma(i),\psi(i,1)} = c_*, \quad 1 \leq i \leq \xi(c_*).$$

Repeat this process on the remaining pairs as many times as needed, until we have chosen the whole permutation $\sigma(i), \psi(i,j)$. After applying transformation (2.10) to the pairs $a_i b_i$, we see that each possible value of a_i comes in succession. Thus we can factorize (2.10) in the form

$$a_1 b_{1,1} \dots b_{1,q_1} \wedge a_2 b_{2,1} \dots b_{2,q_2} \wedge \cdots \wedge a_{m'} b_{m',1} \dots b_{m',q_{m'}} \wedge,$$

where the notation is changed and we need some separators but where $m' \leq m$. (With the new notation $a_1, \dots, a_{m'}$ is a sequence without repetition, whereas in the previous representation (a_1, \dots, a_m) possibly contained repeated values.) Note that the permutation in (2.10) was chosen to maximize sequentially $q_1, q_2, \dots, q_{m'}$. We can then put the sequence $a_1 \dots a_{m'}$ in front to get

$$a_1 \dots a_{m'} \wedge b_{1,1} \dots b_{1,q_1} \wedge b_{2,1} \dots b_{2,q_2} \wedge \cdots \wedge b_{m',1} \dots b_{m',q_{m'}} \wedge$$

We don't lose information here, since we can recover the previous representation by redistributing each a_i in front of the sequence of contexts $b_{i,1} \dots b_{i,q_i}$. The new syntax for the pairs description is the grammar

$$G ::= \{a\} \wedge \{\{b\}\wedge\}$$

instead of $\{ab\}$, so that the syntax for the new sample code is

$$\{\{w|p\}\wedge\}\{a\} \wedge \{\{b\}\wedge\} \quad (2.11)$$

instead of

$$\{\{w\}\wedge\} \quad (2.12)$$

(We know in the new code that the description of the random sets ends before separator \wedge number n and is followed by the description of the pairs.) Let us note that G is represented as parameter α in section (2) and that $\mu(\alpha) = 2^{-|G|}$, where $|G|$ represents the number of bits used to represent G . On the other hand $q_\alpha(\bar{A}) = 2^{-|D|}$, where $D ::= \{\{w|p\}\wedge\}$ is the description of \bar{A} . We now remark that the end $\{\{b\}\wedge\}$ of the new code (2.11) has the same syntax as the initial code (2.12). We can thus compress it in the same way by indexing iteratively the most frequent syntax pair $J_s = \{c, d\}$ as long as it appears at least three times. Here the frequency is computed within the grammar G , or in other words the frequency is the number of different contexts a_i in which fragment c appears along with d within the list $b_{i,1} \dots b_{i,q_i}$.

This analysis is different from what is usually done in statistics, where frequencies are computed within the data set (X_1, \dots, X_n) . Using syntax analysis, we produce a second level of classification based on a kind of context analysis. This step is necessary if we want to put a common (syntax) label on fragments that appear in different images at different locations. To give an example, let us assume that we have some cats (not necessarily the same) appearing in different images, not in the same place either, but lying on a sofa or on a bed. In that case, we can say that we have a fixed context *the sofa* or *the bed* and that the cats appear frequently in those two contexts. Then, our syntax analysis will tend to put a common syntax label *cat* (or more accurately *what appears on beds and sofas*) on those lazy cats sharing the same context.

As a matter of fact, our approach is very close to the modeling of language in which the syntactic category of a word is usually mainly determined by the type of contexts in which it occurs. Indeed, it is usually impossible to guess syntactic categories from words morphology. To draw a parallel with language modeling, one can think of our images (X_1, \dots, X_n) as sentences $(\bar{A}_1, \dots, \bar{A}_n)$ containing sequences of words $\{\{w\}\wedge\}$. However, in our setting the words w belonging to the set \bar{A}_j are not ordered as in language modeling, which makes a major difference. This contextual approach in the case of modeling image patches is not completely new and has been applied to contextual bags of visual words in [Li+11]. However, we use compression both to determine the context and the classification with respect to the context. This differs from classical context modeling that requires to estimate conditional probability distributions in order to determine the context and to perform a contextual classification.

The syntax for the new shorter code for $\{\{b\}\wedge\}$ becomes

$$\{\{b|s\}\wedge\}\{cd\}$$

so that the syntax of the new shorter code for the sample becomes

$$\{\{w|p\}\wedge\}\{a\} \wedge \{\{b|s\}\wedge\}\{cd\}$$

This code is made of two parts. The description $D ::= \{\{w|p\}\wedge\}$ of the random sets $A_{0,i}$, $1 \leq i \leq n$ using a mixture of terminal symbols of type $w \in \llbracket 1, k \rrbracket$ and of non terminal symbols of type $p \in \llbracket k+1, k+m \rrbracket$ and the new grammar

$$G ::= \{a\} \wedge \{\{b|s\}\wedge\}\{cd\} \quad (2.13)$$

specifying how to rewrite each p into a (unique) sequence of terminal symbols. This grammar itself has been compressed using non terminal symbols of type s .

Remark now that G defines a classification function f of its symbols. Let us start with the definition of $f(p)$, where $p \in \llbracket k+1, k+m \rrbracket$ is a non terminal symbol. The grammar G contains a (unique) rule $p ::= ab$ and in the compressed representation of the context of a

$$C_a = \{b' : p' ::= ab' \in G\}$$

either b has been left unchanged, in which case we set $f(p) = b$, or b is produced by a unique non terminal syntax label s , in which case we define $f(p) = s$. When j is a terminal symbol, that is when $j \in \llbracket 1, k \rrbracket$, we set $f(j) = j$. We obtain a classification function

$$f : \llbracket 1, k+m \rrbracket \longrightarrow \llbracket 1, k+m+t \rrbracket$$

where t is the number of syntax pairs cd .

We can now recode the description D using the classification function f . For any $j \in \llbracket 1, k+m \rrbracket$ we can write

$$f^{-1}(f(j)) = \{i_1 < i_2 < \dots < i_\ell\}$$

and define $h(j)$ as the solution of the equation $j = i_{h(j)}$. It is easy to see that

$$j \mapsto (f(j), h(j)), \quad j \in \llbracket 1, k+m \rrbracket$$

is one to one. Therefore we can recode the description D in the form

$$\{\{fh\}\wedge\}$$

More precisely we get

$$D = f_{1,1}h_{1,1} \dots f_{1,r_1}h_{1,r_1} \wedge \dots \wedge f_{n,1}h_{n,1} \dots f_{n,r_n}h_{n,r_n} \wedge .$$

Gathering the values of h following the same value of f in each element of the sample we can factorize the representation of D to obtain a code of type

$$\{\{f\{h\}\wedge\}\wedge\}$$

where sample elements are separated by a double sign $\wedge\wedge$. Namely

$$D = f_{1,1}h_{1,1,1} \dots h_{1,1,q_{1,1}} \wedge \dots \wedge f_{1,r_1}h_{1,r_1,1} \dots h_{1,r_1,q_{1,r_1}} \wedge \wedge \\ \dots \wedge \wedge f_{n,1}h_{n,1,1} \dots h_{n,1,q_{n,1}} \wedge \dots \wedge f_{n,r_n}h_{n,r_n,1} \dots h_{n,r_n,q_{n,r_n}} \wedge \wedge,$$

with a change of indexation. This representation can be split into

$$\{\{f\}\wedge\}\{\{h\}\wedge\}$$

where the values of h corresponding to each value of f have been gathered on the right. Namely, we get now

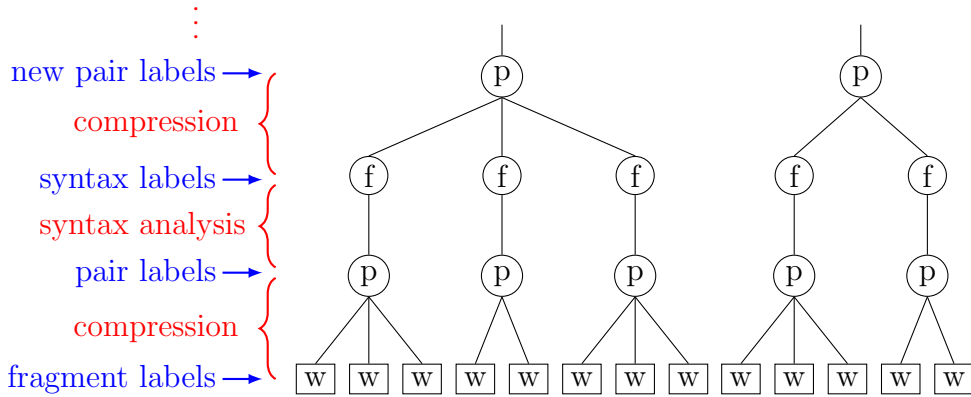
$$D = f_{1,1} \dots f_{1,r_1} \wedge \dots \wedge f_{n,1} \dots f_{n,r_n} \wedge h_{1,1,1} \dots h_{1,1,q_{1,1}} \wedge \dots \wedge h_{n,r_n,1} \dots h_{n,r_n,q_{n,r_n}} \wedge .$$

In this way, we obtain a new sample code of type

$$\{\{f\}\wedge\}\{\{h\}\wedge\}G,$$

where G is as in (2.13). This code is made of a new description D made of a family $R ::= \{\{f\}\wedge\}$ of random sets with syntax labels and of a specification $\{\{h\}\wedge\}$ describing the way to rewrite each syntax label. This is followed by a grammar G describing the meaning of non terminal labels with the help of rewriting rules.

We can now repeat the whole compression process on R to create multiple levels of syntax analysis, resulting in the estimation of a syntax tree for each initial random set \bar{A}_i , $1 \leq i \leq n$.



In the above figure we have represented the set of fragment labels indirectly indexed by a pair label. This is why there may be more than two fragment labels indexed by a single pair label. For instance, if the grammar contains $p_1 ::= w_1 p_2$ and $p_2 ::= w_2 w_3$, then we see that p_1 rewrites to $w_1 w_2 w_3$.

At the end of the syntax analysis stage, we have computed a new shorter binary code T for $(\bar{A}_1, \dots, \bar{A}_n)$. This results in a new lossy code (T, B, C) for the image sample (X_1, \dots, X_n) . The code (T, B, C) defines an approximation (Y_1, \dots, Y_n) of the sample (X_1, \dots, X_n) according to the formula

$$Y_i = \sum_{j \in \bar{A}_i} C_j, \quad 1 \leq i \leq n,$$

where \bar{A}_i can be computed from T . We use an algorithm that ensures that the quadratic distortion is under a given level, namely

$$\bar{P}_{I,S}[(X_{I,S} - Y_{I,S})^2] = n^{-1} d^{-1} \|X - Y\|^2 \leq \alpha.$$

This is more precisely a property of the fragmentation algorithm, that is maintained afterwards, since the following steps are lossless compression operations.

In this coding approach we do not estimate nor select from data a statistical model, and this is precisely what we wanted to avoid, in the hope of obtaining an algorithm that can lead to meaningful results on smaller samples. This algorithm is very simple as it combines mainly two actions: grouping and context analysis. For this reason, it could inspire new biological neuron models describing the way the mammals' brain is processing perceptive inputs. Indeed, using short codes for frequently concomitant events is compatible with the modeling of a Pavlovian or respondent conditioning. Typically we can have in mind the cat running to eat, as soon as she hears the sound of kibbles falling into her dish.

2.5. RELATION WITH STATISTICAL ESTIMATION

Nonetheless, a link can be made with statistical estimation through the following computations.

The algorithm we described in the previous sections produces a code $\theta(X_1, \dots, X_n) = (T, B, C)$ corresponding to an approximation $(Y_1, \dots, Y_n) = f \circ \theta(X_1, \dots, X_n) = g(X_1, \dots, X_n)$ of (X_1, \dots, X_n) . If we assume that T , B and C are coded by binary prefix codes, we get by concatenation a binary prefix code for θ and we can define $|\theta|$ as the code length of θ in bits. We can then define according to the Kraft inequality a sub-probability measure by the formula $Q_\theta = 2^{-|\theta|}$. Since $Y = f(\theta)$, we get also a sub-probability measure on Y , putting $Q_Y = Q_\theta \circ f^{-1}$.

Consider then the measures $\mathbb{P}_{X,Y,V}$ and $Q_{Y,V}$ defined by

$$\begin{aligned} \mathbb{P}_X &= \mathbb{P}_{X_1}^{\otimes n}, & \mathbb{P}_{Y|X} &= \delta_{g(X)}, \\ \mathbb{P}_{V|X,Y} &= \mathcal{N}(X, \sigma^2 I_{nd}), & Q_{V|Y} &= \mathcal{N}(Y, \sigma^2 I_{nd}). \end{aligned}$$

We see that as previously, $V = (V_1, \dots, V_n)$ is a noisy version of X and we can raise the question of the prediction of \mathbb{P}_{V_n} from the sample (V_1, \dots, V_{n-1}) . Note that $\mathbb{P}_{V_1, \dots, V_n} = \mathbb{P}_{V_1}^{\otimes n}$ is an i.i.d. sequence of vector valued random variables. Consider the progressive estimator based on $Q_{Y,V}$ and defined as

$$\overline{Q}_{V_n | V_1, \dots, V_{n-1}} = \frac{1}{n} \sum_{i=1}^n Q_{V_n | V_1, \dots, V_{i-1}}.$$

Using the convexity of the divergence, the fact that Q_{V_1, \dots, V_n} is exchangeable, and the fact that $\mathbb{P}_{V_1, \dots, V_n}$ is i.i.d, we can prove

LEMMA 4

$$\begin{aligned} \frac{1}{d} \mathbb{P}_{V_1, \dots, V_{n-1}} \left[\mathcal{K}(\mathbb{P}_{V_n}, \overline{Q}_{V_n | V_1, \dots, V_{n-1}}) \right] &\leq \frac{1}{nd} \mathcal{K}(\mathbb{P}_{V_1, \dots, V_n}, Q_{V_1, \dots, V_n}) \\ &\leq \frac{1}{nd} \mathcal{K}(\mathbb{P}_{Y_1, \dots, Y_n}, Q_{Y_1, \dots, Y_n}) + \frac{\alpha}{2\sigma^2} \leq \frac{1}{nd} \mathcal{K}(\mathbb{P}_\theta, Q_\theta) + \frac{\alpha}{2\sigma^2}. \end{aligned}$$

Therefore, if we can find a coding distribution Q_θ and a distortion level α depending on the sample size n such that the right-hand side goes to zero when n goes to infinity, we obtain a consistent estimator of \mathbb{P}_{V_n} , the distribution of the random image we sampled from convoluted with a Gaussian noise.

The interest of this lemma is of course mainly theoretical. The idea is not to compute $\overline{Q}_{V_n | V_1, \dots, V_{n-1}}$ in practice, but only to show that in the case where the coding distribution Q_θ is sufficiently efficient to make $\frac{1}{nd} \mathcal{K}(\mathbb{P}_\theta, Q_\theta)$ small, then Q_θ contains enough information to estimate \mathbb{P}_V accurately, since $\overline{Q}_{V_n | V_1, \dots, V_{n-1}}$ is a function of Q_θ . This is an indication that θ is a relevant representation of the information contained in \mathbb{P}_{V_n} and therefore of at least some part of the information contained in \mathbb{P}_{X_n} (since V_n is obtained by adding some noise to X_n).

PROOF. By convexity,

$$\mathbb{P}_{V_1, \dots, V_{n-1}} \left[\mathcal{K}(\mathbb{P}_{V_n}, \overline{Q}_{V_n | V_1, \dots, V_{n-1}}) \right] \leq \frac{1}{n} \mathbb{P}_{V_1, \dots, V_{n-1}} \left(\sum_{i=1}^n \mathcal{K}(\mathbb{P}_{V_n}, Q_{V_n | V_1, \dots, V_{i-1}}) \right).$$

Remark now that

$$Q_{V_n | V_1, \dots, V_{i-1}} = Q_{V_i | V_1, \dots, V_{i-1}},$$

because, Q being exchangeable,

$$Q_{V_1, \dots, V_{i-1}, V_n} = Q_{V_1, \dots, V_i}.$$

Moreover, $\mathbb{P}_{V_n} = \mathbb{P}_{V_i}$, so that

$$\begin{aligned} \mathbb{P}_{V_1, \dots, V_{n-1}} \left(\sum_{i=1}^n \mathcal{K}(\mathbb{P}_{V_n}, Q_{V_n | V_1, \dots, V_{i-1}}) \right) &= \mathbb{P}_{V_1, \dots, V_{n-1}} \left(\sum_{i=1}^n \mathcal{K}(\mathbb{P}_{V_i}, Q_{V_i | V_1, \dots, V_{i-1}}) \right) \\ &= \sum_{i=1}^n \mathbb{P}_{V_1, \dots, V_{i-1}} \left[\mathcal{K}(\mathbb{P}_{V_i | V_1, \dots, V_{i-1}}, Q_{V_i | V_1, \dots, V_{i-1}}) \right] = \mathcal{K}(\mathbb{P}_{V_1, \dots, V_n}, Q_{V_1, \dots, V_n}). \end{aligned}$$

This proves the first inequality of the lemma. To prove the second inequality, remark that

$$\mathcal{K}(\mathbb{P}_V, Q_V) \leq \mathcal{K}(\mathbb{P}_{Y,V}, Q_{Y,V}) = \mathcal{K}(\mathbb{P}_Y, Q_Y) + \mathbb{P}_Y \left[\mathcal{K}(\mathbb{P}_{V|Y}, Q_{V|Y}) \right].$$

But $\mathbb{P}_{V|Y} = \mathbb{P}_{X|Y}(\mathbb{P}_{V|X,Y}) = \mathbb{P}_{X|Y}(\mathbb{P}_{V|X})$. Accordingly,

$$\mathbb{P}_Y \left[\mathcal{K}(\mathbb{P}_{V|Y}, Q_{V|Y}) \right] \leq \mathbb{P}_Y \mathbb{P}_{X|Y} \left[\mathcal{K}(\mathbb{P}_{V|X}, Q_{V|Y}) \right] = \frac{1}{2\sigma^2} \mathbb{P}_{X,Y}(\|X - Y\|^2) \leq \frac{nd\alpha}{2\sigma^2}.$$

This proves the second inequality of the lemma. To prove the third, it is enough to note that, since $Y = f(\theta)$,

$$\mathcal{K}(\mathbb{P}_Y, Q_Y) \leq \mathcal{K}(\mathbb{P}_{\theta,Y}, Q_{\theta,Y}) = \mathcal{K}(\mathbb{P}_\theta, Q_\theta).$$

□

CHAPTER 3

Information k -means and information fragmentation algorithms

3.1. INFORMATION k -MEANS ALGORITHMS

In this part of our work, we will present and study various extensions of the k -means algorithm for their own sake. All these variants offer a mean of estimating the distribution $\mathbb{P}_{\mathbb{P}_{Y|X}}$ of a conditional probability measure $\mathbb{P}_{Y|X}$ from the observation of an i.i.d. sample $\mathbb{P}_{Y|X=X_i}$. These variants include the information fragmentation algorithm forming the first step of the syntax analysis scheme described in the overview.

Consider a couple of random variables $(X, Y) \in \mathfrak{X} \times \mathfrak{Y}$, where \mathfrak{X} and \mathfrak{Y} are complete separable metric spaces, so that we can define regular conditional probability measures. Suppose there exists a reference measure $\nu \in \mathcal{M}_+^1(\mathfrak{Y})$ such that $\mathbb{P}(\mathbb{P}_{Y|X} \ll \nu) = 1$. Define $p_X = \frac{d\mathbb{P}_{Y|X}}{d\nu}$. We are interested in the case where $\mathbb{P}_{Y|X}$ is known and represents a bag of words model. This means that each random sample X is described by a random probability measure $\mathbb{P}_{Y|X}$. In the original bag of words model, \mathfrak{Y} is a set of words, and $\mathbb{P}_{Y|X}$ is the distribution of words in a text X drawn at random from some corpus of texts. Here we include the case where \mathfrak{X} and \mathfrak{Y} can be more general measurable spaces.

We consider the following minimization problem also called information k -means problem

$$\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right),$$

where $\llbracket 1, k \rrbracket = \{1, \dots, k\}$, $\mathbb{L}_{+,1}^1(\nu) = \left\{ q \in \mathbb{L}^1(\nu) : q \geq 0, \int q d\nu = 1 \right\}$ and

$$\mathfrak{K}(q_j, p_X) = \begin{cases} \int q_j \log(q_j/p_X) d\nu, & \int q_j \mathbb{1}(p_X = 0) d\nu = 0, \\ +\infty, & \text{otherwise} \end{cases}$$

is the Kullback divergence between densities. The purpose of this section is to discuss the general properties of the information k -means problem and to build a mathematical

framework and algorithms to perform the minimization. As we have seen in the overview, we chose to study this algorithm rather than the better known k -means divergence algorithm

$$\inf_{q \in \left(\mathbb{L}_{+,1}^1(\nu)\right)^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(p_X, q_j) \right)$$

because of Proposition 2 on page 17, showing that our proposal contains the classical Euclidean k -means as a special case.

Let us state some version of the Bayes rule that will be useful in the following discussion.

LEMMA 5 *Let $\mathbb{P}_{X,Y}$ be a joint distribution defined on the product of two Polish spaces. The following statements are equivalent:*

1. *There exists a measure μ such that $\mathbb{P}_{Y|X} \ll \mu$, \mathbb{P}_X almost surely;*
2. *$\mathbb{P}_{Y|X} \ll \mathbb{P}_Y$, \mathbb{P}_X almost surely;*
3. *$\mathbb{P}_{X,Y} \ll \mathbb{P}_X \otimes \mathbb{P}_Y$;*
4. *$\mathbb{P}_{X|Y} \ll \mathbb{P}_X$, \mathbb{P}_Y almost surely.*

Moreover, they imply the following identities between Radon–Nikodym derivatives:

$$\frac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X \otimes \mathbb{P}_Y)} = \frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} = \frac{d\mathbb{P}_{X|Y}}{d\mathbb{P}_X}.$$

PROOF. To prove that 1. implies 2., it is sufficient to show that $\mathbb{P}_{Y|X} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) = 0$, \mathbb{P}_X almost surely. But when 1. is true

$$\mathbb{P}_{Y|X} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) = \int \mathbb{1} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) \frac{d\mathbb{P}_{Y|X}}{d\mu} d\mu.$$

Thus by the Tonelli-Fubini theorem

$$\begin{aligned} \mathbb{P}_X \left(\mathbb{P}_{Y|X} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) \right) &= \mathbb{P}_X \left(\int \mathbb{1} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) \frac{d\mathbb{P}_{Y|X}}{d\mu} d\mu \right) \\ &= \int \mathbb{1} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) \mathbb{P}_X \left[\frac{d\mathbb{P}_{Y|X}}{d\mu} \right] d\mu \\ &= \int \mathbb{1} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) \frac{d\mathbb{P}_Y}{d\mu} d\mu = 0. \end{aligned}$$

Therefore $\mathbb{P}_{Y|X} \left(\frac{d\mathbb{P}_Y}{d\mu} = 0 \right) = 0$, \mathbb{P}_X almost surely. Obviously 2. implies 1. Now let us show that 2. implies 3. Let f be a bounded measurable function, we have by Fubini's theorem

$$\begin{aligned} \int f d\mathbb{P}_{X,Y} &= \int \left(\int f d\mathbb{P}_{Y|X} \right) d\mathbb{P}_X = \int \left(\int f \frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} d\mathbb{P}_Y \right) d\mathbb{P}_X \\ &= \int f \frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} d(\mathbb{P}_Y \otimes d\mathbb{P}_X), \end{aligned}$$

implying 3. and that \mathbb{P}_X almost surely

$$\frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} = \frac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X \otimes \mathbb{P}_Y)}.$$

We will show now that 3. implies 2. Let f be a bounded measurable function, we have by Fubini's theorem

$$\begin{aligned} \int f d\mathbb{P}_{X,Y} &= \int f \frac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X \otimes \mathbb{P}_Y)} d(\mathbb{P}_X \otimes d\mathbb{P}_Y) \\ &= \int \left(\int f \frac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X \otimes \mathbb{P}_Y)} d\mathbb{P}_Y \right) d\mathbb{P}_X \\ &= \int \left(\int f d\mathbb{P}_{Y|X} \right) d\mathbb{P}_X, \end{aligned}$$

showing that \mathbb{P}_X almost surely $\mathbb{P}_{Y|X} \ll \mathbb{P}_Y$ and

$$\frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} = \frac{d\mathbb{P}_{X,Y}}{d(\mathbb{P}_X \otimes \mathbb{P}_Y)}.$$

The equivalence between 3. and 4. is immediate by interchanging the roles of X and Y . \square

The following lemma will be useful to optimize the information k -means criterion.

LEMMA 6 *Let $\pi \in \mathcal{M}_+^1(\Omega)$ be a probability measure on the measurable space Ω . Let $h : \Omega \rightarrow \mathbb{R} \cup \{+\infty\}$ be a measurable function such that*

$$Z = \int \exp(-h) d\pi < \infty.$$

Let $\pi_{\exp(-h)}$ be the probability measure whose density with respect to π is proportional to $\exp(-h)$ so that

$$\frac{d\pi_{\exp(-h)}}{d\pi} = \frac{\exp(-h)}{Z}.$$

The identity

$$\inf_{\eta \in \mathbb{Z}} \left(\mathfrak{K}(\rho, \pi) + \int \max\{h, \eta\} d\rho \right) = -\log \left(\int \exp(-h) d\pi \right) + \mathfrak{K}(\rho, \pi_{\exp(-h)}) \in \mathbb{R} \cup \{+\infty\}$$

is satisfied for any $\rho \in \mathcal{M}_+^1(\Omega)$ and implies that

$$\inf_{\rho \in \mathcal{M}_+^1(\Omega)} \inf_{\eta \in \mathbb{Z}} \left(\mathfrak{K}(\rho, \pi) + \int \max\{h, \eta\} d\rho \right) = -\log \left(\int \exp(-h) d\pi \right),$$

the minimum being reached when $\rho = \pi_{\exp(-h)}$.

Note that the lemma could also be written as

$$\mathfrak{K}(\rho, \pi) + \int h d\rho = -\log \left(\int \exp(-h) d\pi \right) + \mathfrak{K}(\rho, \pi_{\exp(-h)})$$

if we are willing to follow the convention that

$$\int h \, d\rho = \inf_{\eta \in \mathbb{Z}} \int \max\{h, \eta\} \, d\rho$$

and that $+\infty - \infty = +\infty$.

PROOF. See [Cat04, page 159]. Note that the role of $\eta \in \mathbb{Z}$ in this lemma is only to make sure that the integrals are always well defined in $\mathbb{R} \cup \{+\infty\}$ in the sense that the negative part of the integrand is integrable. When ρ is not absolutely continuous with respect to π , it is also not absolutely continuous with respect to $\pi_{\exp(-h)}$ since $\pi(A) = 0$ if and only if $\pi_{\exp(-h)}(A) = 0$. In this case $\mathfrak{K}(\rho, \pi) = \mathfrak{K}(\rho, \pi_{\exp(-h)}) = +\infty$ and the identity is true, both sides being equal to $+\infty$. When $\rho \ll \pi$, then $\rho \ll \pi_{\exp(-\max\{h, \eta\})}$ and

$$\frac{d\rho}{d\pi_{\exp(-\max\{h, \eta\})}} = Z_\eta \exp(\max\{h, \eta\}) \frac{d\rho}{d\pi},$$

where

$$Z_\eta = \int \exp(-\max\{h, \eta\}) \, d\pi < +\infty.$$

Therefore

$$\mathfrak{K}(\rho, \pi_{\exp(-\max\{h, \eta\})}) = \log(Z_\eta) + \int \left[\max\{h, \eta\} + \log\left(\frac{d\rho}{d\pi}\right) \right] d\rho.$$

By the monotone convergence theorem

$$\lim_{\eta \rightarrow -\infty} Z_\eta = Z \text{ and } \lim_{\eta \rightarrow -\infty} \int \left[\max\{h, \eta\} + \log\left(\frac{d\rho}{d\pi}\right) \right] d\rho = \int \left[h + \log\left(\frac{d\rho}{d\pi}\right) \right] d\rho,$$

since we know that

$$\int \left[\log(Z) + h + \log\left(\frac{d\rho}{d\pi}\right) \right]_- d\rho = \int \log\left(\frac{d\rho}{d\pi_{\exp(-h)}}\right)_- \frac{d\rho}{d\pi_{\exp(-h)}} d\pi_{\exp(-h)} \leq \exp(-1) < +\infty$$

and therefore that

$$\int \left[h + \log\left(\frac{d\rho}{d\pi}\right) \right]_- d\rho < +\infty.$$

This proves that

$$\begin{aligned} \lim_{\eta \rightarrow -\infty} \mathfrak{K}(\rho, \pi_{\exp(-\max\{h, \eta\})}) &= \log(Z) + \int \left[h + \log\left(\frac{d\rho}{d\pi}\right) \right] d\rho = \mathfrak{K}(\rho, \pi_{\exp(-h)}) \\ &= \log(Z) + \inf_{\eta \in \mathbb{Z}} \int \left[\max\{h, \eta\} + \log\left(\frac{d\rho}{d\pi}\right) \right] d\rho \\ &= \log(Z) + \inf_{\eta \in \mathbb{Z}} \left(\int \max\{h, \eta\} \, d\rho + \int \log\left(\frac{d\rho}{d\pi}\right) \, d\rho \right) \\ &= \log(Z) + \inf_{\eta \in \mathbb{Z}} \left(\int \max\{h, \eta\} \, d\rho + \mathfrak{K}(\rho, \pi) \right), \end{aligned}$$

and therefore that

$$\mathfrak{K}(\rho, \pi_{\exp(-h)}) - \log(Z) = \inf_{\eta \in \mathbb{Z}} \left(\mathfrak{K}(\rho, \pi) + \int \max\{h, \eta\} \, d\rho \right)$$

as stated in the lemma. The second statement of the lemma is a consequence of the fact that the Kullback divergence is non negative. \square

LEMMA 7 Let $\mathbb{P}_{X,Y}$ be a joint distribution defined on the product of two Polish spaces. Assume that $\mathbb{P}_X(\mathbb{P}_{Y|X} \ll \mathbb{P}_Y) = 1$. Consider the normalizing constant

$$Z = \mathbb{P}_Y \left(\exp[-\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y})] \right).$$

Obviously, $Z \in [0, 1]$. If $Z = 0$, then

$$\inf_{Q_Y \in \mathcal{M}_+^1(\mathcal{Y})} \mathbb{P}_X[\mathcal{K}(Q_Y, \mathbb{P}_{Y|X})] = +\infty.$$

Otherwise, $Z > 0$ and for any $Q_Y \in \mathcal{M}_+^1(\mathcal{Y})$,

$$\mathbb{P}_X[\mathcal{K}(Q_Y, \mathbb{P}_{Y|X})] = \mathcal{K}(Q_Y, Q_Y^*) + \mathbb{P}_X[\mathcal{K}(Q_Y^*, \mathbb{P}_{Y|X})] = \mathcal{K}(Q_Y, Q_Y^*) + \log(Z^{-1}),$$

where $Q_Y^* \ll \mathbb{P}_Y$ is defined by the relation

$$\begin{aligned} \frac{dQ_Y^*}{d\mathbb{P}_Y} &= Z^{-1} \exp[-\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y})] \\ &= Z^{-1} \exp \left\{ \mathbb{P}_X \left[\log \left(\frac{d\mathbb{P}_{Y|X}}{d\mathbb{P}_Y} \right) \right] \right\}. \end{aligned} \quad (3.1)$$

Consequently

$$\inf_{Q_Y \in \mathcal{M}_+^1(\mathcal{Y})} \mathbb{P}_X[\mathcal{K}(Q_Y, \mathbb{P}_{Y|X})] = \mathbb{P}_X[\mathcal{K}(Q_Y^*, \mathbb{P}_{Y|X})] = \log(Z^{-1}) < \infty,$$

The probability measure Q_Y^* represents the geometric mean of $\mathbb{P}_{Y|X}$ with respect to \mathbb{P}_X .

PROOF. By Lemma 1 on page 9,

$$\mathbb{P}_X[\mathcal{K}(Q_Y, \mathbb{P}_{Y|X})] = \mathcal{K}(\mathbb{P}_X \otimes Q_Y, \mathbb{P}_{X,Y}) = \mathcal{K}(Q_Y, \mathbb{P}_Y) + Q_Y[\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y})]. \quad (3.2)$$

Thus, when (3.2) is finite, $Q_Y \ll \mathbb{P}_Y$ and

$$Q_Y[\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y}) < +\infty] = 1,$$

so that

$$\mathbb{P}_Y[\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y}) < +\infty] > 0,$$

implying that $Z > 0$. Assuming from now on that (3.2) is finite, introduce

$$\mathcal{A} = \left\{ y : \mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y=y}) < +\infty \right\}.$$

From Lemma 6 on page 41 and (3.2), for any $Q_Y \in \mathcal{M}_+^1(\mathcal{A})$,

$$\begin{aligned} \mathbb{P}_X[\mathcal{K}(Q_Y, \mathbb{P}_{Y|X})] &= \underbrace{-\log[\mathbb{P}_Y(\mathcal{A})] - \log \mathbb{P}_{Y|Y \in \mathcal{A}} \left\{ \exp[-\mathcal{K}(\mathbb{P}_X, \mathbb{P}_{X|Y})] \right\}}_{=\log(Z^{-1})} + \mathcal{K}(Q_Y, Q_Y^*) \\ &= \mathbb{P}_X[\mathcal{K}(Q_Y^*, \mathbb{P}_{Y|X})] + \mathcal{K}(Q_Y, Q_Y^*). \end{aligned}$$

Moreover, when $Q_Y(\mathcal{A}) < 1$, $Q_Y \not\ll Q_Y^*$, so that both members are equal to $+\infty$. The identity (3.1) is a consequence of Lemma 5 on page 40. \square

PROPOSITION 8 *The information k -means problem can be expressed as*

$$\begin{aligned}
\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) &= \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \inf_{(q_1, \dots, q_k) \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}, p_X) \right) \\
&= \inf_{(q_1, \dots, q_k) \in (\mathbb{L}_{+,1}^1(\nu))^k} \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}, p_X) \right) \\
&= \inf_{(q_1, \dots, q_k) \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\mathfrak{K}(q_{\ell_q^*(X)}, p_X) \right) \\
&= \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}^{*,\ell}, p_X) \right) \\
&= \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\log(Z_{\ell(X)}^{-1}) \right),
\end{aligned}$$

where $\ell_q^*: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket$ is the best classification function for a fixed $q = (q_1, \dots, q_k)$ defined as

$$\ell_q^*(x) = \arg \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_x), \quad x \in \mathfrak{X},$$

whereas $q_1^{*,\ell}, \dots, q_k^{*,\ell}$ are the best information k -means centers with respect to $\ell(X)$ defined as

$$q_j^{*,\ell} = Z_j^{-1} \exp \left\{ \mathbb{P}_{X | \ell(X)=j} [\log(p_X)] \right\}, \quad j \in \llbracket 1, k \rrbracket,$$

where

$$Z_j = \int \exp \left\{ \mathbb{P}_{X | \ell(X)=j} [\log(p_X)] \right\} d\nu,$$

with the convention that $q_j^{*,\ell}$ can be given any arbitrary value in the case when $Z_j = 0$, the corresponding criterion being in this case infinite. Besides, we have the following Pythagorean identity

$$\mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}, p_X) \right) = \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}^{*,\ell}, p_X) \right) + \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}, q_{\ell(X)}^{*,\ell}) \right).$$

PROOF. This proposition is a straightforward consequence of Lemma 7 applied to $\mathbb{P}_{X, Y | \ell(X)=j}$. \square

Let us state the empirical counterpart of the previous proposition.

COROLLARY 9 *Let X_1, \dots, X_n be an i.i.d sample drawn from \mathbb{P}_X . Then, the empirical version of the information k -means problem tries to partition the observations p_{X_1}, \dots, p_{X_n} into k -clusters, what is expressed here by*

$$\begin{aligned}
\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_{X_i}) &= \inf_{\ell: \llbracket 1, n \rrbracket \rightarrow \llbracket 1, k \rrbracket} \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \frac{1}{n} \sum_{i=1}^n \mathfrak{K}(q_{\ell(i)}, p_{X_i}) \\
&= \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \inf_{\ell: \llbracket 1, n \rrbracket \rightarrow \llbracket 1, k \rrbracket} \frac{1}{n} \sum_{i=1}^n \mathfrak{K}(q_{\ell(i)}, p_{X_i}) = \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \frac{1}{n} \sum_{i=1}^n \mathfrak{K}(q_{\ell_q^*(i)}, p_{X_i}) \\
&= \inf_{\ell: \llbracket 1, n \rrbracket \rightarrow \llbracket 1, k \rrbracket} \frac{1}{n} \sum_{i=1}^n \mathfrak{K}(q_{\ell(i)}^{*,\ell}, p_{X_i}) = \inf_{\ell: \llbracket 1, n \rrbracket \rightarrow \llbracket 1, k \rrbracket} \sum_{j=1}^k \frac{|\ell^{-1}(j)|}{n} \log(Z_j^{-1}),
\end{aligned}$$

where $\ell_q^* : \mathfrak{X} \mapsto \llbracket 1, k \rrbracket$ is the best classification function for a fixed $q = (q_1, \dots, q_k)$ defined as

$$\ell_q^*(i) = \arg \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_{X_i})$$

whereas $q_j^{*,\ell}, j \in \llbracket 1, k \rrbracket$ are the information k -means centers defined as

$$q_j^{*,\ell} = Z_j^{-1} \left(\prod_{i \in \ell^{-1}(j)} p_{X_i} \right)^{1/|\ell^{-1}(j)|},$$

where

$$Z_j = \int \left(\prod_{i \in \ell^{-1}(j)} p_{X_i} \right)^{1/|\ell^{-1}(j)|} d\nu.$$

PROOF. Apply the previous proposition to the empirical measure $\bar{\mathbb{P}}_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ of the sample X_1, \dots, X_n . \square

LEMMA 10 *Let us assume that $\mathbb{P}_X \left(\int p_X^2 d\nu \right) < \infty$. Then, the optimal centers $q_j^{*,\ell}$ defined in the previous lemma verify $q_j^{*,\ell} \in \mathbb{L}^2(\nu)$. Furthermore, in this case*

$$\begin{aligned} \inf \left\{ \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) : q \in \left(\mathbb{L}_{+,1}^1(\nu) \right)^k \right\} \\ = \inf \left\{ \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) : q \in \left(\mathbb{L}_{+,1}^1(\nu) \cap \mathbb{L}^2(\nu) \right)^k \right\}. \end{aligned}$$

PROOF. Apply Jensen's inequality and the Fubini-Tonelli theorem to obtain that $q_j^{*,\ell} \in \mathbb{L}^2(\nu)$. Indeed, for any $j \in \llbracket 1, k \rrbracket$, if $Z_j = 0$, we can pick up any value for $q_j^{*,\ell}$, and in particular a value in $\mathbb{L}^2(\nu)$, in the same way if $\mathbb{P}_X(\ell(X) = j) = 0$, we can make an arbitrary choice for $q_j^{*,\ell}$, otherwise, $Z_j > 0$, and

$$\begin{aligned} \int (q_j^{*,\ell})^2 d\nu &= Z_j^{-2} \int \exp \left\{ 2\mathbb{P}_{X|\ell(X)=j} [\log(p_X)] \right\} d\nu \\ &\leq Z_j^{-2} \mathbb{P}_{X|\ell(X)=j} \left(\int p_X^2 d\nu \right) \leq Z_j^{-2} \mathbb{P}_X(\ell(X) = j)^{-1} \mathbb{P}_X \left(\int p_X^2 d\nu \right) < \infty \end{aligned}$$

Then according to Proposition 8 on the facing page

$$\begin{aligned} \mathbb{P}_X \left[\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right] &= \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left[\mathfrak{K}(q_{\ell(X)}, p_X) \right] \\ &\geq \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left[\mathfrak{K}(q_{\ell(X)}^{*,\ell}, p_X) \right] \geq \inf_{\ell: \mathfrak{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X \left[\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j^{*,\ell}, p_X) \right], \end{aligned}$$

showing that we can restrict the optimization to $q_j \in \mathbb{L}^2(\nu)$. \square

PROPOSITION 11 *Consider the particular case when $\mathfrak{X} = \mathfrak{Y} = \mathbb{R}^p$ and choose as reference measure the Gaussian measure $\nu = \mathfrak{N}_p(0_{\mathbb{R}^p}, \sigma^2 I_p)$. Choose also $\mathbb{P}_{Y|X} = \mathfrak{N}_p(X, \sigma^2 I_p)$ where*

$\sigma > 0$. In this particular setting, the information k -means problem is identical to the classical euclidian k -means problem. Namely

$$\begin{aligned} \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \frac{1}{2\sigma^2} \mathbb{P}_X \left(\|X - \mathbb{P}_{X|\ell(X)}(X)\|_2^2 \right) \\ &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \inf_{\mu_1, \dots, \mu_k \in \mathbb{R}^{p \times k}} \frac{1}{2\sigma^2} \mathbb{P}_X \left(\|X - \mu_{\ell(X)}\|_2^2 \right) \\ &= \inf_{\mu_1, \dots, \mu_k \in \mathbb{R}^{p \times k}} \frac{1}{2\sigma^2} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|_2^2 \right). \end{aligned}$$

PROOF. In this situation, p_X is equal to

$$\begin{aligned} p_X(y) &= \frac{d\mathfrak{N}_p(X, \sigma^2 I_p)}{d\nu}(y) = \exp \left\{ -\frac{1}{2\sigma^2} (\|y - X\|_2^2 - \|y\|_2^2) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (\|X\|_2^2 - 2y^\top X) \right\}. \end{aligned}$$

According to Proposition 8 on page 44 for a given classification function ℓ , the optimal centers in the information k -means problem are given by

$$q_j^{\star, \ell} = Z_j^{-1} \exp \left\{ \mathbb{P}_{X|\ell(X)=j} [\log(p_X)] \right\}, \quad j \in \llbracket 1, k \rrbracket.$$

Accordingly

$$q_j^{\star, \ell}(y) = Z_j^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \mathbb{P}_{X|\ell(X)=j} (\|X\|_2^2 - 2y^\top X) \right\} = \frac{d\mathfrak{N}_p(\mathbb{P}_{X|\ell(X)=j}(X), \sigma^2 I_p)}{d\nu}.$$

Then, computing the Kullback divergence between two multivariate normal distributions, we obtain

$$\mathfrak{K}(q_j^{\star, \ell}, p_X) = \mathfrak{K}(\mathfrak{N}_p(\mathbb{P}_{X|\ell(X)=j}(X), \sigma^2 I_p), \mathfrak{N}_p(X, \sigma^2 I_p)) = \frac{1}{2\sigma^2} \|X - \mathbb{P}_{X|\ell(X)=j}(X)\|_2^2.$$

Thus from Proposition 8 on page 44, the information k -means loss is such that

$$\begin{aligned} \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\mathfrak{K}(q_{\ell(X)}^{\star, \ell}, p_X) \right) \\ &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \frac{1}{2\sigma^2} \mathbb{P}_X \left(\|X - \mathbb{P}_{X|\ell(X)}(X)\|_2^2 \right). \end{aligned}$$

Besides,

$$\begin{aligned} \inf_{\mu_1, \dots, \mu_k \in \mathbb{R}^{p \times k}} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|_2^2 \right) &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \inf_{\mu_1, \dots, \mu_k \in \mathbb{R}^{p \times k}} \mathbb{P}_X \left(\|X - \mu_{\ell(X)}\|_2^2 \right) \\ &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \inf_{\mu_1, \dots, \mu_k \in \mathbb{R}^{p \times k}} \left\{ \mathbb{P}_X \left(\|X - \mathbb{P}_{X|\ell(X)}(X)\|_2^2 \right) + \mathbb{P}_X \left(\|\mathbb{P}_{X|\ell(X)}(X) - \mu_{\ell(X)}\|_2^2 \right) \right\} \\ &= \inf_{\ell: \mathcal{X} \rightarrow \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\|X - \mathbb{P}_{X|\ell(X)}(X)\|_2^2 \right). \end{aligned}$$

□

We are now going to linearize the information k -means algorithm, using the kernel trick. Let us introduce the separable Hilbert space $H = \{(f, x) \in \mathbb{L}^2(\nu) \times \mathbb{R}\}$ equipped with the inner-product

$$\langle h, h' \rangle = \langle h_1, h'_1 \rangle_{\mathbb{L}^2(\nu)} + \mu h_2 h'_2, \quad h = (h_1, h_2), \quad h' = (h'_1, h'_2) \in H,$$

where $\mu > 0$ is a positive real parameter to be chosen afterwards. The associated norm is

$$\|(h_1, h_2)\| = \sqrt{\langle h_1, h_1 \rangle_{\mathbb{L}^2(\nu)} + \mu h_2^2} = \sqrt{\int h_1^2 d\nu + \mu h_2^2}, \quad h = (h_1, h_2) \in H.$$

Define for any constant $B \in \mathbb{R}_+$

$$\Theta_B = \left\{ (q, \mathfrak{K}(q, 1)) : q \in \mathbb{L}_{+,1}^1(\nu) \cap \mathbb{L}^2(\nu), \int q^2 d\nu \leq B^2 \right\} \subset H,$$

this definition being justified by the fact that

$$\mathfrak{K}(q, 1) = \int q \log(q) d\nu \leq \log\left(\int q^2 d\nu\right) < +\infty \quad (3.3)$$

whenever $\int q^2 d\nu < +\infty$.

