Apprentissage à partir de données extrêmes multivariées : application au traitement du langage naturel

Hamid Jalalzai

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAT043

Thèse de doctorat

Learning from Multivariate Extremes: Theory and Application to Natural Language Processing

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 IP Paris Doctoral School (ED IP Paris)
Spécialité de doctorat : Informatique, données, IA

Thèse présentée et soutenue à Palaiseau, le 6 Novembre 2020, par

HAMID JALALZAI

Composition du Jury :

Stéphane Boucheron Professeur, Université de Paris	Président du jury
Pablo Piantanida Professeur, Université Paris-Saclay	Rapporteur
Olivier Wintenberger Professeur, Sorbonne Université	Rapporteur
Anja Janßen Professeure, Otto-von-Guericke University	Examinateur
Matthieu Labeau Maître de conférence, Télécom Paris	Examinateur
Chloé Clavel Professeure, Télécom Paris	Directrice de thèse
Anne Sabourin Maître de conférence, Télécom Paris	Co-directrice de thèse
Stephan Clémençon Professeur, Télécom Paris	Invité
Eric Gaussier Professeur, Université Grenoble Alpes	Invité

Contents

List of Figures

List of Tables

Remerciements

Mes premiers remerciements vont à mes encadrants : Anne, Chloé et Eric. Anne, merci pour ta présence. Je te remercie pour le savoir que tu m'as transmis, travailler avec toi à l'interface entre la théorie des valeurs extrêmes et la théorie de l'apprentissage statistique a été un vrai plaisir. Eric, merci pour ton investissement et ton énergie incroyable. Chloé, merci de m'avoir accompagné durant cette thèse : la liberté et la confiance que tu as su m'accorder n'ont pu qu'accroître ma motivation.

Je suis également extrêmement reconnaissant envers Pablo Piantanida et Olivier Wintenberger de s'être plongés dans mon manuscrit de thèse et de m'avoir ainsi fait profiter de leur expertise. Je suis très reconnaissant envers Anja Janßen, Stephan Clémençon et Matthieu Labeau pour avoir accepté de faire partie de mon jury. J'ai également été honoré que Stéphane Boucheron préside mon jury. Chacune de ces rencontres conforte mon engouement pour les mathématiques. J'ai une pensée toute particulière dans ces remerciements pour mes co-auteurs: Stephan et Johan pour leur bienveillance et leurs conseils précieux mais également Pierre et Rémi qui, j'en suis sûr, mèneront leurs thèses à bien.

Je souhaite remercier toute l'équipe stat' de Télécom qui a pu offrir le cadre idéal pour mener une thèse. Je salue bien sûr tous les thésards en mathématiques appliquées de Télécom rencontrés entre Barrault et Saclay mais également ceux que j'ai pu rencontrer à l'occasion des visites, séminaires et conférences aux quatres coins du monde.

Je remercie également profondément mes *insa*tiables toulousains et toutes les personnes que j'ai rencontrées sur les bancs d'école parisiens ou d'ailleurs, merci pour votre soutien tout au long de cette thèse. Merci en particulier à K. Merci à mes professeurs de lycée, de prépa et d'école qui m'ont accompagné sur la voie des mathématiques et de la recherche et m'ont encouragé dans cette direction. Je salue l'ensemble de la fondation hellénique à Paris, qui aura été ma deuxième maison.

Enfin, je souhaite remercier ma famille pour leur indéfectible soutien. Je pense en premier à mes parents qui m'ont soutenu durant mon cursus scolaire. Je suis reconnaissant envers mon frère qui aura été là avec détermination tout au long de ma thèse ainsi que ma soeur pour sa bienveillance. J'espère un jour voir notre benjamine nous rejoindre, les aînés et moi, parmi les Drs Jalalzai.

List of Symbols

Notation	Description
c.d.f.	cumulative distribution function
r.v.	random variable
i.i.d.	independent and identically distributed
s.t.	subject to
$\emptyset$	empty set
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of nonnegative numbers
$\mathbb{R}^d$	set of d-dimensional real-valued vector
δ_x	Dirac mass at point $x \in \mathbb{R}^d$
$A \setminus B$	relative complement of a set B in A
A^c	complement of a set A
$\lvert A \rvert$	cardinal of a set A
$\bar{F}$	Survival function associated to the *c.d.f.* F
$\mathbb{P}\{\cdot\}$	probability of a random event
$\mathbb{E}\{\cdot\}$	expectation of a random variable
$\mathbb{1}\{\cdot\}$	indicator function of an event
$[\![a, b]\!]$	range of natural numbers between a and b
$\lfloor \cdot \rfloor$	integer part
$[\cdot]_+$	positive part
$Y_{(1)} \leq \ldots \leq Y_{(n)}$	order statistics of $Y_1, \ldots, Y_n$
$\langle \cdot, \cdot \rangle$	inner product
$\lvert\lvert \cdot \rvert\rvert$	an arbitrary norm
$\lvert\lvert \cdot \rvert\rvert_p$	ℓ_p norm

Table 1 – Summary of notations.

Abstract

Extremes surround us and appear in a large variety of data. Natural data like the ones related to environmental sciences contain extreme measurements; in hydrology, for instance, extremes may correspond to floods and heavy rainfalls or on the contrary droughts. Data related to human activity can also lead to extreme situations; in the case of bank transactions, the money allocated to a sale may be considerable and exceed common transactions. The analysis of this phenomenon is one of the basis of fraud detection. Another example related to humans is the frequency of encountered words. Some words are ubiquitous while others are rare. No matter the context, extremes which are rare by definition, correspond to uncanny data. These events are of particular concern because of the disastrous impact they may have. Extreme data, however, are less considered in modern statistics and applied machine learning, mainly because they are substantially scarce: these events are outnumbered –in an era of so-called "*big data*"– by the large amount of classical and *non-extreme data* that corresponds to the bulk of a distribution. Thus, the wide majority of machine learning tools and literature may not be well-suited or even performant on the distributional tails where extreme observations occur.

Through this dissertation, the particular challenges of working with extremes are detailed and methods dedicated to them are proposed. The first part of the thesis is devoted to statistical learning in extreme regions. In Chapter 4, non-asymptotic bounds for the empirical angular measure are studied. Here, a pre-established anomaly detection scheme via minimum volume set on the sphere, is further improved. Chapter 5 addresses empirical risk minimization for binary classification of extreme samples. The resulting non-parametric analysis and guarantees are detailed. The approach is particularly well suited to treat new samples falling out of the convex envelop of encountered data. This extrapolation property is key to designing new embeddings achieving label preserving data augmentation. Chapter 6 focuses on the challenge of learning the latter heavy-tailed (and to be precise *regularly varying*) representation from a given input distribution. Empirical results show that the designed representation allows better classification performance on extremes and leads to the generation of coherent sentences. Lastly, Chapter 7 analyses the dependence structure of multivariate extremes. By noticing that extremes tend to concentrate on particular *clusters* where features tend to be recurrently large simulatenously, we define an optimization problem that identifies the aformentioned subgroups through weighted means of features.

Résumé

Les *extrêmes* apparaissent dans une grande variété de données. Par exemple, concernant les données hydrologiques, les extrêmes peuvent correspondre à des inondations, des moussons voire des sécheresses. Les données liées à l'activité humaine peuvent également conduire à des situations extrêmes, dans le cas des transactions bancaires, le montant alloué à une vente peut être considérable et dépasser les transactions courantes. Un autre exemple lié à l'activité humaine est la fréquence des mots utilisés : certains mots sont omniprésents alors que d'autres sont très rares. Qu'importe le contexte applicatif, les extrêmes qui sont rares par définition, correspondent à des données particulières. Ces événements sont notamment alarmants au vu de leur potentiel impact désastreux. Cependant, les données extrêmes sont beaucoup moins considérées dans les statistiques modernes ou les pratiques courantes d'apprentissage machine, principalement car elles sont considérablement sous représentées : ces événements se retrouvent noyés - à l'ère du "*big data*" - par une vaste majorité de données classiques et non extrêmes. Ainsi, la grande majorité des outils d'apprentissage machine qui se concentrent naturellement sur une distribution dans son ensemble peut être inadaptée sur les queues de distribution où se trouvent les observations extrêmes.
Dans cette thèse, les défis liés aux extrêmes sont détaillés et l'accent est mis sur le développement de méthodes dédiées à ces données. La première partie se consacre à l'apprentissage statistique dans les régions extrêmes. Dans le chapitre 4, des garanties non asymptotiques sur l'erreur d'estimation de la mesure angulaire empirique sont étudiées et permettent d'améliorer des méthodes de détection d'anomalies par minimum volume set sur la sphère. En particulier, le problème de la minimisation du risque empirique pour la classification binaire dédiée aux échantillons extrêmes est traitée au chapitre 5. L'analyse non paramétrique et les garanties qui en résultent sont détaillées. L'approche est adaptée pour traiter de nouveaux échantillons se trouvant hors de l'enveloppe convexe formée par les données rencontrées. Cette propriété d'extrapolation est l'élément clé et charnière nous permettant de concevoir de nouvelles représentations conservant un label donné et d'ainsi augmenter la quantité de données. Le chapitre 6 se concentre sur l'apprentissage de cette représentation à queue lourde (pour être précis, à *variation régulière*) à partir d'une distribution d'entrée. Les illustrations montrent une meilleure classification des extrêmes et conduit à la génération de phrases cohérentes. Enfin, le chapitre 7 propose d'analyser la structure de dépendance des extrêmes multivariés. En constatant que les extrêmes se concentrent au sein de groupes où les variables explicatives ont tendance à prendre –de manière récurrente– de grandes valeurs simultanément ; il en résulte un problème d'optimisation visant à identifier ces sous-groupes grâce à des moyennes pondérées des composantes.

Part I

Introduction & Preliminaries

Chapter 1

Summary of the Dissertation

1.1 Motivation

In Fisher (1936), Fisher introduces the *Iris dataset* to illustrate Latent Discriminant Analysis. This dataset contains multivariate data related to three species of iris flowers and describes four measurements from each flower sample: the length and the width of both sepals and petals in centimers. If one focuses on the distribution of the sepal's width, Figure 1.1 (bottom) shows that the mass of the empirical distribution is concentrated around the distributional median. In Figure 1.1 (top), four *outliers* are visible on the boxplot. The three outlier samples on the right (*respectively* the sample on the left) are flowers with sepals' width larger (*respectively* smaller) than 1.5 times the third (*respectively* first) quantile to which is added (*respectively* substracted) the inter quantile range.

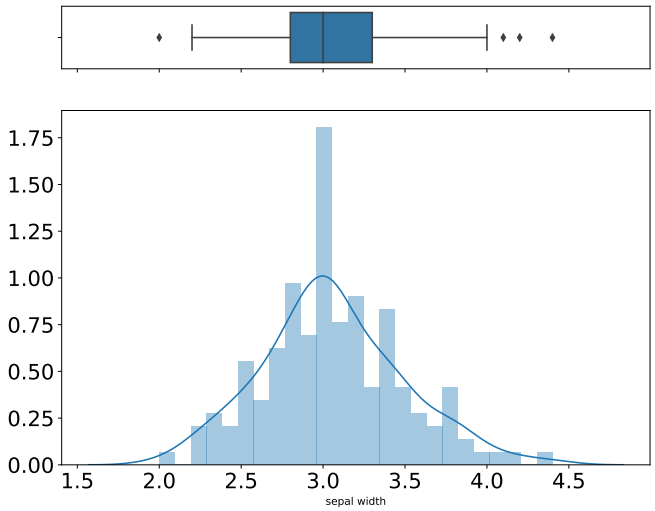


Figure 1.1 – Iris data - Sepal's width distribution and boxplot.

In most textbooks, the common procedure within data pre-processing imply removing the outliers from the input data (or *train data*) to reduce the noise and improve the performance for future inference of any desired model. The point of removing the outliers is to ensure that the focus is set solely on the bulk of samples. Even by not removing extreme samples, most machine learning algorithms neglect the extremes/outliers because of their scarcity. Neglecting the samples distant

from the central bulk can lead to erroneous conclusions. In order to contextualize the work built throughout this dissertation we study a simple example. The example, shown below, consists of two concentric circles with different radia. Bivariate data are labeled differently depending on circle size. Samples from the inner circle are labeled $+1$ (orange dots) while the ones from the outter circle are labeled -1 (blue dots). A handful of samples following a bivariate Cauchy are added to the training observations with label $+1$ (orange dots with a black edge). A regularized tree classifier is trained on this dataset. The background color (orange or blue) is given by the model's predictions. The predictions associated with this model are not as performant on the Cauchy samples as they are for the bulk data. This example shows that, in the setting of binary classification (which will be defined clearly in Section 1.2.3), the samples considered extreme are neglected.

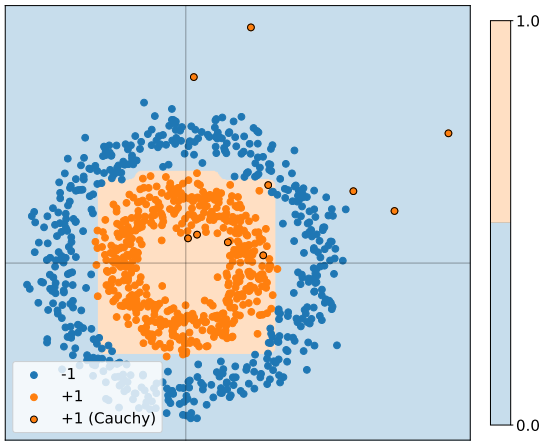


Figure 1.2 – Bivariate centered circles with corresponding prediction from a tree classifier.

Two conclusions can be drawn from this introductory illustration; first, the notion of extreme samples or outliers is not absolute: it is something relative to the main trend of the data. Second, in the context of binary classification, extreme samples may be neglected in the inference process. This can result in unadapted models which may lead to dramatic consequences for classification tasks related to environmental sciences or finance and insurance.

The contributions of this dissertation are introduced in the following sections. First, we provide a brief overview of relevant frameworks. Second, major points of the manuscript's chapters are summarized and contextualized with respect to these frameworks.

1.2 Statistical Framework

We start off by introducing the objects and notions used along this dissertation.

1.2.1 Multivariate Extremes & Regular Variation

Chapter 2 details at length the framework related to working with regularly varying data. We gather in this section the key element. In short, regularly varying random vectors are characterized by their *c.d.f.* and their distributional tails. These tails must be heavier than an exponential distribution (*heavy-tails distributions*) and asymptotically behave like a power law *i.e. regularly vary.*

1.2.2 Text Representation and Analysis

Chapter 3 provides an overview of the evolution of text representation and goes through properties of modern text representation. In a second step, common invariances used in machine learning are evoked.

1.2.3 Statistical Learning - Empirical Risk Minimization

In this section, we want to highlight the paradigm of empirical risk minimization (ERM in abbreviated form) which is a central element in predictive learning and in the rest of this dissertation. In the standard setup, following Devroye and collab. (1996), given Z a random variable (r.v) valued in $\mathcal{Z}$ with distribution P and a measurable loss function ℓ valued in $\mathbb{R}_+$, ERM is designed to minimize the following quantity

$$R_P(\theta) = \mathbb{E}_P[\ell(\theta, Z)], \tag{1.1}$$

known as the *risk* associated to a given parameter $\theta \in \Theta$. In general, Θ is a class of functions or hypothesis $\mathcal{G}$ which may, for example, be valued in a finite and discrete set $\mathcal{Y}$ in the context of *classification* or $\mathbb{R}$ in the *regression* setting.

In the case of classification Z decomposes as a pair (X, Y) valued in $\mathcal{X} \times \mathcal{Y}$ and ℓ is the 0/1 loss. In common learning problems, $\mathcal{X}$ is a subset of $\mathbb{R}^p$ with $p \geq 1$ and X is denoted as the *feature vector* while Y is defined as the *label*. P corresponds to their joint probability distribution. The risk defined in Equation 1.1 then rewrites as

$$R_P(g) = \mathbb{P}\left\{g(X) \neq Y\right\}, \tag{1.2}$$

for any $g \in \mathcal{G}$ valued in $\mathcal{Y}$.

Friedman and collab. (2001, p 21) show that the minimizer of the risk R_P is given by the classifier $x \mapsto \arg\max_{y \in \mathcal{Y}} \mathbb{P}\left\{Y = y \mid X = x\right\}$ for all $x \in \mathcal{X}$. Chapter 2 provides further details in the case of binary classification *i.e.* when $\mathcal{Y} = \{-1, 1\}$.

In most practical situations, the distribution P is unknown and learning must be performed on a proxy of R_P such as its empirical counterpart $\widehat{R}_P$. The empirical counterpart is a M-estimation, which can be penalized, and relies on n independent and identically distributed (*i.i.d*) samples $Z_1, \ldots, Z_n$ drawn from P and is defined as,

$$\widehat{R}_P(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i).$$

For instance, in the case of classification defined above, the empirical version of Equation 1.2 is given by

$$\widehat{R}_P(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{g(X_i) \neq Y_i\},$$

where $\{(X_i, Y_i)\}_{i=1}^n$ correspond to n *i.i.d.* copies of the random pair (X, Y).

Concentration inequalities (see Boucheron and collab. (2013)) are broad tools, used in this dissertation to measure or bound $\sup_{\theta\in\Theta} |\widehat{R}_P(\theta) - R_P(\theta)|$, the maximal deviation between $\widehat{R}_P$ and R_P.

1.3 Statistical Learning in Extreme Regions

In this dissertation, statistical learning is illustrated on two major frameworks known as *supervised* and *unsupervised* learning. Supervised learning corresponds to the learning framework that relies on labeled data (one may consider the random pair (X, Y) from Section 1.2.3) where the inference consists of mapping an input to a desired output based on training data corresponding to a pair given by an *observation* and its *associated label*. Binary classification is the illutration of supervised learning studied in this dissertation. In contrast, unsupervised learning corresponds to the framework where the label is no longer provided and the focus is set on investigating inner properties of the data such as homogeneous structures –known as *clusters* for clustering tasks or detection of anomalies–.

1.3.1 Concentration Bounds for the Empirical Angular Measure with Application to MV Sets Estimation

This section is a summary of Chapter 4 based on Clémençon and collab. (2021). This chapter first describes concentration bounds for the empirical angular measure. We then apply this approach to a statistical learning application related to anomaly detection *via* minimum volume sets.

In the standard multivariate extreme value theory setup, $X = (X_1, \ldots, X_d)$ is a continuous random vector with probability distribution P and marginal cumulative distribution functions $F_j(u) = \mathbb{P}\{X_j \leq u\}$, $u \in \mathbb{R}$, and a key assumption is that the distribution of the coordinate-wise maximum of independent copies of the r.v. X lies in the domain of attraction of an extreme value distribution, see *e.g.* De Haan and Ferreira (2007a). This assumption implies in particular that, after standardization $V_j = (1 - F_j(X_j))^{-1}$ of each component of the r.v. X into unit-Pareto margins, one obtains a standard regularly varying random vector $V = (V_1, \ldots, V_d)$ with tail index equal to one: $v\mathbb{P}\{V_j > v\} = 1$, for all $v > 1$. The working assumption in Chapter 4, referred to as the *multivariate regular variation* hypothesis, is that a similar homogeneity property holds jointly at extreme levels,

i.e. that there exists a positive Radon measure μ on the starred positive orthant $E = [0, \infty)^d \setminus \{0\}$ such that

$$t\mathbb{P}\left\{t^{-1}V \in A\right\} \to \mu(A) \tag{1.3}$$

for all Borel-measurable sets A bounded away from the origin $0 = (0, \ldots, 0)$ such that $\mu(\partial A) = 0$, denoting by ∂B the boundary of any Borelian subset $B \subset \mathbb{R}^d$. The measure μ is usually referred to as the exponent measure and determines the distribution of the most extreme observations. The limit measure μ is a homogeneous Radon measure on E, *i.e.* $\mu(\lambda \cdot) = \lambda^{-1}\mu(\cdot)$, for all $\lambda > 0$, whose margins are standardized in the sense that:

$$\forall y \in (0, \infty),\ \forall j \in \{1, \ldots, d\},\ \ \mu(\{x = (x_1, \ \ldots, \ x_d) \in E : x_j \geq y\}) = y^{-1}.$$

In this chapter, we denote by $\|x\| = \max\{|x_1|, \ \ldots, \ |x_d|\}$ the ℓ_∞-norm of any vector $x = (x_1, \ \ldots, \ x_d)$ in $\mathbb{R}^d$. Let $\mathbb{S} = \{x \in [0, \infty)^d : ||x|| = 1\}$ denote the unit sphere restricted to the positive orthant and consider the mapping $\theta : E \to \mathbb{S}$ that assigns to any vector $x \in E$ its angle $\theta(x) = x/\|x\|$. The angular measure Φ is then defined as the push-forward measure of the restriction of μ to $\mathbb{S}$ by θ: for any Borel set $A \subset \mathbb{S}$, we have

$$\Phi(A) = \mu(\mathcal{C}_A) \quad \text{where} \quad \mathcal{C}_A = \{x \in E : \|x\| \geq 1, \ \theta(x) \in A\}. \tag{1.4}$$

The empirical angular measure $\widehat{\Phi}$ is then naturally defined as the empirical distribution of a fraction of the angles $\theta(\widehat{V}_i)$, those corresponding to the k-largest values among the $\|X_i\|$'s namely. The asymptotic study of its accuracy is limited to the two dimensional case so far.

In Clémençon and collab. (2021), nonasymptotic analysis of the empirical angular measure $\widehat{\Phi}$, concentration inequalities are established for the maximal deviations

$$\sup_{A \in \mathcal{A}} |\widehat{\Phi}(A) - \Phi(A)| \tag{1.5}$$

over specific classes $\mathcal{A}$ of Borelian subsets of $\mathbb{S}$. In addition, the authors of Clémençon and collab. (2021) propose a truncated version $\widetilde{\Phi}$ of the empirical angular measure estimator, which discards in contrast the very largest observations, for which it is harder to control the angular stochastic error. The concentration properties of the estimator are essentially preserved, despite the truncation step. It is noteworthy that the results obtained hold true whatever the dimension d, although the upper bounds on the estimation error deteriorate at a linear rate as the dimension d increases. The technical analysis essentially combines the use of framing sets, just like in Einmahl and collab. (2001) and Einmahl and Segers (2009) although they are defined in a different manner here, with concentration inequalities adapted to rare events borrowed from Goix and collab. (2015). The concentration bounds are next applied to a statistical learning problem concerning unsupervised anomaly detection. Following the steps of Thomas and collab. (2017a), anomalies are assumed to be 'unusual' extreme observations here. In this framework, we show that the learning task can be formulated as Minimum Volume set (MV-set) estimation on the sphere $\mathbb{S}$, namely as the statistical recovery of measurable subsets Ω of $\mathbb{S}$ containing a given (very large) fraction α of the

'normal' statistical population of angles $\theta(V)$ with minimum Lebesgue measure: any extreme observation X with angle $\theta(V)$ lying outside the region Ω is then considered as an anomaly. In this case as well, the concentration results established here for the empirical angular measure enables us to get statistical guarantees for empirical MV-set estimation on $\mathbb{S}$, on which this anomaly detection procedure is grounded.

1.3.2 On Binary Classification in Extreme Regions

This section is a summary of Chapter 5 based on Jalalzai and collab. (2018) which is dedicated to binary classification in extreme regions.

Classification can be considered as the flagship problem in statistical learning as it covers a wide range of practical applications and its probabilistic theory can be extended to some extent to various other prediction problems. We recall the standard setup from Section 1.2.3: (X, Y) is a random pair defined on a certain probability space with (unknown) joint probability distribution P, where the (output) r.v. Y is a binary label, taking its values in $\{-1, +1\}$ say, and X models some information, valued in $\mathbb{R}^d$ and hopefully useful to predict Y. In this context, the goal pursued is generally to build a classifier $g : \mathbb{R}^d \to \{-1, +1\}$ minimizing the probability of error $L_P(g) = \mathbb{P}\{Y \neq g(X)\}$. The Empirical Risk Minimization paradigm consists in considering solutions g_n of the minimization problem $\min_{g \in \mathcal{G}} \widehat{L}_n(g)$, where $\widehat{L}_n(g) = (1/n)\sum_{i=1}^n \mathbb{1}\{Y_i \neq g(X_i)\}$ is a statistical estimate of the risk $L(g)$.

Because extreme observations X, *i.e.* observations whose norm $\|X\|$ exceeds some large threshold $t > 0$, are rare and thus underrepresented in the training dataset $\mathcal{D}_n$, classification errors in these regions of the input space may have a negligible impact on the global prediction error of $\widehat{g}_n$. Using the total probability formula, one may indeed write

$$\begin{aligned} L_P(g) =& \mathbb{P}\{\|X\| > t\}\mathbb{P}\{Y \neq g(X) \mid \|X\| > t\} + \\ & \mathbb{P}\{\|X\| \leq t\}\mathbb{P}\{Y \neq g(X) \mid \|X\| \leq t\}. \end{aligned} \tag{1.6}$$

Hence, due to the extremely small order of magnitude of $\mathbb{P}\{\|X\| > t\}$ and of its empirical counterpart, there is no guarantee that the standard ERM strategy produces an optimal classifier on the extreme region $\{x : \ \|x\| > t\}$. In other words the quantity $\mathbb{P}\{Y \neq \widehat{g}_n(X) \mid \|X\| > t\}$ may not be nearly optimal, whereas in certain practical applications (*e.g.* finance, insurance, environmental sciences, aeronautics safety), accurate prediction in extreme regions is crucial.

The purpose of Chapter 5 is to investigate the problem of building a classifier such that the first term of the decomposition (1.6) is asymptotically minimum as $t \to +\infty$. We thus consider the conditional probability of error, which quantity is next referred to as the *classification risk above level* t, given by

$$L_t(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\},$$

We assume that the distributions of X given $Y = +/-1$ are both multivariate regularly varying with both tail index equal to 1. This assumption results from the

common standardization of input vectors X to Pareto margins. The influence of the standardization is first ignored for the sake of simplicity and is later discussed in the chapter, through the concentration bounds studied in Chapter 4. We prove that $\min_g L_t(g)$ converges to a quantity denoted by L_∞^* and referred to as the *asymptotic risk in the extremes*, as $t \to \infty$. It is also shown that this limit can be interpreted as the minimum classification error related to a (non observable) random pair (X_∞, Y_∞), whose distribution P_∞ corresponds to the limit of the conditional distribution of (X, Y) given $\|X\| > t$, for an appropriate normalization of X, as $t \to \infty$. With respect to the goal set above we next investigate the performance of minimizer $\widehat{g}_{n,\tau}$ of an empirical version of the risk $L_{P_{t_\tau}}$, where t_τ is the $(1-\tau)$ quantile of the r.v. $\|X\|$ and $\tau \ll 1$. The computation of $\widehat{g}_{n,\tau}$ involves solely the angular components of the $k \stackrel{\text{def}}{=} \lfloor n\tau \rfloor$ input observations with largest norm, and the minimization is performed over a collection of angular classifiers $\mathcal{G}_S$ (*i.e.* defined on the sphere S associated to the considered norm $||\cdot||$) of finite VC dimension. Based on a variant of the VC inequality tailored to low probability regions, rate bounds for the deviation $L_t(\widehat{g}_{n,\tau}) - L_\infty^*$ are established, of order $O_{\mathbb{P}}(1/\sqrt{n\tau})$.

The theoretical results are also illustrated by experiments based on synthetic and real data, as illustrated in Figure 1.3, where in the case of synthetic data (following a Logistic distribution), the test performance of a regular classifier are worse than the performance of an angular classifier defined on the sphere associated to the ℓ_1 norm denoted as the simplex. Both classifiers are selected from the class of Random Forests (RF) (see Breiman (2001)). The experimental settings are detailed at length in the dedicated section in Chapter 5.

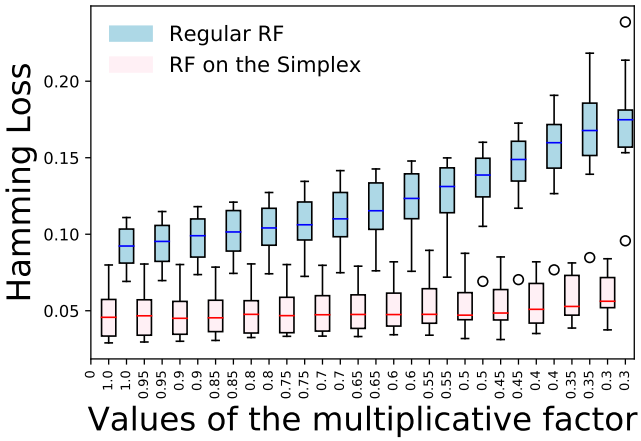


Figure 1.3 – Synthetic data - test loss of RF on the simplex and regular RF depending on the multiplicative factor.

This framework proposes using angular classifiers to perform classification in extreme regions, because it leads to improved classification for new samples that fall outside the convex envelop of the training set. This extrapolation ability is leveraged to perform data augmentation in Chapter 6 whose purpose is to learn a heavy-tailed (and more precisely regularly varying) representation based on any input distribution.

1.4 Learning a Heavy-Tailed Representation & Subspace Clustering in Extremes

We now focus on two chapters: the first one is designed to learn a heavy-tailed representation and the second introduces preliminary work to select subgroups of features which may be large simultaneously.

1.4.1 Heavy-tailed Representations, Text Polarity Classification & Data Augmentation

This section is a summary of Chapter 6 based on Jalalzai and collab. (2020) which is dedicated to learning a heavy-tailed distribution from an input distribution in a classification framework and details a way to leverage the resulting representation to augment text datasets.

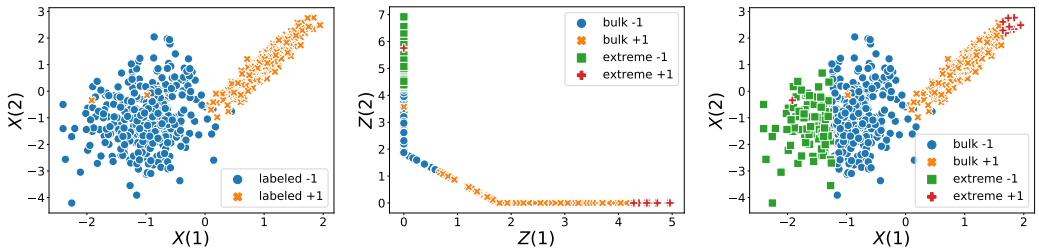
Representing the meaning of natural language in a mathematically grounded way is a scientific challenge that has received increasing attention with the explosion of digital content and text data in the last decade. Relying on the richness of contents, several embeddings have been proposed Devlin and collab. (2018); Peters and collab. (2018); Radford and collab. (2018) with demonstrated efficiency for the considered tasks when trained on massive datasets. However, none of these embeddings take into account the fact that word frequency distributions are heavy tailed Baayen (2002); Church and Gale (1995); Mandelbrot (1953), so that extremes are naturally present in texts. Similarly, Babbar and collab. (2014) shows that, contrary to image taxonomies, the underlying distributions for words and documents in large scale textual taxonomies are also heavy tailed. Exploiting this information, several studies, as Clinchant and Gaussier (2010); Madsen and collab. (2005), were able to improve text mining applications by accurately modeling the tails of textual elements.
In chapter 6, we rely on the framework of multivariate extreme value analysis, based on extreme value theory (EVT) which focuses on the distributional tails. A major advantage of the framework related to EVT in the case of labeled data is that classification on the tail regions may be performed using the angle $\Theta(x) = \|x\|^{-1}x$ only, (see Chapter 5). The main idea here is to take advantage of the scale invariance for two tasks regarding sentiment analysis of text data:

1. Improved classification of extreme inputs, compared to classical methods on the same samples which represent a non negligible proportion of the data (namely, 25% in our experiments),

2. Label preserving data augmentation, as the most probable label of an input x is unchanged by multiplying x by $\lambda > 1$.

The present chapter builds upon the methodological framework proposed in Chapter 5 for classification in extreme regions. However, there is no reason to assume that the previously mentioned text embeddings satisfy the required regularity assumptions. The aim of the present work is to extend Chapter 5's

methodology to datasets which do not satisfy the corresponding assumptions, in particular to text datasets embedded by state of the art techniques. This is achieved by the algorithm *Learning a Heavy Tailed Representation* (in short **LHTR**) which learns a transformation mapping the input data X onto a random vector Z which does satisfy the required assumptions. The transformation, denoted φ, is obtained through an adversarial strategy Goodfellow and collab. (2016) as illustrated in the case of simulated bidimensional data as illustrated in Figure 1.4.



(a) (b) (c)

Figure 1.4 – Figure 1.4a: Bivariate samples X_i in the input space. Figure 1.4b: Latent space representation $Z_i = \varphi(X_i)$. Extremes of each class are selected in the latent space. Figure 1.4c: X_i's in the input space with extremes from each class selected in the latent space.

Our second contribution in Chapter 6 is a novel data augmentation mechanism **GENELIEX** which takes advantage of the scale invariance properties of Z to generate synthetic sequences that keep invariant the attribute of the original sequence. Label preserving data augmentation is an effective solution to the data scarcity problem and is an efficient pre-processing step for moderate dimensional datasets Wang and Perez (2017); Wei and Zou (2019). Adapting these methods to NLP problems remains a challenging issue. The problem consists in constructing a transformation h such that for any sample x with label $y(x)$, the generated sample $h(x)$ would remain label consistent: $y\Big(h(x)\Big) = y(x)$ Ratner and collab. (2017). The dominant approaches for text data augmentation rely on word level transformations such as synonym replacement, slot filling, swap deletion Wei and Zou (2019) using external resources such as wordnet Miller (1995). Linguistic based approaches can also be combined with vectorial representations provided by language models Kobayashi (2018). However, to the best of our knowledge, building a vectorial transformation without using any external linguistic resources remains an open problem. In this work, as the label $y\Big(h(x)\Big)$ is unknown as soon as $h(x)$ does not belong to the training set, we address this issue by learning both an embedding φ and a classifier g satisfying a relaxed version of the problem above mentioned, namely $\forall \lambda \geq 1$,

$$g\Big(h_\lambda(\varphi(x))\Big) = g\Big(\varphi(x)\Big).$$

h_λ is chosen as the homothety with scale factor λ, $h_\lambda(x) = \lambda x$ which will appear coherent with the scale invariance property from Definition 5 in Chapter 2. In Chapter 6, we work with output vectors issued by current state of the art embeddings Devlin and collab. (2018) but we emphasize that the proposed

methodology could equally be applied using any other representation as input (refer to Chapter 3 for further details about text representation). The considered cutting edge embedding does not satisfy the regularity properties required by EVT (refer to Chapter 2 for further elements). Besides, there is no reason why a classifier g trained on such embedding would be scale invariant, *i.e.* would satisfy for a given sequence u, embedded as x, $g(h_\lambda(x)) = g(x)\ \forall\lambda \geq 1$. On the classification task, we demonstrate on two datasets of sentiment analysis that the embedding learnt by **LHTR** on top of pretrained embeddings is indeed following a heavy-tailed distribution. Besides, a classifier trained on the embedding learnt by **LHTR** outperforms the same classifier trained on the input embedding. On the dataset augmentation task, quantitative and qualitative experiments demonstrate the ability of **GENELIEX** to generate new sequences while preserving labels as illustrated in the following example

input	i'm not eating here!
generated sentence ($\lambda = 1$)	i don't eat here.
generated sentence ($\lambda = 1.1$)	i don't eat here!
generated sentence ($\lambda = 1.3$)	i'm not going to eat here!
generated sentence ($\lambda = 1.5$)	i will never going to eat here!

Table 1.1 – **GENELIEX** outputs examples.

Consequently, the overall idea of a text input being extreme is discussed and relevant inner properties of extremes in text are shown. Empirical results provide further explanation of extreme in a text embedded with **LHTR**.

1.4.2 Subspace Clustering for Multivariate Extremes

This section is a summary of Chapter 7 based on Jalalzai and Leluc (2020) which introduces preliminary work to the analysis of features that recurrently have the potential of being large simultaneously.

Clustering is essential for exploratory data mining, data structure analysis and a common technique for statistical data analysis. It is widely used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Many clustering approaches exist with different intrinsic notions of what a cluster is. In the standard setup, the goal is to group objects into subsets, known as clusters, such that objects within a given cluster are more related to one another than the ones from a different cluster. Clustering is already quite well-known (see Bishop (2006); Friedman and collab. (2001) and references therein) conversely to Extreme Value Theory (EVT) which is gaining interest in the machine learning community. It was used in anomaly detection Clifton and collab. (2011); Goix and collab. (2016); Roberts (1999); Thomas and collab. (2017b), classification Jalalzai and collab. (2018, 2020); Vignotto and Engelke (2018) or clustering Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019); Janssen and collab. (2020) when dedicated to the most extreme regions of the sample space.

Scaling up multivariate EVT is a major challenge when addressing high-dimensional learning tasks. Most multivariate extreme value models have been designed to handle moderate dimensional problems, *e.g.*, with dimension $p \leq 10$. For larger dimensions, simplifying modeling choices are needed, stipulating for instance that only some predefined subgroups of components may be concomitant extremes, or, on the contrary, that all must be Sabourin and Naveau (2014); Stephenson (2009). This curse of dimensionality can be explained, in the context of extreme values analysis, by the relative scarcity of extreme data, the computational complexity of the estimation procedure and, in the parametric case, by the fact that the dimension of the parameter space usually grows with that of the sample space.

Recalling the framework of Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019), the goal of this chapter is to present a novel optimization-based approach for clustering extremes in a multivariate setup. Given $N \geq 1$ *i.i.d* copies $X_1, \ldots, X_N$ of a heavy-tailed random variable $X = (X^1, \ldots, X^p)$, we want to identify clusters of features $K \subset [\![1, p]\!]$ such that the variables $\{X^j : j \in K\}$ may be large while the other variables X^j for $j \notin K$ simultaneously remain small. The idea of the resulting optimization problem essentially boils down to finding $w \in \mathbb{R}^p$ belonging to a well chosen subset of the $\mathbb{R}^p$-simplex for any extreme input $X \in \mathbb{R}^p_+$:

$$||X||_1 \approx \langle X, w \rangle.$$

Figure 1.5 illustrates three examples of relevant subsets of the $\mathbb{R}^3$-simplex.

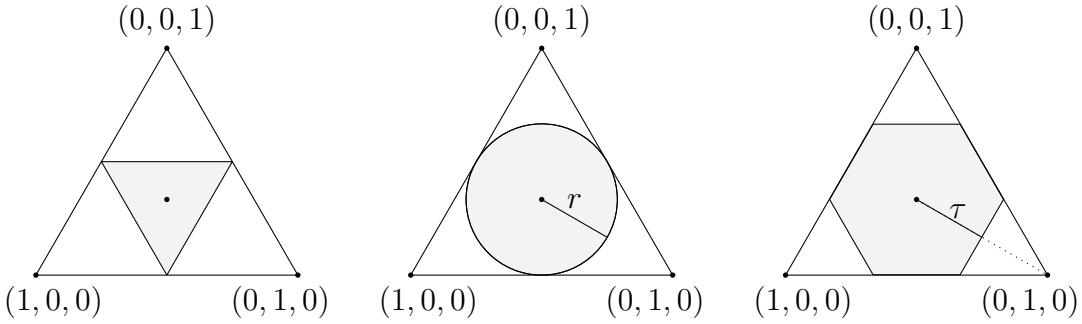


Figure 1.5 – Simplex of $\mathbb{R}^3$ with various subsets of interest.

Up to approximately 2^p combinations of extreme features are possible and contributions such as Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019); Engelke and Hitz (2018); Goix and collab. (2016) tend to identify a smaller number of simultaneous extreme features. Dimensional reduction methods such as principal components analysis and derivatives Cooley and Thibaud (2019); Drees and Sabourin (2019); Tipping and Bishop (1999); Wold and collab. (1987) can be designed to find a lower dimensional subspace where extremes tend to concentrate. Another way of identifying the clusters of features that may jointly be large is to select combinations of extreme features, in the spirit of archetypes defined in Cutler and Breiman (1994). Following this path, the idea of the present chapter is to decompose the ℓ_1-norm of a positive input sample as a weighted mean of its features. Several EVT contributions

are aimed at assessing a sparse support of multivariate extremes Chiapino and Sabourin (2016); De Haan and Ferreira (2007b); Engelke and Ivanovs (2020); Meyer and Wintenberger (2019).

The contributions of this chapter are:

1. we study at length different subsets on the probability simplex,
2. we present a novel optimization-based approach to perform clustering of extreme features in the multivariate set-up,
3. we show how to leverage the obtained clusters to detect outliers within the extreme regions in the context of anomaly detection.

We close this chapter by performing some numerical experiments to highlight the performance of the studied approach on two statistical learning problems: anomaly detection in the extremes and extreme features clustering.

1.5 Outline and Contributions of the Thesis

This dissertation is organized as follows. Part I summarizes the main concepts and contributions of this dissertation and relates them with the relevant backgrounds on both natural language processing and extreme value theory.

- Chapter 2 provides a quick introduction to extreme value theory and invites the reader to focus on the key elements of the literature used along this dissertation.
- Chapter 3 provides an overview on text analysis and current state-of-the-art approaches to text representation.

Part II is devoted to statistical learning in extreme regions.

- Chapter 4 describes a non asymptotic bound for the angular measure and illustrates it in an unsupervised framework related to anomaly detection via minimum-volume sets. This chapter relies on the article in preparation Clémençon and collab. (2021).
- Chapter 5 presents the problem of binary classification in the extremes. This is the cornerstone of this thesis as it is a pivot point to learn a suitable representation for data augmentation. This chapter is based on Jalalzai and collab. (2018).

Part III deals with problems that build upon from statistical learning as

- Chapter 6 exploits the structure of the classifier designed for extreme regions and focuses on learning a heavy-tailed representation by means of a binary classification task. The resulting dilation invariant representation is leveraged to expand datasets with a label consistent approach. This chapter corresponds to the work from Jalalzai and collab. (2020).
- Chapter 7 considers preliminary work concerning the problem of finding relevant directions to represent multivariate extremes. The subspace clustering method relies on the assumption that extreme events are due to particular coordinates' subgroups. This chapter rests on Jalalzai and Leluc (2020).

Part IV concludes the thesis and illuminates perspectives set forth within these contributions.

1.6 References

Baayen, R. H. 2002, *Word frequency distributions*, vol. 18, Springer Science & Business Media. 26

Babbar, R., C. Metzig, I. Partalas, E. Gaussier and M.-R. Amini. 2014, ■On power law distributions in large-scale taxonomies■, *ACM SIGKDD Explorations Newsletter*, vol. 16, n° 1, p. 47–56. 26

Bishop, C. M. 2006, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc. 28

Boucheron, S., G. Lugosi and P. Massart. 2013, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press. 22

Breiman, L. 2001, ■Random forests■, *Machine learning*, vol. 45, n° 1, p. 5–32. 25

Chautru, E. and collab.. 2015, ■Dimension reduction in multivariate extreme value analysis■, *Electronic journal of statistics*, vol. 9, n° 1, p. 383–418. 28, 29

Chiapino, M. and A. Sabourin. 2016, ■Feature clustering for extreme events analysis, with application to extreme stream-flow data■, in *International Workshop on New Frontiers in Mining Complex Patterns*, Springer, p. 132–147. 28, 29, 30

Chiapino, M., A. Sabourin and J. Segers. 2019, ■Identifying groups of variables with the potential of being large simultaneously■, *Extremes*, vol. 22, n° 2, p. 193–222. 28, 29

Church, K. W. and W. A. Gale. 1995, ■Poisson mixtures■, *Natural Language Engineering*, vol. 1, n° 2, p. 163–190. 26

Clémençon, S., H. Jalalzai, A. Sabourin and J. Segers. 2021, ■Concentration bounds for the empirical angular measure with statistical learning applications■, *arXiv preprint arXiv:2104.03966*. 22, 23, 31

Clifton, D. A., S. Hugueny and L. Tarassenko. 2011, ■Novelty detection with multivariate extreme value statistics■, *Journal of signal processing systems*, vol. 65, n° 3, p. 371–389. 28

Clinchant, S. and E. Gaussier. 2010, ■Information-based models for ad hoc ir■, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, p. 234–241. 26

Cooley, D. and E. Thibaud. 2019, ■Decompositions of dependence for high-dimensional extremes■, *Biometrika*, vol. 106, n° 3, p. 587–604. 29

Cutler, A. and L. Breiman. 1994, ■Archetypal analysis■, *Technometrics*, vol. 36, n° 4, p. 338–347. 29

De Haan, L. and A. Ferreira. 2007a, *Extreme value theory: an introduction*, Springer Science & Business Media. 22

De Haan, L. and A. Ferreira. 2007b, *Extreme value theory: an introduction*, Springer Science & Business Media. 30

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova. 2018, ■Bert: Pre-training of deep bidirectional transformers for language understanding■, *arXiv preprint arXiv:1810.04805*. 26, 27

Devroye, L., L. Györfi and G. Lugosi. 1996, *A Probabilistic Theory of Pattern Recognition*, Applications of mathematics : stochastic modelling and applied probability, U.S. Government Printing Office. 21

Drees, H. and A. Sabourin. 2019, ■Principal component analysis for multivariate extremes■, *arXiv preprint arXiv:1906.11043*. 29

Einmahl, J. H., L. De Haan and V. I. Piterbarg. 2001, ■Nonparametric estimation of the spectral measure of an extreme value distribution■, *Annals of Statistics*, p. 1401–1423. 23

Einmahl, J. H. and J. Segers. 2009, ■Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution■, *The Annals of Statistics*, p. 2953–2989. 23

Engelke, S. and A. S. Hitz. 2018, ■Graphical models for extremes■, *arXiv preprint arXiv:1812.01734*. 29

Engelke, S. and J. Ivanovs. 2020, ■Sparse structures for multivariate extremes■, *arXiv preprint arXiv:2004.12182*. 30

Fisher, R. A. 1936, ■The use of multiple measurements in taxonomic problems■, *Annals of eugenics*, vol. 7, n° 2, p. 179–188. 19

Friedman, J., T. Hastie and R. Tibshirani. 2001, *The elements of statistical learning*, Springer series in statistics Springer, Berlin. 21, 28

Goix, N., A. Sabourin and S. Clémençon. 2015, ■Learning the dependence structure of rare events: a nonasymptotic study■, in *Proceedings of the International Conference on Learning Theory, COLT'15*. 23

Goix, N., A. Sabourin and S. Clémençon. 2016, ■Sparse representation of multivariate extremes with applications to anomaly ranking■, in *Artificial Intelligence and Statistics*, p. 75–83. 28, 29

Goodfellow, I., Y. Bengio and A. Courville. 2016, *Deep Learning*, MIT Press. http://www.deeplearningbook.org. 27

Jalalzai, H., S. Clémençon and A. Sabourin. 2018, ■On binary classification in extreme regions■, in *Advances in Neural Information Processing Systems*, p. 3092–3100. 24, 28, 31

Jalalzai, H., P. Colombo, C. Clavel, E. Gaussier, G. Varni, E. Vignon and A. Sabourin. 2020, ■Heavy-tailed representations, text polarity classification & data augmentation■, *arXiv preprint arXiv:2003.11593*. 26, 28, 31

Jalalzai, H. and R. Leluc. 2020, ■Informative Clusters for Multivariate Extremes■, *arXiv e-prints*, arXiv:2008.07365. 28, 31

Janssen, A., P. Wan and collab.. 2020, ■k-means clustering of extremes■, *Electronic Journal of Statistics*, vol. 14, n° 1, p. 1211–1233. 28

Kobayashi, S. 2018, ■Contextual augmentation: Data augmentation by words with paradigmatic relations■, *arXiv preprint arXiv:1805.06201*. 27

Madsen, R. E., D. Kauchak and C. Elkan. 2005, ■Modeling word burstiness using the dirichlet distribution■, in *Proceedings of the 22nd international conference on Machine learning*, p. 545–552. 26

Mandelbrot, B. 1953, ■An informational theory of the statistical structure of language■, *Communication theory*, vol. 84, p. 486–502. 26

Meyer, N. and O. Wintenberger. 2019, ■Sparse regular variation■, *arXiv preprint arXiv:1907.00686*. 30

Miller, G. A. 1995, ■Wordnet: a lexical database for english■, *Communications of the ACM*, vol. 38, n° 11, p. 39–41. 27

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer. 2018, ■Deep contextualized word representations■, in *Proc. of NAACL*. 26

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever. 2018, ■Improving language understanding by generative pre-training■, *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*. 26

Ratner, A. J., H. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré. 2017, ■Learning to compose domain-specific transformations for data augmentation■, in *Advances in neural information processing systems*, p. 3236–3246. 27

Roberts, S. 1999, ■Novelty detection using extreme value statistics■, *Vision, Image and Signal Processing, IEE Proceedings* -, vol. 146, n° 3, p. 124–129. 28

Sabourin, A. and P. Naveau. 2014, ■Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization■, *Comput. Stat. Data Anal.*, vol. 71, p. 542–567. 29

Stephenson, A. 2009, ■High-dimensional parametric modelling of multivariate extreme events■, *Australian & New Zealand Journal of Statistics*, vol. 51, p. 77–88. 29

Thomas, A., S. Clemencon, A. Gramfort and A. Sabourin. 2017a, ■Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere■, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 54, édité par A. Singh and J. Zhu, PMLR, Fort Lauderdale, FL, USA, p. 1011–1019. URL http://proceedings.mlr.press/v54/thomas17a.html. 23

Thomas, A., S. Clemencon, A. Gramfort and A. Sabourin. 2017b, ■Anomaly detection in extreme regions via empirical mv-sets on the sphere.■, in *AISTATS*, p. 1011–1019. 28

Tipping, M. E. and C. M. Bishop. 1999, ■Probabilistic principal component analysis■, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, nº 3, p. 611–622. 29

Vignotto, E. and S. Engelke. 2018, ■Extreme value theory for open set classification–gpd and gev classifiers■, *arXiv preprint arXiv:1808.09902*. 28

Wang, J. and L. Perez. 2017, ■The effectiveness of data augmentation in image classification using deep learning■, *Convolutional Neural Networks Vis. Recognit.* 27

Wei, J. and K. Zou. 2019, ■Eda: Easy data augmentation techniques for boosting performance on text classification tasks■, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 6383–6389. 27

Wold, S., K. Esbensen and P. Geladi. 1987, ■Principal component analysis■, *Chemometrics and intelligent laboratory systems*, vol. 2, nº 1-3, p. 37–52. 29

Chapter 2

Preliminaries on Extreme Value Theory

Chapter abstract

This chapter provides a general background and overview on Extreme Value Theory. It highlights relevant tools and knowledge required to Chapters 4, 5, 6 and 7. Most results and definitions are borrowed from Beirlant and collab. (2006); Embrechts and collab. (2013); Foss and collab. (2011); Mikosch (1999); Nair and collab. (2020); Resnick (1987).

Extremes correspond to observations which appear further down in the tails of heavy-tailed distributions. As Mikosch (1999) states, there is no singular definition of a *heavy-tailed* distribution as these distributions are encountered in a wide variety of fields. Yet, these definitions agree that a heavy tailed distribution is a type of model characterizing deviations of extremes from the majority of data. In this dissertation, we focus on the maxima of random variables, order statistics close to the maximum and distributions with regularly varying tails.

2.1 Univariate Extreme Value Theory

We first define stable distributions and the notion of domain of attraction to introduce required elements to study extreme values.

2.1.1 Stable Distributions & Domain of Attraction

Definition 1 (Stable Random Variable)**.** *A random variable X (or the associated distribution) is said to be* stable *if for X_1, X_2, iid copies of X, and any choice of non-negative constants c_1, c_2, there exists $a(c_1, c_2) > 0$ and $b = b(c_1, c_2) \in \mathbb{R}$ such that the following result in law holds:*

$$c_1 X_1 + c_2 X_2 \stackrel{\mathcal{L}}{=} aX + b. \tag{2.1}$$

By induction on Equation 2.1, this definition extends to the case where $n > 2$ iid copies $(X_i)_{i=1}^n$ $\sum_{i=1}^n X_i \overset{\mathcal{L}}{=} a_n X + b_n$.

Definition 2 (Domain of attraction)**.** *Let $(X_i)_{i=1}^n$ be n independent copies of a random variable X with distribution F. X (or the associated distribution) is said to belong to the* domain of attraction *of a distribution G if there exists $a_n > 0, b_n \in \mathbb{R}$, such that*

$$a_n^{-1}\left(\sum_{i=1}^n X_i + b_n\right) \xrightarrow[n\to\infty]{\mathcal{L}} G. \tag{2.2}$$

holds.

Equation 2.2 from Definition 2, may rewrite as, for any given x belonging to the domain of G,

$$\mathbb{P}\left\{a_n^{-1}\left(\sum_{i=1}^n X_i + b_n\right) \leq x\right\} \xrightarrow[n\to\infty]{} G(x).$$

2.1.2 Extreme Value Distributions

Now that definitions of Domain of Attraction and Stable Random Variable are set, one can note that they are designed for $\sum_{i\leq n} X_i$, the sum of the terms of the sequence $(X_i)_{i\leq n}$. Similar results can be obtained when studying the maxima of the sequence $(X_i)_{i=1}^n$. Consider $M_n = \max_{i\leq n}(X_i)$ the maximum of the sequence $(X_i)_{i=1}^n$.

Definition 3 (Max-Stable distribution)**.** *Any non-degenerate random variable X (or the associated distribution) is said to be* max-stable *if there exist deterministic constants $a_n > 0$, $b_n \in \mathbb{R}$ such that holds:*

$$M_n \overset{\mathcal{L}}{=} a_n X + b_n. \tag{2.3}$$

First, based on Definition 1, max stability can be seen as a analog definition of stability not for the sum of the terms of the sequence $(X_i)_{i=1}^n$ but for M_n, defined as the maximum of the sequence $(X_i)_{i=1}^n$. Second, if Equation 2.3 holds, one can rewrite it as

$$a_n^{-1}\big(M_n - b_n\big) \overset{\mathcal{L}}{=} X.$$

Up to switching the sign of b_n, the latter equality recalls Definition 2. Therefore, it would relevant to study the possible domain of attraction dedicated to the maximum of max-stable distribution.

Definition 4 (Maximum Domain of Attraction)**.** *A random variable X (or the associated distribution) is said to be in the* maximum *domain of attraction of H if there exists $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

$$a_n^{-1}\big(M_n - b_n\big) \xrightarrow[n\to\infty]{\mathcal{L}} H.$$

The latter equation rewrites as, for any x belonging to the domain of definition of H

$$\mathbb{P}\left\{a_n^{-1}\Big(M_n - b_n\Big) \leq x\right\} \xrightarrow[n\to\infty]{} H(x).$$

The following theorem provides the three possible distributional limits for maxima of max-stable distributions and distributions lying in the domain of attraction of an *extreme value distribution*.

Theorem 1 (Fisher–Tippett–Gnedenko)**.** *For any sequence* $(X_i)_{i=1}^n$*, if there exist* $a_n > 0$*,* $b_n \in \mathbb{R}$ *and a non-degenerate distribution* H *(i.e.* H *does not reduce to a Dirac mass) such that*

$$a_n^{-1}\Big(M_n - b_n\Big) \xrightarrow[n\to\infty]{\mathcal{L}} H \tag{2.4}$$

then, up to rescaling the input variable, H *belongs to one of the three following extreme value types:*

- *Gumbel :* $H(x) = \exp\left(-e^{-x}\right)$ *for* $x \in \mathbb{R}$*,*
- *Fréchet :* $H(x) = \exp\left(-x^{-\alpha}\right)$ *for* $x > 0$ *and* $H(x) = 0$ *otherwise,*
- *Weibull :* $H(x) = \exp\left(-(-x)^{\alpha}\right)$ *if* $x < 0$ *and* $H(x) = 1$ *otherwise,*

with $\alpha > 0$*.*

Figure 2.1 illustrates the probability density functions for the three limiting distributions.

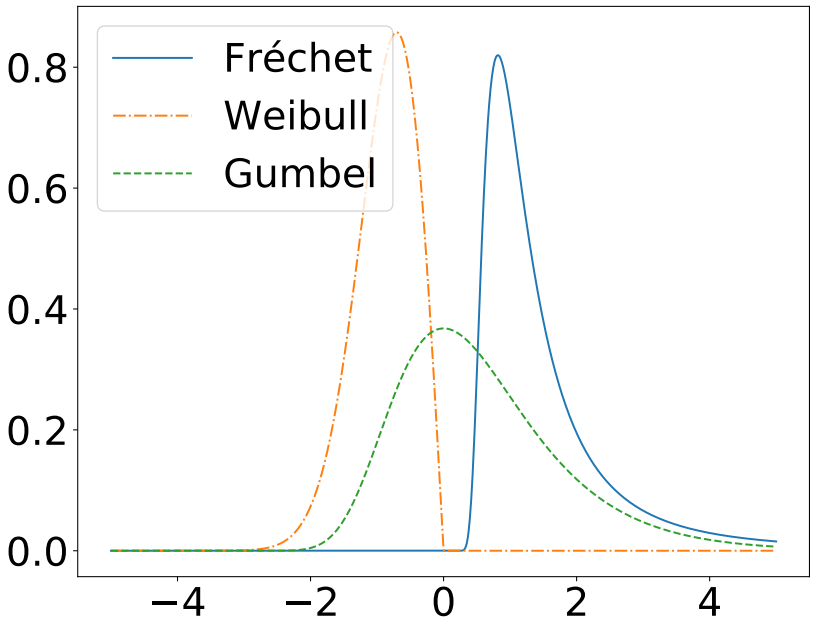


Figure 2.1 – probability Density functions with α set to 2.

This theorem can be seen as an analogous for maxima of the Central Limit Theorem, recall that for X a random variable with finite second moment, then for $a_n = \sqrt{n}$ and $b_n = n\mathbb{E}\{X\}$ then $\frac{\sum_{i=1}^n X_i - b_n}{a_n} \xrightarrow[n\to\infty]{\mathcal{L}} Z$, with Z being a standard Gaussian distribution.

