The Angular Measure for Multivariate Extremes: Concentration Bounds, Goodness-of-Fit Tests, and Generative Methods

Stéphane Lhaut

Thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Sciences

August 2025

Institute of Statistics, Biostatistics and Actuarial Sciences
Louvain Institute of Data Analysis and Modeling
Université catholique de Louvain
Louvain-la-Neuve
Belgium

Thesis Committee:

Pr. Johan **Segers** (Advisor)	UCLouvain/ISBA, Belgium
Pr. Anna **Kiriliouk**	UNamur/naXys, Belgium
Pr. Rainer **von Sachs**	UCLouvain/ISBA, Belgium
Pr. Karim **Barigou**	UCLouvain/ISBA, Belgium
Pr. Anne **Sabourin**	Université Paris Cité/MAP5, France

The Angular Measure for Multivariate Extremes: Concentration Bounds, Goodness-of-Fit Tests, and Generative Methods
by Stéphane Lhaut

ISBA
Université catholique de Louvain
Voie du Roman Pays, 20
1348 Louvain-la-Neuve
Belgium

This work was partially supported by the F.R.S-FNRS.

On ne peut comprendre un processus en l'interrompant.
La compréhension doit rejoindre le cheminement
du processus et cheminer avec lui.
FRANK HERBERT

Preamble

This thesis is concerned with the analysis of multivariate extreme values. In a broad sense, this means that, after modelling a multivariate phenomenon of interest as a random variable, we are concerned about the statistical analysis of the tail of the resulting probability distribution. The main difficulty comes from the tail dependence structure which can (almost) be arbitrary. Many possible tools exist to describe this dependence and one of the most popular approaches is obtained by considering the so-called *angular measure* [36] of the probability distribution (after standardization of its margins). The angular measure describes the probability of the phenomenon of interest being in a certain direction (angle) of the space given that it is sufficiently far from a predefined origin point. Inference on this measure is the theme of this work.

Chapter 1: Extreme Value Analysis: An Introduction. In this introductory chapter, we formalize the notion of the tail analysis we aim to conduct using the theory of regular variation of [74]. We explain the key challenges related to the univariate and multivariate frameworks. A precise definition of the angular measure is provided.

Chapter 2: Uniform Concentration Bounds for Frequencies of Rare Events. This chapter provides a set of Vapnik–Chervonenkis type concentration inequalities that are specially designed for the analysis of rare events, and, in particular, extreme events. No extreme value theory is used here, and the applicability of these results is not limited to this framework.

Chapter 3: Concentration Bounds for the Empirical Angular Measure with Statistical Learning Applications. We apply the results of the previous chapter to the problem of empirically estimating the angular measure in a space of arbitrary dimension. The main result is a finite-sample bound on the deviation from the true angular measure, uniformly on a possibly infinite class of subsets of the sphere. Applications to the statistical learning problems of binary classification and minimum-volume set estimation are discussed.

Chapter 4: Bivariate Asymptotic Theory and Testing for the Angular Measure. We leave the general framework to focus on the bivariate situation, from a large sample perspective. Extending previous asymptotic results on

the angular measure in this case, we develop a goodness-of-fit procedure for testing general parametric models that are popular in extremes. Our main results provide a simple and efficient way of testing these models together with an asymptotically justified parametric bootstrap to approach the limit distribution of the test statistic. Multiple testing issues are also discussed.

Chapter 5: Wasserstein–Aitchison Generative Adversarial Networks for Angular Measures. Of a different nature, generative methods based on Generative Adversarial Networks (GANs) and variants thereof are becoming popular for estimating dependences in high-dimensional spaces. We exploit and adapt these methods to the problem of estimating an angular measure in a space of arbitrary dimension. In particular, we explain how the special constraints of this measure are enforced in the generative networks. Among other things, this relies on equipping the simplex with an orthonormal basis, the resulting object being known as the Aitchison simplex from compositional data analysis. The method's performance is illustrated on simulated and real financial data.

Chapter 6: Concluding Remarks and Perspectives. This chapter summarises the thesis and proposes extensions and perspectives on the theory and methods that have been introduced.

The main contributions of this thesis are contained in the following papers.

- **Stéphane Lhaut**, Anne Sabourin and Johan Segers. *Uniform concentrations bounds for frequencies of rare events.* Statistics and Probability Letters, 189:109610, 2022.
- Stéphan Clémençon, Hamid Jalalzai, **Stéphane Lhaut**, Anne Sabourin, and Johan Segers. *Concentration bounds for the empirical angular measure with statistical learning applications.* Bernoulli, 29(4):2797–2827, 2023.
- **Stéphane Lhaut** and Johan Segers. *An asymptotic expansion of the empirical angular measure for bivariate extremal dependence.* In Matteo Barigozzi, Siegfried Hörmann, and Davy Paindaveine, editors, Recent Advances in Econometrics and Statistics. Springer, November 2024.
- **Stéphane Lhaut** and Johan Segers. *Testing parametric models for the angular measure for bivariate extremes.* arXiv preprint arXiv:2411.12673 (submitted to Extremes).
- **Stéphane Lhaut**, Holger Rootzén and Johan Segers. *Wasserstein Aitchison GAN for angular measures of multivariate extremes.* arXiv preprint arXiv:2504.21438 (submitted to Extremes).

Acknowledgments

Ce document est le fruit de quatre années de travail, que j'ai eu la chance de passer sous la supervision du Professeur Johan Segers. Je retire énormément de nos discussions, et ce depuis longtemps puisque tu me supervisais déjà pour mon travail de fin de bachelier (ou de license, pour les nombreux Français que j'ai eu la chance de rencontrer). Mon goût pour les probabilités vient presque uniquement de ton enseignement et je suis ravi d'avoir pu collaborer si longtemps avec toi après avoir réussi l'examen de ton cours en juin 2019 ! Merci aussi pour m'avoir ensigné, directement ou indirectement, le goût pour les choses bien faites et l'envie de se dépasser, mathématiquement et ailleurs.

Au-delà de la recheche, l'environnement de travail qu'offre l'ISBA est, je pense, assez unique de part son dynamisme et la variété des gens qu'on y rencontre. Par exemple, mon voisin de bureau, Gabriel Bailly, avec qui nous cohabitons professionnellement depuis 3 ans, et dont le travail porte sur l'étude statistique d'objets vivants dans des espaces non-Euclidiens: un sujet que, vous en conviendrez, peu d'entre nous à l'ISBA souhaiterions tenter d'approcher. Merci pour ton écoute et ton enthousiasme lorsque je m'emballais au tableau dans des explications, souvent peu convaincantes, sur un problème mathématique que je venais de lire ou sur une idée farfelue sur laquelle j'esssayais de te lancer. Notre cohabitation fut un plaisir !

Lorsque je sortais ma tête de mon bureau et des problèmes mathématiques, sans doute un peu trop souvent, j'avais la chance de pouvoir parler avec des collègues sympathiques et enthousiastes. Je pense en particulier à Quentin Le Coënt et Aurèle Bartoloméo pour les nombreuses soirées à Louvain-la-plage, à Fanny Hoogstoel grâce à qui je suis sans doute meilleur au kicker et un peu moins MAFY, à Alexandre Jacquemain chez qui nous attendons toujours de revenir faire un BBQ, à Jean-Loup (attention de bien prononcer le "p") Dupret pour nos discussions sur le Tour de France et ta bonne ambiance communicative, à Benjamin Deketelaere pour être un super colocataire et pongiste engagé, à Lise Léonard pour être toujours motivée et partager un niveau à peu près similaire au mien au karaoké, à Anas Mourahib pour tous ces voyages et rencontres lors de nos conférences, à Morine Delhelle pour sa joie de vivre (surtout après un cocktail), à Mathilde Foulon pour les nombreuses discussions/séances de psychologie dans mon bureau, à Hugo Brunet pour être Hugo, et à Victor Dujardin pour m'avoir semé dans la côte du Malaise durant les 5 miles de LLN. Une pensée spéciale va à mon père spirituel Eugen Pircalabelu

pour avoir été toujours à l'écoute, chaleureux et un collègue formidable à tout point de vue. Thank you for those very nice years! Plus généralement, merci à tous ceux avec qui nous avons partagé des moments agréables, ce sont ces moments qui ont rendu un tel travail possible.

En tant qu'assistant à l'ISBA, nous avons la chance de pouvoir collaborer avec l'équipe du SMCS, composée de gens formidables. Merci en particulier à Benjamin Colling pour nos nombreuses réunions du vendredi après-midi sur LSTAT2040 qui terminent en Orval au Café des Halles, à Alain Guillet pour les nombreuses pauses improvisées lors de tes ballades au premier et à Laura Vermeren avec qui nous parviendrons peut-être un jour à percer le secret des Elementals !

L'équipe administrative de l'ISBA est formée de gens disponibles et toujours prêts à aider. Merci en particulier à Nancy Guillaume, Sophie Malali, Tatiana Regout and Marguerite-Marie Hanon pour votre gentillesse.

Je souhaite vivement remercier les Professeurs Anna Kiriliouk, Rainer von Sachs, Karim Barigou et Anne Sabourin d'avoir accepté de faire partie de mon jury, et plus généralement pour leur aide précieuse durant mon parcours doctoral.

Merci à Papa, Maman et mes frères pour m'avoir soutenu durant ces quatre années qui ont été remplies de joie mais aussi de moments difficiles. Merci à mes amis Dominique, Mathias et Albert pour tous les bons moments qu'on a partagés.

Last but not least, merci à Lara Wautier, le L, avec qui j'ai le plaisir de partager ma vie et un superbe appartement, toujours animé, particulièrement depuis que nous le partageons également avec notre petite Golden, Nixie. Je suis heureux à vos côtés.

Contents

Preamble i

Acknowledgments iii

Table of Contents v

1 Extreme Value Analysis: An Introduction 1
- 1.1 Extrapolation and regular variation . . . 1
- 1.2 Univariate risks . . . 5
- 1.3 Multivariate risks . . . 8

2 Uniform Concentration Bounds for Frequencies of Rare Events 15
- 2.1 Introduction . . . 15
- 2.2 Inequalities for rare events . . . 17
- 2.3 Working directly with the maximal deviation . . . 18
- 2.4 Working with the expected maximal deviation . . . 21
- 2.5 Comparison of the bounds . . . 23
- 2.A An improvement of the Vapnik–Chervonenkis inequality . . . 24
- 2.B Vapnik–Chervonenkis inequality for relative deviations . . . 27
- 2.C Proof of Theorem 2.3.2 . . . 28

3 Concentration Bounds for the Empirical Angular Measure with Statistical Learning Applications 33
- 3.1 Learning from multivariate extremes . . . 34
- 3.2 Upper tail dependence . . . 38
  - 3.2.1 Regular variation and exponent measure . . . 38
  - 3.2.2 Angular measure . . . 39
  - 3.2.3 Data standardization and ranks . . . 39
  - 3.2.4 The empirical angular measure . . . 40
- 3.3 Concentration inequalities . . . 41
  - 3.3.1 Framing random sets . . . 41
  - 3.3.2 Concentration bounds for the empirical angular measure 43
  - 3.3.3 Examples of collections of subsets of the sphere . . . 47
- 3.4 Applications to statistical learning . . . 48
  - 3.4.1 Minimum-volume set estimation . . . 48
  - 3.4.2 Classification in extreme regions . . . 50

3.5 Simulation experiments . . . 55
3.6 Conclusion . . . 58
3.A Concentration inequality for rare events . . . 59
3.B Auxiliary results . . . 61
3.C Proof of Theorem 3.3.1 . . . 63
3.D Proofs of Remark 3.3.3 . . . 69
3.E Proofs of examples . . . 78
3.F Proofs of classification application . . . 80
3.G Proofs of simulations experiments . . . 84