LEMMA 12 *Assume that $\text{ess sup}_X \int \log(p_X)^2 d\nu < \infty$ and $\text{ess sup}_X \int p_X^2 d\nu < \infty$. Remark first that the smallest information ball containing the support of \mathbb{P}_{p_X} satisfies*

$$\begin{aligned} \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \text{ess sup}_X \mathfrak{K}(q, p_X) &\leq \text{ess sup}_X \mathfrak{K}(1, p_X) \\ &= \text{ess sup}_X \int \log(p_X^{-1}) d\nu \leq \text{ess sup}_X \left(\int \log(p_X)^2 d\nu \right)^{1/2} < \infty. \end{aligned}$$

Define $B = \text{ess sup}_X \left(\int p_X^2 d\nu \right)^{1/2} \exp \left[\inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \text{ess sup}_X \mathfrak{K}(q, p_X) \right] < \infty$ and consider the random variable

$$W = (-\log(p_X), \mu^{-1}) \in H.$$

The following two minimization problems are equivalent

$$\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) = \inf_{\theta \in \Theta_B^k} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle_H \right).$$

PROOF. Let $B' = \text{ess sup}_X \left(\int p_X^2 d\nu \right)^{1/2}$ and $C = \text{ess sup}_X \left(\int \log(p_X)^2 d\nu \right)^{1/2}$. First let us remark that under the hypothesis of the lemma, the information k -means criterion is finite. Indeed,

$$\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) \leq \mathbb{P}_X \left(\mathfrak{K}(1, p_X) \right) = \mathbb{P}_X \left(\int \log(p_X^{-1}) d\nu \right) \leq C < \infty.$$

Now, for any classification function $\ell : \mathfrak{X} \mapsto \llbracket 1, k \rrbracket$ for which the criterion is finite, we know from Lemma 10 on page 45 that $q_j^{\star, \ell} \in \mathbb{L}^2(\nu)$ and we can remark that

$$\mathfrak{K}(q_j^{\star, \ell}, p_X) = \langle \theta_j^{\star, \ell}, W \rangle_H$$

where $\theta_j^{\star, \ell} = (q_j^{\star, \ell}, \mathfrak{K}(q_j^{\star, \ell}, 1))$. So, it is sufficient to conclude the proof to show that $\theta_j^{\star, \ell} \in \Theta_B$. As in the proof of Lemma 10 on page 45,

$$\int (q_j^{\star, \ell})^2 d\nu \leq Z_j^{-2} \mathbb{P}_{X|\ell(X)=j} \left(\underbrace{\int p_X^2 d\nu}_{\leq B'^2} \right) \leq Z_j^{-2} B'^2, \quad j \in \llbracket 1, k \rrbracket.$$

By Jensen's inequality, for any $j \in \llbracket 1, k \rrbracket$,

$$\begin{aligned} Z_j &= \sup_{q \in \mathbb{L}_{+,1}^1(\nu)} \int q \exp \left\{ \mathbb{P}_{X|\ell(X)=j} \left[\log(p_X/q) \right] \right\} d\nu \\ &\geq \sup_{q \in \mathbb{L}_{+,1}^1(\nu)} \exp \left\{ \mathbb{P}_{X|\ell(X)=j} \left[\int q \log(p_X/q) \right] \right\} \\ &= \exp \left\{ - \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \mathbb{P}_{X|\ell(X)=j} \left[\mathfrak{K}(q, p_X) \right] \right\}. \end{aligned}$$

Hence

$$Z_j^{-1} \leq \exp \left\{ \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \operatorname{ess\,sup}_X \mathfrak{K}(q, p_X) \right\} \leq \exp(C).$$

Therefore

$$\left(\int (q_j^{\star, \ell})^2 d\nu \right)^{1/2} \leq B' \exp \left[\inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \operatorname{ess\,sup}_X \mathfrak{K}(q, p_X) \right] = B \leq B' \exp(C) < \infty,$$

proving that $B < \infty$ and that $\theta_j^{\star, \ell} = (q_j^{\star, \ell}, \mathfrak{K}(q_j^{\star, \ell}, 1)) \in \Theta_B$, which concludes the proof. \square

PROPOSITION 13 *Under the hypotheses of the previous lemma there exists an optimal quantizer $\theta^\star \in \Theta_B^k$ minimizing the k -means risk, that is such that*

$$\mathbb{E} \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^\star, W \rangle \right) = \inf_{\theta \in \Theta_B^k} \mathbb{E} \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right).$$

PROOF. This follows the proof of Theorem 3.2 in [Fis10]. Remark first that for any $w \in H$,

$$\begin{aligned} H^k &\longrightarrow \mathbb{R} \\ \theta &\longmapsto \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, w \rangle \end{aligned}$$

is weakly continuous, since, by definition of the weak topology of H , $\theta \mapsto \langle \theta_j, w \rangle$ are, and taking a finite minimum is a continuous operation. Note that

$$\|\Theta_B\| = \sup_{\theta \in \Theta_B} \|\theta\| \leq \sqrt{B^2 + \mu \log(B^2)^2} < +\infty,$$

according to equation (3.3) on page 47. Therefore Θ_B is weakly relatively compact. Let $(\theta_n)_{n \in \mathbb{N}}$ be a bounded sequence in H^k , converging weakly to θ . By the dominated convergence theorem

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_{n,j}, W \rangle \right) &= \mathbb{P}_W \left(\lim_{n \rightarrow \infty} \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_{n,j}, W \rangle \right) \\ &= \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right), \end{aligned}$$

since $\left| \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_{n,j}, W \rangle \right| \leq \|\theta_n\| \|W\|_\infty$, where $\|W\|_\infty = \text{ess sup} \|W\| < +\infty$. Thus

$$\mathcal{R} : \theta \mapsto \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right)$$

is weakly continuous on the weak closure $\bar{\Theta}_B^k$ of the weakly relatively compact set Θ_B^k . Therefore, it reaches its minimum $\tilde{\theta}$ on $\bar{\Theta}_B^k$. Remark that, since

$$\mathcal{K}(q, 1) = \sup_{h \in \mathbb{L}^2(\nu)} \int h q \, d\nu - \log \left(\int \exp(h) \, d\nu \right),$$

according to the Donsker Varadhan representation, the function $q \mapsto \mathcal{K}(q, 1)$ defined on $\mathbb{L}^2(\nu) \cap \mathbb{L}_{1,+}(\nu)$ is weakly lower semicontinuous. Indeed, it is a supremum of weakly continuous function. Accordingly, its epigraph is weakly closed. As Θ_B belongs to this epigraph, its weak closure also belongs to it. This implies that for each $j \in \llbracket 1, k \rrbracket$, $\tilde{\theta}_j$ belongs to it, so that $\tilde{\theta} = ((q_j, y_j), j \in \llbracket 1, k \rrbracket)$, where $y_j \geq \mathcal{K}(q_j, 1)$. Let us put $\theta_\star = ((q_j, \mathcal{K}(q_j, 1)), j \in \llbracket 1, k \rrbracket)$. By monotonicity of \mathcal{R} with respect to y_j , the corresponding coefficient of W being positive,

$$\inf_{\theta \in \Theta_B^k} \mathcal{R}(\theta) = \inf_{\theta \in \bar{\Theta}_B^k} \mathcal{R}(\theta) = \mathcal{R}(\tilde{\theta}) \geq \mathcal{R}(\theta_\star).$$

Since $\theta_\star \in \Theta_B^k$, the reverse inequality also holds and $\mathcal{R}(\theta_\star) = \inf_{\theta \in \bar{\Theta}_B^k} \mathcal{R}(\theta)$. Note that we used the weak topology, since the unit ball of H is not strongly compact when the dimension of H is infinite. \square

3.2. INFORMATION FRAGMENTATION

We are now going to set information k -means into a broader context. This will lead us to propose modified criteria and more general representations, for bag of words, and more generally for random signals, observed through a statistical sample.

We consider as in the previous sections a couple of random variables $(X, Y) \in \mathfrak{X} \times \mathfrak{Y}$ such that $\mathbb{P}(\mathbb{P}_{Y|X} \ll \nu) = 1$ for some dominating measure $\nu \in \mathfrak{M}_+^1(\mathfrak{Y})$. We assume that \mathfrak{X} and \mathfrak{Y} are Polish spaces, so that regular conditional probability measures exist.

3.2.1. RECALL OF THE INFORMATION k -MEANS SETTING. In information k -means, we were trying to find a random variable $W = \ell(X) \in \llbracket 1, k \rrbracket$ and a conditional probability measure $Q_{Y|W}$ minimizing

$$\mathbb{P}_{X,W} \left[\mathcal{K}(Q_{Y|W}, \mathbb{P}_{Y|X}) \right].$$

3.2.2. FIRST GENERALIZATION: ESTIMATING A JOINT DISTRIBUTION. Instead of estimating the conditional probability measure $\mathbb{P}_{Y|X}$, we may be willing to estimate the joint distribution $\mathbb{P}_{X,Y}$ by some simpler measure $Q_{X,Y}$. If we are willing to use the same simplification as in information k -means, we are led to assume that

$$Q_{Y|X} = Q_{Y|W}, \quad (3.4)$$

where $W = \ell(X) \in \llbracket 1, k \rrbracket$ is a measurable function of X . Meanwhile, the first marginal Q_X can be left unconstrained. Note that condition (3.4) implies that

$$Q_{X,Y|W} = Q_{X|W} Q_{Y|W},$$

since $Q_{X,Y|W} = Q_{X|W} Q_{Y|X,W}$ and $Q_{Y|X,W} = Q_{Y|X} = Q_{Y|W}$ from (3.4).

If we still want to use an information criterion, we are led to consider the minimization of

$$\mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}).$$

From Lemma 1 on page 9, it decomposes into

$$\begin{aligned} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) &= \mathfrak{K}(Q_X, \mathbb{P}_X) + Q_X[\mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X})] \\ &= \mathfrak{K}(Q_X, \mathbb{P}_X) + Q_{X,W}[\mathfrak{K}(Q_{Y|W}, \mathbb{P}_{Y|X})]. \end{aligned}$$

Remark first that if, instead of leaving Q_X unconstrained, we let $Q_X = \mathbb{P}_X$, we get that $Q_{X,W} = \mathbb{P}_{X,W}$ (since $W = \ell(X)$), and we fall back on the information k -means problem.

Let us see now what happens if we let Q_X be free and use the optimal value of this distribution. From Lemma 6 on page 41, we obtain

$$\begin{aligned} \inf_{Q_X} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) &= \inf_{Q_X} \left\{ \mathfrak{K}(Q_X, \mathbb{P}_X) + Q_X[\mathfrak{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X})] \right\} \\ &= -\log \left\{ \mathbb{P}_X \left[\exp \left(-\mathfrak{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right) \right] \right\}, \end{aligned}$$

where the optimal value of Q_X is absolutely continuous with respect to \mathbb{P}_X , with density

$$\frac{dQ_X}{d\mathbb{P}_X} = Z^{-1} \exp \left[-\mathfrak{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right]. \quad (3.5)$$

PROPOSITION 14 *Consider the modified k -means criterion*

$$\mathfrak{C}(\ell : \mathfrak{X} \rightarrow \llbracket 1, k \rrbracket, \mu_j \in \mathfrak{M}_+^1(\mathfrak{Y}), 1 \leq j \leq k) = -\log \left\{ \mathbb{P}_X \left[\exp \left(-\mathfrak{K}(\mu_{\ell(X)}, \mathbb{P}_{Y|X}) \right) \right] \right\}.$$

For a given set of k probability measures $\mu_j \in \mathfrak{M}_+^1(\mathfrak{Y})$, $1 \leq j \leq k$, the optimal classification function ℓ is

$$\ell^*(X) = \arg \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}).$$

For a given classification function $\ell : \mathbf{X} \rightarrow \llbracket 1, k \rrbracket$, we can lower the criterion introducing Q_X defined by its density

$$\frac{dQ_X}{dP_X} = Z^{-1} \exp \left[-\mathfrak{K}(\mu_{\ell(X)}, P_{Y|X}) \right]$$

and

$$\frac{d\mu'_j}{d\nu} = Z^{-1} \exp \left\{ Q_{X|\ell(X)=j} \left[\log \left(\frac{dP_{Y|X}}{d\nu} \right) \right] \right\}, \quad 1 \leq j \leq k.$$

We obtain that

$$\mathfrak{C}(\ell, \mu'_j, 1 \leq j \leq k) = \mathfrak{C}(\ell, \mu_j, 1 \leq j \leq k) - Q_X \left[\mathfrak{K}(\mu_{\ell(X)}, \mu'_{\ell(X)}) \right].$$

Thus we have a descent algorithm that can reach a local minimum of the criterion and can play the role that Lloyd's algorithm plays for the previous information k -means criterion.

3.2.3. INFORMATION FRAGMENTATION. So far, we have made the hypotheses that $W = \ell(X)$ and that

$$Q_{Y|X} = Q_{Y|W}, \quad (3.6)$$

meaning that under the probability Q the conditional probability measures $Q_{Y|X=x}$ are equal for all $x \in \ell^{-1}(j)$. In this setting, Q is a good approximation of P when $Q_{Y|W=j}$ is a good approximation of $P_{Y|X=x}$ for each $x \in \ell^{-1}(j)$.

Moreover, hypothesis (3.6) can equivalently be written in the two following forms:

$$\begin{aligned} Q_{Y|X,W} &= Q_{Y|W}, \\ \text{or equivalently} \quad Q_{X,Y|W} &= Q_{X|W} Q_{Y|W}. \end{aligned} \quad (3.7)$$

These two last formulations are equivalent without any hypothesis on W , since it is always the case that $Q_{X,Y|W} = Q_{X|W} Q_{Y|X,W}$. On the other hand they may not be equivalent to (3.6) when W is not assumed to be a measurable function of X .

In the current section, we will relax the constraint on W to $W = \ell(X, Y) \in \llbracket 1, k \rrbracket$, in other words we will assume that W is a measurable function of (X, Y) , or that $\sigma(W) \subset \sigma(X, Y)$. This means that, instead of giving a label W to each bag of words X , we now give labels to each word Y of each bag of words X . In the mean time, we will keep hypothesis (3.7) on Q .

In this new setting,

$$Q_{Y|X} = Q_{W|X} (Q_{Y|X,W}) = Q_{W|X} (Q_{Y|W}).$$

This means that under Q the conditional probability measure $Q_{Y|X}$ describing the content of the signal X is a mixture of k fragments $Q_{Y|W=j}$, $1 \leq j \leq k$. We see therefore that the power of approximation of this new model is greater than the previous one. While in the previous model each bag of words $P_{Y|X}$ had to be close to a model $Q_{Y|W}$ taking only k possible values, in the new model each bag of words $P_{Y|X}$ has to be close to a linear combination of k fragments $Q_{Y|W}$.

Another important interpretation of the new model comes from the identity

$$Q_{X,Y} = Q_W(Q_{X|W} \otimes Q_{Y|W}).$$

It shows that we are looking for an approximation of $\mathbb{P}_{X,Y}$ by a mixture of k product distributions.

If we were to achieve this with the usual EM algorithm for mixture estimation, we would try to minimize

$$\inf_{W,Q} \mathfrak{K}(\mathbb{P}_{X,Y}, Q_{X,Y})$$

(or rather to decrease this criterion iteratively). This would fit into the usual framework of statistical inference that is to find a model that can predict with the highest possible probability all observed data configurations.

Here, in connection with the information k -means algorithm, we raise the question of minimizing the reverse divergence

$$\inf_{W,Q} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}).$$

In this framework, instead of looking for a model that can predict all the observations, we look for a model whose predictions can all be observed with the highest possible probability. This is closer to what is held for a valid theory in experimental sciences. In physics, for instance, a model is considered to be valid if all its predictions can be observed. On the other hand, a model that can predict all observations at the price of also predicting events that cannot be observed would be considered as false or not relevant.

Let us now state the equivalent of Proposition 14 on page 50, that is let us describe the generalization of Lloyd's algorithm to information fragmentation.

PROPOSITION 15 *Consider k centers $\mu_j \in \mathfrak{M}_+^1(\mathfrak{Y})$, $1 \leq j \leq k$. Let*

$$\mathfrak{T} = \left\{ A \subset \llbracket 1, k \rrbracket : \mu_i \perp \mu_j, i \neq j \in A \right\}$$

be the set of tilings by mutually singular probability measures μ_j . The partial minimum

$$\inf_{Q_{X,Y}} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y})$$

taken on all probability measures $Q_{X,Y}$, such that for some measurable function $\ell : \mathfrak{X} \times \mathfrak{Y} \rightarrow \llbracket 1, k \rrbracket$,

$$Q_{X,Y} \left[Q_{Y|X, \ell(X,Y)} = \mu_{\ell(X,Y)} \right] = 1 \tag{3.8}$$

is equal to

$$\begin{aligned} & - \sup_{\ell} \log \left\{ \mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j \left(\ell_X^{-1}(j) \right) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right\} \\ & = - \log \mathbb{P}_X \left(\sup_{A \in \mathfrak{T}} \left\{ \sum_{j \in A} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right\} \right), \end{aligned}$$

where

$$\begin{aligned}\ell_x : \mathbf{Y} &\rightarrow \llbracket 1, k \rrbracket \\ y &\mapsto \ell(x, y).\end{aligned}$$

For any given choice of ℓ , putting $W = \ell(X, Y)$, the optimum in $Q_{W, X}$ is reached when

$$\frac{dQ_{W, X}}{dP_{W, X}} = Z^{-1} \exp \left[-\mathfrak{K}(\mu_W, P_{Y|W, X}) \right]. \quad (3.9)$$

Symmetric formulas apply when we exchange the role of X and Y . They describe the optimization on all probability measures $Q_{X, Y}$ such that

$$Q_{X, Y} \left[Q_{X|Y, \ell(X, Y)} = \rho_{\ell(X, Y)} \right] = 1,$$

where $\rho_j \in \mathfrak{M}_+^1(\mathfrak{X})$, $1 \leq j \leq k$ are given. Note that in this case, the optimal value of $Q_{W, Y}$ is given by

$$\frac{dQ_{W, Y}}{dP_{W, Y}} = Z^{-1} \exp \left[-\mathfrak{K}(\rho_W, P_{X|W, Y}) \right],$$

and that we can use the identity

$$\mathfrak{K}(\rho_W, P_{X|W, Y}) = \mathfrak{K}(\rho_W, P_{X|W}) - \rho_W \left[\log \left(\frac{dP_{Y|W, X}}{dP_{Y|W}} \right) \right], \quad (3.10)$$

to avoid computing $P_{X|Y, W}$ explicitly.

Iterating these two optimization steps, we reach a local minimum for the optimization on all probability measures $Q_{X, Y}$, such that for some measurable function $\ell : \mathfrak{X} \times \mathbf{Y} \rightarrow \llbracket 1, k \rrbracket$

$$Q_{X, Y} \left[Q_{X|Y, \ell(X, Y)} = Q_{X|\ell(X, Y)} \otimes Q_{Y|\ell(X, Y)} \right] = 1.$$

PROOF. Assume for the moment that the classification function ℓ is fixed. Note that

$$Q_{Y|X, \ell(X, Y)=j} = Q_{Y|X, Y \in \ell_X^{-1}(j)}.$$

Let us set

$$Q_{Y|X, Y \in \ell_X^{-1}(j)} = \mu_j,$$

whenever $\mu_j[\ell_X^{-1}(j)] = 1$, and give it any arbitrary value otherwise. Let us set $Q_{W, X}$ as described in equation (3.9). This defines $Q_{X, Y}$. Remark that condition (3.8) is satisfied. Indeed

$$\frac{dQ_{W|X}}{dP_{W|X}}(j) = Z_X^{-1} \exp \left[-\mathfrak{K}(\mu_j, P_{Y|X, W=j}) \right],$$

so that when $\mu_j[\ell_X^{-1}(j)] < 1$, $\mathfrak{K}(\mu_j, P_{Y|X, W=j}) = \mathfrak{K}(\mu_j, P_{Y|X, Y \in \ell_X^{-1}(j)}) = +\infty$ and therefore $Q_{W|X}(j) = Q(\ell(X, Y) = j | X) = 0$. Thus

$$Q_{Y|X} \left[\ell(X, Y) \in \left\{ j : \mu_j[\ell_X^{-1}(j)] = 1 \right\} \right] = 1,$$

implying condition (3.8) according to the construction of $Q_{Y|X, \ell(X,Y)=j}$.

From the decomposition formula for the Kullback divergence (Lemma 1 on page 9), any probability measure $Q'_{X,Y}$ satisfying condition (3.8) is such that

$$\mathfrak{K}(Q'_{X,Y}, \mathbb{P}_{X,Y}) = \mathfrak{K}(Q'_{X,Y,W}, \mathbb{P}_{X,Y,W}) = \mathfrak{K}(Q'_{X,W}, \mathbb{P}_{X,W}) + Q'_{X,W} \left[\mathfrak{K}(\mu_W, \mathbb{P}_{Y|X,W}) \right].$$

Therefore, according to Lemma 6,

$$\inf_{Q'_{X,W}} \mathfrak{K}(Q'_{X,Y}, \mathbb{P}_{X,Y}) = \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = -\log \left\{ \mathbb{P}_{X,W} \left[\exp \left(-\mathfrak{K}(\mu_W, \mathbb{P}_{Y|X,W}) \right) \right] \right\}.$$

Remark that

$$\frac{d\mathbb{P}_{Y|X,W}}{d\mathbb{P}_{Y|X}} = \frac{d\mathbb{P}_{W|X,Y}}{d\mathbb{P}_{W|X}},$$

so that

$$\begin{aligned} \mathfrak{K}(\mu_W, \mathbb{P}_{Y|X,W}) &= \mathfrak{K}(\mu_W, \mathbb{P}_{Y|X}) - \mu_W \left[\log \left(\frac{d\mathbb{P}_{Y|X,W}}{d\mathbb{P}_{Y|X}} \right) \right] \\ &= \mathfrak{K}(\mu_W, \mathbb{P}_{Y|X}) - \mu_W \left[\log \left(\frac{d\mathbb{P}_{W|X,Y}}{d\mathbb{P}_{W|X}} \right) \right]. \end{aligned}$$

Accordingly,

$$\begin{aligned} &\mathbb{P}_{W|X} \left[\exp \left(-\mathfrak{K}(\mu_W, \mathbb{P}_{Y|X,W}) \right) \right] \\ &= \mathbb{P}_{W|X} \left\{ \exp \left\{ -\mathfrak{K}(\mu_W, \mathbb{P}_{Y|X}) + \mu_W \left[\log \left(\frac{d\mathbb{P}_{W|X,Y}}{d\mathbb{P}_{W|X}} \right) \right] \right\} \right\} \\ &= \sum_{j=1}^k \mathbb{1} \left[\mathbb{P}_{Y|X}(\ell_X^{-1}(j)) > 0 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) + \mu_j \left[\log(\mathbb{1}(\ell(X,Y) = j)) \right] \right] \\ &= \sum_{j=1}^k \mathbb{1} \left[\mathbb{P}_{Y|X}(\ell_X^{-1}(j)) > 0 \right] \mathbb{1} \left[\mu_j(\ell_X^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \\ &= \sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_X^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right], \end{aligned}$$

because when $\mu_j(\ell_X^{-1}(j)) = 1$ and $\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) < +\infty$, necessarily $\mu_j \ll \mathbb{P}_{Y|X}$, so that $\mathbb{P}_{Y|X}(\ell_X^{-1}(j)) > 0$. Therefore

$$\begin{aligned} &\inf \left\{ \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) : Q_{X,Y} \text{ satisfies (3.8) for some } \ell \right\} \\ &= -\sup_{\ell} \log \left[\mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_X^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right]. \end{aligned}$$

For a given classification function ℓ , consider

$$A_X(\ell) = \left\{ j : \mu_j(\ell_X^{-1}(j)) = 1 \right\}.$$

The measures $\{\mu_j, j \in A_X(\ell)\}$ being concentrated on disjoint sets, they are by definition mutually singular, so that $A_X(\ell) \in \mathfrak{J}$. Therefore

$$\begin{aligned} \sup_{\ell} \log \left[\mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_X^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right] \\ = \sup_{\ell} \log \left[\mathbb{P}_X \left(\sum_{j \in A_X(\ell)} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right] \\ \leq \log \left[\mathbb{P}_X \left(\sup_{A \in \mathfrak{J}} \left\{ \sum_{j \in A} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right\} \right) \right]. \end{aligned}$$

On the other hand, when the measures $\{\mu_j, j \in A\}$ are mutually singular, that is when $A \in \mathfrak{J}$, we can find disjoint measurable sets $\{B_j(A) \subset \mathfrak{Y}, j \in A\}$, such that $\mu_j(B_j(A)) = 1, j \in A$, and therefore we can find a measurable function $\ell_A : \mathfrak{Y} \rightarrow \llbracket 1, k \rrbracket$ such that $\mu_j(\ell_A^{-1}(j)) = 1$, for any $j \in A$. Let

$$A_x = \arg \max_{A \in \mathfrak{J}} \sum_{j \in A} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X=x}) \right], \quad x \in \mathfrak{X}.$$

Let us define ℓ^* by the formula $\ell_x^* = \ell_{A_x}$, $x \in \mathfrak{X}$. Since

$$x \mapsto \sum_{j \in A} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X=x}) \right]$$

is measurable, $x \mapsto A_x$ is measurable and therefore $x \mapsto \ell_{A_x}$ is measurable, taking its values in a finite set of measurable functions, implying that ℓ^* is measurable. Remark that

$$\begin{aligned} \log \left[\mathbb{P}_X \left(\sup_{A \in \mathfrak{J}} \left\{ \sum_{j \in A} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right\} \right) \right] &= \log \left[\mathbb{P}_X \left(\sum_{j \in A_X} \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right] \\ &\leq \log \left[\mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_{A_X}^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right] \\ &= \log \left[\mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_X^{*-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right] \\ &\leq \sup_{\ell} \log \left[\mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_j(\ell_X^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right] \right) \right]. \end{aligned}$$

This proves that

$$\begin{aligned} \inf \left\{ \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) : Q_{X,Y} \text{ satisfies (3.8) for some } \ell \right\} \\ = -\log \left[\mathbb{P}_X \left(\sup_{A \in \mathfrak{J}} \left\{ \sum_{j \in A} \mathfrak{K}(\mu_j, \mathbb{P}_{Y|X}) \right\} \right) \right]. \end{aligned}$$

To prove identity (3.10), it is enough to remark that

$$\begin{aligned}
\mathfrak{K}(\rho_W, \mathbb{P}_{X|Y,W}) &= \rho_W \left[\log \left(\frac{d\rho_W}{d\mathbb{P}_{X|Y,W}} \right) \right] \\
&= \rho_W \left[\log \left(\frac{d\rho_W}{d\mathbb{P}_{X|W}} \right) \right] - \rho_W \left[\log \left(\frac{d\mathbb{P}_{X|Y,W}}{d\mathbb{P}_{X|W}} \right) \right] \\
&= \rho_W \left[\log \left(\frac{d\rho_W}{d\mathbb{P}_{X|W}} \right) \right] - \rho_W \left[\log \left(\frac{d\mathbb{P}_{Y|X,W}}{d\mathbb{P}_{Y|W}} \right) \right].
\end{aligned}$$

□

3.3. SIGNAL FRAGMENTATION

Consider a random signal $X : \Omega \rightarrow \mathbb{R}^d$. Let us assume that $S : \Omega \rightarrow \llbracket 1, d \rrbracket$ and $V : \Omega \rightarrow \mathbb{R}$ are such that

$$\mathbb{P}_{S,V|X} = \mathbb{P}_{S|X} \mathcal{N}(X_S, \sigma^2),$$

where $\text{supp}(\mathbb{P}_{S|X}) = \llbracket 1, d \rrbracket$ and where σ is a positive real parameter. Note that X is a function of $\mathbb{P}_{S,V|X}$, given by the identity

$$X_s = \mathbb{P}_{V|X, S=s}(V), \quad 1 \leq s \leq d.$$

In other words, $\mathbb{P}_{S,V|X}$ is a lossless representation of the signal X . This representation is not unique, since we can choose the smoothing parameter σ and the site distribution $\mathbb{P}_{S|X}$ as we please to represent the same X . Usually we will take $\mathbb{P}_{S|X}$ to be the uniform probability measure on $\llbracket 1, d \rrbracket$, this choice being independent from X .

We are looking for labels $W = \ell(X, S) \in \llbracket 1, k \rrbracket$ and for a probability measure $Q \in \mathfrak{M}_+^1(\Omega)$ such that $Q_{X,S,V|W} = Q_{X|W} Q_{S,V|W}$, Q_W almost surely, minimizing

$$\mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}).$$

The process $Q_{X,S,V}$ can be understood as a patch process approximating $\mathbb{P}_{X,S,V}$. Indeed,

$$Q_{S,V|X} = Q_{W|X}(Q_{S,V|W}),$$

meaning that the signal $Q_{S,V|X}$ is a mixture of the k patches $Q_{S,V|W=j}$, $1 \leq j \leq k$.

PROPOSITION 16 (GENERALIZED k -MEANS ALGORITHM) *Consider k centers $\rho_j \in \mathfrak{M}_+^1(\mathbb{R}^d)$, $1 \leq j \leq k$. Define*

$$\mathfrak{T}_2 = \left\{ B \subset \llbracket 1, k \rrbracket : \rho_i \perp \rho_j, i \neq j \in B \right\},$$

the set of tilings by mutually singular probability measures ρ_j . The partial minimum

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V})$$

taken on all probability measures Q , such that for some measurable function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$

$$Q \left[Q_{X|S,V,\ell(X,S)} = \rho_{\ell(X,S)} \right] = 1. \quad (3.11)$$

is equal to

$$\begin{aligned} & - \sup_{\ell} \log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1} \left[\rho_j(\ell_S^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right) \\ & = - \log \mathbb{P}_S \left(\sup_{B \in \mathcal{T}_2} \sum_{j \in B} \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right), \end{aligned}$$

where

$$\begin{aligned} \ell_s : \mathbb{R}^d & \rightarrow \llbracket 1, k \rrbracket \\ x & \mapsto \ell(x, s). \end{aligned}$$

For any choice of ℓ , and in particular for the optimal one, putting $W = \ell(X, S)$, the optimum in $Q_{W,S,V}$ is reached when

$$\frac{dQ_{W,S,V}}{d\mathbb{P}_{W,S} \otimes \lambda_V} = Z^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right] g_{\sigma, \rho_W(X_S)}(V), \quad (3.12)$$

where λ_V is the Lebesgue measure on \mathbb{R} and

$$g_{\sigma, m}(v) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left(-\frac{(v-m)^2}{2\sigma^2} \right).$$

In particular, for the optimal choice of $Q_{W,S,V}$, $Q_{V|S,W} = \mathcal{N}(\rho_W(X_S), \sigma^2)$ is a Gaussian probability measure and

$$\frac{dQ_{S|W}}{d\mathbb{P}_{S|W}} = Z_W^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right].$$

On the other hand, consider k centers $\mu_{S,V}^{(j)} \in \mathfrak{M}_+^1(\llbracket 1, d \rrbracket \times \mathbb{R})$, $1 \leq j \leq k$ such that

$$\mu_{V|S}^{(j)} = \mathcal{N}(\mu_{V|S}^{(j)}(V), \sigma^2), \quad 1 \leq j \leq k.$$

Define

$$\mathfrak{T}_1 = \left\{ A \subset \llbracket 1, k \rrbracket : \mu_S^{(i)} \perp \mu_S^{(j)}, i \neq j \in A \right\},$$

the set of tilings by mutually singular probability measures $\mu_S^{(j)}$ (or equivalently by mutually singular probability measures $\mu_{S,V}^{(j)}$).

The partial minimum

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V})$$

taken on all probability measures $Q \in \mathfrak{M}_+^1(\Omega)$ such that, for some measurable function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$

$$Q \left[Q_{S,V|X, \ell(X,S)} = \mu_{S,V}^{(\ell(X,S))} \right] = 1, \quad (3.13)$$

is equal to

$$\begin{aligned}
& - \sup_{\ell} \log \mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_S^{(j)} \left(\ell_X^{-1}(j) \right) = 1 \right] \right. \\
& \quad \times \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \Bigg) \\
& = - \log \mathbb{P}_X \left(\sup_{A \in \mathfrak{T}_1} \sum_{j \in A} \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \right),
\end{aligned}$$

where

$$\begin{aligned}
\ell_x : \llbracket 1, d \rrbracket &\rightarrow \llbracket 1, k \rrbracket \\
s &\mapsto \ell(x, s).
\end{aligned}$$

For any value of ℓ , and in particular for the optimal one, considering $W = \ell(X, S)$, the minimum in $Q_{X,W}$ is reached when

$$\frac{dQ_{X,W}}{d\mathbb{P}_{X,W}} = Z^{-1} \exp \left\{ -\mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) - \mu_S^{(W)} \left[\left(\mu_{V|S}^{(W)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\}. \quad (3.14)$$

Alternating these two partial optimization steps, we can converge to a local minimum for the optimization problem

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}),$$

where the infimum is taken over probability measures $Q \in \mathfrak{M}_+^1(\Omega)$ satisfying, for some measurable classification function $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$,

$$Q \left[Q_{X,S,V|\ell(X,S)} = Q_{X|\ell(X,S)} \otimes Q_{S,V|\ell(X,S)} \right] = 1. \quad (3.15)$$

PROOF. Assume for the time being that the classification function ℓ is fixed and note that

$$Q_{X|S,V,\ell(X,S)=j} = Q_{X|S,V,X \in \ell_S^{-1}(j)}.$$

Let us set

$$Q_{X|S,V,\ell(X,S)=j} = \rho_j$$

whenever $\rho_j(\ell_S^{-1}(j)) = 1$ and give it any arbitrary value otherwise. Set $W = \ell(X, S)$ and define $Q_{S,V,W}$ by equation (B.4) on page 145. This defines $Q_{X,S,V}$. Condition (B.3) on page 145 is satisfied. Indeed,

$$\frac{dQ_{W|S}}{d\mathbb{P}_{W|S}} = Z_S^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S) / (2\sigma^2) \right].$$

Remark that

$$\mathbb{P}_{X|S,W=j} = \mathbb{P}_{X|S,X \in \ell_S^{-1}(j)},$$

so that $\mathbb{P}_{X|S,W=j}(\ell_S^{-1}(j)) = 1$. Therefore, when $\rho_j(\ell_S^{-1}(j)) < 1$, ρ_j is not absolutely continuous with respect to $\mathbb{P}_{X|S,W=j}$, $\mathfrak{K}(\rho_j, \mathbb{P}_{X|S,W=j}) = +\infty$ and $Q_{W|S}(j) = 0 = Q(\ell(X, S) =$

$j|S)$. Thus $\rho_W(\ell_S^{-1}(W)) = 1$, Q almost surely, and therefore $Q_{X|S,V,\ell(X,S)} = \rho_{\ell(X,S)}$, Q almost surely, as required by condition (B.3) on page 145.

Let us prove that

$$\frac{dQ_{S,V,W}}{dP_{S,V,W}} = Z^{-1} \exp \left[-\mathfrak{K}(\rho_W, P_{X|S,V,W}) \right]. \quad (3.16)$$

Remark that

$$\begin{aligned} \mathfrak{K}(\rho_W, P_{X|W,S,V}) &= \mathfrak{K}(\rho_W, P_{X|W,S}) - \rho_W \left[\log \left(\frac{dP_{X|W,S,V}}{dP_{X|W,S}} \right) \right] \\ &= \mathfrak{K}(\rho_W, P_{X|W,S}) - \rho_W \left[\log \left(\frac{dP_{V|X,W,S}}{dP_{V|W,S}} \right) \right] \\ &= \mathfrak{K}(\rho_W, P_{X|W,S}) - \rho_W \left[\log \left(\frac{dP_{V|X,S}}{dP_{V|W,S}} \right) \right], \quad (\text{since } W = \ell(X, S)). \end{aligned}$$

Hence

$$\begin{aligned} &\exp \left[-\mathfrak{K}(\rho_W, P_{X|S,V,W}) \right] \\ &= \exp \left\{ -\mathfrak{K}(\rho_W, P_{X|W,S}) + \rho_W \left[\log \left(\frac{dP_{V|X,S}}{dP_{V|W,S}} \right) \right] \right\} \frac{d\lambda_V}{dP_{V|W,S}} \\ &= \frac{d\lambda_V}{dP_{V|W,S}} \exp \left[-\mathfrak{K}(\rho_W, P_{X|W,S}) - \rho_W [(V - X_S)^2]/(2\sigma^2) - \log(\sigma\sqrt{2\pi}) \right] \\ &= \frac{d\lambda_V}{dP_{V|W,S}} \exp \left[-\mathfrak{K}(\rho_W, P_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right] \times \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(V - \rho_W(X_S))^2}{2\sigma^2} \right] \\ &= \frac{d\lambda_V}{dP_{V|W,S}} \exp \left[-\mathfrak{K}(\rho_W, P_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right] g_{\sigma, \rho_W(X_S)}(V). \quad (3.17) \end{aligned}$$

Thus, from the definition of $Q_{S,V,W}$,

$$\frac{dQ_{W,S,V}}{dP_{W,S} \otimes \lambda_V} = Z^{-1} \frac{dP_{V|W,S}}{d\lambda_V} \exp \left[-\mathfrak{K}(\rho_W, P_{X|S,V,W}) \right],$$

proving (3.16).

According to the decomposition formula for the divergence (Lemma 1 on page 9), for any probability measure $Q'_{X,S,V}$ satisfying condition (B.3) on page 145, for the same classification function ℓ as Q ,

$$\mathfrak{K}(Q'_{S,V,X}, P_{S,V,X}) = \mathfrak{K}(Q'_{S,V,W}, P_{S,V,W}) + Q'_{S,V,W} \left[\mathfrak{K}(\rho_W, P_{X|S,V,W}) \right],$$

so that from Lemma 6 on page 41 and from (3.16),

$$\inf_{Q'_{S,V,W}} \mathfrak{K}(Q'_{S,V,X}, P_{S,V,X}) = -\log \left\{ P_{S,V,W} \left[\exp \left[-\mathfrak{K}(\rho_W, P_{X|S,V,W}) \right] \right] \right\}$$

is reached when $Q'_{X,S,V} = Q_{X,S,V}$. Moreover, from (3.17),

$$\mathbb{P}_{V|S,W} \left[\exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|S,V,W}) \right] \right] = \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|S,W}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right].$$

Thus

$$\begin{aligned} \inf_{Q'_{S,V,X}} \mathfrak{K}(Q'_{S,V,X}, \mathbb{P}_{S,V,X}) &= -\log \mathbb{P}_{S,W} \left[\exp \left(-\mathfrak{K}(\rho_W, \mathbb{P}_{X|S,W}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right) \right] \\ &= -\log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{P}_{W|S}(j) \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S,W=j}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right). \end{aligned}$$

Remark now that whenever $\mathbb{P}_{W|S}(j) > 0$,

$$\begin{aligned} \mathfrak{K}(\rho_j, \mathbb{P}_{X|S,W=j}) &= \mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \rho_j \left[\log \left(\frac{d\mathbb{P}_{X|S,W=j}}{d\mathbb{P}_{X|S}} \right) \right] \\ &= \mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \rho_j \left[\log \left(\frac{\mathbb{P}_{W|X,S}(j)}{\mathbb{P}_{W|S}(j)} \right) \right], \end{aligned}$$

so that

$$\exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S,W=j}) \right] = \frac{\mathbb{1}[\rho_j(\ell_S^{-1}(j)) = 1]}{\mathbb{P}_{W|S}(j)} \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) \right]$$

Therefore,

$$\begin{aligned} \inf_{Q'_{S,V,X}} \mathfrak{K}(Q'_{S,V,X}, \mathbb{P}_{S,V,X}) &= -\log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1}[\mathbb{P}_{W|S}(j) > 0] \mathbb{1}[\rho_j(\ell_S^{-1}(j)) = 1] \right. \\ &\quad \left. \times \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right) \\ &= -\log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1}[\rho_j(\ell_S^{-1}(j)) = 1] \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right), \end{aligned}$$

because $\mathbb{P}_{W|S}(j) = \mathbb{P}_{X|S}(\ell_S^{-1}(j)) > 0$ when $\rho_j(\ell_S^{-1}(j)) = 1$ and $\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) < \infty$ and consequently $\rho_j \ll \mathbb{P}_{X|S}$.