The three distributions are encapsulted in the Generalized Extreme Value Distribution

$$G(x) = \exp\left(-\left[1+\gamma x\right]_+^{-1/\gamma}\right), \qquad \text{with } \gamma \in \mathbb{R}.$$

In the case where $\gamma > 0$, G is a Fréchet distribution. In the case where $\gamma = 0$, G is called a Gumbel distribution and is light tailed. Finally if $\gamma < 0$, G is referred to as a bounded tail Weibull distribution.

Relying on results related to subexponential distributions and regular variation deferred to Section 2.3, Equation 2.4 is equivalent to

$$\lim_{n\to\infty} n\mathbb{P}\left\{\frac{X-b_n}{a_n} \geq x\right\} = -\log H(x) \tag{2.5}$$

for any continuity point $x \in \mathbb{R}$ of G.

In a nutshell, this result relies on subexponential results and the following equivalence when $n \to \infty$

$$-\log\left(\mathbb{P}\left\{\frac{M_n-b_n}{a_n} < x\right\}\right) \sim n\left(1-\mathbb{P}\left\{\frac{X-b_n}{a_n} < x\right\}\right) = n\left(\mathbb{P}\left\{\frac{X-b_n}{a_n} \geq x\right\}\right).$$

as $\mathbb{P}\{X \leq a_n x + b_n\} \to 1$ when $n \to \infty$. The maximum's behavior is asymptotically characterized by a parametric class. Moreover, the behavior of the maximum is mostly related to the behavior of upper data points (or high order statistics in a multivariate setting).

2.1.3 Univariate Regular Variation

The class of regularly varying functions and regularly varying distributions is critical to this dissertation.

Definition 5 (scale invariance)**.** *Let $\bar{F}$ be the survival function associated to a c.d.f F. F is said to be scale invariant if there exists $c > 0$ and a continuous and positive function h such that*

$$\bar{F}(\lambda x) = h(\lambda)\bar{F}(x),$$

for all (λ, x) verifying $\lambda x > c$.

The parameter λ in Definition 5 can be interpreted as a *change of scale* for the considered units. This implies that the shape of $\bar{F}$ remains unchanged up to a multiplicative factor $h(\lambda)$. One can show that F is scale invariant if and only if $\bar{F}$ has a power law tail.

Definition 6. *(Regular variation* Karamata *(1933)) A positive measurable function f is regularly varying with index $\alpha \in \mathbb{R}$, denoted as $f \in \mathcal{RV}_\alpha$, if*

$$\lim_{t\to+\infty} f(tx)/f(t) = x^\alpha \qquad \forall x > 0.$$

In the case where $\alpha = 0$, f is considered as slowly varying.

The notion of regular variation is defined for a random variable X when the function of interest is the distributional tail of X.

Definition 7. *(Univariate regular variation) A non-negative random variable X is regularly varying with tail index $\alpha \geq 0$ if its right distribution tail $x \mapsto \mathbb{P}\{X > x\}$ is regularly varying with index $-\alpha$, i.e.,*

$$\lim_{t \to +\infty} \mathbb{P}\{X > tx \mid X > t\} = x^{-\alpha} \qquad \text{for all } x > 1.$$

Note that belonging to the domain of attraction of Fréchet distributions is equivalent to being regularly varying (see Basrak and collab. (2002)). Jessen and Mikosch (2006) and Basrak and collab. (2002) provide further characterizations and properties of regularly varying distributions. In particular, equivalence between a multivariate random vector being regular varying and the univariate regular variation of all linear combinations of the components of such a vector is demonstrated. This property will be relevant for Chapter 7. Regularly varying functions are homogeneous of factor α. Moreover a random variable X is regularly varying with index α if $\bar{F}(x) = x^{-\alpha}L(x)$ where $L(x)$ is a slowly varying function, so regularly varying functions are asymptotically scale invariant. The scale invariance (or homogeneity) of regularly varying distributions is a key element in the remainder of this dissertation.

Figure 2.2 (top) illustrates that in a case of a Pareto distribution, the survival function's shape is unchanged with respect to the scale of the input, which is not verified for a standard exponential distribution, Figure 2.2 (bottom). Formally, in the case of the Pareto distribution with parameter $\alpha > 0$, the scale invariance of the survival function is clear since,

$$\bar{F}(\lambda x) = \frac{1}{(\lambda x)^{\alpha}} = \frac{1}{\lambda^{\alpha}} \bar{F}(x).$$

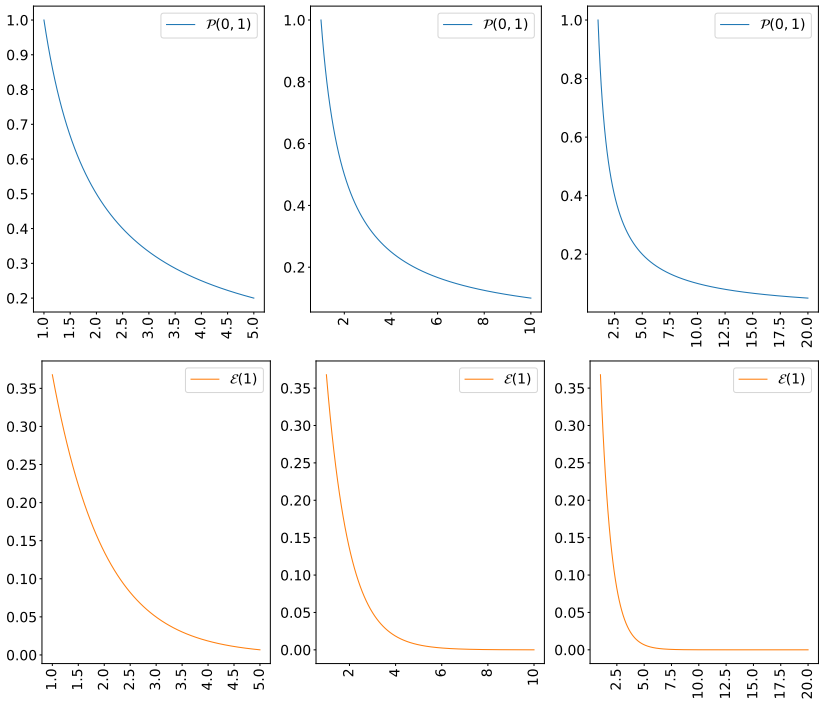


Figure 2.2 – Evolution of $1 - F(x)$ unit Pareto $\mathcal{P}(0,1)$ (top) and a standard Exponential $\mathcal{E}(1)$ (bottom) on varying ranges.

2.2 From Univariate to Multivariate Extremes

The notions and definitions from Section 2.1 can be generalized from real valued random variables to multivariate random variables as the previous definitions no longer hold since $\mathbb{R}^p$ is not an ordered set. In such a case, the notion of maxima must be transcribed to the study of multivariate vectors. We rely on the vague convergence of measures (Resnick, 1987, Section 3.4). Let $X_1, \dots, X_n$ be n copies of a p-dimensional random vector $X = (X^1, \dots, X^p)$ with distribution F, defined as for all $x \in \mathbb{R}^p$, $F(x) = \mathbb{P}\{X^1 < x^1, \dots, X^p < x^p\}$. The notion of maximum is extended to a p-variate vector $M_n = \left(\max_{i \leq n}(X_i^1), \dots, \max_{i \leq n}(X_i^p)\right)$, corresponding to n-th componentwise maximum.

Definition 8 (Multivariate Maximum Domain of Attraction)**.** *A multivariate random variable $X \in \mathbb{R}^p$ (or the associated multivariate distribution), is said to be in the* multivariate maximum *domain of attraction of H if there exists two p-variate sequences $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$, where for all $n \in \mathbb{N}$ elements of a_n are positive and elements of b_n belong to $\mathbb{R}$, such that*

$$a_n^{-1}\left(M_n - b_n\right) \xrightarrow[n \to \infty]{\mathcal{L}} H.$$

where H is a non degenerate multivariate distribution.

This definition implies that the margins $(H^1, \dots, H^p)$ of the limit multivariate distribution H have a univariate extreme value distribution. Although, H alone can no longer be defined through a unique parametric family of distributions. It can be defined through a limiting measure. This major difference leads to the defintion of multivariate regularly varying distributions.

2.2.1 Multivariate Regular Variation

Definition 9. *(Multivariate regular variation) A random vector $X \in \mathbb{R}_+^p$ is regularly varying with tail index $\alpha \geq 0$ if there exist $f \in \mathcal{RV}_{-\alpha}$ and a nonzero Radon measure μ on $E = \mathbb{R}_+^d \setminus \{0\}$ such that*

$$f(t)^{-1}\mathbb{P}\left\{t^{-1}X \in A\right\} \xrightarrow[t\to\infty]{} \mu(A),$$

where $A \subset E$ is any Borel set such that $0 \notin \partial A$ and $\mu(\partial A) = 0$.

The limiting multivariate distribution H from Definition 8 can be characterized by the following equality (Resnick (1987)) for a given $x_0 \in \mathbb{R}^p$

$$H(x) \overset{\text{def}}{=} \begin{cases} \exp\left(\mu([x_0, x]^{\complement})\right) & \text{if } x \geq x_0 \\ 0 & \text{otherwise} \end{cases}$$

where $[x_0, x]^{\complement} = \cup_{i=1}^{p}\left\{u \in [x_0, +\infty] \setminus \{x_0\} : u_j > x_j\right\}$ and μ known as *the exponential distribution*, is a Radon measure defined on $[x_0, +\infty] \setminus \{x_0\}$ is a central element of multivariate regular variation. μ is homogeneous of factor $-\alpha$, $\forall s > 0$, and any Borel set A verifying assumptions of Definition 8, $\mu(sA) = s^{-\alpha}\mu(A)$.

This latter homogeneity is of remarkable importance since it can be leveraged to obtain a pseudo-polar decomposition of μ. For a given $x \in [0, +\infty] \setminus \{0\}$, one can consider the bijective function:

$$\begin{aligned} T_{(r,\theta)} : [0, +\infty]^p \setminus \{0\} &\to (0, +\infty] \times \mathbb{S} \\ x &\mapsto \left(||x||, \frac{x}{||x||}\right), \end{aligned}$$

where $||\cdot||$ is any norm on $\mathbb{R}^p$ and $\mathbb{S}$ is the associated sphere. From this polar decomposition, results a measure Φ defined on the positive orthant of the unit sphere $\mathbb{S}$ associated to the norm $||\cdot||$. Φ, known as the *angular measure* is defined as

$$\Phi(B) \overset{\text{def}}{=} \mu\Big(\{x : ||x|| > 1, \frac{x}{||x||} \in B\}\Big),$$

for any set $B \subset \mathbb{S}$. Φ encapsultes the structure of the tail and is of major importance for the study of the dependence structure of inputs' features. From the homogeneity of μ of degree α,

$$\begin{aligned} \mu(\{x : ||x|| > t, \frac{x}{||x|| \in B}\}) &= \mu(t\{x : ||x|| > 1, \frac{x}{||x|| \in B}\}) \\ &= t^{-\alpha}\Phi(B) \end{aligned}$$

Working with multivariate extremes imply in one case knowing the multivariate sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ from Definition 8 or knowing the tail index α of the exponent measure. One way to work around finding the degree of homogeneity α is to standardize the marginals distributions. Its results that the standardized random variable is regularly varying with tail index equal to 1 *i.e.* that $a_n = n$ and $b_n = 0$.

2.2.2 Marginal Standardization

Similarly to the study of copulas, the natural and preliminary step is to standardize each margin to a known distribution. The choice of the distribution is arbitrary. (Beirlant and collab., 2006, Section 8.2.2) illustrate standardization to unit Fréchet margins and another common standardization is to transform all margins to unit Pareto Resnick (1987). The Pareto standardization is defined as follow for any $x \in \mathbb{R}^p$:

$$T(x_j) \mapsto V_j \overset{\text{def}}{=} \frac{1}{1 - F_j(x_j)}, \qquad \text{for all } j \in \{1, \dots, p\}. \tag{2.6}$$

In practice, the marginal distribution F_j is replaced by its empirical counterpart $\widehat{F_j}$.

$$\widehat{T}(x_j) \mapsto \widehat{V_j} \overset{\text{def}}{=} \frac{1}{1 - \widehat{F_j}(x_j)}, \qquad \text{for all } j \in \{1, \dots, p\}. \tag{2.7}$$

Remark 1. *In practice, in order to avoid division by zero, the empirical cumulative distribution relying on n data samples $\{X_1, \dots, X_n\}$ is reweighted by $n/n+1$ as follow*

$$\widehat{T}(x_j) = \frac{1}{1 - \frac{n}{n+1}\frac{1}{n}\sum_{i \leq n} \mathbb{1}\{X_j \leq x_j\}}.$$

The influence of the empirical standardization resulting from Equation 2.7 over the real standardization from Equation 2.6 is later discussed in Chapter 4. Standardization leads to setting the tail index α equal to 1. For the remainder of this dissertation, the Pareto standardization will be performed on multivariate data $x \in \mathbb{R}^p$ and the resulting vector $V \overset{\text{def}}{=} T(x) = (V^1, \dots, V^p)$ valued in $(1, \infty)^p$ has unit Pareto marginals. In this case, the definition of V being regularly varying may be rewritten:

$$t\mathbb{P}\left\{t^{-1}V \in A\right\} \xrightarrow[t \to \infty]{} \mu(A).$$

for any Borel set A in $[0, \infty]^p \setminus \{0\}$ such that $\mu(\partial A) = 0$ and $0 \notin \partial A$.

2.3 Heavy-Tailed Distributions

When discovering extreme value theory, one may notice that many will call a distribution heavy-tailed, while meaning regularly varying, although the two are not necessarily the same or interchangeable. The purpose of this section is to further explore the principles of heavy-tailed distributions. Most definitions are extracted from Foss and collab. (2011). Among the larger class of *leptokurtic* distributions, heavy-tailed distributions is the class of distributions whose tails are heavier than the exponential distribution. From this point forward, we will focus on the right tail. Definitions and properties can be extended to the left tail by symmetry. Figure 2.3, illustrates the densities of three common distributions: the Pareto, Exponential and Normal densities. The Pareto distribution (green) is heavy-tailed, while the Normal density (blue) is not.

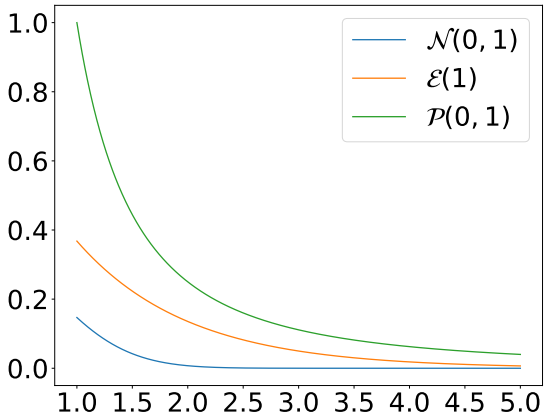


Figure 2.3 – Density plot of univariate standard Normal $\mathcal{N}(0,1)$, standard Exponential $\mathcal{E}(1)$ and standard Pareto $\mathcal{P}(0,1)$ on [1, 5].

Definition 10 (heavy-tail distribution, Foss and collab. (2011))**.** *A real-valued random variable X with distribution F is said to be heavy-tailed if*

$$\int_{-\infty}^{+\infty} e^{tx} F(dt) = +\infty, \qquad \forall x > 0.$$

A consequence of the previous definition is

$$\lim_{t\to\infty} e^{tx} \mathbb{P}\{X > t\} = +\infty, \qquad \forall x > 0.$$

The complementary set of heavy-tailed distributions –*i.e.* the set of distributions failing to be heavy-tailed– is the set of *light-tailed distributions*. Within the set of heavy-tailed distribution, relevant subsets based inner properties are detailed below.

Definition 11 (long-tailed function)**.** *A function f is long tailed if and only if*

$$\lim_{t\to\infty} \frac{f(t+y)}{f(t)} = 1 \qquad \forall y \in \mathbb{R}.$$

It results that a random variable $X \in \mathbb{R}$ is long tailed if and only if

$$\lim_{t\to\infty} \mathbb{P}\{X > x + t \mid X > t\} = 1.$$

A way to understand long-tailed distributions is to consider that if a long-tailed random variable reaches a large value, the random variable will probably reach even larger values. Lemma 2.17 in Foss and collab. (2011) states that any function that is long-tailed is necessarily heavy-tailed although the contrapositive statement is not true. In practical applications distributions are not only long-tailed but possess the additional regularity property of *subexponentiality* which is also known as the catastrophe principle in environmental fields Boxma and Zwart (2007); Nair and collab. (2020) or single big jump in risk theory (George, 2017, Chapter 8).

Definition 12 (subexponential distribution)**.** *A distribution F is said to be subexponential if*

$$\lim_{t\to\infty} \frac{(1-F)^{*n}(t)}{(1-F)(t)} = n \qquad \forall n \in \mathbb{N}, n \geq 2.$$

where we denote by F^{*n}, the n-fold convolution of the distribution F with itself.

Remark 2. *Analysis of the operation of convolution and resulting tail behavior is beyond the scope of this chapter but the interested reader may refer to (*Foss and collab.*, 2011, Chapter 3).*

An interpretation of the latter definition is that, given n independent random variables $X_1, \ldots, X_n$ with common subexponential distribution F,

$$\mathbb{P}\{X_1 + \ldots + X_n > t\} \sim \mathbb{P}\{\max(X_1, \ldots, X_n) > t\} \qquad \text{when } t \to \infty \tag{2.8}$$

"$\sim$" representing in this definition that the ratio of the quantities surrounding this symbol converges to 1. Equation 2.8 translates that the maximum of the sequence is a major contribution to the sequence sum. For a given threshold t, exceedance of t by the sum essentially boils down to exceedance of t by the maximum of the sequence. Lemma 3.2 in Foss and collab. (2011) states that subexponential distributions on $\mathbb{R}^+$ are long-tailed and the converse is not true. Note that the definition of subexponential distribution in Extreme Value Theory collides with the definition from Wainwright (2015), which is more common in the machine learning literature (*e.g.* Duchi and collab. (2012)).

In the case of the sequence $(X_i)_{i\in\{1,\ldots,n\}}$ being composed of n independent copies of X being regularly varying, as defined in Definition 6, one can go further:

$$\begin{aligned}\mathbb{P}\{X_1 + \ldots + X_n > t\} &\sim \mathbb{P}\{\max(X_1, \ldots, X_n) > t\} \\ &\sim n\mathbb{P}\{X_1 > t\} \qquad \text{when } t \to \infty.\end{aligned}$$

As defined earlier in Definition 6 the notion of *regular variation* is a natural way for modelling power law behaviors that appears in various fields of probability theory. We saw that regular variation can be interpreted as an asymptotic scale invariance and such result is a hinge element motivating Chapter 6. From Corollary 1.3.6 and Remark 1.3.7 from Foss and collab. (2011), we can determine that the class of regularly varying distributions is included in the class of subexponential distributions. Figure 2.4 illustrates the sequence of class inclusions from tail properties defined above.

Figure 2.4 illustrates the different classes of distributions and helps the reader to remember that regularly varying distributions are a small within the larger class of heavy-tailed distributions. For instance, Pareto distributions are regularly varying, but log-normal distributions while not regularly varying still belong to the class of subexponential distributions (Embrechts and collab., 2000, p 131-132).

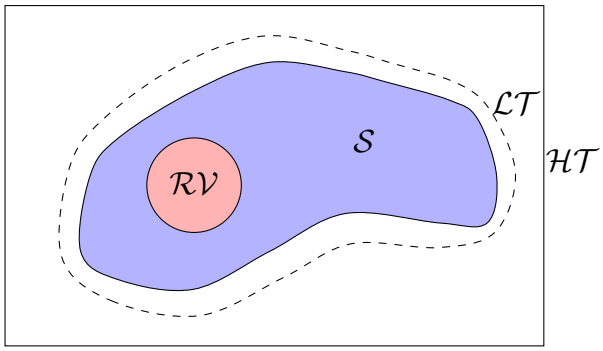


Figure 2.4 – Illustration of the class of regularly varying distributions ($\mathcal{RV}$), included in the class of subexponential distributions ($\mathcal{S}$), included in the class of long-tailed ($\mathcal{LT}$) distributions, included in the class of heavy-tailed distributions ($\mathcal{HT}$).

2.4 References

Basrak, B., R. A. Davis and T. Mikosch. 2002, ■A characterization of multivariate regular variation■, *Annals of Applied Probability*, p. 908–920. 41

Beirlant, J., Y. Goegebeur, J. Segers and J. L. Teugels. 2006, *Statistics of extremes: theory and applications*, John Wiley & Sons. 37, 44

Boxma, O. and B. Zwart. 2007, ■Tails in scheduling■, *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, nº 4, p. 13–20. 45

Duchi, J. C., P. L. Bartlett and M. J. Wainwright. 2012, ■Randomized smoothing for stochastic optimization■, *SIAM Journal on Optimization*, vol. 22, nº 2, p. 674–701. 46

Embrechts, P., L. de Haan and X. Huang. 2000, ■Modelling multivariate extremes■, *Extremes and integrated risk management*, p. 59–67. 46

Embrechts, P., C. Klüppelberg and T. Mikosch. 2013, *Modelling extremal events: for insurance and finance*, vol. 33, Springer Science & Business Media. 37

Foss, S., D. Korshunov, S. Zachary and collab.. 2011, *An introduction to heavy-tailed and subexponential distributions*, vol. 6, Springer. 37, 44, 45, 46

George, K. D. 2017, *Risk Theory: A Heavy Tail Approach*, # N/A. 45

Jessen, H. A. and T. Mikosch. 2006, ■Regularly varying functions■, *Publications de L'institut Mathematique*, vol. 80, nº 94, p. 171–192. 41

Karamata, J. 1933, ■Sur un mode de croissance régulière. théorèmes fondamentaux■, *Bulletin de la Société Mathématique de France*, vol. 61, p. 55–62. 40

Mikosch, T. 1999, *Regular variation, subexponentiality and their applications in probability theory*, Eindhoven University of Technology. 37

Nair, J., A. Wierman and B. Zwart. 2020, ■The fundamentals of heavy tails: Properties, emergence, and estimation■, *Preprint, California Institute of Technology*. 37, 45

Resnick, S. 1987, *Extreme Values, Regular Variation, and Point Processes*, Springer Series in Operations Research and Financial Engineering. 37, 42, 43, 44

Wainwright, M. 2015, ■Basic tail and concentration bounds■, *URl: https://www. stat. berkeley. edu/.../Chap2_TailBounds_Jan22_2015. pdf (visited on 12/31/2017)*. 46

Chapter 3

Preliminaries on Text Analysis

Chapter abstract

This chapter provides a brief recap on natural language processing and the methods dedicated to text representation. This chapter summarises common methods and sets the framework for Chapter 6. It is not designed to be an exhaustive introduction to the field of natural language processing but dwells on representation of textual content as it is an almost routine step for modern text-related machine learning tasks.

Text analysis and representation of text correspond to fields aimed at designing textual embedding. Finding the best suited representation of text data is of major importance in various fields such as sentiment analysis, speech to text, translations and other modern natural language processing tasks. The Deep Learning revolution Goodfellow and collab. (2016) improved state-of-the-art performance for many fields including natural language processing. First, this chapter introduces existing approaches to text analysis and representation in order to depict relevant statistical properties of modern text representation. In a second step common invariances in machine learning are evoked. The objectif of the chapter is to introduce the required and relevant knowledge for Chapter 6 which delineates mathematically funded methods to extend properties and attributes of cutting edge text representation.

3.1 Evolution of Text Representation

Representing the meaning of natural language in a mathematically grounded way is a scientific challenge. One of the first approaches to embed text data was the *time-frequency inverse document frequency* method (in short TFIDF). The TFIDF measures the relative importance of a given word in a document given a corpus of various documents. The high dimensional resulting embedding is then a central element to perform downstream machine learning tasks. The word frequency distributions are heavy-tailed Baayen (2002); Church and Gale (1995); Man-

DELBROT (1953) in common corpuses and are often modeled as following a Zipf distribution: the frequency of any word is inversely proportional to its rank given the corpus. This yields that the most common word occurs twice as often as the second most common one and three times more common than the third most common word, and so on. Zipf law can be used as an introductory example of heavy-tailed distributions. Figure 3.1 illustrates the frequency of words in the IMdB dataset (which is a common dataset for text analysis containing movie reviews). Similarly, BABBAR and collab. (2014) shows that, contrary to image taxonomies, the underlying distributions for words and documents in large scale textual taxonomies are also heavy-tailed. Exploiting this information, several studies, as CLINCHANT and GAUSSIER (2010); MADSEN and collab. (2005), were able to improve text mining applications by accurately modeling the tails of textual elements. As the TFIDF relies on words frequency, it is likely to be a multivariate heavy-tailed representation.

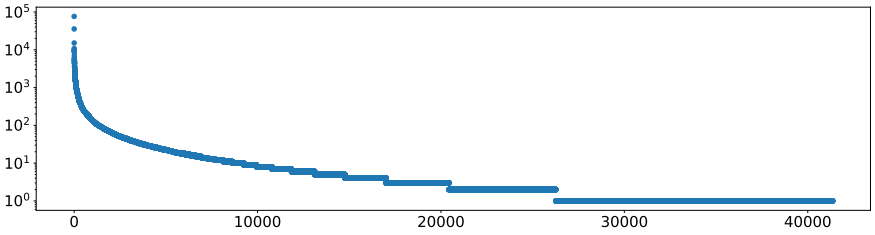


Figure 3.1 – Evolution of the frequency of words (IMdB dataset).

With the recent explosion in data resulting from the rise and success of social platforms or online shopping and streaming websites and concomitant advances in computational power, neural networks have emerged as powerful mean yielding state-of-the-art performance in various fields including text processing and image analysis. Several embeddings issued by various models (such as CER and collab. (2018); DEVLIN and collab. (2018); PETERS and collab. (2018); RADFORD and collab. (2018)) demonstrated a significant performance increase for various tasks when trained on massive datasets. The models can be summarized into two families:

1. Contextual word embeddings (BENGIO and collab. (2003); PETERS and collab. (2018)) such as BERT and derivatives (DEVLIN and collab. (2018); LIU and collab. (2019); YANG and collab. (2019)) which are designed to learn a contextual representation of a word wherein the embedding depends on the sentence where it appears.

2. classical word embeddings (JOULIN and collab. (2016); MIKOLOV and collab. (2013a,b); PENNINGTON and collab. (2014); TURIAN and collab. (2010)), which exploit the *Pointwise Mutual Information* matrix (BOUMA (2009); CHURCH and HANKS (1990)) to learn a dense representation of words by means of matrix approximation (*e.g.* matrix factorization).

The release pace and size of state-of-the-art models have progressively increased over time, as illustrated in Figure 3.2 borrowed from SANH and collab. (2019). No

existing contextual word embeddings, however, account the fact that extremes are naturally present in texts, contrary to the TFIDF representation. It results that the notion of extreme in textual data depends on the chosen representation embedding the underlying text. In the remaining of this dissertation we work with contextual word embeddings.

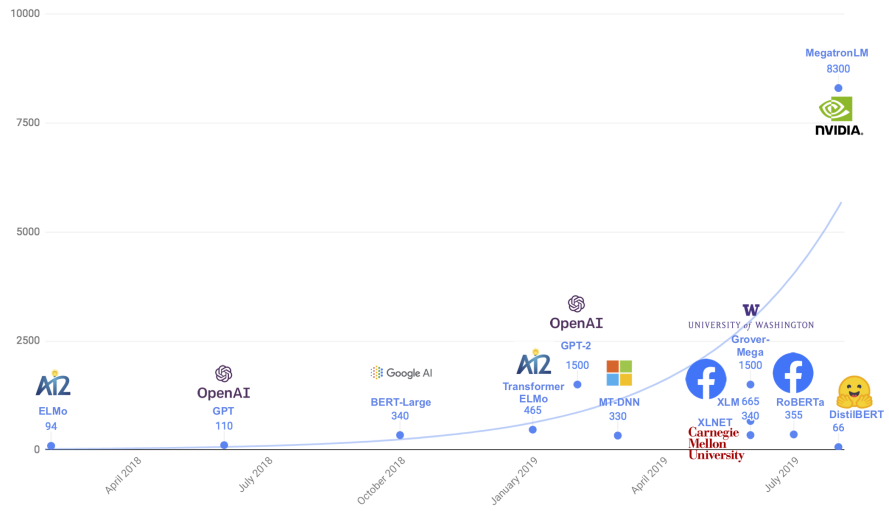


Figure 3.2 – Time evolution of the number of parameters in recent language models (Sanh and collab. (2019)).

3.2 Properties of Text Representation & Common Invariance

Finding the best tailored representation for multiple tasks is a major challenge and modern embeddings show interesting latent semantic properties. Mikolov and collab. (2013b) introduce the *compositionality* property showing that (linear) relations exist between the embeddings of word pairs as illustrated with the well known equation:

$$(\boldsymbol{king} - \boldsymbol{man}) + \boldsymbol{woman} = \boldsymbol{queen}$$

This leads us to treat languages, once embedded, like vector spaces with various mathematical properties and paves the way to miscellaneous contributions designed to further analyse the constructed embeddings Drozd and collab. (2016). The authors of Allen and collab. (2019); Allen and Hospedales (2019) nuance the results mentioned above and further detail the notions of analogies when working with embeddings.
Note that in more recent word embeddings, the resulting vectors contain different information (gender, syntaxe) from previous frequency based representations.Hence, the more recent embeddings provide a paradigm change as the information provided by the representation changed.

3.3 Common Invariance in Machine Learning

A model's invariance represents its aptitude for keeping the model's output constant while change occur to the model's input. The invariance property encapsultes various types of transformations, such as homothetic changes, symmetries and rotations. The concept of invariance is particularly well-suited for image processing, *e.g.* the nature of the content in an image does not depend on the scale or of the image. Such transformations are common in data augmentation Ratner and collab. (2017). In Figure 3.3, the label associated with the image *i.e.* the class of the image, remains the same for different transformations. Various techniques have been developed for label image processing as illustrated in Simard and collab. (1998) for an array of image transformations. Directly transcribing these methods to natural language processing tasks is not effortless and most methods mainly focus on textual data (*e.g.* Kalchbrenner and collab. (2014); Kobayashi (2018)) rather than focusing on text embeddings.

One may also refer to recent publications related to adversarial learning designed to lure models and change models' output/inference while applying minimum change to the original input (Szegedy and collab. (2013)). In the following section, we quickly review adversarial learning and adversarial autoencoders, a relevant tool used in Chapter 6.

(a) original input.

(b) original input after rotation.

(c) original input after dilation.

(d) original input after axial symmetry.

Figure 3.3 – Figures (3.3b, 3.3c, 3.3d) correspond to different tranformations applied to the original Figure 3.3a.

3.4 Adversarial Learning

Adversarial networks, form a system where two neural networks are competing. A first model G, called the generator, generates samples as close as possible to the input dataset. A second model D, called the discriminator, aims at distinguishing samples produced by the generator from the input dataset. The goal of the generator is to maximize the probability of the discriminator making a mistake. Hence, if P_{input} is the distribution of the input dataset then the adversarial network intends to minimize the distance (as measured by the Jensen-Shannon divergence) between the distribution of the generated data P_G and P_{input}. In short, the problem is a minmax game with value function $V(D, G)$

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim P_{\text{input}}}[\log D(x)] + \mathbb{E}_{z \sim P_G}[\log \Big(1 - D(G(z))\Big)].$$

Auto-encoders and derivations Fard and collab. (2018); Goodfellow and collab. (2016); Laforgue and collab. (2018) form a subclass of neural networks whose purpose is to build a suitable representation by learning encoding and decoding functions which capture the core properties of the input data. An adversarial auto-encoder (see Makhzani and collab. (2015)) is a specific kind of auto-encoders where the encoder plays the role of the generator of an adversarial network. Thus the latent code is forced to follow a given distribution while containing information relevant to reconstructing the input. In chapter 6, a similar adversarial encoder constrains the encoded representation to be heavy-tailed.

3.5 References

Allen, C., I. Balazevic and T. Hospedales. 2019, ■What the vec? towards probabilistically grounded embeddings■, in *Advances in Neural Information Processing Systems*, p. 7467–7477. 51

Allen, C. and T. Hospedales. 2019, ■Analogies explained: Towards understanding word embeddings■, *arXiv preprint arXiv:1901.09813*. 51

Baayen, R. H. 2002, *Word frequency distributions*, vol. 18, Springer Science & Business Media. 49

Babbar, R., C. Metzig, I. Partalas, E. Gaussier and M.-R. Amini. 2014, ■On power law distributions in large-scale taxonomies■, *ACM SIGKDD Explorations Newsletter*, vol. 16, nº 1, p. 47–56. 50

Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin. 2003, ■A neural probabilistic language model■, *Journal of machine learning research*, vol. 3, nº Feb, p. 1137–1155. 50

Bouma, G. 2009, ■Normalized (pointwise) mutual information in collocation extraction■, *Proceedings of GSCL*, p. 31–40. 50

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar and collab.. 2018, ■Universal sentence encoder■, *arXiv preprint arXiv:1803.11175*. 50

Church, K. and P. Hanks. 1990, ■Word association norms, mutual information, and lexicography■, *Computational linguistics*, vol. 16, nº 1, p. 22–29. 50

Church, K. W. and W. A. Gale. 1995, ■Poisson mixtures■, *Natural Language Engineering*, vol. 1, nº 2, p. 163–190. 49

Clinchant, S. and E. Gaussier. 2010, ■Information-based models for ad hoc ir■, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, p. 234–241. 50

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova. 2018, ■Bert: Pre-training of deep bidirectional transformers for language understanding■, *arXiv preprint arXiv:1810.04805*. 50

Drozd, A., A. Gladkova and S. Matsuoka. 2016, ■Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen■, in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, p. 3519–3530. 51

Fard, M. M., T. Thonet and E. Gaussier. 2018, ■Deep k-means: Jointly clustering with k-means and learning representations■, *arXiv preprint arXiv:1806.10069*. 54

Goodfellow, I., Y. Bengio and A. Courville. 2016, *Deep Learning*, MIT Press. http://www.deeplearningbook.org. 49, 54

Joulin, A., E. Grave, P. Bojanowski and T. Mikolov. 2016, ■Bag of tricks for efficient text classification■, *arXiv preprint arXiv:1607.01759*. 50

Kalchbrenner, N., E. Grefenstette and P. Blunsom. 2014, ■A convolutional neural network for modelling sentences■, *arXiv preprint arXiv:1404.2188*. 52

Kobayashi, S. 2018, ■Contextual augmentation: Data augmentation by words with paradigmatic relations■, *arXiv preprint arXiv:1805.06201*. 52

Laforgue, P., S. Clémençon and F. d'Alché Buc. 2018, ■Autoencoding any data through kernel autoencoders■, *arXiv preprint arXiv:1805.11028*. 54

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov. 2019, ■Roberta: A robustly optimized bert pretraining approach■, *arXiv preprint arXiv:1907.11692*. 50

Madsen, R. E., D. Kauchak and C. Elkan. 2005, ■Modeling word burstiness using the dirichlet distribution■, in *Proceedings of the 22nd international conference on Machine learning*, p. 545–552. 50

Makhzani, A., J. Shlens, N. Jaitly, I. Goodfellow and B. Frey. 2015, ■Adversarial autoencoders■, *arXiv preprint arXiv:1511.05644*. 54

Mandelbrot, B. 1953, ■An informational theory of the statistical structure of language■, *Communication theory*, vol. 84, p. 486–502. 49

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean. 2013a, ■Distributed representations of words and phrases and their compositionality■, in *Advances in neural information processing systems*, p. 3111–3119. 50

Mikolov, T., W.-t. Yih and G. Zweig. 2013b, ■Linguistic regularities in continuous space word representations■, in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, p. 746–751. 50, 51

Pennington, J., R. Socher and C. D. Manning. 2014, ■Glove: Global vectors for word representation■, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 50

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer. 2018, ■Deep contextualized word representations■, in *Proc. of NAACL*. 50

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever. 2018, ■Improving language understanding by generative pre-training■, *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*. 50

Ratner, A. J., H. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré. 2017, ■Learning to compose domain-specific transformations for data augmentation■, in *Advances in neural information processing systems*, p. 3236–3246. 52

Sanh, V., L. Debut, J. Chaumond and T. Wolf. 2019, ■Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter■, *arXiv preprint arXiv:1910.01108*. 5, 50, 51

Simard, P. Y., Y. A. LeCun, J. S. Denker and B. Victorri. 1998, ■Transformation invariance in pattern recognition—tangent distance and tangent propagation■, in *Neural networks: tricks of the trade*, Springer, p. 239–274. 52

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus. 2013, ■Intriguing properties of neural networks■, *arXiv preprint arXiv:1312.6199*. 52

Turian, J., L. Ratinov and Y. Bengio. 2010, ■Word representations: a simple and general method for semi-supervised learning■, in *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 384–394. 50

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le. 2019, ■Xlnet: Generalized autoregressive pretraining for language understanding■, *arXiv preprint arXiv:1906.08237*. 50

Part II

Statistical Learning in Extreme Regions

Chapter 4

Concentration Bounds for the Empirical Angular Measure with Application to MV Sets Estimation

Chapter abstract

As emphasized in Chapter 2, in multivariate extreme value theory, the angular measure Φ on the unit sphere characterizes the (first order) dependence structure of the components of any heavy-tailed multivariate random variable $\mathbf{X}$. Insofar as it carries most of the information related to the extremal behaviour of the random variable $\mathbf{X}$, the statistical recovery of the angular measure is of crucial importance in many applications. In the common situation when $\mathbf{X}$'s components have different tail indices, the *rank-transformation* offers a convenient and popular way of standardizing i.i.d. observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, copies of the random variable $\mathbf{X}$ in order to build an empirical version $\widehat{\Phi}$ of the angular measure based on the $k \leq n$ most extreme observations. However, the resulting structure of the empirical functional $\widehat{\Phi}$ is complex, due to the strong dependence of the k terms averaged to form it, and the study of its concentration properties is challenging.

This chapter aims to present and illustrate the influence of the marginal standardization to unit Pareto relying on nonasymptotic bounds for the maximal deviations $\sup_{A \in \mathcal{A}} |\widehat{\Phi}(A) - \Phi(A)|$ over classes $\mathcal{A}$ of Borelian subsets of the unit sphere of controlled complexity. In addition, we study a variant of the classic angular measure estimator, based on the observations of intermediate order of magnitude rather than on the extremes. The bounds are used as a key leverage for a statistical learning application related and anomaly detection in extreme regions, *via* minimum-volume sets estimation on the sphere, in order to obtain generalization guarantees for decision rules learnt by means of the empirical risk minimization principle. The theoretical results are also supported by illustrative numerical experiments.

4.1 Introduction

Estimation and prediction problems related to the extremal behaviour of a multivariate random vector $\mathbf{X}$, taking its values in $\mathbb{R}^d$ say, is of crucial importance in a wide variety of applications, ranging from risk assessment in environmental sciences to the analysis of 'weak signals' in machine-learning for instance. In the standard multivariate extreme value theory (MEVT in short) setup, $\mathbf{X} = (X_1, \ldots, X_d)$ is a continuous random vector with probability distribution P and marginal cumulative distribution functions $F_j(u) = \mathbb{P}\{X_j \leq u\}$, $u \in \mathbb{R}$, and a key assumption is that the distribution of the coordinate-wise maximum of independent copies of the r.v. $\mathbf{X}$ lies in the domain of attraction of an extreme value distribution, see *e.g.* De Haan and Ferreira (2007). This assumption implies in particular that, after standardization $V_j = (1 - F_j(X_j))^{-1}$ of each component of the r.v. $\mathbf{X}$ into unit-Pareto margins, one obtains a standard regularly varying random vector $\mathbf{V} = (V_1, \ldots, V_d)$ with tail index equal to one: $v\mathbb{P}\{V_j > v\} = 1$, for all $v > 1$. The working assumption in this chapter —detailed in Chapter 2—, referred to as the *multivariate regular variation* hypothesis, is that a similar homogeneity property holds jointly at extreme levels, *i.e.* that there exists a positive Radon measure μ on the starred positive orthant $E = [0, \infty)^d \setminus \{\mathbf{0}\}$ such that

$$t\mathbb{P}\left\{t^{-1}\mathbf{V} \in A\right\} \to \mu(A) \tag{4.1}$$

for all Borel-measurable sets A bounded away from the origin $\mathbf{0} = (0, \ldots, 0)$ and such that $\mu(\partial A) = 0$, denoting by ∂B the boundary of any Borelian subset $B \subset \mathbb{R}^d$. The measure μ is usually referred to as the exponent measure and determines the distribution of the most extreme observations. Equation (4.1) guarantees that the weak convergence of $\mu_t = t\mathbb{P}\{t^{-1}\mathbf{V} \in A\}$ towards μ as t tends to infinity in the space $\mathcal{M}_0(E)$ of positive Radon measures on E holds true in the following sense: $\int f \, \mathrm{d}\mu_t \to \int f \, \mathrm{d}\mu$ as $t \to \infty$ for any bounded and continuous function f on $[0, \infty)^d$ that vanishes in a neighborhood of the origin, see Hult and Lindskog (2006) for details. The limit measure μ is a homogeneous Radon measure on E, *i.e.* $\mu(\lambda \cdot) = \lambda^{-1}\mu(\cdot)$, for all $\lambda > 0$, whose margins are standardized in the sense that:

$$\forall y \in (0, \infty), \ \forall j \in \{1, \ldots, d\}, \quad \mu(\{x = (x_1, \ldots, x_d) \in E : x_j \geq y\}) = y^{-1}.$$

Here and throughout, by $\|x\| = \max\{|x_1|, \ldots, |x_1|\}$ is meant the ℓ_∞-norm of any vector $x = (x_1, \ldots, x_d)$ in $\mathbb{R}^d$. Although the concepts introduced below generalize to any norm on $\mathbb{R}^d$ such that the vectors of the canonical basis have all unit norm, the extension of the results of this chapter to any such norm requires significant additional work, as shall be discussed later, and is left for future research. Let $\mathbb{S} = \{x \in [0, \infty)^d : \|x\| = 1\}$ denote the unit sphere restricted to the positive orthant and consider the mapping $\theta : E \to \mathbb{S}$ that assigns to any vector $x \in E$ its angle $\theta(x) = x/\|x\|$. The angular measure Φ is then defined as the push-forward measure of the restriction of μ to $\mathbb{S}$ by θ: for any Borel set $A \subset \mathbb{S}$, we have

$$\Phi(A) = \mu(\mathcal{C}_A) \quad \text{where} \quad \mathcal{C}_A = \{x \in E : \|x\| \geq 1, \ \theta(x) \in A\}. \tag{4.2}$$

The exponent measure μ is fully determined by the angular measure, insofar as we have, for any Borel set $A \subset \mathbb{S}$, $(\mu \circ \Psi^{-1})\{r > u, \theta \in A\} = u^{-1}\Phi(A)$, denoting

by $\Psi(x) = (\|x\|, \theta(x))$ the polar-coordinate transformation on E. In other terms for non-negative Borel measurable functions f on $[0,\infty)^d \setminus \{0\}$, we have

$$\int_{[0,\infty)^d\setminus\{0\}} f(x)\,\mathrm{d}\mu(x) = \int_0^\infty \int_{\mathbb{S}} f(r\theta)\,\mathrm{d}\Phi(\theta)\,\frac{\mathrm{d}r}{r^2}. \tag{4.3}$$

Nonparametric estimation of the angular measure Φ has been investigated from an asymptotic perspective in Einmahl and collab. (2001) and Einmahl and Segers (2009). The authors make use of the empirical versions $\widehat{F}_j$ of the marginal *c.d.f.*'s based on a sample $\{\mathbf{X}_i = (X_{i,1},\ \ldots,\ X_{i,d}) :\ i = 1,\ \ldots,\ n\}$ of $n \geq 1$ independent copies of the r.v. $\mathbf{X}$ in order to form the rank transformed variables $\widehat{\mathbf{V}}_i = (1 - \widehat{F}_1(X_{i,1}),\ \ldots,\ 1 - \widehat{F}_d(X_{i,d}))$, $i \in \{1,\ \ldots,\ n\}$, whose common distribution can be considered as a reasonable substitute for that of $\mathbf{V}$. The empirical angular measure $\widehat{\Phi}$ is then naturally defined as the empirical distribution of a fraction of the angles $\theta(\widehat{\mathbf{V}}_i)$, those corresponding to the k-largest values among the $\|\mathbf{X}_i\|$'s namely. The asymptotic study of its accuracy is limited to the two dimensional case so far. Extension to the multivariate case is far from straightforward because controlling the fluctuations of $\widehat{\Phi}$ on a class of Borelian sets boils down to controlling those of the 'ideal empirical distribution' (the one that would be based on the $\mathbf{V}_i$'s if the latter were observable) on a collection of random sets, due to the rank transform. In order to derive asymptotic results from empirical process theory, these random sets must be framed in between two deterministic framing sets, whose geometry may be complex. This is the main barrier to an extension of these results in dimension larger than two.

It is the goal of this chapter to illustrate a nonasymptotic analysis of the empirical angular measure $\widehat{\Phi}$. Precisely, we first illustrate the relevance of concentration inequalities for the maximal deviations

$$\sup_{A\in\mathcal{A}} |\widehat{\Phi}(A) - \Phi(A)| \tag{4.4}$$

over specific classes $\mathcal{A}$ of Borelian subsets of $\mathbb{S}$. Secondly, we study the influence of the marginal standardization to unit Pareto in a statistical learning application. In addition, we study a truncated version $\widetilde{\Phi}$ of the empirical angular measure estimator, which discards in contrast the very largest observations, for which it is harder to control the angular stochastic error. The concentration properties of the estimator are essentially preserved, in spite of the truncation step. The technical analysis essentially combines the use of framing sets, just like in Einmahl and collab. (2001) and Einmahl and Segers (2009) although they are defined in a different manner here, with concentration inequalities adapted to rare events borrowed from Goix and collab. (2015). The concentration bounds are next used to study at length a statistical learning problem: binary classification in extreme regions. While the standard Empirical Risk Minimization (ERM) approach, the main paradigm of statistical learning theory, tends to ignore the predictive performance of classifier candidates in regions of the input space of lowest density, we focus on the extreme classification risk, *i.e.* the probability of error in extreme regions of the input space, in contrast. By means of the control of the fluctuations of the empirical angular measure, we establish generalization bounds for classifiers obtained by minimizing the empirical classification error based on the most extreme input observations only, *i.e.* empirical extreme risk minimizers.

The chapter is organized as follows. In Section 4.2, the main notations are set out and crucial notions pertaining to multivariate EVT are briefly recalled. The main results related to the nonasymptotic analysis of the empirical angular measure are formulated in Section 4.3, while a truncated version of the angular measure estimator is introduced and studied in Section 4.3.3. Section 4.4 illustrates the significance of our nonasymptotic results on a statistical learning problem related to anomaly detection. Numerical experiments are presented in Section 4.5. Some technical details are deferred to the Section 4.6.

4.2 Background and Preliminaries

As a first go, we set out the main notations used in the subsequent analysis and recall the usual data standardization procedures in MEVT as described in Chapter 2. Equipped with these notions, we next rigorously describe the empirical angular measure and define its truncated variant.

4.2.1 Notations - Data Standardization in Multivariate EVT

Here and throughout, the coordinates of any vector $x \in \mathbb{R}^d$ are denoted by $x_1, \ldots, x_d$. For all $(s,t) \in \mathbb{R}^2$, we use the notation $s \vee t = \max\{s, t\}$ and $s \wedge t = \min\{s, t\}$. The positive part of any real number z is denoted by $z_+ = z \vee 0$. Unless otherwise specified, binary relations and operators (*e.g.* addition, minimum) between two vectors or between a vector and a scalar are understood componentwise. For instance, for any a, b in $\mathbb{R}^d$, $a \preceq b$ means that $a_j \leq b_j$ for $j = 1, \ldots, d$ and $a \wedge 1 = (a_1 \wedge 1, \ldots, a_d \wedge 1)$. For a, b in $\mathbb{R}^d$ such that $a \preceq b$, $[a,b] = \{x \in \mathbb{R}^d, a \preceq x \preceq b\}$ is a rectangle in $\mathbb{R}^d$ and for any non empty subset $A \subset \mathbb{R}^d$ and $t > 0$, we define the rescaled set $tA = \{ta, a \in A\}$. The floor and ceiling functions are respectively denoted by $z \in \mathbb{R} \mapsto \lfloor z \rfloor$ and $z \in \mathbb{R} \mapsto \lceil z \rceil$. By $\mathcal{B}(x, \varepsilon) = \{y \in \mathbb{S};\ \|x - y\| \leq \varepsilon\}$ is meant the intersection between $\mathbb{S}$ and the closed ball with center $x \in \mathbb{R}^d$ and radius $\varepsilon > 0$ related to the ℓ_∞-norm.

On (empirical) standardization of the marginals. The marginal transformation to unit Pareto margins is denoted by $v : \mathbb{R}^d \to [1, \infty]^d$, with

$$v(x) = \left(\frac{1}{1 - F_1(x_1)}, \ldots, \frac{1}{1 - F_d(x_d)} \right) \text{ for all } x \in \mathbb{R}^d, \tag{4.5}$$

with the convention $1/0 = \infty$. Equipped with this notation, we have $\mathbf{V} = v(\mathbf{X})$. Since the distribution P of the heavy-tailed random variable $\mathbf{X}$ under study is unknown, it is replaced in practice by its empirical version based on a sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ of $n \geq 1$ copies of $\mathbf{X}$, namely $P_n(\,\cdot\,) = (1/n)\sum_{i=1}^n \mathbb{1}\{\mathbf{X}_i \in \cdot\}$. In particular, the marginal distributions $F_j(t)$ are substituted with their statistical counterparts

$$\widehat{F}_j(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_{i,j} \leq t\}, \text{ for } t \in \mathbb{R} \text{ and } 1 \leq j \leq d, \tag{4.6}$$

in order to mimic the standardizations aforementioned. Define $\widehat{\mathbf{V}}_i = \hat{v}(\mathbf{X}_i)$, where $\hat{v}(x) = (\hat{v}_1(x_1), \ldots, \hat{v}_d(x_d))$ for all for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and

$$\hat{v}_j(x_j) = \frac{1}{1 - \frac{n}{n+1}\widehat{F}_j(x_j)}, \tag{4.7}$$

for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. Observe incidentally that the factor $n/(n+1)$ in (4.7) serves to avoid division by zero when $x_j \geq \max(X_{1j}, \ldots, X_{nj})$. Then, define

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\widehat{\mathbf{V}}_i},$$

the empirical measure of the pseudo-observations $\widehat{\mathbf{V}}_1, \ldots, \widehat{\mathbf{V}}_n$, which can legitimately be considered as an estimator of the distribution of the $\mathbf{V}_i$'s. Equipped with this notations, the empirical version of the angle $\theta(x) = \|x\|^{-1}x = \|v(x)\|^{-1}v(x)$ of any vector $x \in \mathbb{R}^d$ is $\widehat{\theta}(x) = \|\hat{v}(x)\|^{-1}\hat{v}(x)$.

4.2.2 The Empirical Angular Measure - Problem Statement

The exponent measure μ being defined as a limit in (4.1), an empirical version of it can be obtained at distance $t = n/k$, where $k \in (0, n]$ such that $1 \ll k \ll n$ (in a sense that will be specified later), by computing

$$\widehat{\mu}(B) = \tfrac{n}{k}\widehat{P}_n(\tfrac{n}{k}B) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\left\{\widehat{\mathbf{V}}_i \in (n/k)B\right\} \tag{4.8}$$

for Borel sets $B \subset [0, \infty)^d \setminus \{0\}$. Similarly, since $\Phi(A) = \lim_{t\to\infty} \Phi_t(A)$ for any Borelian subset $A \subset \mathbb{S}$ such that $\mu(\partial \mathcal{C}_A) = 0$, the empirical angular measure is

$$\widehat{\Phi}(A) = \widehat{\mu}(\mathcal{C}_A) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\left\{\widehat{\mathbf{V}}_i \in (n/k)\mathcal{C}_A\right\}. \tag{4.9}$$

As defined in Einmahl and collab. (2001) or Einmahl and Segers (2009), it is simply the empirical version of the finite distance angular measure

$$\Phi_t(A) = t\mathbb{P}\left\{t^{-1}\mathbf{V} \in \mathcal{C}_A\right\} \tag{4.10}$$

at level $t = n/k$. The accuracy of this estimator has been investigated in an asymptotic framework stipulating that $k = k(n)$ tends to infinity as $n \to \infty$ so that $k = o(n)$ in Einmahl and collab. (2001) and Einmahl and Segers (2009).

In Section 4.3.3, a nonasymptotic study of the accuracy of an alternative estimator is proposed, referred to as the *truncated empirical angular measure*, which is closely related to the exponent measure of doubly truncated cones. For $M > 1$ and any measurable subset $A \subset \mathbb{S}$, define the doubly truncated cone

$$\mathcal{C}_A^M = \{tA, \quad 1 \leq t < M\}$$

and set by convention $\mathcal{C}_\varnothing^M = \varnothing$. Notice that $\mathcal{C}_A^M = \mathcal{C}_A \setminus M\mathcal{C}_A$, so that, by homogeneity,

$$\mu\left(\mathcal{C}_A^M\right) = \mu(\mathcal{C}_A) - \mu(M\mathcal{C}_A) = \left(\frac{M-1}{M}\right)\mu(\mathcal{C}_A).$$

On the basis of the equation above, we define the truncated version of $\widehat{\Phi}$

$$\widehat{\Phi}_M(A) = \frac{M}{M-1}\widehat{\mu}\left(\mathcal{C}_A^M\right) = \frac{M}{M-1}\frac{1}{k}\sum_{i=1}^{n}\mathbb{1}\left\{\widehat{\mathbf{V}}_i \in (n/k)\mathcal{C}_A^M\right\}, \tag{4.11}$$

for all $M > 1$. The rationale behind the truncation of the empirical angular measure can be grasped as follows: the k/M most extreme observations among the k largest ones (with respect to each coordinate) are discarded because, as shall be seen in Section 4.3, the variability induced by replacing the $\mathbf{V}_i$'s with the $\widehat{\mathbf{V}}_i$'s for the latter cannot be controlled.

In this chapter, to establish generalization bounds for classification in extreme regions, we rely on probability bounds for the uniform deviations

$$\sup_{A\in\mathcal{A}}|\widehat{\Phi}(A) - \Phi(A)| \text{ and } \sup_{A\in\mathcal{A}}|\widehat{\Phi}_M(A) - \Phi(A)|,$$

where the suprema are taken over a class $\mathcal{A}$ of Borel subsets A of $\mathbb{S}$ under appropriate assumptions. Such bounds are proved in Section 4.3. They are next used to analyze two statistical learning problems involving multivariate extreme data. They are finally supported by numerical experiments, confirming that the replacement of the $\mathbf{V}_i$'s with their empirical counterparts does not damage the statistical recovery of the angular measure, in spite of the strong dependence structure exhibited by the $\widehat{\mathbf{V}}_i$'s.

4.3 Angular Measure Estimation - Concentration Bounds

The main results of Clémençon and collab. (2020) are formulated in this section. We first investigate the concentration properties of the empirical angular measure (4.9) and turn next to those of the alternative estimator (4.11), the analysis proposed highlighting the advantages offered by the double truncation.

4.3.1 Empirical Angular Process - The Framing Approach

In this subsection, we sketch the proof of Clémençon and collab. (2020) to obtain a non asymptotic upper bound for the uniform deviation $\sup_{A\in\mathcal{A}}|\widehat{\Phi}(A) - \Phi(A)|$ of estimated angular probabilities from their true values, the supremum being taken over a class $\mathcal{A}$ of Borelian subsets of the unit sphere $\mathbb{S}$. For the sake of clarity, before we list the technical assumptions required to state it in a rigorous manner. Observe that one may express the empirical angular measure of a set A in $\mathcal{A}$, namely $\Phi(A) = \mu(\mathcal{C}_A) = (n/k)\mu((n/k)\mathcal{C}_A)$, as the raw empirical distribution of the random set

$$\widehat{\Gamma}_A = \hat{v}^{-1}(\tfrac{n}{k}\mathcal{C}_A) = \left\{x \in [0,\infty)^d : \ \|\hat{v}(x)\|_\infty \geq \tfrac{n}{k},\ \theta(\hat{v}(x)) \in A\right\}, \tag{4.12}$$

up to a scaling factor:

$$\widehat{\Phi}(A) = \tfrac{n}{k}\widehat{P}_n(\tfrac{n}{k}\mathcal{C}_A) = \tfrac{n}{k}P_n(\hat{v}^{-1}(\tfrac{n}{k}\mathcal{C}_A)) = \tfrac{n}{k}P_n(\widehat{\Gamma}_A).$$

Using the homogeneity property of the measure μ, one may write for all $A \in \mathcal{A}$,

$$\widehat{\Phi}(A) - \Phi(A) = \tfrac{n}{k} P_n(\widehat{\Gamma}_A) - \mu(\mathcal{C}_A).$$

Following in the footsteps of the approach developed in Einmahl and collab. (2001) and Einmahl and Segers (2009), the rationale behind the uniform control of the deviation above is as follows. For any $A \in \mathcal{A}$, one can construct two nested deterministic sets $\Gamma_A^- \subset \Gamma_A^+$ framing the cone $\mathcal{C}_A$, *i.e.* such that $\Gamma_A^- \subset \mathcal{C}_A \subset \Gamma_A^+$, and such that the event

$$\mathcal{E}_A = \left\{ (n/k)\Gamma_A^- \subset \widehat{\Gamma}_A \subset (n/k)\Gamma_A^+ \right\} \tag{4.13}$$

occurs with high probability. Indeed, on the event (4.13), an upper bound for the signed error can be then obtained as follows:

$$\begin{aligned}\widehat{\Phi}(A) - \Phi(A) &= \tfrac{n}{k} P_n(\widehat{\Gamma}_A) - \mu(\mathcal{C}_A) \leq \tfrac{n}{k} P_n(\tfrac{n}{k}\Gamma_A^+) - \mu(\Gamma_A^-) \\ &\leq \tfrac{n}{k} |P_n(\tfrac{n}{k}\Gamma_A^+) - P(\tfrac{n}{k}\Gamma_A^+)| + |\tfrac{n}{k} P(\tfrac{n}{k}\Gamma_A^+) - \mu(\Gamma_A^+)| + \mu(\Gamma_A^+ \setminus \Gamma_A^-).\end{aligned}$$

A lower bound can be derived in a similar way, yielding on $\mathcal{E}_A$,

$$\begin{aligned}|\widehat{\Phi}(A) - \Phi(A)| \leq & \max_{B \in \{\Gamma_A^+, \Gamma_A^-\}} \tfrac{n}{k} |P(\tfrac{n}{k}B) - \mu(B)| && \text{(bias term)} \\ & + \max_{B \in \{\Gamma_A^+, \Gamma_A^-\}} \tfrac{n}{k} |P_n(\tfrac{n}{k}B) - P(\tfrac{n}{k}B)| && \text{(stochastic error)} \\ & + \mu(\Gamma_A^+ \setminus \Gamma_A^-) && \text{(framing gap).}\end{aligned} \tag{4.14}$$

In the next subsection we introduce technical assumptions, enabling the construction of the framing sets and the control of the probability of occurence of $\cap_{A \in \mathcal{A}} \mathcal{E}_A$ in particular, so as to control each of these three terms uniformly over $A \in \mathcal{A}$. More specifically, under appropriate complexity assumptions for the class $\mathcal{A}$ and that composed of the framing sets, the stochastic error term can be uniformly bounded by means of the concentration inequality for empirical processes over collections of sets of extreme values established in Goix and collab. (2015).