4 Bivariate Asymptotic Theory and Testing for the Angular Measure 87
4.1 Introduction . . . 88
4.2 Regular variation and the angular measure . . . 91
4.2.1 Regular variation . . . 91
4.2.2 Angular measure . . . 92
4.2.3 Maximum empirical Euclidean likelihood estimator . . . 93
4.3 Asymptotic functional expansion of the empirical angular measure . . . 96
4.3.1 Weak convergence of tail empirical processes . . . 96
4.3.2 Asymptotic expansions . . . 97
4.4 Goodness-of-fit tests for parametric models . . . 102
4.4.1 Test statistic . . . 102
4.4.2 Parameter estimator and empirical stable tail dependence function . . . 102
4.4.3 Limit distribution of test statistic and estimation of critical values . . . 104
4.5 Simulations and finite-sample performance . . . 106
4.5.1 Logistic model . . . 106
4.5.2 Hüsler–Reiss model . . . 110
4.6 Testing the Hüsler–Reiss model for river discharge data . . . 113
4.7 Conclusion . . . 115
4.A Proof of weak convergence results . . . 116
4.B Link between densities of Λ and Φ_p . . . 123
4.C Proof of Theorem 4.3.1 . . . 124
4.C.1 Reduction to $\theta \in [0, \pi/4]$ . . . 125
4.C.2 Weak convergence of a tail empirical process . . . 127
4.C.3 Asymptotic expansions of the three terms . . . 128
4.D Proof of Corollary 4.3.1 . . . 138
4.E Proof of Theorem 4.4.1 . . . 139
4.F Proof of Theorem 4.4.2 . . . 143
4.G Proof of Theorem 4.4.3 . . . 144

4.H Sampling from the asymptotic distribution of the test statistic 159

5 Wasserstein–Aitchison Generative Adversarial Networks for angular measures 163
5.1 Introduction 164
5.2 Background 165
5.2.1 A brief literature survey 165
5.2.2 Regular variation and extreme value analysis 166
5.2.3 Wasserstein Generative Adversial Networks 172
5.2.4 Aitchison coordinates 175
5.3 Combining GAN, Aitchison coordinates and Extremes 177
5.4 Numerical experiments 183
5.4.1 Performance and computation 183
5.4.2 Simulated data: logistic dependence structure 186
5.4.3 Financial data: daily returns of industry portfolios . . 189
5.5 Conclusions 193
5.A Proof of Proposition 5.2.1 194
5.B The Aitchison simplex 195
5.C Angular measure with respect to another norm 197
5.D Implementation of competing methods 198
5.E Additional figures 200

6 Concluding Remarks and Perspectives 203

Bibliography 207

Extreme Value Analysis: An Introduction

1

As opposed to "standard" statistical approaches which are focused on modelling the bulk of the data distribution, Extreme Value Analysis (EVA) is focusing on providing accurate models for the tail of the data distribution. The core motivation of EVA is that, even though no historical data point lies in a tail region, such events may well happen in the future. If the data represents some financial loss or some stress applied to a physical system, such events may well have disastrous consequences! Therefore, we are interested in obtaining a general and robust theory for the possible tail behaviours of data distributions.

In this world, nothing comes for free. As such, carrying an EVA requires hypotheses. Of course, we are interested in obtaining the most general results and therefore using the least restrictive assumptions. The key mathematical concepts underlying the ones we use in EVA is *regular variation* of measures (and functions). Many formulations and frameworks are possible to introduce this concept. Here, we choose to follow the path of considering the involved measures in the space $\mathcal{M}_0$ as introduced in [74] since we think that this leads to a more elegant formulation than the traditional one developed, e.g., by Resnick [107] using vague convergence of positive Radon measures on an artificially compactified space. Implications for statistical analyses are identical and this choice merely reflects the author's preferences.

1.1 Extrapolation and regular variation

Extrapolation. Consider a positive multivariate risk

$$X = (X_1, \ldots, X_d)$$

with values in $\mathbb{E} \stackrel{\text{def}}{=} [0, \infty)^d$. Denote by P_X the law of X on $(\mathbb{E}, \mathcal{B}(\mathbb{E}))$, where $\mathcal{B}(\mathbb{E})$ denotes the Borel σ-algebra on $\mathbb{E}$ and let $\mathbb{E}_0 \stackrel{\text{def}}{=} \mathbb{E} \setminus \{(0, \ldots, 0)\}$. Assume we have collected historical data on this risk, that is, we are given a random sample $X_1, \ldots, X_n$ distributed as P_X. We will always assume that the observations $X_i = (X_{i1}, \ldots, X_{id})$ are independent. It could be that, in this sample, no (or few) observation has been observed in a given Borel set $B \in \mathcal{B}(\mathbb{E})$ for which we are interested in obtaining an assessment of $P_X(B)$, which we

think to be nonzero. For example, we may have $B = tC$ where $C \in \mathcal{B}(\mathbb{E})$ and $tC \overset{\text{def}}{=} \{tx : x \in C\}$ for a large $t > 0$. A simple computation based on the empirical measure $\widehat{P}_X \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ would be useless in this case. Two approaches are possible:

I. either we assume that $P_X = P_\theta$ for some finite-dimensional parameter θ and use the historical data to provide an estimate $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of the unknown parameter. In particular, this leads to the estimate $P_{\widehat{\theta}}(B)$ of the probability of interest;

II. or we assume some homogeneity on P_X, at least in its tail part, so that we may relate the tail probability $P_X(tC)$ to the more "central" probability $P_X(C)$ for which an empirical estimation $\widehat{P}_X(C)$ is possible.

The first approach implies that we are able to provide a parametric model for the whole distribution P_X. This is quite restrictive, even more in the case $d > 1$ which is the main topic of this thesis. EVA looks to avoid strong assumptions on the data generating process and follows a guiding principle of Vladimir Vapnik (born in 1936) in [124, Section 1.9]:

When solving a given problem, try to avoid solving a more general problem as an intermediate step.

In this context, this means that when trying to model the tail of X, one should avoid modelling the whole distribution P_X as an intermediate step. As such, we choose to follow the second approach. The simplest form this path could take is to assume that the measure P_X is *positive homogenous* of order $-\alpha < 0$, that is

$$P_X(tC) = t^{-\alpha} P_X(C), \qquad \forall t > 1. \tag{1.1}$$

Of course, this is too restrictive in practice as nearly all familiar distribution P_X *do not* satisfy (1.1). Consequently, we need to be more cautious and provide a more general working assumption. This assumption is regular variation. We start by introducing it for functions on $\mathbb{R}$ as it will be useful for the one-dimensional case $d = 1$ described below, and it is a necessary step towards defining regular variation of measures on $\mathbb{E}$.

Regular variation of functions. A function $F : (0, \infty) \to (0, \infty)$ is called regularly varying (at infinity) when it is asymptotically positively homogeneous.

Definition 1.1.1 (Regular variation of functions)**.** *A measurable $F : (0, \infty) \to (0, \infty)$ is called regularly varying with index $\rho \in \mathbb{R}$ if*

$$\lim_{x \to \infty} \frac{F(tx)}{F(x)} = t^\rho, \qquad \forall t > 0.$$

This is denoted by $F \in \mathrm{RV}_\rho$. *If* $\rho = 0$, *we say that* F *is* slowly varying.

Definition 1.1.1 formalises the idea that F satisfies

$$F(tx) \sim t^\rho F(x), \qquad \text{as } x \to \infty.$$

It also directly follows from Definition 1.1.1 that if $F \in \mathrm{RV}_\rho$, it can be written as

$$F(x) = L(x)x^\rho, \qquad \forall x \in (0, \infty) \tag{1.2}$$

where $L(x) \stackrel{\text{def}}{=} F(x)/x^\rho$ is slowly varying, that is, F behaves roughly like a power function at infinity.

Examples of slowly varying functions include functions L such that $L(x) \to C$ for some constant $C > 0$ as $x \to \infty$, logarithmic functions like $L(x) = \log(1 + x)$ or compositions of it like $L(x) = \log(\log(1 + x))$. Functions like $x > 0 \mapsto \exp(x)$ or $x > 0 \mapsto \sin(x)$ are *not* slowly varying.

In the case of one-dimensional risks $X \in (0, \infty)$ described below, we will see that assuming that the data distribution P_X is regularly varying is equivalent to assuming that its survival function

$$\overline{F}(x) \stackrel{\text{def}}{=} \mathrm{P}(X > x), \qquad x > 0, \tag{1.3}$$

satisfies $\overline{F} \in \mathrm{RV}_{-\alpha}$ for some $\alpha > 0$.

Regular variation of measures. Before introducing the key concept underlying our extrapolation approach, we need some definitions. For $x \in \mathbb{E}$ and $r > 0$, we will use the notation

$$B_r(x) \stackrel{\text{def}}{=} \{y \in \mathbb{E} : \|y - x\| \leq r\},$$

where $\|\cdot\|$ denotes some norm on $\mathbb{R}^d$. We will consider the space of measures

$$\mathcal{M}_0(\mathbb{E}) \stackrel{\text{def}}{=} \Big\{\mu \text{ measure on } (\mathbb{E}_0, \mathcal{B}(\mathbb{E}_0)) \mid \forall r > 0, B \in \mathcal{B}(\mathbb{E}_0) \text{ with } B \subseteq \mathbb{E}_0 \setminus B_r(0),\ \mu(B) < \infty\Big\}$$

and the spaces of functions

$$C_b(\mathbb{E}) \stackrel{\text{def}}{=} \Big\{f : \mathbb{E} \to \mathbb{R} \mid f \text{ is continuous and bounded}\Big\}$$

and

$$C_0(\mathbb{E}) \stackrel{\text{def}}{=} C_b(\mathbb{E}_0) \cap \Big\{f : \mathbb{E}_0 \to \mathbb{R} \mid \exists r > 0,\ f(x) = 0\ \forall x \in B_r(0)\Big\}$$

Regular variation relies on the following notion of convergence [74, Theorem 2.1].

Definition 1.1.2 (Convergence of measures in $\mathcal{M}_0(\mathbb{E})$)**.** *Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be measures in $\mathcal{M}_0(\mathbb{E})$. We say that μ_n converges to μ in the space $\mathcal{M}_0(\mathbb{E})$ when*

$$\lim_{n\to\infty} \int_{\mathbb{E}} f(x)\mathrm{d}\mu_n(x) = \int_{\mathbb{E}} f(x)\mathrm{d}\mu(x), \qquad \forall f \in C_0(\mathbb{E}).$$

This is denoted by $\mu_n \xrightarrow{\mathcal{M}_0} \mu$.

As for weak convergence of probability measures, we have a Portmanteau theorem which will be useful in interpreting the regular variation assumption [74, Theorem 2.4]. For a set $B \subseteq \mathbb{E}$, we will denote its closure by $\overline{B}$ and its boundary by ∂B.

Theorem 1.1.1 (Portmanteau theorem for $\mathcal{M}_0(\mathbb{E})$ convergence)**.** *Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be measures in $\mathcal{M}_0(\mathbb{E})$. The following statements are equivalent.*

(i) $\mu_n \xrightarrow{\mathcal{M}_0} \mu$

(ii) $\lim_{n\to\infty} \mu_n(B) = \mu(B)$ *for any* $B \in \mathcal{B}(\mathbb{E})$ *such that* $\mu(\partial B) = 0$ *and* $0 \notin \overline{B}$.

Many formulation exists for the regular variation assumption. The most useful one for us is the following [74, Theorem 3.1].