Reasoning as in the proof of Proposition 15 on page 52, we conclude that

$$\begin{aligned} \inf_{\ell} \inf_{Q'_{S,V,X}} \mathfrak{K}(Q'_{S,V,X}, \mathbb{P}_{S,V,X}) &= -\sup_{\ell} \log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1}[\rho_j(\ell_S^{-1}(j)) = 1] \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right) \\ &= -\log \mathbb{P}_S \left(\sup_{B \in \mathcal{J}_2} \sum_{j \in B} \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S) \right] \right). \end{aligned}$$

The proof of the second half of the proposition is in the same spirit. We construct an optimal solution for a given classification function ℓ setting

$$Q_{S,V|X,\ell(X,S)=j} = \mu_{S,V}^{(j)}$$

when $\mu_S^{(j)}(\ell_X^{-1}(j)) = 1$, and any value otherwise. We complete the definition of $Q_{X,S,V}$ defining $Q_{X,W}$ as in equation (B.6) on page 146. We then remark that $\mathbb{P}_{S|X,W=j} = \mathbb{P}_{S|X,S \in \ell_X^{-1}(j)}$, so that when $\mu_S^{(j)}(\ell_X^{-1}(j)) < 1$, $\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X,W=j}) = +\infty$, implying that

$$Q_{X,S} \left[\mu_S^{(\ell(X,S))}(\ell_X^{-1}(\ell(X,S))) = 1 \right] = 1,$$

and therefore condition (3.13) on page 57 is satisfied. Let us prove now that

$$\frac{dQ_{X,W}}{d\mathbb{P}_{X,W}} = Z^{-1} \exp \left[-\mathfrak{K}(\mu_{S,V}^{(W)}, \mathbb{P}_{S,V|X,W}) \right]. \quad (3.18)$$

Indeed,

$$\begin{aligned} \mathfrak{K}(\mu_{S,V}^{(W)}, \mathbb{P}_{S,V|X,W}) &= \mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) + \mu_S^{(W)} \left[\mathfrak{K}(\mu_{V|S}^{(W)}, \mathbb{P}_{V|S,X,W}) \right] \\ &= \mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) + \mu_S^{(W)} \left[\mathfrak{K}(\mu_{V|S}^{(W)}, \mathbb{P}_{V|S,X}) \right], \quad (\text{since } W = \ell(X,S)) \\ &= \mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) + \mu_S^{(W)} \left[(\mu_{V|S}^{(W)}(V) - X_S)^2 / (2\sigma^2) \right]. \end{aligned}$$

From the decomposition property of the divergence,

$$\begin{aligned} \mathfrak{K}(Q'_{X,S,V}, \mathbb{P}_{X,S,V}) &= \mathfrak{K}(Q'_{X,S,V,W}, \mathbb{P}_{X,S,V,W}) \\ &= \mathfrak{K}(Q'_{X,W}, \mathbb{P}_{X,W}) + Q'_{X,W} \left[\mathfrak{K}(Q'_{S,V|X,W}, \mathbb{P}_{S,V|X,W}) \right] \\ &= \mathfrak{K}(Q'_{X,W}, \mathbb{P}_{X,W}) + Q'_{X,W} \left[\mathfrak{K}(\mu_{S,V}^{(W)}, \mathbb{P}_{S,V|X,W}) \right], \end{aligned}$$

so that according to equation (3.18) and Lemma 6 on page 41, Q is optimal for ℓ fixed, and

$$\begin{aligned} \inf_{Q'} \mathfrak{K}(Q'_{X,S,V}, \mathbb{P}_{X,S,V}) &= -\log \mathbb{P}_{X,W} \left[\exp \left[-\mathfrak{K}(\mu_{S,V}^{(W)}, \mathbb{P}_{S,V|X,W}) \right] \right] \\ &= -\log \mathbb{P}_{X,W} \left[\exp \left\{ -\mathfrak{K}(\mu_S^{(W)}, \mathbb{P}_{S|X,W}) - \mu_S^{(W)} \left[(\mu_{V|S}^{(W)}(V) - X_S)^2 / (2\sigma^2) \right] \right\} \right] \\ &= -\log \mathbb{P}_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_S^{(j)}(\ell_X^{-1}(j)) = 1 \right] \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[(\mu_{V|S}^{(j)}(V) - X_S)^2 / (2\sigma^2) \right] \right\} \right), \end{aligned}$$

where the last identity is proved as in the case of the first half of the proposition. As in the proof of Proposition 15 on page 52, we conclude that

$$\begin{aligned} \inf_{\ell} \inf_{Q'} \mathfrak{K}(Q'_{X,S,V}, \mathbb{P}_{X,S,V}) &= -\log \mathbb{P}_X \left(\sup_{A \in \mathcal{I}_1} \sum_{j \in A} \exp \left\{ -\mathfrak{K}(\mu_S^{(j)}, \mathbb{P}_{S|X}) - \mu_S^{(j)} \left[(\mu_{V|S}^{(j)}(V) - X_S)^2 / (2\sigma^2) \right] \right\} \right). \end{aligned}$$

□

CHAPTER 4

PAC-Bayesian bounds for information k -means and information fragmentation

4.1. A PAC-BAYESIAN BOUND FOR INFORMATION k -MEANS

In this section, we consider the setting described in Section 3.1 on page 39, and more specifically in Proposition 13 on page 48.

We first derive a non-asymptotic dimension free bound for the representation of the problem in the separable Hilbert space H . Then, we translate this result to the case of the original information k -means risk.

Our proofs are based on the following PAC-Bayesian lemma.

LEMMA 17 *Consider two measurable spaces \mathcal{T} and \mathcal{W} , a prior probability measure $\pi \in \mathcal{M}_+^1(\mathcal{T})$ defined on \mathcal{T} , and a measurable function $h : \mathcal{T} \times \mathcal{W} \rightarrow \mathbb{R}$. Let $W \in \mathcal{W}$ be a random variable and let (W_1, \dots, W_n) be a sample made of n independent copies of W . Let λ be a positive real parameter.*

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \exp \left[\sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})} \sup_{\eta \in \mathbb{N}} \left\{ \int \min \left\{ \eta, -\lambda \sum_{i=1}^n h(\theta', W_i) \right. \right. \right. \right. \\ \left. \left. \left. - n \log \left[\mathbb{P}_W \exp[-\lambda h(\theta', W)] \right] \right\} d\rho(\theta') - \mathfrak{K}(\rho, \pi) \right\} \right] \right\} \leq 1. \quad (4.1)$$

Consequently, for any $\delta \in]0, 1[$, with probability at least $1 - \delta$,

$$\sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})} \sup_{\eta \in \mathbb{N}} \left\{ \int \min \left\{ \eta, -\lambda \sum_{i=1}^n h(\theta', W_i) \right. \right. \\ \left. \left. - n \log \left[\mathbb{P}_W \exp[-\lambda h(\theta', W)] \right] \right\} d\rho(\theta') - \mathfrak{K}(\rho, \pi) \right\} \leq \log(\delta^{-1}). \quad (4.2)$$

Note that the role of η in this formula is to give a meaning to the integration with respect to ρ in all circumstances.

PROOF. We follow here the same arguments as in the proof of Proposition 1.7 in [Giu15]. Remark that the supremum in ρ can be restricted to the case when $\mathfrak{K}(\rho, \pi) < \infty$, and recall that in this case $\rho \ll \pi$ and $\mathfrak{K}(\rho, \pi) = \int \log\left(\frac{d\rho}{d\pi}(\theta')\right) d\rho(\theta')$. Note also that

$$\int \mathbb{1}\left(\frac{d\rho}{d\pi}(\theta') > 0\right) d\rho(\theta') = \int \mathbb{1}\left(\frac{d\rho}{d\pi}(\theta') > 0\right) \frac{d\rho}{d\pi}(\theta') d\pi(\theta') = \int \frac{d\rho}{d\pi}(\theta') d\pi(\theta') = 1.$$

Applying Jensen's inequality, we get

$$\begin{aligned} & \exp\left\{\sup_{\rho \in \mathfrak{M}_+^1(\mathcal{T})} \sup_{\eta \in \mathbb{N}} \int \min\left\{\eta, -\lambda \sum_{i=1}^n h(\theta', W_i) \right. \right. \\ & \quad \left. \left. - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} d\rho(\theta') - \mathfrak{K}(\rho, \pi)\right\} \\ & \leq \sup_{\eta \in \mathbb{N}} \sup_{\substack{\rho \in \mathfrak{M}_+^1(\mathcal{T}) \\ \mathfrak{K}(\rho, \pi) < \infty}} \int \exp\left\{\min\left\{\eta, -\lambda \sum_{i=1}^n h(\theta', W_i) \right. \right. \\ & \quad \left. \left. - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} \frac{d\rho}{d\pi}(\theta')^{-1} d\rho(\theta')\right\} \\ & = \sup_{\eta \in \mathbb{N}} \sup_{\substack{\rho \in \mathfrak{M}_+^1(\mathcal{T}) \\ \mathfrak{K}(\rho, \pi) < \infty}} \int \exp\left\{\min\left\{\eta, \right. \right. \\ & \quad \left. \left. -\lambda \sum_{i=1}^n h(\theta', W_i) - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} \mathbb{1}\left(\frac{d\rho}{d\pi}(\theta') > 0\right) d\pi(\theta')\right\} \\ & \leq \sup_{\eta \in \mathbb{N}} \int \exp\left\{\min\left\{\eta, -\lambda \sum_{i=1}^n h(\theta', W_i) - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} d\pi(\theta')\right\} \\ & \quad \stackrel{\text{monotone convergence}}{=} \int \exp\left\{-\lambda \sum_{i=1}^n h(\theta', W_i) - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} d\pi(\theta'). \end{aligned}$$

Let us put

$$\begin{aligned} Y' &= \sup_{\rho \in \mathfrak{M}_+^1(\mathcal{T})} \sup_{\eta \in \mathbb{N}} \left\{ \int \min\left\{\eta, -\lambda \sum_{i=1}^n h(\theta', W_i) \right. \right. \\ & \quad \left. \left. - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} d\rho(\theta') - \mathfrak{K}(\rho, \pi) \right\} \text{ and} \\ Y &= \log \int \exp\left\{-\lambda \sum_{i=1}^n h(\theta', W_i) - n \log\left[\mathbb{P}_W \exp[-\lambda h(\theta', W)]\right]\right\} d\pi(\theta'). \end{aligned}$$

We just proved that $Y' \leq Y$. Moreover, Y is measurable, according to Fubini's theorem for non-negative functions. Therefore Y is a random variable. Note that we did not prove that Y' itself is measurable. Remark now that

$$\mathbb{P}_{W_1, \dots, W_n}[\exp(Y)]$$

$$\begin{aligned}
&= \mathbb{P}_{W_1, \dots, W_n} \int \exp \left\{ -\lambda \sum_{i=1}^n h(\theta', W_i) - n \log \left[\mathbb{P}_W \exp[-\lambda h(\theta', W)] \right] \right\} d\pi(\theta'), \\
&\stackrel{\text{Fubini}}{=} \int \mathbb{P}_{W_1, \dots, W_n} \exp \left\{ -\lambda \sum_{i=1}^n h(\theta', W_i) - n \log \left[\mathbb{P}_W \exp[-\lambda h(\theta', W)] \right] \right\} d\pi(\theta') \\
&= \int \left(\mathbb{1} \left(\mathbb{P}_W \left[\exp(-\lambda h(\theta', W)) \right] < +\infty \right) \prod_{i=1}^n \frac{\mathbb{P}_{W_i}[\exp(-\lambda h(\theta', W_i))]}{\mathbb{P}_W[\exp(-\lambda h(\theta', W))]} \right) d\pi(\theta') \leq 1,
\end{aligned}$$

proving the first part of the lemma. From Markov's inequality,

$$\mathbb{P}(Y \geq \log(\delta^{-1})) \leq \delta \mathbb{P}_{W_1, \dots, W_n}[\exp(Y)] \leq \delta.$$

Consequently $\mathbb{P}(Y \leq \log(\delta^{-1})) \geq 1 - \delta$. We have proved that the non necessarily measurable event $Y' \leq \log(\delta^{-1})$ contains the measurable event $Y \leq \log(\delta^{-1})$ whose probability is at least $1 - \delta$. \square

LEMMA 18 *Let W be a bounded random vector in a separable Hilbert space H . Let $\Theta \subset H$ be a bounded set of parameters. Define $\|\Theta\| = \sup\{\|\theta\| : \theta \in \Theta\}$ and $\|W\|_\infty = \text{ess sup}\|W\|$.*

Assume that

$$\mathbb{P}\left(\inf_{\theta \in \Theta} \langle \theta, W \rangle \geq 0\right) = 1. \quad (4.3)$$

Let W_1, \dots, W_n be a statistical sample made of n independent copies of W . Consider any number of centers $k \geq 2$, any sample size $n \geq 8k/\log(k)$ and any probability level $\delta \geq \exp(-n \log(k))$. With probability at least $1 - \delta$, for any $\theta \in \Theta^k$,

$$\mathbb{P}_W\left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) \leq \bar{\mathbb{P}}_W\left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) + \left(\sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n}\right)^{1/4}\right) \|\Theta\| \|W\|_\infty,$$

where $\bar{\mathbb{P}}_W = \frac{1}{n} \sum_{i=1}^n \delta_{W_i}$ is the empirical measure of the sample. In expectation

$$\mathbb{P}_{W_1, \dots, W_n} \left[\sup_{\theta \in \Theta^k} \left(\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) - \bar{\mathbb{P}}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) \right) \right] \leq \left(\frac{8k \log(k)}{n} \right)^{1/4} \|\Theta\| \|W\|_\infty,$$

Note that condition (4.3) is not essential. Its role is only to provide slightly better constants.

PROOF. Let $\Phi : H \rightarrow \ell_2 \subset \mathbb{R}^{\mathbb{N}}$ be an isometry (obtained by considering an orthonormal basis of H). Let us consider an infinite sequence of normal random variables $\varepsilon \sim \mathcal{N}(0, 1)^{\otimes \mathbb{N}} \in \mathfrak{M}_+^1(\mathbb{R}^{\mathbb{N}})$, independent from the random variable X , and therefore from W . For any two random variables $U, V : \Omega \rightarrow \mathbb{R}^{\mathbb{N}}$, let us define the following extension of the inner product

$$\langle U, V \rangle = \begin{cases} \lim_{s \rightarrow \infty} \sum_{t=0}^s U_t V_t, & \text{when } \limsup_{s \rightarrow \infty} \sum_{t=0}^s U_t V_t = \liminf_{s \rightarrow \infty} \sum_{t=0}^s U_t V_t \in \mathbb{R}, \\ 0, & \text{otherwise.} \end{cases} \quad (4.4)$$

Note that the fact that U and V are measurable implies that $\langle U, V \rangle : \Omega \rightarrow \mathbb{R}$ is measurable. Remark nonetheless that this extension of the inner product is not bilinear, due to the introduction of a condition depending on the existence of the limit. Note that our construction is related to the Gaussian process

$$G(f, g) = \langle \Phi(f) + \beta^{-1/2} \varepsilon, \Phi(g) \rangle, \quad (f, g) \in H^2. \quad (4.5)$$

For any $\theta \in H$, introduce $\rho_\theta = \mathbb{P}_{\Phi(\theta) + \beta^{-1/2} \varepsilon} \in \mathcal{M}_+^1(\mathbb{R}^\mathbb{N})$, and put, for any $\theta \in H^k$, $\rho_\theta = \bigotimes_{j=1}^k \rho_{\theta_j} \in \mathcal{M}_+^1((\mathbb{R}^\mathbb{N})^k)$. Note that for any $\theta, \tilde{\theta} \in H$,

$$\mathcal{K}(\rho_\theta, \rho_{\tilde{\theta}}) = \sum_{i \in \mathbb{N}} \mathcal{K}(\mathbb{P}_{\Phi(\theta)_i + \beta^{-1/2} \varepsilon_i}, \mathbb{P}_{\Phi(\tilde{\theta})_i + \beta^{-1/2} \varepsilon_i}) = \sum_{i \in \mathbb{N}} \frac{\beta}{2} [\Phi(\theta)_i - \Phi(\tilde{\theta})_i]^2 = \frac{\beta \|\theta - \tilde{\theta}\|^2}{2}.$$

In the same way, for any $\theta, \tilde{\theta} \in H^k$,

$$\mathcal{K}(\rho_\theta, \rho_{\tilde{\theta}}) = \frac{\beta}{2} \sum_{j=1}^k \|\theta_j - \tilde{\theta}_j\|^2 = \frac{\beta}{2} \|\theta - \tilde{\theta}\|^2.$$

We will use a PAC-Bayesian inequality deduced from Lemma 17 on page 63 to be stated later. It requires to upper bound the following expressions, that can be seen as perturbations of exponential moments :

$$M(\theta, \lambda) = \int \log \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] d\rho_\theta(\theta'), \quad \theta \in \Theta^k, \quad \lambda > 0.$$

Note that $M(\theta, \lambda)$ being the expectation with respect to ρ_θ of a non positive random variable is always well defined in $\mathbb{R}_- \cup \{-\infty\}$. We can use Jensen's inequality to move the perturbation inside the logarithm, and Fubini's theorem for non-negative functions, to move it inside the expectation, obtaining

$$M(\theta, \lambda) \leq \log \mathbb{P}_W \left[\int \exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) d\rho_\theta(\theta') \right].$$

We can then linearize the minimum in j , remarking that

$$\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \leq \inf_{\alpha \geq 1} \left(\sum_{j=1}^k \exp \left(-\alpha \lambda \langle \theta'_j, \Phi(W) \rangle \right) \right)^{1/\alpha}.$$

Note that

$$\rho_\theta \circ (\theta'_j \mapsto \langle \theta'_j, \Phi(W) \rangle)^{-1} = \mathcal{N}(\langle \theta_j, W \rangle, \|W\|^2/\beta),$$

meaning that under the probability measure $d\rho_\theta(\theta')$, the random variable $\langle \theta'_j, \Phi(W) \rangle$ (where we use the extension of the scalar product) is a scalar Gaussian random variable with mean $\langle \theta_j, W \rangle$ (where we use the scalar product in H) and variance $\|W\|^2/\beta$. Using Jensen's inequality again and the formula giving the Laplace transform of a Gaussian measure, we deduce that

$$\begin{aligned}
M(\theta, \lambda) &\leq \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \left(\sum_{j=1}^k \int \exp \left(-\alpha \lambda \langle \theta'_j, \Phi(W) \rangle \right) d\rho_\theta(\theta') \right)^{1/\alpha} \right] \\
&= \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \left(\sum_{j=1}^k \exp \left(-\alpha \lambda \langle \theta_j, W \rangle + \frac{\alpha^2 \lambda^2}{2\beta} \|W\|^2 \right) \right)^{1/\alpha} \right] \\
&\leq \log \mathbb{P}_W \exp \left(-\lambda \min_{j \in [1, k]} \langle \theta_j, W \rangle + \inf_{\alpha \geq 1} \frac{\alpha \lambda^2}{2\beta} \|W\|^2 + \frac{\log(k)}{\alpha} \right) \\
&\leq \log \mathbb{P}_W \exp \left(-\lambda \min_{j \in [1, k]} \langle \theta_j, W \rangle + \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\| \right) \\
&\leq \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \log \mathbb{P}_W \exp \left(-\lambda \min_{j \in [1, k]} \langle \theta_j, W \rangle \right),
\end{aligned}$$

under the condition that $\alpha^2 = \frac{2\beta \log(k)}{\lambda^2 \|W\|^2} \geq 1$ almost surely, that is

$$\lambda^2 \|W\|_\infty^2 \leq 2\beta \log(k). \quad (4.6)$$

Notice that

$$\mathbb{P}_W \left(-\lambda \min_{j \in [1, k]} \langle \theta_j, W \rangle \in [-\lambda \|\Theta\| \|W\|_\infty, 0] \right) = 1.$$

Therefore, according to Hoeffding's lemma,

$$\log \mathbb{P}_W \exp \left(-\lambda \min_{j \in [1, k]} \langle \theta_j, W \rangle \right) \leq -\lambda \mathbb{P}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) + \frac{\lambda^2}{8} \|\Theta\|^2 \|W\|_\infty^2.$$

Indeed, for any measurable function f such that $\mathbb{P}(a \leq f(W) \leq b) = 1$, a Taylor expansion with integral remainder of $u \mapsto \log \mathbb{P}_W \left(\exp(uf(W)) \right)$ between 0 and 1 gives

$$\log \mathbb{P}_W \left(\exp(f(W)) \right) = \mathbb{P}_W(f(W)) + \int_0^1 (1-u) \mathbf{Var}_u(f(W)) du,$$

where, putting

$$\mathbb{P}_{W|u}(h(W)) = \frac{\mathbb{P}_W \left(\exp[uf(W)] h(W) \right)}{\mathbb{P}_W \left(\exp[uf(W)] \right)},$$

$$\mathbf{Var}_u(f(W)) = \mathbb{P}_{W|u} \left[\left(f(W) - \mathbb{P}_{W|u} f(W) \right)^2 \right] \leq \mathbb{P}_{W|u} \left[\left(f(W) - (a+b)/2 \right)^2 \right] \leq \frac{(b-a)^2}{4}.$$

We refer to [Cat14] for more details concerning the proof of Hoeffding's lemma and related bounds. We get

$$M(\theta, \lambda) \leq -\lambda \mathbb{P}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) + \frac{\lambda^2}{8} \|\Theta\|^2 \|W\|_\infty^2 + \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty. \quad (4.7)$$

Let us apply the PAC-Bayesian inequality of Lemma 17 on page 63 to our problem. Choose $\pi = \rho_{\tilde{\theta}}$ for some deterministic value $\tilde{\theta}$ of the parameter (that we will set equal to zero later). We obtain that with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta_B^k} \sup_{\eta \in \mathbb{N}} \int \min \left\{ \eta, -\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle \right. \\ \left. - n \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} \right\} d\rho_\theta(\theta') - \frac{\beta}{2} \|\theta - \tilde{\theta}\|^2 \leq \log(\delta^{-1}).$$

When θ' is distributed according to ρ_θ , $(\langle \theta'_j, \Phi(W_i) \rangle, j \in \llbracket 1, k \rrbracket) \sim \bigotimes_{j=1}^k \mathcal{N}(\langle \theta_j, W_i \rangle, \beta^{-1} \|W_i\|^2)$ is a Gaussian vector in \mathbb{R}^k with independent components and it is elementary from this remark to check that $\min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle$ is integrable with respect to ρ_θ , for any (fixed) value of $W_i \in H$. Therefore, from what we proved already the negative part of

$$-\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle - n \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\}$$

is integrable with respect to ρ_θ , so that the integral of this expression with respect to ρ_θ is well defined in $\mathbb{R} \cup \{+\infty\}$ and the monotone convergence theorem applied to its positive part ensures that

$$\begin{aligned} & \sup_{\eta \in \mathbb{N}} \int \min \left\{ \eta, -\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle - n \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} \right\} d\rho_\theta(\theta') \\ &= \int \left\{ -\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle - n \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} \right\} d\rho_\theta(\theta') \\ &= -\lambda \sum_{i=1}^n \int \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W_i) \rangle d\rho_\theta(\theta') - n \int \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} d\rho_\theta(\theta') \\ &\geq -\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \int \langle \theta'_j, \Phi(W_i) \rangle d\rho_\theta(\theta') - n \int \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} d\rho_\theta(\theta') \\ &= -\lambda \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W_i \rangle - \underbrace{n \int \log \left\{ \mathbb{P}_W \left[\exp \left(-\lambda \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right] \right\} d\rho_\theta(\theta')}_{= M(\theta, \lambda)}. \end{aligned}$$

Thus, combining the upper bound (4.7) on page 67 with Lemma 17 on page 63, we get with probability at least $1 - \delta$, for any $\theta \in \Theta^k$,

$$\begin{aligned} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) &\leq \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W_i \rangle \\ &\quad + \frac{\lambda}{8} \|\Theta\|^2 \|W\|_\infty^2 + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{\beta \|\theta - \tilde{\theta}\|^2 + 2 \log(\delta^{-1})}{2n\lambda}. \end{aligned}$$

Choose $\tilde{\theta} = 0$. We obtain, with probability at least $1 - \delta$, for any $\theta \in \Theta^k$,

$$\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) - \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W_i \rangle$$

$$\begin{aligned}
&\leq \frac{\lambda}{8} \|\Theta\|^2 \|W\|_\infty^2 + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{2n\lambda} \\
&\leq \sqrt{\frac{1}{4n} (k\beta \|\Theta\|^2 + 2 \log(\delta^{-1}))} \|\Theta\| \|W\|_\infty + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty \\
&\leq \left(\sqrt{\frac{\log(\delta^{-1})}{2n}} + \sqrt{\frac{k\beta \|\Theta\|^2}{4n}} \right) \|\Theta\| \|W\|_\infty + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty \\
&\leq \left(\sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n} \right)^{1/4} \right) \|\Theta\| \|W\|_\infty,
\end{aligned}$$

where we have taken

$$\begin{aligned}
\beta &= \sqrt{\frac{8n \log(k)}{k}} \|\Theta\|^{-2}, \\
\text{and } \lambda &= 2 \sqrt{\frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \\
&= 2 \sqrt{\sqrt{\frac{8k \log(k)}{n}} + \frac{2 \log(\delta^{-1})}{n}} \|\Theta\|^{-1} \|W\|_\infty^{-1}.
\end{aligned}$$

Our computations require condition (4.6) on page 67 that translates as

$$\begin{aligned}
\sqrt{\frac{2k \log(k)}{n}} + \frac{\log(\delta^{-1})}{n} &\leq \log(k) \sqrt{\frac{n \log(k)}{2k}} \\
&\iff \delta \geq \exp \left[-n \log(k) \sqrt{\frac{n \log(k)}{2k}} \left(1 - \frac{2k}{n \log(k)} \right) \right].
\end{aligned}$$

The condition is for example satisfied when $k \geq 2$, $n \geq 8k/\log(k)$ and $\delta \geq \exp(-n \log(k))$.

□

LEMMA 19 *Let W be a bounded random vector in a separable Hilbert space H . Let $\Theta \subset H$ be a bounded set of parameters. Define $\|\Theta\| = \sup\{\|\theta\| : \theta \in \Theta\}$ and $\|W\|_\infty = \text{ess sup}\|W\|$. Let W_1, \dots, W_n be a statistical sample made of n independent copies of W . Consider any number $k \geq 2$ of centers, any sample size $n \geq 2k/\log(k)$ and any probability level $\delta \geq \exp(-n \log(k))$.*

Let $\bar{\mathbb{P}}_W$ be the expectation with respect to the empirical measure $\frac{1}{n} \sum_{i=1}^n \delta_{W_i}$.

With probability at least $1 - \delta$, for any $\theta \in \Theta^k$,

$$\begin{aligned}
&\mathbb{P}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) - \inf_{\theta \in \Theta^k} \mathbb{P}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) \\
&\leq \bar{\mathbb{P}}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) - \inf_{\theta \in \Theta^k} \bar{\mathbb{P}}_W \left(\min_{j \in [1, k]} \langle \theta_j, W \rangle \right) \\
&\quad + \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \right) \|\Theta\| \|W\|_\infty.
\end{aligned}$$

Consequently, if

$$\widehat{\theta}(W_1, \dots, W_n) \in \arg \min_{\theta \in \Theta^k} \bar{\mathbb{P}}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right)$$

is an empirical minimizer, with probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \widehat{\theta}_j, W \rangle \mid W_1, \dots, W_n \right) &\leq \inf_{\theta \in \Theta^k} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) \\ &\quad + \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + 2 \left(\frac{2k \log(k)}{n} \right)^{1/4} \right) \|\Theta\| \|W\|_\infty. \end{aligned}$$

PROOF. In this proof, we will bound

$$(\mathbb{P}_W - \bar{\mathbb{P}}_W) \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle \right),$$

where $\bar{\mathbb{P}}_W$ is the expectation with respect to the empirical measure $\frac{1}{n} \sum_{i=1}^n \delta_{W_i}$, and where

$$\theta_j^* \in \arg \min_{\theta \in \bar{\Theta}^k} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right),$$

where $\bar{\Theta}$ is the weak closure of Θ . We will use the same line of proof as in the case of the previous proposition. With the same notation, we need to bound the quantities

$$M(\theta, \lambda) = \int \log \mathbb{P}_W \left[\exp \left(\lambda \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle \right) \right) \right] d\rho_\theta(\theta'), \quad \theta \in \Theta^k, \quad \lambda > 0.$$

Acting exactly as in the previous proof, we get that

$$\begin{aligned} M(\theta, \lambda) &\leq \log \mathbb{P}_W \exp \left(\lambda \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) + \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\| \right) \\ &\leq \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \log \mathbb{P}_W \exp \left(\lambda \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) \right). \end{aligned}$$

under condition (4.6) on page 67. Remark that \mathbb{P}_W almost surely

$$-\|\Theta\| \|W\|_\infty \leq \lambda \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) \leq \|\Theta\| \|W\|_\infty.$$

Apply Hoeffding's inequality to deduce that

$$M(\theta, \lambda) \leq \lambda \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) + \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{\lambda^2}{2} \|\Theta\|^2 \|W\|_\infty^2.$$

Operating in the same way as in the previous proof, we obtain that with probability at least $1 - \delta$ for any $\theta \in \Theta^k$,

$$\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) - \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle \right) - \bar{\mathbb{P}}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) + \bar{\mathbb{P}}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle \right)$$

$$\begin{aligned}
&\leq \frac{\lambda}{2} \|\Theta\|^2 \|W\|_\infty^2 + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{\beta \|\theta\|^2 + 2 \log(\delta^{-1})}{2n\lambda} \\
&\leq \sqrt{\frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \|\Theta\| \|W\|_\infty + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty \\
&\leq \sqrt{\frac{2 \log(\delta^{-1})}{n}} \|\Theta\| \|W\|_\infty + \sqrt{\frac{k\beta}{n}} \|\Theta\|^2 \|W\|_\infty + \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty \\
&\leq \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + 2 \left(\frac{2k \log(k)}{n} \right)^{1/4} \right) \|\Theta\| \|W\|_\infty.
\end{aligned}$$

In these inequalities, we have chosen

$$\begin{aligned}
\beta &= \sqrt{\frac{2n \log(k)}{k}} \|\Theta\|^{-2} \\
\text{and } \lambda &= \sqrt{\frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \\
&= \sqrt{\sqrt{\frac{2k \log(k)}{n}} + \frac{2 \log(\delta^{-1})}{n}} \|\Theta\|^{-1} \|W\|_\infty^{-1}.
\end{aligned}$$

Condition (4.6) on page 67 reads

$$\begin{aligned}
\sqrt{\frac{2k \log(k)}{n}} + \frac{2 \log(\delta^{-1})}{n} &\leq 2 \log(k) \sqrt{\frac{2n \log(k)}{k}} \\
&\iff \delta \geq \exp \left(-n \log(k) \left(\sqrt{\frac{2n \log(k)}{k}} - \sqrt{\frac{k}{2n \log(k)}} \right) \right).
\end{aligned}$$

It is satisfied when $k \geq 2$, $n \geq 2k / \log(k)$ and $\delta \geq \exp(-n \log(k))$. \square

4.1.1. EXPLICIT BOUND IN THE INFORMATION k -MEANS SETTING. As proved in Lemma 12 on page 47, we can use the previous bounds in the kernel space H to analyze the original information k -means minimizer.

PROPOSITION 20 *Assume that*

$$\text{ess sup}_X \left(\int p_X^2 d\nu \right) < +\infty \quad \text{and} \quad \text{ess sup}_X \left(\int \log(p_X)^2 d\nu \right) < +\infty.$$

Consider the information radius

$$R = \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \text{ess sup}_X \mathcal{K}(q, p_X)$$

and the bounds

$$B = \text{ess sup}_X \left(\int p_X^2 d\nu \right)^{1/2} \exp(R)$$

$$\text{and } C = \operatorname{ess\,sup}_X \left(\int \log(p_X)^2 \, d\nu \right)^{1/2}.$$

Note that $R \leq \operatorname{ess\,sup}_X \mathfrak{K}(1, p_X) \leq C$.

Introduce the parameter space

$$\mathfrak{Q}_B = \left\{ q \in \mathbb{L}_{+,1}^1(\nu) \cap \mathbb{L}^2(\nu) : \int q^2 \, d\nu \leq B^2 \right\}.$$

Given (X_1, \dots, X_n) , a sample made of n independent copies of X , with probability at least $1 - \delta$, for any $q \in \mathfrak{Q}_B^k$,

$$\begin{aligned} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) &\leq \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_{X_i}) \\ &\quad + \left(\sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n} \right)^{1/4} \right) (BC + 2 \log(B)). \end{aligned}$$

Consider an empirical risk minimizer $\hat{q}(X_1, \dots, X_n) \in \mathfrak{Q}_B^k$ satisfying

$$\hat{q} \in \arg \min_{q \in \mathfrak{Q}_B^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right).$$

With probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(\hat{q}, p_X) \right) &\leq \inf_{q \in \left(\mathbb{L}_{+,1}^1(\nu) \right)^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) \\ &\quad + \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \right) (BC + 2 \log(B)). \end{aligned}$$

PROOF. The proof consists in applying Lemmas 12 on page 47, 18 on page 65 and 19 on page 69. Introduce $\theta = (q, \mathfrak{K}(q, 1)) \in H$ and $W = (-\log(p_X), \mu^{-1}) \in H$. Let

$$\Theta = \{ \theta \in H : q \in \mathfrak{Q}_B \}.$$

Remark that

$$\|W\|_\infty^2 = \operatorname{ess\,sup}_X \|\log(p_X)\|_\nu^2 + \mu^{-1} \text{ and that } \|\theta\|^2 = \|q\|_\nu^2 + \mu \mathfrak{K}(q, 1)^2.$$

According to Jensen's inequality,

$$\mathfrak{K}(q, 1) = \int q \log(q) \, d\nu \leq \log \left(\int q^2 \, d\nu \right) = \log(\|q\|_\nu^2).$$

Thus $\|\Theta\|_H^2 \leq B^2 + \mu \log(B^2)^2$. Accordingly

$$\begin{aligned} \|\Theta\|^2 \|W\|_\infty^2 &\leq (C^2 + \mu^{-1}) (B^2 + \mu \log(B^2)^2) \\ &= B^2 C^2 + \log(B^2)^2 + \mu^{-1} B^2 + \mu C^2 \log(B^2)^2 = (BC + 2 \log(B))^2, \end{aligned}$$

taking $\mu = \frac{B}{C \log(B^2)}$. The proposition then follows from Lemmas 18 on page 65 and 19 on page 69, in view of Lemma 12 on page 47. \square

4.1.2. CLASSICAL k -MEANS QUANTIZATION IN A SEPARABLE HILBERT SPACE. In this section, we investigate how to adapt the previous proofs in the case of the classical k -means risk in order to obtain non-asymptotic dimension-free bounds. Let us consider a separable Hilbert space $(\mathfrak{X}, \|\cdot\|)$. Let (X_1, \dots, X_n) be n independent copies of a random vector $X \sim \mathbb{P}_X$ such that $\|X\| \leq B$, almost surely with $B > 0$ and introduce the codebook (or k centers) $\mu = (\mu_1, \dots, \mu_k) \in \mathfrak{X}^k$. In this framework, the k -means risk is given by

$$\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right),$$

and its empirical counterpart is

$$\frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \|X_i - \mu_j\|^2.$$

Let us consider the closed ball

$$\mathfrak{B} = \{\mu \in \mathfrak{X} : \|\mu\| \leq B\},$$

then we have the following lemma

LEMMA 21

$$\inf_{\mu_1, \dots, \mu_k \in \mathfrak{X}^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) = \inf_{\mu_1, \dots, \mu_k \in \mathfrak{B}^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right).$$

PROOF. Recall that, for a fixed $\ell : \mathfrak{X} \rightarrow \llbracket 1, k \rrbracket$, we have

$$\mathbb{P}_X \left(\|X - \mu_{\ell(X)}\|^2 \right) = \mathbb{P}_X \left(\|X - \mathbb{P}_{X|\ell(X)}(X)\|^2 \right) + \mathbb{P}_X \left(\|\mathbb{P}_{X|\ell(X)}(X) - \mu_{\ell(X)}\|^2 \right).$$

Thus optimal centers μ_j^* are given by conditional means $\mathbb{P}_{X|\ell(X)=j}(X)$. Notice that, by Jensen's inequality, $\|\mathbb{P}_{X|\ell(X)=j}(X)\| \leq \mathbb{P}_{X|\ell(X)=j}(\|X\|) \leq B$, since $\mathbb{P}_X(\|X\| \leq B) = 1$. Therefore we can restrict the optimization to the case when $\mu_j \in \mathfrak{B}$. \square

Remark also that

$$\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) = \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|\mu_j\|^2 - 2\langle \mu_j, X \rangle + \|X\|^2 \right).$$

In view of this, we can introduce the new separable Hilbert space $H = \mathfrak{X} \times \mathbb{R}^2$ endowed with the inner product

$$\langle h, h' \rangle_H = 2\langle h_1, h'_1 \rangle_{\mathfrak{X}} + B^{-2}h_2 h'_2 + B^2h_3 h'_3.$$

Define

$$\begin{aligned} \theta_j &= (\mu_j, \|\mu_j\|^2, 1) \in \Theta \subset H \\ \text{where } \Theta &= \{(\mu, \|\mu\|^2, 1) : \mu \in \mathfrak{B}\} \\ \text{and } W &= (-X, B^2, B^{-2}\|X\|^2) \in H. \end{aligned}$$

We obtain that $\langle \theta_j, W \rangle = \|X - \mu_j\|^2$, so that

$$\inf_{\mu \in \mathcal{X}^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) = \inf_{\theta \in \Theta^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right),$$

with $\langle \theta_j, W \rangle \geq 0$. Therefore, we are in a situation where we can apply Proposition 18 on page 65 and Proposition 19 on page 69.

Remark that

$$\begin{aligned} \|W\|^2 &= 2\|X\|^2 + B^2 + B^{-2}\|X\|^4, \\ \|\theta\| &= 2\|\mu\|^2 + B^{-2}\|\mu\|^4 + B^2, \end{aligned}$$

so that $\|\Theta\| \|W\|_\infty \leq 4B^2$.

PROPOSITION 22 *Let $X_i, 1 \leq i \leq n$ be a sample made of n independent copies of X . Assume that $k \geq 2$, $n \geq 2k \log(k)$ and $\delta \geq \exp(-n/4)$. With probability at least $1 - \delta$, for any $\mu = (\mu_1, \dots, \mu_k) \in \mathcal{B}^k$,*

$$\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) \leq \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \|X_i - \mu_j\|^2 + 4B^2 \left(\sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n} \right)^{1/4} \right).$$

With probability at least $1 - \delta$, for any $\mu \in \mathcal{B}^k$,

$$\begin{aligned} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) - \inf_{\mu \in \mathcal{B}^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) \\ \leq \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) - \inf_{\mu \in \mathcal{B}^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) \\ + 4B^2 \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \right), \end{aligned}$$

where $\bar{\mathbb{P}}_X = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. In particular, if

$$\hat{\mu}(X_1, \dots, X_n) \in \arg \min_{\mu \in \mathcal{B}^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right),$$

with probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \hat{\mu}_j\|^2 \right) \leq \inf_{\mu \in \mathcal{X}^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) \\ + 4B^2 \left(\sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \right). \end{aligned}$$

4.1.3. DISCUSSION ABOUT THE BOUNDS. We get a bound for the excess risk in the classical k -means setting, and more generally in the information k -means setting, of order

$$\mathcal{O} \left(\frac{\log(k)k}{n} \right)^{1/4}. \quad (4.8)$$

Remark that the information k -means setting is a new extension of the classical k -means setting, where we minimize the Kullback divergence with respect to its first argument. Minimization with respect to the second argument is done instead in [BDG04], within the general framework of Bregman divergence. Besides, [Fis10] develops the approach of [BDG04] using Bregman divergence and provides a non asymptotic bound for the excess risk of the same order as in [BDL08], that is

$$\mathfrak{O}\left(\frac{k}{\sqrt{n}}\right). \quad (4.9)$$

Moreover, [BFL20] proves also a deviation bound of the same order as in [BDL08] in the case of robust k -means with Bregman divergence. However, the dependence in k is not explicit and depends on a constant based on the characteristics of the underlying data distribution. We will prove later on faster bounds of order

$$\mathfrak{O}\left(\log\left(\frac{n}{k}\right)\sqrt{\frac{k \log(k)}{n}}\right). \quad (4.10)$$

Note that (4.9) goes to zero whenever k^2/n goes to zero, whereas (4.10) goes to zero whenever $k \log(k) \log(\log(k))/n$ goes to zero and that (4.8) goes to zero whenever $k \log(k)/n$ goes to zero, so that, although (4.8) is slower than (4.10), it provides a slightly better consistency condition.

In addition, [Lev13] derives a bound of order

$$\mathfrak{O}\left(\frac{k^3}{n}\right),$$

with a better dependence in n and a worse dependence in k , under margin conditions. Margin conditions are beyond the scope of the present study, where we focus on the use of the k -means criterion as a vector quantization tool, a situation where margin conditions are unlikely to hold true. In this perspective, equation (4.8) is the most significant bound, although it is the slowest one with respect to n alone, since it provides the best sufficient condition for consistency (namely that $k \log(k)/n$ tends to zero).

4.2. A BOUNDED CRITERION FOR INFORMATION k -MEANS

In this section, we still consider a couple of random variables $(X, Y) : \Omega \rightarrow \mathfrak{X} \times \mathfrak{Y}$, where \mathfrak{X} and \mathfrak{Y} are two Polish spaces. We assume that for some reference probability measure ν , $\mathbb{P}(\mathbb{P}_{Y|X} \ll \nu) = 1$. Remark that the sigma algebra associated with \mathfrak{Y} is countably generated, so that $\mathbb{L}^2(\nu)$ is separable (see prop 3.4.5 in [Coh]). We let $p_X = \frac{d\mathbb{P}_{Y|X}}{d\nu}$. For any family of probability densities $q_j \in \mathbb{L}_{+,1}^1$, $1 \leq j \leq k$, we consider a classification function

$$\ell(X) \in \arg \min_{j \in [1,k]} \mathfrak{K}(q_j, p_X)$$

and the model

$$\mathfrak{Q}_q = \left\{ Q \in \mathfrak{M}_+^1(\Omega) : Q_{Y|X} \ll \nu \text{ and } \frac{dQ_{Y|X}}{d\nu} = q_{\ell(X)} \right\}.$$

LEMMA 23 For any family $q = (q_1, \dots, q_k) \in (\mathbb{L}_{+,1}^1(\nu))^k$ of k centers, define the bounded criterion

$$\mathfrak{C}(q) = 1 - \exp\left[-\inf_{Q \in \mathfrak{Q}_q} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y})\right].$$

It can be expressed as

$$\mathfrak{C}(q) = 1 - \mathbb{P}_X\left(\exp\left[-\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X)\right]\right).$$

PROOF. Combining Lemma 1 on page 9 and Lemma 6 on page 41, we get

$$\begin{aligned} \inf_{Q \in \mathfrak{Q}_q} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) &= \inf_{Q \in \mathfrak{Q}_q} \left\{ \mathfrak{K}(Q_X, P_X) + Q_X[\mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X})] \right\} \\ &= \min_{q_1, \dots, q_k} \left\{ -\log \mathbb{P}_X\left(\exp\left[-\mathfrak{K}(q_{\ell(X)}, p_X)\right]\right) \right\} = -\log \mathbb{P}_X\left(\exp\left[-\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X)\right]\right). \end{aligned}$$

This completes the proof. \square

From the Aronszajn theorem ([Aro50]), there is a mapping $\Psi : \mathbb{L}^2(\nu) \rightarrow H$ to a separable Hilbert space H such that

$$\langle \Psi(f), \Psi(g) \rangle_H = \exp\left(\int fg \, d\nu\right), \quad f, g \in \mathbb{L}^2(\nu).$$

Notice that H is separable since $K(f, g) = \exp\left(\int fg \, d\nu\right)$ is a continuous kernel (see lemma 4.33 in [CS08]). Introduce a positive real parameter μ to be chosen afterwards and remark that

$$\exp\left(-\mathfrak{K}(q_j, p_X)\right) = \left\langle \exp(-\mathfrak{K}(q_j, 1)) \Psi(\mu q_j), \Psi(\mu^{-1} \log(p_X)) \right\rangle_H$$

Accordingly, if we let $\theta_j = \exp(-\mathfrak{K}(q_j, 1)) \Psi(\mu q_j)$ and $W = \Psi(\mu^{-1} \log(p_X))$, we obtain that

$$\mathfrak{C}(q) = 1 - \mathbb{P}_X\left(\max_{j \in \llbracket 1, k \rrbracket} \underbrace{\langle \theta_j, W \rangle_H}_{\in [0,1]}\right).$$

To proceed, we need the following lemma in kernel space H .

LEMMA 24 Consider a separable Hilbert space H and a random vector $W \in H$. Let W_1, \dots, W_n be a sample made of n independent copies of W and let $\Theta \subset H$ be a bounded set of parameters.

Assume that $\mathbb{P}_W(\langle \theta, W \rangle \in [0, 1], \theta \in \Theta) = 1$.