4.3.2 A Bound for the Maximal Deviations of $\widehat{\Phi}$

The main result of Clémençon and collab. (2020) is stated in this subsection. It requires that the following hypotheses are fulfilled.

Assumption 1 (Subsets of the sphere)**.** *The class $\mathcal{A}$ is a collection of non-empty Borel sets of $\mathbb{S}$, containing a dense countable subset, with the following properties:*

(i) There exists $\tau \in (0, 1)$ such that

$$\forall A \in \mathcal{A}, \qquad A \subset \{\theta \in \mathbb{S} : \min(\theta_1, \ldots, \theta_d) > \tau\}. \tag{4.15}$$

(ii) For any $A \in \mathcal{A}$ and $\varepsilon > 0$, there exist Borel subsets $A_+(\varepsilon)$ and $A_-(\varepsilon)$ of $\mathbb{S}$ such that

$$\bigcup_{x \in A_-(\varepsilon)} \mathcal{B}(x, \varepsilon) \subset A \text{ and } \bigcup_{x \in A} \mathcal{B}(x, \varepsilon) \subset A_+(\varepsilon).$$

In addition, there exists $c > 0$ such that

$$\forall A \in \mathcal{A}, \forall \varepsilon > 0, \qquad \Phi\Big(A_+(\varepsilon) \setminus A_-(\varepsilon)\Big) \leq c\,\varepsilon. \tag{4.16}$$

(iii) The class $\mathcal{C} = \{A_\sigma(\varepsilon) : \sigma \in \{-,+\},\ \varepsilon > 0\}$ *is of finite* VC *dimension* $V_\mathcal{C}$.

We point out that the existence of a countable collection $\mathcal{A}_0 \subset \mathcal{A}$ such that for every $A \in \mathcal{A}$ there exists a sequence $A_n \in \mathcal{A}_0$ such that $\lim_{n\to\infty} \mathbb{1}\{x \in A_n\} = \mathbb{1}\{x \in A\}$ for every $x \in \mathbb{R}^d$ (the class of indicators $\{\mathbb{1}\{\cdot \in A\} : A \in \mathcal{A}\}$ is pointwise measurable (VAN DER VAART and WELLNER, 1996, Example 2.3.4)) ensures the measurability of countable unions and suprema taken over $\mathcal{A}$. Condition (i) stipulates that the elements of the class $\mathcal{A}$ are bounded away from the $2^d - 1$ faces of the sphere. Though it may considered as restrictive at first glance, we point out however that maximal deviations of the empirical angular measure over classes of Borel subsets of a given face of the sphere correspond to maximal deviations of the empirical estimator of the angular measure of the corresponding (heavy-tailed) marginal of the original r.v. X. Regarding condition (ii), observe that $A_-(\varepsilon) \subset A \subset A_+(\varepsilon)$ for all $A \in \mathcal{A}$ and $\varepsilon > 0$. As shall be illustrated by the examples given later, many classes of subsets of $\mathbb{S}$ satisfy this assumption, required to build the framing sets mentioned in the previous subsection.

In order to deal with the estimation error stemming from the use of the marginal empirical distribution functions in (4.7), we frame the cones $\mathcal{C}_A$ in between a slightly smaller and larger sets, built from the $A_\sigma(\varepsilon)$'s. For $A \in \mathcal{A}$, $\sigma \in \{-,+\}$ and $r,\ h > 0$, define the sets

$$\Gamma_A^\sigma(r,h) = \left\{x \in [0,\infty)^d :\ \|x\| \geq \tfrac{1}{r},\ \theta(x) \in A_\sigma(h\|x\|)\right\},$$

Observe that, for all $A \in \mathcal{A}$ and $h > 0$, we have $\mathcal{C}_A \subset \Gamma_A^+(r,h)$ as soon as $r \geq 1$ and $\Gamma_A^-(r,h) \subset \mathcal{C}_A$ when $r \leq 1$. The upper confidence bound at level $1-\delta$ for the maximal deviation (4.4) stated in the theorem below is derived from (4.14) with framing sets $\Gamma_A^+ = \Gamma_A^+(r_+,h)$ and $\Gamma_A^- = \Gamma_A^-(r_-,h)$ for specific choices of r_+, r_- and h.

Theorem 2 (CLÉMENÇON and collab. (2020)). *Suppose that Assumption 1 is fulfilled. Then, for any* $\delta \in (0,1)$, *we have with probability at least* $1-\delta$*:*

$$\sup_{A\in\mathcal{A}} |\widehat{\Phi}(A) - \Phi(A)| \leq \sup_{A\in\mathcal{A},\ \sigma\in\{+,-\}} \left|\tfrac{n}{k} P(\tfrac{n}{k}\Gamma_A^\sigma) - \mu(\Gamma_A^\sigma)\right| + C\left(\sqrt{\frac{d(1+\Delta)V_\mathcal{C}\log((d+1)/\delta)}{k}} + \frac{\log\left((d+1)/\delta\right)}{k}\right) + (2d + 3c(\log(d/(3c)) - \log(\Delta) + 1))\,\Delta, \quad (4.17)$$

where $\Delta = C\sqrt{\log((d+1)/\delta)/(\rho k)} + C\log((d+1)/\delta)/k$ *and* C *is a universal constant, as soon as* $n > (3 \vee 6c)/\tau$, $3 \vee 6c < k < \tau n$ *and* $k/n < \rho < \tau$.

The technical proof is given in Section 4.6. It involves the framing sets defined by $h = 3\Delta$, $r_- = 1 + 1/n - 1/k - \Delta$ and $r_+ = 1 + 1/n - 1/k + \Delta$. Observe that k is chosen sufficiently large so that $(2/k) \leq \Delta < (1 - 1/k) \wedge (1/(3c))$ and $\rho/(1-\Delta\rho) \leq \tau$. In particular, we have $r_- \leq 1 \leq r_+$. One may then show that the framing gap is bounded by $(2d + 3c(\log(d/(3c)) - \log(\Delta) + 1))\,\Delta$ for any $A \in \mathcal{A}$, while the estimation error is uniformly bounded by

$$C\left(\sqrt{\frac{d(1+\Delta)V_\mathcal{C}\log((d+1)/\delta)}{k}} + \frac{\log\left((d+1)/\delta\right)}{k}\right),$$

with probability larger than $1-\delta$. The constant C is that appearing in the concentration inequality for empirical processes over collections of sets of extreme values proved in Goix and collab. (2015), which is applied here to the collection $\mathcal{F} = \{(n/k)\Gamma_A^\sigma : \ \sigma \in \{+,-\}, \ A \in \mathcal{A}\}$, of VC dimension less than $V_\mathcal{C}$ and used to control the stochastic error term. For fixed ρ and δ, when ignoring the bias term, the bound on the right hand side of (4.17) is of order

$$3c\Delta \log(1/\Delta) + \mathrm{o}\left(\frac{\log(k)}{\sqrt{k}}\right) = \frac{3}{2} cC \sqrt{\frac{\log(1/\delta)}{\rho}} \frac{\log(k)}{\sqrt{k}} + \mathrm{o}\left(\frac{\log(k)}{\sqrt{k}}\right), \qquad (4.18)$$

as $k \to \infty$. Interestingly, the angular part of the gap term dominates the error term.

4.3.3 Concentration Bounds for the Truncated Estimator of Φ

We now extend the concentration result established in the previous subsection to the truncated version (4.11) of the empirical angular measure, which discards the most extreme observations (a fraction of the largest $\|\widehat{V}_i\|$'s) because of the variability they induce in the estimation procedure. The approach is the same as that used for (4.9) except that, this time, we need to build framing sets for the doubly truncated cones $\mathcal{C}_A^M$, $M > 1$. For all $A \in \mathcal{A}$, $\sigma \in \{-,+\}$ and $h > 0$ and $0 \leq s < r$, consider the sets

$$\Gamma_A^\sigma(r, s; h) = \left\{ x \in [0,\infty)^d : \ \tfrac{1}{r} \leq \|x\| < \tfrac{1}{s}, \ \theta(x) \in A_\sigma(h\|x\|) \right\},$$

where $1/s$ is to be read as ∞ if $s = 0$. In order to derive an upper confidence bound for the maximal deviations of the estimator (4.11) at level $1-\delta$ from a 'bias-variance-gap' decomposition similar to (4.14), we use the framing $\Gamma_A^{M,-} \subset \mathcal{C}_A^M \subset \Gamma_A^{M,+}$, where the framing sets are defined as

$$\begin{aligned} \Gamma_A^{M,-} &= \Gamma_A^-\Big(1 + \tfrac{1}{n} - \tfrac{1}{k} - \Delta, (1+\tfrac{1}{n})\tfrac{1}{M} - \tfrac{1}{k} + \Delta; 3\Delta\Big), \\ \Gamma_A^{M,+} &= \Gamma_A^+\Big(1 + \tfrac{1}{n} - \tfrac{1}{k} + \Delta, ((1+\tfrac{1}{n})\tfrac{1}{M} - \tfrac{1}{k} - \Delta)_+; 3\Delta\Big), \end{aligned} \qquad (4.19)$$

choosing k and Δ as in the previous subsection.

A slight modification of the proof of Theorem 2 then permits to establish the following result.

Theorem 3 (Clémençon and collab. (2020)). *Let $M > 1$ and suppose that Assumption 1 is fulfilled. Then, for all $\delta \in (0,1)$, we have with probability larger than $1-\delta$:*

$$\begin{aligned} \sup_{A \in \mathcal{A}} |\widehat{\Phi}_M(A) - \Phi(A)| \leq{}& \tfrac{M}{M-1} \sup_{A \in \mathcal{A},\, \sigma \in \{-,+\}} |\tfrac{n}{k} P(\tfrac{n}{k}\Gamma_A^{M,\sigma}) - \mu(\Gamma_A^{M,\sigma})| \\ &+ \tfrac{M}{M-1} C \left(\sqrt{\frac{d(1+\Delta)V_\mathcal{C} \log((d+1)/\delta)}{k}} + \frac{\log((d+1)/\delta)}{k} \right) \\ &+ \tfrac{M}{M-1} \Big(4d\Delta + 3c\Delta \log\Big(M \wedge d/(3c\Delta)\Big) + \Big(3c\Delta - d/M\Big)_+ \Big), \end{aligned} \qquad (4.20)$$

where Δ, k, n and ρ are chosen as in Theorem 2.

See Section 4.6 for the technical proof. For fixed $M > 1$, ρ and δ, ignoring the bias, the bound on the right hand side of (4.20) is of order

$$\frac{M}{M-1}\left(C\sqrt{\frac{dV_{\mathcal{C}}\log(1/\delta)}{k}} + (4d + 3c\log(M))\Delta\right) + \mathrm{o}\left(\frac{1}{\sqrt{k}}\right)$$
$$= \frac{M}{M-1}C\sqrt{\frac{\log(1/\delta)}{k}}\left(\sqrt{dV_{\mathcal{C}}} + (4d + 3c\log(M))\sqrt{\rho^{-1}}\right) + \mathrm{o}\left(\frac{1}{\sqrt{k}}\right),$$

as $k \to \infty$. This is to be compared with (4.18): the factor $\log(k)$ has disappeared, in return for a factor $\log(M)$. Setting $M = \sqrt{k}$ formally yields (4.18) again.

4.4 Statistical Learning Application: Anomaly Detection through Minimum Volume Set Estimation

It is the purpose of this section to illustrate how the concentration results previously stated are proved useful to establish sound nonasymptotic guarantees for the validity of certain statistical learning procedure relying on the empirical angular measure and recently introduced in the literature. The problem studied in this chapter is of a learning procedure, beyond classic empirical risk minimization, the Minimum Volume set estimation methodology in order to perform (unsupervised) anomaly detection among extreme observations, as described in Thomas and collab. (2017). The concept of *minimum volume set* (MV-set in abbreviated form) has been introduced in the seminal contribution Einmahl and Mason (1992) (see also Polonik (1997)) in order to extend the notion of quantile of a univariate probability distribution to the multivariate framework and describe regions where a random vector $\mathbf{X}$ valued in a measurable space $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$ takes its values with highest/smallest probability. Let ν be a σ-finite measure of reference on $\mathcal{X}$ equipped with its Borel σ-algebra and fix $\alpha \in (0, 1)$. A MV-set Ω_α^* of mass at least α is any solution of the constrained minimization problem $\min_\Omega \nu(\Omega)$ subject to $\mathbb{P}\{\mathbf{X} \in \Omega\} \geq \alpha$, where the minimum is taken over all measurable subsets Ω of $\mathcal{X}$. Unsupervised novelty/anomaly detection methods can be based on the statistical recovery of MV-sets: for a large value of α, abnormal observations are those which belong to the complementary set $\mathcal{X} \setminus \Omega_\alpha^*$. In the case where the distribution $F(dx)$ of the r.v. X is absolutely continuous w.r.t. ν with density $f(x) = (dF/d\nu)(x)$ and under the assumption that the r.v. $f(\mathbf{X})$ is bounded and has a continuous distribution $F_f(dx)$, it can be shown that the superlevel set $\{x \in \mathcal{X} : \ f(x) \geq F_f^{-1}(1-\alpha)\}$ is the unique solution of the constrained optimization problem above, denoting by $K^{-1}(u) = \inf\{t \in \mathbb{R} : \ K(t) \geq u\}$ the generalized inverse of any cumulative distribution function K on $\mathbb{R}$. From a statistical perspective, estimates $\widehat{\Omega}_\alpha^*$ of minimum volume sets are built from a training dataset $\mathbf{X}_1, \ \ldots, \ \mathbf{X}_n$ composed of $n \geq 1$ independent copies of the generic r.v. $\mathbf{X}$ by replacing in the constrained optimization problem the unknown probability measure P with an empirical version, the raw empirical distribution $P_n = (1/n)\sum_{i=1}^n \delta_{\mathbf{X}_i}$ typically, and restricting optimization to a collection $\mathcal{A}$ of

borelian subsets of $\mathcal{X}$, supposed rich enough to include all density superlevel sets (or reasonable approximants of the latter). In Polonik (1997), functional limit results are derived for the generalized empirical quantile process $\{\nu(\widehat{\Omega}^*_\alpha) - \lambda^*(\alpha)\}$ under certain assumptions for the class $\mathcal{A}$ (stipulating in particular that $\mathcal{A}$ is a Glivenko-Cantelli class for $F(dx)$). In Scott and Nowak (2006), it is proposed to replace the level α by $\alpha - \psi$ where ψ plays the role of tolerance parameter and is picked so as to be of the same order as the supremum $\sup_{\Omega\in\mathcal{A}} |P_n(\Omega) - P(\Omega)|$ roughly, complexity of the class $\mathcal{A}$ being controlled by the VC dimension or by means of the concept of Rademacher averages, in order to establish rate bounds at $n < +\infty$ fixed. Alternatively, so-termed *plug-in* techniques, consisting in computing first an estimate $\widehat{f}$ of the density f and considering next superlevel sets $\{x \in \mathcal{X} :\ \widehat{f} \geq t\}$ of the resulting estimator have been investigated in several papers, among which Tsybakov (1997) or Rigollet and Vert (2009) for instance. Such an approach however yields significant computational issues even for moderate values of the dimension, inherent to the curse of dimensionality phenomenon.

Minimum volume set estimation on the sphere. Relying on the notion of MV-set, an approach to (unsupervised) anomaly detection in extreme regions has been introduced in Thomas and collab. (2017). It is implemented as follows. We place ourselves in the framework described in section 4.2 and consider a regularly varying r.v. $\mathbf{X}$ with angular measure Φ. supposed to be absolutely continuous w.r.t. the $(d-1)$-dimensional Hausdorff measure λ on $\mathbb{S}$ with density ϕ. As Φ describes the least/most probable directions $\theta(\mathbf{X})$ of extremes $\mathbf{X} > t$, a legitimate technique to detect abnormal data among extremes consists in trying to recover MV-sets of Φ with large mass $\alpha \in (0,\ \Phi(\mathbb{S}))$ and reference measure λ, *i.e.* in solving the problem

$$\min_{\Omega\subset\mathbb{S},\ \text{Borelian}} \lambda(\Omega) \text{ subject to } \Phi(\Omega) \geq \alpha, \tag{4.21}$$

and pin extreme observations $\mathbf{X}$ with angle $\theta(\mathbf{X})$ in the complement of angular MV-sets as anomalies. Under the additional hypothesis that $\Phi(\theta(\mathbf{X}))$ is essentially bounded, the problem (4.21) has a unique solution $\Omega^*_\alpha = \{\theta \in \mathbb{S} :\ \phi(\theta) \geq F_\Phi^{-1}(\Phi(\mathbb{S}) - \alpha)\}$, where $F_\Phi(t) = \Phi(\{\theta \in \mathbb{S} :\ \phi(\theta) \leq t\})$. Of course, minimization is restricted to a class $\mathcal{A}$ of Borel subsets of $\mathbb{S}$ in practice, the unknown angular measure Φ is replaced with the empirical estimator (4.9) and the level α by $\alpha - \psi$, where ψ is a certain tolerance parameter. As proposed in Scott and Nowak (2006), a natural approach, referred to as MV-ERM therein, consists in taking as tolerance parameter an upper confidence bound at level $1 - \delta \in (0, 1)$ for $\sup_{A\in\mathcal{A}} |\Phi(A) - \widehat{\Phi}(A)|$. Building on the concentration bound stated in Theorem 2, the result below provides performance guarantees for empirical MV-sets on $\mathbb{S}$ at confidence level $1 - \delta$, when the tolerance parameter is greater than the bound

on the right hand side of (4.17), namely

$$\psi(\delta) \geq \sup_{A\in\mathcal{A},\, \sigma\in\{+,-\}} \left|\tfrac{n}{k} P(\tfrac{n}{k}\Gamma_A^\sigma) - \mu(\Gamma_A^\sigma)\right| + C\left(\sqrt{\frac{d(1+\Delta)V_{\mathcal{C}}\log((d+1)/\delta)}{k}} + \frac{\log\left((d+1)/\delta\right)}{k}\right) + (2d + 3c(\log(d/(3c)) - \log(\Delta) + 1))\,\Delta. \quad (4.22)$$

Theorem 4. *Suppose that the class $\mathcal{A}$ satisfies Assumption 1 . Let $\delta \in (0,1)$ and $\alpha > 0$ a fixed mass level. Consider a tolerance parameter $\psi(\delta)$ such that (4.22) holds true. Then, for any solution $\hat{A}$ of the empirical angular* MV*-set problem*

$$\min\left\{\lambda(A) :\ A \in \mathcal{A},\, \widehat{\Phi}(A) \geq \alpha - \psi(\delta)\right\}, \quad (4.23)$$

we simultaneously have, with probability at least $1-\delta$,

$$\lambda(\hat{A}) \leq \min\{\lambda(A) :\ A \in \mathcal{A},\, \Phi(A) \geq \alpha\} \text{ and } \Phi(\hat{A}) \geq \alpha - 2\psi(\delta).$$

The proof is straightforward and deferred to Section 4.6. Of course, a similar result can be obtained for empirical MV-sets built from the truncated empirical measure (4.11) with $M > 1$, namely solutions of

$$\min\left\{\lambda(A) :\ A \in \mathcal{A},\, \widehat{\Phi}_M(A) \geq \alpha - \psi(\delta)\right\}, \quad (4.24)$$

by taking the tolerance level $\psi(\delta)$ larger than the bound on the right hand side of (4.20). In addition, the angular measure Φ can be replaced by the penultimate or sub-asymptotic version Φ_t at $t = n/k$. The bias term in the error bound then disappears.

4.5 Illustrative Numerical Experiments

Our experiments aim at (i) investigating the influence of the rank-transformation upon the empirical angular measure $\widehat{\Phi}$, (ii) comparing the latter with its truncated version $\widehat{\Phi}_M$.

Regarding goals $(i), (ii)$ we compare the performance of the two versions $\widehat{\Phi}$ and $\widehat{\Phi}_M$ of the empirical angular measure together with a pseudo baseline $\tilde{\Phi}$ obtained with probability integral transformed data $V_i = v(X_i)$, $\tilde{\Phi}(A) = \tilde{\mu}(\mathcal{C}_A)$ with $\tilde{\mu}(B) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{V_i \in n/kB\}$, *i.e.*

$$\tilde{\Phi}(A) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{\theta(V_i) \in A, \|V_i\|_\infty \geq n/k\}. \quad (4.25)$$

Notice that $\tilde{\Phi}(A)$ is only observable in general when the marginal distributions F_j are known. In our first experiment (Section 4.5.1), using simulated data, we consider as a performance measure the supremum error $\sup_{A\in\mathcal{A}} |\Phi'(A) - \Phi(A)|$ with $\Phi' \in \{\widehat{\Phi}, \widehat{\Phi}_M, \tilde{\Phi}\}$ and $\mathcal{A}$ is a class of sets composed of hyper-rectangles on the unit cube. In our second experiment (Section 5.4.2) using simulated data again, we compare the estimators $\widehat{\Phi}$ and $\widehat{\Phi}_M$ in a classification context. More precisely, we compare the classification score of three classifiers obtained by minimization of the empirical risks $\widehat{L}^\tau$ and $\widehat{L}^{\tau,M}$

4.5.1 Supremum error of the empirical measures

The goal of this experiment is to assess the influence of the rank transformation and the truncation on the error $\sup_{A\in\mathcal{A}}|\Phi'(A)-\Phi(A)|$, for $\Phi' \in \{\widehat{\Phi}, \widehat{\Phi}^M, \widetilde{\Phi}\}$ in a setting where $\Phi(A)$ can be approached with arbitrary precision by Monte-Carlo sampling, the transformation v to Pareto margins based on the marginal distributions F_j can be computed, and the bias term in the upper bounds of Theorems 2 and 3 is zero by construction, so that the error only stems from the terms named stochastic error and gap (*resp.* $\text{stochastic error}_{\text{M}}$ and gap_{M}) in the two statements.

Experimental setting.

We consider $i.i.d.$ copies $X_i, i \geq 1$ of a random vector $X = R\Theta$ where R and Θ are independent, R follows a unit-Pareto distribution and Θ follows a symmetric Dirichlet distribution on the unit simplex $\mathbb{S}_{L^1}$ with parameter $(\nu, \ldots, \nu)$, for some concentration parameter $\nu > 0$. The following lemma, which proof is deferred to Section 4.6, shows that this setting matches our purposes, even though the marginal distributions F_j are only known on $[1, \infty)$.

Lemma 1. *Let R be a unit-Pareto distributed random variable and let Θ be a centered random variable valued on $\mathbb{S}_{L^1}$ independent from R satisfying $\mathbb{E}(\Theta_j) = 1/d$, $j \leq d$. Let $X = R\Theta$. Then*

1. *The restriction to $[1,\infty)^d$ of the transformation v to unit Pareto margin writes*

$$\forall x \in [1,\infty)^d, \quad v(x) = dx.$$

Conversely for any $x \in [0,\infty)^d \setminus \{0\}$,

$$v(x) \in (d,\infty)^d \Rightarrow x \in (1,\infty) \text{ and } v(x) = dx.$$

2. *For any angular set $A \subset \mathbb{S}$ such that the closure of A is included in $(0,\infty)^d$ (*i.e.* A is bounded away from the boundaries of the positive orthant) and such that $\Phi(\partial A) = 0$, the angular measure defined by (4.2) is related to the distribution of X as follows:*

$$\Phi(A) = d\mathbb{P}\{X \in \mathcal{C}_A\}$$

3. *For all $t > d/\tau$ and $A \subset \mathbb{S}$ satisfying (4.15) such that $\Phi(\partial A) = 0$,*

$$\Phi(A) = t\mathbb{P}\{V \in t\mathcal{C}_A\}$$

An immediate consequence of Lemma 1 is that for $X = R\Theta$ as described above, and for $A \subset \mathbb{S}$ satisfying (4.15), the following facts hold true:

1. For n, k such that $n/k > d/\tau$, given an independent sample $\{(R_i, \Theta_i), i \leq n\}$ distributed as (R, Θ) and letting $X_i = R_i\Theta_i$, the pseudo-empirical estimator $\widetilde{\Phi}(A)$ based on the X_i's and defined in (4.25) writes

$$\tilde{\Phi}(A) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\{\Theta_i/\|\Theta_i\|_\infty \in A\,,\, dR_i\|\Theta_i\|_\infty \geq n/k\} \tag{4.26}$$

2. $\Phi(A)$ may be approached with arbitrary precision by the Monte-Carlo estimator

$$\begin{aligned}\Phi_{\mathrm{MC}}(A) &= \frac{d}{N}\sum_{i=1}^{N}\mathbb{1}\{X_i \in \mathcal{C}_A\} \\ &= \frac{d}{N}\sum_{i=1}^{N}\mathbb{1}\{\Theta_i/\|\Theta_i\|_\infty \in A\,,\, R_i\|\Theta_i\|_\infty \geq 1\}\end{aligned} \tag{4.27}$$

which variance is less that $d^2/(4N)$, where $X_i = R_i\Theta_i$ and $\{(R_i, \Theta_i), i \leq N\}$ is an independent sample distributed as (R, Θ).

3. For n, k such that $n/k > d/\tau$, we have

$$\Phi(A) = \frac{n}{k}\mathbb{P}\left\{V \in \frac{n}{k}\mathcal{C}_A\right\}, \tag{4.28}$$

so that the bias term in theorems 2 and 3 is null.

In this experiment we consider the cases $d = 2$ and $d = 5$ and we choose the class $\mathcal{A}$ as the finite class of hyper-rectangles on the sphere $\mathbb{S}$ forming a regular grid on $\mathbb{S}_\tau$, with side length $h = (1-\tau)/S$ with $S = 5$ and $\tau = 0.1$, *i.e.* each set $A \in \mathcal{A}$ is of the kind $A = \{x \in \mathbb{R}^d : x_j \in A_j, \forall 1 \leq j \leq d\}$, where $A_{j_0} = \{1\}$ for some $j_0 \in \{1, \ldots, d\}$, and for $j \neq j_0$, $A_j = (\tau + i_j h, \tau + (i_j + 1)h)$ for some $i_j \in \{0, \ldots, S-1\}$. The concentration parameter of the Dirichlet distribution for Θ is set to $\nu = 10$.

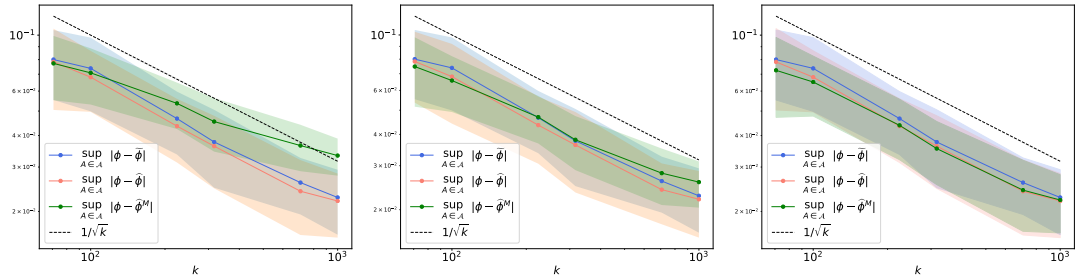
We set the Monte-Carlo sample size to $N = 5 \cdot 10^8$ so that $\Phi_{\mathrm{MC}}(A)$ has standard deviation less than $d/\sqrt{4N} \leq d \times 2.24 \cdot 10^{-5}$. As shall be seen below, the latter is negligible compared to the supremum deviations of $\widetilde{\Phi}, \widehat{\Phi}$ and $\widehat{\Phi}_M$ from Φ_{MC}. Thus for ease of notation we identify Φ and Φ_{MC}.

Results.

We consider varying sample sizes $n \in \{5\cdot 10^3, 10^4, 5\cdot 10^4, 10^5, 5\cdot 10^5, 10^6\}$. For each sample size n, the integer k is set to $\sqrt{n}$. Figure 4.1 shows the mean logarithmic supremum errors $\log(\sup_{A\in\mathcal{A}}|\Phi' - \Phi|)$ as a function of $\log(k)$, for $\Phi' \in \{\widehat{\Phi}, \widehat{\Phi}^M, \widetilde{\Phi}\}$, averaged over 50 independent experiments in dimension $d = 2$. Three values of M are considered. They are chosen so that respectively 10%, 5% and 2% of the set $\{\widehat{V}_i : \|\widehat{V}_i\| \geq n/k, i \leq n\}$ are discarded.

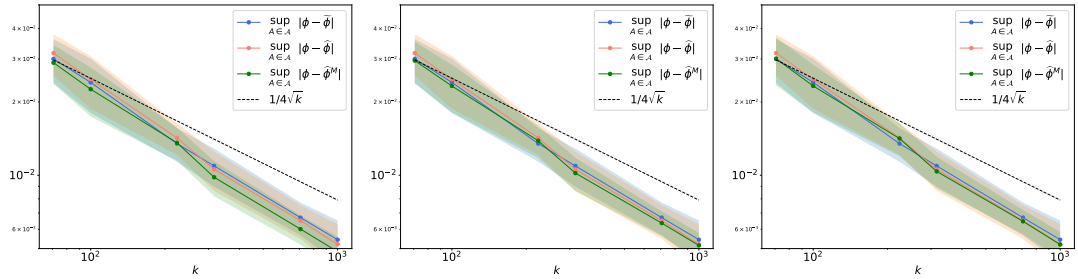
The results gathered in Figure 4.1 confirm that the error of all three estimators, viewed as a function of k, decreases at rate $1/\sqrt{k}$, which indicates that our upper bounds, viewed as functions of k, may be sharp (up to multiplicative constants). Perhaps surprisingly, the pseudo-empirical estimator $\widetilde{\Phi}$ (using knowledge of margins) is – to a small extent – outperformed by the classical version $\widehat{\Phi}$ using the rank transformation, even though the gap between the two is moderate. This empirical finding is consistent with the empirical copula literature, *e.g.* Genest and Segers (2010) show that for positively associated pairs (which is the case here), the variance of the empirical copula based on rank-transformed data is less than that of the empirical copula using knowledge of margins, *i.e.* using the true marginal distributions F_j instead of the empirical ones.

The only situation where the truncated estimator $\widehat{\Phi}_M$ has significately worse performance than the two other is that of a low dimension, and large sample size (Figure 4.2a). For higher dimensions and smaller sample sizes the three estimators have comparable performance.



(a) (b) (c)

Figure 4.1 – Errors $\sup_{A\in\mathcal{A}} |\Phi'(A) - \Phi(A)|$, $\Phi' \in \{\widehat{\Phi}, \widehat{\Phi}_M, \widetilde{\Phi}\}$ in dimension $d = 2$, as a function of k with $k = \sqrt{n}$ on log scales. Only the error curve for $\widehat{\Phi}_M$ varies between Figures 4.1a, 4.1b and 4.1c where M is respectively set so that $10\%, 5\%$ and 2% of extremes are discarded. Colored intervals represent standard deviations of the errors.



(a) (b) (c)

Figure 4.2 – Errors $\sup_{A\in\mathcal{A}} |\Phi'(A) - \Phi(A)|$, $\Phi' \in \{\widehat{\Phi}, \widehat{\Phi}_M, \widetilde{\Phi}\}$ in dimension $d = 5$, as a function of k with $k = \sqrt{n}$ on log scales. Only the error curve for $\widehat{\Phi}_M$ varies between Figures 4.2a, 4.2b and 4.2c where M is respectively set so that $10\%, 5\%$ and 2% of extremes are discarded. Colored intervals represent standard deviations of the errors. Note that the dotted line has equation $y = 1/(4\sqrt{k})$, the factor 4 allowing for a better visualization of the other curves.

4.5.2 Anomaly detection through Minimum Volume Set Estimation

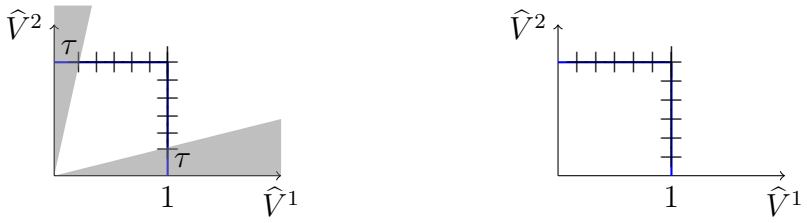


Figure 4.3 – (Left) Illustration of a paving of $\mathbb{S}_\tau$ with 5 rectangles with size $(1-\tau)/5$. Gray cones are unobserved regions (in accordance with constraint (4.15)). (Right) Illustration of a paving of $\mathbb{S}_\tau$ with 6 rectangles with size $1/6$.

Anomaly detection is a natural application of MV-sets estimation. Indeed, given a minimum volume set of mass $0 < \alpha < 1$ for a density f on a sample space $\mathcal{X}$, one can declare as an anomaly any new candidate point $x_0 \in \mathcal{X}$ not belonging to

that set. In the case where $\mathcal{X}$ is compact, doing so is equivalent to performing a Neyman-Pearson test of $H_0 : X \sim f$ against the alternative that X is uniformly distributed on $\mathcal{X}$. In anomaly detection, it is common to use 'scoring functions' $s(x)$, reflecting the degree of normality of a candidate point x. Such algorithms construct an empirical scoring function $\hat{s}$ using the training set and declare as abnormal any new point x_0 such that $\hat{s}(x_0) < s_0$ for some threshold s_0 chosen so as to satisfy type-I error constraints. Any scoring function s^* which is a non-decreasing transform of the probability density function f of the normal instances is optimal in the sense of Neyman Pearson for the test described above, because the associated normality regions are minimum-volume sets for the density of the normal class. It is thus natural to construct a scoring function $\hat{s}$ so that the level sets of $\hat{s}$ are empirical MV-sets. In a multivariate extreme values setting, Thomas and collab. (2017) introduce a scoring function $\hat{s}$ designed for anomaly detection in extreme regions and constructed *via* MV-sets estimation. They first define the angular component $\hat{s}_\theta$ of $\hat{s}$ on $\mathbb{S}$ according to the general principle described above, so that the smaller $\hat{s}_\theta(\theta)$, the more abnormal the direction θ in extreme regions. In practice $\hat{s}_\theta$ is chosen as being piecewise constant on a partition $(B_i, i \leq M)$ of the unit sphere such that all B_i's have the same Lebesgue measure (as illustrated in Figure 4.3 (Right)), namely for $\theta \in B_i$, $\hat{s}_\theta(x)$ is proportional to he number of training angular samples belonging to B_i. The scoring function on the whole extreme region $\{x : \|x\| > t\}$ for some large t is then defined as

$$\hat{s}(x) = (1/\|\widehat{v}(x)\|^2) \cdot \hat{s}_\theta(\hat{\theta}(x)).$$

The scoring function $\hat{s}$ is not purely angular since only considering angular MV-sets does not yield an optimal decision function as the density of the largest observations includes a radial part.

Experimental Setting for Anomaly Detection through MV-Set Estimation.

The goal of this experiment is to assess the influence of the parameter τ on anomaly detection *via* MV-Set estimation. We compare the performance of $\hat{s}$ trained over all the extreme samples from the train dataset $\mathcal{T}_{\text{train}}$ (see Figure 4.3 (Right)) and $\hat{s}_\tau$ which training set is restricted to extreme instances which angle belongs to $\mathbb{S}_\tau$ (see Figure 4.3 (Left)) In order to keep the two methods comparable, anomalies are set so that they appear $\mathbb{S}_\tau$. The performance of $\hat{s}$ and $\hat{s}_\tau$ is measured in terms of Area Under the ROC curve (ROC AUC) since the latter is equal to the probability that a given scoring function will rank a randomly chosen anomaly sample higher than a randomly chosen normal sample Our experimental set-up generalizes the setting described in Section 4.5.1 to the anomaly detection framework. Consider a random variable $X = R\Theta$ where R follows a standard Pareto distribution and Θ follows a Dirichlet distribution on the simplex $\mathbb{S}_{L^1}$ with concentration parameter ν. R and Θ are independent. Anomalies only appear in the test set $\mathcal{T}_{\text{test}}$, the proportion of anomalies among the test data is defined as p_{anomaly}, the remaining test data are samples generated as X. The anomalies are uniformly distributed over the region $\{x : n/k \leq \|\hat{v}(x)\| \leq \max_{x \in \mathcal{T}_{\text{train}}}(\|\hat{v}(x)\|)\}$. The test set $\mathcal{T}^\tau_{\text{test}}$ is restricted to test samples (both anomalies and non anomalies) with angular component lying in $\mathbb{S}_\tau$ so that the performance of $\hat{s}_\tau$ can be compared to the one of $\hat{s}$.

Results.

The dimension d of the problem is set to 5. Before discarding the samples which do not belong to $\mathbb{S}_\tau$, the train and test sets comprise 10^5 samples each. The proportion p_{anomaly} is set to 1%. We choose the class $\mathcal{A}$ as a the finite class of hyper-rectangles on the sphere $\mathbb{S}$ forming a regular grid on $\mathbb{S}_\tau$, with side length $h = (1-\tau)/S$ with $S = 7$. The concentration parameter of the Dirichlet distribution for Θ is set to $\nu = 3$, τ ranges between 0 and 0.1. Figure 4.4 gathers boxplots of ROC AUC's obtained over 30 independently simulated datasets. The dashed vertical line for $\tau = 0.08$ shows the value of τ beyond which the performance of $\hat{s}$ and $\hat{s}_\tau$ are significantly different based on a Kolmogorov-Smirnov test. This figure illustrates that as long as τ remains small but non zero, no significant difference is observed between $\hat{s}_\tau$ and $\hat{s}$: the loss of information induced by τ is not significant while the theoretical guarantees are valid. Figure 4.5 illustrates the evolution of the average number of train samples discarded while increasing τ. This second figure bears out that, in our setting, for small values of τ the number of discarded samples is negligible.

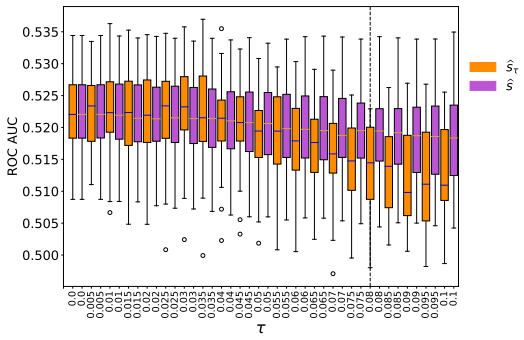


Figure 4.4 – Evolution of ROC AUC on $\mathcal{T}^\tau_{\text{test}}$ for scoring functions $\hat{s}_\tau$ and $\hat{s}$ with varying values of τ.

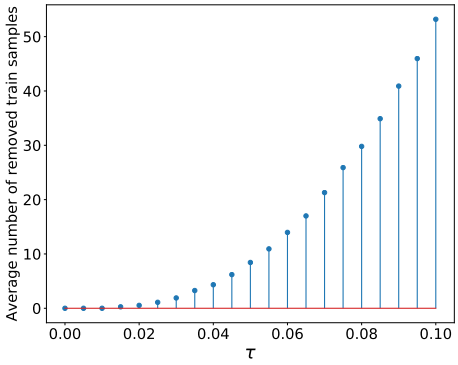


Figure 4.5 – Evolution of the average number of removed train samples with τ.

4.6 Proofs

Auxiliary Results

We start with stating auxiliary results that are used in the proof of the main theorems.

Theorem 5 (Theorem 1 in Goix and collab. (2015)). *Let P_n denote the empirical distribution of an independent random sample $\xi_1, \ldots, \xi_n$ from a distribution P on some measurable space. Let $\mathcal{A}$ be a VC-class of measurable subsets of the state space with VC-dimension $V_\mathcal{A}$. Let B be a measurable subset containing $\bigcup_{A\in\mathcal{A}} A$ and write $p = P(B)$. There is an absolute constant $C > 0$ such that, for all $\delta \in (0,1)$, on an event with probability at least $1-\delta$, we have*

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq C\left(\sqrt{pV_\mathcal{A} n^{-1}\log(1/\delta)} + n^{-1}\log(1/\delta)\right).$$

In Goix and collab. (2015), the constant C is not made explicit. Just before their Lemma 14, there is a reference to Koltchinskii (2006) providing bounds on the expectation of a symmetrized supremum, but from that source, the constant looks near impossible to trace. An alternative route to obtain a value for the constant is via Theorems 1.16–17 in Lugosi (2002), giving an explicit bound for the expectation of a symmetrized supremum in terms of an integral over covering numbers of the class of sets. In turn, the covering numbers of such a class can be bounded explicitly in terms of its VC dimension. In this way, the constant C can be made explicit but its value is most likely going to be far from optimal, as the cited bounds are not sharp, in particularly not the one in Theorem 1.16 in Lugosi (2002). We prefer the current write-up in terms of a generic C. Should a sharp value for C be found in the future, it can be substituted as is in our results.

Lemma 2. *Let $\|\cdot\|$ be a norm on a real vector space and write $\theta(z) = z/\|z\|$ for non-zero z. For non-zero vectors x and y, we have*

$$\|\theta(x) - \theta(y)\| \leq 2\frac{\|x - y\|}{\|x\| \vee \|y\|}.$$

Proof. Since $\theta(\,\cdot\,)$ is scale-invariant, we can divide both x and y by $\|x\| \vee \|y\|$ without changing the two sides of the inequality. For the sake of the proof we can thus assume that $\|x\| = 1 \geq \|y\| > 0$, in which case we need to show that

$$\|x - \theta(y)\| \leq 2\|x - y\|.$$

By the triangle inequality, we have

$$\|x - \theta(y)\| \leq \|x - y\| + \|y - \theta(y)\|$$

and

$$\|y - \theta(y)\| = \left|1 - \tfrac{1}{\|y\|}\right| \|y\| = |\|y\| - 1| = |\|y\| - \|x\|| \leq \|y - x\|. \qquad \square$$

Proof of Theorem 2

As a first go, observe that we can assume that the heavy-tailed r.v. $\mathbf{X}$ has unit-Pareto margins without any loss of generality. Indeed, the empirical exponent and angular measures $\widehat{\mu}$ and $\widehat{\Phi}$ only depend on the sample $X_1, \ldots, X_n$ through the transformed data

$$\widehat{F}_j(X_{ij}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{X_{tj} \leq X_{ij}\}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. The sum of indicators is equal to the rank of X_{ij} within $X_{1j}, \ldots, X_{nj}$. As F_j is continuous, the ranks of $X_{1j}, \ldots, X_{nj}$ are with probability one equal to those of $V_{1j}, \ldots, V_{nj}$, where $V_{ij} = 1/(1 - F_j(X_{ij}))$. Hence, even though the margins $F_1, \ldots, F_d$ are unknown, the fact that $\widehat{\mu}$ and $\widehat{\Phi}$ are rank statistics implies that for the sake of the proof, we may and will henceforth assume that the marginals $F_1, \ldots, F_d$ are unit-Pareto and that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample of P.

Recall that, for any $A \in \mathcal{A}$ and $\varepsilon > 0$, we have

$$\forall A \in \mathcal{A}, \qquad \Gamma_A^- \subset \mathcal{C}_A \subset \Gamma_A^+.$$

We shall construct an event $\mathcal{E}_1$ with probability at least $1-\delta$, on which we have

$$\forall A \in \mathcal{A}, \qquad \tfrac{n}{k}\Gamma_A^- \subset \widehat{\Gamma}_A \subset \tfrac{n}{k}\Gamma_A^+. \tag{4.29}$$

Then, on the event $\mathcal{E}_1$, the decomposition (4.14) of the estimation error holds true. We treat the three terms involved in the decomposition in turn. At the end, we construct the event $\mathcal{E}_1$ with the required properties.

Stochastic error. We apply Theorem 5 to the collection

$$\mathcal{F} = \{\tfrac{n}{k}\Gamma_A^\sigma : \ \sigma \in \{-,+\}, \ A \in \mathcal{A}\}, \tag{4.30}$$

which is of finite VC dimension $V(\mathcal{F}) \leq V_{\mathcal{C}}$. For every $A \in \mathcal{A}$, we have

$$\tfrac{n}{k}\Gamma_A^- \subset \tfrac{n}{k}\Gamma_A^+ \subset \left\{x \in [0,\infty)^d : \ \|x\| \geq \tfrac{n}{k}\tfrac{1}{1+\Delta}\right\}.$$

Since the margins of P are unit-Pareto, it follows that the probability p appearing in Theorem 5 applied to $\mathcal{F}$ is bounded by

$$P\left[\bigcup_{j=1}^d \left\{x \in [0,\infty)^d : x_j \geq \tfrac{n}{k}\tfrac{1}{1+\Delta}\right\}\right] \leq d(1+\Delta)\tfrac{k}{n}.$$

As a consequence, on an event $\mathcal{E}_2$ with probability at least $1-\delta/(d+1)$, we have,

$$\begin{aligned}
&\sup_{A\in\mathcal{A},\, \sigma\in\{-,+\}} \tfrac{n}{k}|P_n(\tfrac{n}{k}\Gamma_A^\sigma) - P(\tfrac{n}{k}\Gamma_A^\sigma))| \\
&\leq \tfrac{n}{k} C\left(\sqrt{d(1+\Delta)\tfrac{k}{n}V(\mathcal{F})n^{-1}\log((d+1)/\delta)} + n^{-1}\log((d+1)/\delta)\right) \\
&= C\left(\sqrt{d(1+\Delta)V(\mathcal{F})k^{-1}\log((d+1)/\delta)} + k^{-1}\log((d+1)/\delta)\right).
\end{aligned}$$

This is the term labelled error in the statement of Theorem 2. Since $\mathcal{E}_1$ and $\mathcal{E}_2$ have probabilities at least $1-d\delta/(d+1)$ and $1-\delta/(d+1)$, respectively, the event $\mathcal{E}_1 \cap \mathcal{E}_2$ has probability at least $1-\delta$.

Bias term. Taking the supremum over $A \in \mathcal{A}$ immediately yields the bias term in the statement of Theorem 2.

Framing gap. For $A \in \mathcal{A}$, any point $x \in \Gamma_A^+ \setminus \Gamma_A^-$ has either a norm $\|x\|_\infty$ in between certain two bounds or an angle $\theta(x)$ contained in a set of the form $A_+(\varepsilon) \setminus A_-(\varepsilon)$ for some $\varepsilon > 0$. Specifically,

$$\begin{aligned}
\mu(\Gamma_A^+ \setminus \Gamma_A^-) \leq\ & \mu\left(\left\{x \in [0,\infty)^d : \ (1+\tfrac{1}{n}-\tfrac{1}{k}+\Delta)^{-1} \leq \|x\| < (1+\tfrac{1}{n}-\tfrac{1}{k}-\Delta)^{-1}\right\}\right) \\
&+ \mu\left(\left\{x \in [0,\infty)^d : \ \|x\| \geq 1, \ \theta(x) \in A_+(3\Delta\|x\|) \setminus A_-(3\Delta\|x\|)\right\}\right).
\end{aligned} \tag{4.31}$$

The first term on the right-hand side in (4.31) is equal to

$$\left((1+\tfrac{1}{n}-\tfrac{1}{k}+\Delta) - (1+\tfrac{1}{n}-\tfrac{1}{k}-\Delta)\right)\mu(\{x \in [0,\infty)^d : \|x\| \geq 1\}) = 2\Delta\Phi(\mathbb{S}).$$

The second term on the right-hand side in (4.31) can be computed via the product representation (4.3) of μ in polar coordinates: in view of (4.16) in Assumption 1, the result is

$$\int_1^\infty \Phi\Big(A_+(3\Delta r) \setminus A_-(3\Delta r)\Big)\frac{\mathrm{d}r}{r^2} \le \int_1^\infty \min\{\Phi(\mathbb{S}), 3c\Delta r\}\frac{\mathrm{d}r}{r^2} = 3c\Delta(1 + \log\Phi(\mathbb{S}) - \log(3c\Delta)),$$

since $\int_1^\infty \min(b, ar)\frac{\mathrm{d}r}{r^2} = a(\log(b/a) + 1)$ for $a \in (0, b]$ and $3c\Delta \le 1 \le \Phi(\mathbb{S})$ by assumption.

The resulting upper bound in (4.31) does not depend on $A \in \mathcal{A}$. Bounding $\Phi(\mathbb{S})$ by d, we obtain the term labelled gap in the statement of Theorem 2.

Construction of the event $\mathcal{E}_1$. We still need to construct an event $\mathcal{E}_1$ with probability at least $1 - d\delta$ on which the inclusions (4.29) hold. To do this, we apply Theorem 5 to each of the collections

$$\mathcal{F}_j = \Big\{\{x \in [0,\infty)^d : x_j > \tfrac{n}{k}y\} : \ y \in [\rho, \infty)\Big\}, \qquad j = 1, \dots, d.$$

Fix $j = 1, \dots, d$ and let $P_{n,j} = n^{-1}\sum_{i=1}^n \delta_{X_{ij}}$. Each set in the collection $\mathcal{F}_j$ is a subset of $\{x \in [0,\infty)^d : x_j > \frac{n}{k}\rho\}$, whose P-probability is $p = \frac{k}{n}\rho^{-1}$. The class $\mathcal{F}_j$ has VC dimension 1. By Theorem 5, there exists an event $\mathcal{E}_{1,j}$ with probability at least $1 - \delta/(d+1)$ on which

$$\sup_{x_j \ge \frac{n}{k}\rho} \Big|P_{n,j}((x_j,\infty)) - x_j^{-1}\Big| \le C\left(\sqrt{\tfrac{k}{n}\rho^{-1}n^{-1}\log((d+1)/\delta)} - n^{-1}\log((d+1)/\delta)\right) = n^{-1}C\left(\sqrt{k\rho^{-1}\log((d+1)/\delta)} + \log((d+1)/\delta)\right) = \tfrac{k}{n}\Delta. \tag{4.32}$$

Since

$$\hat{v}_j(x_j) = \frac{1}{1 - \frac{n}{n+1}P_{n,j}((-\infty, x_j])} = \frac{n+1}{nP_{n,j}((x_j,\infty)) + 1},$$

we have on the event $\mathcal{E}_{1,j}$ the bounds

$$\forall x_j \ge \tfrac{n}{k}\rho, \qquad \frac{n+1}{nx_j^{-1} + k\Delta + 1} \le \hat{v}_j(x_j) \le \frac{n+1}{(nx_j^{-1} - k\Delta)_+ + 1}. \tag{4.33}$$

Moreover, since $\hat{v}_j$ is monotone, we have on $\mathcal{E}_{1,j}$ the inequalities

$$\forall x_j \le \tfrac{n}{k}\rho, \qquad \hat{v}_j(x_j) \le \hat{v}_j(\tfrac{n}{k}\rho) \le \frac{n+1}{k(\rho^{-1} - \Delta) + 1} \le \frac{n}{k}\frac{\rho}{1 - \rho\Delta} \le \frac{n}{k}\tau. \tag{4.34}$$

Let $\mathcal{E}_1 = \bigcap_{j=1}^d \mathcal{E}_{1,j}$, the probability of which is at least $1 - d\delta/(d+1)$, as required. We need to show that on $\mathcal{E}_1$, the inclusions (4.29) hold. To do so, we proceed in steps. Throughout, we work on $\mathcal{E}_1$.

Step 1: Restriction to $(\frac{n}{k}\rho, \infty)^d$. — If $x \in [0,\infty)^d$ is such that $\|\hat{v}(x)\|_\infty \ge \frac{n}{k}$ but there exists $j = 1, \dots, d$ with $x_j \le \frac{n}{k}\rho$, then, by (4.34), we have on $\mathcal{E}_1$ the bound

$$\theta_j(\hat{v}(x)) = \frac{\hat{v}_j(x_j)}{\|\hat{v}(x)\|} \le \tau$$

and thus, by Assumption 1, necessarily $\theta(\hat{v}(x)) \notin A$ for all $A \in \mathcal{A}$. Hence, on $\mathcal{E}_1$, we have

$$\widehat{\Gamma}_A = \left\{ x \in (\tfrac{n}{k}\rho, \infty)^d : \ \|\hat{v}(x)\| \geq \tfrac{n}{k}, \ \theta(\hat{v}(x)) \in A \right\}. \tag{4.35}$$

Step 2: Radial framing. — On $\mathcal{E}_1$, we have

$$\forall x \in [0,\infty)^d, \qquad \|x\| \geq \frac{n}{k} \frac{1}{1 + \frac{1}{n} - \frac{1}{k} - \Delta} \implies \|\hat{v}(x)\| \geq \frac{n}{k}. \tag{4.36}$$

Indeed, for such x, there exists $j = 1, \ldots, d$ such that $x_j \geq \frac{n}{k} \frac{1}{1+1/n-1/k-\Delta}$ and thus, on $\mathcal{E}_1$,

$$\hat{v}_j(x_j) \geq \hat{v}_j(\tfrac{n}{k} \tfrac{1}{1+1/n-1/k-\Delta}) \geq \frac{n+1}{n\frac{k}{n}(1 + \frac{1}{n} - \frac{1}{k} - \Delta) + k\Delta + 1} = \frac{n+1}{k(1+\frac{1}{n})} = \frac{n}{k}.$$

Furthermore, on $\mathcal{E}_1$, we have

$$\forall x \in [\tfrac{n}{k}\rho, \infty)^d, \qquad \|\hat{v}(x)\| \geq \frac{n}{k} \implies \|x\| \geq \frac{n}{k} \frac{1}{1 + \frac{1}{n} - \frac{1}{k} + \Delta}. \tag{4.37}$$

Indeed, for such x, there exists $j = 1, \ldots, d$ such that $\hat{v}_j(x_j) \geq \frac{n}{k}$ and thus, by (4.33), also

$$\frac{n+1}{(nx_j^{-1} - k\Delta)_+ + 1} \geq \frac{n}{k},$$

which implies

$$x_j \geq \frac{n}{k} \frac{1}{1 + \frac{1}{n} - \frac{1}{k} + \Delta}.$$

Recall $\widehat{\Gamma}_A$ in (4.12). In view of (4.35), it follows that, on $\mathcal{E}_1$, for all $A \in \mathcal{A}$, we have

$$\begin{aligned} \widehat{\Gamma}_A &\supset \left\{ x \in (\tfrac{n}{k}\rho, \infty)^d : \ \|x\| \geq \frac{n}{k} \frac{1}{1 + \frac{1}{n} - \frac{1}{k} - \Delta}, \ \theta(\hat{v}(x)) \in A \right\}, \\ \widehat{\Gamma}_A &\subset \left\{ x \in (\tfrac{n}{k}\rho, \infty)^d : \ \|x\| \geq \frac{n}{k} \frac{1}{1 + \frac{1}{n} - \frac{1}{k} + \Delta}, \ \theta(\hat{v}(x)) \in A \right\}. \end{aligned} \tag{4.38}$$

Step 3: Angular framing. — We will show that, on $\mathcal{E}_1$, for any $x \in [\frac{n}{k}\rho, \infty)^d$ and $A \in \mathcal{A}$,

$$\theta(x) \in A_-(3\tfrac{k}{n}\Delta\|x\|) \implies \theta(\hat{v}(x)) \in A \implies \theta(x) \in A_+(3\tfrac{k}{n}\Delta\|x\|). \tag{4.39}$$

In combination with (4.38), this will show the inclusions (4.29) and thus finish the proof.

Fix such x and A. Put $\varepsilon = 3\frac{k}{n}\Delta\|x\|$. We consider two cases: $\varepsilon < 1$ and $\varepsilon \geq 1$.

If $\varepsilon \geq 1$, the two implications in (4.39) are trivially fulfilled: Since the $\|\cdot\|$-diameter of $\mathbb{S}$ is equal to 1, we have $A_-(\varepsilon) = \varnothing$ (as $\mathbb{S} \setminus A$ is not-empty) while $A_+(\varepsilon) = \mathbb{S}$.