Definition 1.1.3 (Regular variation of measures)**.** *We say that a measure $\mu \in \mathcal{M}_0(\mathbb{E})$ is regularly varying if one of the two following equivalent conditions is satisfied:*

(i) there exists a nonzero measure $\nu \in \mathcal{M}_0(\mathbb{E})$ and a function $b \in \mathrm{RV}_{-\alpha}$ for some $\alpha > 0$ such that

$$t\mu(b(t)\cdot) \xrightarrow{\mathcal{M}_0} \nu(\cdot), \qquad t \to \infty;$$

(ii) there exists a nonzero measure $\nu \in \mathcal{M}_0(\mathbb{E})$ and a function $c \in \mathrm{RV}_{\alpha}$ for some $\alpha > 0$ such that

$$c(t)\mu(t\cdot) \xrightarrow{\mathcal{M}_0} \nu(\cdot), \qquad t \to \infty.$$

Coming back to our statistical problem of estimating $P_X(tC)$ for some large $t > 0$, we understand that Definition 1.1.3 can play a key role. Indeed, assuming $P_X \in \mathcal{M}_0(\mathbb{E})$ is regularly varying, we get the following approximation from condition (ii):

$$P_X(tC) \approx \frac{\nu(C)}{c(t)}, \qquad t \to \infty,$$

provided C satisfies $\nu(\partial C) = 0$ and $0 \notin \overline{C}$. Assuming this relation is exact for t large enough, we get an estimate by computing an estimate of the unknown

measure ν and the number $c(t)$. Dealing with $c(t)$ won't represent a big challenge as it follows from the proof of [74, Theorem 3.1] that we may pick

$$c(t) = \frac{1}{\mu(\mathbb{E} \setminus B_0(t))}, \qquad t > 0, \tag{1.4}$$

and this will be estimated empirically on the sample for some (intermediate) large $t > 0$. Statistical estimation of the measure ν represents the main difficulty.

The problem is not fully unstructured as any measure ν satisfying one of the equivalent statements of Definition 1.1.3 is automatically positive homogeneous [74, Theorem 3.1]. That is, for any $B \in \mathcal{B}(\mathbb{E})$ and $t > 0$, we must have

$$\nu(tB) = t^{-\alpha}\nu(B) \tag{1.5}$$

for the same $\alpha > 0$ such that $c \in \mathrm{RV}_\alpha$. Indeed, we have for $t, s > 0$, by point (ii) of Definition 1.1.3,

$$\lim_{s\to\infty} c(ts)\mu(tsB) = \nu(B),$$

and, at the same time, in view of Definition 1.1.1,

$$\lim_{s\to\infty} c(ts)\mu(tsB) = \lim_{s\to\infty} \frac{c(ts)}{c(s)} c(s)\mu(s(tB)) = t^{\alpha}\nu(tB),$$

showing (1.5). In some sense, the somewhat unrealistic assumption in (1.1) becomes true for the tail measure ν.

In the next section, we start by explaining the estimation problem in dimension $d = 1$ which is much easier as no dependence structure is involved and the measure ν takes a simple form. Next, we move on to $d > 1$ which is the main topic of this thesis.

1.2 Univariate risks

Let X be a random variable with values in $[0, \infty)$. Assuming that its distribution P_X is regularly varying leads to a particularly simple form for ν due to the homogeneity property (1.5).

Proposition 1.2.1 (One-dimensional regular variation)**.** *Let $\mu \in \mathcal{M}_0([0, \infty))$ be regularly varying, i.e., there exists a nonzero measure $\nu \in \mathcal{M}_0([0, \infty))$ and a regularly varying function $b : t > 0 \mapsto b(t) \in (0, \infty)$ such that*

$$t\mu(b(t)\cdot) \xrightarrow{\mathcal{M}_0} \nu(\cdot), \qquad t \to \infty.$$

There exists $\alpha > 0$ and $c > 0$ such that for any $x > 0$,

$$\nu((x, \infty)) = cx^{-\alpha}. \tag{1.6}$$

As the class of intervals $\{(x,\infty) : x > 0\}$ forms a π-system generating $\mathcal{B}((0,\infty))$, it follows from an extension of the Dynkin's π–λ theorem [74, Theorem 2.8] that the measure ν is fully determined by (1.6).

Proof. Recall from (1.5) that for any $x > 0$ and any $t > 0$

$$\nu(t(x,\infty)) = t^{-\alpha}\nu((x,\infty))$$

for some $\alpha > 0$. In particular, picking $x = 1$, we get

$$\nu\left((t,\infty)\right) = t^{-\alpha}\nu\left((1,\infty)\right),$$

for any $t > 0$. Setting $c \overset{\text{def}}{=} \nu\left((1,\infty)\right)$, we get the result. □

In the previous proof, we implicitly assume that $\nu((1,\infty)) > 0$, an assumption that we will always assume to be satisfied. In (1.6), it is always possible to have $c = 1$ by changing the function b as a regularly varying function is not sensitive to multiplication by a constant. Consequently, we have shown that assuming the regular variation of $X \in (0,\infty)$ is equivalent to existence of a function $b(\cdot)$, with $b(t) \to \infty$ as $t \to \infty$, and such that for all $x > 0$,

$$\lim_{t\to\infty} t\,\mathsf{P}\left(\frac{X}{b(t)} > x\right) = x^{-\alpha}$$

for some $\alpha > 0$. It is now a standard argument [107, Theorem 3.6] that this requirement is equivalent to the survival function $\overline{F}$ of X defined in (1.3) satisfying $\overline{F} \in \mathrm{RV}_{-\alpha}$.

Regular variation of the risk's survival function is equivalent to the risk being roughly distributed as a Pareto random variable in the tail, modulo a term which does not change too quickly at infinity. This assumption is of course more general than assuming that X is Pareto distributed itself. Examples of families of distributions P_X satisfying such an hypothesis are: Pareto, Burr, Fisher, inverse Gamma, Fréchet, Student, etc. See [13, Table 2.1] for other examples.

From the statistical perspective, the hypothesis $\overline{F} \in \mathrm{RV}_{-\alpha}$ is crucial as it tells us that as $x \to \infty$,

$$\mathsf{P}(X > tx) \approx t^{-\alpha}\, P(X > x), \tag{1.7}$$

for any $t > 0$. This precisely means that extrapolation is possible. If one is interested in obtaining an estimate of $\mathsf{P}(X > y)$ for some large $y > 0$ where no empirical data is available above y, we just rewrite $y = t \cdot y/t$ for some large intermediate $0 < t < y$ such that $\mathsf{P}(X > y/t)$ can be evaluated empirically

by $\widehat{P}_X((y/t,\infty))$ and, provided we are able to provide an estimator $\widehat{\alpha}$ for the unknown parameter $\alpha > 0$, we deduce the estimate

$$\widehat{\mathrm{P}(X > y)} = t^{-\widehat{\alpha}}\, \widehat{P}_X((y/t,\infty))$$

for the probability of interest. The resulting semi-parametric estimate is the core of what is called in EVA the *Peaks-Over-Threshold* method, or POT in short.

It is clear from the previous reasoning that in this one-dimensional situation, the key question is: how to obtain the estimator $\widehat{\alpha}$? One way to do so is to use a simple plug-in estimator in the regular variation assumption. This leads to the popular *Hill estimator*.

Hill estimator. The estimator of $\alpha > 0$ is based on a simple estimate of the tail measure v in the regular variation assumption. Recall that P_X is supposed to satisfy

$$t\, P_X\,(b(t)\cdot) \xrightarrow{\mathcal{M}_0} v(\cdot)$$

where, as suggested by (1.4), we may pick $b(t) = c^{-1}(t) = (1/\overline{F})^{-1}(t) = F^{-1}(1-1/t)$ for $t > 1$. The inverse has to be considered in a generalized sense if F is not strictly increasing. Considering $t = n/k$ with $k = k(n)$ such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ and sample replacement leads to the following estimator

$$\widehat{v}_n(\cdot) \overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\left\{\frac{X_i}{X_{n-k:n}} \in \cdot\right\}$$

for the measure v, since $\widehat{F}^{-1}(1-k/n) = X_{n-k:n}$ where $X_{1:n} < X_{2:n} < \ldots < X_{n:n}$ represents the ordered random sample from P_X. Showing that $\widehat{v}_n$ converges weakly to v as a random element of $\mathcal{M}_0([0,\infty))$ is outside the scope of this introduction, but see [107, Chapter 4] for details in the language of Radon measures.

From the Portmanteau Theorem 1.1.1, we expect to have, for $x > 0$,

$$\widehat{v}_n((x,\infty)) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\left\{\frac{X_i}{X_{n-k:n}} > x\right\} \approx x^{-\alpha}.$$

Multiplying both sides by x^{-1} and integrating with respect to x on $(1,\infty)$ leads to

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{n} \int_1^\infty \mathbb{I}\left\{\frac{X_i}{X_{n-k:n}} > x\right\} \frac{\mathrm{d}x}{x} &= \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right) \\ &\approx \int_1^\infty x^{-(1+\alpha)}\mathrm{d}x = \alpha^{-1}.\end{aligned}$$

The Hill estimator of α is therefore defined as

$$\widehat{\alpha}_H \overset{\text{def}}{=} \left(\frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right) \right)^{-1}. \tag{1.8}$$

Consistency of this estimator follows from the behaviour of the random measure $\widehat{v}_n$ [107, Theorem 4.2]. See [34] for its asymptotic normality.

We now have all the key ingredients to perform tail extrapolation under regular variation based on (1.7). Of course, the one-dimensional situation is far from being solved.

- How to verify that the distribution P_X of our data is regularly varying so that extrapolation may take place?
- How to choose k in (1.8)? This is a standard bias/variance trade-off inherent to EVA methods: picking k larger allows for a large effective sample size, therefore reducing the variance of the estimate, but this comes at the price of including more central order statistics $X_{n-i+1:n}$, therefore increasing the bias of the estimation of the tail parameter α. Many techniques exist for choosing k, often by visual inspection of the so-called *Hill plot* or variants of it; however, in many data applications, these plots are uninformative and the choice remains a difficult question.
- Allowing for a more general framework than the regular variation introduced here permits to deal with more general tail behaviours than the ones discussed here. Chapter 5 will introduce the precise assumptions.

As we are mainly concerned with multivariate risks X and use the regular variation in a different manner, we will not try to answer those questions here nor to explore the huge amount of probabilistic and statistical considerations related to one-dimensional extremes. Many excellent textbooks exist on that topic, see [13, 34, 106, 107] for example.

1.3 Multivariate risks

Recall $\mathbb{E} = [0, \infty)^d$ and $\mathbb{E}_0 = \mathbb{E} \setminus \{(0, \dots, 0)\}$. Consider a risk $X = (X_1, \dots, X_d)$ with values in $\mathbb{E}$. As explained before, we are interested in estimating P_X in tail regions of its support where few or no historical data points are to be found. As such, we rely on the regular variation assumption introduced in Definition 1.1.3.

An unfortunate consequence of the regular variation assumption is that, if it is assumed directly on P_X, it implies that all margins X_j and X_k of the risk are *tail equivalent* [107, Remark 6.1], that is

$$\lim_{t \to \infty} \frac{\mathrm{P}(X_j > t)}{\mathrm{P}(X_k > t)} = \lambda_{jk} \in (0, \infty)$$

for any $j, k \in \{1, \dots, d\}$. This is a direct consequence of Definition 1.1.3 and Proposition 1.2.1. Indeed, defining $B_j = [0, \infty)^{j-1} \times [x, \infty) \times [0, \infty)^{d-j} \in \mathcal{B}(\mathbb{E})$ for some $x > 0$, we have, from Theorem 1.1.1, that there exists a regularly varying function $c : (0, \infty) \to (0, \infty)$ and a nonzero measure $\nu \in \mathcal{M}_0(\mathbb{E})$ such that

$$\lim_{t \to \infty} c(t)\, \mathrm{P}(X_j > tx) = \nu_j((x, \infty)).$$

where ν_j is j-th margin of ν, which we assume again to satisfy $\nu_j((1, \infty)) > 0$. From Proposition 1.2.1, we have

$$\nu_j((x, \infty)) = c_j x^{-\alpha_j},$$

for some $c_j > 0$ and $\alpha_j > 0$. Since this holds for any $j \in \{1, \dots, d\}$ and any $x > 0$, we get

$$\lim_{t \to \infty} \frac{\mathrm{P}(X_j > t)}{\mathrm{P}(X_k > t)} = \lim_{t \to \infty} \frac{c(t)\mathrm{P}(X_j > t)}{c(t)\mathrm{P}(X_k > t)} = \frac{c_j}{c_k} \stackrel{\text{def}}{=} \lambda_{jk},$$

showing tail equivalence of X_j and X_k.