For any number of centers $k \geq 2$, any sample size $n \geq 4k\|\Theta\|^2\|W\|_\infty^2/\log(k)$ and any probability level $\delta \geq \exp(-\log(k)n/\sqrt{2})$, with probability at least $1 - \delta$, for any $\theta \in \Theta^k$,

$$\bar{\mathbb{P}}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) \leq \mathbb{P}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) + \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n}\right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2}.$$

For any number of centers $k \geq 2$, any sample size $n \geq k\|\Theta\|^2\|W\|_\infty^2/\log(k)$, any probability level $\delta \geq \exp(-n\log(k)/\sqrt{2})$, any non random family of centers $\theta^* \in \Theta^k$, with probability at least $1 - \delta$, for any $\theta \in \Theta^k$

$$(\bar{\mathbb{P}}_W - \mathbb{P}_W)\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle - \max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle\right) \leq \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{32 k \log(k)}{n}\right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2}.$$

Consequently if

$$\hat{\theta}(W_1, \dots, W_n) \in \arg \max_{\theta \in \Theta^k} \bar{\mathbb{P}}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right)$$

with probability at least $1 - \delta$,

$$\begin{aligned} \sup_{\theta \in \Theta^k} \mathbb{P}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) &\leq \mathbb{P}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \hat{\theta}_j, W \rangle\right) \\ &\quad + \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{32 k \log(k)}{n}\right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2}. \end{aligned}$$

PROOF. The proof is almost the same as the proofs of Propositions 18 on page 65 and 19 on page 69. Reasoning in the same way, we see that

$$\begin{aligned} \int \log \mathbb{P}_W \exp\left(\lambda \max_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, W \rangle\right) d\rho_\theta(\theta') &\leq \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \log \mathbb{P}_W \exp\left(\lambda \max_{j \in \llbracket 1, k \rrbracket} \underbrace{\langle \theta_j, W \rangle}_{\in [0,1]}\right) \\ &\leq \lambda \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \lambda \mathbb{P}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) + \frac{\lambda^2}{8}, \end{aligned}$$

under condition (4.6) on page 67. Remarking that moreover

$$\int \bar{\mathbb{P}}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta'_j, \Phi(W) \rangle\right) d\rho_\theta(\theta') \geq \bar{\mathbb{P}}_W\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right),$$

we deduce that

$$\begin{aligned} (\bar{\mathbb{P}}_W - \mathbb{P}_W)\left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle\right) &\leq \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{\lambda}{8} + \frac{k\beta\|\Theta\|^2 + 2 \log(\delta^{-1})}{2n\lambda} \\ &\leq \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \sqrt{\frac{k\beta\|\Theta\|^2 + 2 \log(\delta^{-1})}{4n}} \\ &\leq \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n}\right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2}, \end{aligned}$$

where we have chosen

$$\begin{aligned} \beta &= \sqrt{\frac{8n \log(k)}{k}} \|\Theta\|^{-1} \|W\|_\infty, \\ \lambda &= 2 \sqrt{\frac{k\beta\|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \\ &= 2 \sqrt{\sqrt{\frac{8k \log(k)}{n}} \|\Theta\| \|W\|_\infty + \frac{2 \log(\delta^{-1})}{n}}. \end{aligned}$$

Accordingly condition (4.6) on page 67 reads

$$\begin{aligned} \sqrt{\frac{8k \log(k)}{n}} \|\Theta\| \|W\|_\infty + \frac{2 \log(\delta^{-1})}{n} &\leq \frac{1}{2} \log(k) \sqrt{\frac{8n \log(k)}{k}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \\ \iff \delta &\geq \exp\left(-n \log(k) \sqrt{\frac{n \log(k)}{2k}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \left(1 - \frac{2k}{n \log(k)} \|\Theta\|^2 \|W\|_\infty^2\right)\right). \end{aligned}$$

Therefore the condition is satisfied when $k \geq 2$, $n \geq 4k \|\Theta\|^2 \|W\|_\infty^2 / \log(k)$ and $\delta \geq \exp(-n \log(k) / \sqrt{2})$. This completes the proof of the first part of the lemma. To prove the second part, remark in the same way that

$$\begin{aligned} (\bar{\mathbb{P}}_W - \mathbb{P}_W) \left(\underbrace{\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle - \max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle}_{\in [-1, 1]} \right) \\ \leq \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \frac{\lambda}{2} + \frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{2n\lambda} \\ \leq \sqrt{\frac{2 \log(k)}{\beta}} \|W\|_\infty + \sqrt{\frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \\ \leq \sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2}, \end{aligned}$$

where we have set

$$\begin{aligned} \beta &= \sqrt{\frac{2 \log(k) n}{k}} \|W\|_\infty \|\Theta\|^{-1} \\ \text{and } \lambda &= \sqrt{\frac{k\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n}} \\ &= \sqrt{\sqrt{\frac{2k \log(k)}{n}} \|\Theta\| \|W\|_\infty + \frac{2 \log(\delta^{-1})}{n}}. \end{aligned}$$

We have also to satisfy condition (4.6) on page 67 that reads

$$\begin{aligned} \sqrt{\frac{2k \log(k)}{n}} \|\Theta\| \|W\|_\infty + \frac{2 \log(\delta^{-1})}{n} &\leq 2 \log(k) \sqrt{\frac{2n \log(k)}{k}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \\ \iff \delta &\geq \exp\left(-n \log(k) \sqrt{\frac{2n \log(k)}{k}} \|\Theta\|^{-1} \|W\|_\infty^{-1} \left(1 - \frac{k}{2n \log(k)} \|\Theta\|^2 \|W\|_\infty^2\right)\right). \end{aligned}$$

It is satisfied when $k \geq 2$, $n \geq k \|\Theta\|^2 \|W\|_\infty^2 / \log(k)$ and $\delta \geq \exp(-n \log(k) / \sqrt{2})$. \square

We need now an equivalent of Lemma 12 on page 47 for the bounded criterion defined in Lemma 23 on page 76.

LEMMA 25 *Assume that $\text{ess sup}_X \int p_X^2 d\nu < +\infty$ and $\text{ess sup}_X \int \log(p_X)^2 d\nu < +\infty$. Consider the information radius*

$$R = \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \text{ess sup}_X \mathfrak{K}(q, p_X)$$

and the bounds

$$B = \operatorname{ess\,sup}_X \left(\int p_X^2 d\nu \right)^{1/2} \exp(R),$$

$$C = \operatorname{ess\,sup}_X \left(\int \log(p_X)^2 d\nu \right)^{1/2}$$

Note that $R \leq \operatorname{ess\,sup}_X \mathfrak{K}(1, p_X) \leq C < +\infty$, so that $B < +\infty$. Consider the set

$$\mathfrak{B} = \{q \in \mathbb{L}_{+,1}^1(\nu) : \int q^2 d\nu \leq B^2\}$$

The minimization of the bounded criterion $\mathfrak{C}(q)$ defined in Lemma 23 on page 76 can be restricted to \mathfrak{B} , in the sense that

$$\inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathfrak{C}(q) = \inf_{q \in \mathfrak{B}^k} \mathfrak{C}(q).$$

PROOF. Remark that

$$\begin{aligned} \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathfrak{C}(q) &= \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \left\{ 1 - \exp \left[- \inf_{Q_X} \left(\mathfrak{K}(Q_X, \mathbb{P}_X) + Q_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) \right) \right] \right\} \\ &= 1 - \exp \left\{ - \inf_{Q_X} \left[\mathfrak{K}(Q_X, \mathbb{P}_X) + \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} Q_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) \right] \right\} \end{aligned}$$

and apply Lemma 12 on page 47 to restrict the minimization to $q \in \mathfrak{B}^k$. \square

We have seen that

$$\mathfrak{C}(q) = 1 - \mathbb{P}_X \left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle_H \right),$$

where $\theta_j = \exp(-\mathfrak{K}(q_j, 1)) \Psi(\mu q_j)$ and $W = \Psi(\mu^{-1} \log(p_X))$. Note that for any $q \in \mathfrak{B}^k$,

$$\begin{aligned} \|\theta_j\|^2 &\leq \|\Psi(\mu q_j)\|^2 = \exp(\mu^2 \|q_j\|_\nu^2) \leq \exp(\mu^2 B^2) \\ \text{and } \|W\|^2 &= \exp(\mu^{-2} \|\log(p_X)\|_\nu^2) \leq \exp(\mu^{-2} C^2), \\ \text{so that } \|\theta_j\|^2 \|W\|^2 &\leq \exp(\mu^2 B^2 + \mu^{-2} C^2) = \exp(2BC), \end{aligned}$$

if we choose $\mu = \sqrt{C/B}$. In view of Lemma 24 on page 76, this leads to

PROPOSITION 26 *In the situation described at the beginning of this section, consider a sample X_1, \dots, X_n made of n independent copies of X . For any family of k probability densities $q \in (\mathbb{L}_{+,1}^1(\nu))^k$, consider the bounded criterion $\mathfrak{C}(q)$ defined in Lemma 23 on page 76 and its empirical counterpart*

$$\bar{\mathfrak{C}}(q) = 1 - \bar{\mathbb{P}}_X \left(\exp \left[- \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right] \right).$$

Assume that $k \geq 2$, $n \geq 4k \exp(2BC) / \log(k)$ and $\delta \geq \exp(-\log(k)n/\sqrt{2})$. With probability at least $1 - \delta$, for any $q \in \mathfrak{B}^k$,

$$\mathfrak{C}(q) \leq \bar{\mathfrak{C}}(q) + \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n} \right)^{1/4} \exp(BC/2).$$

Let $q^* \in \mathcal{B}^k$ be a non random family of centers, and assume that $k \geq 2$, $n \geq k \exp(2BC)/\log(k)$, and $\delta \geq \exp(-n \log(k)/\sqrt{2})$. With probability at least $1 - \delta$, for any $q \in \mathcal{B}^k$,

$$\mathfrak{C}(q) - \mathfrak{C}(q^*) \leq \bar{\mathfrak{C}}(q) - \bar{\mathfrak{C}}(q^*) + \sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \exp(BC/2).$$

Consequently, if

$$\hat{q}(X_1, \dots, X_n) \in \arg \min_{q \in \mathcal{B}^k} \bar{\mathfrak{C}}(q),$$

with probability at least $1 - \delta$,

$$\mathfrak{C}(\hat{q}) \leq \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathfrak{C}(q) + \sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4} \exp(BC/2).$$

4.3. A BOUNDED CRITERION FOR THE EUCLIDEAN k -MEANS

In this section we are given a random vector $X \in \mathbb{R}^d$ and a sample X_1, \dots, X_n made of n independent copies of X . Introduce the random variable Y whose joint distribution with X is given by $\mathbb{P}_{Y|X} = \mathcal{N}(X, \sigma^2 I_d)$, and consider a family of k centers μ_1, \dots, μ_k . Consider a nearest neighbour classification function $\ell : \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket$ that satisfies therefore $\|x - \mu_{\ell(x)}\| = \min_{j \in \llbracket 1, k \rrbracket} \|x - \mu_j\|$. Consider the conditional probability measure $Q_{Y|X}$ defined as $Q_{Y|X} = \mathcal{N}(\mu_{\ell(X)}, \sigma^2 I_d)$. Remark that the k -means Euclidean criterion can be written as

$$\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 = 2\sigma^2 \mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X}).$$

Inspired by this identity, we can introduce another loss function with the help of the model

$$\mathfrak{Q}_\mu = \left\{ Q_{X,Y} : Q_{Y|X} = \mathcal{N}(\mu_{\ell(X)}, \sigma^2 I_d) \right\}.$$

It is based on the identity

$$\begin{aligned} \inf_{Q \in \mathfrak{Q}_\mu} \mathfrak{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) &= -\log \mathbb{P}_X \left(-\mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X}) \right) \\ &= -\log \mathbb{P}_X \exp \left(-\frac{1}{2\sigma^2} \min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right). \end{aligned}$$

In view of this, one may be incited to introduce the loss function

$$\mathfrak{C}(\mu) = 1 - \exp \left(- \inf_{Q \in \mathfrak{Q}_\mu} \mathfrak{K}(Q_{Y|X}, \mathbb{P}_{Y|X}) \right) = \mathbb{P}_X \left[\underbrace{1 - \exp \left(-\frac{1}{2\sigma^2} \min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right)}_{\in [0,1]} \right]. \quad (4.11)$$

It is the expectation of a loss function ranging in the unit interval and the corresponding empirical risk function is

$$\bar{\mathfrak{C}}(\mu) = \bar{\mathbb{P}}_X \left[1 - \exp \left(-\frac{1}{2\sigma^2} \min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right) \right].$$

Observing also that

$$\mathfrak{C}(\mu) = 1 - \max_{j \in \llbracket 1, k \rrbracket} \mathbb{P}_X \left(\exp \left(-\frac{1}{2\sigma^2} \|X - \mu_j\|^2 \right) \right),$$

from the Aronszajn theorem, there is a mapping $\Psi : \mathbb{R}^d \rightarrow H$ to a separable Hilbert space H such that

$$\begin{aligned} \mathfrak{C}(\mu) &= 1 - \mathbb{P}_X \left(\max_{j \in \llbracket 1, k \rrbracket} \langle \Psi(X), \Psi(\mu_j) \rangle \right) \\ &= 1 - \mathbb{P}_X \left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right), \end{aligned}$$

where $\theta_j = \Psi(\mu_j)$ and $W = \Psi(X)$. Due to the Gaussian kernel, it is interesting to remark that Ψ maps the whole Euclidean space \mathbb{R}^d to the unit sphere of H . Therefore, we will not need any boundedness or even integrability assumption on X to be in a situation to apply Lemma 24 on page 76.

PROPOSITION 27 *Consider any $k \geq 2$, any $n \geq 4k/\log(k)$ and any $\delta \geq \exp(-n \log(k)/\sqrt{2})$. With probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,*

$$\mathfrak{C}(\mu) \leq \bar{\mathfrak{C}}(\mu) + \sqrt{\frac{\log(\delta^{-1})}{2n}} + \left(\frac{8k \log(k)}{n} \right)^{1/4}.$$

Assume now that $k \geq 2$, $n \geq k/\log(k)$ and $\delta \geq \exp(-n \log(k)/\sqrt{2})$. For any non random family of centers $\mu^ \in \mathbb{R}^{d \times k}$, with probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,*

$$\mathfrak{C}(\mu) - \mathfrak{C}(\mu^*) \leq \bar{\mathfrak{C}}(\mu) - \bar{\mathfrak{C}}(\mu^*) + \sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4}.$$

Consequently, if

$$\hat{\mu}(X_1, \dots, X_n) \in \arg \min_{\mu \in \mathbb{R}^{d \times k}} \bar{\mathfrak{C}}(\mu),$$

with probability at least $1 - \delta$,

$$\mathfrak{C}(\hat{\mu}) \leq \inf_{\mu \in \mathbb{R}^{d \times k}} \mathfrak{C}(\mu) + \sqrt{\frac{2 \log(\delta^{-1})}{n}} + \left(\frac{32 k \log(k)}{n} \right)^{1/4}.$$

Note that the risk $\mathfrak{C}(\mu)$ can be defined by the right-hand side of equation (4.11) in the case when X belongs to a separable Hilbert space and that in this case also Lemma 24 applies and Proposition 27 holds true, although in this case the definition of $\mathbb{P}_{Y|X}$ requires the construction of a Gaussian process similar to (4.5) on page 66.

4.4. PAC-BAYESIAN BOUNDS FOR INFORMATION FRAGMENTATION

In this section, we consider a random signal $X \in \mathbb{R}^d$, a statistical sample X_1, \dots, X_n made of n independent copies of X and a set of k fragments $\mu_j \in \mathbb{R}^d$, $1 \leq j \leq k$ that we would like to optimize to get the best possible approximation of X in terms of quantization.

More precisely, letting B_j be the support of μ_j , defined as

$$B_j = \{s \in \llbracket 1, d \rrbracket : \mu_{j,s} \neq 0\}, \quad 1 \leq j \leq k,$$

introduce for each $K \leq k$ the family of subsets

$$\mathcal{T}_{\mu,K} = \left\{ A \subset \llbracket 1, k \rrbracket : |A| \leq K \text{ and } B_i \cap B_j = \emptyset, i \neq j \in A \right\}.$$

We are willing to approximate X by $\sum_{j \in A} \mu_j$ for the best possible choice of $A \in \mathcal{T}_{\mu,K}$ depending on X . This can be seen as a structured instance of k -means, where the set of centers is

$$\sum_{j \in A} \mu_j, \quad A \in \mathcal{T}_{\mu,K}.$$

Some care should be taken though, due to the fact that $\mathcal{T}_{\mu,K}$ depends on μ and is therefore not a constant size index set. Besides, we will take advantage of the special structure of this fragment k -means problem to derive specific generalization bounds.

Let us first address the fact that $\mathcal{T}_{\mu,K}$ depends on μ . To circumvent this, we will describe the possible values of $\mathcal{T}_{\mu,K}$ first, and then we will describe the range of values of μ that corresponds to each value of $\mathcal{T}_{\mu,K}$. In other words we will condition our description of the parameter space on the value of $\mathcal{T}_{\mu,K}$.

Remark that the value of $\mathcal{T}_{\mu,K}$ depends only on the symmetric graph of the intersections of the supports

$$\mathcal{J}_\mu = \left\{ (i, j) \in \llbracket 1, k \rrbracket^2 : B_i \cap B_j \neq \emptyset \right\}.$$

Therefore, we can describe the problem in the following way. Consider the set \mathcal{G} of symmetric and reflexive graphs on the set of vertices $\llbracket 1, k \rrbracket$ (we mean by reflexive that the graph contains the diagonal of $\llbracket 1, k \rrbracket^2$). There are $2^{k(k-1)/2}$ such graphs (there are $\binom{k}{2}$ distinct pairs of vertices, so we can make $2^{\binom{k}{2}}$ possible connections), so that

$$\log(|\mathcal{G}|) = \frac{k(k-1)}{2} \log(2).$$

For each $g \in \mathcal{G}$, and each maximum number of components $K \leq k$, introduce the set of subsets

$$\mathcal{T}_{g,K} = \left\{ A \subset \llbracket 1, k \rrbracket : |A| \leq K \text{ and } A^2 \cap g = \mathbf{diag}(A^2) \right\}.$$

(In other words the set of disconnected subsets of at most K vertices.) Consider the fragment model

$$\mathcal{M}_g = \left\{ \mu \in \mathbb{R}^{d \times k} : \mathcal{J}_\mu \subset g \right\}.$$

We see that $\mathcal{T}_{g,K} \subset \mathcal{T}_{\mu,K}$, that $\mathcal{J}_{\mathcal{J}_\mu} = \mathcal{J}_\mu$ and that $\mu \in \mathcal{M}_{\mathcal{J}_\mu}$. Therefore

$$\left\{ \left(\sum_{j \in A} \mu_j, A \in \mathcal{T}_{\mu,K} \right) : \mu \in \mathbb{R}^{d \times k} \right\} \subset \left\{ \left(\sum_{j \in A} \mu_j, A \in \mathcal{T}_{g,K} \right) : g \in \mathcal{G}, \mu \in \mathcal{M}_g \right\}.$$

Let us now describe possible risk functions for our problem. We assume in the following discussion that the parameters k and K are fixed. The most obvious risk is the Euclidean criterion

$$\mathfrak{C}_1(g, \mu) = \frac{1}{d} \mathbb{P}_X \left(\min_{A \in \mathfrak{T}_{g,K}} \left\| X - \sum_{j \in A} \mu_j \right\|^2 \right), \quad g \in \mathfrak{G}, \mu \in \mathfrak{M}_g.$$

We normalize by $1/d$ to scale nicely with the dimension of the signal. Introducing the representation of the problem by the triplet of random variables (X, S, V) , where $\mathbb{P}_{S|X}(s) = 1/d$ and $\mathbb{P}_{V|S,X} = \mathcal{N}(X_S, \sigma^2)$, we can rewrite \mathfrak{C}_1 as

$$\begin{aligned} \mathfrak{C}_1(g, \mu) &= \mathbb{P}_X \left[\min_{A \in \mathfrak{T}_{g,K}} \mathbb{P}_S \left((X_S - \sum_{j \in A} \mu_{j,S})^2 \right) \right] \\ &= 2\sigma^2 \mathbb{P}_X \left[\min_{A \in \mathfrak{T}_{g,K}} \mathbb{P}_S \left(\mathfrak{K}(Q_{V|S}^{(\mu, A)}, \mathbb{P}_{V|X,S}) \right) \right] \\ &= 2\sigma^2 \inf_{Q \in \mathfrak{Q}_1(g, \mu)} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}), \end{aligned}$$

$$\text{where } Q_{V|S}^{(\mu, A)} = \mathcal{N}\left(\sum_{j \in A} \mu_{j,S}, \sigma^2\right),$$

$$\text{and where } \mathfrak{Q}_1(g, \mu) = \left\{ Q_{X,S,V} : Q_{X,S} = \mathbb{P}_{X,S}, Q_{V|X,S} = Q_{V|S}^{(\mu, A(X))}, A(X) \in \mathfrak{T}_{g,K} \right\}.$$

Based on this interpretation, we can define smaller risk functions by relaxing the constraint that $Q_{X,S} = \mathbb{P}_{X,S}$. For each $\mu \in \mathfrak{M}_g$ and each $A \in \mathfrak{T}_{g,K}$, choose a classification function

$$\ell_{\mu,A} : \llbracket 1, d \rrbracket \longrightarrow A$$

such that

$$\text{supp}(\mu_j) \subset \ell_{\mu,A}^{-1}(j), \quad j \in A.$$

Note that

$$\sum_{j \in A} \mu_{j,S} = \mu_{\ell_{\mu,A}(S), S}, \quad s \in \llbracket 1, d \rrbracket.$$

Define

$$B_j = \bigcap_{\substack{A \in \mathfrak{T}_{g,K} \\ : j \in A}} \ell_{\mu,A}^{-1}(j), \quad j \in \llbracket 1, k \rrbracket.$$

and remark that

$$\text{supp}(\mu_j) \subset B_j, \quad j \in \llbracket 1, k \rrbracket.$$

Let us introduce the models

$$\mathfrak{Q}_2(g, \mu) = \left\{ Q_{X,S,V} : Q_{S|X} = \mathbb{P}_S, Q_{V|X,S} = Q_{V|S}^{(A(X), \mu)}, A(X) \in \mathfrak{T}_{g,K} \right\}, \quad (Q_X \text{ is free})$$

$$\begin{aligned} \mathfrak{Q}_3(g, \mu) &= \left\{ Q_{X,S,V} : Q_{S|X, \ell(\mu, A(X), S)=j} = \mathbb{P}_{S|S \in B_j}, Q_{V|X,S} = Q_{V|S}^{(A(X), \mu)}, A(X) \in \mathfrak{T}_{g,K} \right\}, \\ &\quad (Q_{X, \ell(\mu, A(X), S)} \text{ is free}) \end{aligned}$$

$$\mathfrak{Q}_4(g, \mu) = \left\{ Q_{X,S,V} : Q_{S|X} \left(\bigcup_{j \in A(X)} B_j \right) = 1, Q_{V|X,S} = Q_{V|S}^{(A(X), \mu)}, A(X) \in \mathfrak{T}_{g,K} \right\},$$

$(Q_{X,S} | S \in \bigcup_{j \in A(X)} B_j \text{ is free})$ and the risk functions

$$\mathfrak{C}_t(g, \mu) = 2\sigma^2 \left[1 - \exp \left(- \inf_{Q \in \mathfrak{Q}_t(g, \mu)} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}) \right) \right], \quad t \in \llbracket 2, 4 \rrbracket.$$

LEMMA 28 *For any $g \in \mathfrak{G}$ and any $\mu \in \mathfrak{M}_g$,*

$$\mathfrak{C}_2(g, \mu) \leq \mathfrak{C}_1(g, \mu)$$

and

$$\mathfrak{C}_4(g, \mu) \leq \mathfrak{C}_3(g, \mu).$$

When the fragmentation is complete in the sense that

$$\bigcup_{j \in A} B_j = \llbracket 1, d \rrbracket, \quad A \in \overline{\mathfrak{T}}_{g,K},$$

where

$$\overline{\mathfrak{T}}_{g,K} = \left\{ A \in \mathfrak{T}_{g,K} : A \subset A' \in \mathfrak{T}_{g,K} \Rightarrow A' = A \right\}$$

are the maximal sets of $\mathfrak{T}_{g,K}$, then

$$\mathfrak{C}_4(g, \mu) \leq \mathfrak{C}_3(g, \mu) \leq \mathfrak{C}_2(g, \mu) \leq \mathfrak{C}_1(g, \mu).$$

PROOF. When the fragmentation is complete

$$\mathfrak{Q}_1(g, \mu) \subset \mathfrak{Q}_2(g, \mu) \subset \mathfrak{Q}_3(g, \mu) \subset \mathfrak{Q}_4(g, \mu),$$

implying that

$$\begin{aligned} \mathfrak{C}_4(g, \mu) \leq \mathfrak{C}_3(g, \mu) \leq \mathfrak{C}_2(g, \mu) &\leq 2\sigma^2 \left[1 - \exp \left(- \inf_{Q \in \mathfrak{Q}_1(g, \mu)} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}) \right) \right] \\ &\leq 2\sigma^2 \inf_{Q \in \mathfrak{Q}_1(g, \mu)} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}) = \mathfrak{C}_1(g, \mu). \end{aligned}$$

In the general case, the central inclusion $\mathfrak{Q}_2(g, \mu) \subset \mathfrak{Q}_3(g, \mu)$ does not hold and we cannot compare \mathfrak{C}_1 or \mathfrak{C}_2 with \mathfrak{C}_3 or \mathfrak{C}_4 . \square

LEMMA 29 *For any $g \in \mathfrak{G}$ and any $\mu \in \mathfrak{M}_g$,*

$$\begin{aligned} \mathfrak{C}_2(g, \mu) &= 2\sigma^2 \mathbb{P}_X \left[1 - \max_{A \in \mathfrak{T}_{g,K}} \exp \left(- \frac{1}{2\sigma^2} \mathbb{P}_S \left[\left(X_S - \sum_{j \in A} \mu_{j,S} \right)^2 \right] \right) \right], \\ \mathfrak{C}_3(g, \mu) &= 2\sigma^2 \mathbb{P}_X \left[1 - \max_{A \in \mathfrak{T}_{g,K}} \sum_{j \in A} \mathbb{P}_S(B_j) \exp \left(- \frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} \left[(X_S - \mu_{j,S})^2 \right] \right) \right], \\ \mathfrak{C}_4(g, \mu) &= 2\sigma^2 \mathbb{P}_X \left[1 - \max_{A \in \mathfrak{T}_{g,K}} \mathbb{P}_S \left(\sum_{j \in A} \mathbb{1}(S \in B_j) \exp \left[- \frac{1}{2\sigma^2} (X_S - \mu_{j,S})^2 \right] \right) \right]. \end{aligned}$$

We define the empirical counterparts of our three risk functions as

$$\bar{\mathcal{C}}_1(g, \mu) = \bar{\mathbb{P}}_X \left[\min_{A \in \mathcal{T}_{g,K}} \mathbb{P}_S \left((X_S - \sum_{j \in A} \mu_{j,S})^2 \right) \right], \quad (4.12)$$

$$\bar{\mathcal{C}}_2(g, \mu) = 2\sigma^2 \bar{\mathbb{P}}_X \left[1 - \max_{A \in \mathcal{T}_{g,K}} \exp \left(-\frac{1}{2\sigma^2} \mathbb{P}_S \left[(X_S - \sum_{j \in A} \mu_{j,S})^2 \right] \right) \right], \quad (4.13)$$

$$\bar{\mathcal{C}}_3(g, \mu) = 2\sigma^2 \bar{\mathbb{P}}_X \left[1 - \max_{A \in \mathcal{T}_{g,K}} \sum_{j \in A} \mathbb{P}_S(B_j) \exp \left(-\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} \left[(X_S - \mu_{j,S})^2 \right] \right) \right], \quad (4.14)$$

$$\bar{\mathcal{C}}_4(g, \mu) = 2\sigma^2 \bar{\mathbb{P}}_X \left[1 - \max_{A \in \mathcal{T}_{g,K}} \mathbb{P}_S \left(\sum_{j \in A} \mathbb{1}(S \in B_j) \exp \left[-\frac{1}{2\sigma^2} (X_S - \mu_{j,S})^2 \right] \right) \right]. \quad (4.15)$$

We put $\mathcal{C}_t(\mu) = \mathcal{C}_t(\mathcal{J}_\mu, \mu)$ and consider the four optimization problems

$$\inf \left\{ \mathcal{C}_t(g, \mu) : g \in \mathcal{G}, \mu \in \mathcal{M}_g \right\} = \inf \left\{ \mathcal{C}_t(\mu) : \mu \in \mathbb{R}^{d \times k} \right\}, \quad 1 \leq t \leq 4,$$

based on the observation of the sample (X_1, \dots, X_n) , assuming that the data distribution \mathbb{P}_X is unknown to the statistician.

We can readily state generalization bounds for the first two criteria, \mathcal{C}_1 and \mathcal{C}_2 , using results for the k -means algorithm for each value of g and taking a union bound. By union bound, we mean that if $\mathfrak{B}(g, \delta)$ denotes an excess risk bound (whether for \mathcal{C}_1 or \mathcal{C}_2), for a fixed $g \in \mathcal{G}$, such that $\mathbb{P}(\mathfrak{B}(g, \delta)) \geq 1 - \delta$. Then, by the union bound we get that

$$\mathbb{P} \left(\bigcap_{g \in \mathcal{G}} \mathfrak{B}(g, \delta/|\mathcal{G}|) \right) \geq 1 - \sum_{g \in \mathcal{G}} \mathbb{P}(\Omega \setminus \mathfrak{B}(g, \delta/|\mathcal{G}|)) \geq 1 - \delta,$$

where in our case $|\mathcal{G}| = 2^{k(k-1)/2}$.

Proceeding in this way and using Propositions 22 on page 74 and 27 on page 81 leads to the following proposition.

PROPOSITION 30 *Define $\bar{\mathcal{C}}_j(\mu) = \bar{\mathcal{C}}_j(\mathcal{J}_\mu, \mu)$. Assume that*

$$\mathbb{P}_X \left(\sup_{s \in \llbracket 1, d \rrbracket} |X_s| \leq B \right) = 1.$$

Consider any $k \geq 2$ and any $\delta \geq \exp(-n/4)$. With probability at least $1 - \delta$, for any $\mu \in [-B, B]^{d \times k}$,

$$\mathcal{C}_1(\mu) \leq \bar{\mathcal{C}}_1(\mu) + 4B^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{4n}} + \left(\frac{8|\mathcal{T}_{\mu,K}| \log(|\mathcal{T}_{\mu,K}|)}{n} \right)^{1/4} \right).$$

For any non random set of k fragments $\mu^ \in [-B, B]^{d \times k}$, with probability at least $1 - \delta$, for any $\mu \in [-B, B]^{d \times k}$,*

$$\begin{aligned} \mathcal{C}_1(\mu) - \mathcal{C}_1(\mu^*) - \bar{\mathcal{C}}_1(\mu) + \bar{\mathcal{C}}_1(\mu^*) \\ \leq 4B^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32|\mathcal{T}_{\mu,K}| \log(|\mathcal{T}_{\mu,K}|)}{n} \right)^{1/4} \right). \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in [-B, B]^{d \times k}} \bar{\mathfrak{C}}_1(\mu) + 4B^2 \left(\frac{32 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4},$$

with probability at least $1 - \delta$,

$$\mathfrak{C}_1(\hat{\mu}) \leq \inf_{\mu \in \mathbb{R}^{d \times k}} \mathfrak{C}_1(\mu) + 4B^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4} \right).$$

PROPOSITION 31 Consider any $k \geq 2$ and any $\delta \geq \exp(-n/4)$. With probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,

$$\mathfrak{C}_2(\mu) \leq \bar{\mathfrak{C}}_2(\mu) + 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{4n}} + \left(\frac{8 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4} \right).$$

For any non random set of k fragments $\mu^* \in \mathbb{R}^{d \times k}$, with probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,

$$\begin{aligned} \mathfrak{C}_2(\mu) - \mathfrak{C}_2(\mu^*) - \bar{\mathfrak{C}}_2(\mu) + \bar{\mathfrak{C}}_2(\mu^*) \\ \leq 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4} \right). \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathbb{R}^{d \times k}} \bar{\mathfrak{C}}_2(\mu) + 2\sigma^2 \left(\frac{32 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4},$$

with probability at least $1 - \delta$,

$$\mathfrak{C}_2(\hat{\mu}) \leq \inf_{\mu \in \mathbb{R}^{d \times k}} \mathfrak{C}_2(\mu) + 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 |\mathfrak{T}_{\mu, K}| \log(|\mathfrak{T}_{\mu, K}|)}{n} \right)^{1/4} \right).$$

The previous bounds depend on the factor $|\mathfrak{T}_{\mu, K}|$, that is of order 2^k , and therefore too large in many situations. To get rid of this term, we will use the following lemma, that is specific to the fragmentation setting.

LEMMA 32 Let $W = (W_j, 1 \leq j \leq k)$ be a random vector in the product H^k , where H is a separable Hilbert space (that we can take as being ℓ_2 if we want). Consider a sample $(W^{(1)}, \dots, W^{(n)})$ made of n independent copies of W . Consider a bounded parameter set $\Theta \subset H^k$ and a set \mathfrak{T} of subsets of $\llbracket 1, k \rrbracket$. Assume that

$$\mathbb{P}_W \left(\sum_{j \in A} \langle \theta_j, W_j \rangle \in [a, b], \quad A \in \mathfrak{T}, \theta \in \Theta \right) = 1.$$

Consider the risk

$$\mathfrak{C}(\theta) = \mathbb{P}_W \left(\min_{A \in \mathfrak{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right), \quad \theta \in \Theta,$$

and its empirical counterpart

$$\bar{\mathfrak{C}}(\theta) = \bar{\mathbb{P}}_W \left(\min_{A \in \mathfrak{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right), \quad \theta \in \Theta$$

Put

$$K(\mathfrak{T}) = \max_{A \in \mathfrak{T}} |A|, \quad \|\Theta\| = \sup_{\theta \in \Theta} \left(\sum_{j=1}^k \|\theta_j\|^2 \right)^{1/2} \text{ and } \|W\|_\infty = \max_{j \in \llbracket 1, k \rrbracket} \operatorname{ess\,sup}_{\mathbb{P}_W} \|W_j\|.$$

Assume that $|\mathfrak{T}| \geq 3$ and that $\delta \geq \exp[-n \log(|\mathfrak{T}|)/\sqrt{2}]$. With probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\mathfrak{C}(\theta) \leq \bar{\mathfrak{C}}(\theta) + \sqrt{\frac{\log(\delta^{-1})}{2n}}(b-a) + \left(\frac{8K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2} (b-a)^{1/2}.$$

Consider any non random value of the parameter $\theta^* \in \Theta$. With probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned} \mathfrak{C}(\theta) - \mathfrak{C}(\theta^*) &\leq \bar{\mathfrak{C}}(\theta) - \bar{\mathfrak{C}}(\theta^*) \\ &\quad + \sqrt{\frac{2 \log(\delta^{-1})}{n}}(b-a) + \left(\frac{32 K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2} (b-a)^{1/2}. \end{aligned}$$

Consequently, if

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \bar{\mathfrak{C}}(\theta),$$

with probability at least $1 - \delta$,

$$\mathfrak{C}(\hat{\theta}) \leq \inf_{\theta \in \Theta} \mathfrak{C}(\theta) + \sqrt{\frac{2 \log(\delta^{-1})}{n}}(b-a) + \left(\frac{32 K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2} (b-a)^{1/2}.$$

In expectation

$$\mathbb{P}_{W^{(1)}, \dots, W^{(n)}} \left(\mathfrak{C}(\hat{\theta}) \right) \leq \inf_{\theta \in \Theta} \mathfrak{C}(\theta) + \left(\frac{32 K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2} (b-a)^{1/2}.$$

PROOF. The proof follows the same line as Lemma 18 on page 65 and 19 on page 69. To prove the first part of the lemma, we need to bound

$$M(\theta, \lambda) = \int \log \left[\mathbb{P}_W \left(\exp \left(-\lambda \min_{A \in \mathfrak{T}} \sum_{j \in A} \langle \theta'_j, \Phi(W_j) \rangle \right) \right) \right] d\rho_\theta(\theta'), \quad \theta \in \Theta, \lambda > 0.$$

and to apply Lemma 17 on page 63. First remark that

$$M(\theta, \lambda) \leq \int \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \left(\sum_{A \in \mathfrak{T}} \exp \left(-\alpha \lambda \sum_{j \in A} \langle \theta'_j, \Phi(W_j) \rangle \right) \right)^{1/\alpha} \right] d\rho_\theta(\theta').$$

Then use Jensen's inequality and Fubini's theorem to move the integration with respect to ρ_θ inside to get

$$M(\theta, \lambda) \leq \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \left(\sum_{A \in \mathfrak{T}} \int \exp \left(-\alpha \lambda \sum_{j \in A} \langle \theta'_j, \Phi(W_j) \rangle \right) d\rho_\theta(\theta') \right)^{1/\alpha} \right]$$

$$\begin{aligned}
&= \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \left(\sum_{A \in \mathcal{T}} \exp \left(-\alpha \lambda \sum_{j \in A} \langle \theta_j, W_j \rangle + \frac{\alpha^2 \lambda^2}{2\beta} \sum_{j \in A} \|W_j\|^2 \right) \right)^{1/\alpha} \right] \\
&\leq \log \mathbb{P}_W \left[\inf_{\alpha \geq 1} \exp \left(\log(|\mathcal{T}|)/\alpha - \lambda \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle + \frac{\alpha \lambda^2}{2\beta} K(\mathcal{T}) \|W\|_\infty^2 \right) \right] \\
&= \lambda \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} + \log \mathbb{P}_W \left[\exp \left(-\lambda \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) \right].
\end{aligned}$$

The above inequalities require that $\alpha \geq 1$, that is

$$\lambda^2 K(\mathcal{T}) \|W\|_\infty^2 \leq 2\beta \log(|\mathcal{T}|). \quad (4.16)$$

Remembering Hoeffding's lemma (used in the course of the proof of Lemma 18 on page 65), we get

$$M(\theta, \lambda) \leq \lambda \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} - \lambda \mathbb{P}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) + \frac{\lambda^2}{8} (b-a)^2.$$

Thus according to Lemma 17 on page 63, with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{P}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) &\leq \int \bar{\mathbb{P}}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta'_j, \Phi(W_j) \rangle \right) d\rho_\theta(\theta') \\
&\quad + \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} + \frac{\lambda}{8} (b-a)^2 + \frac{\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{2n\lambda}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\int \bar{\mathbb{P}}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta'_j, \Phi(W_j) \rangle \right) d\rho_\theta(\theta') \\
&\leq \bar{\mathbb{P}}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \int \langle \theta'_j, \Phi(W_j) \rangle d\rho_\theta(\theta') \right) = \bar{\mathbb{P}}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right).
\end{aligned}$$

Take

$$\lambda = 2 \sqrt{\frac{\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{n(b-a)^2}}$$

to obtain

$$\begin{aligned}
&(\mathbb{P}_W - \bar{\mathbb{P}}_W) \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) \\
&\leq \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} + (b-a) \sqrt{\frac{\beta \|\Theta\|^2 + 2 \log(\delta^{-1})}{4n}} \\
&\leq \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} + (b-a) \sqrt{\frac{\beta \|\Theta\|^2}{4n}} + (b-a) \sqrt{\frac{\log(\delta^{-1})}{2n}}.
\end{aligned}$$

Then choose

$$\beta = \|W\|_\infty \|\Theta\|^{-1} (b-a)^{-1} \sqrt{8nK(\mathcal{T}) \log(|\mathcal{T}|)}$$

to get

$$\begin{aligned} & (\mathbb{P}_W - \overline{\mathbb{P}}_W) \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) \\ & \leq \sqrt{\frac{\log(\delta^{-1})}{2n}} (b-a) + \left(\frac{8K(\mathcal{T}) \log(|\mathcal{T}|)}{n} \right)^{1/4} \|\Theta\|^{1/2} \|W\|_\infty^{1/2} (b-a)^{1/2}. \end{aligned}$$

and

$$\lambda = \frac{2}{(b-a)} \sqrt{\frac{\|W\|_\infty \|\Theta\|}{(b-a)}} \sqrt{\frac{8K(\mathcal{T}) \log(|\mathcal{T}|)}{n}} + \frac{2 \log(\delta^{-1})}{n}.$$

We have to satisfy condition (4.16) on page 88 that reads

$$\begin{aligned} & \frac{\|W\|_\infty \|\Theta\|}{(b-a)} \sqrt{\frac{8K(\mathcal{T}) \log(|\mathcal{T}|)}{n}} + \frac{2 \log(\delta^{-1})}{n} \leq \frac{\log(|\mathcal{T}|)}{2} \sqrt{\frac{8n \log(|\mathcal{T}|)}{K(\mathcal{T})}} \frac{(b-a)}{\|W\|_\infty \|\Theta\|} \\ & \iff \delta \geq \exp \left(-n \log(|\mathcal{T}|) \sqrt{\frac{n \log(|\mathcal{T}|)}{2K(\mathcal{T})}} \frac{(b-a)}{\|W\|_\infty \|\Theta\|} \left(1 - \frac{2K(\mathcal{T}) \|W\|_\infty^2 \|\Theta\|^2}{n \log(|\mathcal{T}|) (b-a)^2} \right) \right). \end{aligned}$$

The condition is satisfied when $|\mathcal{T}| \geq 2$,

$$n \geq \frac{4K(\mathcal{T}) \|W\|_\infty^2 \|\Theta\|^2}{\log(|\mathcal{T}|) (b-a)^2} \quad (4.17)$$

and $\delta \geq \exp[-n \log(|\mathcal{T}|)/\sqrt{2}]$. When $|\mathcal{T}| \geq 3$ and condition (4.17) is not satisfied, the generalization bound is larger than $b-a$ and therefore trivially true. Thus we can drop condition (4.17) if we assume that $|\mathcal{T}| \geq 3$.

To prove the second part of the lemma, we have got to study

$$M(\theta, \lambda) = \int \log \left[\mathbb{P}_W \left(\exp \left(\lambda \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j^*, W_j \rangle - \lambda \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta'_j, W_j \rangle \right) \right) \right] d\rho_\theta(\theta'), \quad \theta \in \Theta, \lambda > 0.$$

We can remark that \mathbb{P}_W almost surely

$$\left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j^*, W_j \rangle - \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta'_j, W_j \rangle \right) \in [a-b, b-a]$$

and proceed as in the first part of the proof to establish that

$$M(\theta, \lambda) \leq \lambda \|W\|_\infty \sqrt{\frac{2K(\mathcal{T}) \log(|\mathcal{T}|)}{\beta}} + \lambda \mathbb{P}_W \left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j^*, W_j \rangle - \min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) + \frac{\lambda^2}{2} (b-a)^2.$$

So the situation is the same as in the first part with a variance term increased by a factor four, and except for this small difference the computations are the same. \square

To apply Lemma 32 on page 86, we need to put the risk functions \mathfrak{C}_t , $t \in \{1, 3, 4\}$ in a suitable form, that is to express the different risks in terms of a sum of inner products of parameters θ_j and random vectors W_j belonging to some separable Hilbert space. Let us notice that we will not be able to express criterion \mathfrak{C}_2 in this form, since the summation over $j \in A$ is inside the exponential.