The interesting case is thus $\varepsilon < 1$. By Lemma 2, we have

$$\|\theta(\hat{v}(x)) - \theta(x)\| \leq 2\frac{\|\hat{v}(x) - x\|}{\|x\|}.$$

Further, by (4.33), for all $j = 1, \ldots, d$, since $nx_j^{-1} - k\Delta \geq n \cdot 3\frac{k}{n}\Delta - k\Delta > 0$,

$$\begin{aligned}|\hat{v}_j(x_j) - x_j| &= \max_{\sigma\in\{-1,+1\}} \left| \frac{n+1}{nx_j^{-1} + \sigma k\Delta + 1} - x_j \right| \\ &= x_j \max_{\sigma\in\{-1,+1\}} \left| \frac{n+1}{n + (\sigma k\Delta + 1)x_j} - 1 \right| \\ &= x_j \max_{\sigma\in\{-1,+1\}} \frac{|1 - (\sigma k\Delta + 1)x_j|}{n + (\sigma k\Delta + 1)x_j}.\end{aligned}$$

Recall that $x_j \geq \frac{n}{k}\rho \geq 1$ and $k\Delta \geq 2$. For he case $\sigma = +1$, we have

$$\frac{|1 - (+k\Delta + 1)x_j|}{n + (+k\Delta + 1)x_j} \leq \frac{(k\Delta + 1)x_j}{n + (k\Delta + 1)x_j} \leq \frac{3}{2}\frac{k}{n}\Delta x_j.$$

For the case $\sigma = -1$, we also have

$$\frac{|1 - (-k\Delta + 1)x_j|}{n + (-k\Delta + 1)x_j} = \frac{(k\Delta - 1)x_j - 1}{n + (-k\Delta + 1)x_j} \leq \frac{k\Delta x_j}{n - k\Delta x_j} \leq \frac{3}{2}\frac{k}{n}\Delta x_j,$$

since $\varepsilon < 1$ implies $n - k\Delta x_j \geq n - k\Delta\|x\|_\infty \geq n - n/3 = 2n/3$. We conclude that

$$\|\hat{v}(x) - x\| \leq \frac{3}{2}\frac{k}{n}\Delta\|x\|^2$$

and thus

$$\|\theta(\hat{v}(x)) - \theta(x)\| \leq 3\tfrac{k}{n}\Delta\|x\| = \varepsilon.$$

The implications (4.39) now follow by definition of $A_-(\varepsilon)$ and $A_+(\varepsilon)$.

We conclude that, on $\mathcal{E}_1$, the inclusions (4.29) hold, as required. The proof Theorem 2 is complete. □

Proof of Theorem 3

The structure of the proof is identical to the one of the proof of Theorem 2 in Section 4.6.

Reduction to unit-Pareto margins. This part goes through without modification.

Decomposition of the estimation error. Rather than the set $\widehat{\Gamma}_A$ in (4.12), we now write

$$\widehat{\Gamma}_A^M = \hat{v}^{-1}(\tfrac{n}{k}\mathcal{C}_A^M) = \left\{x \in [0,\infty)^d : \ \tfrac{n}{k} \leq \|\hat{v}(x)\|_\infty < \tfrac{n}{k}M, \ \theta(\hat{v}(x)) \in A\right\}, \quad (4.40)$$

so that

$$\begin{aligned}\widehat{\Phi}_M(A) &= \tfrac{M}{M-1}\tfrac{n}{k}\widehat{P}_n(\tfrac{n}{k}\mathcal{C}_A^M) = \tfrac{M}{M-1}\tfrac{n}{k}P_n\left(\hat{v}^{-1}(\tfrac{n}{k}\mathcal{C}_A^M)\right) = \tfrac{M}{M-1}\tfrac{n}{k}P_n(\widehat{\Gamma}_A^M), \\ \Phi(A) &= \tfrac{M}{M-1}\mu(\mathcal{C}_A^M) = \tfrac{M}{M-1}\tfrac{n}{k}\mu(\tfrac{n}{k}\mathcal{C}_A^M).\end{aligned}$$

Recall the sets $\Gamma_{A,1}^{M,-}$ and $\Gamma_{A,1}^{M,+}$ defined in (4.19). We have

$$1+\tfrac{1}{n}-\tfrac{1}{k}-\Delta<1<1+\tfrac{1}{n}-\tfrac{1}{k}+\Delta,$$
$$((1+\tfrac{1}{n})\tfrac{1}{M}-\tfrac{1}{k}-\Delta)_+<\tfrac{1}{M}<(1+\tfrac{1}{n})\tfrac{1}{M}-\tfrac{1}{k}+\Delta.$$

Since moreover $A_-(\varepsilon)\subset A\subset A_+(\varepsilon)$ for any $A\in\mathcal{A}$ and $\varepsilon>0$, we have

$$\forall A\in\mathcal{A},\qquad \Gamma_{A,1}^{M,-}\subset\mathcal{C}_A^M\subset\Gamma_{A,1}^{M,+}.$$

We will show that on the event $\mathcal{E}_1$ constructed in the proof of Theorem 2 also

$$\forall A\in\mathcal{A},\qquad \tfrac{n}{k}\Gamma_{A,1}^{M,-}\subset\widehat{\Gamma}_A^M\subset\tfrac{n}{k}\Gamma_{A,1}^{M,+}. \tag{4.41}$$

Then, on the event $\mathcal{E}_1$, we have

$$\begin{aligned}
&\tfrac{M-1}{M}\Big(\widehat{\Phi}_M(A)-\Phi(A)\Big)\\
&=\tfrac{n}{k}P_n(\widehat{\Gamma}_A^M)-\mu(\mathcal{C}_A^M)\\
&\le\tfrac{n}{k}P_n(\tfrac{n}{k}\Gamma_{A,1}^{M,+})-\mu(\Gamma_{A,1}^{M,-})\\
&\le\tfrac{n}{k}|P_n(\tfrac{n}{k}\Gamma_{A,1}^{M,+})-P(\tfrac{n}{k}\Gamma_{A,1}^{M,+})|+|\tfrac{n}{k}P(\tfrac{n}{k}\Gamma_{A,1}^{M,+})-\mu(\Gamma_{A,1}^{M,+})|+\mu(\Gamma_{A,1}^{M,+}\setminus\Gamma_{A,1}^{M,-}).
\end{aligned}$$

A lower bound for the estimation error can be derived in a similar way, yielding, on $\mathcal{E}_1$,

$$\begin{aligned}
|\widehat{\Phi}_M(A)-\Phi(A)|\le{}&\tfrac{M}{M-1}\max_{B\in\{\Gamma_{A,1}^{M,+},\Gamma_{A,1}^{M,-}\}}\tfrac{n}{k}|P_n(\tfrac{n}{k}B)-P(\tfrac{n}{k}B)| && \text{(stochastic error)}\\
&+\tfrac{M}{M-1}\max_{B\in\{\Gamma_{A,1}^{M,+},\Gamma_{A,1}^{M,-}\}}\tfrac{n}{k}|P(\tfrac{n}{k}B)-\mu(B)| && \text{(bias term)}\\
&+\tfrac{M}{M-1}\mu(\Gamma_{A,1}^{M,+}\setminus\Gamma_{A,1}^{M,-}) && \text{(framing gap).}
\end{aligned}$$

The bias term is the one stated in Theorem 3. Below, we treat the stochastic error and the framing gap, and we show that the inclusions (4.41) hold on $\mathcal{E}_1$.

Stochastic error. We apply Theorem 5 to the collection $\mathcal{F}^M$. For every $A\in\mathcal{A}$, we have

$$\tfrac{n}{k}\Gamma_{A,1}^{M,-}\subset\tfrac{n}{k}\Gamma_{A,1}^{M,+}\subset\Big\{x\in[0,\infty)^d:\|x\|_\infty\ge\tfrac{n}{k}\tfrac{1}{1+\Delta}\Big\}.$$

Since the margins of P are unit-Pareto, it follows that the probability p appearing in Theorem 5 applied to $\mathcal{F}$ is bounded by

$$P\Big[\bigcup_{j=1}^d\Big\{x\in[0,\infty)^d:x_j\ge\tfrac{n}{k}\tfrac{1}{1+\Delta}\Big\}\Big]\le d(1+\Delta)\tfrac{k}{n}.$$

As a consequence, on an event $\mathcal{E}_2^M$ with probability at least $1-\delta$, we have, in view of Assumption 1,

$$\begin{aligned}
&\sup_{A\in\mathcal{A}}\max_{B\in\{\Gamma_{A,1}^{M,+},\Gamma_{A,1}^{M,-}\}}\tfrac{n}{k}|P_n(\tfrac{n}{k}B)-P(\tfrac{n}{k}B)|\\
&\qquad\le C\left(\sqrt{d(1+\Delta)V(\mathcal{F}^M)k^{-1}\log(1/\delta)}+k^{-1}\log(1/\delta)\right).
\end{aligned}$$

Apart from the factor $\frac{M}{M-1}$, this is the term labelled error_M in the statement of Theorem 3. Since $\mathcal{E}_1$ and $\mathcal{E}_2^M$ have probabilities at least $1-d\delta$ and $1-\delta$, respectively, the event $\mathcal{E}_1\cap\mathcal{E}_2^M$ has probability at least $1-(d+1)\delta$.

Framing gap. For $A \in \mathcal{A}$, any point $x \in \Gamma_{A,1}^{M,+} \setminus \Gamma_{A,1}^{M,-}$ has either a norm $\|x\|_\infty$ in between certain two bounds or an angle $\theta(x)$ contained in a set of the form $A_+(\varepsilon) \setminus A_-(\varepsilon)$ for some $\varepsilon > 0$. Specifically,

$$\begin{aligned}
&\mu\Big(\Gamma_{A,1}^{M,+} \setminus \Gamma_{A,1}^{M,-}\Big) \\
&\le \mu\left(\left\{x \in [0,\infty)^d : \Big(1 + \tfrac{1}{n} - \tfrac{1}{k} + \Delta\Big)^{-1} \le \|x\|_\infty \le \Big(1 + \tfrac{1}{n} - \tfrac{1}{k} - \Delta\Big)^{-1}\right\}\right) \\
&\quad + \mu\left(\left\{x \in [0,\infty)^d : \Big((1 + \tfrac{1}{n})\tfrac{1}{M} - \tfrac{1}{k} + \Delta\Big)^{-1} \le \|x\|_\infty \le \Big((1 + \tfrac{1}{n})\tfrac{1}{M} - \tfrac{1}{k} - \Delta\Big)_+^{-1}\right\}\right) \\
&\quad + \mu\left(\left\{x \in [0,\infty)^d : 1 \le \|x\|_\infty \le M,\ \theta(x) \in A_+(3\Delta\|x\|_\infty) \setminus A_-(3\Delta\|x\|_\infty)\right\}\right).
\end{aligned}$$

By homogeneity of μ, both the first and second terms are bounded by $2\Delta\Phi(\mathbb{S}) \le 2d\Delta$. For the third term, the product representation (4.3) of μ in polar coordinates and the inequality (4.16) in Assumption 1 yield the bound

$$\begin{aligned}
\int_1^M \Phi\Big(A_+(3\Delta r) \setminus A_-(3\Delta r)\Big)\frac{\mathrm{d}r}{r^2} &\le \int_1^M \Phi(\mathbb{S}) \wedge (3c\Delta r)\frac{\mathrm{d}r}{r^2} \\
&\le \int_1^M d \wedge (3c\Delta r)\frac{\mathrm{d}r}{r^2} \\
&= 3c\Delta \ln\Big(M \wedge \tfrac{d}{3c\Delta}\Big) + \Big(3c\Delta - \tfrac{d}{M}\Big)_+,
\end{aligned}$$

since for $a \in (0, b]$ we have

$$\int_1^M b \wedge (ar)\frac{\mathrm{d}r}{r^2} = a\ln\Big(M \wedge \tfrac{b}{a}\Big) + \Big(a - \tfrac{b}{M}\Big)_+$$

and since $3c\Delta \le 1 \le \Phi(\mathbb{S})$ by assumption.

Summing the contributions, we get an upper bound on $\mu\Big(\Gamma_{A,1}^{M,+} \setminus \Gamma_{A,1}^{M,-}\Big)$ that does not depend on $A \in \mathcal{A}$. Up to a factor $M/(M-1)$, we obtain the term labelled gap_M in the statement of Theorem 3.

The inclusions (4.41) hold on event $\mathcal{E}_1$. The event $\mathcal{E}_1$ was constructed in the course of the proof of Theorem 2. On that event, the inequalities (4.33) and (4.34) hold.

Step 1: Restriction to $(\frac{n}{k}\rho, \infty)^d$. — By exactly the same reasoning as in the proof of Theorem 2, we conclude that, on $\mathcal{E}_1$, the set $\widehat{\Gamma}_A^M$ in (4.40) satisfies

$$\widehat{\Gamma}_A^M = \Big\{x \in (\tfrac{n}{k}\rho, \infty)^d :\ \tfrac{n}{k} \le \|\hat{v}(x)\|_\infty < \tfrac{n}{k}M,\ \theta(\hat{v}(x)) \in A\Big\}. \tag{4.42}$$

Step 2: Radial framing. — On $\mathcal{E}_1$, apart from (4.36) and (4.37), we also have

$$\forall x \in [0,\infty)^d, \qquad \|x\|_\infty \ge \frac{n}{k}\frac{1}{((1 + \frac{1}{n})\frac{1}{M} - \frac{1}{k} - \Delta)_+} \implies \|\hat{v}(x)\|_\infty \ge \frac{n}{k}M \tag{4.43}$$

[if $(1 + \frac{1}{n})\frac{1}{M} - \frac{1}{k} - \Delta \le 0$, then the condition is void] as well as

$$\forall x \in [\tfrac{n}{k}\rho, \infty)^d, \qquad \|\hat{v}(x)\|_\infty \ge \frac{n}{k}M \implies \|x\|_\infty \ge \frac{n}{k}\frac{1}{(1 + \frac{1}{n})\frac{1}{M} - \frac{1}{k} + \Delta}. \tag{4.44}$$

Indeed, for such x as in (4.43), if $(1+\frac{1}{n})\frac{1}{M} - \frac{1}{k} - \Delta > 0$, then there exists $j = 1, \ldots, d$ such that $x_j \geq \frac{n}{k}\frac{1}{(1+1/n)/M - 1/k - \Delta}$ and thus, on $\mathcal{E}_1$, by (4.33)

$$\hat{v}_j(x_j) \geq \hat{v}_j\Big(\tfrac{n}{k}\tfrac{1}{(1+1/n)/M-1/k-\Delta}\Big) \geq \frac{n+1}{n\frac{k}{n}((1+\frac{1}{n})\frac{1}{M} - \frac{1}{k} - \Delta) + k\Delta + 1} = \frac{n}{k}M.$$

Similarly, for x as in (4.44), we have by (4.33) also

$$\frac{n+1}{(nx_j^{-1} - k\Delta)_+ + 1} \geq \frac{n}{k}M,$$

which, after some algebra, can be shown to imply

$$x_j \geq \frac{n}{k}\frac{1}{(1+\frac{1}{n})\frac{1}{M} - \frac{1}{k} + \Delta}.$$

In view of (4.42) together with the inequalities in this step, we get that, for $\widehat{\Gamma}_A^M$ in (4.42),

$$\begin{aligned}
&\left\{x \in (\tfrac{n}{k}\rho, \infty)^d : \ \frac{n}{k}\frac{1}{1+\frac{1}{n}-\frac{1}{k}-\Delta} \leq \|x\|_\infty < \frac{n}{k}\frac{1}{(1+\frac{1}{n})\frac{1}{M}-\frac{1}{k}+\Delta}, \ \theta(\hat{v}(x)) \in A\right\}, \\
&\subset \widehat{\Gamma}_A^M \\
&\subset \left\{x \in (\tfrac{n}{k}\rho, \infty)^d : \ \frac{n}{k}\frac{1}{1+\frac{1}{n}-\frac{1}{k}+\Delta} \leq \|x\|_\infty < \frac{n}{k}\frac{1}{((1+\frac{1}{n})\frac{1}{M}-\frac{1}{k}-\Delta)_+}, \ \theta(\hat{v}(x)) \in A\right\}.
\end{aligned} \tag{4.45}$$

Step 3: Angular framing. — Again, this step does not involve M. The implications (4.39) hold. In combination with (4.45), we conclude that on $\mathcal{E}_1$, the inclusions (4.41) hold. The proof of Theorem 3 is complete. □

Proof of Theorem 4

By construction of $\hat{A}$, we have

$$\forall A \in \mathcal{A}, \qquad \widehat{\Phi}(A) \geq \alpha - \psi \implies \lambda(\hat{A}) \leq \lambda(A).$$

If $A \in \mathcal{A}$ satisfies $\Phi(A) \geq \alpha$, then $\widehat{\Phi}(A) \geq \Phi(A) - \psi \geq \alpha - \psi$ and thus $\lambda(\hat{A}) \leq \lambda(A)$.

Furthermore, since $\widehat{\Phi}(\hat{A}) \geq \alpha - \psi$, necessarily $\Phi(\hat{A}) \geq \widehat{\Phi}(\hat{A}) - \psi \geq \alpha - 2\psi$.

Proof of Lemma 1

For any $x > 0$ and $1 \leq j \leq d$,

$$\begin{aligned}
1 - F_j(x) &= \mathbb{P}\{R\Theta_j > x\} \\
&= \mathbb{E}\{\mathbb{P}\{R > x/\Theta_j\} \mid \Theta_j\} \\
&= \mathbb{E}\{\min(1, \Theta_j/x)\}.
\end{aligned}$$

Since $\Theta \in \mathbb{S}_1$ we have $\Theta_j \leq 1$ so that for $x \geq 1$, $\Theta_j/x \leq 1$ almost surely. In addition $\mathbb{E}(\Theta_j) = 1/d$, hence

$$1/(1 - F_j(x)) = \begin{cases} dx & \text{if } x \geq 1 \\ \mathbb{E}\{\min(1, \Theta_j/x)\} & \text{if } x < 1, \end{cases}$$

which proves that for $x \in [1, \infty)^d, v(x) = dx \in [d, \infty)^d$. For the second part of statement 1, notice that by monotonicity of F_j, for any $x \in \mathbb{R}_+^d \setminus \{0\}$, the following implication holds:

$$\exists j, x_j \leq 1 \Rightarrow 1/(1 - F_j(x_j)) \leq 1/(1 - F_j(1)) = d \Rightarrow v(x) \notin (d, \infty)^d.$$

which proves the desired statement by contraposition.

Turning to statement 2, remind that $\Phi(A) = \lim_t t\mathbb{P}\{v(X) \in t\mathcal{C}_A\}$. and notice that the assumption on A implies that $\exists \tilde{\tau} > 0$ such that $\min_{j \leq d} \inf\{x_j \ : \ x \in A\} \geq \tilde{\tau}$. Then for $t > d/\tilde{\tau}$ the inclusion $t\ \mathcal{C}_A \subset (d, +\infty)^d$ holds true. Thus for such t, $v(X) \in t\mathcal{C}_A \Rightarrow v(X) = dX$, whence

$$\begin{aligned} t\mathbb{P}\{v(X) \in t\ \mathcal{C}_A\} &= t\mathbb{P}\{dR\Theta \in t\mathcal{C}_A\} \\ &= t\mathbb{E}\Big(\mathbb{1}\{\Theta/\|\Theta\|_\infty \in A\}\mathbb{P}\{dR\|\Theta\|_\infty > t, \mid \Theta\}\Big) \\ &= d\mathbb{E}\Big(\mathbb{1}\{\Theta/\|\Theta\|_\infty \in A\}\mathbb{P}\{R\|\Theta\|_\infty > 1, \mid \Theta\}\Big) \\ &= d\mathbb{P}\{R\Theta \in \mathcal{C}_A\} \\ &= d\mathbb{P}\{X \in \mathcal{C}_A\}\ , \end{aligned}$$

which proves statement 2. Finally, if A satifies condition 4.15, we may take $\tilde{\tau} = \tau$ in the above argument which shows that the function $t \mapsto t\mathbb{P}\{V \in t\mathcal{C}_A\}$ is constant on $[d/\tau, \infty)$, so that it reaches its limit. This proves statement 3 and the proof of Lemma 1 is complete.

4.7 References

Clémençon, S., H. Jalalzai, J. Segers and A. Sabourin. 2020, ■Concentration bounds for the empirical angular measure with statistical learning applications■, in *preparation*. 66, 67, 68, 69

De Haan, L. and A. Ferreira. 2007, *Extreme value theory: an introduction*, Springer Science & Business Media. 62

Einmahl, J. and D. Mason. 1992, ■Generalized quantile process■, *The Annals of Statistics*, vol. 20, p. 1062–1078. 70

Einmahl, J. H., L. De Haan and V. I. Piterbarg. 2001, ■Nonparametric estimation of the spectral measure of an extreme value distribution■, *Annals of Statistics*, p. 1401–1423. 63, 65, 67

Einmahl, J. H. and J. Segers. 2009, ■Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution■, *The Annals of Statistics*, p. 2953–2989. 63, 65, 67

GENEST, C. and J. SEGERS. 2010, ■On the covariance of the asymptotic empirical copula process■, *Journal of Multivariate Analysis*, vol. 101, n° 8, p. 1837–1845. 74

GOIX, N., A. SABOURIN and S. CLÉMENÇON. 2015, ■Learning the dependence structure of rare events: a nonasymptotic study■, in *Proceedings of the International Conference on Learning Theory, COLT'15*. 63, 67, 69, 77, 78

HULT, H. and F. LINDSKOG. 2006, ■Regular variation for measures on metric spaces.■, *Publications de l'Institut Mathematique*, , n° 94. 62

KOLTCHINSKII, V. 2006, ■Local Rademacher complexities and oracle inequalities in risk minimization (with discussion)■, *The Annals of Statistics*, vol. 34, p. 2593–2706. 78

LUGOSI, G. 2002, ■Pattern classification and learning theory■, in *Principles of nonparametric learning (Udine, 2001)*, *CISM Courses and Lect.*, vol. 434, Springer, Vienna, p. 1–56. 78

POLONIK, W. 1997, ■Minimum volume sets and generalized quantile processes■, *Stochastic Processes and their Applications*, vol. 69, n° 1, p. 1–24. 70, 71

RIGOLLET, P. and R. VERT. 2009, ■Fast rates for plug-in estimators of density level sets■, *Bernoulli*, vol. 14, n° 4, p. 1154–1178. 71

SCOTT, C. and R. NOWAK. 2006, ■Learning Minimum Volume Sets■, *Journal of Machine Learning Research*, vol. 7, p. 665–704. 71

THOMAS, A., S. CLEMENCON, A. GRAMFORT and A. SABOURIN. 2017, ■Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere■, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 54, édité par A. Singh and J. Zhu, PMLR, Fort Lauderdale, FL, USA, p. 1011–1019. URL http://proceedings.mlr.press/v54/thomas17a.html. 70, 71, 76

TSYBAKOV, A. 1997, ■On nonparametric estimation of density level sets■, *Annals of Statistics*, vol. 25, p. 948–969. 71

VAN DER VAART, A. W. and J. A. WELLNER. 1996, *Weak Convergence and Empirical Process. With Applications to Statistics*, Springer-Verlag, New York. 68

Chapter 5

On Binary Classification in Extreme Regions

Chapter abstract

In pattern recognition, a random label Y is to be predicted based upon observing a random vector X valued in $\mathbb{R}^d$ with $d \geq 1$ by means of a classification rule with minimum probability of error. In a wide variety of applications, ranging from finance/insurance to environmental sciences through teletraffic data analysis for instance, extreme (*i.e.* very large) observations X are of crucial importance, while contributing in a negligible manner to the (empirical) error however, simply because of their rarity. As a consequence, empirical risk minimizers generally perform very poorly in extreme regions. It is the purpose of this chapter to develop a general framework for classification in the extremes. Precisely, under non-parametric heavy-tail assumptions for the class distributions, we prove that a natural and asymptotic notion of risk, accounting for predictive performance in extreme regions of the input space, can be defined and show that minimizers of an empirical version of a non-asymptotic approximant of this dedicated risk, based on a fraction of the largest observations, lead to classification rules with good generalization capacity, by means of maximal deviation inequalities in low probability regions. Beyond theoretical results, numerical experiments are presented in order to illustrate the relevance of the approach developed.

5.1 Introduction

Because it covers a wide range of practical applications and its probabilistic theory can be extended to some extent to various other prediction problems, binary classification can be considered as the flagship problem in statistical learning. In the standard setup, (X, Y) is a random pair defined on a certain probability space with (unknown) joint probability distribution P, where the (output) r.v. Y is a binary label, taking its values in $\{-1, +1\}$ say, and X models some information, valued in

$\mathbb{R}^d$ and hopefully useful to predict Y. In this context, the goal pursued is generally to build, from a training sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ composed of $n \geq 1$ i.i.d. realizations of (X, Y), a classifier $g : \mathbb{R}^d \to \{-1, +1\}$ minimizing the probability of error $L_P(g) = \mathbb{P}\{Y \neq g(X)\}$. The Empirical Risk Minimization paradigm (ERM in abbreviated form, see *e.g.* Devroye and collab. (1996)) suggests considering solutions g_n of the minimization problem $\min_{g \in \mathcal{G}} \widehat{L}_n(g)$, where $\widehat{L}_n(g)$ is a statistical estimate of the risk $L(g)$. In general the empirical version $\widehat{L}_n(g) = (1/n) \sum_{i=1}^n \mathbb{1}\{Y_i \neq g(X_i)\}$ is considered, denoting by $\mathbb{1}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$. This amounts to replacing P in L_P with the empirical distribution of the (X_i, Y_i)'s. The class $\mathcal{G}$ of predictive rules is supposed to be rich enough to contain a reasonable approximant of the minimizer of L_P, *i.e.* the Bayes classifier $g^*(x) = 2\,\mathbb{1}\{\eta(x) \geq 1/2\} - 1$, where $\eta(X) = \mathbb{P}\{Y = 1 \mid X\}$ denotes the posterior probability.

Because extreme observations X, *i.e.* observations whose norm $\|X\|$ exceeds some large threshold $t > 0$, are rare and thus underrepresented in the training dataset $\mathcal{D}_n$ classification errors in these regions of the input space may have a negligible impact on the global prediction error of $\widehat{g}_n$. Notice incidentally that the threshold t may depend on n, since 'large' should be naturally understood as large w.r.t the vast majority of data previously observed. Using the total probability formula, one may indeed write

$$\begin{aligned} L_P(g) =& \mathbb{P}\{\|X\| > t\}\mathbb{P}\{Y \neq g(X) \mid \|X\| > t\} + \\ & \mathbb{P}\{\|X\| \leq t\}\mathbb{P}\{Y \neq g(X) \mid \|X\| \leq t\}. \end{aligned} \tag{5.1}$$

Hence, due to the extremely small order of magnitude of $\mathbb{P}\{\|X\| > t\}$ and of its empirical counterpart, there is no guarantee that the standard ERM strategy produces an optimal classifier on the extreme region $\{x : \ \|x\| > t\}$. In other words the quantity $\mathbb{P}\{Y \neq \widehat{g}_n(X) \mid \|X\| > t\}$ may not be nearly optimal, whereas in certain practical applications (*e.g.* finance, insurance, environmental sciences, aeronautics safety), accurate prediction in extreme regions is crucial.

The purpose of the subsequent analysis is to investigate the problem of building a classifier such that the first term of the decomposition (5.1) is asymptotically minimum as $t \to +\infty$. We thus consider the conditional probability of error, which quantity is next referred to as the *classification risk above level* t, given by

$$L_t(g) := L_{P_t}(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\}, \tag{5.2}$$

denoting by P_t the conditional distribution of (X, Y) given $\|X\| > t$. In this chapter, we address the issue of learning a classifier g_n whose risk $L_t(g_n)$ is asymptotically minimum as $t \to \infty$ with high probability. In order to develop a framework showing that a variant of the ERM principle tailored to this statistical learning problem leads to predictive rules with good generalization capacities, (non-parametric) distributional assumptions related to the tail behavior of the class distributions F_+ and F_-, the conditional distributions of the input r.v. X given $Y = +/-1$, are required. Precisely, we assume that they are both multivariate regularly varying, which correspond to a large non-parametric class of (heavy-tailed) distributions, widely used in applications where the impact of extreme observations should be enhanced, or at least not neglected. Hence,

under appropriate non-parametric assumptions on F_+ and F_-, as well as on the tail behavior of $\eta(x)$, we prove that $\min_g L_t(g)$ converges to a quantity denoted by L^*_∞ and referred to as the *asymptotic risk in the extremes*, as $t \to \infty$. It is also shown that this limit can be interpreted as the minimum classification error related to a (non observable) random pair (X_∞, Y_∞), whose distribution P_∞ corresponds to the limit of the conditional distribution of (X, Y) given $\|X\| > t$, for an appropriate normalization of X, as $t \to \infty$. With respect to the goal set above we next investigate the performance of minimizer $\widehat{g}_{n,\tau}$ of an empirical version of the risk $L_{P_{t_\tau}}$, where t_τ is the $(1-\tau)$ quantile of the r.v. $\|X\|$ and $\tau \ll 1$. The computation of $\widehat{g}_{n,\tau}$ involves the $\lfloor n\tau \rfloor$ input observations with largest norm, and the minimization is performed over a collection of classifiers of finite VC dimension. Based on a variant of the VC inequality tailored to low probability regions, rate bounds for the deviation $L_t(\widehat{g}_{n,\tau}) - L^*_\infty$ are established, of order $O_{\mathbb{P}}(1/\sqrt{n\tau})$ namely. These theoretical results are also illustrated by preliminary experiments based on synthetic data.

The rest of the chapter is organized as follows. Multivariate extreme value theory (MEVT) notions involved in the framework we develop are described in section 5.2, together with the probabilistic setup we consider for classification in the extremes. A notion of risk tailored to this statistical learning task is also introduced therein. Section 5.3 investigates how to extend the ERM principle in this situation. In particular, probability bounds proving the generalization ability of minimizers of a non-asymptotic approximant of the risk previously introduced are established. Illustrative numerical results are displayed in section 5.4, while several concluding remarks are collected in section 5.5. Some technical details and proofs are deferred to the Supplementary Material.

5.2 Probabilistic Framework - Preliminary Results

We start off with recalling concepts, defined in Chapter 2 pertaining to MEVT and next develop a general framework in order to formulate the problem of binary classification in the extremes in a rigorous manner.

5.2.1 Regularly Varying Random Vector

By definition, heavy-tail phenomena are those which are ruled by very large values, occurring with a far from negligible probability and with significant impact on the system under study. When the phenomenon of interest is described by the distribution of a univariate random variable, the theory of regularly varying functions provides the appropriate mathematical framework for the study of heavy-tailed distributions. One may refer to Resnick (1987) for an excellent account of the theory of regularly varying functions and its application to the study of heavy-tailed distributions. For examples of works where such assumptions are considered in the context of statistical learning, see *e.g.* Brownlees and collab. (2015); Carpentier and Valko (2014); Goix and collab. (2016); Ohannessian and Dahleh (2012); Roos and collab. (2006) or Mendelson (2018). Let

$\alpha > 0$, a random variable X is said to be regularly varying with tail index α if $\mathbb{P}\{X > tx \mid X > t\} \xrightarrow[t\to\infty]{} x^{-\alpha}, \qquad x > 1$. This is the case if and only if there exists a function $b : \mathbb{R}_+ \to \mathbb{R}_+^*$ with $b(t) \to \infty$ such that for all $x > 0$, the quantity $t\mathbb{P}\{X/b(t) > x\}$ converges to some limit $h(x)$ as $t \to \infty$. Then b may be chosen as $b(t) = t^{1/\alpha}$ and $h(x) = cx^{-\alpha}$ for some $c > 0$. Based on this characterization, the heavy-tail model can be extended to the multivariate setup. Consider a d-dimensional random vector $X = (X^{(1)}, \ldots, X^{(d)})$ taking its values in $\mathbb{R}_+^d$. Assume that all the $X^{(j)}$ are regularly varying with index $\alpha > 0$. Then the random *vector* X is said to be regularly varying with tail index α if there exists a non null positive Radon measure μ on the punctured space $E = [0, \infty]^d \backslash \{0\}$ and a function $b(t) \to \infty$ such that for all Borel set $A \subset E$ such that $0 \notin \partial A$ and $\mu(\partial A) = 0$,

$$t\mathbb{P}\{X/b(t) \in A\} \xrightarrow[t\to\infty]{} \mu(A).$$

In such a case, the so-called *exponent measure* μ fulfills the homogeneity property $\mu(tC) = t^{-\alpha}\mu(C)$ for all $t > 0$ and any Borel set $C \subset E$. This suggests a decomposition of μ into a radial component and an angular component Φ. For all $x = (x_1, \ldots, x_d) \in \mathbb{R}_+^d$, set

$$\begin{cases} R(x) = \|x\|, \\ \Theta(x) = \left(\dfrac{x_1}{R(x)}, \ldots, \dfrac{x_d}{R(x)}\right) \in S, \end{cases}$$

where S is the positive orthant of the unit sphere in $\mathbb{R}^d$ for the chosen norm $\|\cdot\|$. The choice of the norm is unimportant as all norms are equivalent in $\mathbb{R}^d$. Define an *angular measure* Φ on S as

$$\Phi(B) = \mu\{r\theta \ : \ \theta \in B, r \geq 1\}, \quad B \subset S, \text{measurable}.$$

The angular measure Φ is finite, and the conditional distribution of $(R(X)/t, \Theta(X))$ given that $R(X) > t$ converges as $t \to \infty$ to a limit which admits the following product decomposition: for $r \geq 1$ and $B \subset S$ such that $\Phi(\partial B) = 0$,

$$\begin{aligned} \lim_{t\to\infty} \mathbb{P}\{R(X)/t > r, \Theta(X) \in B \mid R(X) > t\} &= c\,\mu\{x \ : \ R(x) > r, \Theta(x) \in B\} \\ &= c\,r^{-\alpha}\,\Phi(B), \end{aligned}$$

where $c = \mu\{x : R(x) > 1\}^{-1} = \Phi(S)^{-1}$ is a normalizing constant. Thus $c\Phi$ may be viewed as the limiting distribution of $\Theta(X)$ given that $R(X)$ is large.

Remark 3. *It is assumed above that all marginal distributions are tail equivalent to the Pareto distribution with index α. In practice, the tails of the marginals may be different and it may be convenient to work with marginally standardized variables, that is, to separate the margins $F_j(x_j) = \mathbb{P}\{X^{(j)} \leq x_j\}$ from the dependence structure in the description of the joint distribution of X. Consider the standardized variables $V^{(j)} = 1/(1 - F_j(X^{(j)})) \in [1, \infty]$ and $V = (V^{(1)}, \ldots, V^{(d)})$. Replacing X by V permits to take $\alpha = 1$ and $b(t) = t$.*

5.2.2 Classification in the Extremes - Assumptions, Criterion and Optimal Elements

We place ourselves in the binary classification framework recalled in the introduction. For simplicity, we suppose that X takes its values in the positive orthant $\mathbb{R}^d_+$. The general aim is to build from training data in the extreme region (*i.e.* data points (X_i, Y_i) such that $\|X_i\| > t_n$ for a large threshold value $t_n > 0$) a classifier $g_n(x)$ with risk $L_{t_n}(g_n)$ defined in (5.12) being asymptotically minimum as $t_n \to \infty$. In this purpose, we introduce general assumptions guaranteeing that the minimum risk $L_t(g^*)$ above level t has a limit as $t \to \infty$. Throughout the article, we assume that the class distributions F_+ and F_- are heavy-tailed with same index $\alpha = 1$.

Assumption 2. *For all $\sigma \in \{-, +\}$, the conditional distribution of X given $Y = \sigma 1$ is regularly varying with index 1 and angular measure $\Phi_\sigma(d\theta)$ (respectively, exponent measure $\mu_\sigma(dx)$): for $A \subset [0,\infty]^d \backslash \{0\}$ a measurable set such that $0 \notin \partial A$ and $\mu(\partial A) \neq 0$,*

$$t\mathbb{P}\left\{t^{-1}X \in A \;\middle|\; Y = \sigma\, 1\right\} \xrightarrow[t\to\infty]{} \mu_\sigma(A), \qquad \sigma \in \{-,+\},$$

and for $B \subset S$ a measurable set,

$$\Phi_\sigma(B) = \mu_\sigma\{x \in \mathbb{R}^d_+ : R(x) > 1, \Theta(x) \in B\}, \qquad \sigma \in \{-,+\}.$$

Under the hypothesis above, X's marginal distribution, given by $F = pF_+ + (1-p)F_-$, where $p = \mathbb{P}\{Y = +1\} > 0$, is heavy-tailed as well with index 1. Indeed, we have:

$$t\mathbb{P}\left\{t^{-1}X \in A\right\} \xrightarrow[t\to\infty]{} \mu(A) := p\mu_+(A) + (1-p)\mu_-(A).$$

And similarly

$$\Phi(B) := p\Phi_+(B) + (1-p)\Phi_-(B).$$

Observe also that the limiting class balance can be expressed using the latter asymptotic measures. Indeed, let $\Omega = \{x \in \mathbb{R}^d_+ : \|x\| \leq 1\}$ denote the positive orthant of the unit ball and let Ω^c denote its complementary set in $\mathbb{R}^d_+$. We have:

$$p_t = \mathbb{P}\{Y = +1 \mid \|X\| > t\} = \frac{t\mathbb{P}\{\|X\| > t \mid Y = 1\}\, p}{t\mathbb{P}\{\|X\| > t\}} \xrightarrow[t\to\infty]{} p\frac{\mu_+(\Omega^c)}{\mu(\Omega^c)} = p\frac{\Phi_+(S)}{\Phi(S)} \overset{\text{def}}{=} p_\infty. \tag{5.3}$$

Remark 4. (On Assumption 2) *We point out that only the situation where the supposedly heavy-tailed class distributions F_+ and F_- have the same tail index is of interest. Suppose for instance that the tail index α_+ of F_+ is strictly larger than that of F_-, α_-, that is F_- has heavier tail than F_+. In such a case F is still regularly varying with index $\min\{\alpha_+, \alpha_-\}$ and $p_t \to 0$. In this case, one may straightforwardly see that the classifier predicting always -1 on $\{x \in \mathbb{R}^d_+ : \|x\| > t\}$ is optimal as t increases to infinity.*

Remark 5. (On Assumption 2 (bis)) *As noticed in Remark 3, assuming that $\alpha = 1$ is not restrictive when the marginal distributions are known. In practice however, they must be estimated. In the present analysis, we shall neglect the estimation error arising from their estimation. Relaxing this assumption, as made in e.g.* Goix *and collab. (2017), will be the subject of Section 5.3.1.*

Asymptotic criterion for classification in the extremes. The goal pursued is to construct a classifier g_n, based on the training examples $\mathcal{D}_n$, minimizing the *asymptotic risk in the extremes* given by

$$L_\infty(g) = \limsup_{t\to\infty} L_t(g). \tag{5.4}$$

We also set $L_\infty^* = \inf_{g \text{ measurable}} L_\infty(g)$. It is immediate that any classifier which coincides with the Bayes classifier g^* on the region $t\Omega^c = \{x \in \mathbb{R}_+^d : \|x\| > t\}$ is optimal w.r.t. the distribution P_t. In particular g^* minimizes L_t and the associated risk is

$$L_t^* := L_t(g^*) = \mathbb{E}\left[\min\{\eta(X), 1-\eta(X)\} \mid \|X\| > t\right], \qquad t > 0. \tag{5.5}$$

Thus, for all classifier g, $L_t(g) \geq L_t(g^*)$, and taking the limit superior shows that g^* minimizes L_∞, that is $L_\infty^* = L_\infty(g^*)$.

Optimality. The objective formulated above can be connected with a standard binary classification problem, related to a random pair (X_∞, Y_∞) taking its values in the limit space $\Omega^c \times \{-1, +1\}$, see Theorem 6 below. Let $\mathbb{P}\{Y_\infty = +1\} = p_\infty$ as in (5.3) and define the distribution of X_∞ given that $Y_\infty = \sigma 1$, $\sigma \in \{-, +\}$ as $\mu_\sigma(\Omega^c)^{-1}\mu_\sigma(\,\cdot\,)$. Then for $A \subset \Omega^c$, using (5.3),

$$\begin{aligned}\mathbb{P}\{X_\infty \in A, Y_\infty = +1\} &= \frac{p_\infty \mu_+(A)}{\mu_+(\Omega^c)} = \frac{p\mu_+(A)}{\mu(\Omega^c)} = \frac{p \lim_t t\mathbb{P}\{X \in tA \mid Y = +1\}}{\lim_t t\mathbb{P}\{X \in t\Omega^c\}} \\ &= \lim_{t\to\infty} \mathbb{P}\{X \in tA, Y = +1 \mid \|X\| > t\}.\end{aligned}$$

We denote by P_∞ the joint distribution of (X_∞, Y_∞) thus defined. As shall be seen below, under appropriate and natural assumptions, classifiers with minimum asymptotic risk in the extremes are in 1-to-1 correspondence with solutions of the binary classification problem related to (X_∞, Y_∞). Let ρ be a common dominating measure for Φ_-, Φ_+ on S (ρ does not need to be the Lebesgue measure, take *e.g.* $\rho = \Phi_+ + \Phi_-$). Then denote by φ_+, φ_- respectively the densities of Φ_+, Φ_- w.r.t. ρ. By homogeneity of μ_+, μ_-, the conditional distribution of Y_∞ given $X_\infty = x$ is

$$\begin{aligned}\eta_\infty(x) &\stackrel{\text{def}}{=} \mathbb{P}\{Y_\infty = 1 \mid X_\infty = x\} \\ &= \frac{p_\infty \varphi_+(\Theta(x))/\Phi_+(S)}{p_\infty \varphi_+(\Theta(x))/\Phi_+(S) + (1-p_\infty)\varphi_-(\Theta(x))/\Phi_-(S)} \\ &= \frac{p\varphi_+(\Theta(x))}{p\varphi_+(\Theta(x)) + (1-p)\varphi_-(\Theta(x))}.\end{aligned}$$

Notice that η_∞ is independent of the chosen reference measure ρ and that η_∞ is constant along rays, that is $\eta_\infty(tx) = \eta_\infty(x)$ for (t, x) such that $\min(\|tx\|, \|x\|) \geq 1$.

The optimal classifier for the random pair (X_∞, Y_∞) with respect to the classical risk L_{P_∞} is clearly

$$g^*_\infty(x) = 2\mathbb{1}\{\eta_\infty(x) \geq 1/2\} - 1.$$

Again g^*_∞ is constant along rays on Ω^c and is thus a function of $\Theta(x)$ only. We abusively denote $\eta_\infty(x) = \eta_\infty(\Theta(x))$. The minimum classification error is

$$L^*_{P_\infty} = L_{P_\infty}(g^*_\infty) = \mathbb{E}\left[\min\left\{\eta_\infty(\Theta_\infty), 1 - \eta_\infty(\Theta_\infty)\right\}\right], \tag{5.6}$$

where $\Theta_\infty = \Theta(X_\infty)$. More generally, observe that any class $\mathcal{G}_S$ of classifiers $g : \theta \in S \mapsto g(\theta) \in \{-1, +1\}$ defines a class of classifiers on $\mathbb{R}^d_+$, $x \in \mathbb{R}^d_+ \mapsto g(\Theta(x))$, that shall still be denoted by $\mathcal{G}_S$ for simplicity. The next result claims that, under the regularity hypothesis stated below, the classifier g^*_∞ is optimal for the asymptotic risk in the extremes, that is $L_\infty(g^*_\infty) = \inf_g L_\infty(g)$. We shall also prove that $L_\infty(g^*_\infty) = L^*_{P_\infty}$.

Assumption 3. (Uniform convergence on the sphere of $\eta(tx)$) *The limiting regression function η_∞ is continuous on S and*

$$\sup_{\theta \in S} |\eta(\Theta(t\theta)) - \eta_\infty(\theta)| \xrightarrow[t\to\infty]{} 0$$

Remark 6. (On Assumption 3) *By invariance of η_∞ along rays, Assumption 3 is equivalent to*

$$\sup_{\{x \in \mathbb{R}^d_+ : \|x\| \geq t\}} |\eta(x) - \eta_\infty(x)| \xrightarrow[t\to\infty]{} 0.$$

Assumption 3 is satisfied whenever the probability densities f_+, f_- of F_+, F_- are continuous, regularly varying with limit functions q_+, q_-, and when the convergence is uniform, that is if

$$\lim_{t\to\infty} \sup_{x \in S} |t^{d+1} f_\sigma(tx) - q_\sigma(x)| = 0, \qquad \sigma \in \{+, -\}. \tag{5.7}$$

In such a case q_+, q_- are respectively the densities of μ_+, μ_- with respect to the Lebesgue measure and are continuous, which implies the continuity of φ_+, φ_-. The latter uniform convergence assumption is introduced in De Haan *and* Resnick *(1987) and is used* e.g. *in* Cai *and collab. (2011) in the context of minimum level sets estimation.*

Theorem 6. (Optimal classifiers in the extremes) *Under Assumptions 2 and 3,*

$$L^*_t \xrightarrow[t\to\infty]{} L^*_{P_\infty}. \tag{5.8}$$

Hence, we have: $L^*_\infty = L^*_{P_\infty}$. In addition, the classifier g^*_∞ minimizes the asymptotic risk in the extremes:

$$\inf_{g \text{ measurable}} L_\infty(g) = L_\infty(g^*_\infty) = \mathbb{E}\left\{\min(\eta_\infty(\Theta_\infty), 1 - \eta_\infty(\Theta_\infty))\right\}.$$

Refer to the Supplementary Material for the technical proof. Theorem 6 gives us the form of the optimal classifier in the extremes $g^*_\infty(x) = g^*_\infty(\Theta(x))$, which depends only on the angular component $\Theta(x)$, not the norm $R(x)$. This naturally leads to applying the ERM principle to a collection of classifiers of the form $g(x) = g(\Theta(x))$ on the domain $\{x \in \mathbb{R}^d_+ : \|x\| > t\}$ for $t > 0$ large enough. The next section provides statistical guarantees for this approach.

5.3 Empirical Risk Minimization in the Extremes

Consider a class $\mathcal{G}_S$ of classifiers $g : \theta \in S \mapsto g(\theta) \in \{-1, +1\}$ on the sphere S. It also defines a collection of classifiers on $\mathbb{R}^d_+$, namely $\{g(\Theta(x)) : \ g \in \mathcal{G}_S\}$, which we denote by $\mathcal{G}_S$ for simplicity. Sorting the training observations by decreasing order of magnitude, we introduce the order statistics $\|X_{(1)}\| > \ldots > \|X_{(n)}\|$ and we denote by $Y_{(i)}$ the corresponding sorted labels. Fix a small fraction $\tau > 0$ of extreme observations, and let t_τ be the quantile at level $(1-\tau)$ of the r.v. $\|X\|$: $\mathbb{P}\{\|X\| > t_\tau\} = \tau$. Set $k = \lfloor n\tau \rfloor$ and consider the empirical risk

$$\widehat{L}_k(g) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{Y_{(i)} \neq g(\Theta(X_{(i)}))\} = L_{\widehat{P}_k}(g), \tag{5.9}$$

where $\widehat{P}_k$ denotes the empirical distribution of the truncated training sample $\{(X_i, Y_i) : \ \|X_i\| \geq \|X_k\|, \ i \in \{1, \ \ldots, \ n\}\}$, the statistical version of the conditional distribution P_{t_τ}. We now investigate the performance in terms of asymptotic risk in the extremes L_∞ of the solutions of the minimization problem

$$\min_{g \in \mathcal{G}_S} \widehat{L}_k(g). \tag{5.10}$$

The practical issue of designing efficient algorithms for solving (5.10) is beyond the scope of this chapter. Focus is here on the study of the learning principle that consists in assigning to any very large input value x the likeliest label based on the direction $\Theta(x)$ it defines only (the construction is summarized in Algorithm 1 below). The following result provides an upper bound for the excess of classification error in the domain $t_\tau \Omega^c$ of solutions of (5.10). Its proof, which relies on a maximal deviation inequality tailored to low probability regions, is given in the Supplementary Material.

Theorem 7. *Suppose that the class $\mathcal{G}_S$ is of finite* VC *dimension $V_{\mathcal{G}_S} < +\infty$. Let $\widehat{g}_k$ be any solution of (5.10). Recall $k = \lfloor n\tau \rfloor$. Then, for $\delta \in (0,1)$, $\forall n \geq 1$, we have with probability larger than $1-\delta$:*

$$L_{t_\tau}(\widehat{g}_k) - L^*_{t_\tau} \leq \frac{1}{\sqrt{k}}\left(\sqrt{2(1-\tau)\log(2/\delta)} + C\sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right)$$
$$+ \frac{1}{k}\left(5 + 2\log(1/\delta) + \sqrt{\log(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right) + \left\{\inf_{g \in \mathcal{G}_S} L_{t_\tau}(g) - L^*_{t_\tau}\right\},$$

where C is a constant independent from n, τ and δ.

Remark 7. (On model selection) *Selecting an appropriate model class $\mathcal{G}_S$ is a crucial issue in machine-learning. Following in the footsteps of structured risk minimization, one may use a* VC *bound for $\mathbb{E}[\sup_{g \in \mathcal{G}_S} |\widehat{L}_k(g) - \mathbb{E}[\widehat{L}_k(g)]|]$ as a complexity regularization term to penalize in an additive fashion the empirical risk (5.9). Considering a collection of such models, oracle inequalities guaranteeing the quasi-optimality of the rule minimizing the penalized empirical risk can be then classically established by means of a slight modification of the argument of Theorem 7's proof, see e.g. Chapter 18 in* Devroye *and collab. (1996).*

The upper bound stated above shows that the learning rate is of order $O_{\mathbb{P}}(1/\sqrt{k})$, where k is the actual size of the training data set used to perform approximate empirical risk minimization in the extremes. As revealed by the corollary below, this strategy permits to build a consistent sequence of classifiers for the L_∞-risk, when the fraction $\tau = \tau_n$ decays at an appropriate rate (provided that the model bias can be neglected of course).

Corollary 1. *Suppose that the assumptions of Theorems 6-7 are fulfilled. In addition, assume that the model bias asymptotically vanishes as $\tau \to 0$, i.e.*

$$\inf_{g \in \mathcal{G}_S} L_{t_\tau}(g) - L^*_{t_\tau} \longrightarrow 0 \quad \text{as } \tau \to 0.$$

Then, as soon as $k \to +\infty$ as $n \to \infty$, the sequence of classifiers $(\widehat{g}_k)$ is consistent in the extremes, meaning that we have the convergence in probability:

$$L_\infty\left(\widehat{g}_k\right) \to L^*_\infty \text{ as } n \to \infty.$$

Algorithm 1 (ERM in the extremes)**.**

Input *Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, collection $\mathcal{G}_S$ of classifiers on the sphere, size $k \leq n$ of the training set composed of extreme observations*

1 **Standardization.** *Standardize the input vector by applying the rank-transformation: $\forall i \in \{1, \ldots, n\}$, $\hat{V}_i = \hat{T}(X_i)$, where*

$$\hat{T}(x) = \left(1/\left(1 - \hat{F}_j(x_j)\right)\right)_{j=1,\ldots,d},$$

for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

2 **Truncation.** *Sort the training input observations by decreasing order of magnitude*

$$\|\hat{V}_{(1)}\| \geq \ldots \geq \|\hat{V}_{(n)}\|,$$

and consider the set of extreme training points

$$\left\{(\hat{V}_{(1)}, Y_{(1)}), \ldots, (\hat{V}_{(k)}, Y_{(k)})\right\}.$$

3 **Optimization.** *Compute a solution $\widehat{g}_k(\theta)$ of the minimization problem*

$$\min_{g \in \mathcal{G}_S} \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\left\{Y_{(i)} \neq g\left(\Theta(\hat{V}_{(i)})\right)\right\}$$

Output *The classifier $\widehat{g}_k\left(\Theta\left(\hat{T}(x)\right)\right)$, applicable on the region $\{x : \|\hat{T}(x)\| > \|\hat{V}_{(k)}\|\}$.*

Remark 8 (Choice of k). *Determining the best value of k is a typical challenge of Extreme Value analysis. This is typically a bias/variance trade-off, too large values introduce a bias by taking into account observations which are not large enough, so that their distribution deviates significantly from the limit distribution of extremes. On the other hand, too small values obviously increase the variance of the classifier. See* e.g. Goix *and collab. (2016) or* Goix *and collab. (2017) and the reference therein for a discussion. In practice a possible default choice is $k = \sqrt{n}$, otherwise cross-validation can be performed.*

As a first go, as mentioned in Remark 2 the marginal distribution are assumed to be known for the sake of simplicity. Now, relying on the statistical bounds provided in Chapter 4, we will extend Theorem 7 to generalize the framework to cases where the marginal distributions are unknown and standardization is needed.

5.3.1 Influenze of the Marginal Standardization on Classification in Extreme Regions

Now that the framework of classification in extreme regions is set. The aim of this section is to extend these theoretical results to workcases where the assumption resulting from Remark 2 does not apply *i.e.* to the case where the margins are unknown and standardization *via* rank-transformation is necessary. In this way, we apply the two main results from Chapter 4, theorems 2 and 3, to the problem of binary classification in extreme regions studied above.

Classification framework. We first briefly recall the set-up of Section 5.2-5.3 for the sake of completeness and to adapt notations from Chapter 4 to our setting. In a nutshell, we previously considered (V, Y) a random pair where $Y \in \{0, 1\}$ is the label to be predicted and $V \in \mathbb{R}^d$ is the vector of predictors (features). where the conditional distribution of V given $Y = \sigma 1$, with $\sigma \in \{+, -\}$ is regularly varying with index 1 and normalizing function $b(t) = t^{-1}$ and has limit measure μ_σ: for $B \subset [0, \infty)^d \setminus \{0\}$ a measurable set such that $0 \notin \partial A$ and $\mu(\partial B) = 0$,

$$t\mathbb{P}\left\{t^{-1}V \in B \,\middle|\, Y = \sigma\, 1\right\} \xrightarrow[t\to\infty]{} \mu_\sigma(B), \qquad \sigma \in \{-, +\}. \tag{5.11}$$

The angular measure related to $\mu_\pm$ is denoted by $\Phi_\pm$, so that for $A \subset \mathbb{S}$ a measurable set, $\Phi_\pm(A) = \mu_\pm\{x \in \mathbb{R}^d_+ : \|x\| > 1, \theta(x) \in B\}$. Under this assumption, the marginal distribution of V are also regularly varying with limit measure $\mu = p\mu_+ + (1-p)\mu_-$ and angular measure $\Phi = p\Phi_+ + (1-p)\Phi_-$.

Now, we assume that the observed random pair is (X, Y) and that (5.11) holds with $V = v(X)$, where v is defined via the probability integral transform introduced in Chapter 4 Section 4.2, with F_j the j^{th} marginal distribution of X. We emphasize that F_j involved in our definition of V is not conditioned upon Y. Instead, $F_j = pF_{j+} + (1-p)F_{j-}$, with $p = \mathbb{P}\{Y = 1\}$ and $F_{j\pm}(x) = \mathbb{P}\{X_j \le x \mid Y = \pm 1\}$. In this framework, all the results obtained above concerning the probabilistic properties of (V, Y) apply, however in the remaining of this chapter, V is not observed, only the rank transformed version $\widehat{V}$ is.

We previously introduced the conditional classification risk above level t of a classifier $g : \mathbb{R}^d_+ \setminus \{0\} \to \{\pm 1\}$ defined on the standardized input V,

$$L_t^{\mathrm{cond}}(g) := L_{P_t}(g) = \mathbb{P}\{Y \neq g(V) \mid \|V\| > t\}, \tag{5.12}$$

and its asymptotic version $L_\infty^{\mathrm{cond}}(g) = \limsup_{t\to\infty} L_t^{\mathrm{cond}}(g)$. It is shown that the Bayes risk above level t, $\min_g L_t^{\mathrm{cond}}(g)$ (where the minimum is taken over all measurable classifiers) converges to a quantity $L_\infty^{\mathrm{cond}*}$. The latter may be interpreted as the Bayes risk related to a random pair (V_∞, Y_∞) which distribution is the limit of the joint distribution conditional on V being large, $P_t(\,\cdot\,) = \mathbb{P}\{t^{-1}V \in \cdot, Y \in \cdot \mid \|V\| > t\}$.

Relaxed framework in the non-standard case. In the present context (V is not observed and the standardization function v is unknown), we need to introduce a standardization-dependent version of the risk above level t. For $T : \mathbb{R}_d \to \mathbb{R}^d_+$ a standardization function (typically, $T = v$ or $T = \hat{v}$) and g an angular classifier ($g(v) = g(\theta(v))$), define

$$L_t(g, T) = t\mathbb{P}\left\{g(T(X)) \neq Y, \|T(X)\| \geq t\right\}. \tag{5.13}$$

Thus, $L_t(g, v) = t\mathbb{P}\left\{\|V\| \geq t\right\} L_t^{\mathrm{cond}}(g)$. Here the multiplicative factor $t\mathbb{P}\left\{\|V\| \geq t\right\}$ converges to $\Phi(\mathbb{S})$ and does not change the minimizer in the class $\mathcal{G}$. It is introduced for mathematical convenience, in order to avoid the division by a random quantity in the definition of the empirical version of L_t. Also it follows from the definitions that $L_t(g, v)$ converges to $L_\infty(g) := \Phi(\mathbb{S}) L_\infty^{\mathrm{cond}}(g)$ as $t \to \infty$. Since the error bound on the empirical measure obtained in Chapter 4 Sections 4.3 and 4.3.3 are only valid for angular sets A which are at distance τ from the boundary of the positive orthant (see Assumption 1 in Chapter 4), the empirical counterpart of L_t can only be valid if the transformed points $T(X)$ which angle is too small are excluded from the analysis. With this in mind, consider the subset $\mathbb{S}_\tau$ of $\mathbb{S}$ consisting of angles whose minimum coordinate is not less than τ,

$$\mathbb{S}_\tau = \{\theta \in \mathbb{S} : \forall j \in \{1, \ldots, d\}, \theta_j > \tau\},$$

Given integers $1 < k \leq n$ define the empirical risk $\widehat{L}^\tau(g, T)$ associated with a classifier g and a transformation T, restricted to those points which angle (after transformation) belongs to $\mathbb{S}_\tau$,

$$\widehat{L}^\tau(g, T) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\{g(T(X_i)) \neq Y, \theta(T(X_i)) \in \mathbb{S}_\tau, \|T(X_i)\| \geq n/k\}, \tag{5.14}$$

and denote for brevity $\widehat{L}^\tau(g) = \widehat{L}^\tau(g, \hat{v})$. Notice the difference between the above display and the definition of the empirical risk in previous sections involving the k largest transformed instances instead of all those which norm exceeds n/k. The present definition slightly simplifies the notations in our proofs but does not change the order of magnitude of the number of data retained for ERM, indeed with the choice $T = \hat{v}$ the latter number varies between k and dk. Even though the dimension d could be large, it is considered as fixed in the present analysis.