Copula approach. In practice, marginal risks are often *not* tail equivalent. As such, we avoid assuming that Definition 1.1.3 is satisfied with $\mu = P_X$. Instead, we follow a *copula approach* and separate the dependence analysis from the marginal ones, which can be done using the considerations explained in the previous section and references therein. We now solely focus on the extremal dependence of X which is the main challenge in multivariate EVA and forms the main motivation for this work.

Let

$$V \stackrel{\text{def}}{=} \left(\frac{1}{1 - F_1(X_1)}, \dots, \frac{1}{1 - F_d(X_d)} \right), \tag{1.9}$$

where F_j is the cumulative distribution function of X_j, be the unit Pareto margins version of X. Assuming that F_j is continuous for every $j \in \{1, \dots, d\}$, the joint distribution of V only involves the (unique) copula of X. Hence, the regular variation assumption will be placed on the distribution P_V of V, instead of P_X. Note that, due to the unit Pareto margins of P_V, we have

$$P_V(\mathbb{E} \setminus B_0(t)) \sim ct^{-1}, \qquad \text{as } t \to \infty,$$

for some $c > 0$, so that, by (1.4), we may take $c(t) = t$ in the regular variation assumption on P_V. By Theorem 1.1.1, we get that there exists a nonzero measure $\nu \in \mathcal{M}_0(\mathbb{E})$ such that for any $B \in \mathcal{B}(\mathbb{E})$ with $\nu(\partial B) = 0$ and $0 \notin \overline{B}$,

$$\lim_{t \to \infty} t\, P_V(tB) = \nu(B). \tag{1.10}$$

Such a case is sometimes referred to as *standard regular variation*, see [107, Section 6.5.6]. Note that, from (1.5), we must have

$$\nu(t\cdot) = t^{-1}\nu(\cdot), \qquad \forall t > 0. \tag{1.11}$$

Unlike the case $d = 1$, relation (1.11) does not characterize ν up to a finite-dimensional parameter. Estimation on it has to be done *nonparametrically*.

Angular measure. Mathematically, we will benefit from the fact that the collection of sets

$$C_{r,A} \stackrel{\text{def}}{=} \{x \in \mathbb{E} : \|x\| > r, x/\|x\| \in A\}$$

for $r > 0$ and Borel sets $A \subseteq \mathbb{S} \stackrel{\text{def}}{=} \{x \in \mathbb{E} : \|x\| = 1\}$ such that $\nu(\partial C_{r,A}) = 0$ forms a *convergence determining class* for the convergence of measures in $\mathcal{M}_0(\mathbb{E})$ as introduced in Definition 1.1.2, see [74, Theorem 4.1]. If we denote this class of sets by $\mathcal{A}$, we thus have that regular variation of P_V as in (1.10) is equivalent to

$$\lim_{t\to\infty} t\, P_V(C_{r,A}) = \nu(C_{r,A}), \qquad \forall\, C_{r,A} \in \mathcal{A}. \tag{1.12}$$

Note that, by (1.11),

$$\nu(C_{r,A}) = r^{-1}\nu(C_{1,A}), \qquad \forall\, r > 0 \text{ and } A \subseteq \mathbb{S},$$

so that the main quantity of interest for the analysis of the tail of P_V becomes

$$\Phi(A) \stackrel{\text{def}}{=} \nu(C_{1,A}), \qquad \text{for Borel sets } A \subseteq \mathbb{S} \tag{1.13}$$

such that $\nu(\partial C_{1,A}) = 0$. This defines a (finite) measure Φ on $\mathbb{S}$ which will be called the *angular measure* of V (and therefore, of X). This object is the central quantity of interest in this thesis. The geometry of the problem is illustrated in dimension $d = 2$ for the L_∞-norm in Figure 1.1.

Intuitively, the angular measure captures the extremal dependence of V in a convenient and intuitive way: by (1.10) and (1.13), we have for Borel sets $A \subseteq \mathbb{S}$ such that $\nu(\partial C_{1,A}) = 0$,

$$\Phi(A) = \lim_{t\to\infty} t\, \mathrm{P}\left(\|V\| > t,\ \frac{V}{\|V\|} \in A\right). \tag{1.14}$$

In other words, $\Phi(A)$ computes the (appropriately scaled) probability of finding the "angular" component $V/\|V\|$ of V in A when its radius $\|V\|$ becomes large. Another way to phrase things comes from introducing the "polar coordinates" map T defined by

$$T : x \in \mathbb{E}_0 \mapsto T(x) \stackrel{\text{def}}{=} \left(\|x\|, \frac{x}{\|x\|}\right) \in (0,\infty) \times \mathbb{S}. \tag{1.15}$$

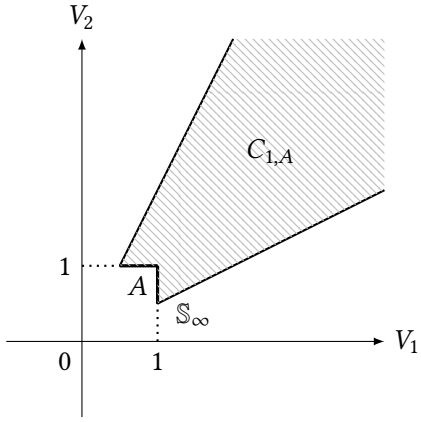


Figure 1.1: The angular measure Φ of the set $A \subset \mathbb{S}_\infty$, the unit sphere associated to the L_∞-norm, corresponds to the measure ν of the cone $C_{1,A}$ associated to A.

Then, Φ can be seen as the angular component of the *push-forward* measure of ν by the map T in the sense that for any $r > 0$ and $a \in \mathbb{S}$,

$$\mathrm{d}(\nu \circ T^{-1})(r, a) = \frac{\mathrm{d}r}{r^2} \otimes \mathrm{d}\Phi(a),$$

or, equivalently, for any ν integrable $f : \mathbb{E} \to \mathbb{R}$,

$$\int_{\mathbb{E}} f(x)\, \mathrm{d}\nu(x) = \int_{\mathbb{S}} \int_0^\infty f(ra)\, \frac{\mathrm{d}r}{r^2} \mathrm{d}\Phi(a). \tag{1.16}$$

It is then easy to see that ν is fully determined by Φ as the radial component of it in polar coordinates has known density, due to the scaling property (1.11). Therefore, inference on Φ is the key point to carry out a multivariate EVA.

Note that not *any* finite measure on $\mathbb{S}$ leads to a valid angular measure. Indeed, since V has unit Pareto margins, we have for $t > 0$,

$$\nu(E_j(t)) = t^{-1}, \qquad \text{where } E_j(t) \overset{\text{def}}{=} \{x \in \mathbb{E} : x_j \geq t\}, \qquad \forall j \in \{1, \ldots, d\}.$$

Picking $f = \mathbb{I}_{E_j(t)}$ in (1.16) leads to

$$t^{-1} = \int_{\mathbb{S}} \int_0^\infty \mathbb{I}(r \geq t a_j^{-1})\, \frac{\mathrm{d}r}{r^2} \mathrm{d}\Phi(a) = t^{-1} \int_{\mathbb{S}} a_j \mathrm{d}\Phi(a),$$

leading to the marginal constraints

$$\int_{\mathbb{S}} a_j \mathrm{d}\Phi(a) = 1, \qquad \forall j \in \{1, \ldots, d\}. \tag{1.17}$$

However, those constraints still leave room for infinitely many choices and no parametric structure is evident from the problem.

Statistical estimation. Relation (1.14) provides the basis for statistical inference on Φ. Indeed, recall that we are given historical observations $X_1, \dots, X_n$ from P_X which we assume to be independent. Since we follow a copula approach and work with standard margins instead of the original distribution, we start transforming the observations to

$$\widehat{V}_i \overset{\text{def}}{=} \left(\frac{1}{1 - \widehat{F}_j(X_{ij})} \right)_{j=1}^{d} = \left(\frac{n+1}{n+1-R_{ij}} \right)_{j=1}^{d}, \qquad i \in \{1, \dots, n\} \tag{1.18}$$

where

$$\widehat{F}_j(t) \overset{\text{def}}{=} \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}\left\{ X_{ij} \leq t \right\}, \qquad t \in \mathbb{R} \tag{1.19}$$

is the (scaled to avoid division by zero in the definition of $\widehat{V}_i$) empirical cumulative distribution function of X_j and R_{ij} is the rank of the observation X_{ij} in the marginal sample $X_{1j}, \dots, X_{nj}$. Starting from the sample in (1.18), one may apply the same idea as in the one-dimensional case and consider the relation (1.14) exact for some large enough data dependent $t > 0$, and replace the unknown P_V by its (approximated) empirical counterpart

$$\widehat{P}_{\widehat{V}} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{\widehat{V}_i}. \tag{1.20}$$

Taking $t = n/k$ for some intermediate $k = k(n)$ such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ leads to the *empirical angular measure*

$$\widehat{\Phi}(A) \overset{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}\left\{ \|\widehat{V}_i\| > \frac{n}{k}, \frac{\widehat{V}_i}{\|\widehat{V}_i\|} \in A \right\}, \qquad A \subseteq \mathbb{S}. \tag{1.21}$$

This estimator and the statistical procedures based on it are studied in the following chapters, under many perspectives.

Note that the main probabilistic challenge related to the study of the finite sample and large sample behaviour of $\widehat{\Phi}$ is that the terms in the sum in (1.20) or (1.21) are *not* independent. This is due to the empirical standardisation in (1.18) which makes the procedure depending on the ranks of the data instead of the original observations.

The marginal constraints (1.17) are not verified by the estimator $\widehat{\Phi}$. Imposing them (at least, in a soft manner) in the estimation process may be desirable and is considered in some of the chapters of this thesis.

Two-dimensional case. In the case $d = 2$, the polar transform T introduced in (1.15) is often replaced by the transformation

$$x = (x_1, x_2) \in \mathbb{E}_0 \mapsto (\|x\|, \arctan(x_2/x_1)) \in (0, \infty) \times [0, \pi/2]$$

as the angular component now truly corresponds to the angle of the vector (x_1, x_2) with respect to the first element of the standard orthonormal basis on $\mathbb{R}^2$. The push-forward measure of ν by this map is also referred to as the angular measure and is denoted by the same letter Φ. This has the advantage of simplifying the analysis as the measure is now supported on a bounded interval of the real line and this will be considered in Chapter 4.

Uniform Concentration Bounds for Frequencies of Rare Events

2

This chapter is based on
Stéphane Lhaut, Anne Sabourin and Johan Segers
Uniform concentrations bounds for frequencies of rare events
Satistics and Probability Letters, 189:109610, 2022.

Abstract

New Vapnik–Chervonenkis type concentration inequalities are derived for the empirical distribution of an independent random sample. Focus is on the maximal deviation over classes of Borel sets within a low probability region. The constants are explicit, enabling numerical comparisons.