Let us start with \mathfrak{C}_1 , assuming that $X \in [-B, B]^d$, so that we can also assume that $\mu \in [-B, B]^{d \times k}$, since shrinking μ to that range decreases the risk. Remark accordingly that

$$\mathfrak{C}_1(g, \mu) = \mathbb{P}_X \left\{ \min_{A \in \mathcal{T}_{g, K}} \left[\mathbb{P}_S \left(\left(\sum_{j \in A} \mu_{j, S} \right)^2 \right) - 2 \sum_{j \in A} \mathbb{P}_S(X_S \mu_{j, S}) \right] \right\} + \mathbb{P}_X \left(\mathbb{P}_S(X_S^2) \right)$$

Moreover, since the supports are disjoint,

$$\left(\sum_{j \in A} \mu_{j, S} \right)^2 = \sum_{j \in A} \mu_{j, S}^2, \quad A \in \mathcal{T}_{g, K}.$$

Thus

$$\mathfrak{C}_1(g, \mu) = \mathbb{P}_X \left(\min_{A \in \mathcal{T}_{g, K}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right) + \mathbb{P}_X \left(\mathbb{P}_S(X_S^2) \right),$$

where $\theta_j = (\mu_j, B^{-1} \mathbb{P}_S(\mu_{j, S}^2)) \in \mathbb{L}^2(\mathbb{P}_S) \times \mathbb{R}$ and $W_j = (-2X, B)$, in the same space. Note that

$$\|\theta_j\|^2 = \mathbb{P}_S(\mu_{j, S}^2) + B^{-2} \mathbb{P}_S(\mu_{j, S}^2)^2 \leq 2B^2 \mathbb{P}_S(\text{supp}(\mu_j)),$$

and that $\|W_j\|^2 = 4\mathbb{P}_S(X_S^2) + B^2 \leq 5B^2$.

Moreover

$$\begin{aligned} -B^2 &\leq -\mathbb{P}_S(X_S^2) \leq \sum_{j \in A} \langle \theta_j, W_j \rangle \\ &\leq \mathbb{P}_S \left[\left(X_S - \sum_{j \in A} \mu_{j, S} \right)^2 \right] - \mathbb{P}_S(X_S^2) \leq \mathbb{P}_S(X_S^2) + 2\mathbb{P}_S \left[\left(\sum_{j \in A} \mu_{j, S} \right)^2 \right] \leq 3B^2. \end{aligned}$$

Consider the model

$$\mathfrak{M}(\mathfrak{S}) = \left\{ \mu \in [-B, B]^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathfrak{S} \right\}$$

and apply Lemma 32 on page 86 to the corresponding parameter set Θ . Taking into account the fact that

$$\|\Theta\|^2 \leq 2B^2 \mathfrak{S},$$

we obtain

PROPOSITION 33 *Consider any $k \geq 3$, any $K \geq 1$, any $\delta \geq \exp(-n \log(k)/\sqrt{2})$ and any $\mathfrak{S} \in [1, k]$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,*

$$\begin{aligned} \mathfrak{C}_1(\mu) - \mathbb{P}_X[\mathbb{P}_S(X_S^2)] &\leq \bar{\mathfrak{C}}_1(\mu) - \bar{\mathbb{P}}_X[\mathbb{P}_S(X_S^2)] \\ &\quad + 4B^2 \left(\sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{4n}} + \left(\frac{5K\mathfrak{S} \log|\mathfrak{T}_{\mu,K}|}{n} \right)^{1/4} \right). \end{aligned}$$

For any non random set of fragments $\mu^* \in \mathfrak{M}(\mathfrak{S})$, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_1(\mu) - \mathfrak{C}_1(\mu^*) - \bar{\mathfrak{C}}_1(\mu) + \bar{\mathfrak{C}}_1(\mu^*) \\ \leq 4B^2 \left(\sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{n}} + \left(\frac{20K\mathfrak{S} \log|\mathfrak{T}_{\mu,K}|}{n} \right)^{1/4} \right). \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{S})} \bar{\mathfrak{C}}_1(\mu) + 4B^2 \left(\frac{20K\mathfrak{S} \log(|\mathfrak{T}_{\mu,K}|)}{n} \right)^{1/4},$$

with probability at least $1 - \delta$,

$$\mathfrak{C}_1(\hat{\mu}) \leq \inf_{\mu \in \mathfrak{M}(\mathfrak{S})} \mathfrak{C}_1(\mu) + 4B^2 \left(\sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{n}} + \left(\frac{20K\mathfrak{S} \log|\mathfrak{T}_{\mu,K}|}{n} \right)^{1/4} \right).$$

Let us now turn our attention to the risk function \mathfrak{C}_3 . Consider a fix set of subsets $\{B_j \subset \llbracket 1, d \rrbracket : 1 \leq j \leq k\}$. Define accordingly

$$\mathfrak{T}_{B,K} = \left\{ A \subset \llbracket 1, k \rrbracket : |A| \leq K, B_i \cap B_j = \emptyset, (i, j) \in A^2, i \neq j \right\}.$$

Consider some separable Hilbert space H (for instance ℓ_2 , the space of square integrable real valued sequences). For each $j \in \llbracket 1, k \rrbracket$, there is a mapping

$$\Psi_j : \mathbb{R}^d \longrightarrow H$$

such that

$$\langle \Psi_j(x), \Psi_j(y) \rangle = \exp \left(-\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} [(x_S - y_S)^2] \right).$$

Introduce the model

$$\mathfrak{M}_B = \{\mu \in \mathbb{R}^{d \times k} : \text{supp}(\mu_j) \subset B_j\}.$$

For any (B, μ) , the risk $\mathfrak{C}_3(B, \mu)$ can be written as

$$\mathfrak{C}_3(B, \mu) = 1 + \mathbb{P}_X \left(\min_{A \in \mathfrak{T}_{B,K}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right),$$

where $\theta_j = \mathbb{P}_S(B_j) \Psi_j(\mu_j)$ and $W_j = -\Psi_j(X)$. Note that

$$\sum_{j=1}^k \|\theta_j\|^2 = \sum_{j=1}^k \mathbb{P}_S(B_j)$$

and that

$$\|W_j\| = 1,$$

while almost surely

$$\sum_{j \in A} \langle \theta_j, W_j \rangle \in [-1, 0].$$

Remark also that $\mathfrak{C}_3(\mu) = \mathfrak{C}_3(\mathfrak{J}_\mu, \mu) = \mathfrak{C}_3(\text{supp}(\mu), \mu)$. Applying Lemma 32 on page 86 gives

PROPOSITION 34 *Consider the model*

$$\mathfrak{M}(\mathfrak{S}) = \left\{ \mu \in \mathbb{R}^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathfrak{S}, |\mathfrak{J}_{\mu, K}| \geq 3 \right\}.$$

Assume that $\delta \geq \exp(-n \log(3)/\sqrt{2})$ and that $\mathfrak{S} \in [1, k]$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\mathfrak{C}_3(\mu) \leq \bar{\mathfrak{C}}_3(\mu) + 2\sigma^2 \left(\sqrt{\frac{kd \log(2) + \log(\delta^{-1})}{2n}} + \left(\frac{8K\mathfrak{S} \log(|\mathfrak{J}_{\mu, K}|)}{n} \right)^{1/4} \right)$$

Consider a non random set of fragments $\mu^* \in \mathfrak{M}(\mathfrak{S})$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_3(\mu) - \mathfrak{C}_3(\mu^*) &\leq \bar{\mathfrak{C}}_3(\mu) - \bar{\mathfrak{C}}_3(\mu^*) \\ &\leq 2\sigma^2 \left(\sqrt{\frac{2kd \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{J}_{\mu, K}|)}{n} \right)^{1/4} \right). \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{S})} \bar{\mathfrak{C}}_3(\mu) + 2\sigma^2 \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{J}_{\mu, K}|)}{n} \right)^{1/4},$$

with probability at least $1 - \delta$,

$$\mathfrak{C}_3(\hat{\mu}) \leq \inf_{\mu \in \mathfrak{M}(\mathfrak{S})} \mathfrak{C}_3(\mu) + 2\sigma^2 \left(\sqrt{\frac{2kd \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{J}_{\mu, K}|)}{n} \right)^{1/4} \right).$$

The strong point of this proposition is that we do not have to assume that the signal X or the fragments μ_j are bounded. The weak point is that the union bound on B_j (that is essentially on the choice of $\text{supp}(\mu_j)$) introduce a dependence of the bound on the dimension d of the signal. Accordingly, when X is a digital image of dimension $d = 10^6$ or so, the proposition is not very meaningful, although it is still interesting within a submodel $\mathfrak{M}(B, \mathfrak{S})$ where B is fixed and accordingly the term $kd \log(2)$ is not present.

To get a generalization bound that is independent of the dimension d and does not require boundedness or integrability assumptions on the signal, we can turn to the smaller and therefore less demanding risk function \mathfrak{C}_4 .

There is a mapping $\Psi : \mathbb{R} \rightarrow H$ of the real line into a reproducing kernel separable Hilbert space H such that

$$\langle \Psi(x), \Psi(y) \rangle_H = \exp\left(-\frac{1}{2\sigma^2}(y - x)^2\right), \quad x, y \in \mathbb{R}.$$

We can then consider $\mathbb{L}^2(H^{\llbracket 1, d \rrbracket}, \mathbb{P}_S)$, another separable Hilbert space whose scalar product is defined as

$$\langle f, g \rangle = \mathbb{P}_S \left(\langle f(S), g(S) \rangle_H \right), \quad f, g : \llbracket 1, d \rrbracket \rightarrow H.$$

Define for any (g, μ) such that $g \in \mathfrak{G}$ and $\mu \in \mathfrak{M}_g$

$$\begin{aligned} \theta_j(s) &= \mathbf{1}(s \in B_j) \Psi(\mu_{j,s}), \\ W_j(s) &= -\Psi(X_s), \quad 1 \leq s \leq d, \quad 1 \leq j \leq k. \end{aligned}$$

We see that

$$\mathfrak{C}_4(g, \mu) = 2\sigma^2 \left[1 + \mathbb{P}_X \left(\min_{A \in \mathfrak{T}_{g,K}} \underbrace{\sum_{j \in A} \langle \theta_j, W_j \rangle}_{\in [-1, 0]} \right) \right].$$

Moreover $\|\theta_j\|^2 = \mathbb{P}_S(B_j)$ and $\|W_j\| = 1$. We obtain in view of Lemma 32 on page 86

PROPOSITION 35 *Define the model*

$$\mathfrak{M}(\mathfrak{S}) = \left\{ \mu \in \mathbb{R}^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathfrak{S}, \quad |\mathfrak{T}_{\mu,K}| \geq 3 \right\},$$

with $\mathfrak{S} \in [1, k]$ and assume that $\delta \geq \exp(-n \log(3)/\sqrt{2})$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\mathfrak{C}_4(\mu) \leq \bar{\mathfrak{C}}_4(\mu) + 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{4n}} + \left(\frac{8 K \mathfrak{S} \log(|\mathfrak{T}_{\mu,K}|)}{n} \right)^{1/4} \right).$$

Moreover, if $\mu^* \in \mathfrak{M}(\mathfrak{S})$ is a non random set of fragments, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_4(\mu) - \mathfrak{C}_4(\mu^*) - \bar{\mathfrak{C}}_4(\mu) + \bar{\mathfrak{C}}_4(\mu^*) \\ \leq 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{T}_{\mu,K}|)}{n} \right)^{1/4} \right). \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{S})} \bar{\mathfrak{C}}_4(\mu) + 2\sigma^2 \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{T}_{\mu,K}|)}{n} \right)^{1/4},$$

with probability at least $1 - \delta$

$$\mathfrak{C}_4(\hat{\mu}) \leq \inf_{\mu \in \mathfrak{M}(\mathfrak{S})} \mathfrak{C}_4(\mu) + 2\sigma^2 \left(\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} + \left(\frac{32 K \mathfrak{S} \log(|\mathfrak{T}_{\mu,K}|)}{n} \right)^{1/4} \right).$$

We see from this proposition that when we work with the robust risk \mathfrak{C}_4 , we can obtain a dimension free generalization bound that does not require any boundedness or integrability condition on the signal X , while \mathfrak{C}_4 still carries a meaningful notion of quantization. Another useful remark to understand the bound is that

$$\begin{aligned} \log(|\mathcal{T}_{\mu,K}|) &\leq \log \left[\sum_{m=0}^K \binom{k}{m} \right] \leq \log \left[\sum_{m=0}^K \frac{k^m}{m!} \right] \\ &\leq \log \left[\left(\frac{k}{K} \right)^K \sum_{m=0}^K \frac{K^m}{m!} \right] \leq K \log \left(\frac{ek}{K} \right). \end{aligned}$$

Thus the complexity term is bounded by

$$\left(\frac{32 K^2 \mathcal{S} \log(ek/K)}{n} \right)^{1/4}.$$

A second important remark we can make is that the upper bounds we obtained are increasing with respect to the term \mathcal{S} , which represents the maximum area covered by the fragments. In that way, if we diminish \mathcal{S} , we obtained a tighter bound. Accordingly, the bounds advocate to some extent the use of the fragmentation algorithm described in section 2.3 on page 29, since the fragmentation algorithm decreases \mathcal{S} at each step.

Besides, one can derive generalization bounds that are uniform with respect to \mathcal{S} , using a union bound over an appropriate grid containing a geometric progression of \mathcal{S} values. This requires to pay a small cost in the bound.

4.5. FASTER BOUNDS

In this section, we derive faster bounds for the different quantization criteria we developed so far. To this aim, we will borrow ideas from the classical chaining method used to upper bound the expected supremum of Gaussian processes (see [BLM13]). However, we will choose a PAC-Bayesian approach to this chaining technique in the sense that we will use a sequence of perturbations of the parameter parametrized by a variance ranging in a geometric grid, which will play the role of the δ -covering sets in the classical chaining argument.

4.5.1. FASTER BOUNDS FOR INFORMATION k -MEANS AND CLASSICAL k -MEANS. We can reach faster speeds than in Lemma 18 on page 65 and the likes through some kind of PAC-Bayesian chaining.

LEMMA 36 *Let W be a random vector in a separable Hilbert space H . Let (W_1, \dots, W_n) be a sample made of n independent copies of W . Let $\Theta \subset H^k$ be a bounded set of parameters. Define*

$$\|\Theta\| = \sup \left\{ \left(\sum_{j=1}^k \|\theta_j\|^2 \right)^{1/2} : \theta \in \Theta \right\}$$

and assume that

$$\mathbb{P}_W \left(\min_{j \in [1,k]} \langle \theta_j, W \rangle \in [a, b] \text{ for all } \theta \in \Theta \right) = 1.$$

For any $k \geq 2$, any $n \geq 2k$ and any $\delta \in]0, 1[$, with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) &\leq \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W_i \rangle \\
&+ \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2 \sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \|W\|_\infty \\
&+ \sqrt{\frac{(\sqrt{2} + 1) \left(k(b-a)^2 + 2 \log(ek) \|W\|_\infty^2 \|\Theta\|^2 \right)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}} (b-a).
\end{aligned}$$

If $\theta^* \in \Theta$ is a non random value of the parameter, with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned}
(\mathbb{P}_W - \bar{\mathbb{P}}_W) \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle - \min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j^*, W \rangle \right) \\
\leq \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2 \sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \|W\|_\infty \\
+ \sqrt{\frac{(\sqrt{2} + 1) \left(k(b-a)^2 + 2 \log(ek) \|W\|_\infty^2 \|\Theta\|^2 \right)}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} (b-a).
\end{aligned}$$

Therefore in the case when

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \bar{\mathbb{P}} \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right),$$

$\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \hat{\theta}_j, W \rangle \right) - \inf_{\theta \in \Theta} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right)$ satisfies the same bound with the same probability.

Moreover, the expected excess risk satisfies

$$\begin{aligned}
\mathbb{P}_{W_1, \dots, W_n} \left[\mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \hat{\theta}_j, W \rangle \right) - \inf_{\theta \in \Theta} \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) \right] \\
\leq \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2 \sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \|W\|_\infty \\
+ \sqrt{\frac{(\sqrt{2} + 1) \left(k(b-a)^2 + 2 \log(ek) \|W\|_\infty^2 \|\Theta\|^2 \right)}{n}}.
\end{aligned}$$

PROOF. Let

$$\rho_{\theta'} | \theta = \mathbb{P}_{\theta_i + \beta^{-1/2} \varepsilon_i, i \in \mathbb{N}}$$

be a Gaussian conditional probability distribution with values in $\mathfrak{M}_+^1(\mathbb{R}^{\mathbb{N}})$, where ε_i , $i \in \mathbb{N}$ is an infinite sequence of independent standard normal random variables. When θ and $\theta' \in \mathbb{R}^{\mathbb{N} \times k}$ are made of k infinite sequences of real numbers, let

$$\rho_{\theta'} | \theta = \bigotimes_{j=1}^k \rho_{\theta'_j} | \theta_j$$

be the tensor product of the previously defined conditional probability distributions. Let W be a random vector in the separable Hilbert space $\ell_2 \subset \mathbb{R}^{\mathbb{N}}$. Consider the functions

$$f(\theta, w) = \min_{j \in [1, k]} \langle \theta_j, w \rangle, \quad \theta \in \mathbb{R}^{\mathbb{N} \times k}, w \in \mathbb{R}^{\mathbb{N}},$$

where the scalar product is extended beyond ℓ_2 as already explained in a measurable but not bilinear way (see equation (4.4) on page 65). Let

$$\bar{f}(\theta, w) = f(\theta, w) - \mathbb{P}_W(f(\theta, W)), \quad \theta \in \mathbb{R}^{\mathbb{N} \times k}, w \in \mathbb{R}^{\mathbb{N}},$$

be the centered loss function. Previously introduced PAC-Bayesian inequalities (see Lemma 17 on page 63) show that

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \exp \sup_{\theta \in \ell_2^k} \left[n\lambda (\mathbb{P}_W - \bar{\mathbb{P}}_W)(\rho_{\theta'|\theta} - \rho_{\theta''|\theta}^2) f(\theta', W) \right. \right. \\ \left. \left. - n\rho_{\theta'|\theta} \left[\log \left(\mathbb{P}_W \left[\exp \left(-\lambda (\delta_{\theta''|\theta'} - \rho_{\theta''|\theta'}) \bar{f}(\theta'', W) \right) \right] \right) \right] - \frac{\beta \|\theta\|^2}{2} \right] \right\} \leq 1.$$

Apply Jensen's inequality and divide by $n\lambda$ to get

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \ell_2^k} \left[(\mathbb{P}_W - \bar{\mathbb{P}}_W)(\rho_{\theta'|\theta} - \rho_{\theta''|\theta}^2) f(\theta', W) \right. \right. \\ \left. \left. - \lambda^{-1} \rho_{\theta'|\theta} \left[\log \left(\mathbb{P}_W \left[\exp \left(-\lambda (\delta_{\theta''|\theta'} - \rho_{\theta''|\theta'}) \bar{f}(\theta'', W) \right) \right] \right) \right] - \frac{\beta \|\theta\|^2}{2n\lambda} \right] \right\} \leq 0.$$

Remark that

$$\begin{aligned} (\delta_{\theta'|\theta} - \rho_{\theta'|\theta}) f(\theta', W) &= \rho_{\theta'|\theta} \left(\min_j \langle \theta_j, W \rangle - \min_j \langle \theta'_j, W \rangle \right) \\ &\leq \rho_{\theta'|\theta} \left(\max_j \langle \theta_j - \theta'_j, W \rangle \right) \leq \sqrt{2 \log(k)/\beta} \|W\|_{\infty}, \end{aligned}$$

where the last inequality follows from the classical maximal inequality of the expectation of the maximum of i.i.d Gaussian random variables (see section 2.5 in [BLM13]).

Writing a symmetric inequality for the opposite, we deduce that

$$\left| (\delta_{\theta'|\theta} - \rho_{\theta'|\theta}) \bar{f}(\theta', W) \right| \leq 2\sqrt{2 \log(k)/\beta} \|W\|_{\infty}. \quad (4.18)$$

Using Hoeffding's inequality, and considering a closed bounded subset $\Theta \subset \ell_2^k$, we deduce that

$$\mathbb{P}_{W_1, \dots, W_n} \left[\sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W)(\rho_{\theta'|\theta} - \rho_{\theta''|\theta}^2) f(\theta', W) \right] \leq \frac{4\lambda}{\beta} \log(k) \|W\|_{\infty}^2 + \frac{\beta \|\Theta\|^2}{2n\lambda}.$$

In view of this, choose

$$\lambda = \frac{\beta \|\Theta\|}{\sqrt{8n \log(k) \|W\|_{\infty}}}$$

and define

$$F = \|W\|_\infty \|\Theta\| \sqrt{\frac{8 \log(k)}{n}}. \quad (4.19)$$

For any integer h ,

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[\left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \left(\rho_{\theta'|\theta}^{2^h} - \rho_{\theta'|\theta}^{2^{h+1}} \right) f(\theta', W) \right] \right\} \leq F.$$

Summing up for $h = 0$ to $H - 1$, where H is to be chosen later, and exchanging \sum_h and \sup_θ , we deduce that

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[\left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \left(\rho_{\theta'|\theta}^{2^H} - \rho_{\theta'|\theta}^{2^H} \right) f(\theta', W) \right] \right\} \leq HF.$$

As we are interested in bounding $(\mathbb{P}_W - \bar{\mathbb{P}}_W) f(\theta, W)$, there remains to upper bound

$$\left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) (\delta_{\theta'|\theta} - \rho_{\theta'|\theta}) f(\theta', W) \quad (4.20)$$

$$\text{and } \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta'|\theta}^{2^H} f(\theta', W), \quad (4.21)$$

or with a change of notation

$$\left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta'|\theta} f(\theta', W). \quad (4.22)$$

An almost sure bound for (4.20) is provided by equation (4.18). To bound (4.22), introduce the influence function

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0, \\ -\log(1 - x + x^2/2), & x \leq 0 \end{cases} \quad (4.23)$$

and put

$$\tilde{f}(\theta, W) = f(\theta, W) - \frac{a+b}{2}.$$

Decompose (4.22) into

$$\begin{aligned} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta'|\theta} f(\theta', W) = & \\ & \rho_{\theta'|\theta} \left[\mathbb{P}_W \tilde{f}(\theta', W) - \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi[\lambda \tilde{f}(\theta', W)] \right) \right] \end{aligned} \quad (4.24)$$

$$+ \rho_{\theta'|\theta} \bar{\mathbb{P}}_W \left[\lambda^{-1} \psi[\lambda \tilde{f}(\theta', W)] - \tilde{f}(\theta', W) \right]. \quad (4.25)$$

In order to bound (4.25), note that from lemma 7.2 in [Cat12]

$$|x - \psi(x)| \leq \frac{x^2}{4(1 + \sqrt{2})}, \quad x \in \mathbb{R}. \quad (4.26)$$

Therefore, from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and the properties of the variance, \mathbb{P}_W almost surely,

$$\rho_{\theta'|\theta} \left[\lambda^{-1} \psi[\lambda \tilde{f}(\theta', W)] - \tilde{f}(\theta', W) \right] \leq \frac{\lambda}{4(1 + \sqrt{2})} \rho_{\theta'|\theta} [\tilde{f}(\theta', W)^2]$$

$$\leq \frac{\lambda}{2(1+\sqrt{2})} \left[\left(\min_j \langle \theta_j, W \rangle - (a+b)/2 \right)^2 + \rho_{\theta' \mid 0} \left(\max_j \langle \theta'_j, W \rangle^2 \right) \right].$$

At this point, it remains to bound the variance term $\rho_{\theta' \mid 0} \left(\max_j \langle \theta'_j, W \rangle^2 \right)$. Let us remark that

$$\rho_{\theta' \mid 0} \circ (\theta'_j \mapsto \langle \theta'_j, W \rangle)^{-1} = \mathcal{N}(0, \|W\|^2/\beta).$$

Next, we need the following maximal inequality.

LEMMA 37 *Let $(\varepsilon_1, \dots, \varepsilon_k)$ be a sequence of Gaussian random variables such that $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$. We have*

$$\mathbb{E} \left(\max_{1 \leq j \leq k} \varepsilon_j^2 \right) \leq 2\sigma^2 \log(ke).$$

PROOF.

$$\begin{aligned} \mathbb{E} \left(\max_{1 \leq j \leq k} \varepsilon_j^2 \right) &= \int_{\mathbb{R}_+} \mathbb{P} \left(\max_{1 \leq j \leq k} \varepsilon_j^2 > t \right) dt \\ &\leq \int_{\mathbb{R}_+} \min \left\{ \sum_{j=1}^k \mathbb{P}(\varepsilon_j^2 > t), 1 \right\} dt \leq \int_{\mathbb{R}_+} \min \left\{ 2k \mathbb{P}(\varepsilon_1 > \sqrt{t}), 1 \right\} dt \\ &\leq \int_{\mathbb{R}_+} \min \left\{ k \exp \left(-\frac{t}{2\sigma^2} \right), 1 \right\} dt \leq 2\sigma^2 \log(k) + \int_{2\sigma^2 \log(k)}^{+\infty} k \exp \left(-\frac{t}{2\sigma^2} \right) dt \\ &\leq 2\sigma^2 \log(k) + 2\sigma^2 = 2\sigma^2 \log(ke). \end{aligned}$$

□

Accordingly, we obtain \mathbb{P}_W almost surely,

$$\begin{aligned} \rho_{\theta' \mid \theta} \left[\lambda^{-1} \psi \left[\lambda \tilde{f}(\theta', W) \right] - \tilde{f}(\theta', W) \right] \\ \leq \frac{\lambda}{2(1+\sqrt{2})} \left[\left(\min_j \langle \theta_j, W \rangle - (a+b)/2 \right)^2 + \rho_{\theta' \mid 0} \left(\max_j \langle \theta'_j, W \rangle^2 \right) \right] \\ \leq \frac{\lambda}{2(1+\sqrt{2})} \left[(b-a)^2/4 + 2 \log(ek) \|W\|_\infty^2 / \beta \right]. \quad (4.27) \end{aligned}$$

The right-hand side of this inequality provides an almost sure upper bound for (4.25). To bound (4.24), or rather the expectation of an exponential moment of (4.24), we can write a PAC-Bayesian bound using the influence function ψ .

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \exp \left[-n \lambda \rho_{\theta' \mid \theta} \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi \left[\lambda \tilde{f}(\theta', W) \right] \right) \right. \right. \\ \left. \left. - n \rho_{\theta' \mid \theta} \left[\log \left(\mathbb{P}_W \left[\exp \left(\psi \left[-\lambda \tilde{f}(\theta', W) \right] \right) \right] \right) \right] - \frac{\beta \|\theta\|^2}{2} \right] \right\} \leq 1. \end{aligned}$$

Remarking that

$$\psi(x) \leq \log(1 + x + x^2/2), \quad x \in \mathbb{R},$$

and removing the exponential according to Jensen's inequality, we obtain

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[\rho_{\theta' | \theta} \left[\mathbb{P}_W \left(\tilde{f}(\theta', W) \right) - \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] \right) \right] - \frac{\lambda}{2} \rho_{\theta' | \theta} \left[\mathbb{P}_W \left(\tilde{f}(\theta', W)^2 \right) \right] \right] \right\} \leq \frac{\beta \|\Theta\|^2}{2}.$$

Using the same maximal inequality in relation to the maximum of squared Gaussian random variables, used in the course of equation (4.27) to bound the variance term, we get

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \rho_{\theta' | \theta} \left[\mathbb{P}_W \left(\tilde{f}(\theta', W) \right) - \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] \right) \right] \right\} \leq \lambda \left[(b-a)^2/4 + 2 \log(ek) \|W\|_\infty^2 / \beta \right] + \frac{\beta \|\Theta\|^2}{2n\lambda}.$$

This provides an upper bound for (4.24). Combining it with the upper bound for (4.25) gives an upper bound for (4.22) that reads

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta' | \theta} f(\theta', W) \right\} \leq \frac{(\sqrt{2}+1)\lambda}{2} \left[(b-a)^2/4 + 2 \log(ek) \|W\|_\infty^2 / \beta \right] + \frac{\beta \|\Theta\|^2}{2n\lambda}.$$

Choosing

$$\lambda = \sqrt{\frac{4\beta \|\Theta\|^2}{(\sqrt{2}+1) \left[(b-a)^2 + 8 \log(ek) \|W\|_\infty^2 / \beta \right] n}}$$

gives

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta' | \theta} f(\theta', W) \right\} \leq \tilde{F}(\beta) \stackrel{\text{def}}{=} \sqrt{\frac{(\sqrt{2}+1) \left(\beta(b-a)^2 + 8 \log(ek) \|W\|_\infty^2 \right) \|\Theta\|^2}{4n}}.$$

Putting everything together,

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) f(\theta, W) \right\} \leq 2\sqrt{2 \log(k)/\beta} \|W\|_\infty + \tilde{F}(2^{-H}\beta) + HF,$$

where F is defined by equation (4.19) on page 97.

Let us choose $\beta = 2n\|\Theta\|^{-2}$ and $H = \lfloor \log(n/k)/\log(2) \rfloor$, so that

$$2^{-H}\beta \leq 4k\|\Theta\|^{-2}.$$

We get

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) f(\theta, W) \right\} \leq \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2\sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \|W\|_\infty$$

$$+ \sqrt{\frac{(\sqrt{2} + 1) \left(k(b - a)^2 + 2 \log(ek) \|W\|_\infty^2 \|\Theta\|^2 \right)}{n}}.$$

The upper deviations from this mean are controled by the extension of Hoeffding's bound called the bounded difference inequality (see section 6.1 and theorem 6.2 in [BLM13]). It gives with probability at least $1 - \delta$

$$\sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W) f(\theta, W) \leq \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W) f(\theta, W) \right\} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} (b - a).$$

This proves the first statement of the lemma. To get the second one, add to the previous inequality

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \left(\bar{\mathbb{P}}_W - \mathbb{P}_W \right) f(\theta^*, W) \right\} = 0$$

to get

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W) \left(f(\theta, W) - f(\theta^*, W) \right) \right\} \\ \leq \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 \log(k)}{n}} + 2 \sqrt{\frac{\log(k)}{n}} \right) \|\Theta\| \|W\|_\infty \\ + \sqrt{\frac{(\sqrt{2} + 1) \left(k(b - a)^2 + 2 \log(ek) \|W\|_\infty^2 \|\Theta\|^2 \right)}{n}}. \end{aligned}$$

and apply the bounded difference inequality to get the deviations. \square

Let us first apply this lemma to the information k -means problem.

PROPOSITION 38 *Assume that*

$$\text{ess sup}_X \left(\int p_X^2 d\nu \right) < +\infty \quad \text{and} \quad \text{ess sup}_X \left(\int \log(p_X)^2 d\nu \right) < +\infty.$$

Consider the information radius

$$R = \inf_{q \in \mathbb{L}_{+,1}^1(\nu)} \text{ess sup}_X \mathcal{K}(q, p_X)$$

and the bounds

$$\begin{aligned} B &= \text{ess sup}_X \left(\int p_X^2 d\nu \right)^{1/2} \exp(R) \\ \text{and } C &= \text{ess sup}_X \left(\int \log(p_X)^2 d\nu \right)^{1/2}. \end{aligned}$$

Introduce the parameter space

$$\mathcal{Q}_B = \left\{ q \in \mathbb{L}_{+,1}^1(\nu) \cap \mathbb{L}^2(\nu) : \int q^2 d\nu \leq B^2 \right\}.$$

Given (X_1, \dots, X_n) , a sample made of n independent copies of X , with probability at least $1 - \delta$, for any $q \in \mathcal{Q}_B^k$,

$$\begin{aligned}
\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) &\leq \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_{X_i}) \\
&+ \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 k \log(k)}{n}} + 2 \sqrt{\frac{k \log(k)}{n}} \right. \\
&\quad \left. + \sqrt{\frac{(\sqrt{2} + 1) k (3 + 2 \log(k))}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}} \right) (BC + 2 \log(B)).
\end{aligned}$$

Consider an empirical risk minimizer $\hat{q}(X_1, \dots, X_n) \in \mathfrak{Q}_B^k$ satisfying

$$\hat{q} \in \arg \min_{q \in \mathfrak{Q}_B^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right).$$

With probability at least $1 - \delta$,

$$\begin{aligned}
\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(\hat{q}, p_X) \mid X_1, \dots, X_n \right) &\leq \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right) \\
&+ \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8 k \log(k)}{n}} + 2 \sqrt{\frac{k \log(k)}{n}} \right. \\
&\quad \left. + \sqrt{\frac{(\sqrt{2} + 1) k (3 + 2 \log(k))}{n}} + \sqrt{\frac{2 \log(\delta^{-1})}{n}} \right) (BC + 2 \log(B)).
\end{aligned}$$

PROOF. Consider the parametrization used in the proof of Proposition 20 on page 71, that is $\theta = (q, \mathfrak{K}(q, 1)) \in H$ and $W = (-\log(p_X), \mu^{-1}) \in H$. This puts the information k -means risk in a suitable form to apply Lemma 36 on page 94. Taking into account that $\|\Theta\| \|W\|_\infty \leq \sqrt{k} (BC + 2 \log(B))$ for a suitable choice of μ ends the proof. \square

We can also apply the previous lemma to the Euclidean k -means algorithm.

PROPOSITION 39 Consider a random vector X in a separable Hilbert space H . Let (X_1, \dots, X_n) be a sample made of n independent copies of X . Consider the ball of radius B

$$\mathfrak{B} = \{x \in H : \|x\| \leq B\}.$$

Assume that $\mathbb{P}(X \in \mathfrak{B}) = 1$ and consider an estimator

$$\hat{\mu} \in \arg \min_{\mu \in H^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \mu_j\|^2 \right).$$

Assume that $n \geq 2k$ and $k \geq 2$. With probability at least $1 - \delta$,

$$\begin{aligned}
\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \hat{\mu}_j\|^2 \right) &\leq \inf_{\mu \in H^k} \mathbb{P}_X \left(\|X - \mu_j\|^2 \right) \\
&+ B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}} \underbrace{\left(\frac{6\sqrt{2}}{\log(2)} + \frac{6}{\log(n/k)} + \frac{1}{\log(n/k)} \sqrt{\frac{2(\sqrt{2} + 1)(17 + 9 \log(k))}{\log(k)}} \right)}_{\leq 12.3}
\end{aligned}$$

$$+ 4B^2 \sqrt{\frac{2 \log(\delta^{-1})}{n}}.$$

Consequently, with probability at least $1 - \delta$,

$$\begin{aligned} \mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \hat{\mu}_j\|^2 \right) &\leq \inf_{\mu \in H^k} \mathbb{P}_X \left(\|X - \mu_j\|^2 \right) \\ &\quad + 16 B^2 \log \left(\frac{n}{k} \right) \sqrt{\frac{k \log(k)}{n}} + 4 B^2 \sqrt{\frac{2 \log(\delta^{-1})}{n}}. \end{aligned}$$

In expectation

$$\mathbb{P}_{X_1, \dots, X_n} \left[\mathbb{P}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|X - \hat{\mu}_j\|^2 \right) \right] \leq \inf_{\mu \in H^k} \mathbb{P}_X \left(\|X - \mu_j\|^2 \right) + 16 B^2 \log \left(\frac{n}{k} \right) \sqrt{\frac{k \log(k)}{n}}.$$

In conclusion a chaining argument yields a dimension free non asymptotic generalization bound that decreases in expectation as $\sqrt{k/n}$ up to logarithmic factors.

PROOF. Remark that

$$\hat{\mu} \in \arg \min_{\mu \in \mathcal{B}^k} \bar{\mathbb{P}}_X \left(\min_{j \in \llbracket 1, k \rrbracket} \|\mu_j\|^2 - 2 \langle X, \mu_j \rangle \right)$$

and apply the lemma to $W = (-2X, \gamma B) \in H \times \mathbb{R}$ and $\theta_j = (\mu_j, \gamma^{-1} \|\mu_j\|^2 B^{-1})$. Note that

$$\|W\|^2 \|\theta_j\|^2 \leq B^4 (4 + \gamma^2) (1 + \gamma^{-2}) = B^4 (5 + \gamma^2 + 4\gamma^{-2}).$$

Choose the optimal value $\gamma^2 = 2$ to get

$$\|W\|^2 \|\Theta\|^2 \leq 9kB^4$$

Remark also that

$$\langle \theta_j, W \rangle \in [-B^2, +3B^2].$$

This gives the first statement of the proposition, according to Lemma 36 on page 94. Since $4B^2$ is a trivial bound, we have now to prove that for any $k \geq 2$ and any $n \geq 2k$,

$$\begin{aligned} \min \left\{ 4, \log \left(\frac{n}{k} \right) \sqrt{\frac{k \log(k)}{n}} \left(\frac{6\sqrt{2}}{\log(2)} + \frac{6}{\log(n/k)} + \frac{1}{\log(n/k)} \sqrt{\frac{2(\sqrt{2} + 1)(17 + 9 \log(k))}{\log(k)}} \right) \right\} \\ \leq 16 \log \left(\frac{n}{k} \right) \sqrt{\frac{k \log(k)}{n}}. \end{aligned}$$

Putting $a = \frac{6\sqrt{2}}{\log(2)}$, $b = 16$, $\rho = n/k$,

$$\eta = 6 + \sqrt{\frac{2(\sqrt{2} + 1)(17 + 9 \log(k))}{\log(k)}},$$

$$f(\rho, k) = \sqrt{\log(k)/\rho} (a \log(\rho) + \eta(k))$$

$$\text{and } g(\rho, k) = b \sqrt{\log(k)/\rho} \log(\rho),$$

we have to prove that

$$\min\{4, f(\rho, k)\} \leq g(\rho, k), \quad \rho \geq 2, k \geq 2.$$

In other words, we have to prove that, when $g(\rho, k) < f(\rho, k)$, then $g(\rho, k) \geq 4$. This can also be written as

$$g(\rho, k) \geq 4, \quad \min\{\rho, k\} \geq 2, \quad g(\rho, k) < f(\rho, k).$$

According to the definitions, this is also equivalent to

$$\log(\rho) - 2\log(\log(\rho)) \leq 2\log(b/4) + \log(\log(k)), \quad \min\{\rho, k\} \geq 2, \quad (b-a)\log(\rho) \leq \eta(k).$$

Since η is decreasing and since $k \mapsto \log(\log(k))$ is increasing, if the statement is true for $k = 2$, it is true for any $k \geq 2$. Thus we have to prove that

$$\log(\rho) - 2\log(\log(\rho)) \leq 2\log(b/4) + \log(\log(2)), \quad \log(2) \leq \log(\rho) \leq \eta(2)/(b-a).$$

Putting $\xi = \log(\rho)$, we have to prove that

$$\xi - 2\log(\xi) \leq 2\log(b/4) + \log(\log(2)), \quad \log(2) \leq \xi \leq \eta(2)/(b-a).$$

Since $\xi \mapsto \xi - 2\log(\xi)$ is convex, it is enough to check the inequality at the two ends of the interval, that is when $\xi \in \{\log(2), \eta(2)/(b-a)\}$, which can be done numerically. More precisely, we have to check that

$$\begin{aligned} & 2\log(b/4) + \log(\log(2)) \\ & - \max\left\{\log(2) - 2\log(\log(2)), \eta(2)/(b-a) - 2\log[\eta(2)/(b-a)]\right\} \geq 0, \end{aligned}$$

and we get numerically that the left-hand side is larger than the minimum of 0.9 and 0.6. \square

4.5.2. FASTER BOUNDS FOR BOTH THE BOUNDED INFORMATION k -MEANS AND THE BOUNDED k -MEANS CRITERION. Let us derive faster generalization bounds concerning the bounded information k -means and the bounded k -means criterion, that are defined respectively in section 4.2 on page 75 and 4.3 on page 80.

PROPOSITION 40 *Let X_1, \dots, X_n be made of n independent copies of X . For any family of k probability densities $q \in (\mathbb{L}_{+,1}^1(\nu))^k$, consider the bounded criterion $\mathfrak{C}(q)$ defined in Lemma 23 on page 76 and its empirical counterpart*

$$\bar{\mathfrak{C}}(q) = 1 - \bar{\mathbb{P}}_X \left(\exp \left[- \min_{j \in \llbracket 1, k \rrbracket} \mathfrak{K}(q_j, p_X) \right] \right).$$

Define the bounds B and C and the set \mathfrak{B} as in Lemma 25 on page 78. Assume that $k \geq 2$ and $n \geq 2k$. With probability at least $1 - \delta$, for any $q \in \mathfrak{B}^k$,

$$\begin{aligned}\mathfrak{C}(q) &\leq \bar{\mathfrak{C}}(q) + \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \right) \exp(BC) \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)k[1+2\log(ek)\exp(2BC)]}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}.\end{aligned}$$

Let $q^* \in \mathfrak{B}^k$ be a non random family of centers, and assume that $k \geq 2$ and $n \geq 2k$. With probability at least $1 - \delta$, for any $q \in \mathfrak{B}^k$,

$$\begin{aligned}\mathfrak{C}(q) - \mathfrak{C}(q^*) &\leq \bar{\mathfrak{C}}(q) - \bar{\mathfrak{C}}(q^*) + \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \right) \exp(BC) \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)k[1+2\log(ek)\exp(2BC)]}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}.\end{aligned}$$

Consequently, if

$$\hat{q}(X_1, \dots, X_n) \in \arg \min_{q \in \mathfrak{B}^k} \bar{\mathfrak{C}}(q),$$

with probability at least $1 - \delta$,

$$\begin{aligned}\mathfrak{C}(\hat{q}) &\leq \inf_{q \in (\mathbb{L}_{+,1}^1(\nu))^k} \mathfrak{C}(q) + \left(\frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \right) \exp(BC) \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)k[1+2\log(ek)\exp(2BC)]}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}.\end{aligned}$$

One can get also a similar bound for the expected excess risk.

PROOF. Remember from section 4.2 on page 75 that the risk $\mathfrak{C}(q)$ can be expressed as

$$\mathfrak{C}(q) = 1 - \mathbb{P}_W \left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) = 1 + \mathbb{P}_W \left(\min_{j \in \llbracket 1, k \rrbracket} \langle -\theta_j, W \rangle \right),$$

where $\theta_j = \exp(-\mathfrak{K}(q_j, 1)) \Psi(\mu_{q_j})$ and $W = \Psi(\mu^{-1} \log(p_X))$. Therefore we can apply Lemma 36 on page 94, using the inequality $\|\theta_j\|^2 \|W\|^2 \leq \exp(2BC)$. \square

Now, let us look at the bounded k -means criterion.

PROPOSITION 41 *Consider the same situation as in Proposition 27 on page 81. Consider any $k \geq 2$ and any $n \geq 2k$. With probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,*

$$\begin{aligned}\mathfrak{C}(\mu) &\leq \bar{\mathfrak{C}}(\mu) + \frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)k(3+2\log(k))}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}.\end{aligned}$$

For any non random family of centers $\mu^* \in \mathbb{R}^{d \times k}$, with probability at least $1 - \delta$, for any $\mu \in \mathbb{R}^{d \times k}$,

$$\begin{aligned} \mathfrak{C}(\mu) - \mathfrak{C}(\mu^*) &\leq \overline{\mathfrak{C}}(\mu) - \overline{\mathfrak{C}}(\mu^*) + \frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \\ &\quad + \sqrt{\frac{(\sqrt{2} + 1)k(3 + 2\log(k))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}. \end{aligned}$$

Consequently, if

$$\hat{\mu}(X_1, \dots, X_n) \in \arg \min_{\mu \in \mathbb{R}^{d \times k}} \overline{\mathfrak{C}}(\mu),$$

with probability at least $1 - \delta$,

$$\begin{aligned} \mathfrak{C}(\hat{\mu}) &\leq \inf_{\mu \in \mathbb{R}^{d \times k}} \mathfrak{C}(\mu) + \frac{\log(n/k)}{\log(2)} \sqrt{\frac{8k \log(k)}{n}} + 2\sqrt{\frac{k \log(k)}{n}} \\ &\quad + \sqrt{\frac{(\sqrt{2} + 1)k(3 + 2\log(k))}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}. \end{aligned}$$

PROOF. From section 4.3 on page 80, we can express the risk as

$$\mathfrak{C}(\mu) = 1 - \mathbb{P}_X \left(\max_{j \in \llbracket 1, k \rrbracket} \langle \theta_j, W \rangle \right) = 1 + \mathbb{P} \left(\min_{j \in \llbracket 1, k \rrbracket} \langle -\theta_j, W \rangle \right),$$

where $\theta_j = \Psi(\mu_j)$ and $W = \Psi(X)$, and where Ψ is the feature map associated to the Gaussian kernel. Accordingly, θ_j and W belongs to the unit ball in the corresponding reproducing kernel Hilbert space H , so that $\|W\|_\infty \leq 1$ and $\|\Theta\| \leq \sqrt{k}$. \square

4.5.3. FASTER BOUNDS FOR INFORMATION FRAGMENTATION. Similarly to what we have done previously, we can also obtain faster bounds for information fragmentation. However, we need first to establish the equivalent of Lemma 36 on page 94 for the information fragmentation risk.