Given a class $\mathcal{G}$ of angular classifiers depending only on the angle θ as above, the ERM strategy that we promote in the present context consists in selecting

$$\hat{g}_k^\tau \in \arg\min_{g \in \mathcal{G}} \widehat{L}^\tau(g).$$

The empirical risk $\widehat{L}^{\tau}(g)$ serves as a proxy for the limit risk restricted to $\mathbb{S}_\tau$,

$$L_\infty^{>\tau}(g) = L_\infty^{>\tau}(g, v) = \lim_{t\to\infty} t\mathbb{P}\left\{g(v(X)) \neq Y, \theta(v(X)) \in \mathbb{S}_\tau, \|v(X)\| > t\right\}. \quad (5.15)$$

That the above limit exists for an angular classifier is a consequence of Lemma 3 below under the additional assumption that $\partial\mathbb{S}_\tau$ and the decision boundaries of g are Φ-null sets, see Assumption 4 below.

The definition of $L_\infty^{>\tau}$ derives naturally from our setting, in that $L_\infty^{>\tau}$ only takes into account those observations which can legitimately enter the statistical analysis as their angular coordinates are uniformly bounded from 0. It will be shown (see Remark 9) that the difference $L_\infty^{>\tau} - L_\infty$ can be uniformly bounded by a multiple of τ under regularity assumptions, namely the existence of a uniformly bounded density for Φ.

Classification risk and angular measure. Consider the following decomposition of the finite distance risk L_t related to $\mathbb{S}_\tau$,

$$\begin{aligned} L_t(g,T) &= L_t^{>\tau}(g,T) + L_t^{<\tau}(g,T) \qquad \text{with} \\ L_t^{>\tau}(g,T) &= t\mathbb{P}\left\{g(T(X)) \neq Y, \theta(T(X)) \in \mathbb{S}_\tau, \|T(X)\| \geq t\right\}\ ; \\ L_t^{<\tau}(g,T) &= t\mathbb{P}\left\{g(T(X)) \neq Y, \theta(T(X)) \notin \mathbb{S}_\tau, \|T(X)\| \geq t\right\} \end{aligned} \quad (5.16)$$

In the particular case where $T = v$ is the transformation to Pareto margins, the following result justifies the definition of $L_\infty^{>\tau}$ (5.15) and provides a useful insight regarding its relation with the angular measures of the two classes. In the sequel, let $\mathbb{S}_g^{\pm}$ denote the two regions of the sphere $\mathbb{S}$ which are respectively labeled by g as $+1$ (*resp.* -1). Recall that $p = \mathbb{P}\{Y = +1\}$. We work hereaftr under the follwing smoothness assumption.

Assumption 4 (smoothness)**.** *τ is chosen in such a way that $\Phi(\partial\mathbb{S}_\tau) = 0$ and the class $\mathcal{G}$ is such that $\Phi(\partial\mathbb{S}_g^+) = \Phi(\partial\mathbb{S}_g^-) = 0$.*

Lemma 3. *If the conditional regular variation property (5.11) and Assumption 4 hold, then for all angular classifier g,*

$$L_t^{>\tau}(g, v) \xrightarrow[t\to\infty]{} L_\infty^{>\tau}(g) := p\Phi^+(\mathbb{S}_g^- \cap \mathbb{S}_\tau) + (1-p)\Phi^-(\mathbb{S}_g^+ \cap \mathbb{S}_\tau).$$

Also,

$$L_t^{<\tau}(g, v) \xrightarrow[t\to\infty]{} L_\infty^{<\tau}(g) := p\Phi^+(\mathbb{S}_g^- \setminus \mathbb{S}_\tau) + (1-p)\Phi^-(\mathbb{S}_g^+ \setminus \mathbb{S}_\tau)$$

Finally,

$$L_t(g, v) \xrightarrow[t\to\infty]{} L_\infty(g) := p\Phi^+(\mathbb{S}_g) + (1-p)\Phi^-(\mathbb{S}_g^+)$$

The proof is deferred to Section 4.6.

Remark 9. *Notice that $L_t^{<\tau}(g, v) \leq t\mathbb{P}\{\theta(V) \notin \mathbb{S}_\tau, \|V\| > t\} \to \Phi(\mathbb{S} \setminus \mathbb{S}_\tau)$. In particular $\sup_g L_\infty^{<\tau}(g) \leq \Phi(\mathbb{S} \setminus \mathbb{S}_\tau)$. If Φ is concentrated on the interior of $\mathbb{S}$ and has bounded density ϕ, we have $\Phi(\mathbb{S} \setminus \mathbb{S}_\tau) = O(\|\phi\|_\infty \tau)$, namely, working with the infinity norm, $\Phi(\mathbb{S} \setminus \mathbb{S}_\tau) \leq \|\phi\|_\infty \tau d$.*

Decomposition of the excess risk. Our main purpose is to obtain an upper bound on the supremum deviation of the empirical risk $\sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g,\hat{v}) - L_\infty(g,v)\right|$. Indeed, assume the existence of $g_\infty^{\mathcal{G}}$ a minimizer of L_∞ over $\mathcal{G}$ (if such a minimizer does not exist, consider a sequence of $1/N$-minimizers and proceed). Denote by g_∞^* the optimal classifier for L_∞ over all possible classifiers. The existence of g_∞^* and its writing as a function of the regression function η_∞ is proved in Theorem 6. Recall that $\hat{g}_k^\tau$ denotes a minimizer of $\widehat{L}^\tau(g,\hat{v})$ over $\mathcal{G}$. The excess risk decomposes classically as follows:

$$\begin{aligned} L_\infty(\hat{g}_k^\tau) - L_\infty(g_\infty^*) &\leq L_\infty(\hat{g}_k^\tau) - \widehat{L}^\tau(\hat{g}_k^\tau) + \underbrace{\widehat{L}^\tau(\hat{g}_k^\tau) - \widehat{L}^\tau(g_\infty^{\mathcal{G}})}_{\leq 0} \\ &\quad \cdots + \widehat{L}^\tau(g_\infty^{\mathcal{G}}) - L_\infty(g_\infty^{\mathcal{G}}) + L_\infty(g_\infty^{\mathcal{G}}) - L_\infty(g_\infty^*) \\ &\leq 2\sup_{g\in\mathcal{G}} |\widehat{L}^\tau(g) - L_\infty(g)| + \mathrm{bias}(\mathcal{G}) \end{aligned} \tag{5.17}$$

where $\mathrm{bias}(\mathcal{G}) = \inf_{g\in\mathcal{G}} L_\infty(g) - L_\infty(g_\infty^*)$ depends on how close to the class $\mathcal{G}$ is the Bayes classifier g_∞^*. In our context, the supremum deviation itself decomposes further, since for all $g \in \mathcal{G}$

$$\begin{aligned} |\widehat{L}^\tau(g) - L_\infty(g)| &= |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g) - L_\infty^{<\tau}(g)| \\ &\leq |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)| + L_\infty^{<\tau}(g) \\ &\leq |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)| + \Phi(\mathbb{S}\setminus\mathbb{S}_\tau) \end{aligned} \tag{5.18}$$

Remark 10 (Choice of τ for classification)**.** *In view of Remark 9, the term $\Phi(\mathbb{S}\setminus\mathbb{S}_\tau)$ may be viewed as an additional bias term which vanishes as $\tau \to 0$. On the other hand, the main result of this section (Theorem 8) shows that the upper bound on $\sup_g |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)|$ grows roughly as $1/\sqrt{\tau}$ as $\tau \to 0$. The choice of τ thus constitutes an additional bias-variance compromise.*

The next result parallels Lemma 3 by relating the empirical risk with the empirical angular measure of the positive and negative classes. In view of the definition for the positive and negative angular measures, define for $A \subset \mathbb{S}$ and $\sigma \in \{-,+\}$

$$\widehat{\Phi}^\sigma(A) = \frac{1}{k^\sigma}\sum_{i=1}^n \mathbb{1}\{Y_i = \sigma 1\}\mathbb{1}\{\theta(\hat{V}_i) \in A,\ \|\hat{V}_i\| \geq n/k\} \tag{5.19}$$

where $k^\sigma = kn^\sigma/n$ and n^σ is the number of pairs such that $Y_i = \sigma 1$, $n^\sigma = \sum_{i\leq n} \mathbb{1}\{Y_i = \sigma 1\}$.

Consider now the type-I and type-II empirical errors, *i.e.* for $\sigma \in \{-,+\}$,

$$\widehat{L}_\sigma^\tau(g,T) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{g(T(X_i)) \neq Y_i, Y_i = \sigma 1, \theta(T(X_i)) \in \mathbb{S}_\tau, \|T(X_i)\| \geq n/k\}.$$

Notice that $\widehat{L}_\sigma^\tau$ can be written as

$$\widehat{L}_+^\tau(g,\hat{v}) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{\theta(\hat{V}_i) \in \mathbb{S}_g^- \cap \mathbb{S}_\tau,\ Y_i = +1\ , \|\hat{V}_i\| \geq n/k\}$$

a similar treatment of $\widehat{L}_+^\tau(g,\hat{v})$ yields immediately:

Lemma 4. *In the case where T is the rank transformation $\hat{v}$, the empirical type-I and type-II errors write as*

$$\widehat{L}^{\tau}_{+}(g,\hat{v}) = \frac{k^{+}}{k}\hat{\Phi}^{+}(\mathbb{S}_g^{-} \cap \mathbb{S}_\tau)$$
$$\widehat{L}^{\tau}_{-}(g,\hat{v}) = \frac{k^{-}}{k}\hat{\Phi}^{-}(\mathbb{S}_g^{+} \cap \mathbb{S}_\tau)$$

In view of the error decomposition (5.17),(5.18), we state our main result in terms of the maximum deviations $\sup_{g\in\mathcal{G}} |\widehat{L}^{\tau}(g,\hat{v}) - L^{\geq\tau}_{\infty}(g)|$, leveraging the results from Section 4.3.

Theorem 8 (Deviations of the empirical tail risk)**.** *Consider the class of sets $\mathcal{A} = \{\mathbb{S}_g^{+} \cap \mathbb{S}_\tau, g \in \mathcal{G}\} \cup \{\mathbb{S}_g^{-} \cap \mathbb{S}_\tau, g \in \mathcal{G}\}$. Under the conditional regular variation assumption (5.11) and the smoothness assumption 4, and if the assumptions of Theorem 2,* i.e. *assumptions 1 relative to the class $\mathcal{A}$ and the marginal angular measure Φ are satisfied, with probability at least $1-(d+2)\delta$:*

$$\sup_{g\in\mathcal{G}} |\widehat{L}^{\tau}(g,\hat{v}) - L^{\geq\tau}_{\infty}(g)| \leq 2(\text{error} + \text{bias II} + \text{gap})$$

where error *and* gap *are the same as in Theorem 2,* i.e.

$$\text{error} = C\left(\sqrt{d(1+\Delta)V_{\mathcal{F}} k^{-1}\log(1/\delta)} + k^{-1}\log(1/\delta)\right),$$
$$\text{gap} = \left(2d + 3c + 3c\ln(\tfrac{d}{3c})\right)\Delta - 3c\Delta\ln(\Delta),$$

with C the constant appearing in Theorem 5, Δ as defined in Assumption 1.

$$\text{bias II} = \sup\Big\{ \Big|\tfrac{n}{k}\mathbb{P}\left\{V \in \tfrac{n}{k}B, Y = \sigma 1\right\} - \mathbb{P}\left\{Y = \sigma 1\right\}\mu^{\sigma}(B)\Big| : B = \Gamma^{+}_{A,1} \text{ or } B = \Gamma^{-}_{A,1} \text{ for some } A \in \mathcal{A}, \sigma \in \{-,+\}\Big\}$$

The proof relies on the relationships between the (empirical) classification risks and the (empirical) angular measure pointed out in Lemmata 3, 4, which imply in particular that for $g \in \mathcal{G}$,

$$|\widehat{L}^{\tau}(g,\hat{v}) - L^{\geq\tau}_{\infty}(g)| \leq \left|\frac{k^{+}}{k}\hat{\Phi}^{+}(\mathbb{S}_g^{-} \cap \mathbb{S}_\tau) - p\Phi^{+}(\mathbb{S}_g^{-} \cap \mathbb{S}_\tau)\right| + \ldots$$
$$\left|\frac{k^{-}}{k}\hat{\Phi}^{-}(\mathbb{S}_g^{+} \cap \mathbb{S}_\tau) - (1-p)\Phi^{-}(\mathbb{S}_g^{+} \cap \mathbb{S}_\tau)\right|$$

The right-hand-side of the latter display is then uniformly upper bounded by adapting the arguments of the proof of Theorem 2, see Section 4.6 for details.

Truncated empirical risk and associated estimator. Consider now a truncated version of the risk where the observations X such that $\|T(X)\| > Mn/k$ are not taken into account, where M is the truncation level defined in Section 4.3.3:

$$\widehat{L}^{\tau,M}(g,T) = \frac{M}{k(M-1)} \sum_{i=1}^{n} \mathbb{1}\{g(T(X_i)) \neq Y_i, \theta(T(X)) \in \mathbb{S}_\tau, \dots \\ \dots n/k \leq \|T(X_i)\| \leq Mn/k\}. \tag{5.20}$$

Denote by $\hat{g}_k^{\tau,M}$ the minimizer of $\widehat{L}^{\tau,M}(g) := \widehat{L}^{\tau,M}(g,\hat{v})$ among the family of angular classifiers $\mathcal{G}$. Following the lines of the previous paragraph it can easily be proven, relying on the arguments from the proof of Therorem 3 relative to the truncated empirical angular measure, that

$$\sup_{g\in\mathcal{G}} |\widehat{L}^{\tau,M}(g,\hat{v}) - L_\infty^{\geq\tau}(g,v)| \leq 2(\mathrm{error}_{\mathrm{M}}\,\mathrm{II} + \mathrm{bias}_{\mathrm{M}}\,\mathrm{II} + \mathrm{gap}_{\mathrm{M}}\,\mathrm{II}),$$

where $\mathrm{error}_{\mathrm{M}}\,\mathrm{II}, \mathrm{bias}_{\mathrm{M}}\,\mathrm{II}, \mathrm{gap}_{\mathrm{M}}\,\mathrm{II}$ have an expression similar (if not identical) to $\mathrm{error}_M, \mathrm{bias}_M, \mathrm{gap}_M$ in Theorem 3. For the sake of concision we do note derive the precise expression for the upper bound and leave the details to the interested reader. Again this concentration bound implies another one on the excess risk $L_\infty(\hat{g}_k^{\tau,M}) - L_\infty(g^*)$. We verify experimentally in Section 5.4 that the performances of $\hat{g}_k^{\tau}$ and $\hat{g}_k^{\tau,M}$ are comparable.

5.4 Illustrative Numerical Experiments

5.4.1 On the importance of a dedicated classifier in extreme regions

The purpose of our experiments is to provide insights into the performance of the classifier $\widehat{g}_k$ on extreme regions constructed *via* Algorithm 1.The training set is ordered as in Step 1 of Algorithm 1. For a chosen k, let $t = \|\hat{T}(X^{\mathrm{train}}_{(k)})\|$, the L_1 norm is used throughout our experiments. The extreme test set $\mathcal{T}$ is the subset of test points such that $\|\hat{T}(X_i^{\mathrm{test}})\| > t$. To approximate of the asymptotic risk in the extremes $L_\infty(\widehat{g}_k)$ and illustrate the generalization ability of the proposed classifier in the extreme region, we consider decreasing subsets of $\mathcal{T}$. Namely denoting $n_{\mathrm{test}} = |\mathcal{T}|$, we keep only the $\lfloor \kappa n_{\mathrm{test}} \rfloor$ largest instances of $\mathcal{T}$ in terms on $\|\hat{T}(X_i^{\mathrm{test}})\|$, for decreasing values of $\kappa \in (0,1]$. This experimental framework is summarized in Figure 5.1, where $\lambda t = \|\hat{T}(X^{\mathrm{test}}_{(\lfloor \kappa n_{\mathrm{test}} \rfloor)})\| \geq t$.

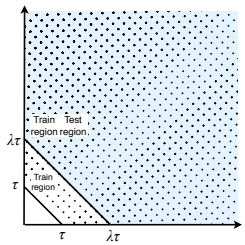


Figure 5.1 – Train set (dotted area) and test set (colored area).

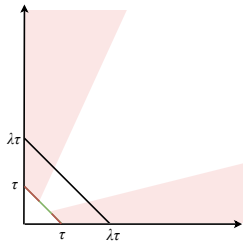


Figure 5.2 – Colored cones correspond to a given label from the classifier on the simplex.

We consider two different classification algorithms for Step 3 in Algorithm 1, namely random forest (RF) and k-nearest neighbors (k-NN), which correspond

to two different classes $\mathcal{G}_S$ of classifiers. For each class $\mathcal{G}_S$, the performance of $\widehat{g}_k$ (which considers only the direction $\Theta(\hat{T}(x))$ of both training and testing data, in other words classifies the projected datasets onto the unit sphere (see Figure 5.2) is compared with that of the classical version of the algorithm (RF or k-NN) taking as input the same training data but without the standardization and truncation steps neither the projection onto the unit sphere. Figures 5.4 and 5.5 summarize the results obtained using RF respectively with a multivariate simulated dataset and with a real world dataset. The simulated dataset is generated from a logistic distribution as described in Stephenson (2003). The positive and negative instances are generated using two different dependency parameters. An example of dataset thus obtained is displayed in Figure 5.3. We report the results obtained with $5 \cdot 10^3$ points for each label for the train set and $5 \cdot 10^4$ points for each label for the test set. $k = 100$ and $\kappa \in [1, 0.3]$. the number of trees for both random forests (in the regular setting and in the setting of Algorithm 1) is set to 200. The number of neighbors for both k-NN's is set to 5.

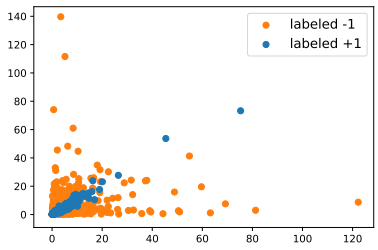


Figure 5.3 – Toy dataset generated from a multivariate logistic distribution projected on $\mathbb{R}^2$.

The real dataset known as Ecoli dataset, introduced in Nakai and Kanehisa (1992), deals with protein localization and contains 336 instances and 8 features. The Supplementary Material gathers additional details concerning the datasets and the tuning of RF and k-NN in our experiments, as well as additional results obtained with the above described datasets and with a simulated dataset from a different distribution.

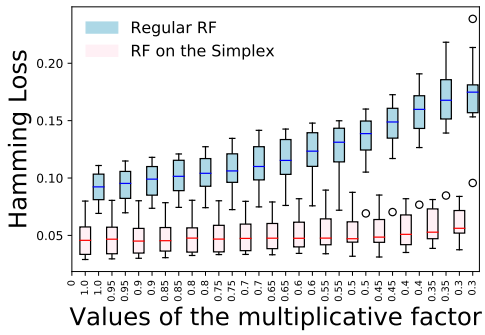


Figure 5.4 – Logistic data - test loss of RF on the simplex and regular RF depending on the multiplicative factor κ.

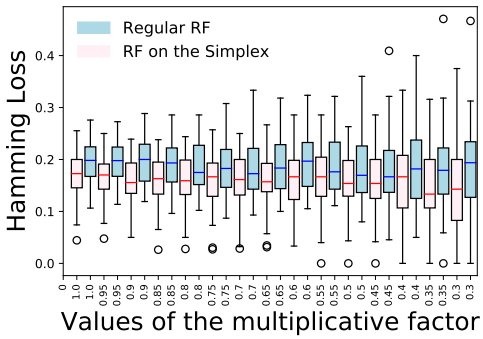


Figure 5.5 – Real data - test loss of RF on the simplex and regular RF depending on the multiplicative factor κ.

Figure 5.4 shows the evolutions of the Hamming losses with decreasing values of $\kappa \in [0.3, 1]$. The boxplots display the losses obtained with 10 independently simulated datasets. For the experiment on the Ecoli dataset (Figure 5.5), one third of the dataset is used as a test set and the rest corresponds to the train set. $k = 100$ and $\kappa \in [0.3, 1]$ (considering smaller values of κ was prevented by data scarcity). The boxplots display the results for different (random) partitions of the data into a train and a test set. In both examples, the loss of the regular classifier is worse (and even increases) when κ decreases whereas the classifier resulting from the proposed approach is better and has a better extrapolation ability.

5.4.2 Marginal Standardization for Binary Classification in Extreme Regions

We place ourselves in the classification framework described in Section 5.3.1, where two angular classifiers $\widehat{g}^{\tau}$ and $\widehat{g}^{\tau,M}$ are proposed, which are issued respectively from the minimization of the traditional and the truncated empirical classification risks $\widehat{L}^{\tau}$ and $\widehat{L}^{\tau,M}$. The aim of this experiment is to compare the performance of the empirical risk minimizers $\widehat{g}^{\tau}$ and $\widehat{g}^{\tau,M}$ in terms of classification score.

Experimental Setting for Classification.

Our experimental set-up generalizes the setting described in Section 4.5.1 to the binary classification framework. Namely, consider a random pair $(X, Y) \in \mathbb{R}_+^d \times \{-1, +1\}$ and for $\sigma \in \{-, +\}$, let us denote by X_σ a random variable following the conditional law of X given that $Y = \sigma 1$. We generate independent samples $(X_{+,i}, +1)$ and $(X_{-,i}, -1)$, $i \geq 1$, where $X_{\sigma,i} = R_{\sigma,i}\Theta_{\sigma,i}$, $R_{\sigma,i}$ follows a standard Pareto distribution, $\Theta_{\sigma,i}$ follows a Dirichlet distribution on the simplex $\mathbb{S}_{L^1}$ with concentration parameter ν_σ and $R_{\sigma,i}, \Theta_{\sigma,i}$ are independent. At the training step for both classifiers entering the comparison we choose (n, k, τ) such that $n\tau/k > d$.

The empirical risks $\widehat{L}^{\tau}$ and $\widehat{L}^{\tau,M}$, are straightforwardly evaluated using their respective definitions (5.14) and (5.20). Notice that introducing the index sets:

$$\begin{aligned} \mathcal{I}^{\hat{v}} &= \{i \leq n : \widehat{V}_i/\|\widehat{V}_i\| \in \mathbb{S}_\tau, \|\widehat{V}_i\| \geq n/k\}, \\ \mathcal{I}^{\hat{v},M} &= \{i \leq n : \widehat{V}_i/\|\widehat{V}_i\| \in \mathbb{S}_\tau, n/k \leq \|\widehat{V}_i\| \leq nM/k\}, \end{aligned} \tag{5.21}$$

we may write

$$\begin{aligned} \widehat{L}^{\tau}(g) &\propto \sum_{i \in \mathcal{I}^{\hat{v}}} \mathbb{1}\{g(\widehat{V}_i/\|\widehat{V}_i\|) \neq Y_i\}, \\ \widehat{L}^{\tau,M}(g) &\propto \sum_{i \in \mathcal{I}^{\hat{v},M}} \mathbb{1}\{g(\widehat{V}_i/\|\widehat{V}_i\|) \neq Y_i\}. \end{aligned}$$

In the present experiment we consider three families of classifiers, namely the classifiers issued from logistic regression, from a decision tree and from a random forest, respectively denoted by $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$. In order to ensure a finite VC-dimension, the maximum depth of a tree is set to 10, both for the classes $\mathcal{G}_2$ and $\mathcal{G}_3$. Notice that random forest classifiers are not strictly speaking empirical risk minimizers but are obtained by aggregation of such classifiers, so that the theory developed in the present work does not apply as is for $\mathcal{G}_3$. We have chosen to include $\mathcal{G}_3$ in the study anyway in view of the popularity of such classifiers for applied purposes.

In this context the two considered estimators $\{\widehat{g}^{\tau}, \widehat{g}^{\tau,M}\}$ are the output of the considered classification algorithm (logistic regression / classification tree / random forest) taking as input at the training step respectively the sets $\{(\widehat{V}_i/\|\widehat{V}_i\|, Y_i)$ for $i \in \mathcal{I}^{\hat{v}}$ (*resp.* for $i \in \mathcal{I}^{\hat{v},M})$

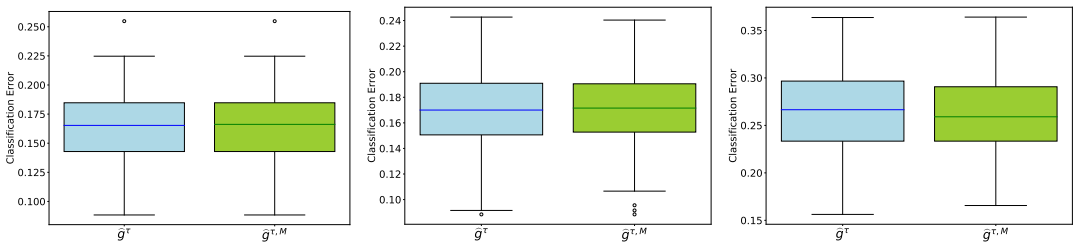
To measure the generalization performance of $\{\widehat{g}^{\tau}, \widehat{g}^{\tau,M}\}$ on tail regions we consider a test set $(X'_i, Y'_i)_{i\leq n}$ generated in the same way as the train set and independent from the latter. We denote by $\widehat{V}'_i = \hat{v}(X'_i)$ the transformed samples using the rank transform $\hat{v}$ issued from the training step. In view of the results from Section 5.3.1 we restrict our attention to test data which angular component belongs to $\mathbb{S}_{\tau}$. Thus we consider the index set:

$$\mathcal{I}^{\tau}_{\text{test}} = \{i \leq n : \widehat{V}'_i/\|\widehat{V}'_i\| \in \mathbb{S}_{\tau}, \|\widehat{V}'_i\| \geq n/k\}\,.$$

The classification test error of a candidate classifier g is thus

$$L^{\tau}_{\text{test}}(g) = \frac{1}{|\mathcal{I}^{\tau}_{\text{test}}|} \sum_{i \in \mathcal{I}^{\tau}_{\text{test}}} \mathbb{1}\{g(\widehat{V}'_i/\|\widehat{V}'_i\|) \neq Y'_i\}\,.$$

Results.
For simplicity we choose n as an even integer and we fix the number of positive and negative instance to $n/2$ in the training and testing set so as to mimic a balanced mixture model with $p = \mathbb{P}\{Y = +1\} = 1/2$. The number of train and test samples is set to $n = 10^5$. The truncation level M is selected so that 5% of the set $\{\|\widehat{V}_i\| : \|\widehat{V}_i\| \geq n/k\}$ are discarded (*i.e.* Mn/k is the 0.95-quantile of the latter set). As in Section 4.5.1, the censoring parameter τ is set to 0.1 and k is set to $\sqrt{n}$. The Dirichlet concentration parameters are chosen as $\nu_+ = 1, \nu_- = 2$. Figure 5.6 displays boxplots of the quantities $L^{\tau}_{\text{test}}(\widehat{g}^{\tau})$ and $L^{\tau}_{\text{test}}(\widehat{g}^{\tau,M})$obtained with 100 independent experiments, for the classes $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$. These results show no significant difference between the performance of $\widehat{g}^{\tau}, \widehat{g}^{\tau,M}$. More precisely, Kolmogorov-Smirnoff tests between the distributions of the two considered classifiers for each class did not allow to reject the null hypothesis of equality between distributions, as the minimum p-values over the three pairs is 0.91. This confirms that the alternative classifier $\widehat{g}^{\tau,M}$ is a reasonable alternative to $\widehat{g}^{\tau}$.



(a) (b) (c)

Figure 5.6 – Classification test error of $\widehat{g}^{\tau}$ and $\widehat{g}^{\tau,M}$ issued from logistic regression (5.6a), random forests (5.6b) and classification trees (5.6c).

5.5 Conclusion

In various applications (*e.g.* safety/security, finance, insurance, environmental sciences), it is of prime importance to predict the response Y of a system when it is impacted by shocks, corresponding to extremely large input values X. In this chapter, we have developed a rigorous probabilistic framework for binary classification in extreme regions, relying on the (nonparametric) theory of regularly varying random vectors, and proved the accuracy of the ERM approach in this context, when the risk functional is computed from extreme observations only. The present contribution may open a new line of research, insofar as progress can be naturally expected in the design of algorithmic learning methods tailored to extreme points (or their projection onto the unit sphere) and statistical issues such as estimation of the minimum risk in the extremes, L_∞^*, remain to be addressed.

5.6 Technical proofs

Proof of Theorem 6

Letting P denote the law of X, in view of expression (5.6) for $L_{P_\infty}^*$,

$$\begin{aligned} L_t(g^*) - L_{P_\infty}^* &= L_t(g^*) - \mathbb{E}\left[\min\{\eta_\infty(X_\infty), 1-\eta_\infty(X_\infty)\}\right] \\ &\leq \frac{t\int_{\{\|x\|>t\}} \min\{\eta(x), 1-\eta(x)\} - \min\{\eta_\infty(x), 1-\eta_\infty(x)\}\,\mathrm{d}P(x)}{t\mathbb{P}\left\{\|X\|>t\right\}} \\ &+ \left(\frac{t\int_{\{\|x\|>t\}} \min(\eta_\infty, 1-\eta_\infty)\,\mathrm{d}P}{t\mathbb{P}\left\{\|X\|>t\right\}} - \frac{\int_{\{\|x\|>1\}} \min(\eta_\infty, 1-\eta_\infty)\,\mathrm{d}\mu}{\mu(\Omega^c)}\right) \\ &:= A + B \end{aligned}$$

The first term is controlled by Assumption 3. Indeed

$$A \leq \sup_{r\geq t}\sup_{\theta\in S} |\eta(r\theta) - \eta_\infty(\theta)|$$

which goes to 0 as $t\to\infty$ under Assumption 3. Now regular variation of F means that for any continuous function h with compact support in $[0,\infty]^d \setminus \{0\}$, (that is with support bounded away from 0), $t\mathbb{E}\left\{h(t^{-1}X)\right\} \to \int h\,\mathrm{d}\mu$, which implies, using the continuity assumption on η_∞, that $B\to 0$ as well.

We now turn to the second assertion of the theorem. Since $L_\infty^* = L_\infty(g^*)$, any classifier $\tilde{g}$ such that

$$\limsup_{t\to\infty} \left\{L_t(\tilde{g}) - L_t(g^*)\right\} = 0 \tag{5.22}$$

minimizes L_∞ as well. We shall thus prove that 5.22 holds for $\tilde{g} = g_\infty^*$. For any classifier of the kind $g(x) = 2\mathbb{1}\{s(x) > 1/2\} - 1$ where s is a scoring function, we have

$$\begin{aligned} L_t(g) &= \frac{\int_{\{\|x\|>t\}} \eta(x)\mathbb{1}\{s(x)<1/2\} + (1-\eta(x))\mathbb{1}\{s(x)>1/2\}\}\,\mathrm{d}P(x)}{\mathbb{P}\left\{\|X\|>t\right\}} \\ &= \mathbb{E}\left\{(2\eta(X)-1)\mathbb{1}\{s(X)<1/2\} \mid \|X\|>t\right\} + \mathbb{E}\left\{1-\eta(X) \mid \|X\|>t\right\}, \end{aligned}$$

thus

$$L_t(g^*_\infty) - L_t(g^*) = \mathbb{E}\left\{(2\eta(X)-1)\left(\mathbb{1}\{\eta_\infty(X) < 1/2\} - \mathbb{1}\{\eta(X) < 1/2\}\right) \mid \|X\| > t\right\}.$$

Let $0 < \epsilon < 1/2$. We may write

$$L_t(g^*_\infty) - L_t(g^*) = \frac{1}{t\mathbb{P}\{\|x\| > t\}}(A + B + C),$$

with

$$A = t\int_{\|x\|>t, \eta_\infty(x)<1/2-\epsilon} (2\eta(x)-1)\left(1 - \mathbb{1}\left\{\eta(x) < 1/2\right\}\right) \mathrm{d}P(x)\,,$$
$$B = t\int_{\|x\|>t, \eta_\infty(x)>1/2+\epsilon} (2\eta(x)-1)\left(-\mathbb{1}\left\{\eta(x) < 1/2\right\}\right) \mathrm{d}P(x)\,,$$
$$C = t\int_{\|x\|>t, |\eta_\infty(x)-1/2|\le\epsilon} (2\eta(x)-1)\left(\mathbb{1}\{\eta_\infty(x) < 1/2\} - \mathbb{1}\{\eta(x) < 1/2\}\right) \mathrm{d}P(x)\,.$$

For $t_0 > 0$ such that $\sup_{\|x\|>t_0} |\eta(x) - \eta_\infty(x)| < \epsilon/2$ (see Remark 6), the integrands in A and B are zero. On the other hand,

$$|C| \le 2\epsilon * 2t\mathbb{P}\{\|X\| > t\}$$

Thus for $t > t_0$, $L_t(g^*_\infty) - L_t(g^*) < 4\epsilon$. Since ϵ is arbitrarily small, the proof is complete.

Proof of Theorem 7

The proof relies on the classical bias/variance risk decomposition which takes the following form for the risk above level t:

$$L_{t_\tau} - L^*_{t_\tau} \le 2\sup_{g\in\mathcal{G}_S} |\widehat{L}_k(g) - L_{t_\tau}(g)| + \inf_{g\in\mathcal{G}_S} L_{t_\tau}(g) - L^*_{t_\tau}. \tag{5.23}$$

The statement of the theorem then immediately derives from the uniform bound on the deviations of $\widehat{L}_k$ of the class $\mathcal{G}_S$ stated in Theorem 9 below.

Theorem 9. *In the setting of Theorem 7, for all $\delta \in (0,1)$, we have with probability $1-\delta$:*

$$\sup_{g\in\mathcal{G}_S} \left|\widehat{L}_k(g) - L_{t_\tau}(g)\right| \le \frac{1}{\sqrt{k}}\left(\sqrt{2(1-\tau)\log(2/\delta)} + C\sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right)$$
$$+ \frac{1}{k}\left(5 + 2\log(1/\delta) + \sqrt{\log(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right)$$

where C is a constant independent of n, τ, δ.

Proof of Theorem 9. Set $k = \lfloor n\tau \rfloor$ throughout. Introduce the pseudo-empirical risk

$$\widetilde{L}_k = \frac{1}{k}\sum_1^n \mathbb{1}\{g(X_i) \ne Y_i, \|X_i\| \ge t_\tau\}.$$

Notice that $\widetilde{L}_k$ is not observed since t_τ is unknown: it serves a useful intermediate quantity in the following excess risk decomposition:

$$\sup_{g\in\mathcal{G}_S}\left|\widehat{L}_k(g)-L_{t_\tau}(g)\right| \le \underbrace{\sup_{g\in\mathcal{G}_S}\left|\widehat{L}_k(g)-\widetilde{L}_k(g)\right|}_{A} + \underbrace{\sup_{g\in\mathcal{G}_S}\left|\widetilde{L}_k(g)-L_{t_\tau}(g)\right|}_{B} \tag{5.24}$$

The remainder of the proof consists in controlling the first term A in the r.h.s. of (5.24) via the Bernstein inequality while the second term B requires a call to a VC inequality for low probability regions. As for the first term,

$$\begin{aligned}
A &\le \frac{1}{k}\left|\sum_1^n \mathbb{1}\{g(X_i)\neq Y_i\}\left(\mathbb{1}\{\|X_i\|\ge X_{(k)}\}-\mathbb{1}\{\|X_i\|\ge t_\tau\}\right)\right| \\
&\le \frac{1}{k}\sum_1^n \left|\mathbb{1}\{\|X_i\|\ge \|X_{(k)}\|\}-\mathbb{1}\{\|X_i\|\ge t_\tau\}\right| \\
&= \frac{1}{k}\sum_1^k \left|1-\mathbb{1}\{\|X_{(i)}\|\ge t_\tau\}\right| + \frac{1}{k}\sum_{k+1}^n \left|\mathbb{1}\{\|X_{(i)}\|\ge t_\tau\}\right| \\
&= \begin{cases} \frac{1}{k}\sum_{k+1}^n \mathbb{1}\{\|X_{(i)}\|\ge t_\tau\} & \text{if } \|X_{(k)}\|\ge t_\tau \\ \frac{1}{k}\sum_1^k \mathbb{1}\{\|X_{(i)}\|< t_\tau\} & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{1}{k}\sum_1^n \mathbb{1}\{\|X_{(i)}\|\ge t_\tau\}-\frac{k}{k} & \text{if } \|X_{(k)}\|\ge t_\tau \\ \frac{1}{k}\sum_1^n \mathbb{1}\{\|X_{(i)}\|< t_\tau\}-\frac{n-k}{k} & \text{otherwise} \end{cases} \\
&= \left|\frac{1}{k}\sum_1^n \mathbb{1}\{\|X_i\|\ge t_\tau\}-1\right| \\
&\le \frac{|S_n-n\tau|}{k}+\frac{1}{k}
\end{aligned}$$

where $S_n=\sum_1^n W_i$ and $W_i=\mathbb{1}\{\|X_i\|>t_\tau\}$. Since $\mathbb{E}\{W_i\}=n\tau$, Bernstein inequality implies, for $y>0$, $\mathbb{P}\{|S_n-n\tau|/k>y\}\le 2\exp\{-(y^2k^2/2)/(n\tau(1-\tau)+yk/3)\}:=\delta$. Solving the latter bound for y yields

$$y=\frac{1}{k}\left(\frac{1}{3}\log(2/\delta)+\sqrt{(1/3\log(2/\delta))^2+2n\tau(1-\tau)\log(2/\delta)}\right).$$

Simplifying the latter bound using that for $a,b>0$, $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, and that $n\tau\le k+1$, we obtain that with probability $1-\delta$,

$$A\le\sqrt{\frac{2}{k}(1-\tau)\log(2/\delta)}+\frac{1}{k}\left(\frac{2}{3}\log(2/\delta)+\sqrt{2(1-\tau)\log(2/\delta)}+1\right) \tag{5.25}$$

We now turn to the second term (B) in (5.24). Write

$$\begin{aligned}
B \le &\underbrace{\frac{n}{k}\sup_{g\in\mathcal{G}_S}\left|\frac{1}{n}\sum\mathbb{1}\{g(X_i)\neq Y_i,\|X_i\|\ge t_\tau\}-\mathbb{P}\{g(X)\neq Y,\|X\|\ge t_\tau\}\right|}_{B_1} \\
&+\underbrace{\left|\frac{1}{\tau}-\frac{n}{k}\right|\mathbb{P}\{g(X)\neq Y,\|X\|\ge t_\tau\}}_{B_2}
\end{aligned}$$

First,

$$B_2 \leq \left|\frac{1}{\tau} - \frac{n}{k}\right| \tau = |(k - n\tau)/k| \leq 1/k. \tag{5.26}$$

Turning to B_1, from Theorem 1 in Goix and collab. (2015), we have that for any class of sets $\mathcal{A}$ on an input space $\mathcal{Z}$ with finite VC dimension $V_{\mathcal{A}}$, if $(Z_i)_{i\leq n}$ are *i.i.d.* copies of a r.v. Z, then

$$\sup_{A\in\mathcal{A}} \left|\mathbb{P}\{Z \in A\} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_i \in A\}\right| \leq C\left(\sqrt{p}\sqrt{\frac{V_{\mathcal{A}}}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right) \tag{5.27}$$

where $p = \mathbb{P}\{Z \in \bigcup_{A\in\mathcal{A}} A\}$ and C is an absolute constant.
By setting $Z = (X, Y)$, $\mathcal{A} = \{\mathcal{A}_g \ : \ g \in \mathcal{G}_{\mathcal{S}}\}$ where $\mathcal{A}_g = \Big\{(x, y),\ g(x) \neq y, \|x\| > t_\tau\Big\}$, so that $p = \tau$, Equation (5.27) becomes:

$$\begin{aligned}
B_2 &\leq \frac{n}{k}\sup_{g\in\mathcal{G}_S}\left|\mathbb{P}\{Z \in A_g\} - \frac{1}{n}\sum \mathbb{1}\{Z_i \in A_g\}\right| \\
&\leq \frac{n}{k}C\left(\sqrt{\tau}\sqrt{\frac{V_{\mathcal{G}_S}}{n}\log(1/\delta)} + 1/n\log(1/\delta)\right) \\
&= C\left(\sqrt{\frac{V_{\mathcal{G}_S}n\tau}{k^2}\log(1/\delta)} + \frac{1}{k}\log(1/\delta)\right) \\
&\leq C\left(\sqrt{\frac{V_{\mathcal{G}_S}}{k}\log(1/\delta)} + \frac{1}{k}\left(\log(1/\delta) + \sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right)\right)
\end{aligned} \tag{5.28}$$

Combining equations (5.25), (5.28) and (5.26), we obtain

$$\begin{aligned}
\sup_{g\in\mathcal{G}_S}\left|\widehat{L}_k(g) - L_{t_\tau}(g)\right| &\leq \sqrt{\frac{2}{k}(1-\tau)\log(2/\delta)} + \frac{1}{k}\left(\frac{2}{3}\log(2/\delta) + \sqrt{2(1-\tau)\log(2/\delta)} + 1\right) \\
&\quad + \frac{1}{k} + C\left(\sqrt{\frac{V_{\mathcal{G}_S}}{k}\log(1/\delta)} + \frac{1}{k}\left(\log(1/\delta) + \sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right)\right) \\
&\leq \frac{1}{\sqrt{k}}\left(\sqrt{2(1-\tau)\log(2/\delta)}C\sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right) \\
&\quad + \frac{1}{k}\left(5 + 2\log(1/\delta) + \sqrt{\log(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right)
\end{aligned}$$

□

Proof of Lemma 3

We prove only the first statement. The proof of the second one follows the same lines and is left to the reader. The third statement is obtained by adding up the left hand right hand sides of the first two.

Decompose $L_t^{>\tau}$ into a type-I and a type-II risk: $L_t^{>\tau}(g, T) = L_{t,+}^{>\tau}(g, T) + L_{t,-}^{>\tau}(g, T)$ with

$$L_{t\sigma}^{>\tau}(g, T) = t\mathbb{P}\left\{g(T(X)) \neq Y, Y = \sigma 1, \theta(T(X)) \in \mathbb{S}_\tau, \|T(X)\| \geq t\right\}.$$

Consider the truncated cones generated by the positively and negatively regions $\mathbb{S}^{\pm}$, $R_g^{\sigma} = \{tv, v \in \mathbb{S}_g^{\pm}, t \geq 1\}$, $\sigma \in \{-, +\}$. Equipped with these notations we may write

$$L_{t,+}^{>\tau}(g,T) = t\mathbb{P}\left\{T(X) \in R_g^-, \theta(T(X)) \in \mathbb{S}_\tau, Y = +1, \|T(X)\| \geq t\right\},$$
$$L_{t,-}^{>\tau}(g,T) = t\mathbb{P}\left\{T(X) \in R_g^+, \theta(T(X)) \in \mathbb{S}_\tau, Y = -1, \|T(X)\| \geq t\right\}.$$

Then notice that

$$\begin{aligned} L_{t,+}^{>\tau}(g,v) &= t\mathbb{P}\left\{V \in R_g^-, \theta(T(X)) \in \mathbb{S}_\tau, \|V\| \geq t, Y = +1\right\} \\ &= pt\mathbb{P}\left\{\theta(V) \in \mathbb{S}_g^- \cap \mathbb{S}_\tau, \|V\| \geq t \;\middle|\; Y = +1\right\} \\ &\to p\Phi^+(\mathbb{S}_g^- \cap \mathbb{S}_\tau), \end{aligned}$$

where the last convergence occurs because of Assumption 4 and the fact that Φ^+ is dominated by Φ. proceeding similarly with $L_t^-(g,v)$, the result follows.

Proof of Theorem 8

Recall from the proof of Lemma 3 the decomposition of $L_\infty^{>\tau}$ into type-I and type-II errors, $L_\infty^{>\tau} = L_{\infty,+}^{>\tau} + L_{\infty,-}^{>\tau}$ with $L_{\infty,+}^{>\tau} = p\Phi^+(\mathbb{S}_g^- \cap \mathbb{S}_\tau)$ and $L_{\infty,-}^{>\tau} = (1-p)\Phi^-(\mathbb{S}_g^+ \cap \mathbb{S}_\tau)$. Recall also the definitions for their empirical counterparts $\widehat{L}_\pm^\tau$ in Lemma 4. For $g \in \mathcal{G}$, the deviations of the empirical risk may be bounded by the sum of the deviations of the two error types,

$$\begin{aligned} |\widehat{L}^\tau(g,\hat{v}) - L_\infty^{>\tau}(g)| &= |\widehat{L}_+^\tau(g,\hat{v}) - L_{\infty,+}^{>\tau}(g) + \widehat{L}_-^\tau(g,\hat{v}) - L_{\infty,-}^{>\tau}(g)| \\ &\leq |\widehat{L}_+^\tau(g,\hat{v}) - Ł_{\infty,+}^{>\tau}(g)| + |\widehat{L}_-^\tau(g,\hat{v}) - L_{\infty,-}^{>\tau}(g)|. \end{aligned} \tag{5.29}$$

Let us focus on the first term of the sum. From Lemmata 3 and 4, recalling that $k^+ = kn^+/n$, we have

$$\begin{aligned} &|\widehat{L}_+^\tau(g,\hat{v}) - L_{\infty,+}^{>\tau}(g)| \\ &\leq \left|\frac{k^+}{k}\hat{\Phi}^+(\mathbb{S}_g^- \cap \mathbb{S}_\tau) - p\Phi^+(\mathbb{S}_g^- \cap \mathbb{S}_\tau)\right|, \end{aligned}$$

which suggests extending the concentration results concerning the empirical measure $\hat{\Phi}$ to its conditional version $\hat{\Phi}^+, \hat{\Phi}^-$. To do so we shall work on the product space $\mathbb{R}^d \times \{-1,1\}$. We first introduce some notations. Let Q be the joint distribution of the pair (X,Y) on $\mathbb{R}^d \times \{-1,1\}$ and let Q_n denote its empirical version, $Q_n = \frac{1}{n}\sum_{i\leq n}\delta_{(X_i,Y_i)}$. As in Section 4.3 we can and will assume that each margin X_j is unit Pareto so that $X = V$. The empirical measure of the rank-transformed data is

$$\widehat{Q}_n = \frac{1}{n}\sum_{i\leq n}\delta_{(\hat{V}_i,Y_i)} = Q_n \circ (\hat{v}, \mathrm{id})^{-1}$$

where id is the identity function mapping (on $\{-1,1\}$).

Finally for define a limit measure on $\mathbb{R}_+^d \setminus \{0\} \times \{-1,1\}$,

$$\nu(B \times \{\sigma 1\}) = \lim t\mathbb{P}\{V \in tB, Y = \sigma 1\} = \mathbb{P}\{Y = \sigma 1\}\,\mu^\sigma(B),$$

which exists from the conditional regular variation assumption (5.11). Notice that ν is homogeneous of order -1 w.r.t. the first component. With these notations and those borrowed from the proof of Theorem 2 (see (4.35) for the definition of $\widehat{\Gamma}_A$), for $A \subset \mathbb{S}$, we have

$$\begin{aligned}\frac{k^+}{k}\widehat{\Phi}^+(A) &= \frac{n}{k}\widehat{Q}_n(\frac{n}{k}\mathcal{C}_A \times \{+1\}) \\ &= \frac{n}{k}Q_n(\hat{v}^{-1}(\frac{n}{k}\mathcal{C}_A) \times \{+1\}) \\ &= \frac{n}{k}Q_n(\widehat{\Gamma}_A \times \{+1\})\end{aligned}$$

and

$$p\Phi^+(A) = \nu(\mathcal{C}_A \times \{+1\})$$

It is shown in the proof of Theorem 2 that under the assumptions of the statement, there exists an event $\mathcal{E}_1$ of probability at least $(1 - d\delta)$ on which $\frac{n}{k}\Gamma^-_{A,1} \subset \widehat{\Gamma}_A \subset \frac{n}{k}\Gamma^+_{A,1}$.

In addition recall that

$$\forall A \in \mathcal{A}, \qquad \Gamma^-_{A,1} \subset \mathcal{C}_A \subset \Gamma^+_{A,1}.$$

Thus on the event $\mathcal{E}_1$, as in the proof of Theorem 2 we can decompose the error as

$$\begin{aligned}\frac{k^+}{k}\widehat{\Phi}^+(A) - p\Phi^+(A) &= \tfrac{n}{k}Q_n(\widehat{\Gamma}_A \times \{+1\}) - \nu(\mathcal{C}_A \times \{+1\}) \\ &\le \tfrac{n}{k}Q_n(\tfrac{n}{k}\Gamma^+_{A,1} \times \{+1\}) - \nu(\Gamma^-_{A,1} \times \{+1\}) \\ &\le \tfrac{n}{k}|Q_n(\tfrac{n}{k}\Gamma^+_{A,1} \times \{+1\}) - Q(\tfrac{n}{k}\Gamma^+_{A,1} \times \{+1\})| \\ &\quad + |\tfrac{n}{k}Q(\tfrac{n}{k}\Gamma^+_{A,1} \times \{+1\}) - \nu(\Gamma^+_{A,1} \times \{+1\})| \\ &\quad + \nu(\Gamma^+_{A,1} \setminus \Gamma^-_{A,1} \times \{+1\}).\end{aligned}$$

A lower bound for the estimation error can be derived in a similar way, yielding, on $\mathcal{E}_1$,

$$\begin{aligned}|\frac{k^+}{k}\widehat{\Phi}^+(A) - p\Phi^+(A)| \le & \max_{B \in \{\Gamma^+_{A,1}, \Gamma^-_{A,1}\}} \tfrac{n}{k}|Q_n(\tfrac{n}{k}B \times \{+1\}) - Q(\tfrac{n}{k}B \times \{+1\})| \quad \text{(stochastic error II)} \\ & + \max_{B \in \{\Gamma^+_{A,1}, \Gamma^-_{A,1}\}} \tfrac{n}{k}|Q(\tfrac{n}{k}B \times \{+1\}) - \nu(B \times \{+1\})| \quad \text{(bias term -II)} \\ & + \nu(\Gamma^+_{A,1} \setminus \Gamma^-_{A,1} \times \{+1\}) \quad \text{(framing gap II).}\end{aligned}$$

We treat the three terms separately, following closely the proof of Theorem 2.

Stochastic error II. Since by construction $Q(B \times \{+1\}) \le P(B)$ the Q-probability of the class $\mathcal{F}' = \mathcal{F} \times \{+1\}$ defined in (4.30) is less than $d(1 + \Delta)\frac{k}{n}$. Also the class $\mathcal{F}'$ has same VC-dimension $V_{\mathcal{F}}$ as $\mathcal{F}$. Thus on an event $\mathcal{E}_2^+$ of probability $1 - \delta$ we again have

$$\begin{aligned}&\sup_{A \in \mathcal{A}} \max_{B \in \{\Gamma^+_{A,1}, \Gamma^-_{A,1}\}} \tfrac{n}{k}|Q_n(\tfrac{n}{k}B \times \{+1\}) - Q(\tfrac{n}{k}B \times \{+1\})| \\ &\le C\left(\sqrt{d(1+\Delta)V_{\mathcal{F}}k^{-1}\log(1/\delta)} + k^{-1}\log(1/\delta)\right).\end{aligned}$$

which is the term error in the statement.

Bias term-II. Taking the supremum over $A \in \mathcal{A}$ immediately yields the bias term in the statement of Theorem 8.

Framing gap-II. As it is the case the proof of Theorem 2, the framing gap in the product space satisfies

$$\begin{aligned}
&\nu(\Gamma^+_{A,1} \setminus \Gamma^-_{A,1} \times \{+1\}) \leq \\
&\nu\left(\left\{x \in [0,\infty)^d : \ (1+\tfrac{1}{n}-\tfrac{1}{k}+\Delta)^{-1} \leq \|x\|_\infty < (1+\tfrac{1}{n}-\tfrac{1}{k}-\Delta)^{-1}\right\} \times \{+1\}\right) \\
&+\nu\left(\left\{x \in [0,\infty)^d : \ \|x\|_\infty \geq 1, \ \theta(x) \in A_+(3\Delta\|x\|_\infty) \setminus A_-(3\Delta\|x\|_\infty)\right\} \times \{+1\}\right)
\end{aligned} \tag{5.30}$$

The first term on the right-hand side of (5.30) is equal to

$$2\Delta\nu(\{x \in [0,\infty)^d : \|x\|_\infty \geq 1\} \times \{+1\}) = 2\Delta p\Phi^+(\mathbb{S}) \leq 2\Delta\Phi(\mathbb{S}),$$

where the latter inequality comes from the decomposition $\Phi = p\Phi^+ + (1-p)\Phi^-$. The second term on the right-hand side in (5.30) can be expressed using the polar decomposition of μ (and thus ν),

$$\begin{aligned}
&\nu\left(\left\{x \in [0,\infty)^d : \ \|x\|_\infty \geq 1, \ \theta(x) \in A_+(3\Delta\|x\|_\infty) \setminus A_-(3\Delta\|x\|_\infty)\right\} \times \{+1\}\right) \\
&= \int_1^\infty p\Phi^+\Big(A_+(3\Delta r) \setminus A_-(3\Delta r)\Big)\frac{\mathrm{d}r}{r^2} \\
&\leq \int_1^\infty \Phi\Big(A_+(3\Delta r) \setminus A_-(3\Delta r)\Big)\frac{\mathrm{d}r}{r^2}
\end{aligned}$$

Now it has been shown in the above mentioned proof that the latter display is less than

$$3c\Delta(1 + \ln\Phi(\mathbb{S}) - \ln(3c\Delta))$$

We thus obtain the same term gap as in Theorem 2

So far we have only treated one of the two terms of the error decomposition (5.29). The second one is treated in the same way and the associated upper bound is identical, which yields the factor 2 in the statement of Theorem 8. The decomposition of $|k^-/k\widehat{\Phi}^-(A) - (1-p)\Phi^-(A)|$ into a stochastic, a bias and framing gap holds true on the same event $\mathcal{E}_1$. The bound on the stochastic error holds true on an event $\mathcal{E}_2^-$ of probability at least $1-\delta$. Thus the upper bound in the statement of Theorem 8 holds true on the intersection $\mathcal{E}_1 \cap \mathcal{E}_2^+ \cap \mathcal{E}_2^-$ which has probability at least $1-(d+2)\delta$.

5.7 Numerical experiments

5.7.1 Synthetic data from the Clover distribution

The Clover distribution introduced in Cai and collab. (2011) has the following density

$$f(x,y) = \begin{cases} \frac{3}{10\pi} r_0^4 (1+r_0^6)^{\frac{-3}{2}} \left(5 + \frac{4(x^2+y^2)^2 - 32x^2y^2}{r_0(x^2+y^2)^{-\frac{3}{2}}}\right) & x^2+y^2 < r_0 \\ \frac{3}{10\pi}\left(\frac{9(x^2+y^2)^2 - 32x^2y^2}{r_0(x^2+y^2)^{-\frac{3}{2}}}\right) & x^2+y^2 \geq r_0 \end{cases} \quad \text{with } r_0 = 1.248$$

Without loss of generality, only the points within the positive orthant are kept. Points labeled -1 are rotated by an angle θ in the counterclockwise direction. Figure 5.7 provides an example of 2D points from both distributions with $\theta = \frac{\pi}{4}$. the number of trees for both random forests (in the regular setting and in the setting of Algorithm 1) is set to 200. The number of neighbors for both k-NN's is set to 5.

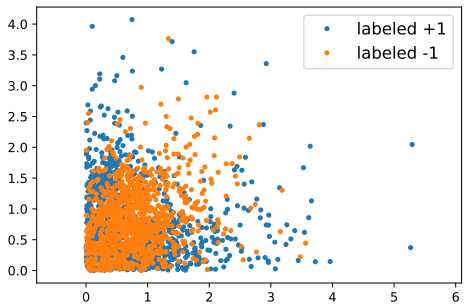


Figure 5.7 – Labeled dataset generated from a Clover distribution and its θ-rotated version.

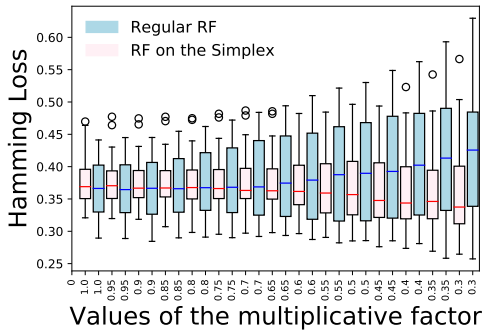


Figure 5.8 – Clover data - test loss of random forest on the simplex and regular random forest depending on the multiplicative factor κ.