2.1 Introduction

Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$. We always let μ_n denote the associated *empirical measure*, i.e.,

$$\mu_n \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \delta_{X_i}.$$

We already know, from [123, Theorem 2], the *Vapnik and Chervonenkis (VC) inequality* which states that

$$\mathrm{P}\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t \right) \leq 4 S_{\mathcal{A}}(2n) \exp(-nt^2/8), \tag{2.1}$$

where $\mathcal{A}$ is a non-empty class of Borel sets on $\mathbb{R}^d$ and

$$S_{\mathcal{A}}(n) \stackrel{\text{def}}{=} \max_{x_1, \ldots, x_n \in \mathbb{R}^d} \left| \left\{ \{x_1, \ldots, x_n\} \cap A : A \in \mathcal{A} \right\} \right|$$

is the shattering coefficient of the class $\mathcal{A}$—for background on VC theory, we refer to [88]. Setting the upper bound equal to $\delta \in (0,1)$ and solving for t yields that, with probability $1-\delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2}{n}\Big[\log(4/\delta) + \log S_{\mathcal{A}}(2n)\Big]}. \tag{2.2}$$

In Appendix 2.A, we show an improvement of this inequality that will be used in Section 2.2: with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n}} + \sqrt{\frac{2}{n}\Big[\log(4/\delta) + \log S_{\mathcal{A}}(2n)\Big]}.$$

Our main goal is to develop bounds with *explicit constants* when working with *low probability regions* (Section 2.2), i.e., it will be assumed that there exists some Borel set $\mathbb{A} \subseteq \mathbb{R}^d$ which contains every set $A \in \mathcal{A}$ and has a "small" probability

$$p \overset{\text{def}}{=} \mu(\mathbb{A}). \tag{2.3}$$

For example, consider an empirical risk minimization problem where we use the empirical risk to classify some data. In that situation, p could be 1% if we are looking to understand how the prediction behaves when the input variables take values above the 99th percentile of the distribution of $\|X\|$, where X has the same distribution as the data in the training set— we refer to [60, Remark 5], [76] and [25, Section 4.2] for classification in extreme regions.

The bounds should get smaller since all the encountered events have a small probability; however, the estimation is more difficult since we have less data in those regions. In this rare events framework, a first improvement of the VC inequality (2.2), based on [6, Theorem 2.1] and [88, Theorem 1.11], is given by

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2p}{n}\Big[\log(12/\delta) + \log S_{\mathcal{A}}(2n)\Big]}. \tag{2.4}$$

The details of the statement and the proof are deferred to Appendix 2.B. Even though the factor p in the square root will improve the bound, this result is still not fully satisfactory. The *effective sample size*, i.e., the expected number of data points in the low probability region, is np rather than n. Therefore, the shattering coefficient involved in the bound seems too large.

This work is heavily motivated by [60] who already introduced a VC type inequality adapted to rare events. However, unlike the classical VC bound, the constants appearing in their result are *not* explicit. A similar remark can be made about [56] where concentration inequalities are also introduced for normalized empirical processes, in a very general setting. In Section 2.3 and

Section 2.4, we discuss different methods that lead to new, explicit, inequalities. Essentially, there are two possibilities: either we directly apply the standard tools of VC theory on the maximal deviation itself (Section 2.3), or we use a variation of the classical *McDiarmid bounded differences inequality* [90, Theorem 3.8] and then we deal with the expectation $\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right]$ (Section 2.4). After deriving the different bounds, we compare them in Section 2.5.

Our aim is not to be exhaustive in the bounds that we derive. Other tools can lead to other inequalities. As an example, in [25, Theorem A.1], we provide a bound on the expectation of the supremum using chaining techniques, which leads to a better asymptotic rate, but which is inaccurate for realistic sample sizes due to the large constants inherent to those methods. Therefore, such results are not presented here. A detailed comparison can be found in [82].

Remark 2.1.1 (Pointwise measurability). To ensure measurability of the supremum appearing in (2.1), a common hypothesis is to assume that the class $\mathcal{A}$ is *pointwise measurable*, as suggested by [122, Example 2.3.4], i.e., that there exists a subclass $\mathcal{A}_0 \subseteq \mathcal{A}$, at most countable, such that for every $A \in \mathcal{A}$, there exists a sequence $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{A}_0$ such that

$$\lim_{n\to\infty} \mathbb{I}_{A_n}(x) = \mathbb{I}_A(x), \quad \text{for every } x \in \mathbb{R}^d.$$

It is easily showed that, under this assumption, $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| = \sup_{A\in\mathcal{A}_0} |\mu_n(A) - \mu(A)|$ and we may replace $\mathcal{A}$ by $\mathcal{A}_0$ everywhere.

2.2 Inequalities for rare events

The main tool to develop adapted bounds is expressed in the following lemma, which already appears in [60, Page 17] and the idea of which is likely to be found in other places in the literature as well, see, e.g., [98, Equation (14.6)] for a result in the same spirit. The proof is elementary and can be found in [82, Chapter 3].

Lemma 2.2.1 (Conditioning trick)**.** *Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. Let $Y_1, \ldots, Y_n$ be an independent random sample with common distribution*

$$\mu_{\mathbb{A}}(\cdot) = \mu(\cdot \mid \mathbb{A}) = \frac{\mu(\cdot \cap \mathbb{A})}{\mu(\mathbb{A})} = \frac{\mu(\cdot \cap \mathbb{A})}{p},$$

and independent of $X_1, \ldots, X_n$. If, for every $k = 1, \ldots, n$ we denote μ_k^Y the empirical measure associated to $Y_1, \ldots, Y_k$ and if $K \sim \mathrm{Bin}(n, p)$ denotes the

number of data points X_i in $\mathbb{A}$, then the following equality in distribution holds

$$\left[(\mu_n(A))_{A\in\mathcal{A}} \,\middle|\, K = k\right] \stackrel{d}{=} \left(\frac{k}{n}\mu_k^Y(A)\right)_{A\in\mathcal{A}},$$

in the sense of equality of finite-dimensional distributions. In particular, under the pointwise measurability assumption (Remark 2.1.1), this equality still holds when considering the supremum over the class $\mathcal{A}$.

This result provides us with a convenient way to adapt classical VC inequalities to our setting: we start by conditioning μ_n on K, then we apply concentration inequalities on the empirical measure μ_K^Y which takes account of the rare nature of the encountered events and finally we integrate out on K. In this last step, we will make use of the *Bernstein's inequality* for binomial random variables [88, Theorem 1.5] which states that for every $t > 0$,

$$\mathsf{P}\big(|K - np| \geq t\big) \leq 2\exp\left(-\frac{t^2}{2(np(1-p) + t/3)}\right). \tag{2.5}$$

When we will be working with the expectation, we shall also make use of another version of the VC inequality (2.2) for the expected maximal deviation:

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right] \leq \sqrt{\frac{2\log(2S_{\mathcal{A}}(2n))}{n}}, \tag{2.6}$$

the proof of which can be found, for example, in [88, Theorem 1.9].

In the bounds that we propose, we will need to define the shattering coefficient $S_{\mathcal{A}}(x)$ for a *real* parameter $x > 0$. It is simply understood that $S_{\mathcal{A}}(x) \stackrel{\text{def}}{=} S_{\mathcal{A}}([x])$, where $[x]$ denotes the integer part of x.

2.3 Working directly with the maximal deviation

Using the improvement of the VC inequality that we show in Appendix 2.A, we derive a first inequality.

Theorem 2.3.1. *Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. If $np \geq 4\log(4/\delta)$, where $\delta \in (0,1)$, then we have, with probability $1-\delta$,*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{2}{3n}\log(4/\delta) + \sqrt{\frac{p}{n}}\left(\sqrt{2\log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log S_{\mathcal{A}}(4np)} + 1\right).$$

Proof of Theorem 2.3.1. Let $t > 0$. If $K = \sum_{i=1}^{n} \mathbb{I}\{X_i \in \mathbb{A}\} \sim \mathrm{Bin}(n, p)$, then by Lemma 2.2.1,

$$\begin{aligned}
&\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{|\mu_n(A) - \mu(A)|}{p} \geq t\right) \\
&\quad = \mathsf{E}\left[\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{|\mu_n(A) - \mu(A)|}{p} \geq t \,\middle|\, K\right)\right] \\
&\quad = \mathsf{E}\left[\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\left|\frac{K}{n}\mu_K^Y(A) - \mu(A)\right|}{p} \geq t \,\middle|\, K\right)\right] \\
&\quad = \mathsf{E}\left[\mathsf{P}\left(\sup_{A\in\mathcal{A}} \left|\frac{K}{np}\mu_K^Y(A) - \mu_{\mathbb{A}}(A)\right| \geq t \,\middle|\, K\right)\right] \\
&\quad \leq \mathsf{E}\left[\mathsf{P}\left(\frac{K}{np} \sup_{A\in\mathcal{A}} \left|\mu_K^Y(A) - \mu_{\mathbb{A}}(A)\right| + \left|\frac{K}{np} - 1\right| \geq t \,\middle|\, K\right)\right],
\end{aligned}$$

where the last inequality follows from the triangle inequality and the fact that $\mu_{\mathbb{A}}(A) \leq 1$ for any $A \in \mathcal{A}$. Let $t = u + s$ for $u, s > 0$. Then, the latter quantity is bounded by

$$\mathsf{P}\left(\left|\frac{K}{np} - 1\right| \geq s\right) + \mathsf{E}\left[\mathsf{P}\left(\frac{K}{np} \sup_{A\in\mathcal{A}} \left|\mu_K^Y(A) - \mu_{\mathbb{A}}(A)\right| \geq u \,\middle|\, K\right) \mathbb{I}\left\{\left|\frac{K}{np} - 1\right| \leq s\right\}\right]. \tag{2.7}$$

Let $\delta \in (0, 1)$. We will pick $u = u(\delta)$ and $s = s(\delta)$ such that each term in (2.7) is bounded by $\delta/2$.

We deal with the first term of (2.7) directly using the Bernstein inequality (2.5),

$$\begin{aligned}
\mathsf{P}\left(\left|\frac{K}{np} - 1\right| \geq s\right) &\leq \mathsf{P}\big(|K - np| \geq nps\big) \\
&\leq 2\exp\left(-\frac{n^2p^2s^2}{2(np(1-p) + nps/3)}\right) \\
&\leq 2\exp\left(-\frac{nps^2}{2(1 + s/3)}\right).
\end{aligned}$$

The positive root s_+ associated with the quadratic equation obtained by equaling this last term to $\delta/2$ satisfies

$$s_+ \leq \frac{2}{3np}\log(4/\delta) + \sqrt{\frac{2}{np}\log(4/\delta)} \stackrel{\text{def}}{=} s(\delta).$$

To deal with the second term of (2.7), we use the improved VC inequality

that we develop in Appendix 2.A in the form (2.14),

$$\mathsf{P}\left(\frac{K}{np}\sup_{A\in\mathcal{A}}\left|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)\right|\geq u\,\middle|\,K\right)$$
$$\leq 4S_{\mathcal{A}}(2K)\exp\left(-\frac{(np)^2u^2}{2K}\left(1-\sqrt{\frac{2K}{(np)^2u^2}}\right)-\frac{1}{4}\right).$$

By monotonicity in K on the region of interest [1], we get

$$\mathsf{E}\left[\mathsf{P}\left(\frac{K}{np}\sup_{A\in\mathcal{A}}\left|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)\right|\geq u\,\middle|\,K\right)\mathbb{I}\left\{\left|\frac{K}{np}-1\right|\leq s\right\}\right]$$
$$\leq 4S_{\mathcal{A}}(2np(1+s))\exp\left(-\frac{npu^2}{2(1+s)}\left(1-\sqrt{\frac{2(1+s)}{npu^2}}\right)-\frac{1}{4}\right).$$

If $np\geq 4\log(4/\delta)$, we have $s(\delta)\leq\frac{1}{6}+\frac{1}{\sqrt{2}}\leq 1$. Hence, using $s=1$ in the latter expression, we find that it equals $\delta/2$ if

$$u=u(\delta)\stackrel{\text{def}}{=}\frac{1}{\sqrt{np}}+2\sqrt{\frac{1}{np}\left[\log(8/\delta)+\log S_{\mathcal{A}}(4np)\right]}.$$

Regrouping the two terms, we obtain that with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}}\frac{|\mu_n(A)-\mu(A)|}{p}$$
$$\leq\frac{2}{3np}\log(4/\delta)+\sqrt{\frac{1}{np}}\left(\sqrt{2\log(4/\delta)}+2\sqrt{\log(8/\delta)+\log S_{\mathcal{A}}(4np)}+1\right).$$

Multiplying both sides by p, we get the result. □

Another possibility consists of symmetrizing the process, based on our improved version of the classical argument (Lemma 2.A.1), before using the conditioning trick. It leads to a simpler bound. The proof is deferred to Appendix 2.C.