LEMMA 42 *Let $W = (W_j, 1 \leq j \leq k)$ be a random vector in the product H^k , where H is a separable Hilbert space (that we can take as being ℓ_2 if we want). Consider a sample $(W^{(1)}, \dots, W^{(n)})$ made of n independent copies of W . Consider a bounded parameter set $\Theta \subset H^k$ and a set \mathfrak{T} of subsets of $\llbracket 1, k \rrbracket$. Assume that*

$$\mathbb{P}_W \left(\sum_{j \in A} \langle \theta_j, W_j \rangle \in [a, b], A \in \mathfrak{T}, \theta \in \Theta \right) = 1.$$

Consider the risk

$$\mathfrak{C}(\theta) = \mathbb{P}_W \left(\min_{A \in \mathfrak{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right), \quad \theta \in \Theta,$$

and its empirical counterpart

$$\overline{\mathfrak{C}}(\theta) = \overline{\mathbb{P}}_W \left(\min_{A \in \mathfrak{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle \right), \quad \theta \in \Theta.$$

Put

$$K(\mathfrak{T}) = \max_{A \in \mathfrak{T}} |A|, \quad \|\Theta\| = \sup_{\theta \in \Theta} \left(\sum_{j=1}^k \|\theta_j\|^2 \right)^{1/2} \text{ and } \|W\|_\infty = \max_{j \in \llbracket 1, k \rrbracket} \text{ess sup}_{\mathbb{P}_W} \|W_j\|.$$

Let \mathfrak{S} be a positive real parameter. Assume that $|\mathcal{T}| \geq 2$ and that $n \geq 2\mathfrak{S}K(\mathcal{T})$. For any $\delta \in]0, 1[$, with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned} \mathfrak{C}(\theta) &\leq \bar{\mathfrak{C}}(\theta) + \left(\frac{\log(n\mathfrak{S}^{-1}K(\mathcal{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathcal{T})\log(|\mathcal{T}|)}{n}} + 2\sqrt{\frac{K(\mathcal{T})\log(|\mathcal{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)\left(\mathfrak{S}(b-a)^2 + 2\log(e|\mathcal{T}|)\|W\|_\infty^2\|\Theta\|^2\right)K(\mathcal{T})}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}(b-a). \end{aligned}$$

If $\theta^* \in \Theta$ is a non random value of the parameter, with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned} \mathfrak{C}(\theta) - \mathfrak{C}(\theta^*) &\leq \bar{\mathfrak{C}}(\theta) - \bar{\mathfrak{C}}(\theta^*) \\ &\quad + \left(\frac{\log(n\mathfrak{S}^{-1}K(\mathcal{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathcal{T})\log(|\mathcal{T}|)}{n}} + 2\sqrt{\frac{K(\mathcal{T})\log(|\mathcal{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)\left(\mathfrak{S}(b-a)^2 + 2\log(e|\mathcal{T}|)\|W\|_\infty^2\|\Theta\|^2\right)K(\mathcal{T})}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}(b-a). \end{aligned}$$

Consequently, if

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \bar{\mathfrak{C}}(\theta),$$

with probability at least $1 - \delta$,

$$\begin{aligned} \mathfrak{C}(\hat{\theta}) &\leq \inf_{\theta \in \Theta} \mathfrak{C}(\theta) + \left(\frac{\log(n\mathfrak{S}^{-1}K(\mathcal{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathcal{T})\log(|\mathcal{T}|)}{n}} + 2\sqrt{\frac{K(\mathcal{T})\log(|\mathcal{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)\left(\mathfrak{S}(b-a)^2 + 2\log(e|\mathcal{T}|)\|W\|_\infty^2\|\Theta\|^2\right)K(\mathcal{T})}{n}} + \sqrt{\frac{2\log(\delta^{-1})}{n}}(b-a) \end{aligned}$$

In expectation

$$\begin{aligned} &\mathbb{P}_{W^{(1)}, \dots, W^{(n)}}(\mathfrak{C}(\hat{\theta})) \\ &\leq \inf_{\theta \in \Theta} \mathfrak{C}(\theta) + \left(\frac{\log(n\mathfrak{S}^{-1}K(\mathcal{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathcal{T})\log(|\mathcal{T}|)}{n}} + 2\sqrt{\frac{K(\mathcal{T})\log(|\mathcal{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ &\quad + \sqrt{\frac{(\sqrt{2}+1)\left(\mathfrak{S}(b-a)^2 + 2\log(e|\mathcal{T}|)\|W\|_\infty^2\|\Theta\|^2\right)K(\mathcal{T})}{n}}. \end{aligned}$$

PROOF. We follow here the same arguments as in the proof of Lemma 36 on page 94, so that we consider the same perturbation

$$\rho_{\theta'} | \theta = \mathbb{P}_{\theta_i + \beta^{-1/2}\varepsilon_i, i \in \mathbb{N}},$$

along with

$$\rho_{\theta'} | \theta = \bigotimes_{j=1}^k \rho_{\theta'_j | \theta_j},$$

in the case where θ and $\theta' \in \mathbb{R}^{\mathbb{N} \times k}$.

Taking a basis, we can assume without loss of generality that $W = (W_1, \dots, W_k)$ is a random vector in the separable Hilbert space $(\ell_2)^k \subset \mathbb{R}^{\mathbb{N} \times k}$. Define the loss function

$$f(\theta, w) = \min_{A \in \mathcal{J}} \sum_{j \in A} \langle \theta_j, w_j \rangle, \quad \theta \in \mathbb{R}^{\mathbb{N} \times k}, w = (w_1, \dots, w_k) \in \mathbb{R}^{\mathbb{N} \times k},$$

where the scalar product is extended beyond ℓ_2 as in equation (4.4) on page 65. Introduce

$$\bar{f}(\theta, w) = f(\theta, w) - \mathbb{P}_W(f(\theta, W)), \quad \theta \in \mathbb{R}^{\mathbb{N} \times k}, w \in \mathbb{R}^{\mathbb{N} \times k}.$$

Since

$$\min_A a_A - \min_A b_A \leq \max_A (a_A - b_A),$$

we see that

$$\begin{aligned} (\delta_{\theta'|\theta} - \rho_{\theta'|\theta})f(\theta', W) &= \rho_{\theta'|\theta} \left(\min_{A \in \mathcal{J}} \sum_{j \in A} \langle \theta_j, W_j \rangle - \min_{A \in \mathcal{J}} \sum_{j \in A} \langle \theta'_j, W_j \rangle \right) \\ &\leq \rho_{\theta'|\theta} \left(\max_{A \in \mathcal{J}} \sum_{j \in A} \langle \theta_j - \theta'_j, W_j \rangle \right). \end{aligned}$$

To bound the right-hand side of the previous inequality, we need a maximal inequality, knowing that

$$\rho_{\theta'|\theta} \circ (\theta'_j \mapsto \sum_{j \in A} \langle \theta_j - \theta'_j, W_j \rangle)^{-1} = \mathcal{N}\left(0, \sum_{j \in A} \|W_j\|^2 / \beta\right).$$

LEMMA 43 *Let \mathcal{J} be a set of subsets of $\llbracket 1, k \rrbracket$ and let $(\varepsilon_A)_{A \in \mathcal{J}}$ be a random process with marginal distributions $\mathbb{P}_{\varepsilon_A} = \mathcal{N}(0, \sigma_A^2)$, $A \in \mathcal{J}$. The expectation of its maximum is bounded by*

$$\mathbb{P}_\varepsilon \left(\max_{A \in \mathcal{J}} \varepsilon_A \right) \leq \sqrt{2 \log(|\mathcal{J}|)} \max_{A \in \mathcal{J}} \sigma_A.$$

PROOF. By Jensen's inequality, for any $t > 0$

$$\begin{aligned} \mathbb{P}_\varepsilon \left(\max_{A \in \mathcal{J}} \varepsilon_A \right) &\leq \frac{1}{t} \log \mathbb{P}_\varepsilon \left(\max_{A \in \mathcal{J}} \exp(t \varepsilon_A) \right) \\ &\leq \frac{1}{t} \log \left(\sum_{A \in \mathcal{J}} \mathbb{P}_\varepsilon \left(\exp(t \varepsilon_A) \right) \right) \leq \frac{1}{t} \log \left(|\mathcal{J}| \exp(t^2 \max_{A \in \mathcal{J}} \sigma_A^2 / 2) \right) \\ &\leq \frac{\log(|\mathcal{J}|)}{t} + \frac{t}{2} \max_{A \in \mathcal{J}} \sigma_A^2. \end{aligned}$$

Minimizing over t ends the proof. \square

Applying this lemma in the particular case where $\sigma_A^2 = \sum_{j \in A} \|W_j\|^2 / \beta$, and remarking that $\max_{A \in \mathcal{J}} \sigma_A^2 \leq K(\mathcal{J}) \|W\|_\infty^2 / \beta$, we get

$$(\delta_{\theta'|\theta} - \rho_{\theta'|\theta})f(\theta', W) \leq \rho_{\theta'|\theta} \left(\max_{A \in \mathcal{J}} \sum_{j \in A} \langle \theta_j - \theta'_j, W_j \rangle \right) \leq \sqrt{2K(\mathcal{J}) \log(|\mathcal{J}|) / \beta} \|W\|_\infty.$$

Accordingly, the same bound holding true for the opposite,

$$\left| \left(\delta_{\theta'|\theta} - \rho_{\theta'|\theta} \right) \bar{f}(\theta', W) \right| \leq 2\sqrt{2K(\mathcal{T}) \log(|\mathcal{T}|)/\beta} \|W\|_\infty.$$

Following the same line of proof as in Lemma 36 on page 94, we obtain

$$\mathbb{P}_{W_1, \dots, W_n} \left[\sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W) (\rho_{\theta'|\theta} - \rho_{\theta'|\theta}^2) f(\theta', W) \right] \leq \frac{4\lambda}{\beta} K(\mathcal{T}) \log(|\mathcal{T}|) \|W\|_\infty^2 + \frac{\beta \|\Theta\|^2}{2n\lambda}.$$

In the same way, choose

$$\lambda = \frac{\beta \|\Theta\|}{\sqrt{8nK(\mathcal{T}) \log(|\mathcal{T}|)} \|W\|_\infty}$$

and define

$$F = \|W\|_\infty \|\Theta\| \sqrt{\frac{8K(\mathcal{T}) \log(|\mathcal{T}|)}{n}}. \quad (4.28)$$

Following the PAC-Bayesian chaining strategy introduced previously, let us consider a sequence of measures $(\rho_{\theta'|\theta}^{2^h})_{h \in \llbracket 0, H-1 \rrbracket}$. For any integer $h \in \llbracket 0, H-1 \rrbracket$,

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[(\mathbb{P}_W - \bar{\mathbb{P}}_W) (\rho_{\theta'|\theta}^{2^h} - \rho_{\theta'|\theta}^{2^{h+1}}) f(\theta', W) \right] \right\} \leq F.$$

Summing over h , we get

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[\sum_{h=0}^{H-1} (\mathbb{P}_W - \bar{\mathbb{P}}_W) (\rho_{\theta'|\theta}^{2^h} - \rho_{\theta'|\theta}^{2^{h+1}}) f(\theta', W) \right] \right\} \\ \leq \mathbb{P}_{W_1, \dots, W_n} \left\{ \sum_{h=0}^{H-1} \sup_{\theta \in \Theta} \left[(\mathbb{P}_W - \bar{\mathbb{P}}_W) (\rho_{\theta'|\theta}^{2^h} - \rho_{\theta'|\theta}^{2^{h+1}}) f(\theta', W) \right] \right\} \leq HF. \end{aligned}$$

Hence, simplifying the telescoping sums on the left-hand side yields

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left[(\mathbb{P}_W - \bar{\mathbb{P}}_W) (\rho_{\theta'|\theta} - \rho_{\theta'|\theta}^{2^H}) f(\theta', W) \right] \right\} \leq HF.$$

As in the proof of Lemma 36 on page 94, since we are dealing with $(\mathbb{P}_W - \bar{\mathbb{P}}_W) f(\theta, W)$, it remains to bound the following two quantities

$$(\mathbb{P}_W - \bar{\mathbb{P}}_W) (\delta_{\theta'|\theta} - \rho_{\theta'|\theta}) f(\theta', W) \quad (4.29)$$

$$\text{and } (\mathbb{P}_W - \bar{\mathbb{P}}_W) \rho_{\theta'|\theta}^{2^H} f(\theta', W). \quad (4.30)$$

For the sake of simplicity, we can rewrite this last quantity as

$$(\mathbb{P}_W - \bar{\mathbb{P}}_W) \rho_{\theta'|\theta} f(\theta', W), \quad (4.31)$$

after a change of notation. Using the influence function ψ introduced in [Cat12] and defined by equation (4.23) on page 97 to decompose

$$(\mathbb{P}_W - \bar{\mathbb{P}}_W) \rho_{\theta'|\theta} f(\theta', W),$$

into a sub-Gaussian part and an other part representing extreme values, we get

$$\begin{aligned} (\mathbb{P}_W - \bar{\mathbb{P}}_W) \rho_{\theta'|\theta} f(\theta', W) = \\ \rho_{\theta'|\theta} \left[\mathbb{P}_W \tilde{f}(\theta', W) - \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] \right) \right] \end{aligned} \quad (4.32)$$

$$+ \rho_{\theta'|\theta} \bar{\mathbb{P}}_W \left[\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] - \tilde{f}(\theta', W) \right], \quad (4.33)$$

where $\tilde{f}(\theta, W)$ is defined in the same way as in the proof of 36 on page 94 as

$$\tilde{f}(\theta, W) = f(\theta, W) - \frac{a+b}{2}.$$

From inequality (4.26) on page 97, the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, and the properties of the variance, we get

$$\begin{aligned} \rho_{\theta'|\theta} \left[\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] - \tilde{f}(\theta', W) \right] &\leq \frac{\lambda}{4(1+\sqrt{2})} \rho_{\theta'|\theta} [\tilde{f}(\theta', W)^2] \\ &\leq \frac{\lambda}{2(1+\sqrt{2})} \left[\left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle - (a+b)/2 \right)^2 + \rho_{\theta'|\theta} \left(\max_{A \in \mathcal{T}} \left(\sum_{j \in A} \langle \theta'_j, W_j \rangle \right)^2 \right) \right]. \end{aligned}$$

Next, we need to bound the expectation of the maximum of the squares of Gaussian random variables. Consider the following adaptation of Lemma 37 on page 98.

LEMMA 44 *Let \mathcal{T} be a set of subsets of $\llbracket 1, k \rrbracket$ and let $(\varepsilon_A)_{A \in \mathcal{T}}$ be a random process with marginal distributions $\mathbb{P}_{\varepsilon_A} = \mathcal{N}(0, \sigma_A^2)$, $A \in \mathcal{T}$. It satisfies*

$$\mathbb{E} \left(\max_{A \in \mathcal{T}} \varepsilon_A^2 \right) \leq 2 \left(\max_{A \in \mathcal{T}} \sigma_A^2 \right) \log(|\mathcal{T}|e).$$

PROOF. It is similar to Lemma 37 on page 98. Indeed,

$$\begin{aligned} \mathbb{P}_\varepsilon \left(\max_{A \in \mathcal{T}} \varepsilon_A^2 \right) &\leq \int_{\mathbb{R}_+} \min \left\{ \sum_{A \in \mathcal{T}} \exp \left(-\frac{t}{2\sigma_A^2} \right), 1 \right\} dt \\ &\leq \int_{\mathbb{R}_+} \min \left\{ |\mathcal{T}| \exp \left(-\frac{t}{2 \max_{A \in \mathcal{T}} \sigma_A^2} \right), 1 \right\} dt = 2 \left(\max_{A \in \mathcal{T}} \sigma_A^2 \right) \log(|\mathcal{T}|e). \end{aligned}$$

□

From this lemma, we obtain, \mathbb{P}_W almost surely,

$$\begin{aligned} \rho_{\theta'|\theta} \left[\lambda^{-1} \psi [\lambda \tilde{f}(\theta', W)] - \tilde{f}(\theta', W) \right] &\leq \frac{\lambda}{2(1+\sqrt{2})} \left[\left(\min_{A \in \mathcal{T}} \sum_{j \in A} \langle \theta_j, W_j \rangle - (a+b)/2 \right)^2 + \rho_{\theta'|\theta} \left(\max_{A \in \mathcal{T}} \left(\sum_{j \in A} \langle \theta_j, W_j \rangle \right)^2 \right) \right] \\ &\leq \frac{\lambda}{2(1+\sqrt{2})} \left[(b-a)^2/4 + 2K(\mathcal{T}) \log(e|\mathcal{T}|) \|W\|_\infty^2 / \beta \right]. \end{aligned} \quad (4.34)$$

This gives a bound for (4.33). It remains to derive an upper bound for (4.32). Here, we just have to follow the same line of reasoning as in the proof of the Lemma 36 on page 94,

which consists in applying a PAC-Bayesian inequality with influence function ψ combined with Jensen's inequality. Using the previous lemma, we are led to

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \rho_{\theta' | \theta} \left[\mathbb{P}_W \left(\tilde{f}(\theta', W) \right) - \bar{\mathbb{P}}_W \left(\lambda^{-1} \psi \left[\lambda \tilde{f}(\theta', W) \right] \right) \right] \right\} \\ \leq \lambda \left[(b-a)^2/4 + 2K(\mathcal{T}) \log(e|\mathcal{T}|) \|W\|_\infty^2 / \beta \right] + \frac{\beta \|\Theta\|^2}{2n\lambda}. \end{aligned}$$

Then, combining this inequality with (4.34), we get

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta' | \theta} f(\theta', W) \right\} \\ \leq \frac{(\sqrt{2} + 1)\lambda}{2} \left[(b-a)^2/4 + 2K(\mathcal{T}) \log(e|\mathcal{T}|) \|W\|_\infty^2 / \beta \right] + \frac{\beta \|\Theta\|^2}{2n\lambda}. \end{aligned}$$

At this point, we choose

$$\lambda = \sqrt{\frac{4\beta \|\Theta\|^2}{(\sqrt{2} + 1) \left[(b-a)^2 + 8K(\mathcal{T}) \log(e|\mathcal{T}|) \|W\|_\infty^2 / \beta \right] n}},$$

which leads to

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) \rho_{\theta' | \theta} f(\theta', W) \right\} \\ \leq \tilde{F}(\beta) \stackrel{\text{def}}{=} \sqrt{\frac{(\sqrt{2} + 1) \left(\beta(b-a)^2 + 8K(\mathcal{T}) \log(e|\mathcal{T}|) \|W\|_\infty^2 \right) \|\Theta\|^2}{4n}}. \end{aligned}$$

Therefore,

$$\mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) f(\theta, W) \right\} \leq 2\sqrt{2K(\mathcal{T}) \log(|\mathcal{T}|)/\beta} \|W\|_\infty + \tilde{F}(2^{-H}\beta) + HF,$$

where F is defined in (4.28). Take

$$H = \left\lfloor \log \left(\frac{n}{K(\mathcal{T})\mathcal{S}} \right) \log(2)^{-1} \right\rfloor$$

and $\beta = 2n\|\Theta\|^{-2}$, so that

$$2^{-H}\beta \leq 4K(\mathcal{T})\mathcal{S}\|\Theta\|^{-2}.$$

We obtain

$$\begin{aligned} \mathbb{P}_{W_1, \dots, W_n} \left\{ \sup_{\theta \in \Theta} \left(\mathbb{P}_W - \bar{\mathbb{P}}_W \right) f(\theta, W) \right\} \\ \leq \left(\frac{\log(n\mathcal{S}^{-1}K(\mathcal{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathcal{T}) \log(|\mathcal{T}|)}{n}} + 2\sqrt{\frac{K(\mathcal{T}) \log(|\mathcal{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ + \sqrt{\frac{(\sqrt{2} + 1) \left(\mathcal{S}(b-a)^2 + 2\log(e|\mathcal{T}|) \|W\|_\infty^2 \|\Theta\|^2 \right) K(\mathcal{T})}{n}}. \end{aligned}$$

Accordingly, from the bounded difference inequality, we get with probability at least $1 - \delta$

$$\begin{aligned} & \sup_{\theta \in \Theta} (\mathbb{P}_W - \bar{\mathbb{P}}_W) f(\theta, W) \\ & \leq \left(\frac{\log(n \mathfrak{S}^{-1} K(\mathfrak{T})^{-1})}{\log(2)} \sqrt{\frac{8K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n}} + 2\sqrt{\frac{K(\mathfrak{T}) \log(|\mathfrak{T}|)}{n}} \right) \|\Theta\| \|W\|_\infty \\ & \quad + \sqrt{\frac{(\sqrt{2} + 1) \left(\mathfrak{S}(b - a)^2 + 2 \log(e|\mathfrak{T}|) \|W\|_\infty^2 \|\Theta\|^2 \right) K(\mathfrak{T})}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}} (b - a). \end{aligned}$$

In that way, we obtain the first part of the lemma. The other part follows in the same manner as in the proof of Lemma 36 on page 94. \square

At this point, we are in a position to apply previous Lemma 42 on page 105 and get faster speeds with respect to n than in Propositions 33 on page 90, 34 on page 92 and 35 on page 93 for the different risks $(\mathfrak{C}_i)_{i \in \{1,3,4\}}$ defined in section 4.4 on page 81.

PROPOSITION 45 *Consider the same setting as in Proposition 33 on page 90. Consider the model*

$$\mathfrak{M}(\mathfrak{S}) = \left\{ \mu \in [-B, B]^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathfrak{S}, |\mathfrak{T}_{\mu,K}| \geq 2 \right\}.$$

Consider any $k \geq 2$, any $K \geq 1$, any $\mathfrak{S} \in [1, k]$, any $n \geq 2\mathfrak{S}K$, and any $\delta \in]0, 1[$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_1(\mu) - \mathbb{P}_X[\mathbb{P}_S(X_S^2)] & \leq \bar{\mathfrak{C}}_1(\mu) - \bar{\mathbb{P}}_X[\mathbb{P}_S(X_S^2)] \\ & \quad + \sqrt{10} B^2 \left(\frac{\log(n \mathfrak{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} \right) \\ & \quad + B^2 \left(\sqrt{\frac{4(\sqrt{2} + 1)(9 + 5 \log(|\mathfrak{T}_{\mu,K}|))\mathfrak{S}K}{n}} + 2\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} \right). \end{aligned}$$

For any non random set of fragments $\mu^ \in \mathfrak{M}(\mathfrak{S})$, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,*

$$\begin{aligned} \mathfrak{C}_1(\mu) - \mathfrak{C}_1(\mu^*) - \bar{\mathfrak{C}}_1(\mu) + \bar{\mathfrak{C}}_1(\mu^*) \\ & \leq \left(\frac{\log(n \mathfrak{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} \right) \sqrt{10} B^2 \\ & \quad + \sqrt{\frac{4(\sqrt{2} + 1)(4 + 5 \log(e|\mathfrak{T}_{\mu,K}|))\mathfrak{S}K}{n}} B^2 + 4\sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}} B^2. \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{S})} \bar{\mathfrak{C}}_1(\mu) + \left(\frac{\log(n \mathfrak{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} \right) \sqrt{10} B^2$$

$$+ \sqrt{\frac{4(\sqrt{2} + 1)(4 + 5 \log(e|\mathcal{T}_{\mu,K}|))\mathfrak{S}K}{n}} B^2,$$

with probability at least $1 - \delta$,

$$\begin{aligned} \mathfrak{C}_1(\hat{\mu}) &\leq \inf_{\mu \in \mathfrak{M}(\mathfrak{S})} \mathfrak{C}_1(\mu) + \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \sqrt{10} B^2 \\ &\quad + \sqrt{\frac{4(\sqrt{2} + 1)(4 + 5 \log(e|\mathcal{T}_{\mu,K}|))\mathfrak{S}K}{n}} B^2 + 4\sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{n}} B^2. \end{aligned}$$

PROOF. In section 4.4 on page 81, the risk \mathfrak{C}_1 has already been written in a suitable form. Therefore the proof follows from Lemma 42 on page 105 and the bounds $\|\Theta\| \leq B\sqrt{2}\mathfrak{S}$, $\|W\|_\infty \leq B\sqrt{5}$ and $K(\mathcal{T}) \leq K$, along with $b - a \leq 4B^2$. \square

Let us state now the faster bounds related to the risk \mathfrak{C}_3 .

PROPOSITION 46 *Consider the model*

$$\mathfrak{M}(\mathfrak{S}) = \left\{ \mu \in \mathbb{R}^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathfrak{S}, |\mathcal{T}_{\mu,K}| \geq 2 \right\},$$

where $\mathfrak{S} \in [1, k]$. Assume that $n \geq 2\mathfrak{S}K$. For any $\delta \in]0, 1[$, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_3(\mu) &\leq \bar{\mathfrak{C}}_3(\mu) + 2\sigma^2 \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \\ &\quad + 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1)(1 + 2 \log(e|\mathcal{T}_{\mu,K}|))\mathfrak{S}K}{n}} + \sigma^2 \sqrt{\frac{2kd \log(2) + 2 \log(\delta^{-1})}{n}}. \end{aligned}$$

Consider a non random set of fragments $\mu^* \in \mathfrak{M}(\mathfrak{S})$. With probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathfrak{S})$,

$$\begin{aligned} \mathfrak{C}_3(\mu) - \mathfrak{C}_3(\mu^*) &\leq \bar{\mathfrak{C}}_3(\mu) - \bar{\mathfrak{C}}_3(\mu^*) \\ &\quad + 2\sigma^2 \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \\ &\quad + 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1)(1 + 2 \log(e|\mathcal{T}_{\mu,K}|))\mathfrak{S}K}{n}} + 2\sigma^2 \sqrt{\frac{2kd \log(2) + 2 \log(\delta^{-1})}{n}}. \end{aligned}$$

Consequently, if

$$\hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{S})} \bar{\mathfrak{C}}_3(\mu) + 2\sigma^2 \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right)$$

$$+ 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1) \left(1 + 2 \log(e|\mathcal{T}_{\mu,K}|)\right) \mathcal{S}K}{n}},$$

with probability at least $1 - \delta$,

$$\begin{aligned} \mathfrak{C}_3(\hat{\mu}) &\leq \inf_{\mu \in \mathfrak{M}(\mathcal{S})} \mathfrak{C}_3(\mu) \\ &\quad + 2\sigma^2 \left(\frac{\log(n \mathcal{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8 \mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2 \sqrt{\frac{\mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \\ &\quad + 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1) \left(1 + 2 \log(e|\mathcal{T}_{\mu,K}|)\right) \mathcal{S}K}{n}} + 2\sigma^2 \sqrt{\frac{2kd \log(2) + 2 \log(\delta^{-1})}{n}}. \end{aligned}$$

PROOF. Considering the formulation of the risk \mathfrak{C}_3 in terms of θ and W given in section 4.4 on page 81 and using the fact that $\|W\|_\infty = 1$, $b - a = 1$ and $\|\Theta\| \leq \mathcal{S}^{1/2}$, we conclude from Lemma 42 on page 105. \square

As discussed in section 4.4, the generalization bounds obtained for the risk \mathfrak{C}_3 depend on the (possibly high) dimension d of the signal. This is a motivation to consider rather the risk \mathfrak{C}_4 , for which we know how to get dimension free bounds.

PROPOSITION 47 *Define the model*

$$\mathfrak{M}(\mathcal{S}) = \left\{ \mu \in \mathbb{R}^{d \times k} : \sum_{j=1}^k \mathbb{P}_S(\text{supp}(\mu_j)) \leq \mathcal{S}, |\mathcal{T}_{\mu,K}| \geq 2 \right\},$$

where $\mathcal{S} \in [1, k]$ and assume that $n \geq 2 \mathcal{S}K$. For any $\delta \in]0, 1[$, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathcal{S})$,

$$\begin{aligned} \mathfrak{C}_4(\mu) &\leq \bar{\mathfrak{C}}_4(\mu) + 2\sigma^2 \left(\frac{\log(n \mathcal{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8 \mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2 \sqrt{\frac{\mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \\ &\quad + 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1) \left(1 + 2 \log(e|\mathcal{T}_{\mu,K}|)\right) \mathcal{S}K}{n}} + \sigma^2 \sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}}. \end{aligned}$$

Moreover, if $\mu^* \in \mathfrak{M}(\mathcal{S})$ is a non random set of fragments, with probability at least $1 - \delta$, for any $\mu \in \mathfrak{M}(\mathcal{S})$,

$$\begin{aligned} \mathfrak{C}_4(\mu) - \mathfrak{C}_4(\mu^*) &- \bar{\mathfrak{C}}_4(\mu) + \bar{\mathfrak{C}}_4(\mu^*) \\ &\leq 2\sigma^2 \left(\frac{\log(n \mathcal{S}^{-1} K^{-1})}{\log(2)} \sqrt{\frac{8 \mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} + 2 \sqrt{\frac{\mathcal{S}K \log(|\mathcal{T}_{\mu,K}|)}{n}} \right) \\ &\quad + 2\sigma^2 \sqrt{\frac{(\sqrt{2} + 1) \left(1 + 2 \log(e|\mathcal{T}_{\mu,K}|)\right) \mathcal{S}K}{n}} + 2\sigma^2 \sqrt{\frac{k(k-1) \log(2) + 2 \log(\delta^{-1})}{n}}. \end{aligned}$$

Consequently, if

$$\begin{aligned} \hat{\mu} \in \arg \min_{\mu \in \mathfrak{M}(\mathfrak{s})} \bar{\mathfrak{C}}_4(\mu) + 2\sigma^2 \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} \right) \\ + 2\sigma^2 \sqrt{\frac{(\sqrt{2}+1)(1+2\log(e|\mathfrak{T}_{\mu,K}|))\mathfrak{S}K}{n}}, \end{aligned}$$

with probability at least $1 - \delta$

$$\begin{aligned} \mathfrak{C}_4(\hat{\mu}) \leq \inf_{\mu \in \mathfrak{M}(\mathfrak{s})} \mathfrak{C}_4(\mu) + 2\sigma^2 \left(\frac{\log(n\mathfrak{S}^{-1}K^{-1})}{\log(2)} \sqrt{\frac{8\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} + 2\sqrt{\frac{\mathfrak{S}K \log(|\mathfrak{T}_{\mu,K}|)}{n}} \right) \\ + 2\sigma^2 \sqrt{\frac{(\sqrt{2}+1)(1+2\log(e|\mathfrak{T}_{\mu,K}|))\mathfrak{S}K}{n}} + 2\sigma^2 \sqrt{\frac{k(k-1)\log(2) + 2\log(\delta^{-1})}{n}}. \end{aligned}$$

PROOF. We can apply Lemma 42 on page 105 to the risk \mathfrak{C}_4 expressed in terms of θ and W . Let us recall that

$$\mathfrak{C}_4(g, \mu) = 2\sigma^2 \left[1 + \mathbb{P}_X \left(\min_{A \in \mathfrak{T}_{g,K}} \underbrace{\sum_{j \in A} \langle \theta_j, W_j \rangle}_{\in [-1,0]} \right) \right],$$

where

$$\begin{aligned} \theta_j(s) &= \mathbf{1}(s \in B_j) \Psi(\mu_{j,s}), \\ W_j(s) &= -\Psi(X_s), \quad 1 \leq s \leq d, \quad 1 \leq j \leq k. \end{aligned}$$

We conclude the proof remarking that $\|\theta_j\|^2 = \mathbb{P}_S(B_j)$, $\|W\|_\infty = 1$, and $\|\Theta\| \leq \mathfrak{S}^{1/2}$. \square

CHAPTER 5

Experiment on digital images

In this section, we describe some very preliminary experiment showing what to expect and hope for. We wrote a program that computes the syntax tree from a training set $(X_1, \dots, X_n) \in \mathbb{R}^{d \times n}$. Each X_i is a 300×300 greyscale image extracted at random from two larger greyscale images (we only kept the green channel from the original jpeg images) and $n = 200$. In order to play with the different parameters and visualize the experimental results in an interactive way, we decided to create a graphical user interface with the help of the `rWidgets2` package of the R language. We created two control panes containing several buttons and sliders to monitor the experiment. A first pane allows to load images and extract samples from them as square windows of the desired size. The positions of the windows can be constrained to a grid and are otherwise random. The grey levels are transformed by

$$f(x) = \log(10^\varepsilon + x), \quad x \in [0, 1],$$

the default value for ε being -2 as seen on the screen shot. This transform is justified by the fact that we will use an additive error model whereas we are more sensitive to light intensity ratios than to light intensity differences. Hence the logarithmic transform.

This is the control pane allowing to pick-up the training set drawn at random from larger scenes.

Prepare sample

×

Load new image

Upper left frame corner :

199

299

Lower right frame corner :

798

1198

Set frame

image size

300

step size

1

Sample size

100

epsilon value

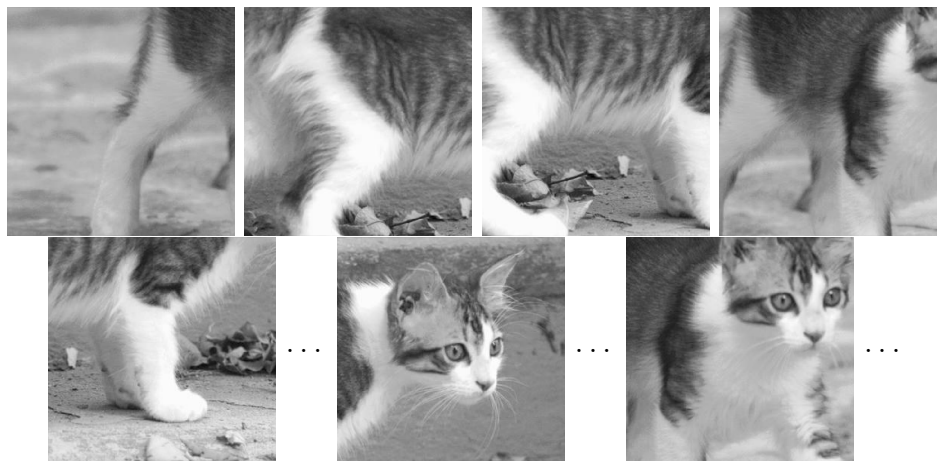
-2

Add to training set

We will show results on a sample drawn from two scenes featuring the same kitten.



The images are drawn from the yellow area of each scene. This is the beginning of the training sample



The fragmentation and syntax analysis are launched from a second control pane.

The parameter $\beta = \text{Beta}$ is used to compute the fragmentation criterion

$$2 \times 10^\beta |B_j| - |A_j|.$$

The parameter $\tau = \text{Threshold value}$ is used to compute the threshold

$$\alpha = \tau \bar{P}_S \left[\text{Var}(\bar{P}_{X_{I,S}|S}) \right]$$

appearing in equation (2.9) on page 30. Pressing the **Compute initial patches** button executes the sample fragmentation into a maximum number of patches prescribed in the text field **Number of patches**. Pressing the **Compute next level patches** button computes the next two levels of the syntax tree. This means that it computes pair labels and deduces from the compression of the pair label list a syntax classification f to be applied to pair labels. Pressing the button many times, we can grow the syntax tree ad libitum, until no new relabelling is observed. See the figure representing the syntax tree on page 36. The text fields **Number of merged labels** and **Number of syntax labels** specify the maximum number of labels to be created at each stage. Visualization is performed with the help of the sliders occupying the second half of the control pane. The first slider, named **Syntax level**, describes which level of the syntax tree is displayed. The second slider, titled **Show patches in image number**, displays all the patches in a given image (at the syntax level indicated by the previous slider). The next slider goes through all the patches present in a given image at a

given level. In other words, it displays one level of the syntax tree of a given image. The last slider, titled **Change image**, gathers all the images containing a given patch label at a given syntax level. The buttons **Show all** and **Show patch** toggle between displaying a particular patch on a green background and displaying all the patches present in one image using a rainbow of colors.

The experiment can be viewed as a crude vision model where images of two scenes are acquired by some sensor or retina through random eye saccades. In our experiment, the sample is shuffled by a random permutation and the locations in each scene of the acquired images are also random.

This is somehow a stress test for a vision model since there is no structure in the image acquisition process, by opposition to what we would get for instance if the image sample was extracted from a video stream.

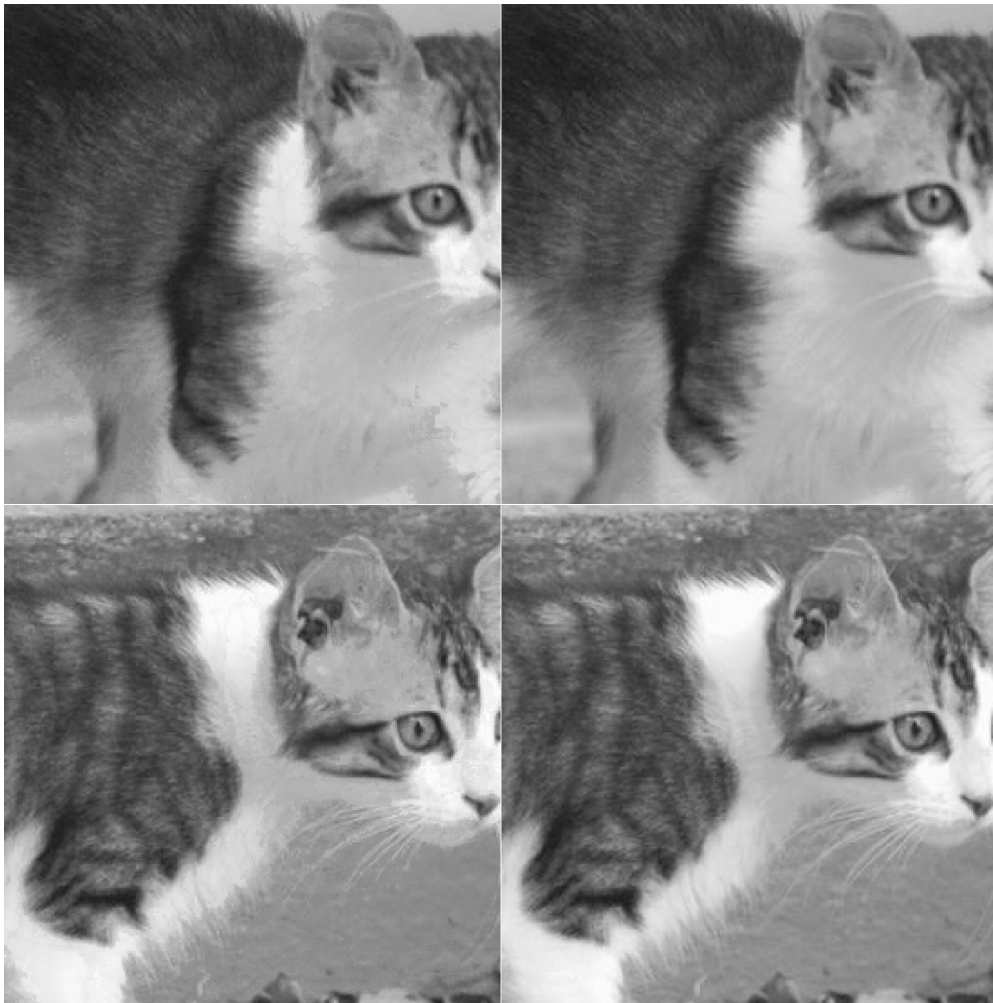
In this situation, two questions come to mind.

1. Will our unsupervised classification algorithm be able to put the same label on images that come from the same scene and are overlapping, so that they have some content in common, but translated. Note that our syntax analysis does not include any explicit translation model. The algorithm only compares pixel values being at the same pixel location in different sample images. The algorithm indeed makes no use of pixel location, that is of the sensor geometry. It is invariant with respect to any arbitrary permutation of the pixels, as long as the same permutation is applied at the same time to all sample images. Thus labelling translated patterns is a challenge. What we hope for is that syntax analysis can solve the problem of invariant pattern classification in a novel way. Syntax analysis describes the world using a mix of logics and compression theory, building up relabelling schemes described by rewriting rules and chosen for their ability to compress information. Chaining rewriting rules can be seen as some sort of crude logical reasoning. So the question is whether this crude logical reasoning can produce a flexible solution to the problem of invariant pattern recognition ? The more classical approach is to model explicitly pattern transformations using geometry or differential geometry and to pose the question of invariant classification with respect to a family of transformations in mathematical terms. Unfortunately, this route leads to daunting mathematical challenges, even in the somehow simplest case of translation invariance.
2. Will syntax analysis be able to put the same label on images containing related patterns coming from different scenes ? For example, in our sample, the kitten's head is present in the two scenes, but slightly rotated and on different backgrounds. Will it be possible to identify those two views of the same head, giving them the same label ?

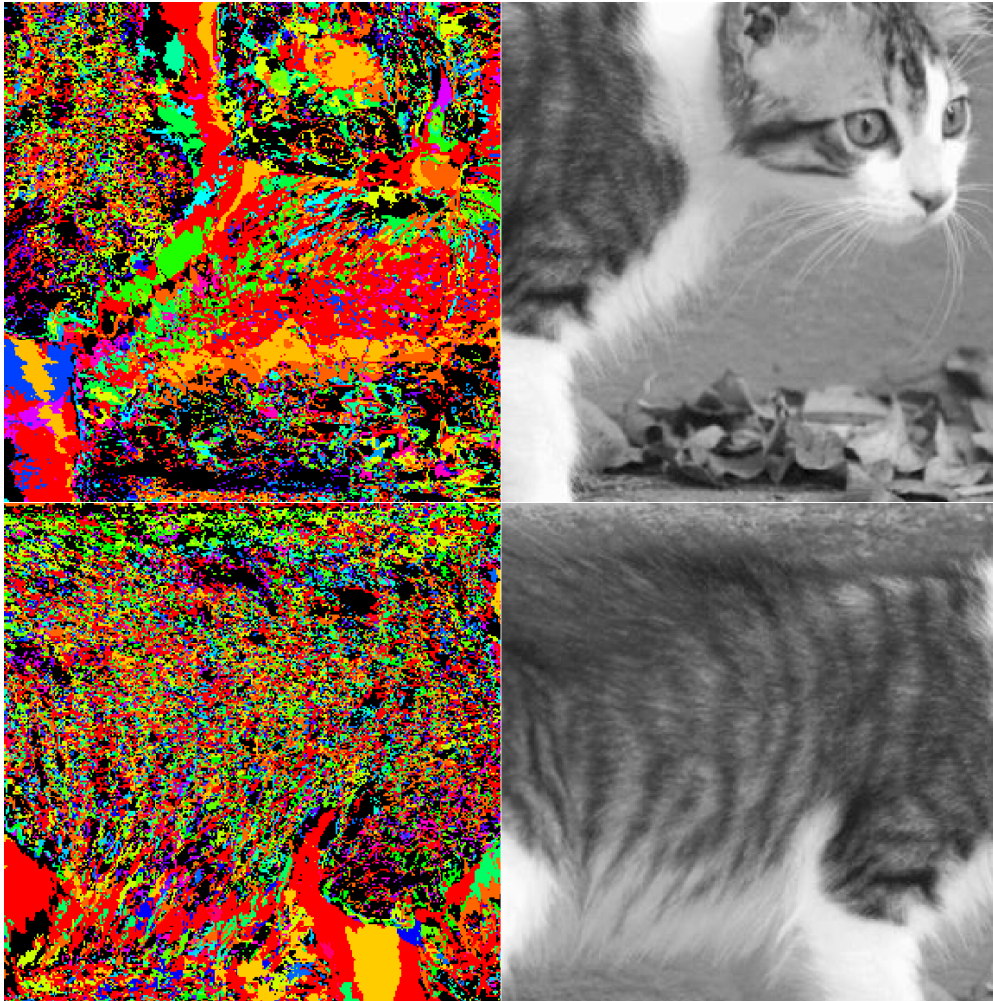
We will see that the answer to both questions is positive. Although more experiments would be needed, to see how stable things are and how the algorithm scales with larger data bases,

this is encouraging.