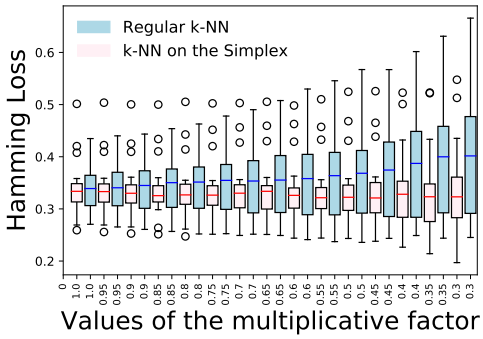


Figure 5.9 – Clover data - test loss of k-NN on the simplex and regular k-NN depending on the multiplicative factor κ.

Figures 5.8 and 5.9 illustrate once again the conclusions from Section 5.4.

5.7.2 Synthetic data from the Logistic distribution

The multivariate logistic model is a widely used model in the context of extreme value analysis, see *e.g.* Coles and Tawn (1991) or Stephenson (2003) for efficient simulation algorithms. The logistic distribution in $\mathbb{R}^d$ with parameter $\delta \in (0, 1]$ has cumulative distribution function

$$F(x) = \exp\left\{ -\left(\sum_{i=1}^{d} x_j^{\frac{-1}{\delta}} \right)^{\delta} \right\}, \quad x \in (0, \infty)^d.$$

For small values of δ, extremes tend to occur simultaneously in all direction, that is, the angular measure concentrates around the center of the positive orthant of

the unit sphere. On the other hand, for δ close to 1, extremes tend to concentrate near the axes, that is only one component at a time is likely to be large. The positive and negative instances are respectively generated according to a logistic distribution with parameter $\delta_+ = 0.1$ and $\delta_- = 0.5$, in dimension $d = 4$. Ten datasets are simulated. Each one is composed of a train set and a test set containing respectively 10^4 and 10^5 instances. k the number of points used for training the classifier in Algorithm 1) is set to 100. The multiplicative factor κ is made to vary between 1 and 0.3. Figure 5.10 displays evolutions of boxplots of Hamming losses depending on the multiplicative factor κ: each box summarizes the distribution of the losses obtained with the 10 considered datasets. It is analogue to Figure 5.4 using k-NN instead of RF, from which similar conclusions can be drawn.

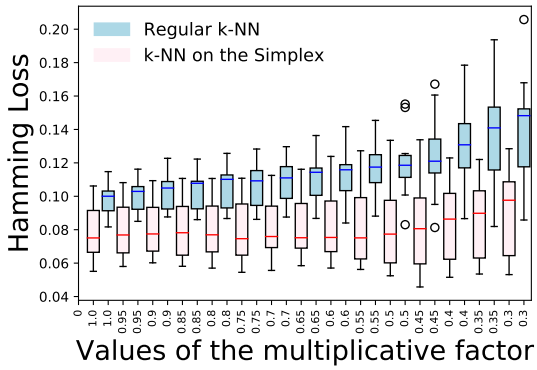
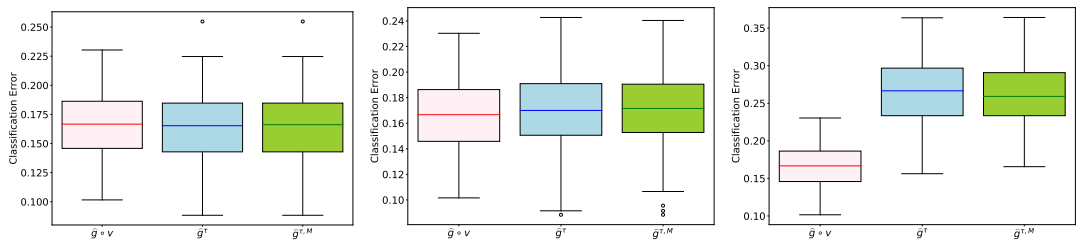


Figure 5.10 – Logistic data - test loss of k-NN on the simplex and regular k-NN depending on the multiplicative factor κ.

5.7.3 Further Numerical Experiments on the Influenze of Marginal Standardization

Figure 5.6 provides a comparison of the classification errors for varying class of classifiers with the empirical standardization and its truncated counterpart. In our experimental framework, since we work in the asymptotic regime, one can directly compute the standardization as detailed in Lemma 1. Therefore Figure 5.11 extends Figure 5.6 as it reports the classification test error with the standardization v for the three class of classifiers mentioned above. In Figure 5.6, we denote by $\hat{g} \circ v$ the classifier trained on samples standardized with v. Note that in order to compare this latter classifier with g^τ and g^τ, M the training and test data are restricted to samples with angular component belonging to $\mathbb{S}_\tau$.



(a) (b) (c)

Figure 5.11 – Classification test error of $\widehat{g} \circ v$, $\widehat{g}^{\tau}$ and $\widehat{g}^{\tau,M}$ issued from logistic regression (5.6a), random forests (5.6b) and classification trees (5.6c).

5.7.4 Real-world data: Ecoli dataset

The Ecoli dataset contains 336 instances. There are 8 features: 7 numerical features and 1 corresponding to a sequence name. In our experimental framework the feature corresponding to the sequence name is dropped: we only work with the other features. The label to be predicted corresponds to the protein localization site among the 8 different localizations possible. In our experiments the labeling is simplified: data labeled *im* are set to 1, all instances labeled differently are set to -1. The classification problem is thus turned into a binary one. Figure 5.12 is the analogue of Figure 5.5 using k-NN instead of RF, from which similar conclusions can be drawn.

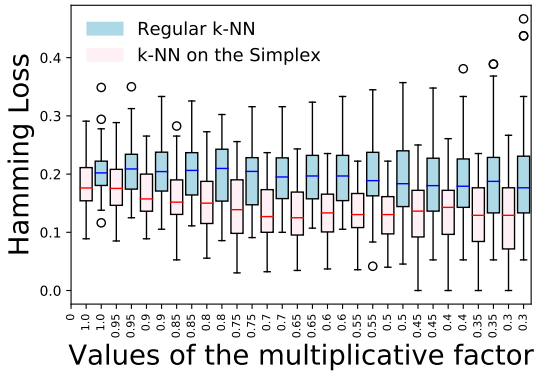


Figure 5.12 – Real data - test loss of regular k-NN depending on the multiplicative factor κ.

Remark 11. *(On the truncation step) In all experiments, the value of k is set to* 100. *Finding k, the number of extreme observations involved in the computation of the empirical risk* (5.9)*, is beyond the scope of this chapter though one could expect better results by improving the selection of k.*

5.8 References

Brownlees, C., E. Joly and G. Lugosi. 2015, ■Empirical risk minimization for heavy-tailed losses■, *Ann. Statist.*, vol. 43, nº 6, p. 2507–2536. 91

Cai, J., J. Einmahl and L. De Haan. 2011, ■Estimation of extreme risk regions under multivariate regular variation■, *The Annals of Statistics*, p. 1803–1826. 95, 113

Carpentier, A. and M. Valko. 2014, ■Extreme bandits■, in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., p. 1089–1097. 91

Coles, S. and J. Tawn. 1991, ■Modeling extreme multivariate events■, *JR Statist. Soc. B*, vol. 53, p. 377–392. 114

De Haan, L. and S. Resnick. 1987, ■On regular variation of probability densities■, *Stochastic processes and their applications*, vol. 25, p. 83–93. 95

Devroye, L., L. Györfi and G. Lugosi. 1996, *A Probabilistic Theory of Pattern Recognition*, Applications of mathematics : stochastic modelling and applied probability, U.S. Government Printing Office. 90, 96

Goix, N., A. Sabourin and S. Clémençon. 2015, ■Learning the dependence structure of rare events: a non-asymptotic study■, in *Conference on Learning Theory*, p. 843–860. 110

Goix, N., A. Sabourin and S. Clémençon. 2016, ■Sparse representation of multivariate extremes with applications to anomaly ranking■, in *Artificial Intelligence and Statistics*, p. 75–83. 91, 98

Goix, N., A. Sabourin and S. Clémençon. 2017, ■Sparse representation of multivariate extremes with applications to anomaly detection■, *Journal of Multivariate Analysis*, vol. 161, p. 12–31. 94, 98

Mendelson, S. 2018, ■Learning without concentration for general loss functions■, *Probability Theory and Related Fields*, vol. 171, nº 1, p. 459–502. 91

Nakai, K. and M. Kanehisa. 1992, ■A knowledge base for predicting protein localization sites in eukaryotic cells■, *Genomics*, vol. 14, nº 4, p. 897–911. 104

Ohannessian, M. I. and M. A. Dahleh. 2012, ■Rare probability estimation under regularly varying heavy tails■, in *Conference on Learning Theory*, p. 21–1. 91

Resnick, S. 1987, *Extreme Values, Regular Variation, and Point Processes*, Springer Series in Operations Research and Financial Engineering. 91

Roos, T., P. Grünwald, P. Myllymäki and H. Tirri. 2006, ■Generalization to unseen cases■, in *Advances in Neural Information Processing Systems 18*, édité par Y. Weiss, B. Schölkopf and J. C. Platt, MIT Press, p. 1129–1136. URL http://papers.nips.cc/paper/2821-generalization-to-unseen-cases.pdf. 91

STEPHENSON, A. 2003, ■Simulating multivariate extreme value distributions of logistic type■, *Extremes*, vol. 6, nº 1, p. 49–59. 104, 114

Part III

Learning a Heavy-Tailed Text Representation & Subspace Clustering in Extremes

Chapter 6

Heavy-tailed Representations, Text Polarity Classification & Data Augmentation

Chapter abstract

The dominant approaches to text representation in natural language rely on learning embeddings on massive corpora which have convenient properties such as compositionality and distance preservation. In this chapter, we develop a novel method to learn a heavy-tailed embedding with desirable regularity properties regarding the distributional tails, which allows to analyze the points far away from the distribution bulk using the framework of multivariate extreme value theory. In particular, following Chapter 5, a classifier dedicated to the tails of the proposed embedding is obtained which exhibits a *scale invariance* property exploited in a novel text generation method for label preserving dataset augmentation. Experiments on synthetic and real text data show the relevance of the proposed framework and confirm that this method generates meaningful sentences with controllable attribute, *e.g.* positive or negative sentiments.

6.1 Introduction

With the explosion of digital content and text data in the last decade, finding the best mathematically grounded way to represent the meaning of natural language is a scientific challenge that has received increasing attention. Relying on the richness of contents, several embeddings have been proposed Devlin and collab. (2018); Peters and collab. (2018); Radford and collab. (2018) with demonstrated efficiency for the considered tasks when learnt on massive datasets. However, none of these embeddings take into account the fact that word frequency distributions are heavy tailed Baayen (2002); Church and Gale (1995); Mandelbrot (1953), so that extremes are naturally present in texts (see also Fig. 6.9a and 6.9b). Similarly, Babbar and collab. (2014) shows that, contrary to image

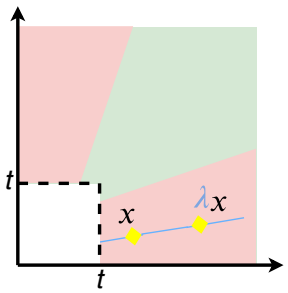


Figure 6.1 – Illustration of an angular classifier g dedicated to extremes $\{x, \|x\|_\infty \geq t\}$ in $\mathbb{R}^2_+$. The red and green truncated cones are respectively labeled as $+1$ and -1 by g.

taxonomies, the underlying distributions for words and documents in large scale textual taxonomies are also heavy tailed. Exploiting this information, several studies, as Clinchant and Gaussier (2010); Madsen and collab. (2005), were able to improve text mining applications by accurately modeling the tails of textual elements.

In this work, we rely on the framework of multivariate extreme value analysis, based on extreme value theory (EVT) which focuses on the distributional tails. EVT is valid under a regularity assumption which amounts to a homogeneity property above large thresholds: the tail behavior of the considered variables must be well approximated by a power law, see Section 6.2 for a rigorous statement. The tail region (where samples are considered as extreme) of the input variable $x \in \mathbb{R}^d$ is of the kind $\{\|x\| \geq t\}$, for a large threshold t. The latter is typically chosen such that a small but non negligible proportion of the data is considered as extreme, namely 25% in our experiments. A major advantage of this framework in the case of labeled data (as detailed in Chapter 5) is that classification on the tail regions may be performed using the angle $\Theta(x) = \|x\|^{-1}x$ only, see Figure 6.1. The main idea behind the present chapter is to take advantage of the scale invariance for two tasks regarding sentiment analysis of text data: *(i)* Improved classification of extreme inputs, *(ii)* Label preserving data augmentation, as the most probable label of an input x is unchanged by multiplying x by $\lambda > 1$.

EVT in a machine learning framework has received increasing attention in the past few years. Learning tasks considered so far include anomaly detection Clifton and collab. (2011); Goix and collab. (2016); Roberts (1999, 2000); Thomas and collab. (2017), anomaly clustering Chiapino and collab. (2019a), unsupervised learning Goix and collab. (2015), online learning Achab and collab. (2017); Carpentier and Valko (2014), dimension reduction and support identification Chiapino and Sabourin (2016); Chiapino and collab. (2019b); Goix and collab. (2017). The present chapter builds upon the methodological framework proposed in Chapter 5 for classification in extreme regions. The goal is to improve the performance of classifiers $\widehat{g}(x)$ issued from Empirical Risk Minimization (ERM) on the tail regions $\{\|x\| > t\}$ Indeed, we argue in Chapter 5, that for very large t, there is no guarantee that $\widehat{g}$ would perform well conditionally to $\{\|X\| > t\}$, precisely because of the scarcity of such examples in the training set. They thus propose to train a specific classifier dedicated to extremes leveraging the probabilistic structure of the tails. We demonstrate the usefulness of their framework with simulated and some real world datasets. However, there is no

reason to assume that the previously mentioned text embeddings satisfy the required regularity assumptions. The aim of the present work is to extend the methodology from Chapter 5 to datasets which do not satisfy their assumptions, in particular to text datasets embedded by state of the art techniques. This is achieved by the algorithm *Learning a Heavy Tailed Representation* (in short **LHTR**) which learns a transformation mapping the input data X onto a random vector Z which does satisfy the aforementioned assumptions. The transformation is learnt by an adversarial strategy Goodfellow and collab. (2016).

In Section 6.7 we propose an interpretation of the extreme nature of an input in both **LHTR** and BERT representations. In a word, these sequences are longer and are more difficult to handle (for next token prediction and classification tasks) than non extreme ones.

Our second contribution is a novel data augmentation mechanism **GENELIEX** which takes advantage of the scale invariance properties of Z to generate synthetic sequences that keep invariant the attribute of the original sequence. Label preserving data augmentation is an effective solution to the data scarcity problem and is an efficient pre-processing step for moderate dimensional datasets Wang and Perez (2017); Wei and Zou (2019). Adapting these methods to NLP problems remains a challenging issue. The problem consists in constructing a transformation h such that for any sample x with label $y(x)$, the generated sample $h(x)$ would remain label consistent: $y\Big(h(x)\Big) = y(x)$ Ratner and collab. (2017). The dominant approaches for text data augmentation rely on word level transformations such as synonym replacement, slot filling, swap deletion Wei and Zou (2019) using external resources such as wordnet Miller (1995). Linguistic based approaches can also be combined with vectorial representations provided by language models Kobayashi (2018). However, to the best of our knowledge, building a vectorial transformation without using any external linguistic resources remains an open problem. In this work, as the label $y\Big(h(x)\Big)$ is unknown as soon as $h(x)$ does not belong to the training set, we address this issue by learning both an embedding φ and a classifier g satisfying a relaxed version of the problem above mentioned, namely $\forall \lambda \geq 1$

$$g\Big(h_\lambda(\varphi(x))\Big) = g\Big(\varphi(x)\Big). \tag{6.1}$$

For mathematical reasons which will appear clearly in Section 6.2.2, h_λ is chosen as the homothety with scale factor λ, $h_\lambda(x) = \lambda x$. In this chapter, we work with output vectors issued by BERT Devlin and collab. (2018). BERT and its variants are currently the most widely used language model but we emphasize that the proposed methodology could equally be applied using any other representation as input. BERT embedding does not satisfy the regularity properties required by EVT (see the results from statistical tests performed in Section 6.9.1) Besides, there is no reason why a classifier g trained on such embedding would be scale invariant, *i.e.* would satisfy for a given sequence u, embedded as x, $g(h_\lambda(x)) = g(x)$ $\forall \lambda \geq 1$. On the classification task, we demonstrate on two datasets of sentiment analysis that the embedding learnt by **LHTR** on top of BERT is indeed following a heavy-tailed distribution. Besides, a classifier trained on the embedding learnt by **LHTR** outperforms the same classifier trained on BERT. On the dataset augmentation task, quantitative and qualitative experiments demonstrate the ability of **GENELIEX** to generate new sequences while preserving labels.

The rest of this chapter is organized as follows. Section 6.2 introduces the necessary background in multivariate extremes. The methodology we propose is detailed at length in Section 6.3. Illustrative numerical experiments on both synthetic and real data are gathered in sections 6.5 and 6.6.

6.2 Background

This section recalls the main notions related to Chapters 2, 5 and briefly brushes through adversarial learning.

6.2.1 Extreme values, heavy tails and regular variation

Extreme value analysis is a branch of statistics which main focus is on events characterized by an unusually high value of a monitored quantity. A convenient working assumption in EVT is *regular variation*. A real-valued random variable X is regularly varying with index $\alpha > 0$, a property denoted as $RV(\alpha)$, if and only if there exists a function $b(t) > 0$, with $b(t) \to \infty$ as $t \to \infty$, such that for any fixed $x > 0$: $t\mathbb{P}\{X/b(t) > x\} \xrightarrow[t\to\infty]{} x^{-\alpha}$. In the multivariate case $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, it is usually assumed that a preliminary component-wise transformation has been applied so that each margin X_j is $RV(1)$ with $b(t) = t$ and takes only positive values. X is *standard multivariate regularly varying* if there exists a positive Radon measure μ on $[0, \infty]^d \backslash \{0\}$

$$t\mathbb{P}\left\{t^{-1}X \in A\right\} \xrightarrow[t\to\infty]{} \mu(A), \tag{6.2}$$

for any Borelian set $A \subset [0, \infty]^d$ which is bounded away from 0 and such that the limit measure μ of the boundary ∂A is zero. For a complete introduction to the theory of Regular Variation, the reader may refer to Resnick (2013). The measure μ may be understood as the limit distribution of tail events. In (6.2), μ is homogeneous of order -1, that is $\mu(tA) = t^{-1}\mu(A)$, $t > 0, A \subset [0, \infty]^d \setminus \{0\}$. This scale invariance is key for our purposes, as detailed in Section 6.2.2. The main idea behind extreme value analysis is to learn relevant features of μ using the largest available data.

6.2.2 Classification in extreme regions

We now recall the classification setup for extremes as introduced in Chapter 5. Let $(X, Y) \in \mathbb{R}^d_+ \times \{-1, 1\}$ be a random pair. We assume standard regular variation for both classes, that is $t\mathbb{P}\{X \in tA \mid Y = \pm 1\} \to \mu_\pm(A)$, where A is as in (6.2). Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and consider the risk of a classifier $g : \mathbb{R}^d_+ \to \{\pm 1\}$ above a radial threshold t,

$$L_t(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\}. \tag{6.3}$$

The goal is to minimize the asymptotic risk in the extremes $L_\infty(g) = \limsup_{t\to\infty} L_t(g)$. Using the scale invariance property of μ, under additional mild regularity assumptions concerning the regression function, namely uniform convergence to the limit

at infinity, one can prove the following result (see Chapter 5, Theorem 6): there exists a classifier $g^\star_\infty$ depending on the pseudo-angle $\Theta(x) = \|x\|^{-1}x$ only, that is $g^\star_\infty(x) = g^\star_\infty\Big(\Theta(x)\Big)$, which is asymptotically optimal in terms of classification risk, *i.e.* $L_\infty(g^\star_\infty) = \inf_{g \text{ measurable}} L_\infty(g)$. Notice that for $x \in \mathbb{R}^d_+ \setminus \{0\}$, the angle $\Theta(x)$ belongs to the positive orthant of the unit sphere, denoted by S in the sequel. As a consequence, the optimal classifiers on extreme regions are based on indicator functions of truncated cones on the kind $\{\|x\| > t, \Theta(x) \in B\}$, where $B \subset S$, see Figure 6.1. We emphasize that the labels provided by such a classifier remain unchanged when rescaling the samples by a factor $\lambda \geq 1$ (*i.e.* $g(x) = g(\Theta(x)) = g(\Theta(\lambda x)), \forall x \in \{x, \|x\| \geq t\}$). The angular structure of the optimal classifier $g^\star_\infty$ is the basis for the following ERM strategy using the most extreme points of a dataset. Let $\mathcal{G}_S$ be a class of angular classifiers defined on the sphere S with finite VC dimension $V_{\mathcal{G}_S} < \infty$. By extension, for any $x \in \mathbb{R}^d_+$ and $g \in \mathcal{G}_S$, $g(x) = g\Big(\Theta(x)\Big) \in \{-1, 1\}$. Given n training data $\{(X_i, Y_i)\}_{i=1}^n$ made of *i.i.d* copies of (X, Y), sorting the training observations by decreasing order of magnitude, let $X_{(i)}$ (with corresponding sorted label $Y_{(i)}$) denote the i-th order statistic, *i.e.* $\|X_{(1)}\| \geq \ldots \geq \|X_{(n)}\|$. The empirical risk for the k largest observations $\widehat{L}_k(g) = \frac{1}{k}\sum_{i=1}^k \mathbf{1}\{Y_{(i)} \neq g(\Theta(X_{(i)}))\}$ is an empirical version of the risk $L_{t(k)}(g)$ as defined in (6.3) where $t(k)$ is a $(1 - k/n)$-quantile of the norm, $\mathbb{P}\{\|X\| > t(k)\} = k/n$. Selection of k is a bias-variance compromise, see Section 6.9 for further discussion. The strategy promoted in Chapter 5 is to use $\widehat{g}_k = \text{argmin}_{g \in \mathcal{G}_S} \widehat{L}_k(g)$, for classification in the extreme region $\{x \in \mathbb{R}^d_+ : \|x\| > t(k)\}$. The following result provides guarantees concerning the excess risk of $\widehat{g}_k$ compared with the Bayes risk above level $t = t(k)$, $L^\star_t = \inf_{g \text{ measurable}} L_t(g)$.

Theorem 10. *(Chapter 5, Theorem 7) If each class satisfies the regular variation assumption (6.2), under an additional regularity assumption concerning the regression function $\eta(x) = \mathbb{P}\{Y = +1 \mid x\}$ (see Equation (6.4) in Section 6.5.3), for $\delta \in (0, 1)$, $\forall n \geq 1$, it holds with probability larger than $1 - \delta$ that*

$$L_{t(k)}(\widehat{g}_k) - L^\star_{t(k)} \leq \frac{1}{\sqrt{k}}\left(\sqrt{2(1 - k/n)\log(2/\delta)} + C\sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right) + \frac{1}{k}\left(5 + 2\log(1/\delta) + \sqrt{\log(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right) + \left\{\inf_{g \in \mathcal{G}_S} L_{t(k)}(g) - L^\star_{t(k)}\right\},$$

where C is a universal constant.

In the present work we do *not* assume that the baseline representation X for text data satisfies the assumptions of Theorem 10. Instead, our goal is is to render the latter theoretical framework applicable by learning a representation which satisfies the regular variation condition given in (6.2), hereafter referred as Condition (6.2) which is the main assumption for Theorem 10 to hold. Our experiments demonstrate empirically that enforcing Condition (6.2) is enough for our purposes, namely improved classification and label preserving data augmentation, see Section 6.5.3 for further discussion.

6.2.3 Adversarial learning

Adversarial networks, introduced in Goodfellow and collab. (2014), form a system where two neural networks are competing. A first model G, called the generator, generates samples as close as possible to the input dataset. A second model D, called the discriminator, aims at distinguishing samples produced by the generator from the input dataset. The goal of the generator is to maximize the probability of the discriminator making a mistake. Hence, if P_{input} is the distribution of the input dataset then the adversarial network intends to minimize the distance (as measured by the Jensen-Shannon divergence) between the distribution of the generated data P_G and P_{input}. In short, the problem is a minmax game with value function $V(D, G)$

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{input}}}[\log D(x)] + \mathbb{E}_{z\sim P_G}[\log\Big(1 - D(G(z))\Big)].$$

Auto-encoders and derivations Fard and collab. (2018); Goodfellow and collab. (2016); Laforgue and collab. (2018) form a subclass of neural networks whose purpose is to build a suitable representation by learning encoding and decoding functions which capture the core properties of the input data. An adversarial auto-encoder (see Makhzani and collab. (2015)) is a specific kind of auto-encoders where the encoder plays the role of the generator of an adversarial network. Thus the latent code is forced to follow a given distribution while containing information relevant to reconstructing the input. In the remaining of this chapter, a similar adversarial encoder constrains the encoded representation to be heavy-tailed.

6.3 Heavy-tailed Text Embeddings

6.3.1 Learning a heavy-tailed representation

We now introduce a novel algorithm *Learning a heavy-tailed representation* (**LHTR**) for text data from high dimensional vectors as issued by pre-trained embeddings such as BERT. The idea behind is to modify the output X of BERT so that classification in the tail regions enjoys the statistical guarantees presented in Section 6.2, while classification in the bulk (where many training points are available) can still be performed using standard models. Stated otherwise, **LHTR** increases the information carried by the resulting vector $Z = \varphi(X) \in \mathbb{R}^{d'}$ regarding the label Y in the tail regions of Z in order to improve the performance of a downstream classifier. In addition **LHTR** is a building block of the data augmentation algorithm **GENELIEX** detailed in Section 6.3.2. **LHTR** proceeds by training an encoding function φ in such a way that *(i)* the marginal distribution $q(z)$ of the code Z be close to a user-specified heavy tailed target distribution p satisfying the regularity condition (6.2); and *(ii)* the classification loss of a multilayer perceptron trained on the code Z be small.

A major difference distinguishing **LHTR** from existing auto-encoding schemes is that the target distribution on the latent space is not chosen as a Gaussian distribution but as a heavy-tailed, regularly varying one. A workable example of such a target is provided in our experiments (Section 6.5). As the Bayes classifier (*i.e.* the optimal one among all possible classifiers) in the extreme

region has a potentially different structure from the Bayes classifier on the bulk (recall from Section 6.2 that the optimal classifier at infinity depends on the angle $\Theta(x)$ only), **LHTR** trains two different classifiers, g^{ext} on the extreme region of the latent space on the one hand, and g^{bulk} on its complementary set on the other hand. Given a high threshold t, the extreme region of the latent space is defined as the set $\{z : \|z\| > t\}$. In practice, the threshold t is chosen as an empirical quantile of order $(1-\kappa)$ (for some small, fixed κ) of the norm of encoded data $\|Z_i\| = \|\varphi(X_i)\|$. The classifier trained by **LHTR** is thus of the kind $g(z) = g^{\text{ext}}(z)\mathbb{1}\{\|z\| > t\} + g^{\text{bulk}}(z)\mathbb{1}\{\|z\| \leq t\}$. If the downstream task is classification on the whole input space, in the end the bulk classifier g^{bulk} may be replaced with any other classifier g' trained on the original input data X restricted to the non-extreme samples (*i.e.* $\{X_i, \|\varphi(X_i)\| \leq t\}$). Indeed training g^{bulk} only serves as an intermediate step to learn an adequate representation φ.

Remark 12. *Recall from Section 6.2.2 that the optimal classifier in the extreme region as $t \to \infty$ depends on the angular component $\theta(x)$ only, or in other words, is scale invariant. One can thus reasonably expect the trained classifier $g^{ext}(z)$ to enjoy the same property. This scale invariance is indeed verified in our experiments (see Sections 6.5 and 6.6) and is the starting point for our data augmentation algorithm in Section 6.3.2. An alternative strategy would be to train an angular classifier,* i.e. *to impose scale invariance. However in preliminary experiments (not shown here), the resulting classifier was less efficient and we decided against this option in view of the scale invariance and better performance of the unconstrained classifier.*

The goal of **LHTR** is to minimize the weighted risk

$$R(\varphi, g^{\text{ext}}, g^{\text{bulk}}) = \rho_1 \mathbb{P}\left\{Y \neq g^{\text{ext}}(Z), \|Z\| \geq t\right\} + \rho_2 \mathbb{P}\left\{Y \neq g^{\text{bulk}}(Z), \|Z\| < t\right\} + \rho_3 \mathfrak{D}(q(z), p(z)),$$

where $Z = \varphi(X)$, $\mathfrak{D}$ is the Jensen-Shannon distance between the heavy tailed target distribution p and the code distribution q, and ρ_1, ρ_2, ρ_3 are positive weights. Following common practice in the adversarial literature, the Jensen-Shannon distance is approached (up to a constant term) by the empirical proxy $\widehat{L}(q,p) = \sup_{D \in \Gamma} \widehat{L}(q,p,D)$, with $\widehat{L}(q,p,D) = \frac{1}{m}\sum_{i=1}^{m} \log D(Z_i) + \log\left(1 - D(\tilde{Z}_i)\right)$, where Γ is a wide class of discriminant functions valued in $[0,1]$, and where independent samples $Z_i, \tilde{Z}_i$ are respectively sampled from the target distribution and the code distribution q. The classifiers g^{ext}, g^{bulk} are of the form $g^{\text{ext}}(z) = 2\mathbb{1}\{C^{\text{ext}}(z) > 1/2) - 1$, $g^{\text{bulk}}(z) = 2\mathbb{1}\{C^{\text{bulk}}(z) > 1/2) - 1$ where $C^{\text{ext}}, C^{\text{bulk}}$ are also discriminant functions valued in $[0,1]$. Following common practice, we shall refer to $C^{\text{ext}}, C^{\text{bulk}}$ as classifiers as well. In the end, **LHTR** solves the following min-max problem $\inf_{C^{\text{ext}}, C^{\text{bulk}}, \varphi} \sup_D \widehat{R}(\varphi, C^{\text{ext}}, C^{\text{bulk}}, D)$ with

$$\widehat{R}(\varphi, C^{\text{ext}}, C^{\text{bulk}}, D) = \frac{\rho_1}{k}\sum_{i=1}^{k} \ell(Y_{(i)}, C^{\text{ext}}(Z_{(i)})) + \frac{\rho_2}{n-k}\sum_{i=k+1}^{n-k} \ell(Y_{(i)}, C^{\text{bulk}}(Z_{(i)})) + \rho_3 \hat{L}(q,p,D),$$

where $\{Z_{(i)} = \varphi(X_{(i)}), i = 1, \ldots, n\}$ are the encoded observations with associated labels $Y_{(i)}$ sorted by decreasing magnitude of $\|Z\|$ (*i.e.* $\|Z_{(1)}\| \geq \cdots \geq \|Z_{(n)}\|$),

$k = \lfloor \kappa n \rfloor$ is the number of extreme samples among the n encoded observations and $\ell(y, C(x)) = -(y \log C(x) + (1-y)\log(1-C(x)), y \in \{0,1\}$ is the negative log-likelihood of the discriminant function $C(x) \in (0,1)$. A summary of **LHTR** and an illustration of its workflow are provided in Section 6.4.1.

6.3.2 A heavy-tailed representation for dataset augmentation

We now introduce **GENELIEX** (Generating Label Invariant sequences from Extremes), a data augmentation algorithm, which relies on the label invariance property under rescaling of the classifier for the extremes learnt by **LHTR**. **GENELIEX** considers input sentences as sequences and follows the seq2seq approach Sutskever and collab. (2014). It trains a Transformer Decoder Vaswani and collab. (2017) G^{ext} on the extreme regions.

For an input sequence $U = (u_1, \ldots, u_T)$ of length T, represented as X_U by BERT with latent code $Z = \varphi(X_U)$ lying in the extreme regions, **GENELIEX** produces, through its decoder G^{ext} M sequences U'_j where $j \in \{1, \ldots, M\}$. The M decoded sequences correspond to the codes $\{\lambda_j Z, j \in \{1, \ldots, M\}\}$ where $\lambda_j > 1$. To generate sequences, the decoder iteratively takes as input the previously generated word (the first word being a start symbol), updates its internal state, and returns the next word with the highest probability. This process is repeated until either the decoder generates a stop symbol or the length of the generated sequence reaches the maximum length ($T_{\max}$). To train the decoder $G^{\text{ext}} : \mathbb{R}^{d'} \to \left[1, \ldots, |\mathcal{V}|\right]^{T_{\max}}$ where $\mathcal{V}$ is the vocabulary on the extreme regions, **GENELIEX** requires an additional dataset $\mathcal{D}_{g_n} = (U_1, \ldots, U_n)$ (not necessarily labeled) with associated representation *via* BERT $(X_{U,1}, \ldots, X_{U,n})$. Learning is carried out by optimising the classical negative log-likelihood of individual tokens ℓ_{gen}. The latter is defined as $\ell_{gen}\Big(U, G^{\text{ext}}(\varphi(X))\Big) \overset{\text{def}}{=} \sum_{t=1}^{T_{\max}} \sum_{v \in \mathcal{V}} \mathbb{1}\{u_t = v\} \log\Big(p_{v,t}\Big)$, where $p_{v,t}$ is the probability predicted by G^{ext} that the t^{th} word is equal to v. A detailed description of the training step of **GENELIEX** is provided in Algorithm 2, see also section 6.4.1 for an illustrative diagram.

Remark 13. *Note that the proposed method only augments data on the extreme regions. A general data augmentation algorithm can be obtained by combining this approach with any other algorithm on the original input data X whose latent code $Z = \varphi(X_U)$ does not lie in the extreme regions.*

6.4 Models

In this section we provide an overview of the algorithms and the detailed pseudocode.

6.4.1 Models Overview

Figure 6.2 provides an overview of the different algorithms proposed in the chapter. Figure 6.2a describes the pipeline for **LHTR** detailed in Algorithm 1. Figure 6.2b

describes the pipeline for the comparative baseline **LHTR**$_1$ where $C^{\text{ext}} = C^{\text{bulk}}$. Figure 6.2c illustrates the pipeline for the baseline classifier trained on BERT. Figure 6.2d describes **GENELIEX** described in Algorithm 2, note that the hatched components are inherited from **LHTR** and are not used in the workflow.

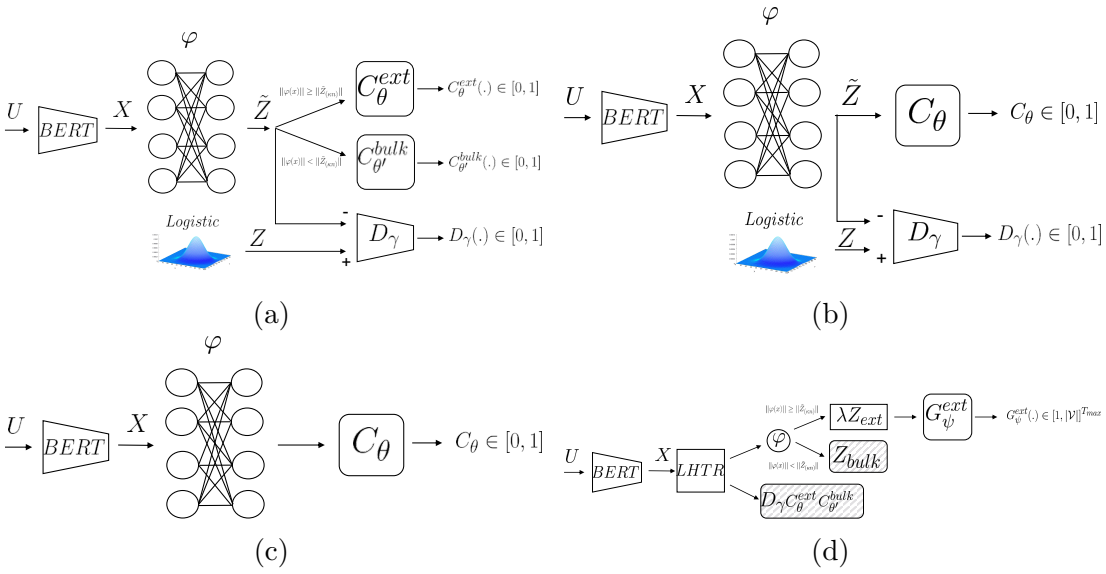


Figure 6.2 – Illustrative pipelines.

Algorithm 2 GENELIEX: training step

INPUT: input of LHTR, $\mathcal{D}_{g_n} = \{U_1, \ldots, U_n\}$
Initialization: parameters of φ_τ, C_θ^{ext}, $C_{\theta'}^{\text{bulk}}$, D_γ and decoder G_ψ^{ext}
Optimization:
φ, C^{ext}, C^{bulk} = LHTR$(\rho_1, \rho_2, \rho_3, \mathcal{D}_n, \kappa, m)$
while ψ not converged **do**
Sample $\{U_1 \ldots, U_m\}$ from the training set $\mathcal{D}_{g_n}$ and define $\tilde{Z}_i = \varphi(X_{U,i})$ for $i \in \{1, \ldots, m\}$.
Sort $\{\tilde{Z}_i\}_{i\in\{1,\ldots,m\}}$ by decreasing order of magnitude $\|\tilde{Z}_{(1)}\| \geq \ldots \geq \|\tilde{Z}_{(m)}\|$.
Update ψ by descending:

$$\mathcal{L}_g^{\text{ext}}(\psi) \stackrel{\text{def}}{=} \frac{\rho_1}{\lfloor \kappa m \rfloor} \sum_{i=1}^{\lfloor \kappa m \rfloor} \ell_{gen.}\Big(U_{(i)}, G_\psi^{\text{ext}}(\tilde{Z}_{(i)})\Big).$$

end while
Compute $\{\tilde{Z}_i\}_{i\in\{1,\ldots,n\}} = \varphi(X_i)_{i\in\{1,\ldots,n\}}$
Sort $\{\tilde{Z}_i\}_{i\in\{1,\ldots,n\}}$ by decreasing order of magnitude $\|\tilde{Z}_{(1)}\| \geq \ldots \|\tilde{Z}_{(k)}\| \geq \ldots \geq \|\tilde{Z}_{(n)}\|$.
OUTPUT: encoder φ, decoder G^{ext} applicable on the region $\{x : \|\varphi(x)\| \geq \|\tilde{Z}_{(\lfloor \kappa n \rfloor)}\|\}$

Algorithm 1 LHTR

INPUT: Weighting coef. $\rho_1, \rho_2, \rho_3 > 0$, Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, batch size m, proportion of extremes κ, heavy tailed prior P_Z.

Initialization: parameters $(\tau, \theta, \theta', \gamma)$ of the encoder φ_τ, classifiers C_θ^{ext}, $C_{\theta'}^{bulk}$ and discriminator D_γ

Optimization:

while $(\tau, \theta, \theta', \gamma)$ not converged **do**

Sample $\{(X_1, Y_1) \ldots, (X_m, Y_m)\}$ from $\mathcal{D}_n$ and define $\tilde{Z}_i = \varphi(X_i)$, $i \leq m$.

Sample $\{Z_1, \ldots, Z_m\}$ from the prior P_Z.

Update γ by ascending:

$$\frac{\rho_3}{m} \sum_{i=1}^{m} \log D_\gamma(Z_i) + \log(1 - D_\gamma(\tilde{Z}_i)).$$

Sort $\{\tilde{Z}_i\}_{i \in \{1,\ldots,m\}}$ by decreasing order of magnitude $||\tilde{Z}_{(1)}|| \geq \ldots \geq ||\tilde{Z}_{(m)}||$.

Update θ by descending:

$$\mathcal{L}^{\text{ext}}(\theta, \tau) \stackrel{\text{def}}{=} \frac{\rho_1}{\lfloor \kappa m \rfloor} \sum_{i=1}^{\lfloor \kappa m \rfloor} \ell\Big(Y_{(i)}, C_\theta^{\text{ext}}(\tilde{Z}_{(i)})\Big).$$

Update θ' by descending:

$$\mathcal{L}^{\text{bulk}}(\theta', \tau) \stackrel{\text{def}}{=} \frac{\rho_2}{m - \lfloor \kappa m \rfloor} \sum_{i=\lfloor \kappa m \rfloor + 1}^{m} \ell\Big(Y_{(i)}, C_{\theta'}^{\text{bulk}}(\tilde{Z}_{(i)})\Big).$$

Update τ by descending:

$$\frac{1}{m} \sum_{i=1}^{m} -\rho_3 \log D_\gamma(\tilde{Z}_i) + \mathcal{L}^{\text{ext}}(\theta, \tau) + \mathcal{L}^{\text{bulk}}(\theta', \tau).$$

end while

Compute $\{\tilde{Z}_i\}_{i \in \{1,\ldots,n\}} = \varphi(X_i)_{i \in \{1,\ldots,n\}}$

Sort $\{\tilde{Z}_i\}_{i \in \{1,\ldots,n\}}$ by decreasing order of magnitude $||\tilde{Z}_{(1)}|| \geq \ldots ||\tilde{Z}_{(\lfloor \kappa n \rfloor)}|| \geq \ldots \geq ||\tilde{Z}_{(n)}||$.

OUTPUT: encoder φ, classifiers C^{ext} for $\{x : ||\varphi(x)|| \geq t := ||\tilde{Z}_{(\lfloor \kappa n \rfloor)}||\}$ and C^{bulk} on the complementary set.

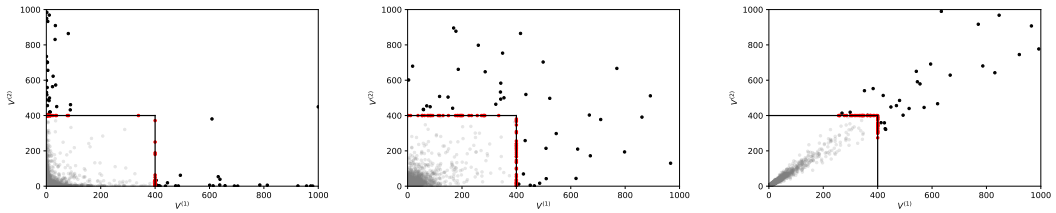
6.5 Experiments : Classification

In our experiments we work with the infinity norm. The proportion of extreme samples in the training step of **LHTR** is chosen as $\kappa = 1/4$. The threshold t defining the extreme region $\{\|x\| > t\}$ in the test set is $t = \|\tilde{Z}_{(\lfloor \kappa n \rfloor)}\|$ as returned by **LHTR**. We denote by $\mathcal{T}_{\text{test}}$ and $\mathcal{T}_{\text{train}}$ respectively the extreme test and train sets thus defined. Classifiers $C^{\text{bulk}}, C^{\text{ext}}$ involved in **LHTR** are Multi Layer Perceptrons (MLP), see Section 6.9.2 for a full description of the architectures.

Heavy-tailed distribution. The regularly varying target distribution is chosen as a multivariate logistic distribution with parameter $\delta = 0.9$, refer to Section 6.5.1 for details and an illustration with various values of δ. This distribution is widely used in the context of extreme values analysis Chiapino and collab. (2019b); Goix and collab. (2016); Thomas and collab. (2017) and differ from the classical logistic distribution.

6.5.1 Logistic distribution

The logistic distribution with dependence parameter $\delta \in (0, 1]$ is defined in $\mathbb{R}^d$ by its *c.d.f.* $F(x) = \exp\left\{-(\sum_{j=1}^{d} x^{(j)\frac{1}{\delta}})^{\delta}\right\}$. Samples from the logistic distribution can be simulated according to the algorithm proposed in Stephenson (2003). Figure 6.3 illustrates this distribution with various values of δ. Values of δ close to 1 yield non concomitant extremes, *i.e.* the probability of a simultaneous excess of a high threshold by more than one vector component is negligible. Conversely, for small values of δ, extreme values tend to occur simultaneously. These two distinct tail dependence structures are respectively called 'asymptotic independence' and 'asymptotic dependence' in the EVT terminology.



(a) near tail independence (b) moderate tail dependence (c) high tail dependence

Figure 6.3 – Illustration of the distribution of the angle $\Theta(X)$ obtained with bivariate samples X generated from a logistic model with different coefficients of dependence ranging from near asymptotic independence Figure 6.3a ($\delta = 0.9$) to high asymptotic dependence Figure 6.3c ($\delta = 0.1$) including moderate dependence Figure 6.3b ($\delta = 0.5$). Non extreme samples are plotted in gray, extreme samples are plotted in black and the angles $\Theta(X)$ (extreme samples projected on the sup norm sphere) are plotted in red. Note that not all extremes are shown since the plot was truncated for a better visualization. However all projections on the sphere are shown.

6.5.2 Toy example: about LHTR

We start with a simple bivariate illustration of the heavy tailed representation learnt by **LHTR**. Our goal is to provide insight on how the learnt mapping φ acts on the input space and how the transformation affects the definition of extremes (recall that extreme samples are defined as those samples which norm exceeds an empirical quantile). Labeled samples are simulated from a Gaussian mixture distribution

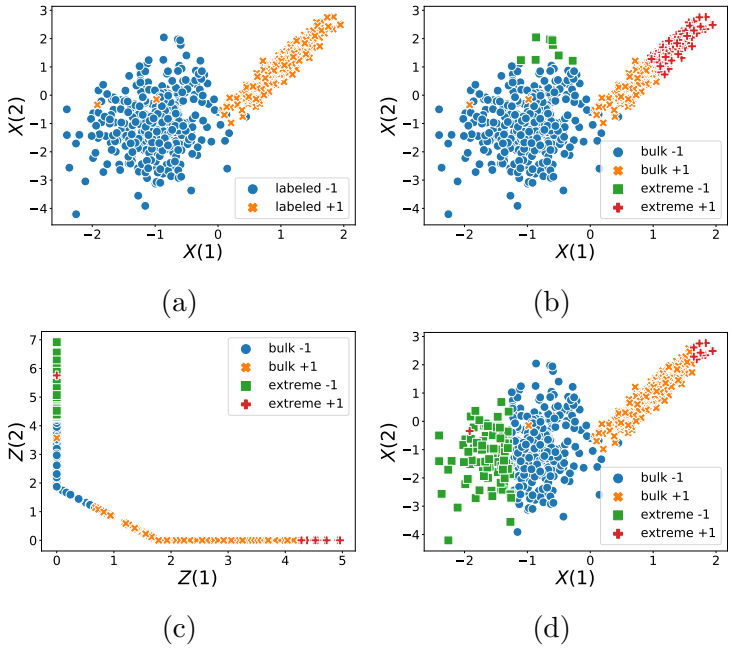


Figure 6.4 – Figure 6.4a: Bivariate samples X_i in the input space. Figure 6.4b: X_i's in the input space with extremes from each class selected in the input space. Figure 6.4c: Latent space representation $Z_i = \varphi(X_i)$. Extremes of each class are selected in the latent space. Figure 6.4d: X_i's in the input space with extremes from each class selected in the latent space.

with two components of identical weight. The label indicates the component from which the point is generated. **LHTR** is trained on 2250 examples and a testing set of size 750 is shown in Figure 6.4. The testing samples in the input space (Figure 6.4a) are mapped onto the latent space *via* φ (Figure 6.4c) In Figure 6.4b, the extreme raw observations are selected according to their norm after a component-wise standardisation of X_i, Indeed, in Figure 6.4b selecting the extreme samples on the input space is not a straightforward step as the two components of the vector are not on the same scale, as introduced in Chapter 2, componentwise standardisation is a natural and necessary preliminary step. Following common practice in multivariate extreme value analysis it was decided to standardise the input data $(X_i)_{i\in\{1,\dots,n\}}$ by applying the rank-transformation:

$$\widehat{T}(x) = \left(1/\left(1-\widehat{F}_j(x)\right)\right)_{j=1,\dots,d}$$

for all $x = (x^1, \dots, x^d) \in \mathbb{R}^d$ where $\widehat{F}_j(x) \overset{\text{def}}{=} \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{X_i^j \leq x\}$ is the j^{th} empirical marginal distribution. Denoting by V_i the standardized variables, $\forall i \in$

$\{1, \dots, n\}, V_i = \widehat{T}(X_i)$. The marginal distributions of V_i are well approximated by standard Pareto distribution, the approximation error comes from the fact that the empirical *c.d.f*'s are used in $\widehat{T}$ instead of the genuine marginal *c.d.f.*'s F_j. After this standardization step, the selected extreme samples are $\{V_i, \|V_i\| \geq V_{(\lfloor \kappa n \rfloor)}\}$.

6.5.3 Enforcing regularity assumptions in Theorem 10

The methodology in the present chapter consists in learning a representation Z for text data *via* **LHTR** satisfying the regular variation condition (6.2). This condition is weaker than the assumptions from Theorem 10 for two reasons: first, it does not imply that each class (conditionally to the label Y) is regularly varying, only that the distribution of Z (unconditionally to the label) is. Second, in Chapter 5, it is additionally required that the regression function $\eta(z) = \mathbb{P}\{Y = +1 \mid Z = z\}$ converges uniformly as $\|z\| \to \infty$. Getting into details, one needs to introduce a limit random pair (Z_∞, Y_∞) which distribution is the limit of $\mathbb{P}\{Y = \cdot, t^{-1}Z \in \cdot \mid \|Z\| > t\}$ as $t \to \infty$. Denote by η_∞ the limiting regression function, $\eta_\infty(z) = \mathbb{P}\{Y_\infty = +1 \mid Z_\infty = z\}$. The required assumption is that

$$\sup_{\{z \in \mathbb{R}^d_+ : \|z\| > t\}} \left|\eta(z) - \eta_\infty(z)\right| \xrightarrow[t \to \infty]{} 0. \tag{6.4}$$

Uniform convergence (6.4) is not enforced in **LHTR** and the question of how to enforce it together with regular variation of each class separately remains open. However, our experiments in sections 6.5 and 6.6 demonstrate that enforcing Condition (6.2) is enough for our purposes, namely improved classification and label preserving data augmentation.

The extreme threshold t is chosen as the 75% empirical quantile of the norm on the training set in the input space. Notice in the latter figure the class imbalance among extremes. In Figure 6.4c, extremes are selected as the 25% samples with the largest norm in the latent space. Figure 6.4d is similar to Figure 6.4b except for the selection of extremes which is performed in the latent space as in Figure 6.4c. On this toy example, the adversarial strategy appears to succeed in learning a code which distribution is close to the logistic target, as illustrated by the similarity between Figure 6.4c and Figure 6.3a. In addition, the heavy tailed representation allows a more balanced selection of extremes than the input representation.

6.5.4 Application to positive *vs.* negative classification of sequences

In this section, we dissect **LHTR** to better understand the relative importance of: *(i)* working with a heavy-tailed representation, *(ii)* training two independent classifiers: one dedicated to the bulk and the second one dedicated to the extremes. In addition, we verify experimentally that the latter classifier is scale invariant, which is neither the case for the former, nor for a classifier trained on BERT input.

Experimental settings. We compare the performance of three models. The baseline **NN model** is a MLP trained on BERT. The second model $\textbf{LHTR}_1$ is a variant of **LHTR** where a single MLP (C) is trained on the output of the encoder φ, using all the available data, both extreme and non extreme ones. The third model

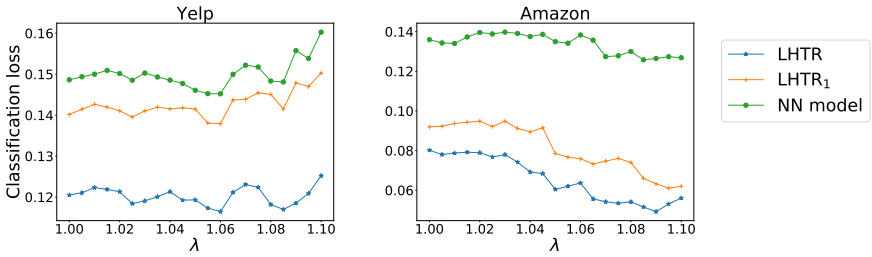


Figure 6.5 – Classification loss of **LHTR**, **LHTR**$_1$ and **NN model** on the extreme test set $\{x \in \mathcal{T}, ||x|| \geq \lambda t\}$ for increasing values of λ (X-axis), on *Yelp* and *Amazon*.

(**LHTR**) trains two separate MLP classifiers C^{ext} and C^{bulk} respectively dedicated to the extreme and bulk regions of the learnt representation φ. All models take the same training inputs, use BERT embedding and their classifiers have identical structure, see Sections 6.4.1 and 6.9.2 for a summary of model workflows and additional details concerning the network architectures.

Comparing **LHTR**$_1$ with **NN model** assesses the relevance of working with heavy-tailed embeddings. Since **LHTR**$_1$ is obtained by using **LHTR** with $C^{\text{ext}} = C^{\text{bulk}}$, comparing **LHTR**$_1$ with **LHTR** validates the use of two separate classifiers so that extremes are handled in a specific manner. As we make no claim concerning the usefulness of **LHTR** in the bulk, at the prediction step we suggest working with a combination of two models: **LHTR** with C^{ext} for extreme samples and any other off-the-shelf ML tool for the remaining samples (*e.g.* **NN model**).

Datasets. In our experiments we rely on two large datasets from *Amazon* (231k reviews) McAuley and Leskovec (2013) and from *Yelp* (1,450k reviews) Liu and collab. (2015); Yu and collab. (2014). Reviews, (made of multiple sentences) with a rating greater than or equal to $^4/_5$ are labeled as $+1$, while those with a rating smaller or equal to $^2/_5$ are labeled as -1. The gap in reviews' ratings is designed to avoid any overlap between labels of different contents.

Results. Figure 6.5 gathers the results obtained by the three considered classifiers on the tail regions of the two datasets mentioned above. To illustrate the generalization ability of the proposed classifier in the extreme regions we consider nested subsets of the extreme test set $\mathcal{T}_{\text{test}}$, $\mathcal{T}^\lambda = \{z \in \mathcal{T}_{\text{test}}, \|z\| \geq \lambda t\}$, $\lambda \geq 1$. For all factor $\lambda \geq 1$, $\mathcal{T}^\lambda \subseteq \mathcal{T}_{\text{test}}$. The greater λ, the fewer the samples retained for evaluation and the greater their norms. On both datasets, **LHTR**$_1$ outperforms the baseline **NN model**. This shows the improvement offered by the heavy-tailed embedding on the extreme region. In addition, **LHTR**$_1$ is in turn largely outperformed by the classifier **LHTR**, which proves the importance of working with two separate classifiers. The performance of the proposed model respectively on the bulk region, tail region and overall, is reported in Table 6.1, which shows that using a specific classifier dedicated to extremes improves the overall performance.

Scale invariance. On all datasets, the extreme classifier g^{ext} verifies Equation (6.1) for each sample of the test set, $g^{\text{ext}}(\lambda Z) = g^{\text{ext}}(Z)$ with λ ranging from 1 to 20, demonstrating scale invariance of g^{ext} on the extreme region. The same experiments conducted both with **NN model** and a MLP classifier trained on BERT and **LHTR**$_1$ show label changes for varying values of λ: none of them are scale invariant. Section 6.9.1 gathers additional experimental details. The scale

Model	*Amazon*			*Yelp*		
	Bulk	Extreme	Overall	Bulk	Extreme	Overall
NN model	**0.085**	0.135	0.098	**0.098**	0.148	0.111
LHTR$_1$	0.104	0.091	0.101	0.160	0.139	0.155
LHTR	0.105	**0.08**	0.0988	0.162	**0.1205**	0.152
Proposed Model	**0.085**	**0.08**	**0,084**	**0.097**	**0.1205**	**0.103**

Table 6.1 – Classification losses on *Amazon* and *Yelp*. 'Proposed Model' results from using **NN model** model for the bulk and **LHTR** for the extreme test sets. The extreme region contains 6.9k samples for *Amazon* and 6.1k samples for *Yelp*, both corresponding roughly to 25% of the whole test set size.

invariance property will be exploited in the next section to perform label invariant generation.

6.6 Experiments : Label Invariant Generation

6.6.1 Experimental Setting

Comparison with existing work. We compare **GENELIEX** with two state of the art methods for dataset augmentation, Wei and Zou (2019) and Kobayashi (2018). Contrarily to these works which use heuristics and a synonym dictionary, **GENELIEX** does not require any linguistic resource. To ensure that the improvement brought by **GENELIEX** is not only due to BERT, we have updated the method in Kobayashi (2018) with a BERT language model (see Section 6.9.3 for details and Table 6.8 for hyperparameters).

Evaluation Metrics. Automatic evaluation of generative models for text is still an open research problem. We rely both on perceptive evaluation and automatic measures to evaluate our model through four criteria (**C1**, **C2**, **C3**,**C4**). **C1** measures Cohesion Crossley and McNamara (2010) (*Are the generated sequences grammatically and semantically consistent?*). **C2** (named Sent. in Table 6.3) evaluates label conservation (*Does the expressed sentiment in the generated sequence match the sentiment of the input sequence?*). **C3** measures the diversity Li and collab. (2015) (corresponding to dist1 or dist2 in Table 6.3[1]) of the sequences (*Does the augmented dataset contain diverse sequences?*). Augmenting the training set with very diverse sequences can lead to better classification performance. **C4** measures the improvement in terms of F1 score when training a classifier (fastText Joulin and collab. (2016)) on the augmented training set (*Does the augmented dataset improve classification performance?*).

Datasets. **GENELIEX** is evaluated on two datasets, a medium and a large one (see Silfverberg and collab. (2017)) which respectively contains 1k and 10k labeled samples. In both cases, we have access to $\mathcal{D}_{g_n}$ a dataset of 80k unlabeled samples. Datasets are randomly sampled from *Amazon* and *Yelp*.

Experiment description. We augment extreme regions of each dataset according to three algorithms: **GENELIEX** (with scaling factor λ ranging from 1 to 1.5),

[1]dist n is obtained by calculating the number of distinct n-grams divided by the total number of generated tokens to avoid favoring long sequences.