Theorem 2.3.2. *Let $X_1,\ldots,X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. If $np\geq 2\log(8/\delta)$, where $\delta\in(0,1)$, then we have, with probability $1-\delta$,*

$$\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\leq\sqrt{\frac{2p}{n}}\left(2\sqrt{\log(8/\delta)+\log S_{\mathcal{A}}(8np)}+1\right).$$

[1] The map $x\in(0,+\infty)\mapsto -x(1-\sqrt{1/x})=-x+\sqrt{x}$ is decreasing on $(1/4,+\infty)$. Our focus is on that region of the real line since, in our situation, $x=\frac{(np)^2u^2}{2K}$ with $u=u(\delta)\geq 1/\sqrt{np}$ and $K\leq np(1+s(\delta))\leq 2np$.

2.4 Working with the expected maximal deviation

The variation of the McDiarmid bounded differences inequality that we use is recalled in [60, Proposition 11]. When combined with [60, Lemma 12], it leads to a convenient Bernstein type concentration inequality for the maximal deviation that is particularly adapted when working with rare events, i.e., for every $t > 0$,

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| - \mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right] \geq t\right) \leq \exp\left(-\frac{nt^2}{4p + 2t/3}\right).$$

Setting the upper bound equal to $\delta \in (0,1)$ and solving for $t > 0$ proves the following result.

Proposition 2.4.1 (Concentration of the maximal deviation)**.** *Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \frac{2}{3n}\log(1/\delta) + 2\sqrt{\frac{p}{n}\log(1/\delta)} + \mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right].$$

Using this inequality, we may directly work with the expectation

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right]$$

to obtain a bound with high probability for $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$. We start by applying the conditioning trick (Lemma 2.2.1) on the expectation.

Lemma 2.4.1. *Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. Then,*

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right] \leq \mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}} |\mu_K^Y(A) - \mu_{\mathbb{A}}(A)|\right] + \sqrt{\frac{p}{n}},$$

where $\mu_{\mathbb{A}}$, μ_K^Y and K are as in Lemma 2.2.1.

Proof of Lemma 2.4.1. By Lemma 2.2.1 and a computation similar to the one in the beginning of the proof of Theorem 2.3.1, we have

$$\begin{aligned}
\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right] &= \mathsf{E}\left[\mathsf{E}\left[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \,\middle|\, K\right]\right] \\
&= \mathsf{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{K}{n}\left(\mu_K^Y(A) - \mu_{\mathbb{A}}(A)\right) + \left(\frac{K}{n} - p\right)\mu_{\mathbb{A}}(A)\right|\right] \\
&\leq \mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}} \left|\mu_K^Y(A) - \mu_{\mathbb{A}}(A)\right|\right] + \mathsf{E}\left[\left|\frac{K}{n} - p\right|\right].
\end{aligned}$$

The second term is easily bounded using Jensen's inequality and the fact that $K \sim \mathrm{Bin}(n,p)$:

$$\left(\mathsf{E}\left[\left|\frac{K}{n}-p\right|\right]\right)^2 \le \mathsf{E}\left[\left(\frac{K}{n}-p\right)^2\right] = \mathsf{Var}\left[\frac{K}{n}-p\right] = \frac{np(1-p)}{n^2} \le \frac{p}{n}. \qquad \square$$

To deal with the remaining expectation $\mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}}|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)|\right]$, we will make use of (2.6). Furthermore, we will also assume that the VC dimension

$$V_{\mathcal{A}} \stackrel{\text{def}}{=} \sup\{n\in\mathbb{N} : S_{\mathcal{A}}(n) = 2^n\}$$

of the class $\mathcal{A}$ is finite to be able to apply the famous *Sauer's Lemma* [21, Lemma 1] which states that for every $n\in\mathbb{N}$,

$$S_{\mathcal{A}}(n) \le (n+1)^{V_{\mathcal{A}}}. \tag{2.8}$$

The purpose of this assumption is that, whenever $V_{\mathcal{A}} < +\infty$, we will be able to use Jensen's inequality which is sharper than the monotonicity arguments underlying every proof in Section 2.3.

Proposition 2.4.2. *Let $X_1,\dots,X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. Then, if $V_{\mathcal{A}} < +\infty$,*

$$\mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}}|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)|\right] \le \sqrt{\frac{2p}{n}\Big[\log 2 + V_{\mathcal{A}}\log(2np+1)\Big]}.$$

Proof of Proposition 2.4.2. By the VC inequality for the expectation (2.6) combined with Sauer's Lemma (2.8) and Jensen's inequality, we have

$$\begin{aligned}
\mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}}\left|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)\right|\right] &= \mathsf{E}\left[\mathsf{E}\left[\frac{K}{n}\sup_{A\in\mathcal{A}}\left|\mu_K^Y(A)-\mu_{\mathbb{A}}(A)\right| \,\middle|\, K\right]\right] \\
&\le \mathsf{E}\left[\frac{K}{n}\sqrt{\frac{2\log\left(2S_{\mathcal{A}}(2K)\right)}{K}}\right] \\
&\le \frac{1}{n}\mathsf{E}\left[\sqrt{2K\Big[\log(2)+V_{\mathcal{A}}\log(2K+1)\Big]}\right] \\
&\le \frac{1}{n}\sqrt{2\mathsf{E}[K]\Big[\log(2)+V_{\mathcal{A}}\log(2\mathsf{E}[K]+1)\Big]} \\
&= \sqrt{\frac{2p}{n}\Big[\log(2)+V_{\mathcal{A}}\log(2np+1)\Big]},
\end{aligned}$$

where we made use of the concavity of the map

$$K \mapsto \sqrt{2K\Big[\log(2)+V_{\mathcal{A}}\log(2K+1)\Big]}$$

(the derivative is clearly decreasing) and the fact that $K\sim\mathrm{Bin}(n,p)$. $\square$

Combining all the ingredients of this section, we finally get a bound on the maximal deviation.

Corollary 2.4.1. *Let $X_1, \dots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. Then, if $V_{\mathcal{A}} < +\infty$, we have, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \le \frac{2}{3n}\log(1/\delta) + \sqrt{\frac{2p}{n}}\left(\sqrt{2\log(1/\delta)} + \sqrt{\log 2 + V_{\mathcal{A}}\log(2np+1)} + \frac{\sqrt{2}}{2}\right). \quad (2.9)$$

2.5 Comparison of the bounds

We obtained various bounds on the maximal deviation $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$ which hold with probability $1-\delta \in (0,1)$. Before comparing them, we summarize the main results of our paper and we recall the bound derived in Theorem 2.B.1:

- Theorem 2.3.1 (working directly on the maximal deviation and symmetrizing the process after the application of the conditioning trick): whenever $np \ge 4\log(4/\delta)$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \le \frac{2}{3n}\log(4/\delta) + \sqrt{\frac{p}{n}}\left(\sqrt{2\log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log S_{\mathcal{A}}(4np)} + 1\right). \quad (2.10)$$

- Theorem 2.3.2 (working directly on the maximal deviation and symmetrizing the process before the application of the conditioning trick): whenever $np \ge 2\log(8/\delta)$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \le \sqrt{\frac{2p}{n}}\left(2\sqrt{\log(8/\delta) + \log S_{\mathcal{A}}(8np)} + 1\right). \quad (2.11)$$

- Corollary 2.4.1 (working on the expected maximal deviation and using Jensen's inequality): whenever $V_{\mathcal{A}} < +\infty$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \le \frac{2}{3n}\log(1/\delta) + \sqrt{\frac{2p}{n}}\left(\sqrt{2\log(1/\delta)} + \sqrt{\log 2 + V_{\mathcal{A}}\log(2np+1)} + \frac{\sqrt{2}}{2}\right). \quad (2.12)$$

- Theorem 2.B.1 (based on [6, Theorem 2.1]): whenever $np \geq 8\log(3/\delta)/3$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2p}{n}\big[\log(12/\delta) + \log S_{\mathcal{A}}(2n)\big]}. \tag{2.13}$$

We observe that the bound (2.12) seems to beat (2.10) in any case. Indeed, both inequalities have the same structure: one term decaying like n^{-1} and one term decreasing like $n^{-1/2}$; however, those terms seem always lower in the concentration bound obtained through the use of McDiarmid's inequality. Even though it could be that the shattering coefficient of $\mathcal{A}$ grows slower than the rate induced by the use of Sauer's Lemma (2.8) in (2.12), the fact that (2.8) enables us to use Jensen's inequality heavily compensates the loss.

The comparison between (2.11) and (2.12) is a bit harder since the structure is not the same anymore. Nevertheless, the coefficient of the leading term in (2.11) seems larger than the one in (2.12). Hence, even though the inequality obtained in Section 2.4 could be worse for small samples due to its additional term, we think that for practical sample sizes, it remains the best bound.

In Figure 2.1, we provide a graphical comparison of our results with the more classical VC inequality for relative deviations (2.13) on the simplest, one-dimensional, VC class that we may think off: $\mathcal{A} = \{(-\infty, t] : t \in \mathbb{R},\ t \leq Q(10^{-3})\}$, where $Q(p) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : F(x) \geq p\}$ is the left-continuous inverse of $F(x) \stackrel{\text{def}}{=} \mu((-\infty, x]))$, the cumulative distribution function associated to μ. In this situation, we easily verify that $S_A(n) = n + 1$ and that the main assumption of our paper (2.3) is satisfied with $p = 10^{-3}$. Note also that $\mathcal{A}$ is pointwise measurable (Remark 2.1.1). We used $\delta = 10^{-2}$.

One clearly observes better performance of our bound (2.12) in this situation, which is valid for every sample size. The bounds coming from Section 2.3 are closer to the older bound; however, it seems that for samples with a size larger than 10^6 data points, they perform a bit better. More importantly, those bounds take into account the reduced effective sample size np in their measure of size complexity $S_{\mathcal{A}}$, which is not the case with (2.13).

2.A An improvement of the Vapnik–Chervonenkis inequality

We show an improvement of the VC inequality [82, Theorem 2.17].

Theorem 2.A.1. *Let $X_1, \ldots, X_n$ be an independent sample with common distribution μ on $\mathbb{R}^d$ and let μ_n be the associated empirical distribution. Consider a non-empty class $\mathcal{A}$ of Borel sets of $\mathbb{R}^d$. Then, for every $\delta \in (0, 1)$, with probability $1 - \delta$, we have*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{1}{2n}} + \sqrt{\frac{2}{n}\big[\log(4/\delta) + \log S_{\mathcal{A}}(2n)\big]}.$$

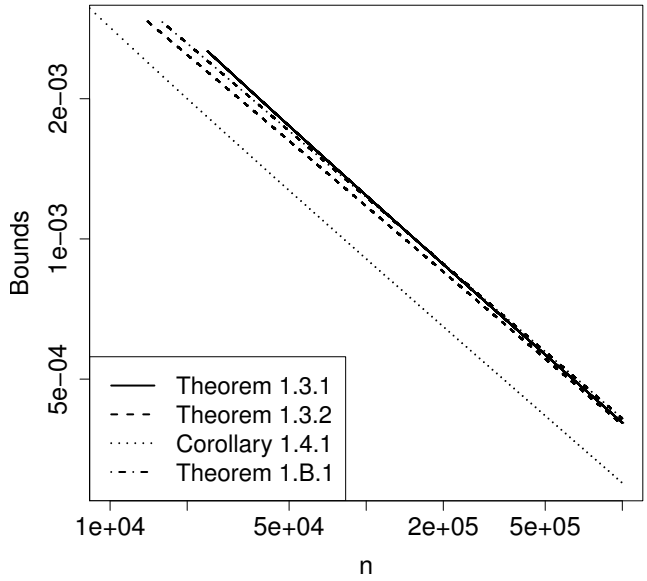


Figure 2.1: Comparison of the bounds for the empirical cumulative distribution function on sufficiently large samples (to satisfy the hypothesis of the different theorems). Axes are on logarithmic scale.