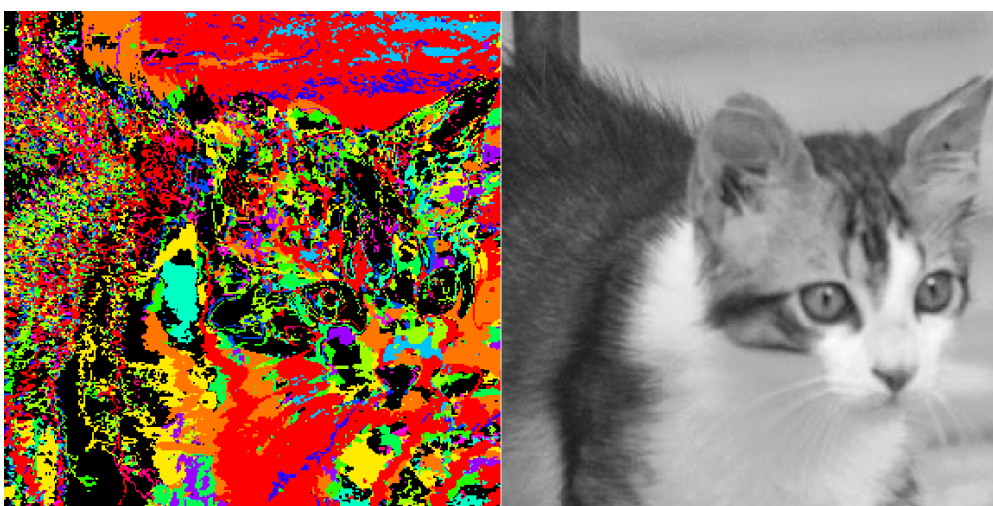
Let us show first the result of the fragmentation algorithm itself. Visually, when $\alpha = 1/10$, the coding distortion does not ruin the image content. We give two examples, on the left we see the patch reconstruction and on the right the original image



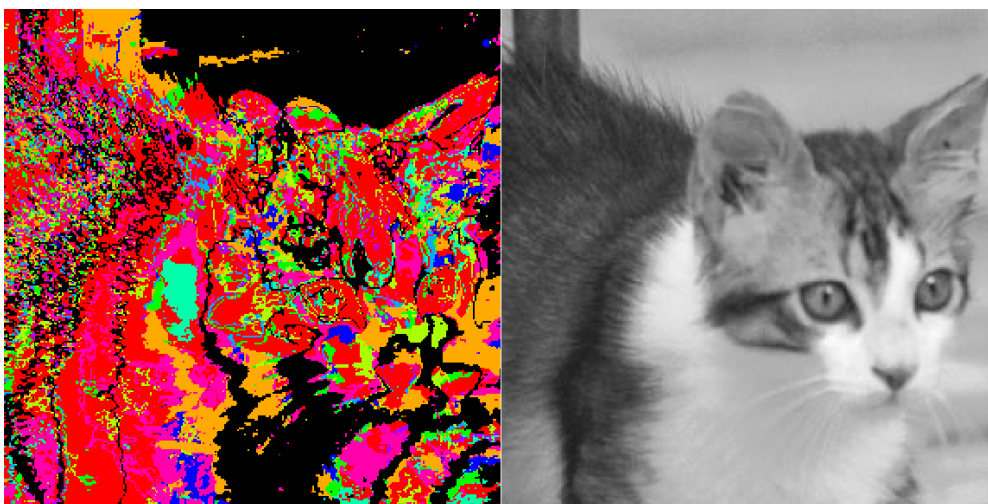
We now show two examples of fragment supports (when we use 1000 fragments for the whole sample). Colors do not match, we use a different rainbow in the two examples, each rainbow indexing only the fragments present in the corresponding image.



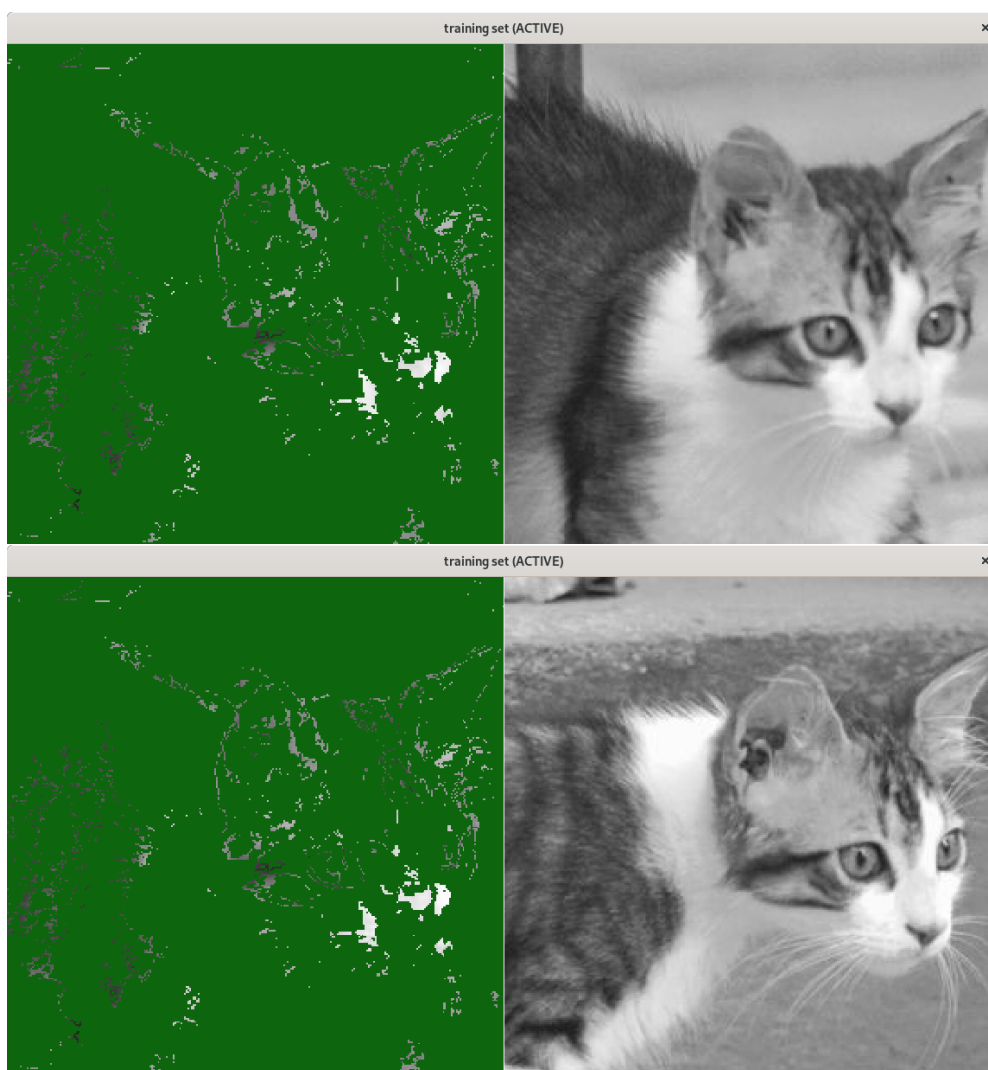
We will now choose an image, and show first its fragments



and then its syntax labels at the highest level (we computed a syntax tree with 2×5 levels, stopping when no new syntax levels are added). The levels are counted as in the diagram on page 36. Two levels are created each time we press the **Compute next level patches** button.



We show all the other images sharing with it a syntax label of the highest level. It turns out that each label is either specific to the reference image, or shared by two images, except for one label that is shared by three images. We start with a label shared by an image drawn from the other scene, since this is maybe the most exciting part of the result

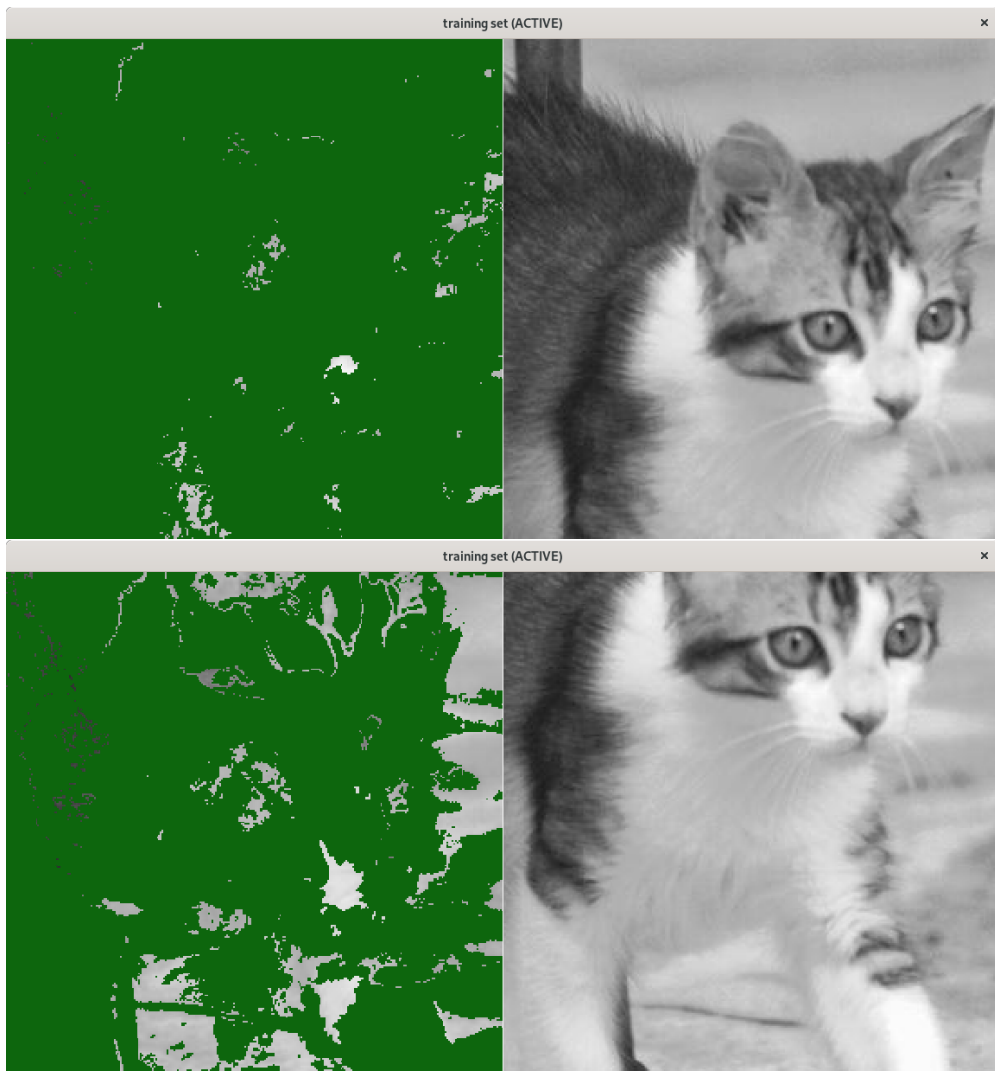


On the left, we can see the patch corresponding to the common label. In this case it has the

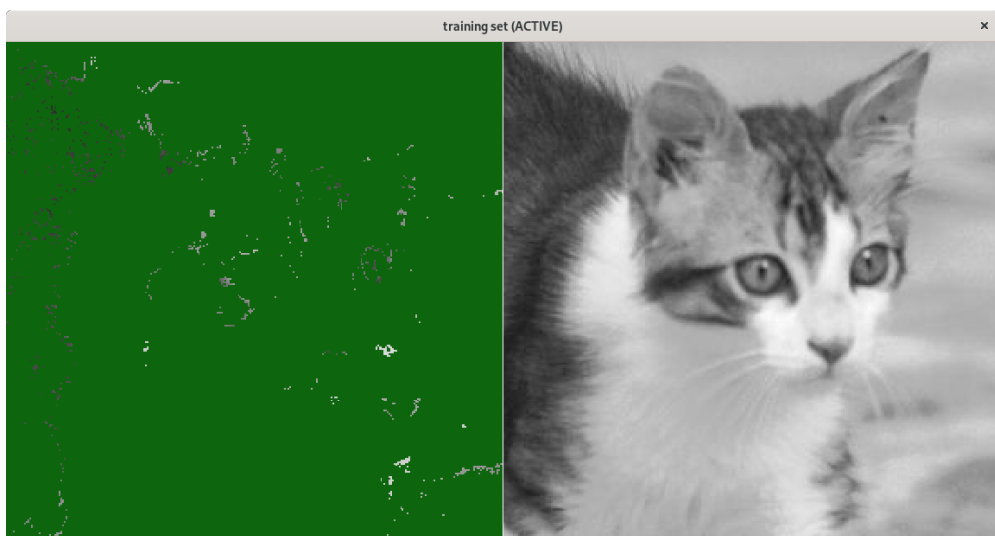
same support on both images, but we will see that it is not always the case (as expected). Let us see now another patch



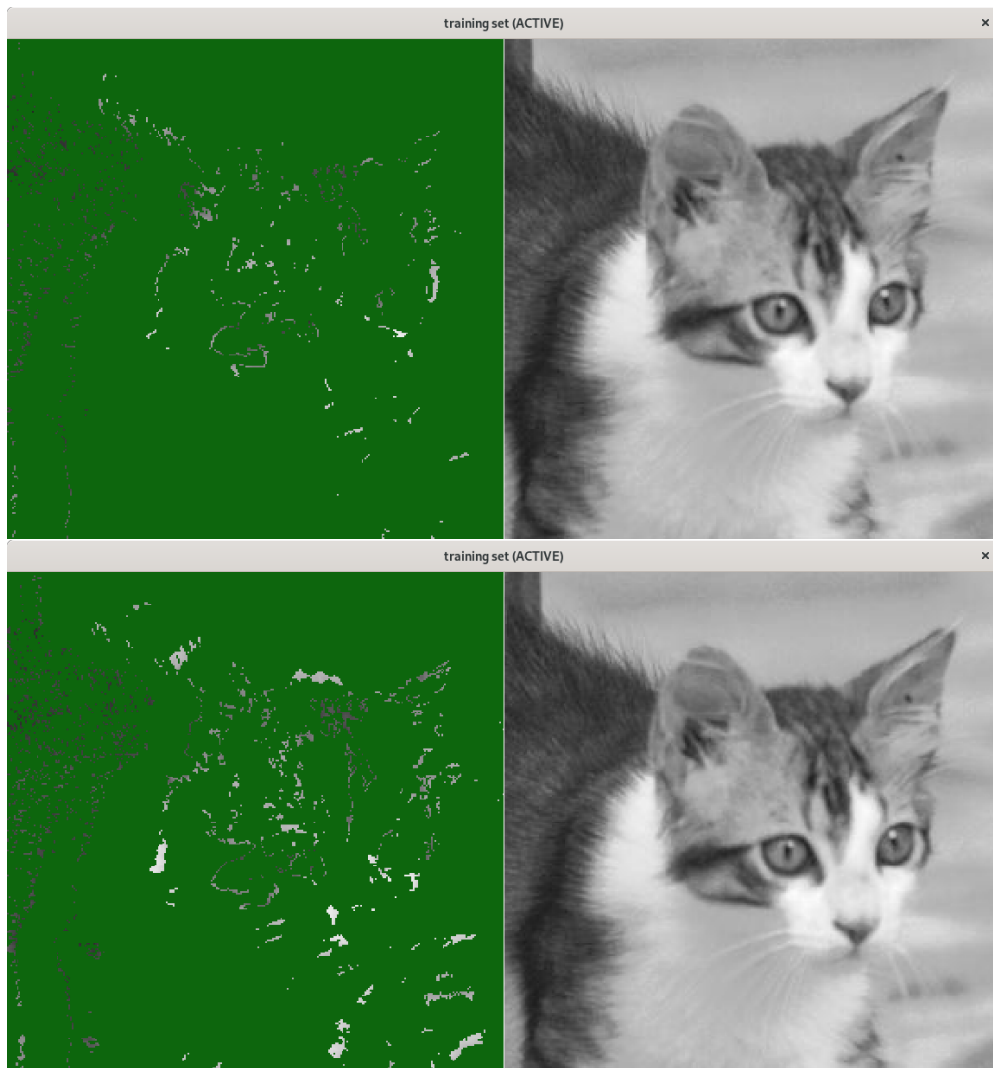
This time the patch has the same syntax label, but not the same support in both images. We can see that the support is somewhat shifted in the direction of the translation. This gives the impression that the syntax analysis performed on top of the fragmentation algorithm brings a positive contribution to the detection of translations. Let us see the other patches. Here again, the patch support changes



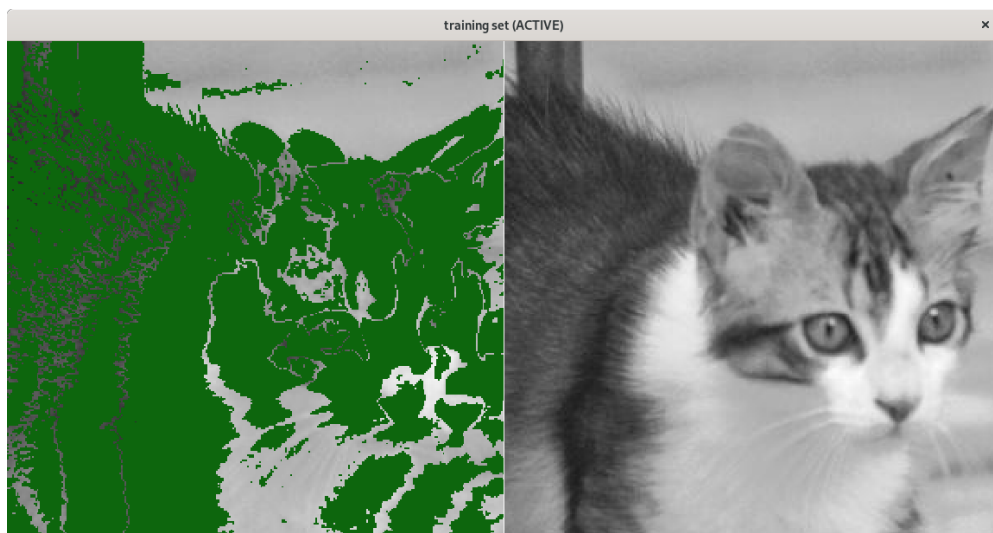
whereas in the next case, it is fixed



so that we do not repeat the original image. It is the same with the next two patches.



Finally, we have a syntax label indexing three images, whose support takes two different values, as shown in the next three pictures.





In conclusion, this small experiment let us hope for much. We have a pattern matching principle that produces visually interesting results. Moreover, it can be applied directly on high dimensional data, here we work with a sample of size 200 in \mathbb{R}^{90000} without being plagued by the curse of dimension. This was predicted at least for the fragmentation part by our dimension free generalization bounds, and it seems to be also the case in practice. That does not mean that, if we were to build a more serious vision algorithm, we could not combine the approach with some preprocessing. Indeed, our algorithm can take as input any vector of measurements, so we could try to extract first a vector of real valued features, like for instance wavelet coefficients. A more thoughtful vision model would also presumably include some sort of multi-scale processing, as well as the use of a retina with a space varying pixel density. Anyhow we preferred to apply a single general purpose treatment to raw data, to show that it can manage to deliver results by itself.

CHAPTER 6

Conclusion

In this thesis, we proposed new algorithms to perform unsupervised clustering and suggested new ideas for signal modeling, with a special focus on digital images. To this aim, we first introduced the information k -means algorithm and developed its mathematical analysis. We showed that it is a generalization of the classical Euclidean k -means criterion and demonstrated its benefits for the clustering of bag of words. Then, based on the interpretation of the information k -means setting as a density estimation framework, we introduced alternative bounded loss functions. Those criteria do not require any integrability assumptions on the sample, and can be regarded as robust variants of the usual k -means loss function.

We put the information k -means algorithm into a broader context that we called information fragmentation. On top of proving generalisation bounds, we justified this algorithm from a data compression perspective, as it computes shorter indices for large and frequent data blocks, similarly to the Lempel Ziv algorithm. We described fragmentation for signals belonging to \mathbb{R}^d . Although it covers the case of digital images, it is in fact much more general, since we do not use pixel geometry through any kind of neighboring relations. More precisely, the same fragments will be computed if we apply any given permutation to the pixel indices of all the images of a given training sample. Thus, the fragmentation algorithm can be applied to any kind of signal, color images, 3D images, video streams, speech and so on.

The fragmentation algorithm is a lossy compression scheme that maps any signal consisting in a vector of real valued measurements to a discrete representation consisting in a set of fragment labels. This mapping is chosen to optimize the compression of a training sample. Compression is then pushed further using our next proposal: a syntax analysis algorithm.

It is made of two stages: grouping and context analysis. The grouping stage performs a lossless compression by merging pairs of labels that appear frequently. This relabeling scheme produces new non terminal symbols and binary context free rewriting rules that we can gather into a context free grammar. The signal representation is changed to a new set of labels, containing non terminal *super fragment* labels that rewrite each into a unique set of fragment labels. The context analysis stage consists in compressing the representation of the

grammar through factoring and grouping of the rewriting rules. It produces a grammar of the grammar made of new rewriting rules for new syntax labels. These syntax labels perform some kind of context analysis drawn by a compression criterion. They induce a classification of the *super fragment* labels into syntax categories. We get a new signal representation using syntax categories. Repeating this scheme as long as we can increase the training sample compression rate, we build a syntax tree producing a hierarchical unsupervised classification of the content of each signal of the training set. Note that the whole scheme is guided by a single criterion, the compression rate and uses essentially two ingredients, grouping and factorization.

This is different from contextual modeling based on conditional probability measures, whose estimation is a difficult statistical inference issue. The choice of a compression scheme rather than a statistical model may be a way to avoid the curse of dimension as suggested by Lemma 4 on page 37. This lemma relates our compression approach to a statistical estimator whose generalization properties depend on the compression rate of the compression scheme.

The rest of our dissertation is devoted to the mathematical justification of the fragmentation algorithm through generalization bounds. These generalization bounds characterize the stability of the fragmentation, assessing that fragments computed from one training set would work almost as well in expectation, and consequently would also work almost as well to represent another independent training set.

Combining PAC-Bayesian lemmas with the kernel trick, we established dimension-free non asymptotic bounds on the excess risk of both k -means, information k -means and information fragmentation. These bounds show that the fragmentation algorithm does not overfit the data as long as

$$\frac{K^2 \mathfrak{S} \log(k/K)}{n}$$

goes to zero, where k is the number of fragments K the maximum number of fragments used to represent a single signal, and \mathfrak{S} is the total number of pixels of the fragments divided by the number of pixels in one image.

We used the same line of proof to obtain generalization bounds of order

$$\mathfrak{O}\left(\frac{k \log(k)}{n}\right)^{1/4}$$

for the classical k -means risk in a separable Hilbert space. In order to improve on the $1/4$ exponent, we blended the chaining method and the PAC-Bayesian technology to get bounds of order

$$\mathfrak{O}\left(\log\left(\frac{n}{k}\right)\left(\frac{k \log(k)}{n}\right)^{1/2}\right)$$

for our various flavours of the k -means criterion, including the classical one and improving in this case on the bound of order

$$\mathfrak{O}\left(\frac{k}{\sqrt{n}}\right)$$

proved in [BDL08].

Finally, we produced experimental results to show what the fragmentation algorithm combined with syntax analysis is capable of, when applied to digital images. The experiment, although preliminary, was very encouraging. First, we worked with 300×300 images, randomly drawn from two larger scenes, without suffering from the curse of dimension. Second, we worked with a sample of size 200, showing that interesting things can be learnt even from small training sets. Third, the algorithm was able to recognize (that is give the same label to) patterns present in the two scenes (a kitten's head) as well as translations of the observation window. This fosters the hope to bring a generic solution to the problem of invariant pattern recognition, where invariance with respect to transformations is understood in a loose sense, without the need to build specific methods dealing with a mathematically well defined subset of pattern transformations.

All this of course is very preliminary. We wanted to transpose syntax analysis from texts to signals, making a realistic proposal for unsupervised signal classification that could be demonstrated on real data. To complete this, we had somehow to stop as soon as we got some success at each step of our research, in order to keep some time to explore the next step. So we are pretty sure everything can be improved. For instance, the optimization strategy both in the fragmentation algorithm and in the syntax analysis is a one pass scheme: we build smaller and smaller fragments, never coming back on past choices, and then build larger and larger super fragments in the syntax phase, whereas it could be worth going up and down more than one time. Also, we presented a batch algorithm, letting open the question of its sequential pendent. Experimenting how the method scales on bigger data sets would also be necessary. The relationship between compression and estimation that Lemma 4 on page 37 hints at would deserve a more precise study. The syntax analysis scheme we propose is a crude first attempt, a more systematic investigation of the use of context free grammars to build compression schemes should be possible.

The use of chaining in PAC-Bayesian bounds is another technical subject that calls for further studies.

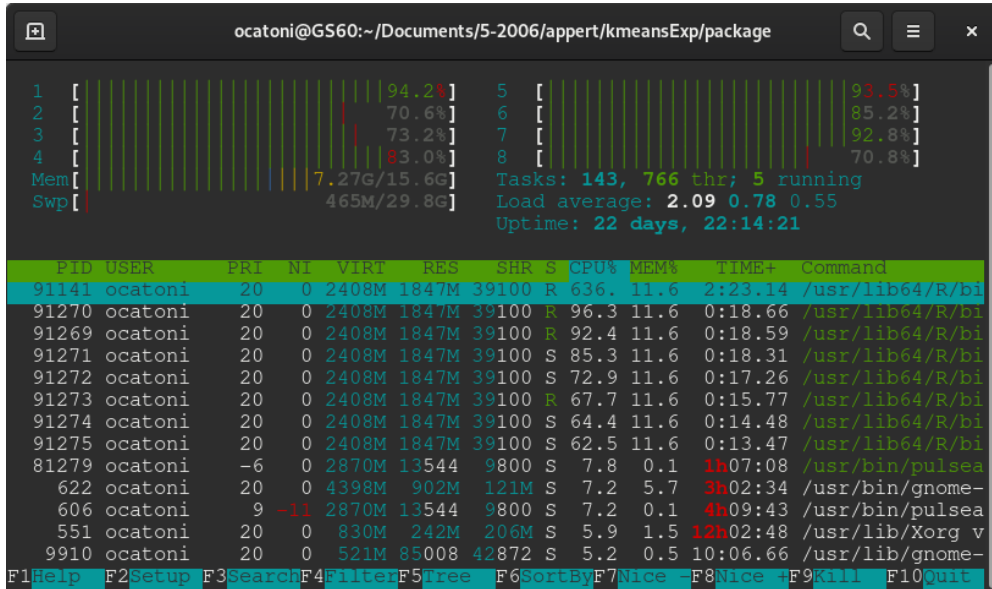
Further research may also include testing the method on other types of data, such as speech, or combinations of different types of data, like video streams with sound.

Our agenda also includes coming back to natural language processing to test the pendent of our syntax analysis algorithm for natural language processing. This would pave the way to combined data analysis, such as images with textual comments.

APPENDIX A

Code highlights

The syntax analysis is coded in R and C++, using the interface provided by the Rcpp package, see [EF11] and [Edd13]. We used the `Rcpp.package.skeleton()` function, to create our own R package. The advantage of this approach is that the communication between R and C++ is made easier and the final code represents a self-contained R package, thus easy to distribute, see [EF13] for more details concerning the creation of R packages with Rcpp. Moreover, the C++ code is parallelized using OpenMP parallelization directives and makes good use of an eight core processor as shown by the result of the linux `htop` command.



```
oatoni@GS60:~/Documents/5-2006/appert/kmeansExp/package
1  [|||||94.2%] 5  [|||||93.5%]
2  [|||||70.6%] 6  [|||||85.2%]
3  [|||||73.2%] 7  [|||||92.8%]
4  [|||||83.0%] 8  [|||||70.8%]
Mem[|||||7.27G/15.6G]
Swp[|||||465M/29.8G]
Tasks: 143, 766 thr; 5 running
Load average: 2.09 0.78 0.55
Uptime: 22 days, 22:14:21

  PID USER      PRI  NI  VIRT   RES   SHR  S  CPU%  MEM%   TIME+  Command
 91141 oatoni    20    0 2408M 1847M 39100 R  63.6 11.6  2:23.14 /usr/lib64/R/bin/R
91270 oatoni    20    0 2408M 1847M 39100 R  96.3 11.6  0:18.66 /usr/lib64/R/bin/R
91269 oatoni    20    0 2408M 1847M 39100 R  92.4 11.6  0:18.59 /usr/lib64/R/bin/R
91271 oatoni    20    0 2408M 1847M 39100 S  85.3 11.6  0:18.31 /usr/lib64/R/bin/R
91272 oatoni    20    0 2408M 1847M 39100 S  72.9 11.6  0:17.26 /usr/lib64/R/bin/R
91273 oatoni    20    0 2408M 1847M 39100 R  67.7 11.6  0:15.77 /usr/lib64/R/bin/R
91274 oatoni    20    0 2408M 1847M 39100 S  64.4 11.6  0:14.48 /usr/lib64/R/bin/R
91275 oatoni    20    0 2408M 1847M 39100 S  62.5 11.6  0:13.47 /usr/lib64/R/bin/R
81279 oatoni    -6    0 2870M 13544 9800 S   7.8  0.1  1h07:08 /usr/bin/pulsea-
622  oatoni    20    0 4398M  902M 121M S    7.2  5.7  3h02:34 /usr/bin/gnome-
606  oatoni    9  -11 2870M 13544 9800 S    7.2  0.1  4h09:43 /usr/bin/pulsea-
551  oatoni    20    0  830M  242M  206M S    5.9  1.5  12h02:48 /usr/lib/Xorg v
9910 oatoni    20    0  521M  85008 42872 S    5.2  0.5  10:06.66 /usr/lib/gnome-
F1Help F2Setup F3Search F4Filter F5Free F6SortBy F7Nice F8Nice + F9Kill F10Quit
```

We give the code of the main C++ functions implementing the fragmentation algorithm, to show how things can be done in practice, when the goal is only to compute the syntax tree. The implementation of the rest of the syntax tree, that is the computation of the pair labels and syntax labels, is quite similar. The main loop `computeLoop()` (line 5) runs the `split()` function (line 25), that is defined on line 32. In order to take less memory, variables and objects are passed to C++ via pointers, and all functions are of type `void`. Notice before the definition of the `computeLoop` function, the typical Rcpp instruction `// [[`

`Rcpp :: export []` that is required to make any compiled C++ function accessible into the R environment. Besides, we adopt an object oriented programming approach creating 2 classes (similar to the `struct` data type in C) named `Patches` and `mergeParam`. For the sake of clarity, we omit the description of the corresponding header file `initialization.h` containing the declaration of the classes. The class `Patches` is associated with the fragmentation step whereas the `mergeParam` is related to the merge and syntax analysis step. For the purpose of brevity, we will only talk about the class `Patches`. This class is composed of several attributes but more importantly contains the method `split()` that is called in the critical loop. The first method is called after the instantiation of a `Patches` object named `parameters` in the `computeLoop` function below. This method is used to initialize the quantity α , mentioned in the introduction of the experiment. Then, comes the call of the `split()` method which tries to find the pair $\{j_{k,1}, j_{k,2}\}$ (noted `max_i` and `max_j` in the code) maximizing the fragmentation criterion, as it is described in section 2.3 on page 29. This is done essentially by computing the fragmentation criterion for each iteration but in a parallel manner due to the compiler directive `#pragma omp parallel for` before the C++ `for` loop. As we already mentioned, this will distribute the computation of the criterion over the different processors. We perform first the computation of the index $j_{k,1}$ (`max_i`) in a first loop and then $j_{k,2}$ (`max_j`) in a second loop. As we discussed in section 2.3 on page 29, this approach avoids the computation of the criterion over all the possible pairs J_k . It is important to notice that we make use of the compiler directive `#pragma omp critical` twice. This compiler directive allows us to execute sequentially a block of code over the different processors. This means that the block of code can only be executed by one processor at a time. In our particular case, our critical block of code will be designed to retrieve the maximum fragmentation criterion (noted `max_crit` in the code), as well as the corresponding index j_k each processor had to compute in his own set of iterations. This provides us a way to reduce all the computations, performed separately over the processors, into one single result. Let us also point out that `max_crit` as well as `max_i` and `max_j` are global variables whereas `my_max_crit`, `my_max_i` and `my_max_j` are local variables private to each processor. Then, the remaining of the code is a matter of updating the storage of the criterion and the pair $\{j_{k,1}, j_{k,2}\}$, along with the matrices A and B representing the sequences of sets $A_{k,j}$ and $B_{k,j}$, $j \in \llbracket 1, k \rrbracket$. For an in-depth look at the code, the reader may look at the GitHub repository <https://github.com/GautierAppert/PatchProcess>.

```

#include "initialization.h"
2
// [[Rcpp::export]]
4
void computeLoop (
6   NumericVector &m,
   NumericVector &v,
8   NumericVector &w,
   NumericVector &C,

```

```

10 IntegerVector &A,
    IntegerVector &B,
12 IntegerVector &Wsx,
    List &paramList)
14 {
    int i, j;

16    // creates a Patches object to pass parameters from R
18    Patches parameters(m, v, w, C, A, B, Wsx, paramList);

20    parameters.computeBeta();
    paramList["thresholdValue"] = parameters.thresholdValue;
22    // loop on patch index
    for ( i = parameters.firstNewLabel- 1; i < parameters.
        lastNewLabel; ++i )
24    {
        parameters.split(i);
26        Rcout << "patch number " << i+1-parameters.shift << " computed
            \n";
        Rcout << "at iteration number " << i+1 << "\n";
28    }
    parameters.output();
30    Rcout << "betaCoeff = " << parameters.betaCoeff << "\n";
}

32 void Patches::split(int iter) {
    int i,j,s;
34    double m_buf, v_buf, weight_buf, square_buf, a, double_buf;
    double crit, max_crit, Cmax_i, Cmax_j;
36    int max_i, count, max_j, int_buf;

38    // computes arg max C
    max_crit = -1;
40    #pragma omp parallel private(i)
    {
42        double my_max_crit = -1;
        int my_max_i;
44        #pragma omp for
        for (i=shift; i<iter-shift; ++i)
46        {
            if ( my_max_crit < C[i] ) {
48                my_max_crit = C[i]; // selection according to the criterion
                .
                my_max_i = i;

```

```

50     }
    }
52 #pragma omp critical
    {
54     if ( max_crit < my_max_crit ) {
        max_crit = my_max_crit;
56     max_i = my_max_i;
    }
58 }
}

60
if (max_crit < 0) {
62     Rcout << "Nothing to split any more.\nLast split = " << iter-
        shift << "\n";
    ++shift;
64     return;
}

66
Rcout << "max_i = " << max_i + 1 << "\n";
68 // computes C for each j != max_i
max_crit = -1;
70 #pragma omp parallel private(j, count, s, weight_buf, crit, m_buf
    , \
    square_buf, v_buf)
72 {
    double my_max_crit = -1;
74     int my_max_j;
    #pragma omp for
76     for (j=shift; j<iter-shift; ++j)
    {
78         if (j == max_i) continue; // we are looking for j != max_i
        count = 0;
80         // compute the intersection between B_max_i and B_j.
        for (s=0; s<d; ++s)
82         {
            count += B[s+d*max_i] * B[s+d*j];
84         }
        if (count == 0) continue;
86
        // compute the sum of weights.
88         weight_buf = w[max_i] + w[j];

90         // compute the criterion for j.
        crit = 0;

```

```

92      // for each pixel do
93      // compute the criterion using the law of total variance.
94      for (s=0; s<d; ++s)
95      {
96          // Test if there is an intersection.
97          if (B[s+d*max_i]*B[s+d*j] == 0) continue;
98
99          // compute the mean of the mean.
100         m_buf = (w[max_i]*m[s+d*max_i]
101             + w[j]*m[s+d*j])/weight_buf;
102         square_buf = m[s+d*max_i] - m_buf;
103
104         // compute the variance of the means.
105         square_buf *= square_buf;
106
107         // compute the means of the variance.
108         v_buf = w[max_i] * ( v[s+d*max_i] + square_buf );
109         square_buf = m_buf - m[s+d*j];
110         square_buf *= square_buf;
111         v_buf += w[j] * ( v[s+d*j] + square_buf);
112         v_buf /= weight_buf;
113         if (v_buf < thresholdValue) {
114             crit += 1;
115         }
116     }
117     crit *= betaCoeff;
118     crit -= weight_buf;
119     if (my_max_crit < crit) {
120         my_max_crit = crit;
121         my_max_j = j;
122     }
123 }
124
125 // get our max_j.
126 #pragma omp critical
127 {
128     if (max_crit < my_max_crit) {
129         max_crit = my_max_crit;
130         max_j = my_max_j;
131     }
132 }
133 }
134 }

```



```

136
137 if (max_crit < 0) {
138     #pragma omp parallel for private(s,double_buf,int_buf)
139     for (s=0;s<d;++s)
140     {
141         double_buf = m[s+d*shift];
142         m[s+d*shift] = m[s+d*max_i];
143         m[s+d*max_i] = double_buf;
144         double_buf = v[s+d*shift];
145         v[s+d*shift] = v[s+d*max_i];
146         v[s+d*max_i] = double_buf;
147         int_buf = B[s+d*shift];
148         B[s+d*shift] = B[s+d*max_i];
149         B[s+d*max_i] = int_buf;
150     }
151     double_buf = w[shift];
152     w[shift] = w[max_i];
153     w[max_i] = double_buf;
154     double_buf = C[shift];
155     C[shift] = C[max_i];
156     C[max_i] = double_buf;
157     #pragma omp parallel for private(i,int_buf)
158     for (i=0;i<n;++i)
159     {
160         int_buf = A[i+n*shift];
161         A[i+n*shift] = A[i+n*max_i];
162         A[i+n*max_i] = int_buf;
163     }
164     ++shift;
165     return;
166 }

167
168 Rcout << "max_j = " << max_j + 1 << "\n";
169 C[iter-shift] = max_crit;
170 Cmax_i = Cmax_j = 0;
171 weight_buf = w[max_i] + w[max_j];
172
173 #pragma omp parallel for private(s, square_buf) reduction(+:
174     Cmax_i, Cmax_j)
175 for (s=0; s<d; ++s) // computes patch number iter
176 {
177
178     // compute patch mean using mean of m[s,i] and m[s,j].
179     m[s + d*(iter-shift)] = (w[max_i]*m[s+d*max_i]

```

```

    + w[max_j]*m[s+d*max_j])/weight_buf;
180
    square_buf = m[s + d*max_i] - m[s + d*(iter-shift)];
182    square_buf *= square_buf;
    v[s+d*(iter-shift)] = w[max_i] * ( v[s+d*max_i] + square_buf );
184    square_buf = m[s+d*max_j] - m[s+d*(iter-shift)];
    square_buf *= square_buf;
186    v[s+d*(iter-shift)] += w[max_j] * ( v[s+d*max_j] + square_buf);
    v[s+d*(iter-shift)] /= weight_buf;
188
    // Update A and B
190    // we need to compute the intersection
    // and check that the variance is under thresholdValue.
192    if ( (B[s+d*max_i]*B[s+d*max_j] > 0) && (v[s+d*(iter-shift)]
        < thresholdValue) )
194    {
        B[s+d*(iter-shift)] = 1;
196        B[s+d*max_i] = 0;
        B[s+d*max_j] = 0;
198    } else {
        B[s+d*(iter-shift)] = 0;
200    }
    if (B[s+d*max_i] > 0) {
202        Cmax_i += 1;
    }
    if (B[s+d*max_j] > 0) {
204        Cmax_j += 1;
206    }
    }
208
    C[max_i] = Cmax_i*betaCoeff - w[max_i];
210    C[max_j] = Cmax_j*betaCoeff - w[max_j];
    w[iter-shift] = w[max_i] + w[max_j];
212    #pragma omp parallel for private(i)
    for (i=0; i<n; ++i)
214    {
        A[i+n*(iter-shift)] = A[i+n*max_i] + A[i+n*max_j];
216    }
    return;
218 }

```


APPENDIX B

Présentation générale

Dans cette thèse, nous cherchons à développer de nouveaux algorithmes pour la classification non supervisée de signaux, en portant un intérêt particulier au cas des images numériques.

Le but de la classification non supervisée, telle que nous la concevons, est de proposer différentes fonctions de classification, dans l'espoir que certaines d'entre elles pourront être utiles en pratique.

Nous prendrons comme point de départ l'algorithme des k -means dans un espace Euclidien. Etant donné un échantillon d'apprentissage $(X_1, \dots, X_n) \in \mathbb{R}^{d \times n}$ et k centres $c_i, 1 \leq i \leq k$, la perte empirique associée à l'algorithme des k -means est donnée par

$$L(c_1, \dots, c_k) = \frac{1}{n} \sum_{i=1}^n \min_{j \in \llbracket 1, k \rrbracket} \|X_i - c_j\|^2 = \inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \bar{\mathbb{P}}_X(\|X - c_{\ell(X)}\|^2),$$

où $\bar{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ représente la mesure empirique associée à l'échantillon. La perte empirique est reliée à la perte théorique

$$\inf_{\ell: \mathbb{R}^d \rightarrow \llbracket 1, k \rrbracket} \mathbb{P}_X(\|X - c_{\ell(X)}\|^2).$$

La première chose que nous allons entreprendre, à partir de là, est de voir ce que nous obtenons en considérant le carré de la norme euclidienne comme la divergence de Kullback de deux gaussiennes. Dans cette interprétation, on ajoute à la variable aléatoire $X \in \mathbb{R}^d$ une seconde variable aléatoire Y telle que $\mathbb{P}_{Y|X} = \mathcal{N}(X, \sigma^2 I_d)$ où I_d est la matrice identité de taille $d \times d$. On obtient que

$$\|X - c_{\ell(X)}\|^2 = 2\sigma^2 \mathcal{K}(Q_{Y|X}, \mathbb{P}_{Y|X}),$$

où $Q_{Y|X} = \mathcal{N}(c_{\ell(X)}, \sigma^2 I_d) = Q_{Y|\ell(X)}$. Il est intéressant de considérer $\mathcal{K}(Q_{Y|X}, \mathbb{P}_{Y|X})$ au lieu de $\mathcal{K}(\mathbb{P}_{Y|X}, Q_{Y|X})$ qui sont égaux dans ce cas, du fait de la propriété suivante.

PROPOSITION 48 *Le critère des k -means euclidiens peut être considéré comme un critère du type information k -means en raison de l'identité suivante*

$$\inf_{c_1, \dots, c_k} \mathbb{P}_X(\|X - c_{\ell(X)}\|^2) = 2\sigma^2 \inf_Q \mathbb{P}_X[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X})].$$

Le point important ici est que nous n'avons pas à restreindre l'infimum à

$$\left\{ Q : Q_{Y|X} = \mathcal{N}(c_{\ell(X)}, \sigma^2 I_d) \right\}.$$

PROOF. La preuve est une conséquence directe de la proposition 11 on page 45. \square

Cette proposition montre que le critère quadratique des k -means est un cas particulier du critère de type information k -means

$$\inf_{\ell: \mathbb{R}^d \rightarrow [1, k]} \inf_Q \mathbb{P} \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right],$$

et de son homologue empirique

$$\inf_{\ell: \mathbb{R}^d \rightarrow [1, k]} \inf_Q \overline{\mathbb{P}} \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right].$$

Ce critère de type information k -means élargit la portée de l'algorithme des k -means de la classification de vecteurs à la classification de distributions. Dans ce cas, l'ensemble de données $X_1, \dots, X_n \sim \mathbb{P}_X^{\otimes n}$ est remplacé par un ensemble de mesures de probabilités conditionnelles p_{X_1}, \dots, p_{X_n} .