Model	*Amazon*				*Yelp*			
	Medium		Large		Medium		Large	
	F1	dist1/dist2	F1	dist1/dist2	F1	dist1/dist2	F1	dist1/dist2
Raw Data	84.0	X	93.3	X	86.7	X	94.1	X
Kobayashi (2018)	85.0	0.10/0.47	92.9	0.14/0.53	87.0	0.15/0.53	94.0	0.14/0.58
Wei and Zou (2019)	85.2	0.11/0.50	93.2	0.14/0.54	87.0	0.15/0.52	**94.2**	0.16/0.59
GENELIEX	**86.3**	**0.14/0.52**	**94.0**	**0.18/0.58**	**88.4**	**0.18/0.62**	**94.2**	**0.16/0.60**

Table 6.2 – Quantitative Evaluation. Algorithms are compared according to **C3** and **C4**. dist1 and dist2 respectively stand for distinct 1 and 2, it measures the diversity of new sequences in terms of unigrams and bigrams. F1 is the F1-score for FastText classifier trained on an augmented labelled training set.

Model	*Amazon*		*Yelp*	
	Sent.	Cohesion	Sent.	Cohesion
Raw Data	83.6	78.3	80.6	0.71
Kobayashi (2018)	**80.0**	**84.2**	82.9	0.72
Wei and Zou (2019)	69.0	67.4	80.0	0.60
GENELIEX	78.4	73.2	**85.7**	**0.77**

Table 6.3 – Qualitative evaluation with three turkers. Sent. stands for sentiment label preservation. The Krippendorff Alpha for Amazon is $\alpha = 0.28$ on the sentiment classification and $\alpha = 0.20$ for cohesion. The Krippendorff Alpha for Yelp is $\alpha = 0.57$ on the sentiment classification and $\alpha = 0.48$ for cohesion.

Kobayashi (2018), and Wei and Zou (2019). For each train set's sequence considered as extreme, 10 new sequences are generated using each algorithm. Section 6.9.3 gathers further details. For experiment **C4** the test set contains 10^4 sequences.

6.6.2 Results

Automatic measures. The results of **C3** and **C4** evaluation are reported in Table 6.2. Augmented data with **GENELIEX** are more diverse than the one augmented with Kobayashi (2018) and Wei and Zou (2019). The F1-score with dataset augmentation performed by **GENELIEX** outperforms the aforementioned methods on Amazon in medium and large dataset and on Yelp for the medium dataset. It equals state of the art performances on Yelp for the large dataset. As expected, for all three algorithms, the benefits of data augmentation decrease as the original training dataset size increases. Interestingly, we observe a strong correlation between more diverse sequences in the extreme regions and higher F1 score: the more diverse the augmented dataset, the higher the F1 score. More diverse sequences are thus more likely to lead to better improvement on downstream tasks (*e.g.* classification).

Perceptive Measures. To evaluate **C1**, **C2**, three turkers were asked to annotate the cohesion and the sentiment of 100 generated sequences for each algorithm and for the raw data. F1 scores of this evaluation are reported in Table 6.3. Grammar evaluation confirms the findings of Wei and Zou (2019) showing that random swaps and deletions do not always maintain the cohesion of the sequence.

In contrast, **GENELIEX** and Kobayashi (2018), using vectorial representations, produce more coherent sequences. Concerning sentiment label preservation, on Yelp, **GENELIEX** achieves the highest score which confirms the observed improvement reported in Table 6.2. On Amazon, turker annotations with data from **GENELIEX** obtain a lower F1-score than from Kobayashi (2018). This does not correlate with results in Table 6.2 and may be explained by a lower Krippendorff Alpha[2] on Amazon ($\alpha = 0.20$) than on Yelp ($\alpha = 0.57$).

Influence of the scaling factor on the linguistic content

Table 6.4 gathers some extreme sequences generated by **GENELIEX** for λ ranging from 1 to 1.5. No major linguistic change appears when λ varies. The generated sequences are grammatically correct and share the same polarity (positive or negative sentiment) as the input sequence. Note that for greater values of λ, a repetition phenomenon appears. The resulting sequences keep the label and polarity of the input sequence but repeat some words Holtzman and collab. (2019).

[2]measure of inter-rater reliability in $[0, 1]$: 0 is perfect disagreement and 1 is perfect agreement.

Input	very sloppy and slow service. when we arrived they told us to sit anywhere but all the tables were still dirty and haven't been cleaned. they didn't bother to ask if we wanted refills on our drinks. we needed an extra plate and didn't get one so my nephew decides to go up to the counter and ask for one because he's hungry. they gave our check when we were still eating. the list can go on and on. i wouldn't recommend this place. go somewhere else for faster and better service. very disappointed
$\lambda = 1.1$	very sloppy and sluggish service. when we got there, they told us to sit anywhere but all the tables were empyt full of dishes and were not cleaned at all. they didn't bother to ask if our drinks would be added. we needed an extra dish and didn't get one, so my cousin decided to go to the counter and ask one because he's hungry. they were going to watch while we were still eating. the list could go on and on. i would not recommend this place. go elsewhere for faster and better service. very very disappointed
$\lambda = 1.2$	services and survivors. when he got there, he told us we were sitting everywhere but all the tables were full of dishes and we didn't wash everything. he never bothered to ask if our drinks would be added. we needed extra food and didn't get one, so my brother decided to go to the locker and ask because he was thirsty. they want to watch it while we eat. the list can be continuous and active. i would not recommend this place. go elsewhere for faster and better service. very disappointed
$\lambda = 1.3$	services and survivors. when he got there, he told us that we were sitting everywhere, but all the tables were full of dishes and we didn't wash everything. he never bothered to ask if our drinks would be added. We needed more food and we didn't get it, so my brother decided to go to the locker and ask because he was thirsty. they want to watch it when we eat. the list can be continuous and active. i would not recommend this place. go faster and faster for better service. very disappointed
Input	visited today with my husband. we were in the firearms section. there were 3 employees in attendance with one customer. my husband ask a question and was ignored. he waited around for another 10 minutes or so. if it had been busy i could understand not receiving help. we left and went elsewhere for our purchases.
$\lambda = 1.1$	visited today with my husband. we were in the firearms section. together with one customer there were 3 employees. my husband asked and was ignored. waited about another 10 minutes. if it was busy, i would understand that i wouldn't get help. we left and went somewhere else because of our purchases.
$\lambda = 1.2$	today she visited with her husband. we were in the gun department. there were 3 employees together with one customer. my husband asked and was ignored. waited another 10 minutes. if he was busy, i would understand that i would not receive help. we went and went somewhere else because of our shopping.
$\lambda = 1.3$	today, she went with her husband. we are in the gun department. there are 3 employees and one customer. my husband rejected me and ignored him. wait another minute. if he has a job at hand, i will understand that i will not get help. we went somewhere else because of our business.

Input	walked in on a friday and got right in. it was exactly what i expected for a thai massage. the man did a terrific job. he was very skilled, working on the parts of my body with the most tension and adjusting pressure as i needed throughout the massage. i walked out feeling fantastic and google eyed.
$\lambda = 1.1$	walked in on a friday and got right in. it was exactly what i expected for a thai massage. the man did a terrific job. he was very skilled, working on the parts of my body with the most tension and adjusting pressure as needed throughout the massage. i walked out feeling fantastic and google eyed.
$\lambda = 1.2$	climb up the stairs and get in. the event that i was expecting a thai massage. the man did a wonderful job. he was very skilled, dealing with a lot of stress and stress on my body parts. i walked out feeling lightly happy and tired.
$\lambda = 1.3$	go up and up. this was the event i was expecting a thai massage. the man did a wonderful job. what this was was an expert, with a lot of stress and stress on my body parts. i walked out feeling lightly happy and tired.
Input	i came here four times during a 3 - day stay in madison. the first two was while i was working - from - home. this place is awesome to plug in, work away at a table, and enjoy a great variety of coffee. the other two times, i brought people who wanted good coffee, and this place delivered. awesome atmosphere. awesome awesome awesome.
$\lambda = 1.1$	i came here four times during a 3-day stay in henderson. the first two were while i was working - from home. this place is great for hanging out, working at tables and enjoying the best variety of coffee. the other two times, i brought in people who wanted a good coffee, and it delivered a place. better environment. really awesome awesome.
$\lambda = 1.2$	i came here four times during my 3 days in the city of henderson. the first two were while i was working - at home. this place is great for trying, working tables and enjoying the best variety of coffee. the other two times, i brought people who wanted good coffee, and it brought me somewhere. good environment. really amazing.
$\lambda = 1.3$	i came here four times during my 3 days in the city of henderson. the first two are when i'm working - at home. this place is great for trying, working tables and enjoying a variety of the best coffees. the other two times, i bring people who want good coffee, and that brings me somewhere. good environment. very amazing.

Table 6.4 – Sequences generated by **GENELIEX** for extreme embeddings implying label (sentiment polarity) invariance for generated Sequence. λ is the scale factor. Two first reviews are negatives, two last reviews are positive.

6.7 Experiments : Extremes in Text

6.7.1 Aim of the experiments

The aim of this section is double: first, to provide some intuition on what characterizes sequences falling in the extreme region of **LHTR**. Second, to investigate the hypothesis that extremes from **LHTR** are input sequences which tend to be harder to model than non extreme ones

Regarding the first aim (*(i) Are there interpretable text features correlated with the extreme nature of a text sample?*, since we characterize extremes by their norm in **LHTR** representation, in practice the question boils down to finding text features which are positively correlated with the norm of the text samples in **LHTR**, which we denote by $\|\varphi(X)\|$ and referred to as the '**LHTR** norm' in the sequel. Preliminary investigations did not reveal semantic features (related to the meaning or the sentiment expressed in the sequence) displaying such correlation. However we have identified two features which are positively correlated both together and with the norm in **LHTR**, namely the sequence length $|U|$ as measured by the number of tokens of the input (recall that in our case an input sequence U is a review composed of multiple sequences), and the norm of the input in BERT representation ('BERT norm', denoted by $\|X\|$).

As for the second question (*(ii) Are **LHTR**'s extremes harder to model?*) we consider the next token prediction loss Bengio and collab. (2003) ('LM loss' in the sequel) obtained by training a language model on top of BERT. The next token prediction loss can be seen as a measure of hardness to model the input sequence. The question is thus to determine whether this prediction loss is correlated with the norm in **LHTR** (or in BERT, or with the sequence length).

6.7.2 Results

Figure 6.6 displays pairwise scatterplots for the four considered variables on *Yelp* dataset (left) and *Amazon* dataset (right). These scatterplot suggest strong dependence for all pairs of variables. For a more quantitative assessment, Figure 6.7 displays the correlation matrices between the four quantities $\|\varphi(X)\|$, $\|X\|$, $|U|$ and 'LM Loss' described above on Amazon and Yelp datasets. Pearson and Spearman two-sided correlation tests are performed on all pairs of variables, both tests having as null hypothesis that the correlation between two variables is zero. For all tests, p-values are smaller than 10^{-16}, therefore null hypotheses are rejected for all pairs.

These results prove that the four considered variables are indeed significantly positively correlated, which answers questions (i) and (ii) above.

Figure 6.8 provides additional insight about the magnitude of the shift in sequence length between extremes in the **LHTR** representation and non extreme samples. Even though the histograms overlap (so that two different sequences of same length may be regarded as extreme or not depending on other factors that are not understood yet), there is a visible shift in distribution for both *Yelp* and *Amazon* datasets, both for the positive and negative class in the classification framework for sentiment analysis. Kolmogorov-Smirnoff tests between the length distributions of the two considered classes for each label were performed, which

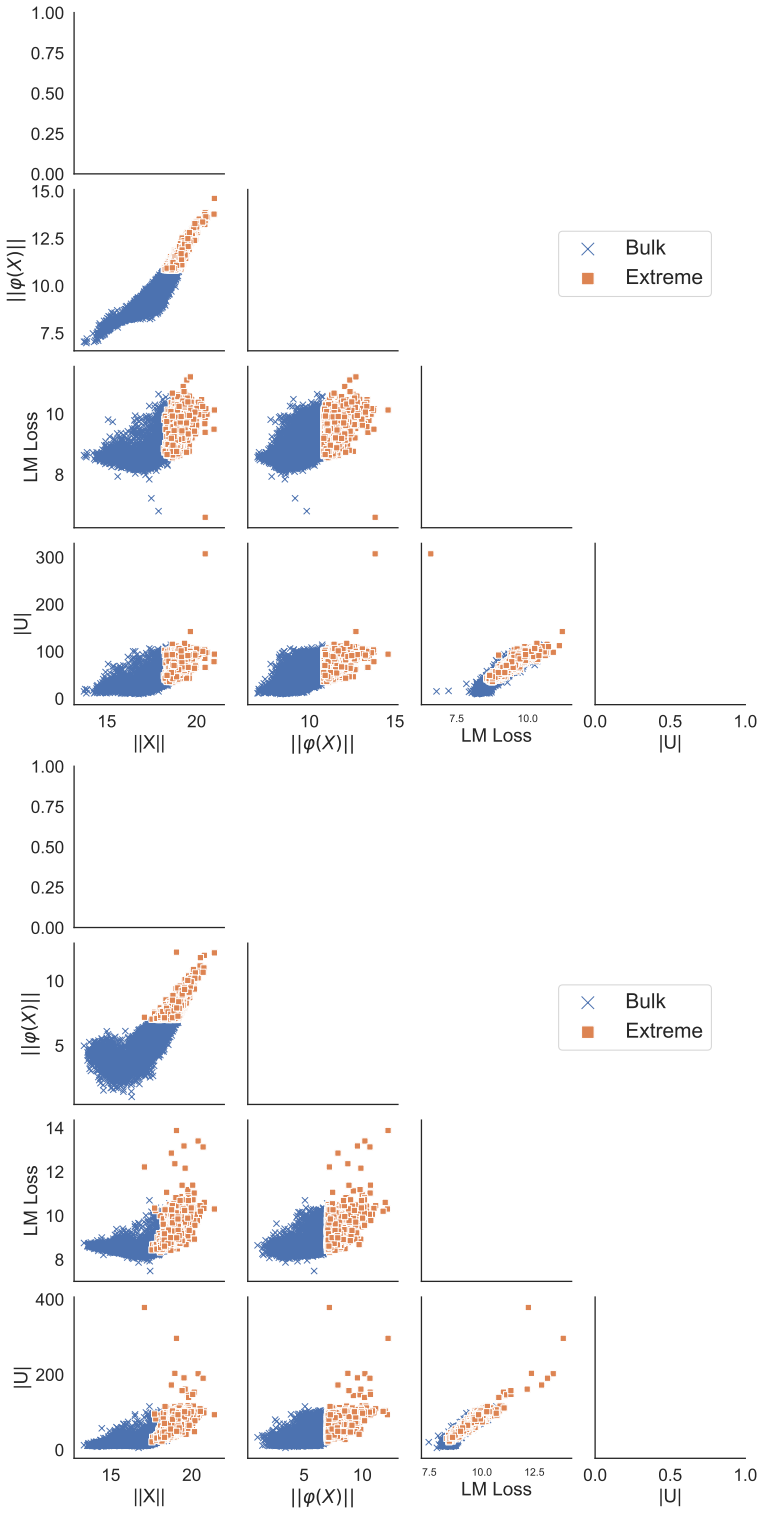


Figure 6.6 – Scatterplots of the four variables 'BERT norm', '**LHTR** norm', 'LM loss' and 'sequence length' on *Yelp* dataset (top) and *Amazon* dataset (bottom).

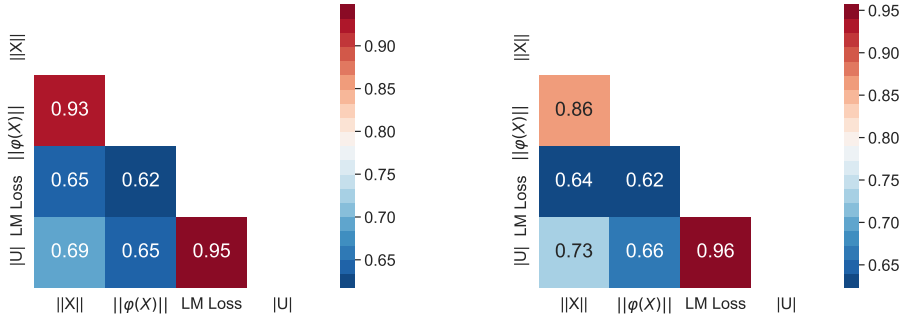
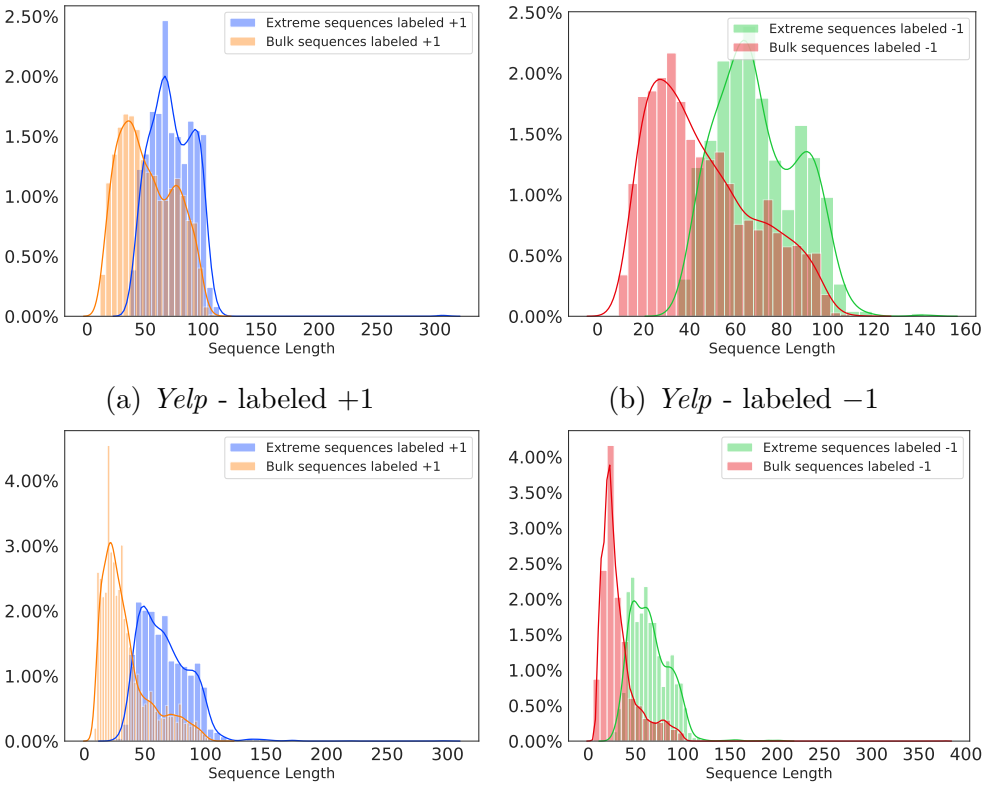


Figure 6.7 – Non diagonal entries of the correlation matrices of the four variables 'BERT norm', '**LHTR** norm', 'LM loss' and 'sequence length' for *Yelp* dataset (left) and *Amazon* dataset (right).

allows us to reject the null hypothesis of equality between distributions, as the maximum p-values is less than 0.05.



(a) *Yelp* - labeled +1

(b) *Yelp* - labeled −1

(c) *Amazon* - labeled +1

(d) *Amazon* - labeled −1

Figure 6.8 – Histograms of the samples' sequence length for *Yelp* dataset (Figure 6.8a and Figure 6.8b) and *Amazon* (Figure 6.8c and Figure 6.8d). The number of sequences in the bulk is approximately 3 times the number of extreme sequences for each dataset 10000 sequences are considered and extreme region contains approximately 3000 sequences .

6.7.3 Experimental conclusions

We summarize the empirical findings of this section:

1. An 'extreme' text sequence in **LHTR** representation is more likely to have a greater length (number of tokens) than a non extreme one.

2. Positive correlation between the BERT norm and the **LHTR** norm indicates that a large sample in the BERT representation is likely to have a large norm in the **LHTR** representation as well: the learnt representation **LHTR** taking BERT as input keeps invariant (in probability) the ordering implied by the norm.

3. A consequence of the two above points is that long sequences tend to have a large norm in BERT.

4. Extreme text samples (regarding the BERT norm or the **LHTR** norm) tend to be harder to model than non-extreme ones.

5. Since extreme texts are harder to model and also somewhat harder to classify in view of the BERT classification scores reported in Table 6.1, there is room for improvement in their analysis and it is no wonder that a method dedicated to extremes *i.e.* relying on EVT such as **LHTR** outperforms the baseline.

6.8 Conclusion

With the deluge of text data resulting from the boom in social networks or streaming platforms and online shopping websites, analysis of natural language processing veered towards various embeddings which showed to be effective ways to achieve state-of-the art results on written benchmarks.The approach promoted in this chapter relies on learning a regularly varying representation by minimizing an empirical proxy of an objective function. The latter writes as a weighted sum of a classification risk and a regularization term penalizing the distance between the representation distribution and a heavy-tailed target. Our experiments show that the obtained representation allows to diminish the classification error compared to models with comparable complexity.The obtained representation used for a text data augmentation task, is competitive with existing data augmentation methods. The attribute invariance under dilation is the key to generate meaningful sentences with prescribed attribute.

6.9 Further experimental material

6.9.1 Scale invariance comparison of BERT and LHTR

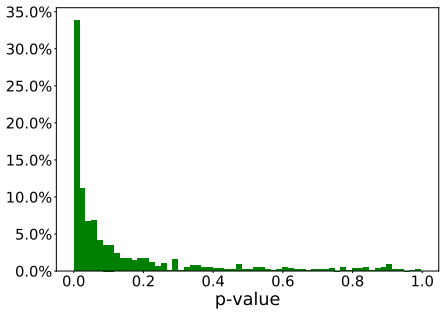
In this section, we compare **LHTR** and BERT and show that the latter is not scale invariant. For this preliminary experiment we rely on labeled fractions of both *Amazon* and *Yelp* datasets respectively denoted as *Amazon small dataset* and *Yelp small dataset* detailed in Kotzias and collab. (2015), each of them containing 1000 sequences from the large dataset. Both datasets are divided at random in a train set $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$. The train set represents $3/4$ of the whole dataset while the

remaining samples represent the test set. We use the hyperparameters reported in Table 6.5.

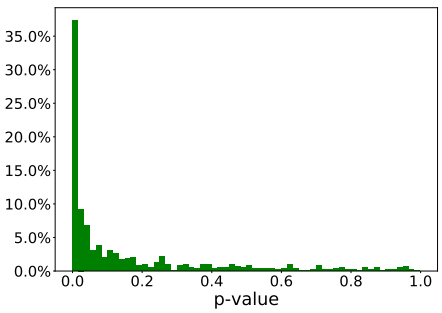
	NN model	**LHTR**$_1$	**LHTR**
Sizes of the layers φ	[768,384,200,50,8,1]	[768,384,200,100]	[768,384,200,150]
Sizes of the layers $C_{\theta'}^{bulk}$	X	[100,50,8,1]	[150,75,8,1]
Sizes of the layers C_{θ}^{ext}	X	X	[150,75,8,1]
ρ_3	X	X	0.001

Table 6.5 – Network architectures for *Amazon small dataset* and *Yelp small dataset* . The weight decay is set to 10^5, the learning rate is set to $5 * 10^{-4}$, the number of epochs is set to 500 and the batch size is set to 64.

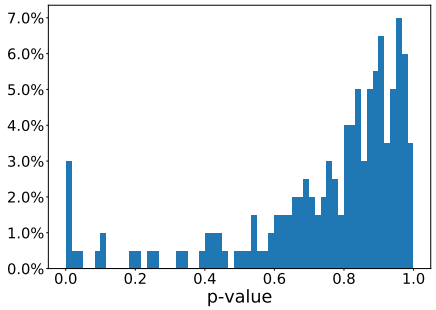
BERT is not regularly varying. In order to show that X is not regularly varying, independence between $\|X\|$ and a margin of $\Theta(X)$ can be tested Coles and Tawn (1994), which is easily done *via* correlation tests. Pearson correlation tests were run on the extreme samples of BERT and **LHTR** embeddings of *Amazon small dataset* and *Yelp small dataset*. The statistical tests were performed between all margins of $\left(\Theta(X_i)\right)_{1\geq i\geq n}$ and $\left(\|X_i\|\right)_{1\geq i\geq n}$.



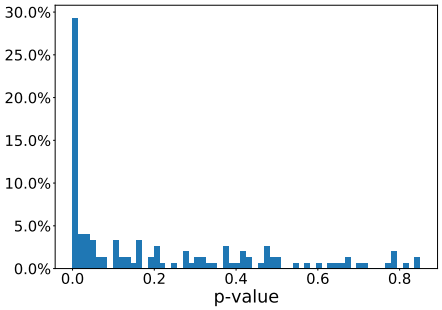
(a) *Yelp small dataset* - BERT



(b) *Amazon small dataset* - BERT



(c) *Yelp small dataset* - **LHTR**



(d) *Amazon small dataset* - **LHTR**

Figure 6.9 – Histograms of the p-values for the non-correlation test between $\left(\Theta(X_i)\right)_{1\geq i\geq n}$ and $\left(\|X_i\|\right)_{1\geq i\geq n}$ on embeddings provided by BERT (Figure 6.9a and Figure 6.9b) or **LHTR** (Figure 6.9c and Figure 6.9d).

Each histogram in Figure 6.9 displays the distribution of the p-values of the correlation tests between the margins X_j and the angle $\Theta(X)$ for $j \in \{1, \dots d\}$, in a

given representation (BERT or **LHTR**) for a given dataset. For both *Amazon small dataset* and *Yelp small dataset* the distribution of the p-values is shifted towards larger values in the representation of **LHTR** than in BERT, which means that the correlations are weaker in the former representation than in the latter. This phenomenon is more pronounced with *Yelp small dataset* than with *Amazon small dataset*. Thus, in BERT representation, even the largest data points exhibit a non negligible correlation between the radius and the angle and the regular variation condition does not seem to be satisfied. As a consequence, in a classification setup such as binary sentiment analysis detailed in Section 6.5.4), classifiers trained on BERT embedding are not guaranteed to be scale invariant. In other words for a representation X of a sequence U with a given label Y, the predicted label $g(\lambda X)$ is not necessarily constant for varying values of $\lambda \geq 1$. Figure 6.10 illustrates this fact on a particular example taken from *Yelp small dataset*. The color (white or black respectively) indicates the predicted class (respectively -1 and $+1$). For values of λ close to 1, the predicted class is -1 but the prediction shifts to class $+1$ for larger values of λ.

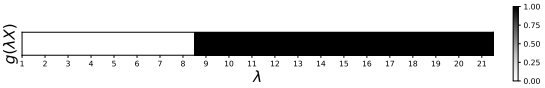
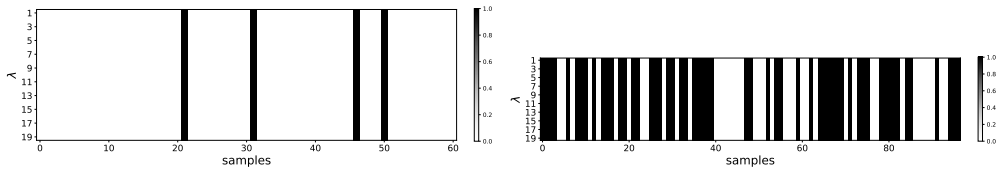


Figure 6.10 – Lack of scale invariance of the classifier trained on BERT: evolution of the predicted label $g(\lambda X)$ from -1 to $+1$ for increasing values of λ, for one particular example X.

Scale invariance of LHTR. We provide here experimental evidence that **LHTR**'s classifier g^{ext} is scale invariant (as defined in Equation (6.1)). Figure 6.11 displays the predictions $g^{\text{ext}}(\lambda Z_i)$ for increasing values of the scale factor $\lambda \geq 1$ and Z_i belonging to $\mathcal{T}_{\text{test}}$, the set of samples considered as extreme in the learnt representation. For any such sample Z, the predicted label remains constant as λ varies, *i.e.* it is scale invariant, $g^{\text{ext}}(\lambda Z) = g^{\text{ext}}(Z)$, for all $\lambda \geq 1$.



(a) *Amazon small dataset* (b) *Yelp small dataset*

Figure 6.11 – Scale invariance of g^{ext} trained on LHTR: evolution of the predicted label $g^{\text{ext}}(\lambda Z_i)$ (white or black for $-1/+1$) for increasing values of λ, for samples Z_i from the extreme test set $\mathcal{T}_{\text{test}}$ from *Amazon small dataset* (Figure 6.11a) and *Yelp small dataset* (Figure 6.11b).

6.9.2 Experimental settings (Classification): additional details

Toy example. For the toy example, we generate 3000 points distributed as a mixture of two normal distributions in dimension two. For training **LHTR**, the

number of epochs is set to 100 with a dropout rate equal to 0.4, a batch size of 64 and a learning rate of $5 * 10^{-4}$. The weight parameter ρ_3 in the loss function (Jensen-Shannon divergence from the target) is set to 10^{-3}. Each component φ, C^{bulk} and C^{ext} is made of 3 fully connected layers, the sizes of which are reported in Table 6.6.

Datasets. For Amazon, we work with the video games subdataset from http://jmcauley.ucsd.edu/data/amazon/. For Yelp Liu and collab. (2015); Yu and collab. (2014), we work with 1,450,000 reviews after that can be found at https://www.yelp.com/dataset.

	Layers' sizes
φ	[2,4,2]
$C^{bulk}_{\theta'}$	[2,8,1]
C^{ext}_{θ}	[2,8,1]

Table 6.6 – Sizes of the successive layers in each component of **LHTR** used in the toy example.

BERT representation for text data. We use BERT pretrained models and code from the library *Transformers* [3]. All models were implemented using Pytorch and trained on a single Nvidia P100. The output of BERT is a $\mathbb{R}^{768}$ vector. All parameters of the models have been selected using the same grid search.

Network architectures. Tables 6.7 report the architectures (layers sizes) chosen for each component of the three algorithms considered for performance comparison (Section 6.5), respectively for the moderate and large datasets used in our experiments. We set $\rho_1 = (1 - \hat{\mathbb{P}}(||Z|| \geq ||Z_{(\lfloor \kappa n \rfloor)}||))^{-1}$ and $\rho_2 = \hat{\mathbb{P}}(||Z|| \geq ||Z_{(\lfloor \kappa n \rfloor)}||)^{-1}$.

	NN model	**LHTR**$_1$	**LHTR**
Sizes of the layers φ	[768,384,200,50,8,1]	[768,384,200,100]	[768,384,200,150]
Sizes of the layers of $C^{bulk}_{\theta'}$	[150,75,8,1]	[100,50,8,1]	[150,75,8,1]
Sizes of the layers of C^{ext}_{θ}	X	X	[150,75,8,1]
ρ_3	X	X	0.01

Table 6.7 – Network architectures for *Amazon dataset* and *Yelp dataset*. The weight decay is set to 10^5, the learning rate is set to $1 * 10^{-4}$, the number of epochs is set to 500 and the batch size is set to 256.

6.9.3 Experiments for data generation

Experimental setting

As mentioned in Section 6.6.1, hyperparameters for dataset augmentation are detailed in Table 6.8. For the Transformer Decoder we use 2 layers with 8 heads, the dimension of the key and value is set to 64 Vaswani and collab. (2017) and the inner dimension is set to 512. The architectures for the models proposed by

[3] https://github.com/huggingface/transformers

	LHTR
Sizes of the layers φ	[768,384,200,150]
Sizes of the layers of $C^{bulk}_{\theta'}$	[150,75,8,1]
Sizes of the layers of C^{ext}_{θ}	[150,75,8,1]
ρ_3	0.01

Table 6.8 – For *Amazon* and *Yelp*, the weight decay is set to 10^5, the learning rate is set to $1 * 10^{-4}$, the number of epochs is set to 100 and the batch size is set to 256.

Wei and Zou (2019) and Kobayashi (2018) are chosen according to the original papers. For a fair comparison with Kobayashi (2018), we update the language model with a BERT model, the labels are embedded in $\mathbb{R}^{10}$ and fed to a single MLP layer. The new model is trained using AdamW Loshchilov and Hutter (2017).

6.10 References

Achab, M., S. Clémençon, A. Garivier, A. Sabourin and C. Vernade. 2017, ■Max k-armed bandit: On the extremehunter algorithm and beyond■, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, p. 389–404. 124

Baayen, R. H. 2002, *Word frequency distributions*, vol. 18, Springer Science & Business Media. 123

Babbar, R., C. Metzig, I. Partalas, E. Gaussier and M.-R. Amini. 2014, ■On power law distributions in large-scale taxonomies■, *ACM SIGKDD Explorations Newsletter*, vol. 16, n° 1, p. 47–56. 123

Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin. 2003, ■A neural probabilistic language model■, *Journal of machine learning research*, vol. 3, n° Feb, p. 1137–1155. 142

Carpentier, A. and M. Valko. 2014, ■Extreme bandits■, in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., p. 1089–1097. 124

Chiapino, M., S. Clémençon, V. Feuillard and A. Sabourin. 2019a, ■A multivariate extreme value theory approach to anomaly clustering and visualization■, *Computational Statistics*, p. 1–22. 124

Chiapino, M. and A. Sabourin. 2016, ■Feature clustering for extreme events analysis, with application to extreme stream-flow data■, in *International Workshop on New Frontiers in Mining Complex Patterns*, Springer, p. 132–147. 124

Chiapino, M., A. Sabourin and J. Segers. 2019b, ■Identifying groups of variables with the potential of being large simultaneously■, *Extremes*, vol. 22, n° 2, p. 193–222. 124, 133

Church, K. W. and W. A. Gale. 1995, ■Poisson mixtures■, *Natural Language Engineering*, vol. 1, n° 2, p. 163–190. 123

Clifton, D. A., S. Hugueny and L. Tarassenko. 2011, ■Novelty detection with multivariate extreme value statistics■, *J Signal Process Syst.*, vol. 65, p. 371–389. 124

Clinchant, S. and E. Gaussier. 2010, ■Information-based models for ad hoc ir■, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, p. 234–241. 124

Coles, S. G. and J. A. Tawn. 1994, ■Statistical methods for multivariate extremes: an application to structural design■, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 43, n° 1, p. 1–31. 146

Crossley, S. and D. McNamara. 2010, ■Cohesion, coherence, and expert evaluations of writing proficiency■, in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 32 (32). 137

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova. 2018, ■Bert: Pre-training of deep bidirectional transformers for language understanding■, *arXiv preprint arXiv:1810.04805*. 123, 125

Fard, M. M., T. Thonet and E. Gaussier. 2018, ■Deep k-means: Jointly clustering with k-means and learning representations■, *arXiv preprint arXiv:1806.10069*. 128

Goix, N., A. Sabourin and S. Clémençon. 2015, ■Learning the dependence structure of rare events: a non-asymptotic study■, in *Conference on Learning Theory*, p. 843–860. 124

Goix, N., A. Sabourin and S. Clémençon. 2016, ■Sparse representation of multivariate extremes with applications to anomaly ranking■, in *Artificial Intelligence and Statistics*, p. 75–83. 124, 133

Goix, N., A. Sabourin and S. Clémençon. 2017, ■Sparse representation of multivariate extremes with applications to anomaly detection■, *Journal of Multivariate Analysis*, vol. 161, p. 12–31. 124

Goodfellow, I., Y. Bengio and A. Courville. 2016, *Deep Learning*, MIT Press. http://www.deeplearningbook.org. 125, 128

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. 2014, ■Generative adversarial nets■, in *Advances in neural information processing systems*, p. 2672–2680. 128

Holtzman, A., J. Buys, M. Forbes and Y. Choi. 2019, ■The curious case of neural text degeneration■, *arXiv preprint arXiv:1904.09751*. 139

Joulin, A., E. Grave, P. Bojanowski and T. Mikolov. 2016, ■Bag of tricks for efficient text classification■, *arXiv preprint arXiv:1607.01759*. 137

Kobayashi, S. 2018, ■Contextual augmentation: Data augmentation by words with paradigmatic relations■, *arXiv preprint arXiv:1805.06201*. 125, 137, 138, 139, 149

Kotzias, D., M. Denil, N. De Freitas and P. Smyth. 2015, ■From group to individual labels using deep features■, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 597–606. 145

Laforgue, P., S. Clémençon and F. d'Alché Buc. 2018, ■Autoencoding any data through kernel autoencoders■, *arXiv preprint arXiv:1805.11028*. 128

Li, J., M. Galley, C. Brockett, J. Gao and B. Dolan. 2015, ■A diversity-promoting objective function for neural conversation models■, *arXiv preprint arXiv:1510.03055*. 137

Liu, J., J. Shang, C. Wang, X. Ren and J. Han. 2015, ■Mining quality phrases from massive text corpora■, in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ACM, p. 1729–1744. 136, 148

Loshchilov, I. and F. Hutter. 2017, ■Decoupled weight decay regularization■, *arXiv preprint arXiv:1711.05101*. 149

Madsen, R. E., D. Kauchak and C. Elkan. 2005, ■Modeling word burstiness using the dirichlet distribution■, in *Proceedings of the 22nd international conference on Machine learning*, p. 545–552. 124

Makhzani, A., J. Shlens, N. Jaitly, I. Goodfellow and B. Frey. 2015, ■Adversarial autoencoders■, *arXiv preprint arXiv:1511.05644*. 128

Mandelbrot, B. 1953, ■An informational theory of the statistical structure of language■, *Communication theory*, vol. 84, p. 486–502. 123

McAuley, J. and J. Leskovec. 2013, ■Hidden factors and hidden topics: understanding rating dimensions with review text■, in *Proceedings of the 7th ACM conference on Recommender systems*, ACM, p. 165–172. 136

Miller, G. A. 1995, ■Wordnet: a lexical database for english■, *Communications of the ACM*, vol. 38, nº 11, p. 39–41. 125

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer. 2018, ■Deep contextualized word representations■, in *Proc. of NAACL*. 123

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever. 2018, ■Improving language understanding by generative pre-training■, *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*. 123

Ratner, A. J., H. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré. 2017, ■Learning to compose domain-specific transformations for data augmentation■, in *Advances in neural information processing systems*, p. 3236–3246. 125

Resnick, S. I. 2013, *Extreme values, regular variation and point processes*, Springer. 126

Roberts, S. 1999, ■Novelty detection using extreme value statistics■, *IEE P-VIS IMAGE SIGN*, vol. 146, p. 124–129. 124

Roberts, S. 2000, ■Extreme value statistics for novelty detection in biomedical data processing■, *IEE P-SCI MEAS TECH*, vol. 147, p. 363–367. 124

Silfverberg, M., A. Wiemerslage, L. Liu and L. J. Mao. 2017, ■Data augmentation for morphological reinflection■, *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, p. 90–99. 137

Stephenson, A. 2003, ■Simulating multivariate extreme value distributions of logistic type■, *Extremes*, vol. 6, nº 1, p. 49–59. 133

Sutskever, I., O. Vinyals and Q. V. Le. 2014, ■Sequence to sequence learning with neural networks■, in *Advances in neural information processing systems*, p. 3104–3112. 130

Thomas, A., S. Clémençon, A. Gramfort and A. Sabourin. 2017, ■Anomaly detection in extreme regions via empirical mv-sets on the sphere.■, in *AISTATS*, p. 1011–1019. 124, 133

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin. 2017, ■Attention is all you need■, in *Advances in neural information processing systems*, p. 5998–6008. 130, 148

Wang, J. and L. Perez. 2017, ■The effectiveness of data augmentation in image classification using deep learning■, *Convolutional Neural Networks Vis. Recognit.* 125

Wei, J. and K. Zou. 2019, ■Eda: Easy data augmentation techniques for boosting performance on text classification tasks■, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 6383–6389. 125, 137, 138, 149

Yu, X., X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick and J. Han. 2014, ■Personalized entity recommendation: A heterogeneous information network approach■, in *Proceedings of the 7th ACM international conference on Web search and data mining*, ACM, p. 283–292. 136, 148

Chapter 7

Subspace Clustering for Multivariate Extremes

Chapter abstract

Capturing the dependence structure of multivariate extreme data is a major challenge in many fields involving the management of risks that come from multiple sources, *e.g.*, portfolio monitoring, environmental risk management, insurance and anomaly detection. The present chapter presents preliminary work and develops a novel optimization-based approach called MEXICO, standing for *Multivariate EXtreme Informative Clustering by Optimization*. It aims at exhibiting a sparsity pattern within the dependence structure of extremes. This is achieved by estimating some disjoint clusters of features that tend to be large simultaneously through an optimization method on the probability simplex. This dimension reduction technique can be applied to statistical learning tasks such as feature clustering and anomaly detection or can be used as a preprocessing step for tasks mentioned in Chapter 5-6. Numerical experiments provide strong empirical evidence of the relevance of our approach.

7.1 Introduction

Clustering is essential for exploratory data mining, data structure analysis and a common technique for statistical data analysis. It is widely used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Many clustering approaches exist with different intrinsic notions of what a cluster is. In the standard setup, the goal is to group objects into subsets, known as clusters, such that objects within a given cluster are more related to one another than the ones from a different cluster. Clustering is already quite well-known (see Bishop (2006); Friedman and collab. (2001) and references therein) conversely to Extreme Value Theory (EVT) which currently gain interest in the machine learning community that has been used in anomaly detection Clifton and collab. (2011); Goix and collab. (2016); Roberts (1999); Thomas and collab. (2017),

classification Jalalzai and collab. (2018, 2020); Vignotto and Engelke (2018) or clustering Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019); Janssen and collab. (2020) when dedicated to the most extreme regions of the sample space.

Scaling up multivariate EVT is a major challenge when addressing high-dimensional learning tasks. Most multivariate extreme value models have been designed to handle moderate dimensional problems, *e.g.*, with dimension $p \leq 10$. For larger dimensions, simplifying modeling choices are needed, stipulating for instance that only some predefined subgroups of components may be concomitant extremes, or, on the contrary, that all must be Sabourin and Naveau (2014); Stephenson (2009). This curse of dimensionality can be explained, in the context of extreme values analysis, by the relative scarcity of extreme data, the computational complexity of the estimation procedure and, in the parametric case, by the fact that the dimension of the parameter space usually grows with that of the sample space.

Recalling the framework of Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019), the goal of this chapter is to present a novel optimization-based approach for clustering extremes in a multivariate setup. Given $N \geq 1$ *i.i.d* copies $X_1, \ldots, X_N$ of a heavy-tailed random variable $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$, we want to identify clusters of features $K \subset [\![1, p]\!]$ such that the variables $\{X^j : j \in K\}$ may be large while the other variables X^j for $j \notin K$ simultaneously remain small. Up to approximately 2^p combinations of extreme features are possible and contributions such as Chautru and collab. (2015); Chiapino and Sabourin (2016); Chiapino and collab. (2019); Engelke and Hitz (2018); Goix and collab. (2016) tend to identify a smaller number of simultaneous extreme features. Dimensional reduction methods such as principal components analysis and derivatives Cooley and Thibaud (2019); Drees and Sabourin (2019); Tipping and Bishop (1999); Wold and collab. (1987) can be designed to find a lower dimensional subspace where extremes tend to concentrate. Another way of identifying the clusters of features that may jointly be large is to select combinations of extreme features, in the spirit of archetypes defined in Cutler and Breiman (1994). Following this path, the idea of the present chapter is to decompose the ℓ_1-norm of a positive input sample as a weighted sum of its features.

Several EVT contributions are aimed at assessing a sparse support of multivariate extremes Chiapino and Sabourin (2016); De Haan and Ferreira (2007); Engelke and Ivanovs (2020); Meyer and Wintenberger (2019). A broader scope of contributions related to the work detailed in this chapter ranges from compressed sensing Candès and collab. (2006); Candes and collab. (2006); Tsaig and Donoho (2006) and matrix factorization Lee and Seung (2001); Şimşekli and collab. (2015) to group sparsity Devijver and collab. (2015); Simon and collab. (2013); Yuan and Lin (2006).

The contributions of this chapter are: *(i)* following contribution laid out by Niculae and collab. (2018), we study at length different subsets on the probability simplex, *(ii)* we present a novel optimization-based approach to perform clustering of extreme features in the multivariate set-up with respected regularity property and *(iii)* we show how to leverage the obtained clusters to detect outliers within

the extreme regions in the context of anomaly detection.

The chapter is organized as follows, in Section 7.2 we begin by introducing the EVT background, multivariate set-up and problem of interest. Then we present in Section 7.3 our optimization-based approach along with its specific details concerning the projection step onto the probability simplex. Section 7.4 is dedicated to applications in statistical learning, namely *feature clustering* and *anomaly detection*. We perform some numerical experiments in Section 7.5 to highlight the performance of our method and we finally conclude in Section 7.6. Proofs, technical details and additional results can be found in the supplementary material.

7.2 Probabilistic Framework

In this section, we briefly recall the main concepts and definitions from Chapter 2. Extreme Value Theory (EVT) develops models for learning the unusual rather than the usual, in order to provide a reasonable assessment of the probability of occurrence of rare events. This section introduces the mathematical framework and classical tool such as standardization for the analysis of multivariate extremes.

Mathematical background. The notion of *regular variation* is a natural way for modelling power law behaviors that appear in various fields of probability theory. In this chapter, we shall focus on the dependence and regular variation of random variables and random vectors. We refer to Resnick (1987) for an excellent account of heavy-tailed distributions and the theory of regularly varying functions.

Notations. The following notations are used throughout this chapter: $\mathcal{M}_{n,p}([1,+\infty[)$ is the set of $n \times p$ matrices valued in $[1,+\infty[$. A_p^m is the set of $p \times m$ matrices valued in $[0,1]$ where the sum of elements of any column equals 1. Any matrix of A_p^m is called a *mixture matrix*. For any $M = (M_i^j) \in \mathcal{M}_{n,p}(\mathbb{R})$, for $i \in [\![1,n]\!]$ (*resp.* $j \in [\![1,p]\!]$), let e_i (*resp.* e^j) denote the vector of the canonical basis such that $e_i M = M_i$ (*resp.* $Me^j = M^j$) where M_i corresponds to the i-th line of M (*resp.* M^j corresponds to the j-th column). Let $E = [0,\infty]^p \backslash \{0\}$ and $\Omega_{p,||\cdot||} = \{x \in \mathbb{R}_+^p : \|x\| \leq 1\}$ the ball associated to the norm $||\cdot||$ and its complementary set $\Omega_{p,||\cdot||}^c = \mathbb{R}_+^p \backslash \Omega_{p,||\cdot||}$, let S denote the sphere associated to $\|\cdot\|$ and for $V \in \mathbb{R}^p$ and $K \subset [\![1,p]\!]$, write $V^{(K)} = (V^j \mathbb{1}_{j \in K})$. Denote by Γ the Euler function.

Definition 13. *(Regular variation Karamata (1933)) A positive measurable function g is regularly varying with index $\alpha \in \mathbb{R}$, notation $g \in \mathcal{R}_\alpha$ if $\lim_{x \to +\infty} g(tx)/g(x) = t^\alpha$ for all $t > 0$.*

The notion of regular variation is defined for a random variable X when the function of interest is the distribution tail of X.

Definition 14. *(Univariate regular variation) A non-negative random variable X is regularly varying with tail index $\alpha \geq 0$ if its right distribution tail $x \mapsto \mathbb{P}\{X > x\}$ is regularly varying with index $-\alpha$, i.e., $\lim_{x \to +\infty} \mathbb{P}\{X > tx \mid X > x\} = t^{-\alpha}$ for all $t > 1$.*

This definition can be extended to the multivariate setting where the topology of the probability space is involved. We rely on the vague convergence of measures (Resnick, 1987, Section 3.4) and consider the following definition (Resnick, 1986, p.69).

Definition 15. *(Multivariate regular variation) A random vector $X \in \mathbb{R}_+^p$ is regularly varying with tail index $\alpha \geq 0$ if there exists $g \in \mathcal{R}_{-\alpha}$ and a nonzero Radon measure μ on E such that*

$$g(t)^{-1}\mathbb{P}\left\{t^{-1}X \in A\right\} \xrightarrow[t\to\infty]{} \mu(A),$$

where $A \subset E$ is any Borel set such that $0 \notin \partial A$ and $\mu(\partial A) = 0$.

Standardization. Let F be the joint cumulative distribution function *(c.d.f)* of X and F^j be the marginal *c.d.f* of X^j with $j \in [\![1, p]\!]$. The tails of the marginals may differ and it is convenient, in practice, to work with marginally standardized variables. In other words, we separate the margins from the dependence structure in the description of the joint distribution of X to compare the different features X^j. For that purpose, we consider the Pareto scaling $T : \mathbb{R}^p \mapsto \mathbb{R}_+^p$ defined by

$$\forall x \in \mathbb{R}^p, \forall j \in [\![1, p]\!], \quad T^j(x^j) = 1/\Big(1 - F^j(x^j)\Big) \in [1, +\infty]. \tag{7.1}$$

This transformation produces a vector $V \overset{\text{def}}{=} T(X) = \Big(T(X^1), \cdots, T(X^p)\Big)$ where each marginal follows a Pareto distribution, *i.e.*, $\mathbb{P}\{V^j > t\} = t^{-1}$ for all $t > 1$. In such a case, a simple choice for f is $f(t) = t^{-1}$ and $V \in \mathbb{R}_+^p$ is regularly varying with tail index $\alpha = 1$ (see (Jalalzai and collab., 2018, Remark 1)). The limiting measure μ is homogeneous and known as the *exponent measure*.

In practice, the marginal functions F^j are unknown and need to be approximated. Assume that $N \geq 1$ i.i.d copies $X_1, \ldots, X_N$ of a heavy-tailed random variable $X \in \mathbb{R}^p$ are available. When the margins are unknown, the Pareto scaling can be approximated by the rank transformation $\widehat{T} : \mathbb{R}^p \mapsto \mathbb{R}_+^p$ relying on the empirical marginal *c.d.f* denoted by $\widehat{F}^j(x) = (1/N + 1)\sum_{i=1}^N \mathbb{1}\{X_i^j \leq x\}$, with $x \in \mathbb{R}$. The empirical version $\widehat{V}$ of V from Equation (7.1) is defined by

$$\widehat{V}_i = \widehat{T}(X_i) = \left(\frac{1}{1 - \widehat{F}^1(X_i^1)}, \ldots, \frac{1}{1 - \widehat{F}^p(X_i^p)}\right), \qquad \forall i \in [\![1, N]\!]. \tag{7.2}$$

The extreme region is then selected by choosing all vectors with norm larger than a fixed threshold $t > 0$, *i.e.*, the extreme data are vectors $\widehat{V}$ such that $||\widehat{V}|| > t$, yielding to n samples considered as extremes. The Euclidian space $\mathbb{R}^p$ being of finite dimension, all norms are equivalent and the choice of the norm does not matter for the limit measure definition Beirlant and collab. (2006b).

Note that this rank standardization is commonly used in multivariate EVT to study the dependence structure of extremes (see Beirlant and collab. (2006a) and references therein) and avoids any further marginal distributions assumptions. The resulting feature variables of (7.2) are not independent and the remaining goal is to discover the dependence structure of standardized extremes.

Problem statement. We consider a vector $V \in \mathbb{R}_+^p$ whose features come from a mixture of extreme values and we would like to find clusters of features that

get large together. We seek $m \geq 2$ clusters $K_1, \ldots, K_m$ with $m < p$ such that all features in a same subset may be large together. Unit sets are not relevant for clustering so we assume that each cluster is of size at least 2. We also want clusters that are disjoint, *i.e.*, for all $i \neq j, K_i \cap K_j = \emptyset$. This choice is motivated to reach a representation of interest, *e.g.*, diversity for portfolio in finance or clusters for smart grids in wireless technologies. In the remaining of this chapter, $V \in \mathcal{M}_{n,p}([1, +\infty[)$ corresponds to all the samples $X_1, \ldots, X_N$ after the rank standardization and selection of extremes. With this notation, we have n samples $V_1, \ldots, V_n \in \mathbb{R}_+^p$ that are *i.i.d.* copies of the vector V and for all $i \in [\![1, n]\!]$, we search a subset K of features such that the ℓ_1-norms of V_i and its restriction $\widetilde{V}_i = V_i^{(K)}$ are almost equal $\|\widetilde{V}_i\|_1 \approx \|V_i\|_1$.

7.3 Optimization Problem

Features mixtures. In order to recover the clusters, we consider mixtures of the components of each sample. The true number of clusters is unknown so we can only have a guess and search for m clusters. We consider the probability simplex defined on the positive orthant $\mathbb{R}_+^p$ by

$$\Delta_p = \{x \in \mathbb{R}_+^p, x_1 + \ldots + x_p = 1\},$$

and let $W \in A_p^m$ with $m < p$ be a *mixture matrix*. We denote by $\widetilde{V} = VW \in \mathcal{M}_{n,m}(\mathbb{R}_+)$ the transformed matrix. The following theorem –in the footsteps of Engelke and collab. (2019)– ensures the preservation of the regularly varying behavior and points out the behavior of the limiting measure.

Theorem 11. *Let $V = T(X) \in \mathbb{R}_+^p$ coming from a Pareto scaling and $W \in A_p^m$ a mixture matrix with $1 < m \leq p$. Then the transformed vector $\widetilde{V} = VW \in \mathbb{R}_+^m$ is regularly varying with tail index $\alpha = 1$. Denote μ (resp. $\widetilde{\mu}$) the limiting measure of V (resp. $\widetilde{V}$), then we have*

$$(1/m)\widetilde{\mu}(\Omega^c_{m,||\cdot||_1}) \leq \widetilde{\mu}(\Omega^c_{m,||\cdot||_\infty}) \leq \mu(\Omega^c_{p,||\cdot||_1}).$$

Remark 14. *(Selection of m) In view of Theorem 11, in practice, the required dimension $m < p$ can be seen as the smallest value m such that the empirical version of $\widetilde{\mu}(\Omega_m^c)$ is arbitrarly close to the empirical version of $\mu(\Omega_p^c)$. In that way, the m selected clusters provide a good enough representation of the dependency between features.*

Loss function. Each column W^j for $j \in [\![1, m]\!]$ is modelling a mixture of components and represents a cluster K_j. For any sample $V_i, i \in [\![1, n]\!]$, we want to find a mixture that gives a good approximation in ℓ_1-norm, *i.e.*, we seek a column $j \in [\![1, m]\!]$ for which $\widetilde{V}_i^j$ is the closest to $||V_i||_1$. In other words, we need to find $j \in [\![1, m]\!]$ in order to minimize the score function γ defined as

$$\forall (i, j) \in [\![1, n]\!] \times [\![1, m]\!], \quad \gamma(W, V_i, j) = ||V_i||_1 - \widetilde{V}_i^j.$$

For each sample V_i, we need to minimize the loss function defined by

$$\mathcal{L}(W, V_i) = \min_{1 \leq j \leq m} \gamma(V_i, W, j) = \min_{1 \leq j \leq m} \left(||V_i||_1 - \widetilde{V}_i^j \right).$$

The optimization problem consists in finding a mixture matrix $W^\star$ minimizing the global loss

$$W^\star \in \underset{W \in A_p^m}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(W, V_i). \tag{7.3}$$

Note that A_p^m is a closed and bounded set hence compact Bourbaki (2007) thus there exists at least one solution which can be reached. Equation (7.3) is composed of two minimization problems and can be rewritten as

$$W^\star \in \underset{W \in A_p^m}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le j \le m} \widetilde{V}_i^j.$$

The index of the column representing a good mixture can be defined with the mapping

$$\varphi : [\![1, n]\!] \to [\![1, m]\!], \quad \varphi(i) = \underset{1 \le j \le m}{\arg\max}\, \widetilde{V}_i^j$$

and the optimization problem becomes

$$W^\star \in \underset{W \in A_p^m}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} (VW)_i^{\varphi(i)} = \underset{W \in A_p^m}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} e_i (VW) e^{\varphi(i)}. \tag{7.4}$$

Simple example. To have a better understanding of the optimization problem, we consider a simple example to show how the matrix W is recovering the different clusters. Assume that the vector $V \in \mathbb{R}_+^p$ is exactly coming from a mixture of m disjoint clusters $K_1, \dots, K_m$ and for each sample V_i, there exists K_j such that $\|V_i^{(K_j)}\|_1 = \|V_i\|_1$. For all $j \in [\![1, m]\!]$, denote $U^j \in [0,1]^p$ the uniform vector with support K_j, *i.e.*, $U^j = (1/|K_j|)^{(K_j)}$. A solution to the optimization problem is given by any column-permutation of the matrix $W^\star$ whose columns are the vectors U^j. Indeed, the transformed data matrix is $\widetilde{V} = VW$ and for any sample V_i that comes from a cluster K_j, we have

$$\forall l \neq j, \quad \widetilde{V}_i^j = V_i U^j = V_i^{(K_j)} U^j \ge V_i U^l = \widetilde{V}_i^l.$$

Taking $\varphi(i) = \arg\max_{1 \le l \le m} \widetilde{V}_i^l$ exactly recovers the cluster of index $j = \varphi(i)$. In the case where the large features of the different sample V_i are all equal, then the columns of the mixture matrix W tend exactly to uniform vectors with restricted support. Now, if one of the large feature is slightly bigger than the other then the associated column of W tends to a vertex of the simplex.