The order of the bound is the same as in the original VC bound (2.2). However, the constant in front of the square root is almost halved. The proof relies on a variation of a classical *symmetrization* argument [21, Lemma 2 p. 185].

Lemma 2.A.1 (Symmetrization)**.** *Let $X_1, \dots, X_n, X'_1, \dots, X'_n$ be an independent sample with common distribution μ on $\mathbb{R}^d$. Let μ_n denote the empirical distribution associated to $X_1, \dots, X_n$ and let μ'_n denote the empirical distribution associated to $X'_1, \dots, X'_n$. Consider a non-empty class $\mathcal{A}$ of Borel sets of $\mathbb{R}^d$. If $t > 0$ and $a \in (0,1)$ are such that $4(1-a)^2 nt^2 \geq 2$, then*

$$\left.\begin{array}{l} \mathrm{P}\big(\sup_{A\in\mathcal{A}}(\mu_n(A) - \mu(A)) \geq t\big) \\ \mathrm{P}\big(\sup_{A\in\mathcal{A}}(\mu(A) - \mu_n(A)) \geq t\big) \end{array}\right\} \leq 2\mathrm{P}\left(\sup_{A\in\mathcal{A}}(\mu_n(A) - \mu'_n(A)) \geq at\right).$$

The sample $X'_1, \dots, X'_n$ is often referred to as a "ghost sample". The choice $a = 1/2$ gives back the initial symmetrization. However, as we will see, this more flexible result will give a way to substantially improve the classical bound.

Proof of Lemma 2.A.1. Let $A \in \mathcal{A}$ and $t > 0$. We first observe that

$$\mathbb{I}\{\mu_n(A) - \mu(A) \geq t\}\mathbb{I}\{\mu'_n(A) - \mu(A) \leq (1-a)t\} \leq \mathbb{I}\{\mu_n(A) - \mu'_n(A) \geq at\}.$$

Integrating both sides with respect to the ghost sample, we get

$$\mathbb{I}\{\mu_n(A) - \mu(A) \geq t\}\mathsf{P}'\big(\mu'_n(A) - \mu(A) \leq (1-a)t\big) \leq \mathsf{P}'\big(\mu_n(A) - \mu'_n(A) \geq at\big),$$

where we used the notation P' to emphasize that the integration was realized only with respect to $X'_1, \ldots, X'_n$. By Bienaymé–Tchebychev inequality and hypothesis, we deduce

$$\begin{aligned}
\mathsf{P}'\big(\mu'_n(A) - \mu(A) \leq (1-a)t\big) &= 1 - \mathsf{P}'\big(\mu'_n(A) - \mu(A) \geq (1-a)t\big) \\
&\geq 1 - \frac{\mathrm{Var}[\mu'_n(A)]}{(1-a)^2t^2} = 1 - \frac{\mathrm{Var}\left[\mathbb{I}\{X'_1 \in A\}\right]}{n(1-a)^2t^2} \\
&\geq 1 - \frac{1}{4n(1-a)^2t^2} \geq 1/2,
\end{aligned}$$

where we used the fact that the variance of $\mathbb{I}\{X'_1 \in A\}$ is bounded by $1/4$ (since $\max_{p\in[0,1]}\{p(1-p)\} = 1/4$). We obtain

$$\mathbb{I}\{\mu_n(A) - \mu(A) \geq t\} \leq 2\mathsf{P}'\big(\mu_n(A) - \mu'_n(A) \geq at\big).$$

Since this relation holds for every $A \in \mathcal{A}$, we finally get that

$$\begin{aligned}
\mathbb{I}\{\sup_{A\in\mathcal{A}} \mu_n(A) - \mu(A) \geq t\} &= \sup_{A\in\mathcal{A}} \mathbb{I}\{\mu_n(A) - \mu(A) \geq t\} \\
&\leq 2 \sup_{A\in\mathcal{A}} \mathsf{P}'\big(\mu_n(A) - \mu'_n(A) \geq at\big) \\
&\leq 2\mathsf{P}'\big(\sup_{A\in\mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at\big).
\end{aligned}$$

Taking the expectation with respect to the sample $X_1, \ldots, X_n$, it follows that

$$\mathsf{P}\big(\sup_{A\in\mathcal{A}} (\mu_n(A) - \mu(A)) \geq t\big) \leq 2\mathsf{P}\big(\sup_{A\in\mathcal{A}} (\mu_n(A) - \mu'_n(A)) \geq at\big).$$

The other inequality has a similar proof. □

Proof of Theorem 2.A.1. Let $nt^2 \geq 1/2$. Following the same arguments as the ones used to prove the classical VC inequality [82, Theorem 2.14] and using our symmetrization, we show that for any $a \in (0,1)$, whenever $4(1-a)^2nt^2 \geq 2$, we have

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \geq t\right) \leq 4S_{\mathcal{A}}(2n)\exp\left(-a^2nt^2/2\right).$$

Choosing the largest $a \in (0,1)$ such that this relation holds, i.e., $a = a(n,t) = 1 - \sqrt{1/(2nt^2)}$, we get

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \geq t\right) \leq 4S_{\mathcal{A}}(2n) \exp\left\{-\frac{nt^2}{2}\left(1 - \sqrt{\frac{2}{nt^2}}\right) - \frac{1}{4}\right\}. \tag{2.14}$$

Setting the upper bound equal to $\delta \in (0,1)$ and solving for t leads to a quadratic equation for t whose positive solution corresponds to the bound proposed in the theorem.

If $nt^2 < 1/2$, the bound (2.14) is trivial. □

2.B Vapnik–Chervonenkis inequality for relative deviations

Theorem 2.B.1 (VC inequality for relative deviations)**.** *Let $X_1, \ldots, X_n$ be an independent random sample with common distribution μ on $\mathbb{R}^d$ and μ_n be the associated empirical measure. Let $\mathcal{A}$ be a non-empty class of Borel sets on $\mathbb{R}^d$. Let $\mathbb{A}$ and p be as in* (2.3)*. If $np \geq (8/3)\log(3/\delta)$, where $\delta \in (0,1)$, then we have, with probability $1-\delta$,*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2p}{n}\left[\log(12/\delta) + \log S_{\mathcal{A}}(2n)\right]}.$$

Proof of Theorem 2.B.1. It follows from [88, Theorem 1.11] that

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t\right) \leq 4S_{\mathcal{A}}(2n)\exp(-nt^2/4) \tag{2.15}$$

and

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{\mu_n(A)}} \geq t\right) \leq 4S_{\mathcal{A}}(2n)\exp(-nt^2/4), \tag{2.16}$$

for every $t > 0$. Therefore, it is convenient to decompose our probability of interest as follows

$$\begin{aligned}
\mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{|\mu(A) - \mu_n(A)|}{\sqrt{p}} \geq t\right) &\leq \mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} \geq t\right) \\
&\quad + \mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{p}} \geq t\right) \\
&\leq 4S_{\mathcal{A}}(2n)\exp(-nt^2/4) + \mathsf{P}\left(\mu_n(\mathbb{A}) \geq 2p\right) \\
&\quad + \mathsf{P}\left(\sup_{A\in\mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{\mu_n(A)}} \geq t/\sqrt{2}\right) \\
&\leq 4S_{\mathcal{A}}(2n)\exp(-nt^2/4) + \mathsf{P}\left(\mu_n(\mathbb{A}) \geq 2p\right) \\
&\quad + 4S_{\mathcal{A}}(2n)\exp(-nt^2/8),
\end{aligned} \tag{2.17}$$

where we made use of (2.15) and (2.16). By hypothesis, the second term of (2.17) is bounded by $\delta/3$. Indeed, by Bernstein's inequality (2.5),

$$\mathsf{P}\left(\mu_n(\mathbb{A}) \geq 2p\right) = \mathsf{P}\left(K - np \geq np\right) \leq \exp\left(-\frac{3}{8}np\right),$$

where $K \sim \mathrm{Bin}(n, p)$ as in Lemma 2.2.1. Choosing

$$t = t(\delta) = 2\sqrt{\frac{2}{n}\left[\log(12/\delta) + \log S_{\mathcal{A}}(2n)\right]},$$

we get that the first term and last term of (2.17) are also bounded by $\delta/3$. The result follows. □

2.C Proof of Theorem 2.3.2

Proof of Theorem 2.3.2. Let $a \in (0, 1)$ and $t > 0$. By symmetrization (Lemma 2.A.1), if $4(1-a)^2nt^2 \geq 2$,

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \geq t\right) \leq 4\mathsf{P}\left(\sup_{A\in\mathcal{A}} (\mu_n(A) - \mu_n'(A)) \geq at\right),$$

where μ_n' is the empirical measure associated to $X_1', \ldots, X_n'$, an independent ghost sample with common distribution μ and independent of $X_1, \ldots, X_n$. We easily verify that, since $\mu(A) \leq p$ for all $A \in \mathcal{A}$, the symmetrization remains true under the weaker hypothesis that

$$(1-a)^2nt^2 \geq 2p. \tag{2.18}$$

It suffices to adapt the bound on the variance $\mathsf{Var}[\mathbb{I}\{X_i \in A\}]$ in the proof of Lemma 2.A.1.

Let $\tilde{K} = \sum_{i=1}^n \mathbb{I}\{X_i \in \mathbb{A} \text{ or } X_i' \in \mathbb{A}\} = \sum_{i=1}^n \mathbb{I}\{(X_i, X_i') \in \tilde{\mathbb{A}}\}$, where $\tilde{\mathbb{A}} = (\mathbb{A} \times \mathbb{R}^d) \cup (\mathbb{R}^d \times \mathbb{A})$. Then, if $(Y_i, Y_i')_{i=1}^n$ is an independent sample, also independent of $(X_i, X_i')_{i=1}^n$, with common distribution given by the conditional

distribution of (X_1, X_1') when it lies in $\tilde{\mathbb{A}}$, we have by Lemma 2.2.1

$$\begin{aligned}
&\mathsf{P}\left(\sup_{A\in\mathcal{A}}(\mu_n(A)-\mu_n'(A))\geq at\right)\\
&\qquad=\mathsf{P}\left(\sup_{A\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}\{X_i\in A\}-\mathbb{I}\{X_i'\in A\}\right)\geq at\right)\\
&\qquad=\mathsf{E}\left[\mathsf{P}\left(\sup_{A\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}\{X_i\in A\}-\mathbb{I}\{X_i'\in A\}\right)\geq at\,\middle|\,\tilde{K}\right)\right]\\
&\qquad=\mathsf{E}\left[\mathsf{P}\left(\frac{\tilde{K}}{n}\sup_{A\in\mathcal{A}}\frac{1}{\tilde{K}}\sum_{i=1}^{\tilde{K}}\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right)\geq at\,\middle|\,\tilde{K}\right)\right]\\
&\qquad=\mathsf{E}\left[\mathsf{P}\left(\sup_{A\in\mathcal{A}}\sum_{i=1}^{\tilde{K}}\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right)\geq nat\,\middle|\,\tilde{K}\right)\right].
\end{aligned}$$