Par exemple, dans le domaine de la fouille de textes, les histogrammes représentant la fréquence des mots, appelés *sacs de mots* sont souvent utilisés pour représenter des documents. De même dans le domaine de la vision par ordinateur les images peuvent être représentées par des histogrammes de caractéristiques visuelles. Il convient de noter que les sacs de mots visuels sont souvent issus du résultat d'un algorithme de clustering utilisé au préalable, typiquement l'algorithme des k -means appliqué à un ensemble de patches d'images pour créer un dictionnaire de caractéristiques locales. Pour plus de détails, on peut se référer à [Tsa12].

Dans le cadre de l'information k -means, on essaie d'approcher $\mathbb{P}_{Y|X}$ par $Q_{Y|\ell(X)}$. Ceci est une invitation à considérer comme variante de la même idée l'approximation de la distribution jointe $\mathbb{P}_{X,Y}$ par $Q_{X,Y}$ telle que $Q_{Y|X} = Q_{Y|\ell(X)}$.

D'après le théorème de désintégration, on note que

$$\mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = \mathcal{K}(Q_X, \mathbb{P}_X) + Q_X [\mathcal{K}(Q_{Y|X}, \mathbb{P}_{Y|X})].$$

Ainsi, considérant plus spécifiquement le modèle

$$\mathcal{Q} = \left\{ Q_{X,Y} : Q_X = \mathbb{P}_X, Q_{Y|X} = Q_{Y|\ell(X)}, \ell(X) \in \{1, \dots, k\} \right\},$$

l'information k -means peut être exprimé comme une projection appelée information projection

$$\inf_{Q_{X,Y} \in \mathcal{Q}} \mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = \inf_{\ell: \mathcal{X} \rightarrow [1, k]} \inf_{Q_{Y|\ell(X)} \in \mathcal{M}_+^1(\mathcal{Y})} \mathbb{P}_X \left[\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right].$$

L'information projection, aussi appelée I-projection [Csi75], consiste à projeter une mesure de probabilités P sur un ensemble \mathcal{Q} de distributions de probabilités, en résolvant

$$\inf_{Q \in \mathcal{Q}} \mathcal{K}(Q, P).$$

Ce concept apparaît également dans le théorème de Sanov [Csi84] qui fournit une borne sur la probabilité que la mesure empirique $\bar{\mathbb{P}}_n$ appartienne à un ensemble de distributions de probabilités \mathcal{Q} , soit de manière informelle

$$-\log\left(\mathbb{P}_X^{\otimes n}(\bar{\mathbb{P}}_n \in \mathcal{Q})\right) \sim n \inf_{Q \in \mathcal{Q}} \mathcal{K}(Q, \mathbb{P}_X).$$

La différence entre l'estimation par maximum de vraisemblance, qui peut être écrite comme

$$\hat{\theta}_{\text{MLE}} \in \arg \min_{\theta \in \Theta} \mathcal{K}(\bar{\mathbb{P}}_n, Q_\theta), \quad (\text{B.1})$$

au moins lorsque l'espace d'états est fini, et la I-projection réside dans le fait que la divergence de Kullback Leibler n'est pas symétrique. En d'autres termes, l'estimation par maximum de vraisemblance a tendance à surestimer le support de la distribution des données, tandis que la I-projection a tendance à la sous-estimer. La différence en termes d'estimation du support est très bien illustrée dans le cas gaussien dans [Bis06], figure 10.2 et 10.3, chap 10. D'ailleurs, on peut voir que (B.1) équivaut à la maximisation de l'espérance d'une fonction de perte

$$\hat{\theta}_{\text{MLE}} \in \arg \max_{\theta \in \Theta} \bar{\mathbb{P}}_n \left(\log \left(\frac{dQ_\theta}{d\nu} \right) \right),$$

où ν est une mesure dominante ($Q_\theta \ll \nu$, pour tout $\theta \in \Theta$), alors que sa contrepartie théorique s'écrit

$$\theta_{\text{MLE}}^* \in \arg \max_{\theta \in \Theta} \mathbb{P} \left(\log \left(\frac{dQ_\theta}{d\nu} \right) \right).$$

De la même manière, nous proposerons une fonction de perte pour l'estimation du paramètre de classification $\ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket$ équivalant à la minimisation du critère de l'information k -means. En effet, à partir du lemme 1 et du lemme 6, on peut remarquer que

$$\inf_{Q_{X,Y} : Q_{Y|X} = Q_{Y|\ell(X)}} \mathcal{K}(Q_{X,Y}, \mathbb{P}_{X,Y}) = -\log \sup_{Q_{X,Y} : Q_{Y|X} = Q_{Y|\ell(X)}} \left\{ \mathbb{P}_X \left[\exp \left(-\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X}) \right) \right] \right\}.$$

Cela montre que la minimisation du critère de l'information k -means est liée à la minimisation de l'espérance d'une fonction de perte $\gamma_\ell(X)$

$$\ell^* \in \arg \min_{\ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket} \mathbb{P}_X(\gamma_\ell(X)), \quad (\text{B.2})$$

où $\gamma_\ell(X) = 1 - \exp\left(-\mathcal{K}(Q_{Y|\ell(X)}, \mathbb{P}_{Y|X})\right)$. Cette fonction de perte est complètement observée (nous supposons que $\mathbb{P}_{Y|X}$ est connue) et joue le rôle de $\log\left(\frac{dQ_\theta}{d\nu}\right)$ dans l'approche par maximum de vraisemblance.

Notez que la fonction de perte $\gamma_\ell(X)$ appartient à l'intervalle $[0, 1]$, puisque la divergence de Kullback est toujours positive. Par la même occasion, nous étudierons l'excès de risque

$$\mathbb{P}_X(\gamma_{\hat{\ell}}(X)) - \mathbb{P}_X(\gamma_{\ell^*}(X)),$$

où

$$\hat{\ell} \in \arg \min_{\ell : \mathcal{X} \mapsto \llbracket 1, k \rrbracket} \bar{\mathbb{P}}_X(\gamma_\ell(X)).$$

De plus, l'information projection apparaît dans de nombreux algorithmes d'apprentissage automatique, notamment dans les méthodes variationnelles bayésiennes (VB) pour l'inférence bayésienne. En effet, les méthodes VB essaient d'approcher une distribution a posteriori par I-projection sur une famille donnée de distributions. Ces méthodes représentent une alternative aux méthodes de Monte Carlo par chaînes de Markov qui sont généralement plus lentes. On pourra se référer à [Bis06], [BKM17], [AR20] et [ARC16] pour plus de détails sur ce sujet. Il s'avère que les méthodes VB représentent également un outil attrayant pour le clustering non supervisé, en particulier pour le calcul des auto-encodeurs variationnels (VAE), voir [Doe16] pour une revue complète sur le sujet. Cela apparaît également dans les auto-encodeurs variationnels pour graphe permettant d'effectuer du clustering de noeuds dans un graphe, voir [Sal+19b] et [Sal+19a].

Il convient de souligner que le clustering de distributions de probabilités (conditionnelles) utilisant la divergence de Kullback comme mesure de similarités ou d'autres critères d'information n'est pas un nouveau sujet. Cela a été très largement utilisé pour labelliser des documents, et en particulier pour le clustering de mots permettant d'extraire des caractéristiques ou bien de réduire la dimension de l'espace sous-jacent. Par exemple, [PTL02] introduit ce qu'il appelle le clustering distributionnel consistant à regrouper les noms d'un texte par rapport à la distribution conditionnelle du verbe sachant le nom. Le regroupement est effectué en mesurant la divergence de Kullback entre les distributions conditionnelles sachant les noms et les centroïdes des distributions associés. La distribution centroïde est définie comme une moyenne intra-cluster de distributions conditionnelles minimisant la moyenne de la divergence de Kullback.

Cependant, dans le cas de l'information k -means, nous suivrons une approche différente. Nous effectuons le regroupement en minimisant la divergence de Kullback par rapport à son premier argument, ce qui conduit à des centroïdes très différents, calculés comme des moyennes géométriques de distributions conditionnelles. Cela représente à notre connaissance une nouveauté par rapport à la littérature existante.

Le regroupement des distributions conditionnelles dans [PTL02] est une version particulière d'un problème plus général appelé *information bottleneck*, voir [TPB01]. Dans la suite, nous verrons que l'information k -means est en fait une variante d'un problème de clustering plus général appelé information fragmentation.

Par ailleurs, dans [Dhi+03], les auteurs proposent un algorithme du type k -means qui diminue une fonction de perte basée sur l'entropie de Jensen-Shannon, exprimée aussi comme une perte d'information mutuelle, conduisant à des centroïdes définis comme des moyennes pondérées de distributions conditionnelles.

En particulier, ils montrent que leur critère d'entropie peut s'exprimer comme un critère de type k -means utilisant la divergence de Kullback comme mesure de distorsion. Plus

formellement, leur critère s'écrit sous la forme

$$\inf_{\ell: \mathcal{Q} \rightarrow \llbracket 1, k \rrbracket} \inf_{q_1, \dots, q_k} \sum_{j=1}^k \sum_{i \in \ell^{-1}(j)} \pi_i \mathcal{K}(p_i, q_j),$$

où \mathcal{Q} est un ensemble de distributions de probabilités discrètes et $\pi_i > 0$ représente certains poids associés à la distribution p_i .

On peut remarquer ici que les centroïdes sont calculés en minimisant la divergence de Kullback par rapport au deuxième argument, de sorte que cela conduit à calculer les centroïdes comme

$$q_j^* = \sum_{i \in \ell^{-1}(j)} \frac{\pi_i p_i}{\sum_{i \in \ell^{-1}(j)} \pi_i},$$

avec ℓ fixé et calculer la meilleure fonction de classification comme

$$\ell^*(i) = \arg \min_{j \in \llbracket 1, k \rrbracket} \mathcal{K}(p_i, q_j^*).$$

De la même manière, [Cao+13] et [Wu12] sont allés plus loin dans cette direction, en étudiant ce qu'ils appellent *Info K-means*. En particulier, [Cao+13] propose un nouvel algorithme pour traiter les problèmes computationnels engendrés par l'Info K -means, en particulier les problèmes qui découlent du calcul de la divergence de Kullback en grande dimension. Ils ont appliqué cet algorithme pour faire du clustering d'images numériques représentant 11 paysages différents. Ils ont préalablement prétraité les images en extrayant des caractéristiques visuelles et en quantifiant ces caractéristiques, afin de considérer chaque image comme un sac de mots visuels. Ensuite, ils ont regroupé les images en utilisant l'Info- K means en prenant $K = 11$ et ont obtenu des résultats prometteurs en retrouvant la partition d'origine donnée par les types de paysages. Il convient de préciser que lorsque nous effectuerons nos expériences, nous n'utiliserons aucune étape de prétraitement sur l'échantillon telle que la sélection de caractéristiques par une méthode de quantification vectorielle. Nous appliquerons directement l'algorithme de l'information fragmentation sur les images numériques originales et cela représente un point fort de notre approche. En outre, on se référera aussi à [Jia+11], qui propose un Algorithme de k -médoïdes pour diminuer une perte du type k -means avec comme mesure de distorsion la divergence de Kullback dans le cas discret et continu, et fournit en plus un estimateur de la divergence de Kullback dans le cas continu. En utilisant les idées de [Dhi+03], [BDG04] présentent un cadre général des k -means basé sur la divergence de Bregman. Les auteurs montrent que ces critères peuvent être minimisés de manière itérative. La distance de Bregman englobe de nombreuses mesures de distorsion telles que la distance euclidienne, la divergence de Kullback, la perte logistique et bien d'autres. Cependant, dans le cas de Kullback, la minimisation est effectuée par rapport au second argument, et non au premier comme dans notre cas.

Pour en revenir à l'information k -means, nous avons vu que si nous choisissons librement la distribution Q_X à la place de \mathbb{P}_X , nous obtenions la fonction de perte bornée donnée par

l'équation (B.2). Dans le cas quadratique, on obtient le critère

$$\inf_{\ell} \mathbb{P}_X \left[1 - \exp \left(-\frac{1}{2\sigma^2} \|X - c_{\ell(X)}\|^2 \right) \right] = \mathbb{P}_X \left[1 - \exp \left(-\frac{1}{2\sigma^2} \min_{j \in \llbracket 1, k \rrbracket} \|X - c_j\|^2 \right) \right].$$

Nous verrons que nous pourrions établir des bornes de généralisation pour ce type de critères robustes sous des hypothèses beaucoup plus faibles que celles imposées par notre premier critère.

Jusqu'à présent, nous avons décrit l'extension des k -means euclidiens au cadre de l'information k -means. La prochaine extension dont nous aimerions parler est celle des k -means à l'information fragmentation. Dans le cas des k -means, nous utilisons un centre/centroïde dans \mathbb{R}^d pour représenter des points voisins. Ce type de classification manque de finesse. On ne s'attend pas vraiment à ce que des images entières (ou plus généralement des signaux entiers) soient semblables. On aimerait plutôt étiqueter *des parties* d'images. Cela peut se faire en étiquetant les pixels de chaque image avec des étiquettes différentes, produisant une *fragmentation* de chacune des images en différentes zones. Dans le cadre habituel des k -means quadratiques, nous proposons d'approcher X par

$$Y = \sum_{j \in A_X} c_j, \text{ où } \llbracket 1, d \rrbracket = \bigsqcup_{j \in A_X} \text{supp}(c_j).$$

Ici A_X remplace $\ell(X)$ et contient les étiquettes des composantes du signal $X \in \mathbb{R}^d$. Ce cadre peut être vu comme une généralisation des k -means euclidiens classiques correspondant à $A_X = \{\ell(X)\}$. Le critère quadratique devient

$$\mathbb{P}_X \left(\left\| X - \sum_{j \in A_X} c_j \right\|^2 \right).$$

En introduisant les projecteurs orthogonaux π_j sur le sous-espace vectoriel engendré par le support de c_j , on peut réécrire le critère comme

$$\mathbb{P}_X \left(\sum_{j \in A_X} \|\pi_j(X) - c_j\|^2 \right) = \sum_{j=1}^k \mathbb{P}_X \left(\mathbf{1}(X \in A_j) \|\pi_j(X) - c_j\|^2 \right),$$

où $A_j = \{x \in \mathbb{R}^d : j \in A_x\}$.

Pour décrire l'information fragmentation en terme d'information projection nous devons introduire une nouvelle variable aléatoire S qui décrit l'emplacement du pixel. Plus précisément, nous remplaçons la représentation Y de X par deux variables aléatoires $S \in \llbracket 1, d \rrbracket$ et $V \in \mathbb{R}$, définies par

$$\mathbb{P}_{S|X} = \frac{1}{d} \sum_{j=1}^d \delta_j \text{ and } \mathbb{P}_{V|X, S=j} = \mathcal{N}(X_j, \sigma^2).$$

Ceci étant dit, nous pouvons décrire le critère quadratique des k -means comme

$$\inf_Q \mathbb{P}_{X, S} \left[\mathcal{K}(Q_{V|S, \ell(X)}, \mathbb{P}_{V|S, X}) \right],$$

tandis que le critère d'entropie globale est

$$\begin{aligned} \inf_{Q: Q_{S,V|X} = \mathbb{P}_S Q_{V|S, \ell(X)}} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}) \\ = -\log \sup_{Q: Q_{S,V|X} = \mathbb{P}_S Q_{V|S, \ell(X)}} \left\{ \mathbb{P}_X \left[\exp \left(-\mathfrak{K}(Q_{S,V|\ell(X)}, \mathbb{P}_{S,V|X}) \right) \right] \right\}. \end{aligned}$$

Pour obtenir le critère de fragmentation quadratique, on a juste besoin d'étendre la fonction de classification ℓ en la faisant dépendre aussi de S , l'emplacement du pixel. On obtient

$$\inf_Q \mathbb{P}_{X,S} \left[\mathfrak{K}(Q_{V|S, \ell(X,S)}, \mathbb{P}_{V|X,S}) \right].$$

Le critère d'entropie globale correspondant est

$$\inf_{Q: Q_{S,V|X, \ell(X,S)} = Q_{S,V|\ell(X,S)}} \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}).$$

Il possède les propriétés intéressantes suivantes

PROPOSITION 49 (CRITÈRE DE FRAGMENTATION GLOBAL) *Considérons k centres $\rho_j \in \mathfrak{M}_+^1(\mathbb{R}^d)$, $1 \leq j \leq k$. Définissons*

$$\mathfrak{T}_2 = \left\{ B \subset \llbracket 1, k \rrbracket : \rho_i \perp \rho_j, i \neq j \in B \right\},$$

l'ensemble des pavages (éventuellement partiels) par des mesures de probabilités mutuellement singulières ρ_j . Le minimum partiel

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V})$$

portant sur toutes les mesures de probabilités Q , telles qu'il existe une fonction mesurable $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$ telle que

$$Q \left[Q_{X|S,V, \ell(X,S)} = \rho_{\ell(X,S)} \right] = 1 \tag{B.3}$$

est égal à

$$\begin{aligned} -\sup_{\ell} \log \mathbb{P}_S \left(\sum_{j=1}^k \mathbb{1} \left[\rho_j(\ell_S^{-1}(j)) = 1 \right] \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right) \\ = -\log \mathbb{P}_S \left(\sup_{B \in \mathfrak{T}_2} \sum_{j \in B} \exp \left[-\mathfrak{K}(\rho_j, \mathbb{P}_{X|S}) - \mathbf{Var}_{\rho_j}(X_S)/(2\sigma^2) \right] \right), \end{aligned}$$

où

$$\begin{aligned} \ell_s : \mathbb{R}^d &\rightarrow \llbracket 1, k \rrbracket \\ x &\mapsto \ell(x, s). \end{aligned}$$

Pour tout choix de ℓ , et en particulier pour le choix optimal, considérant $W = \ell(X, S)$, l'optimum en $Q_{W,S,V}$ est atteint quand

$$\frac{dQ_{W,S,V}}{d\mathbb{P}_{W,S} \otimes \lambda_V} = Z^{-1} \exp \left[-\mathfrak{K}(\rho_W, \mathbb{P}_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2) \right] g_{\sigma, \rho_W(X_S)}(V), \tag{B.4}$$

où λ_V est la mesure de Lebesgue sur \mathbb{R} et

$$g_{\sigma,m}(v) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(v-m)^2}{2\sigma^2}\right).$$

En particulier, pour le choix optimal de $Q_{W,S,V}$, $Q_{V|S,W} = \mathcal{N}(\rho_W(X_S), \sigma^2)$ est une mesure de probabilités gaussienne et

$$\frac{dQ_{S|W}}{dP_{S|W}} = Z_W^{-1} \exp\left[-\mathcal{K}(\rho_W, P_{X|W,S}) - \mathbf{Var}_{\rho_W}(X_S)/(2\sigma^2)\right].$$

D'autre part, considérons k centres $\mu_{S,V}^{(j)} \in \mathcal{M}_+^1(\llbracket 1, d \rrbracket \times \mathbb{R})$, $1 \leq j \leq k$ tels que

$$\mu_{V|S}^{(j)} = \mathcal{N}(\mu_{V|S}^{(j)}(V), \sigma^2), \quad 1 \leq j \leq k.$$

Définissons

$$\mathcal{T}_1 = \left\{ A \subset \llbracket 1, k \rrbracket : \mu_S^{(i)} \perp \mu_S^{(j)}, i \neq j \in A \right\},$$

l'ensemble des pavages par des mesures de probabilités mutuellement singulières $\mu_S^{(j)}$ (ou de manière équivalente par des probabilités mutuellement singulières $\mu_{S,V}^{(j)}$).

Le minimum partiel

$$\inf_Q \mathcal{K}(Q_{X,S,V}, P_{X,S,V})$$

pris sur toutes les mesures de probabilités $Q \in \mathcal{M}_+^1(\Omega)$ telles que, pour une fonction mesurable $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$,

$$Q\left[Q_{S,V|X,\ell(X,S)} = \mu_{S,V}^{(\ell(X,S))}\right] = 1, \quad (\text{B.5})$$

est égal à

$$\begin{aligned} & -\sup_{\ell} \log P_X \left(\sum_{j=1}^k \mathbb{1} \left[\mu_S^{(j)} \left(\ell_X^{-1}(j) \right) = 1 \right] \right. \\ & \quad \times \exp \left\{ -\mathcal{K}(\mu_S^{(j)}, P_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \Bigg) \\ & = -\log P_X \left(\sup_{A \in \mathcal{T}_1} \sum_{j \in A} \exp \left\{ -\mathcal{K}(\mu_S^{(j)}, P_{S|X}) - \mu_S^{(j)} \left[\left(\mu_{V|S}^{(j)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\} \right), \end{aligned}$$

où

$$\begin{aligned} \ell_x &: \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket \\ s &\mapsto \ell(x, s). \end{aligned}$$

Pour toute valeur de ℓ , et en particulier pour la valeur optimale, considérant $W = \ell(X, S)$, le minimum en $Q_{X,W}$ est atteint quand

$$\frac{dQ_{X,W}}{dP_{X,W}} = Z^{-1} \exp \left\{ -\mathcal{K}(\mu_S^{(W)}, P_{S|X,W}) - \mu_S^{(W)} \left[\left(\mu_{V|S}^{(W)}(V) - X_S \right)^2 / (2\sigma^2) \right] \right\}. \quad (\text{B.6})$$

En alternant ces deux opérations d'optimisation partielle, nous pouvons converger vers un minimum local du problème d'optimisation

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}),$$

où l'infimum est pris sur les mesures de probabilités $Q \in \mathfrak{M}_+^1(\Omega)$ satisfaisant, pour une fonction de classification mesurable $\ell : \mathbb{R}^d \times \llbracket 1, d \rrbracket \rightarrow \llbracket 1, k \rrbracket$,

$$Q \left[Q_{X,S,V | \ell(X,S)} = Q_{X | \ell(X,S)} \otimes Q_{S,V | \ell(X,S)} \right] = 1. \quad (\text{B.7})$$

Pour la preuve, voir la proposition 16. La seconde partie de la proposition montre que le critère d'entropie peut être interprété comme l'espérance par rapport à \mathbb{P}_X d'un risque. Cette espérance peut être estimée par une espérance par rapport à la mesure empirique $\bar{\mathbb{P}}_X$. La dernière partie de la proposition décrit le pendant de l'algorithme de Lloyd, dans le cas où on remplace la mesure inconnue \mathbb{P}_X par la mesure empirique $\bar{\mathbb{P}}_X$.

Maintenant que nous avons un critère pour la fragmentation, nous avons besoin d'un algorithme exploitant ce critère.

Nous utiliserons le critère

$$\inf_Q \mathfrak{K}(Q_{X,S,V}, \mathbb{P}_{X,S,V}),$$

pour définir une fonction de distorsion. Soit X_1, \dots, X_n un échantillon d'apprentissage i.i.d. Nous pouvons représenter son contenu par la distribution

$$\begin{aligned} \bar{\mathbb{P}}_{I,S,V} &= \left(\frac{1}{n} \sum_{i=1}^n \delta_i \right) \mathbb{P}_{S,V | X=X_I} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \delta_i \right) \left(\frac{1}{d} \sum_{j=1}^d \delta_j \right) \mathcal{N}(X_{I,S}, \sigma^2) \\ &= \frac{1}{nd} \sum_{i=1}^n \sum_{j=1}^d \delta_{i,j} \mathcal{N}(X_{i,j}, \sigma^2). \end{aligned}$$

Ici I est un indice aléatoire à valeurs dans l'intervalle $\llbracket 1, n \rrbracket$. Nous voyons immédiatement que (X_1, \dots, X_n) est une fonction de $\bar{\mathbb{P}}$, puisque

$$X_{i,s} = \bar{\mathbb{P}}_{V | S=s, I=i}(V), \quad i \in \llbracket 1, n \rrbracket, \quad s \in \llbracket 1, d \rrbracket.$$

Considérons un dictionnaire fini $\mathfrak{C} \subset \mathbb{R}$, par exemple $\mathfrak{C} = \{m2^{-8} : m \in \llbracket 0, 255 \rrbracket\}$ qui permet de coder des intensités lumineuses à valeurs dans l'intervalle unité $[0, 1]$ sur huit bits comme c'est souvent le cas. Pour toute fonction de classification

$$\ell : \llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \longrightarrow \llbracket 1, k \rrbracket$$

définie par

$$\ell^{-1}(j) = A_j \times B_j, \quad 1 \leq j \leq k,$$

où $(A_j \times B_j, 1 \leq j \leq k)$ est une partition de $\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket$ et toute famille de centres $(C_j, 1 \leq j \leq k) \in \mathfrak{C}^{d \times k}$, où $\text{supp}(C_j) \subset B_j$, définissons le paramètre

$$\theta = (A_j, B_j, C_j)_{j=1}^k$$

et le modèle correspondant

$$\begin{aligned} \mathfrak{Q}_\theta = \Big\{ Q_{I,S,V} \in \mathfrak{M}_+^1(\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \times \mathbb{R}) \\ : Q_{S,V|I, (I,S) \in A_j \times B_j} = \mathbb{P}_{S|S \in B_j} \mathcal{N}(C_{j,S}, \sigma^2), j \in \llbracket 1, k \rrbracket \Big\}, \end{aligned}$$

où nous rappelons que $\mathbb{P}_S = \frac{1}{d} \sum_{s=1}^d \delta_s$ est connu et est la mesure uniforme sur l'emplacement des pixels. Pour faire un parallèle entre \mathfrak{Q}_θ et le modèle (B.5) défini dans la proposition 49, on peut remarquer que \mathfrak{Q}_θ détermine

$$\mu_S^{(j)} = \mathbb{P}_{S|S \in B_j} \text{ et } \mu_{V|S}^{(j)} = \mathcal{N}(C_{j,S}, \sigma^2).$$

En d'autres termes, $\mathfrak{Q}_\theta \subset \mathfrak{Q}_\ell$, où

$$\mathfrak{Q}_\ell = \left\{ Q_{I,S,V} \in \mathfrak{M}_+^1(\llbracket 1, n \rrbracket \times \llbracket 1, d \rrbracket \times \mathbb{R}) : Q_{I,S,V|\ell(I,S)} = Q_{I|\ell(I,S)} \otimes Q_{S,V|\ell(I,S)} \right\}.$$

Nous définissons la distorsion $D(\theta)$ de la représentation de l'échantillon (X_1, \dots, X_n) par le paramètre θ comme

$$\begin{aligned} D(\theta) &= \inf_{Q \in \mathfrak{Q}_\theta} \left\{ \mathcal{K}(Q_{I,S,V}, \bar{\mathbb{P}}_{I,S,V}) \right\} \\ &= -\log \bar{\mathbb{P}}_I \left(\sum_{j=1}^k \mathbb{1}(I \in A_j) \mathbb{P}_S(B_j) \exp \left\{ -\frac{1}{2\sigma^2} \mathbb{P}_{S|S \in B_j} \left[(X_{I,S} - Y_{I,S})^2 \right] \right\} \right), \end{aligned}$$

conformément à la proposition 49.

Notons que cela fait sens d'optimiser en $Q \in \mathfrak{Q}_\theta$, puisque la quantification de X_i , donnée par

$$Y_{i,s} = Q_{V|S=s, I=i}(V) = \sum_{j=1}^k \mathbb{1}(i \in A_j) C_{j,s}, \quad 1 \leq i \leq n, \quad 1 \leq s \leq d,$$

ne dépend pas de $Q \in \mathfrak{Q}_\theta$, mais seulement de θ . En fait, elle ne dépend pas de $Q_{I,S}$ si bien que nous aurions pu optimiser encore plus dans la définition de $D(\theta)$.

Remarquons que cette notion de distorsion satisfait

$$\inf_{Q \in \mathfrak{Q}_\ell} \mathcal{K}(Q_{I,S,V}, \bar{\mathbb{P}}_{I,S,V}) \leq D(\theta) \leq \bar{\mathbb{P}}_{I,S} \left[(X_{I,S} - Y_{I,S})^2 \right] = \frac{1}{nd} \|X_1^n - Y_1^n\|^2.$$

Etant donnée une distribution de codage $q(\theta)$ et un niveau de distorsion acceptable $\eta \geq 0$, l'algorithme de fragmentation calcule une représentation avec perte $\hat{\theta}(X_1, \dots, X_n)$ dont la distorsion vérifie $D(\hat{\theta}) \leq \eta$ et dont la longueur de code idéale $-\log(q(\hat{\theta}))$ est aussi faible que possible.

Nous utiliserons une probabilité de codage de la forme

$$q(\theta) = q(A, B, C) = q(A) q(B, C).$$

A la suite de cette étape de fragmentation, conduisant au calcul de $\hat{\theta}$, nous effectuerons une analyse syntaxique visant à remplacer le code idéal $q(A)$ par un code plus efficace $\tilde{q}(A)$. Cette amélioration suit l'approche bayésienne de Shtarkov. Plus précisément, nous considérons une famille $q_\alpha(A)$ de distributions de codage dépendant d'un nouveau paramètre α et d'une loi a priori $\mu(\alpha)$, et nous améliorons $q(A)$ en considérant

$$\tilde{q}(A) = \max_{\alpha} \mu(\alpha) q_\alpha(A).$$

Comme

$$\bar{q}(A) = \sum_{\alpha} \mu(\alpha) q_\alpha(A)$$

est une mesure de probabilités, \tilde{q} est une sous-probabilité, et donc une mesure de codage valide. D'un point de vue bayésien, $\tilde{q}(A)$ peut aussi être considéré comme un estimateur du maximum de vraisemblance a posteriori (MAP). Introduisons

$$\hat{\alpha} \in \arg \max_{\alpha} \mu(\alpha) q_\alpha(\hat{A}).$$

Nous voyons que $\tilde{q}(\hat{A}) = \mu(\hat{\alpha}) q_{\hat{\alpha}}(\hat{A})$ et que $\hat{\alpha}$ est une fonction de l'échantillon (X_1, \dots, X_n) , puisque c'est le cas de \hat{A} . Quand nous détaillerons la construction de $\hat{\alpha}$ nous verrons que nous effectuons une forme d'analyse syntaxique conduisant en particulier au calcul d'un arbre syntaxique pour chaque image X_i , $1 \leq i \leq n$ de l'échantillon.

References

- [AR20] Pierre Alquier and James Ridgway. “Concentration of tempered posteriors and of their variational approximations.” In: *Ann. Statist.* 48.3 (June 2020), pp. 1475–1497. DOI: [10.1214/19-AOS1855](https://doi.org/10.1214/19-AOS1855). URL: <https://doi.org/10.1214/19-AOS1855>.
- [ARC16] Pierre Alquier, James Ridgway, and Nicolas Chopin. “On the properties of variational approximations of Gibbs posteriors.” In: *Journal of Machine Learning Research* 17.236 (2016), pp. 1–41. URL: <http://jmlr.org/papers/v17/15-290.html>.
- [Aro50] Nachman Aronszajn. “Theory of reproducing kernels.” In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.
- [BDG04] Arindam Banerjee, Inderjit Dhillon, and Joydeep Ghosh. “Clustering with Bregman Divergences.” In: *Journal of Machine Learning Research* 6 (June 2004). DOI: [10.1137/1.9781611972740.22](https://doi.org/10.1137/1.9781611972740.22).
- [BDL08] Gérard Biau, L. Devroye, and G. Lugosi. “On the performance of clustering in Hilbert spaces.” In: *IEEE Transactions on Information Theory* 54.2 (2008), pp. 781–790. URL: <https://hal.archives-ouvertes.fr/hal-00290855>.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.
- [BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians.” In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: [10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773). eprint: <https://doi.org/10.1080/01621459.2017.1285773>. URL: <https://doi.org/10.1080/01621459.2017.1285773>.
- [BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [BFL20] Claire Bréchet, Aurélie Fischer, and Clément Levrard. “Robust Bregman Clustering.” working paper or preprint. Apr. 2020. URL: <https://hal.archives-ouvertes.fr/hal-01948051>.

- [Cao+13] Jie Cao, Zhiang Wu, Junjie Wu, and Wenjie Liu. “Towards information-theoretic K-means clustering for image indexing.” In: *Signal Processing* 93.7 (2013), pp. 2026–2037.
- [Cat04] O. Catoni. *Statistical learning theory and stochastic optimization. Ecole d’été de probabilités de Saint-Flour XXXI-2001*. Collection : Lecture notes in mathematics n°1851. Springer, 2004, pp. viii–272. URL: <https://hal.archives-ouvertes.fr/hal-00104952>.
- [Cat12] Olivier Catoni. “Challenging the empirical mean and empirical variance: A deviation study.” In: *Ann. Inst. H. Poincaré Probab. Statist.* 48.4 (Nov. 2012), pp. 1148–1185. DOI: [10.1214/11-AIHP454](https://doi.org/10.1214/11-AIHP454). URL: <https://doi.org/10.1214/11-AIHP454>.
- [Cat14] Olivier Catoni. *Lecture notes for the IFCAM Summer School on Applied Mathematics, Indian Institute of Science*. Bangalore, 2014.
- [CG17] Olivier Catoni and Ilaria Giulini. “Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression.” In: *arXiv preprint arXiv:1712.02747* (2017).
- [CG18] Olivier Catoni and Ilaria Giulini. “Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector.” In: *arXiv preprint arXiv:1802.04308* (2018).
- [CS08] Andreas Christmann and Ingo Steinwart. “Support vector machines.” In: (2008).
- [Coh] Donald L Cohn. *Measure theory*. Springer.
- [Csi75] I. Csiszar. “ I -Divergence Geometry of Probability Distributions and Minimization Problems.” In: *Ann. Probab.* 3.1 (Feb. 1975), pp. 146–158. DOI: [10.1214/aop/1176996454](https://doi.org/10.1214/aop/1176996454). URL: <https://doi.org/10.1214/aop/1176996454>.
- [Csi84] Imre Csiszar. “Sanov Property, Generalized I -Projection and a Conditional Limit Theorem.” In: *Ann. Probab.* 12.3 (Aug. 1984), pp. 768–793. DOI: [10.1214/aop/1176993227](https://doi.org/10.1214/aop/1176993227). URL: <https://doi.org/10.1214/aop/1176993227>.
- [Dhi+03] Dhillon, Inderjit S., Mallela, Subramanyam, Kumar, and Rahul Kumar. “A divisive information theoretic feature clustering algorithm for text classification.” In: *J. Mach. Learn. Res.* 3 (Jan. 2003), pp. 1265–.
- [Doe16] Carl Doersch. “Tutorial on Variational Autoencoders.” In: *ArXiv abs/1606.05908* (2016).
- [Edd13] Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. ISBN 978-1-4614-6867-7. New York: Springer, 2013. DOI: [10.1007/978-1-4614-6868-4](https://doi.org/10.1007/978-1-4614-6868-4).
- [EF11] Dirk Eddelbuettel and Romain François. “Rcpp: Seamless R and C++ Integration.” In: *Journal of Statistical Software* 40.8 (2011), pp. 1–18. DOI: [10.18637/jss.v040.i08](https://doi.org/10.18637/jss.v040.i08). URL: <http://www.jstatsoft.org/v40/i08/>.

- [EF13] Dirk Eddelbuettel and Romain François. “Writing a package that uses Rcpp.” In: (2013). URL: <https://cran.r-project.org/web/packages/Rcpp/vignettes/Rcpp-package.pdf>.
- [Fis10] Aurélie Fischer. “Quantization and clustering with Bregman divergences.” In: *Journal of Multivariate Analysis* 101 (Oct. 2010), pp. 2207–2221. DOI: [10.1016/j.jmva.2010.05.008](https://doi.org/10.1016/j.jmva.2010.05.008).
- [Giu15] Ilaria Giulini. “Generalization bounds for random samples in Hilbert spaces.” PhD thesis. Paris, Ecole normale supérieure, 2015.
- [Jia+11] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin. “Clustering uncertain data based on probability distribution similarity.” In: *IEEE Transactions on Knowledge and Data Engineering* 25.4 (2011), pp. 751–763.
- [Lev13] Clément Levrard. “Non Asymptotic Bounds for Vector Quantization in Hilbert Spaces.” 30 pages, technical proofs are omitted and can be found in the related unpublished paper “Margin conditions for vector quantization”. Oct. 2013. URL: <https://hal.archives-ouvertes.fr/hal-00877564>.
- [Lev15] Clément Levrard. “Nonasymptotic bounds for vector quantization in Hilbert spaces.” In: *Ann. Statist.* 43.2 (Apr. 2015), pp. 592–619. DOI: [10.1214/14-AOS1293](https://doi.org/10.1214/14-AOS1293). URL: <https://doi.org/10.1214/14-AOS1293>.
- [Li+11] Teng Li, Tao Mei, Soo-Ok Kweon, and Xian-Sheng Hua. “Contextual Bag-of-Words for Visual Categorization.” In: *IEEE Trans. Circuits Syst. Video Techn.* 21 (Apr. 2011), pp. 381–392. DOI: [10.1109/TCSVT.2010.2041828](https://doi.org/10.1109/TCSVT.2010.2041828).
- [Llo06] S. Lloyd. “Least Squares Quantization in PCM.” In: *IEEE Trans. Inf. Theor.* 28.2 (Sept. 2006), pp. 129–137. ISSN: 0018-9448. DOI: [10.1109/TIT.1982.1056489](https://doi.org/10.1109/TIT.1982.1056489). URL: <http://dx.doi.org/10.1109/TIT.1982.1056489>.
- [Mai14] Thomas Mainguy. “Markov Substitute Processes : a statistical model for linguistics.” Theses. Université Pierre et Marie Curie - Paris VI, Dec. 2014. URL: <https://tel.archives-ouvertes.fr/tel-01127344>.
- [MPS07] Pascal Massart, Jean Picard, and École d’été de probabilités de Saint-Flour. “Concentration inequalities and model selection.” In: 2007.
- [PTL02] Fernando Pereira, Naftali Tishby, and Lillian Lee. “Distributional Clustering Of English Words.” In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* (May 2002). DOI: [10.3115/981574.981598](https://doi.org/10.3115/981574.981598).
- [Sal+19a] Guillaume Salha, Romain Hennequin, Viet Anh Tran, and Michalis Vazirgiannis. “A Degeneracy Framework for Scalable Graph Autoencoders.” In: *28th International Joint Conference on Artificial Intelligence (IJCAI)*. 2019.

- [Sal+19b] Guillaume Salha, Stratis Limnios, Romain Hennequin, Viet Anh Tran, and Michalis Vazirgiannis. “Gravity-Inspired Graph Autoencoders for Directed Link Prediction.” In: *ACM International Conference on Information and Knowledge Management (CIKM)*. 2019.
- [TPB01] Naftali Tishby, Fernando Pereira, and William Bialek. “The Information Bottleneck Method.” In: *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* 49 (July 2001).
- [Tri16] Bertrand Clarke Tri Le. *Using the Bayesian Shtarkov solution for predictions*. Computational Statistics and Data Analysis, 2016.
- [Tsa12] Chih-Fong Tsai. “Bag-of-Words Representation in Image Annotation: A Review.” In: *International Scholarly Research Notices* 2012 (2012), pp. 1–19.
- [Wu12] Junjie Wu. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.

Titre : Information k -means, fragmentation et analyse syntaxique. Une nouvelle approche de l'apprentissage non supervisé.

Mots clés : Apprentissage non supervisé, Classification, Compression de données, bornes PAC-Bayésiennes, Chaînage, Critère des k -means.

Résumé : Le critère de l'information k -means étend le critère des k -means en utilisant la divergence de Kullback comme fonction de perte. La fragmentation est une généralisation supplémentaire permettant l'approximation de chaque signal par une combinaison de fragments.

Nous proposons un nouvel algorithme de fragmentation pour les signaux numériques se présentant comme un algorithme de compression avec perte.

A l'issue de ce traitement, chaque signal est représenté par un ensemble aléatoire de labels, servant d'entrée à une procédure d'analyse syntaxique, conçue comme un algorithme de compression sans perte.

Cet algorithme, fondé sur deux principes appliqués itérativement, la factorisation et le réétiquetage de configurations fréquentes, produit pour chaque signal un arbre syntaxique fournissant une classification hiérarchique des composantes du signal.

Nous avons testé la méthode sur des images en niveaux de gris, sur lesquelles il a été possible de détecter des configurations translatées ou transformées par une rotation. Ceci donne l'espoir d'apporter une réponse à la reconnaissance invariante par transformations fondée sur un critère de com-

pression très général.

D'un point de vue mathématique, nous avons prouvé deux types de bornes. Tout d'abord, nous avons relié notre algorithme de compression à un estimateur implicite d'un modèle statistique lui aussi implicite, à travers un lemme, prouvant que le taux de compression et le niveau de distorsion de l'un sont reliés à l'excès de risque de l'autre. Ce résultat contribue à expliquer la pertinence de nos arbres syntaxiques.

Ensuite, nous établissons des bornes de généralisation non asymptotiques et indépendantes de la dimension pour les différents critères des k -means et critères de fragmentation que nous avons introduits. Nous utilisons pour cela des inégalités PAC-Bayésiennes appliquées dans des espaces de Hilbert à noyau reproduisant.

Par exemple dans le cas des k -means classiques, nous obtenons une borne en $\mathcal{O}(k \log(k)/n)^{1/4}$ qui fournit la meilleure condition suffisante de consistance, à savoir que l'excès de risque tend vers zéro quand $k \log(k)/n$ tend vers zéro. Grâce à une nouvelle méthode de chaînage PAC-Bayésien, nous prouvons aussi une borne en $\mathcal{O}(\log(n/k) \sqrt{k \log(k)/n})$.

Title : Information k -means, fragmentation and syntax analysis. A new approach to unsupervised machine learning.

Keywords : Unsupervised machine learning, Clustering, Data compression, PAC-Bayesian bounds, Chaining, k -means criterion

Abstract : Information k -means is a new mathematical framework that extends the classical k -means criterion, using the Kullback divergence as a distortion measure. The fragmentation criterion is an even broader extension where each signal is approximated by a combination of fragments instead of a single center.

Using the fragmentation criterion as a distortion measure, we propose a new fragmentation algorithm for digital signals, conceived as a lossy data compression scheme.

Based on the output of the fragmentation algorithm, where each signal is described as a random set of labels, we describe a new syntax model, conceived as a lossless data compression scheme.

Our syntax analysis is based on two principles : factorization and relabeling of frequent patterns. It is an iterative scheme, decreasing at each step as much as possible the length of the representation of the training set. It produces for each signal a syntax tree, providing a multi-level classification of the signal components.

We tested the method on grey level digital images, where it was possible to label successfully translated patterns and rotated patterns. This lets us hope that transformation invari-

ant pattern recognition could be approached in a flexible way using a general purpose data compression criterion.

From a mathematical point of view, we derived two kinds of generalization bounds. First we defined an implicit estimator based on an implicit statistical model, related to our lossy data compression scheme. We proved a lemma relating the data compression rate and the distortion level of the compression algorithm with the excess risk of the statistical estimator. This explains why our syntax trees may be meaningful.

Second, combining PAC-Bayesian lemmas with the kernel trick, we proved non asymptotic dimension-free generalization bounds for the various information k -means and information fragmentation criteria we introduced.

For instance, in the special case of the classical k -means criterion, we get a non asymptotic dimension free generalization bound of order $\mathcal{O}(k \log(k)/n)^{1/4}$ that gives the best sufficient consistency condition, namely that the excess risk goes to zero when $k \log(k)/n$ goes to zero. Using a new kind of PAC-Bayesian chaining, we also proved a bound of order $\mathcal{O}(\log(n/k) \sqrt{k \log(k)/n})$.