Problem relaxation. One can directly solve the linear program (7.4) but this formulation suffers from drawbacks. First, the solution provided tends to be very sparse since it would belong to a vertex of the simplex. Then it involves the search of the mapping φ among all the possible combinations which can be prohibited when n or p increases. Thus, one can solve a relaxed version of (7.4) by introducing another matrix of mixtures $Z \in A_m^n$. The relaxed problem is

$$(W^\star, Z^\star) \in \underset{(W,Z) \in A_p^m \times A_m^n}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} V_i W Z^i. \tag{7.5}$$

Optimization problem. We recognize the trace operator which is linear and can define an objective function $f : A_p^m \times A_m^n \to \mathbb{R}$ that we need to maximize:

$$\begin{cases} (W^\star, Z^\star) \in \arg\max_{(W,Z)} f(W, Z) \\ f(W, Z) = Tr(VWZ)/n \end{cases}$$

The objective function f is bilinear in finite dimension hence continuous. Since maxization occurs on compact sets, there is at least one solution $(W^\star, Z^\star)$. However, it is not unique since any column-permutation of $W^\star$ along with the associated row-permutation of $Z^\star$ is also a valid solution.

Regularization. The constraint of disjoint clusters can be satisfied by forcing the columns of the mixture matrix W to be orthogonal, *i.e.*, for all $i < j$, $\langle W^i, W^j \rangle = 0$. This yields a penalized version of the objective function with a regularization parameter $\lambda > 0$

$$\begin{cases} (W^\star, Z^\star) \in \arg\max_{(W,Z)} f_\lambda(W, Z) \\ f_\lambda(W, Z) = Tr(VWZ)/n - \lambda \sum_{i<j} \langle W^i, W^j \rangle \end{cases}$$

with partial derivatives given by

$$\begin{cases} \nabla_Z f_\lambda(W, Z) & = (VW)^T/n \\ \nabla_W f_\lambda(W, Z) & = (ZV)^T/n - \lambda \widetilde{W}, \quad \widetilde{W}^j = \sum_{i<j} W^i. \end{cases}$$

Update rule. The optimization problem can be addressed using an alternate scheme by computing projected gradient ascent at each iteration

$$\begin{cases} W_{k+1} & = \Pi_{\mathcal{S}} \left(W_k + \delta_k^W \nabla_W f_\lambda(W_k, Z_k) \right) \\ Z_{k+1} & = \Pi_{\Delta_m} \left(Z_k + \delta_k^Z \nabla_Z f_\lambda(W_{k+1}, Z_k) \right) \end{cases} \tag{7.6}$$

where $\Pi_{\mathcal{S}}(\cdot), \Pi_{\triangle_m}(\cdot)$ are respectivetly the projection of each column onto a convex set $\mathcal{S} \subset \Delta_p$ and onto the probability simplex Δ_m. The learning rates δ_k^W, δ_k^Z are step sizes found by backtracking line search.

Projection step on $\mathcal{S}$. In order to recover clusters that are not unit sets, we want to avoid the vertices of the simplex. Thus, we perform a projection step $\Pi_{\mathcal{S}}(\cdot)$ of each column of W onto a convex set $\mathcal{S}$. Several choices are to be considered, as illustrated in Figure 7.1. Denote $\bar{x} = (1/p, \ldots, 1/p)$ the barycenter of the probability simplex Δ_p and consider the following subsets:

(i) ℓ_1 incircle: the coordinate permutations of $(0, 1/(p-1), \ldots, 1/(p-1))$ are the centers of the faces of Δ_p and they define a reversed and scaled simplex $\mathcal{S}_p^{\ell_1}$.

(ii) ℓ_2 incircle: consider the euclidian ball $B_{2,p}(\bar{x}, r) = \{x \in \mathbb{R}^p | \|x - \bar{x}\|_2 \leq r\}$. The radius value $r_p = 1/\sqrt{p(p-1)}$ yields the ℓ_2 inscribed ball of Δ_p along with $\mathcal{S}_p^{\ell_2} = \Delta_p \cap B_{2,p}(\bar{x}, r_p)$.

(iii) Mexican set: The previous subsets do not scale well as the dimension grows and we shall discuss some theoretical results to see that their hypervolumes become very small. To escape from the curse of dimensionality, we consider the convex set where we cut off the vertices using a threshold τ of the distance $L = \|\bar{x} - e_j\|_2 = \sqrt{(p-1)/p}$ between the barycenter and a vertex. It is also the

intersection of the simplex Δ_p and an ℓ_∞ ball. We call this subset the Mexican set $\mathcal{S}_p^\tau$ defined as

$$\mathcal{S}_p^\tau = \left\{ x \in \Delta_p | \max_{1\leq j\leq p} \left\langle x - \bar{x}, \frac{e_j - \bar{x}}{\|e_j - \bar{x}\|_2} \right\rangle \leq \tau \right\}.$$

Define $r_\infty^p(\tau) = 1 - (1-\tau)(p-1)/p$ then we also have the relation

$$\mathcal{S}_p^\tau = \Delta_p \cap B_{\infty,p}(\bar{x}, \tau L) = \Delta_p \cap B_{\infty,p}(0, r_\infty^p(\tau)).$$

The projection onto the simplex is a well-studied subject Chen and Ye (2011); Condat (2016); Daubechies and collab. (2008); Duchi and collab. (2008). For the projection onto the intersection of convex sets, one can perform a naive approach of alternate projections Gubin and collab. (1967) or some refinements using the idea of Dykstra's algorithm Boyle and Dykstra (1986); Bregman and collab. (2003); Dykstra (1983).

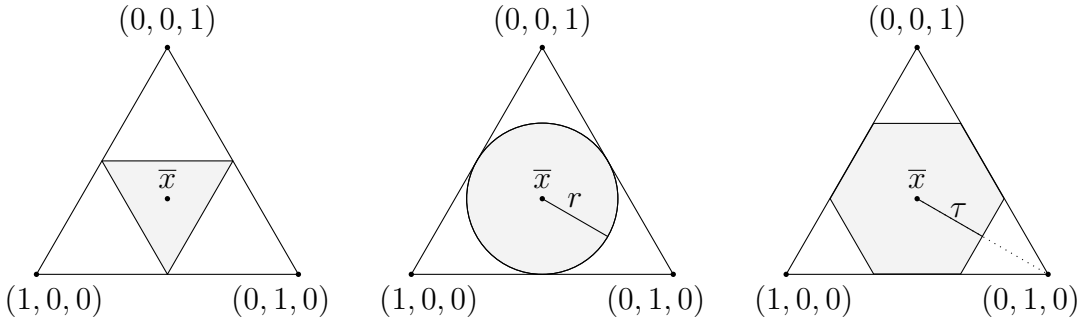


Figure 7.1 – Simplex of $\mathbb{R}^3$ with $\mathcal{S}_3^{\ell_1}$ (left), $\mathcal{S}_3^{\ell_2}$(center) and the Mexican set $\mathcal{S}_3^\tau$(right).

Theorem 12. *(Volumes and ratios) Consider the probability simplex Δ_p and the different manifolds $\mathcal{S}_p^{\ell_1}, \mathcal{S}_p^{\ell_2}, \mathcal{S}_p^\tau$. For any bounded set $\mathcal{D} \subset \mathbb{R}^p$, define its hypervolume $\mathcal{V}ol(\mathcal{D})$ and its ratio $\rho(\mathcal{D})$ as*

$$\mathcal{V}ol(\mathcal{D}) = \int_{\mathbb{R}^p} \mathbb{1}_{\mathcal{D}}(x)\mathrm{d}x, \quad \rho(\mathcal{D}) = \mathcal{V}ol(\mathcal{D})/\mathcal{V}ol(\Delta_p).$$

Denote Γ the Euler function, with $\Gamma(p) = (p-1)!$, we have,

Manifold	*Symbol*	*Hypervolume $\mathcal{V}ol(\mathcal{S})$*	*Ratio $\rho(\mathcal{S})$*
Simplex	Δ_p	$\frac{\sqrt{p}}{\Gamma(p)}$	1
ℓ_1-ball	$\mathcal{S}_p^{\ell_1}$	$\frac{\sqrt{p}}{\Gamma(p)}\left(\frac{1}{(p-1)^{(p-1)}}\right)$	$\frac{1}{(p-1)^{(p-1)}}$
ℓ_2-ball	$\mathcal{S}_p^{\ell_2}$	$\frac{\sqrt{p}}{\Gamma(p)}\left(\frac{\Gamma(p)}{\Gamma\left(\frac{p+1}{2}\right)}\frac{\pi^{(p-1)/2}}{\sqrt{p^p(p-1)^{(p-1)}}}\right)$	$\frac{\Gamma(p)}{\Gamma\left(\frac{p+1}{2}\right)}\frac{\pi^{(p-1)/2}}{\sqrt{p^p(p-1)^{(p-1)}}}$
Mexican set	$\mathcal{S}_p^\tau$	$\frac{\sqrt{p}}{\Gamma(p)}\left(1 - p(1-\tau)^{(p-1)}\left(\frac{p-1}{p}\right)^{(p-1)}\right)$	$1 - p\left[(1-\tau)\left(\frac{p-1}{p}\right)\right]^{(p-1)}$

Moreover, when the dimension grows $p \to +\infty$ and for a fixed $\tau \in (0,1)$, we have $\rho(\mathcal{S}_p^{\ell_1}) \to 0$, $\rho(\mathcal{S}_p^{\ell_2}) \to 0$ and $\rho(\mathcal{S}_p^\tau) \to 1$.

Remark 15. *(Selection of τ) With a high reduction of the probability simplex, the vertices are avoided but the clusters are more difficult to discriminate since the Mexican set tends to the barycenter of the simplex. This trade-off motivates the choice of the threshold τ and Figure 7.2 shows the evolution of the ratios ρ for the different subsets.*

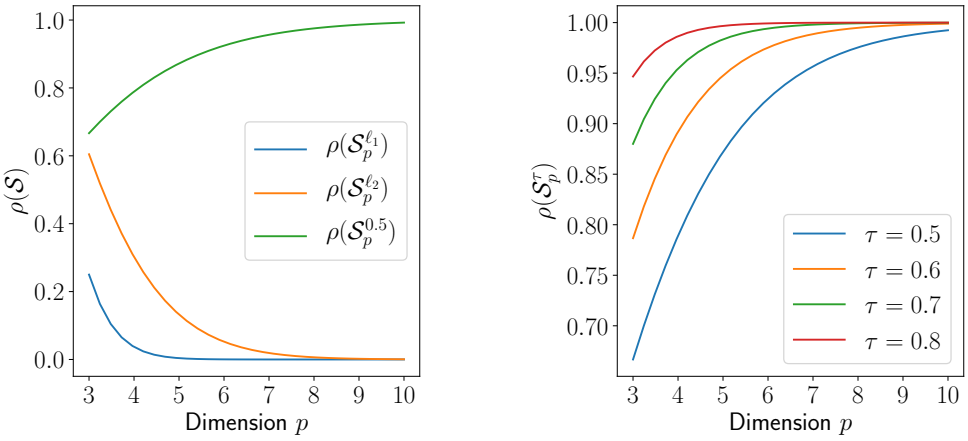


Figure 7.2 – Evolutions of ratio volumes $\rho(\mathcal{S})$ with dimension p.

7.4 Statistical Learning Applications

Starting from random matrices $(W_0, Z_0) \in A_p^m \times A_m^n$, the algorithm of Equation (7.6) returns a pair of matrices (W_{mex}, Z_{mex}) that are of great interest to analyze the dependence structure of the extreme data. On the one hand, the mixture matrix W_{mex} gives insights about the different clusters of features that are large simultaneously. On the other hand, the matrix Z_{mex} gives information about the probability of belonging to each cluster. Indeed, those matrices are trained on the data matrix V so that each column W_{mex}^j represents a cluster K_j and for each sample $V_i, i \in [\![1, n]\!]$, the j^{th}-row of the column Z_{mex}^i is the confidence of belonging to the cluster K_j.

Features Clustering. Consider the features clustering task where we receive a new extreme sample $V_{\text{new}} \in \mathbb{R}_+^p$ and need to predict the cluster where its large features are drawn. To assess the dependency structure of this new sample, one can compute the transformed sample $\widetilde{V}_{\text{new}}$ and assign the predicted cluster by

$$\widetilde{V}_{\text{new}} = V_{\text{new}} W_{mex}, \quad \text{Pred}(V_{\text{new}}) = \underset{1 \le j \le m}{\arg\max}\, \tilde{V}_{\text{new}}^j.$$

Anomaly Detection. Consider now the anomaly detection task where we receive a new extreme sample $V_{\text{new}} \in \mathbb{R}_+^p$ and need to predict whether it is an anomaly or not. One can look at the score function evaluated at the new sample $\gamma\left(W_{mex}, V_{\text{new}}, \varphi_{\text{new}}\right)$ where $\varphi_{\text{new}} = \arg\max_{1 \le j \le m} \tilde{V}_{\text{new}}^j$. If this score is small then it means that the dependency structure of V_{new} is well captured by the mixture $W_{mex}^{\varphi_{\text{new}}}$ and the behavior is rather normal that unusual. Similarly, a high value of this score means that V_{new} cannot be well explained by any mixture of W_{mex} and therefore it is more likely to be an outlier. Based on that remark, it is easy to make a prediction for the behavior of the extreme sample V_{new} using a decreasing function of the score value.

Algorithm 2 (Mexico: training and applications)**.**

Require *Training data* $(X_1, \ldots, X_N), 0 < m < p, \lambda > 0$ *and threshold* $k(= \sqrt{N})$.

1 **Standardization.** *Standardize data* $(\widehat{V}_1, \ldots, \widehat{V}_N)$ *with rank transformation* (7.2).

2 **Truncation.** *Compute extreme regions* $\mathcal{I} = \{i \in [\![1, N]\!], ||\widehat{V}_i|| \geq N/k\}$.

3 **Optimization.** *Compute* $(W_{mex}, Z_{mex}) \in \arg\max_{(W,Z)} f_\lambda(W, Z)$ *using update rule* (7.6).

4 *(**Clustering**) Return cluster* φ_0 *or*
*(**Anomaly Detection**) Return score* $\gamma\left(W_{mex}, V_{new}, \varphi_0\right)$.

Remark 16. *(On selection of* k *in Algorithm 2) Determining* k *is a central bias variance trade-off of Extreme Value analysis (See e.g.* Goix *and collab. (2016) and references therein). As* k *gets too large, a bias is induced by taking into account observations which do not necessarily behave as extremes: their distribution deviates significantly from the limit distribution of extremes. On the other hand, too small values lead to an increase of the algorithm's variance. In practice, a conventional choice is* $k = \sqrt{N}$.

7.5 Numerical Experiments

To compare the performance of our algorithm against state-of-the-art methods, we focus on popular machine learning tasks for extreme events: features clustering and anomaly detection. We shall consider various dimensions up to a big data framework where the dimension p is relatively large compared to the number n of samples. Since the margins distributions are unknown, we apply the rank transformation as described in Algorithm 2. For ease of reproducibility, the code is available in the supplementary material.

7.5.1 Feature Clustering

Recently, Janssen and collab. (2020) explored how the spherical k-means algorithm can be applied in the analysis of only the extremal observations from a data set. We perform a benchmark of this method versus our algorithm MEXICO on simulated data from logistic distribution in a high-dimensional setting. Given the knowledge of the ground truth class assignments of the samples, it is possible to define some intuitive metric using conditional entropy analysis. In particular, Rosenberg and Hirschberg (2007) define the following desirable objectives for any cluster assignment: Homogeneity (H), each cluster contains only members of a single class; Completeness (C), all members of a given class are assigned to the same cluster; v-Measure (v-M): the harmonic mean of Homogeneity and Completeness, which is actually equivalent to the mutual information. The parameter configuration is the following: dimension $p \in \{75, 100, 150, 200\}$, number of train samples $n_{\text{train}} = 1000$ and test samples $n_{\text{test}} = 100$. We used metrics implemented

by *Scikit-Learn* Pedregosa and collab. (2011). The results, obtained over 100 independently simulated dataset for each value of p, are gathered in Table 7.1, where the column dedicated to MEXICO transcribes the best results between projection method with Dykstra's algorithm and alternating projection. Both methods are detailed in the supplementary material. For each dimension p, bold characters in Table 7.1 indicate the best method when results are statistically significant.

p	Spherical-Kmeans Janssen and collab. (2020)			MEXICO		
	H	C	v-M	H	C	v-M
75	0.950±0.034	0.972±0.024	0.961±0.027	**0.978**±0.025	0.976±0.024	**0.977**±0.024
100	0.943±0.031	0.967±0.024	0.955±0.026	**0.978**±0.020	**0.979**±0.021	**0.978**±0.020
150	0.940± 0.026	0.962±0.020	0.951±0.022	**0.976**±0.015	**0.980**±0.013	**0.978**±0.014
200	0.940±0.018	0.962±0.014	0.951±0.015	**0.970**±0.015	**0.975**±0.012	**0.972**±0.013

Table 7.1 – Comparison of Homogeneity (H), Completeness (C) and v-Measure (v-M) from prediction scores for SphericalKmeans and Mexico on simulated data with different dimension p.

7.5.2 Anomaly Detection

We perform a comparison of three algorithms: Isolation Forest Liu and collab. (2008), Damex Goix and collab. (2017) and Mexico. The algorithms are trained and tested on the same datasets, the test set being restricted to extreme regions. Five reference AD datasets are considered: shuttle, forestcover, http, SF and SA. The setting is detailed in Table 7.2. The experiments are performed in a semi-supervised framework where the training set consists of normal data only. More details about the preprocessing and additional results are available in the supplementary material.

Dataset	Size	Anomalies	τ	λ
SF	73 237	3298 (4.5%)	0.8	10
SA	100 655	3377 (3.4%)	0.7	5
http	58 725	2209 (3.8%)	0.5	10
shuttle	49 097	3511 (7.2%)	0.7	5
forestcover	286 048	2747 (0.9%)	0.7	5

Table 7.2 – Description of each dataset and hyperparameters of Mexico for anomaly detection.

The results of means and standard deviations, obtained over 100 runs, are gathered in Table 7.3 and reveal the good performance of our approach.

Dataset	iForest Liu and collab. (2008)		Damex Goix and collab. (2016)		Mexico	
	ROC-AUC	AP	ROC-AUC	AP	ROC-AUC	AP
SF	0.381±0.086	0.393±0.081	0.710±0.031	0.650±0.034	**0.892**±0.013	**0.812**±0.016
SA	0.886±0.032	0.879±0.031	0.982±0.002	0.938±0.012	0.983±0.031	**0.950**±0.011
http	0.656±0.094	0.658±0.099	0.996±0.002	0.968±0.009	0.997±0.002	**0.972**±0.012
shuttle	0.970±0.020	0.826±0.055	0.990±0.003	0.864±0.026	0.990±0.003	0.864±0.037
forestcover	0.654±0.096	0.894±0.037	0.762±0.008	0.893±0.010	**0.863**±0.015	**0.958**±0.006

Table 7.3 – Comparison of Area Under Curve of Receiver Operating Characteristic (ROC-AUC) and Average Precision (AP) from prediction scores of each method on different anomaly detection datasets.

7.6 Conclusion

Understanding the impact of shocks, *i.e.*, extremely large input values on systems is of critical importance in diverse fields, *e.g.*, security, finance, environmental sciences, epidemiology. In this chapter, we have developed a preliminary methodological framework for clustering in extreme regions, relying on the non-parametric theory of regularly varying random vectors, and illustrated its performance for both feature clustering and anomaly detection on simulated and real data. Our approach does not scan all the multiple possible subsets and outperforms existing algorithms. From a broader perspective, extreme data may have dramatic consequences and any clustering algorithm in such a setting should be used with great caution. Note that the purpose of MEXICO is to provide informative clusters of features although no guarantees on its robustness are provided so far. Clustering rare events is the cornerstone of many applications and may have huge social consequences, *e.g.*, a river dam failure in hydrological sciences or a miscarriage of justice when dealing with homeland security. Finally, recovering clusters of data concerning sick patients at an early stage of a global pandemic is the key to slow down the resulting epidemic. In this way, future work will focus on the statistical properties and guarantees of the developed algorithm by further exploring links with kernel methods.

7.7 Proofs of Theorems

Section 7.7 gathers the proofs of the two theorems and Section 7.8 is dedicated to numerical experiments: the preprocessing of the data and additional results. In Section 7.9, we present further numerical experiments with a visualization of the clusters found by the algorithm.

7.7.1 Proof of Theorem 11

Proof. Assume that $V = T(X) \in \mathbb{R}^p_+$ is coming from a Pareto scaling. Then each marginal V^j for $j = 1, \ldots, p$ follows a Pareto distribution and V is a regularly varying random vector with tail index 1 (see Jalalzai and collab. (2018)). Using the characterization of (Basrak and collab., 2002), we have the following equivalence between the behavior of the vector and its components

$$(V \text{ is regularly varying}) \iff (\forall u \in \mathbb{R}^p, \langle u, V \rangle \text{ is univariate regularly varying})$$

Each column $\widetilde{V}^j$ of $\widetilde{V} = VW$ is given by the linear combination $\widetilde{V}^j = \sum_{k=1}^p V^k W_k^j$. Therefore, any linear combination of the form $\langle \widetilde{u}, \widetilde{V} \rangle$ with $\widetilde{u} \in \mathbb{R}^m$ is actually a linear combination of the form $\langle u, V \rangle$. Indeed, we have for $\widetilde{u} \in \mathbb{R}^m$,

$$\langle \widetilde{u}, \widetilde{V} \rangle = \sum_{j=1}^m \widetilde{u}_j \widetilde{V}^j = \sum_{j=1}^m \widetilde{u}_j \left(\sum_{k=1}^p V^k W_k^j \right) = \sum_{k=1}^p \left(\sum_{j=1}^m \widetilde{u}_j W_k^j \right) V^k.$$

Because V is regularly varying then any linear combination of the form $\langle \widetilde{u}, \widetilde{V} \rangle$ is univariate regularly varying, which exactly means, using the equivalence, that $\widetilde{V}$ is a regularly varying random vector. To find the tail index of the transformed vector, we rely on the following Lemma.

Lemma 5. *Let $V \in \mathbb{R}_+^p$ be a random vector with each component following a Pareto distribution and $W \in A_p^m$ a mixture matrix. Then each marginal of the transformed vector VW is regularly varying with tail index 1.*

Proof. Following Lemma 3.9 from Jessen and Mikosch (2006), let $\mathcal{A}_{W^j}$ denote the set $\{x, \langle W^j, x \rangle > 1\}$ where W^j is the j-th column of W, we want to show that $\mu(\mathcal{A}_{W^j}) > 0$. Let $\epsilon = \max_{1 \leq i \leq p} W_i^j$. It follows that $\epsilon > 0$ otherwise $W^j = 0$ and it would not belong to Δ^p. Let $i^* = \arg\max_{i \leq p} W_i^j$ *i.e.* $\epsilon = W_{i^*}^j$. As $W_{i^*}^j$ and V_{i^*} are positive,

$$\begin{aligned} \langle W^j, V \rangle &\geq W_{i^*}^j V_{i^*} \\ &\geq \epsilon V_{i^*}. \end{aligned}$$

Therefore, $\{\epsilon V_{i^*} \geq t\} \subset \{\langle W^j, V \rangle \geq t\}$ and $t\mathbb{P}\{\epsilon V_{i^*} \geq t\} \leq t\mathbb{P}\{\langle W^j, V \rangle \geq t\}$. By taking the limit on both sides of the inequality, we obtain that $0 < \epsilon \leq \mu(\mathcal{A}_{W^j})$ and we conclude $\widetilde{V}$ is regularly varying with tail index 1. □

Since the random vectors V and $\widetilde{V}$ are regularly varying, we have the existence of nonzero Radon measures μ and $\widetilde{\mu}$ that are independent of the considered norm (see Beirlant and collab. (2006b)). Moreover, in virtue of Lemma 5, the tail indexes of V and $\widetilde{V}$ are equal to 1. Consider the complementary of the unit sphere, defined by $\Omega_{m,||\cdot||}^c = \{x \in \mathbb{R}_+^m, \|x\| > 1\}$. We have by definition

$$\widetilde{\mu}(\Omega_{m,||\cdot||}^c) = \lim_{t \to \infty} t\mathbb{P}\left\{ t^{-1}\widetilde{V} \in \Omega_{m,||\cdot||}^c \right\} = \lim_{t \to \infty} t\mathbb{P}\left\{ ||\widetilde{V}|| > t \right\}.$$

Using that $(1/m)||\widetilde{V}||_1 \leq ||\widetilde{V}||_\infty = \max_{1 \leq j \leq m} \widetilde{V}^j = \max_{1 \leq j \leq m} \left(\sum_{k=1}^p V^k W_k^j \right)$ and $W_k^j \in [0,1]$, we have

$$(1/m)\mathbb{P}\left\{ ||\widetilde{V}||_1 > t \right\} \leq \mathbb{P}\left\{ ||\widetilde{V}||_\infty > t \right\} = \mathbb{P}\left\{ \max_{1 \leq j \leq m} \widetilde{V}^j > t \right\} \leq \mathbb{P}\left\{ \sum_{k=1}^p V^k > t \right\}.$$

We recognize the ℓ_1-norm of the random vector $V \in \mathbb{R}_+^p$ and obtain

$$\forall t > 1, \quad (1/m)t\mathbb{P}\left\{ ||\widetilde{V}||_1 > t \right\} \leq t\mathbb{P}\left\{ ||\widetilde{V}||_\infty > t \right\} \leq t\mathbb{P}\left\{ ||V||_1 > t \right\}.$$

Taking the limit $t \to \infty$ on the inequalities provides the desired result $(1/m)\widetilde{\mu}(\Omega_{m,||\cdot||_1}^c) \leq \widetilde{\mu}(\Omega_{m,||\cdot||_\infty}^c) \leq \mu(\Omega_{p,||\cdot||_1}^c)$. □

7.7.2 Proof of Theorem 12

Proof. First, recall the hypervolume of the p-simplex with side length a and the hypervolume of the Euclidian ball of radius R in dimension p,

$$\mathcal{V}ol(\Delta_p, a) = \frac{\sqrt{p}}{(p-1)!}\left(\frac{a}{\sqrt{2}}\right)^{p-1}, \quad \mathcal{V}ol\left(B_{2,p}(0,R)\right) = \frac{\pi^{p/2}R^p}{\Gamma\left(\frac{p}{2}+1\right)}.$$

Probability simplex Δ_p. The probability simplex we consider has a side length of $a = \sqrt{2}$ which gives the value of $\mathcal{V}ol(\Delta_p)$.

ℓ_1*-incircle.* Regarding the ℓ_1-ball, it is the scaled simplex whose side length is given by the distance between two face centers of Δ_p. This length is equal to $\sqrt{2}/(p-1)$ and we deduce the volume $\mathcal{V}ol(\mathcal{S}_p^{\ell_1})$.

ℓ_2*-incircle.* For the ℓ_2-ball, denote $\mathcal{B} = (e_1, \ldots, e_p)$ the canonical basis and let $x \in \mathcal{S}_p^{\ell_2}, x = \sum_{i=1}^p \langle x, e_i\rangle e_i = \sum_{i=1}^p x_i e_i$. The vector $e_p^{'} = \sqrt{p}\bar{x} = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$ is unitary and orthogonal to the simplex Δ_p with $\Delta_p \subset Span(e_p^{'})^{\perp}$. We have $\langle x, e_p^{'}\rangle = 0$ and we can complete the vector $e_p^{'}$ into an orthonormal basis $\mathcal{B}' = (e_1^{'}, \ldots, e_p^{'})$ with $P = \mathcal{P}_{\mathcal{B},\mathcal{B}'}$ and $x = \sum_{i=1}^p \langle x, e_i\rangle e_i = \sum_{i=1}^{p-1} \langle x, e_i^{'}\rangle e_i^{'}$. The hypervolume is invariant by translation so we make the projection of $\mathcal{S}_p^{\ell_2}$ onto $\mathbb{R}^{p-1}$ to see that

$$\mathcal{V}ol(\mathcal{S}_p^{\ell_2}) = \mathcal{V}ol\left(B_{2,p-1}(0, r_p)\right),$$

with $r_p = 1/\sqrt{p(p-1)}$ the radius of the ℓ_2 inscribed ball of Δ_p. This gives the value of $\mathcal{V}ol(\mathcal{S}_p^{\ell_2})$.

Mexican set. Finally for the Mexican set, we cut off with a threshold τ the length $L = \sqrt{(p-1)/p}$ between the barycenter $\bar{x}$ and a vertex e_i. We get p smaller simplices and the volume we want is nothing but the difference between the volume of the simplex Δ_p and p times the volume of a small simplex. To compute the hypervolume of one small simplex, we need to find its side length, knowing that its height is $(1-\tau)L$. We find a side length equal to $\sqrt{2}(1-\tau)(p-1)/p$ and can conclude for the value $\mathcal{V}ol(\mathcal{S}_p^{\tau})$. □

We present in Figure 7.3 the evolution of the ratio $\rho(\mathcal{S}^{\tau})$ of the Mexican set for different values of threshold and dimension.

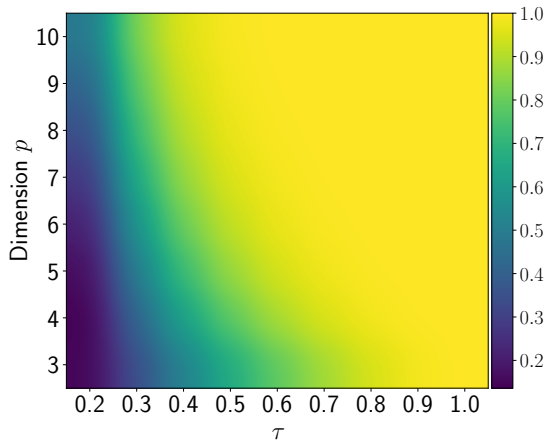


Figure 7.3 – Evolution of $\rho(\mathcal{S}^\tau)$ with varying values of (τ, p).

7.8 Numerical experiments details

7.8.1 Additional results Feature Clustering

We present the full results of the performance of MEXICO regarding the feature clustering task. The projection step is either performed using alternating projections based on the method POCS (Projection Onto Convex Sets) or with the more elaborate technique Dykstra.

p	Spherical-Kmeans Janssen and collab. (2020)			MEXICO (POCS)		
	H	C	v-M	H	C	v-M
75	0.950±0.034	0.972±0.024	0.961±0.027	**0.978**±0.025	0.976±0.024	**0.977**±0.024
100	0.943±0.031	0.967±0.024	0.955±0.026	**0.976**±0.020	**0.979**±0.021	**0.976**±0.020
150	0.940± 0.026	0.962±0.020	0.951±0.022	**0.973**±0.015	**0.977**±0.013	**0.975**±0.014
200	0.940±0.018	0.962±0.014	0.951±0.015	**0.970**±0.015	**0.975**±0.012	**0.972**±0.013

Table 7.4 – Comparison of Homogeneity (H), Completeness (C) and v-Measure (v-M) from prediction scores for Spherical Kmeans and Mexico with alternating projections on simulated data with different dimension p.

p	Spherical-Kmeans Janssen and collab. (2020)			MEXICO (Dykstra)		
	H	C	v-M	H	C	v-M
75	0.950±0.034	0.972±0.024	0.961±0.027	**0.977**±0.025	0.975±0.024	**0.976**±0.024
100	0.943±0.031	0.967±0.024	0.955±0.026	**0.978**±0.020	**0.979**±0.021	**0.978**±0.020
150	0.940± 0.026	0.962±0.020	0.951±0.022	**0.976**±0.015	**0.980**±0.013	**0.978**±0.014
200	0.940±0.018	0.962±0.014	0.951±0.015	**0.967**±0.015	**0.972**±0.012	**0.970**±0.013

Table 7.5 – Comparison of Homogeneity (H), Completeness (C) and v-Measure (v-M) from prediction scores for Spherical Kmeans and Mexico with Dykstra projection on simulated data with different dimension p.

7.8.2 Anomaly detection, real world data preprocessing

We present the details about the preprocessing of the real world datasets.

Dataset	iForest Liu and collab. (2008)	Damex Goix and collab. (2016)	Mexico (POCS)	Mexico (Dykstra)
SF	0.381±0.086	0.710±0.031	**0.892**±0.013	0.710±0.030
SA	0.886±0.032	0.982±0.002	0.981±0.006	0.983±0.031
http	0.656±0.094	0.996±0.002	0.995±0.005	0.997±0.002
shuttle	0.970±0.020	0.990±0.003	0.990±0.003	0.989±0.003
forestcover	0.654±0.096	0.762±0.008	**0.863**±0.015	0.851±0.008

Table 7.6 – Comparison of Area Under Curve of Receiver Operating Characteristic (ROC-AUC)from prediction scores of each method on different anomaly detection datasets.

The shuttle dataset is the fusion of the training and testing datasets available in the UCI repository Lichman (2013). The data have 9 numerical attributes, the first one being time. Labels from 7 different classes are also available. Class 1 instances are considered as normal, the others as anomalies. We use instances from all different classes but class 4, which yields an anomaly ratio (class 1) of 7.2%.

In the forestcover data, also available at UCI repository (Lichman (2013)), the normal data are the instances from class 2 while instances from class 4 are anomalies, other classes are omitted, so that the anomaly ratio for this dataset is 0.9%.

The last three datasets belong to the KDD Cup 99 dataset (KDDCup (1999); Tavallaee and collab. (2009), produced by processing the tcpdump portions of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset, created by MIT Lincoln Lab Lippmann and collab. (2000). The artificial data was generated using a closed network and a wide variety of hand-injected attacks (anomalies) to produce a large number of different types of attack with normal activity in the background. Since the original demonstrative purpose of the dataset concerns supervised AD, the anomaly rate is very high (80%), which is unrealistic in practice, and inappropriate for evaluating the performance on realistic data. We thus take standard preprocessing steps in order to work with smaller anomaly rates.

For datasets SF and http we proceed as described in Yamanishi and collab. (2004): SF is obtained by picking up the data with positive logged-in attribute, and focusing on the intrusion attack, which gives an anomaly proportion of 4.5%. The dataset http is a subset of SF corresponding to a third feature equal to 'http'. Finally, the SA dataset is obtained as in Eskin and collab. (2002) by selecting all the normal data, together with a small proportion (3.4%) of anomalies.

We present the full results of the performance of MEXICO regarding the anomaly detection task. The projection step is either performed using alternating projections based on the method POCS (Projection Onto Convex Sets) or with the more elaborate technique Dykstra.

Dataset	iForest Liu and collab. (2008)	Damex Goix and collab. (2016)	Mexico (POCS)	Mexico (Dykstra)
SF	0.393±0.081	0.650±0.034	**0.812**±0.016	0.661±0.031
SA	0.879±0.031	0.938±0.012	0.940±0.031	0.950±0.011
http	0.658±0.099	0.968±0.009	0.972±0.012	0.971±0.008
shuttle	0.826±0.055	0.864±0.026	0.864±0.037	0.818±0.024
forestcover	0.894±0.037	0.893±0.010	0.958±0.006	0.954±0.004

Table 7.7 – Comparison of Average Precision (AP) from prediction scores of each method on different anomaly detection datasets.

7.9 Further Numerical Experiments

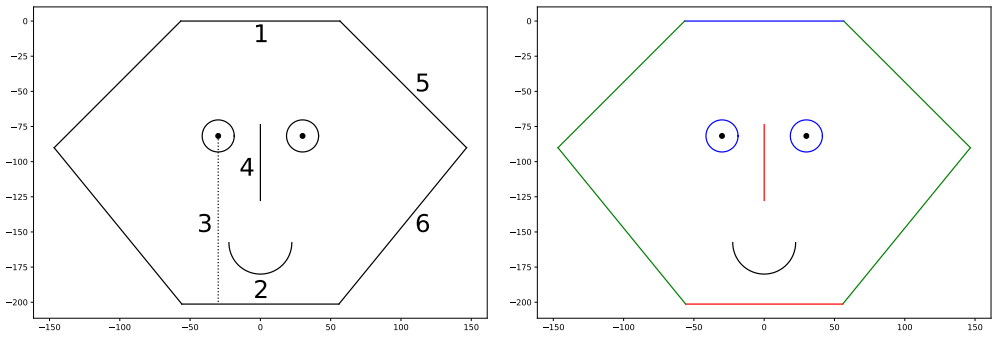
The authors of Cutler and Breiman (1994) provide an archetypal analysis of the Swiss Army dataset. This dataset consists of 6 head dimensions from 200 Swiss soldiers. The data was gathered to construct face masks for the Swiss army. Few samples of the dataset are presented in Table 7.8.

The first measurement (MFB) corresponds to the width of the face just above the eyes. The second feature (BAM) corresponds to the width of the face just below the mouth. The third measurement (TFH) is the distance from the top of the nose to the chin. The fourth feature (LGAN) is the length of the nose. The fifth measurement (LTN) is the distance from the ear to the top of the head while the sixth (LTG) is the distance from the ear to the bottom of the face. For a better visualization of the dataset, we made simple drawings of the different samples. Figure 7.4a illustrates the 6 measurements.

id	MFB	BAM	TFH	LGAN	LTN	LTG
0	113.2	111.7	119.6	53.9	127.4	143.6
1	117.6	117.3	121.2	47.7	124.7	143.9
2	112.3	124.7	131.6	56.7	123.4	149.3
3	116.2	110.5	114.2	57.9	121.6	140.9
4	112.9	111.3	114.3	51.5	119.9	133.5

Table 7.8 – Extract of the Swiss Army dataset.

A question that naturally rises is to figure out subgroups of face features that get large simultaneously. Mexico algorithm performed on the standardized dataset provides the following groups of features : $\{5, 6\}$(green), $\{1, 3\}$(blue) and $\{2, 4\}$(red), as illustrated in Figure 7.4b.



(a) (b)

Figure 7.4 – Illustration of the 6 measurements (a) and subgroups that tend to be large simulatenously (b).

7.10 References

Basrak, B., R. A. Davis and T. Mikosch. 2002, ■A characterization of multivariate regular variation■, *Annals of Applied Probability*, p. 908–920. 166

Beirlant, J., Y. Goegebeur, J. Segers and J. Teugels. 2006a, *Statistics of extremes: theory and applications*, John Wiley & Sons. 158

Beirlant, J., Y. Goegebeur, J. Segers and J. L. Teugels. 2006b, *Statistics of extremes: theory and applications*, John Wiley & Sons. 158, 167

Bishop, C. M. 2006, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc. 155

Bourbaki, N. 2007, *Topologie générale: Chapitres 1 à 4*, Springer Science & Business Media. 160

Boyle, J. P. and R. L. Dykstra. 1986, ■A method for finding projections onto the intersection of convex sets in hilbert spaces■, in *Advances in order restricted statistical inference*, Springer, p. 28–47. 162

Bregman, L. M., Y. Censor, S. Reich and Y. Zepkowitz-Malachi. 2003, ■Finding the projection of a point onto the intersection of convex sets via projections onto half-spaces■, *Journal of Approximation Theory*, vol. 124, n° 2, p. 194–218. 162

Candès, E. J., J. Romberg and T. Tao. 2006, ■Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information■, *IEEE Transactions on information theory*, vol. 52, n° 2, p. 489–509. 156

Candes, E. J., J. K. Romberg and T. Tao. 2006, ■Stable signal recovery from incomplete and inaccurate measurements■, *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, n° 8, p. 1207–1223. 156

Chautru, E. and collab.. 2015, ■Dimension reduction in multivariate extreme value analysis■, *Electronic journal of statistics*, vol. 9, n° 1, p. 383–418. 156

CHEN, Y. and X. YE. 2011, ■Projection onto a simplex■, *arXiv preprint arXiv:1101.6081*. 162

CHIAPINO, M. and A. SABOURIN. 2016, ■Feature clustering for extreme events analysis, with application to extreme stream-flow data■, in *International Workshop on New Frontiers in Mining Complex Patterns*, Springer, p. 132–147. 156

CHIAPINO, M., A. SABOURIN and J. SEGERS. 2019, ■Identifying groups of variables with the potential of being large simultaneously■, *Extremes*, vol. 22, nº 2, p. 193–222. 156

CLIFTON, D. A., S. HUGUENY and L. TARASSENKO. 2011, ■Novelty detection with multivariate extreme value statistics■, *J Signal Process Syst.*, vol. 65, p. 371–389. 155

CONDAT, L. 2016, ■Fast projection onto the simplex and the ℓ_1 ball■, *Mathematical Programming*, vol. 158, nº 1, doi: 10.1007/s10107-015-0946-6, p. 575–585, ISSN 1436-4646. URL https://doi.org/10.1007/s10107-015-0946-6. 162

COOLEY, D. and E. THIBAUD. 2019, ■Decompositions of dependence for high-dimensional extremes■, *Biometrika*, vol. 106, nº 3, p. 587–604. 156

CUTLER, A. and L. BREIMAN. 1994, ■Archetypal analysis■, *Technometrics*, vol. 36, nº 4, p. 338–347. 156, 171

DAUBECHIES, I., M. FORNASIER and I. LORIS. 2008, ■Accelerated projected gradient method for linear inverse problems with sparsity constraints■, *journal of fourier analysis and applications*, vol. 14, nº 5-6, p. 764–792. 162

DE HAAN, L. and A. FERREIRA. 2007, *Extreme value theory: an introduction*, Springer Science & Business Media. 156

DEVIJVER, E. and collab.. 2015, ■Finite mixture regression: a sparse variable selection by model selection for clustering■, *Electronic journal of statistics*, vol. 9, nº 2, p. 2642–2674. 156

DREES, H. and A. SABOURIN. 2019, ■Principal component analysis for multivariate extremes■, *arXiv preprint arXiv:1906.11043*. 156

DUCHI, J., S. SHALEV-SHWARTZ, Y. SINGER and T. CHANDRA. 2008, ■Efficient projections onto the l 1-ball for learning in high dimensions■, in *Proceedings of the 25th international conference on Machine learning*, ACM, p. 272–279. 162

DYKSTRA, R. L. 1983, ■An algorithm for restricted least squares regression■, *Journal of the American Statistical Association*, vol. 78, nº 384, p. 837–842. 162

ENGELKE, S., R. DE FONDEVILLE and M. OESTING. 2019, ■Extremal behaviour of aggregated data with an application to downscaling■, *Biometrika*, vol. 106, nº 1, p. 127–144. 159

ENGELKE, S. and A. S. HITZ. 2018, ■Graphical models for extremes■, *arXiv preprint arXiv:1812.01734*. 156

ENGELKE, S. and J. IVANOVS. 2020, ■Sparse structures for multivariate extremes■, *arXiv preprint arXiv:2004.12182*. 156

ESKIN, E., A. ARNOLD, M. PRERAU, L. PORTNOY and S. STOLFO. 2002, *A Geometric Framework for Unsupervised Anomaly Detection*, Springer US, p. 77–101. 170

FRIEDMAN, J., T. HASTIE and R. TIBSHIRANI. 2001, *The elements of statistical learning*, Springer series in statistics Springer, Berlin. 155

GOIX, N., A. SABOURIN and S. CLÉMENÇON. 2016, ■Sparse representation of multivariate extremes with applications to anomaly ranking■, in *Artificial Intelligence and Statistics*, p. 75–83. 155, 156, 164, 166, 170, 171

GOIX, N., A. SABOURIN and S. CLÉMENÇON. 2017, ■Sparse representation of multivariate extremes with applications to anomaly detection■, *Journal of Multivariate Analysis*, vol. 161, p. 12–31. 165

GUBIN, L., B. T. POLYAK and E. RAIK. 1967, ■The method of projections for finding the common point of convex sets■, *USSR Computational Mathematics and Mathematical Physics*, vol. 7, n° 6, p. 1–24. 162

JALALZAI, H., S. CLÉMENÇON and A. SABOURIN. 2018, ■On binary classification in extreme regions■, in *Advances in Neural Information Processing Systems*, p. 3092–3100. 156, 158, 166

JALALZAI, H., P. COLOMBO, C. CLAVEL, E. GAUSSIER, G. VARNI, E. VIGNON and A. SABOURIN. 2020, ■Heavy-tailed representations, text polarity classification & data augmentation■, *arXiv preprint arXiv:2003.11593*. 156

JANSSEN, A., P. WAN and collab.. 2020, ■k-means clustering of extremes■, *Electronic Journal of Statistics*, vol. 14, n° 1, p. 1211–1233. 156, 164, 165, 169

JESSEN, H. A. and T. MIKOSCH. 2006, ■Regularly varying functions■, *Publications de l'Institut Mathematique*, vol. 80, n° 94, p. 171–192. 167

KARAMATA, J. 1933, ■Sur un mode de croissance régulière. théorèmes fondamentaux■, *Bulletin de la Société Mathématique de France*, vol. 61, p. 55–62. 157

KDDCUP. 1999, ■The third international knowledge discovery and data mining tools competition dataset■, . 170

LEE, D. D. and H. S. SEUNG. 2001, ■Algorithms for non-negative matrix factorization■, in *Advances in neural information processing systems*, p. 556–562. 156

LICHMAN, M. 2013, ■UCI machine learning repository■, URL http://archive.ics.uci.edu/ml. 170

Lippmann, R., J. W. Haines, D. Fried, J. Korba and K. Das. 2000, ■Analysis and results of the 1999 darpa off-line intrusion detection evaluation■, in *RAID*, Springer, p. 162–182. 170

Liu, F. T., K. M. Ting and Z.-H. Zhou. 2008, ■Isolation forest■, in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, p. 413–422. 165, 166, 170, 171

Meyer, N. and O. Wintenberger. 2019, ■Sparse regular variation■, *arXiv preprint arXiv:1907.00686*. 156

Niculae, V., A. F. Martins, M. Blondel and C. Cardie. 2018, ■Sparsemap: Differentiable sparse structured inference■, *arXiv preprint arXiv:1802.04223*. 156

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg and collab.. 2011, ■Scikit-learn: Machine learning in Python■, *JMLR*, vol. 12, p. 2825–2830. 165

Resnick, S. 1987, *Extreme Values, Regular Variation, and Point Processes*, Springer Series in Operations Research and Financial Engineering. 157, 158

Resnick, S. I. 1986, ■Point processes, regular variation and weak convergence■, *Advances in Applied Probability*, vol. 18, n° 1, p. 66–138. 158

Roberts, S. 1999, ■Novelty detection using extreme value statistics■, *IEE P-VIS IMAGE SIGN*, vol. 146, p. 124–129. 155

Rosenberg, A. and J. Hirschberg. 2007, ■V-measure: A conditional entropy-based external cluster evaluation measure■, in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, p. 410–420. 164

Sabourin, A. and P. Naveau. 2014, ■Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization■, *Comput. Stat. Data Anal.*, vol. 71, p. 542–567. 156

Simon, N., J. Friedman, T. Hastie and R. Tibshirani. 2013, ■A sparse-group lasso■, *Journal of computational and graphical statistics*, vol. 22, n° 2, p. 231–245. 156

Stephenson, A. 2009, ■High-dimensional parametric modelling of multivariate extreme events■, *Australian & New Zealand Journal of Statistics*, vol. 51, p. 77–88. 156

Tavallaee, M., E. Bagheri, W. Lu and A. Ghorbani. 2009, ■A detailed analysis of the kdd cup 99 data set■, in *IEEE CISDA*, vol. 5, p. 53–58. 170

Thomas, A., S. Clemencon, A. Gramfort and A. Sabourin. 2017, ■Anomaly detection in extreme regions via empirical mv-sets on the sphere.■, in *AISTATS*, p. 1011–1019. 155

Tipping, M. E. and C. M. Bishop. 1999, ■Probabilistic principal component analysis■, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, n° 3, p. 611–622. 156

Tsaig, Y. and D. L. Donoho. 2006, ■Extensions of compressed sensing■, *Signal processing*, vol. 86, n° 3, p. 549–571. 156

Vignotto, E. and S. Engelke. 2018, ■Extreme value theory for open set classification–gpd and gev classifiers■, *arXiv preprint arXiv:1808.09902*. 156

Wold, S., K. Esbensen and P. Geladi. 1987, ■Principal component analysis■, *Chemometrics and intelligent laboratory systems*, vol. 2, n° 1-3, p. 37–52. 156

Yamanishi, K., J.-I. Takeuchi, G. Williams and P. Milne. 2004, ■On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms■, *Data Mining and Knowledge Discovery*, vol. 8, n° 3, p. 275–300. 170

Yuan, M. and Y. Lin. 2006, ■Model selection and estimation in regression with grouped variables■, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, n° 1, p. 49–67. 156

Şimşekli, U., A. Liutkus and A. T. Cemgil. 2015, ■Alpha-stable matrix factorization■, *IEEE Signal Processing Letters*, vol. 22, n° 12, p. 2289–2293. 156

Part IV

Conclusion & Perspectives

7.11 Conclusion

Extreme data arise in a wide variety of statistical and machine learning applications. Despite this ubiquity, most classical multivariate analysis methods tend to neglect this type of data due to its scarcity. This thesis is devoted to the study and practical use of learning with –and from– extremes in a multivariate context. The contributions presented in this dissertation are the following:
First, we studied a nonasymptotic bound for the maximal deviation of the empirical angular measure. A variant (*i.e.* truncated counterpart) of the empirical angular measure was introduced. We illustrated the resulting non-asymptotic bounds in an unsupervised statistical learning problem related to anomaly detection through minimum volume sets on the sphere. A second statistical learning problem related to classification in the extremes was studied in depth, in the first instance, the influence of the standardization (which is a common preprocessing step) is neglected but is reinforced in a second phase. We studied the relevance of the empirical risk minimization dedicated to the extremes and suggested an algorithm to build a performant and extreme-dedicated classifier whereby extrapolation can be performed to samples lying on the edge of the input space. A non-asymptotic bound for the excess risk in the extremes is derived but an estimation of the minimum risk in the extremes is a relevant problem that remains to be addressed.
In the second part of the dissertation, we studied problems beyond regular statistical learning. Although it was a well-studied problem in previous text representation based on word frequencies, there is low interest in the distribution of modern text representations. We take this opportunity to bridge the gap between well-known text representations and more recent text embeddings derived from deep learning models by providing an algorithm to build a heavy-tailed representation using an adversarial approach. The resulting representation leads to a natural mechanism to augment datasets while preserving the label of any given input. This mechanism provides a way to generate data solely by dilation of the embedding. Lastly, we evaluated the dependence structure of extremes in high dimension to find groups of variables which may be large simultaneously by rewriting this latter problem as an optimization problem. The preliminary work and resulting algorithm shows state-of-the-art performance in extreme regions on subspace clustering (*i.e.* feature clustering) and detection of anomaly. This method remains well-suited when working in a high dimensional space.
The work detailed in this thesis provides both new theoretical, practical and methodological elements: guarantees in the form of generalization results suggest original approaches to well-known problems adapted to extreme regions. In that regard, the richness of topics appearing in this thesis demonstrate the imperative for modern machine learning to leverage existing theory to achieve improved performance and understanding as illustrated here in several applications.

7.12 Perspectives

Several points discussed in this thesis can be further developed, as mentioned throughout the thesis. This section depicts some of this dissertation author's research perspectives related to the problems tackled in this thesis.

From Binary to Multiclass Classification and Regression in Extreme Regions. The framework of binary classification can be extended to multiclass classification as most machine learning applications dealing with classification contain multiclass labels. One may expect the main results establised in Chapter 5 to hold when dealing with a multiclass label up to adapting the assumptions. To go beyond discrete labels, regression in extremes is also a relevant problem when the target label is continuous. The considered loss is no longer the classification loss but is a new and central element.

Data driven partition of the sphere for anomaly detection. Chapter 4 illustrates the detection of anomalies *via* minimum volume sets on the sphere. The partition of the ℓ_∞ sphere is designed and set prior to any data analysis. As a first go, it is still valid. However, a data driven and adaptive partition would be better-suited as it would increase analysis speed in high dimensional problems and performance in the densest regions of the sphere.

Heavy Tailed Representation - from Polarity Classification to Topic Modelling and Text Summarization. Chapter 6 bridges the gap between extreme value theory and natural language processing by establishing a heavy-tailed representation. The resulting representation leads to a label preserving text generation. In the reverse direction, summarization of textual content could be performed by exploiting shorter contents that share the same label (*i.e.* angle). One may also mention that labelling data is a costly step in modern machine learning and semi-supervised alternatives of **`GENELIEX`** and **`LHTR`** would better fit for industrial needs.

Interpretation of Mixtures of Features in Subspace Clustering. Dimension reduction or selection must be performed with care. As the directions selected by the algorithm detailed in Chapter 7 lead to finding and selecting the directions which support extreme features, it may subtly alter or belittle relevant features of the observations for downstream tasks. Re-weighting the features in the objective function may help address this. A second resulting problem would be to extend the approach developed in the chapter to more complex data.

Tail induced sparcity in Neural Networks. Performance of deep learning models often goes together with deeper architecture (see Chapter 3). Recent contributions are designed to reduce the depth of such models with a minor loss in performance by removing unnecessary weights *i.e.* weights close to zero while keeping the largest values. Understanding the dependence structure of the largest weights and the corresponding neurons in the architecture may be a relevant research question.

ECOLE
DOCTORALE

Titre : Apprentissage issue de Données Extrêmes Multivariées : Théorie et Application au Traitement Naturelle du Langage

Mots clés : Statistiques, Apprentissage Machine, Extrêmes Multivariées, Traitement Naturel du Langage

Résumé : Les *extrêmes* apparaissent dans une grande variété de données. Par exemple, concernant les données hydrologiques, les extrêmes peuvent correspondre à des inondations, des moussons voire des sécheresses. Les données liées à l'activité humaine peuvent également conduire à des situations extrêmes, dans le cas des transactions bancaires, le montant alloué à une vente peut être considérable et dépasser les transactions courantes. Un autre exemple lié à l'activité humaine est la fréquence des mots utilisés: certains mots sont omniprésents alors que d'autres sont très rares. Qu'importe le contexte applicatif, les extrêmes qui sont rares par définition, correspondent à des données particulières. Ces événements sont notamment alarmants au vu de leur potentiel impact désastreux. Cependant, les données extrêmes sont beaucoup moins considérées dans les statistiques modernes ou les pratiques courantes d'apprentissage machine, principalement car elles sont considérablement sous représentées : ces événements se retrouvent noyés - à l'ère du "*big data*" - par la vaste majorité de données classiques et non extrêmes. Ainsi, la grande majorité des outils d'apprentissage machine qui se concentrent naturellement sur une distribution dans son ensemble peut être inadaptée sur les queues de distribution où se trouvent les observations extrêmes.
Dans cette thèse, les défis liés aux extrêmes sont détaillés et l'accent est mis sur le développement de méthodes dédiées à ces données. La première partie se consacre à l'apprentissage statistique dans les régions extrêmes. Dans le chapitre 4, des garanties non asymptotiques sur l'erreur d'estimation de la mesure angulaire empirique sont étudiées et permettent d'améliorer des méthodes de détection d'anomalies par minimum volume set sur la sphère. En particulier, le problème de la minimisation du risque empirique pour la classification binaire dédiée aux échantillons extrêmes est traitée au chapitre 5. L'analyse non paramétrique et les garanties qui en résultent sont détaillées. L'approche est adaptée pour traiter de nouveaux échantillons se trouvant hors de l'enveloppe convexe formée par les données rencontrées. Cette propriété d'extrapolation est l'élément clé et charnière nous permettant de concevoir de nouvelles représentations conservant un label donné et d'ainsi augmenter la quantité de données. Le chapitre 6 se concentre sur l'apprentissage de cette représentation à queue lourde (pour être précis, à *variation régulière*) à partir d'une distribution d'entrée. Les illustrations montrent une meilleure classification des extrêmes et conduit à la génération de phrases cohérentes. Enfin, le chapitre 7 propose d'analyser la structure de dépendance des extrêmes multivariés. En constatant que les extrêmes se concentrent au sein de groupes où les variables explicatives ont tendance à prendre de manière récurrente de grandes valeurs simultanément; il résulte un problème d'optimisation visant à identifier ces sous-groupes grâce à des moyennes pondérées des composantes.

ECOLE
DOCTORALE

Title : Learning from Multivariate Extremes: Theory and Application to Natural language Processing

Keywords : Statistics, Machine Learning, Multivariate Extremes, Natural Language Processing

Abstract : *Extremes* surround us and appear in a large variety of data. Natural data like the ones related to environmental sciences contain extreme measurements; in hydrology, for instance, extremes may correspond to floods and heavy rainfalls or on the contrary droughts. Data related to human activity can also lead to extreme situations; in the case of bank transactions, the money allocated to a sale may be considerable and exceed common transactions. The analysis of this phenomenon is one of the basis of fraud detection. Another example related to humans is the frequency of encountered words. Some words are ubiquitous while others are rare. No matter the context, extremes which are rare by definition, correspond to uncanny data. These events are of particular concern because of the disastrous impact they may have. Extreme data, however, are less considered in modern statistics and applied machine learning, mainly because they are substantially scarce: these events are outnumbered –in an era of so-called "*big data*"– by the large amount of classical and *non-extreme data* that corresponds to the bulk of a distribution. Thus, the wide majority of machine learning tools and literature may not be well-suited or even performant on the distributional tails where extreme observations occur.

Through this dissertation, the particular challenges of working with extremes are detailed and methods dedicated to them are proposed. The first part of the thesis is devoted to statistical learning in extreme regions. In Chapter 4, non-asymptotic bounds for the empirical angular measure are studied. Here, a pre-established anomaly detection scheme via minimum volume set on the sphere, is further improved. Chapter 5 addresses empirical risk minimization for binary classification of extreme samples. The resulting non-parametric analysis and guarantees are detailed. The approach is particularly well suited to treat new samples falling out of the convex envelop of encountered data. This extrapolation property is key to designing new embeddings achieving label preserving data augmentation. Chapter 6 focuses on the challenge of learning the latter heavy-tailed (and to be precise *regularly varying*) representation from a given input distribution. Empirical results show that the designed representation allows better classification performance on extremes and leads to the generation of coherent sentences. Lastly, Chapter 7 analyses the dependence structure of multivariate extremes. By noticing that extremes tend to concentrate on particular *clusters* where features tend to be recurrently large simulatenously, we define an optimization problem that identifies the aforementioned subgroups through a weighted means of features.

Institut Polytechnique de Paris
91120 Palaiseau, France