Randomizing using *Rademacher random variables*, we have

$$\sum_{i=1}^{\tilde{K}}\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right)\overset{\mathcal{L}}{=}\sum_{i=1}^{\tilde{K}}\sigma_i\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right),$$

where $\tilde{K}$ is fixed and $\sigma_1,\ldots,\sigma_{\tilde{K}}$ are i.i.d. Rademacher random variables independent of $(Y_1, Y_1'),\ldots,(Y_n, Y_n')$, i.e., for every $i\in\{1,\ldots,\tilde{K}\}$,

$$\mathsf{P}(\sigma_i=1)=\mathsf{P}(\sigma_i=-1)=1/2$$

Conditioning on $(Y_i, Y_i')_{i=1}^n$ and applying Hoeffding's inequality [88, Theorem 1.2] to each of the $S_{\mathcal{A}}(2\tilde{K})$ possible vectors $(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\})$ that can arise as A ranges over $\mathcal{A}$, we obtain

$$\begin{aligned}
&\mathsf{P}\left(\sup_{A\in\mathcal{A}}\sum_{i=1}^{\tilde{K}}\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right)\geq nat\,\middle|\,\tilde{K}\right)\\
&=\mathsf{E}_Y\left[\mathsf{P}_\sigma\left(\sup_{A\in\mathcal{A}}\sum_{i=1}^{\tilde{K}}\sigma_i\left(\mathbb{I}\{Y_i\in A\}-\mathbb{I}\{Y_i'\in A\}\right)\geq nat\,\middle|\,\tilde{K}, Y_1, Y_1',\ldots,Y_n,Y_n'\right)\right]\\
&\leq\min\left\{1, S_{\mathcal{A}}(2\tilde{K})\exp\left(-\frac{a^2(nt)^2}{2\tilde{K}}\right)\right\},
\end{aligned}$$

where the last inequality follows from the fact that the expectation E_Y is taken only with respect to the random variables $Y_1, Y_1',\ldots,Y_n,Y_n'$. We deduce

$$\mathsf{P}\left(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\geq t\right)\leq 4\mathsf{E}\left[\min\left\{1, S_{\mathcal{A}}(2\tilde{K})\exp\left(-\frac{a^2(nt)^2}{2\tilde{K}}\right)\right\}\right]. \tag{2.19}$$

To be able to deal with this last quantity, we consider whether $\tilde{K}$ is less than $2np(1+s)$ or not, where $s > 0$ is to be determined. By monotonicity, it gives

$$\mathsf{E}\left[\min\left\{1, S_{\mathcal{A}}(2\tilde{K})\exp\left(-\frac{a^2(nt)^2}{2\tilde{K}}\right)\right\}\right] \leq \mathsf{P}\big(\tilde{K} \geq 2np(1+s)\big) + S_{\mathcal{A}}(4np(1+s))\exp\left(-\frac{na^2t^2}{4p(1+s)}\right).$$

Since $a \in (0,1)$ is not fixed, we may, given n and t, choose the largest $a < 1$ such that the condition (2.18) is respected. It is given by $a = a(n,t) = 1 - \sqrt{2p/(nt^2)}$. For such an a, we have

$$\mathsf{E}\left[\min\left\{1, S_{\mathcal{A}}(2\tilde{K})\exp\left(-\frac{a^2(nt)^2}{2\tilde{K}}\right)\right\}\right] \leq \mathsf{P}\big(\tilde{K} \geq 2np(1+s)\big) + S_{\mathcal{A}}(4np(1+s))\exp\left(-\frac{nt^2 - 2\sqrt{2np}t + 2p}{4p(1+s)}\right). \quad (2.20)$$

Let $\delta \in (0,1)$. We choose $s = s(\delta)$ such that the first term of (2.20) is bounded $\delta/8$. Then, we choose $t = t(\delta)$ such that the second term equals $\delta/8$. As a consequence of (2.19), we will deduce that $\mathsf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \geq t(\delta)\right) \leq \delta$, or, equivalently, $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq t(\delta)$ with probability at least $1-\delta$.

By Bernstein's inequality (2.5), since $\tilde{K} \sim \mathrm{Bin}(n, \tilde{p})$ with $\tilde{p} = p(2-p)$, we have

$$\begin{aligned}
\mathsf{P}\big(\tilde{K} \geq 2np(1+s)\big) &= \mathsf{P}\big(\tilde{K} - n\tilde{p} \geq 2np(1+s) - n\tilde{p}\big) \\
&= \mathsf{P}\big(\tilde{K} - n\tilde{p} \geq np(2s+p)\big) \\
&\leq \exp\left(-\frac{(np)^2(2s+p)^2}{2(n\tilde{p}(1-\tilde{p}) + np(2s+p)/3)}\right) \\
&\leq \exp\left(-\frac{np(2s+p)^2}{2((2-p) + (2s+p)/3)}\right) \\
&\leq \exp\left(-\frac{nps^2}{1+s/3}\right).
\end{aligned}$$

Equaling this expression to $\delta/8$, we obtain a quadratic equation in s whose positive solution is given by

$$s_+ = s(\delta) \stackrel{\text{def}}{=} \frac{\log(8/\delta)}{3np} + \sqrt{\frac{\log(8/\delta)}{np}}.$$

Since we assumed $np \geq 2\log(8/\delta)$, we have $s(\delta) \leq \frac{1}{6} + \frac{1}{\sqrt{2}} \leq 1$. Hence, by taking $s = 1$, the first term of (2.20) is bounded by $\delta/8$.

For such a choice of s, the second term of (2.20) becomes

$$S_{\mathcal{A}}(8np)\exp\left(-\frac{nt^2-2\sqrt{2np}t+2p}{8p}\right).$$

This last expression equals $\delta/8$ if

$$t=t(\delta)\stackrel{\text{def}}{=}2\sqrt{\frac{2p}{n}\left[\log(8/\delta)+\log S_{\mathcal{A}}(8np)\right]}+\sqrt{\frac{2p}{n}}. \qquad \square$$

Concentration Bounds for the Empirical Angular Measure with Statistical Learning Applications

3

This chapter is based on
Stéphan Clémençon, Hamid Jalalzai, Stéphane Lhaut, Anne Sabourin, and Johan Segers
Concentration bounds for the empirical angular measure with statistical learning applications
Bernoulli, 29(4):2797–2827, 2023.

Abstract

The angular measure on the unit sphere characterizes the first-order dependence structure of the components of a random vector in extreme regions and is defined in terms of standardized margins. Its statistical recovery is an important step in learning problems involving observations far away from the center. In the common situation that the components of the vector have different distributions, the rank transformation offers a convenient and robust way of standardizing data in order to build an empirical version of the angular measure based on the most extreme observations. However, the study of the sampling distribution of the resulting empirical angular measure is challenging. It is the purpose of the paper to establish finite-sample bounds for the maximal deviations between the empirical and true angular measures, uniformly over classes of Borel sets of controlled combinatorial complexity. The bounds are valid with high probability and, up to logarithmic factors, scale as the square root of the effective sample size. The bounds are applied to provide performance guarantees for two statistical learning procedures tailored to extreme regions of the input space and built upon the empirical angular measure: binary classification in extreme regions through empirical risk minimization and unsupervised anomaly detection through minimum-volume sets of the sphere.

3.1 Learning from multivariate extremes

Estimation and prediction problems regarding the extremal behaviour of a d-dimensional random vector $X = (X_1, \ldots, X_d)$ are of key importance for risk assessment in finance, insurance, engineering and environmental sciences and, recently, for the analysis of weak signals in machine learning. A standard assumption to model the joint upper tail of X is that its distribution lies in the maximal domain of attraction of a multivariate extreme value distribution. This assumption comprises two parts:

(i) the marginal distributions of X belong to the maximal domains of attraction of some univariate extreme value distributions;

(ii) after marginal transformation of X through the probability integral transform or a variation thereof, the joint distribution of the transformed vector belongs to the maximal domain of attraction of a multivariate extreme value distribution with pre-specified margins.

Under the side assumption that the marginal distributions of X are continuous, point (ii) above only involves the copula of X. Point (ii) can be imposed on its own, that is, independently of the assumptions on the margins in point (i), and this is what we will do in this paper.

The angular measure for multivariate extremes. The multivariate extreme value distribution in point (ii) is determined by a finite Borel measure, Φ, with support contained in the intersection of $[0, \infty)^d$ with the unit sphere on $\mathbb{R}^d$ with respect to some norm. This so-called angular measure, originally called spectral measure in [36], describes the first-order dependence structure of joint extremes of X and is rooted in the theory of multivariate regular variation [106]. Since then, it has been recognized that the modelling of extremal dependence may require finer assumptions than the traditional maximal domain of attraction condition, leading for instance to conditional extreme value models and the theory of hidden regular variation; see for instance [30, 128] and the references therein. Here, we focus on inference on the angular measure Φ_p with respect to some L_p-norm, for $p \in [1, \infty]$.

Inference on the angular measure is an important problem in extreme value analysis. It plays a part in the construction of confidence intervals for rare event probabilities [33, 37]. It lies at the basis for a test of the hypothesis that a bivariate distribution is in the maximum domain of attraction of an extreme value distribution [41]. It serves to model the action of a covariate on the extremal dependence of a baseline distribution through a density ratio model [31, 23]. The angular density is also at the basis of an estimator of bivariate tail quantile regions [40]. Bounds for probabilities of joint excesses

over high thresholds that are robust against misspecification of the angular measure are derived in [50]. The angular measure underlies techniques to find groups of variables exhibiting extremal dependence, with dimension reduction and sparse representations as objectives [24, 91, 81]; see [51] for a review. In [77], the spherical k-means algorithm applied to a sample from the angular measure yields prototypes of extremal dependence.

Furthermore, the angular measure is helpful for solving supervised and unsupervised learning tasks for sample points far away from the center of the distribution. In the spirit of principal component analysis, the eigendecomposition of the Gram matrix of the angular measure yields low-dimensional summaries of extremal dependence [29, 38]. Anomalous data can be detected from unusual combinations of variables being large simultaneously [59] or from their lack of membership of minimum-volume sets of the unit sphere containing a large fraction of the total mass of the angular measure [120]. Binary classification in extreme regions can be performed on the basis of the differences between the intra-class angular measures of the explanatory variables [76].

The empirical angular measure. For a given dimension d, the collection of all angular measures is subject only to some moment constraints stemming from the marginal standardization of X but does not form a parametric family. The usual considerations in favor of and against the use of parametric versus non-parametric methods therefore apply. Many parametric models have been proposed [26, 27] and new ones continue to be invented, with a growing emphasis on the use of flexible models almost of a semi-parametric nature [17, 113, 112].

Our focus is on non-parametric estimation and inference via the empirical angular measure, $\widehat{\Phi}_p$, introduced in [42] and generalized to every L_p-norms in [47], although already alluded to some years earlier [33, 37]. Given a random sample from an unknown distribution, the marginal standardization mentioned in point (ii) above is done by means of the empirical cumulative distribution functions. As a result, the estimator depends on the data only through the ranks. On the one hand, the use of ranks makes the method invariant with respect to marginal scales and reduces sensitivity to outliers. On the other hand, the additional dependence stemming from the ranks greatly complicates the study of the sampling distribution of the estimator.

In fact, the asymptotic distribution of the empirical angular measure is known only in the bivariate case [42, 47]. In higher dimensions, only its consistency has been established [77, Proof of Proposition 3.3]. This situation stands in contrast with the rank-based estimator of the stable tail dependence function, the asymptotic normality of which is known in any dimension [45, 11]. The reason why the treatment of the empirical angular measure is so

much more difficult is that it involves the empirical copula evaluated at sets of which the boundaries are not parallel to the coordinate axes. As a consequence, the usual argument to deal with the marginal empirical distribution functions via the functional delta method breaks down. Even outside the extreme value context, the asymptotic normality of the empirical copula in even a single non-rectangular set is to this day an open problem.

Variations on the empirical angular measure enforce the aforementioned moment constraints through empirical or Euclidean likelihoods [47, 32] as well as a folding technique [62]. Other variations exploit more specific assumptions on the marginal distributions to estimate them differently than by the empirical distribution functions, for instance by fitting generalized Pareto distributions to the univariate tails [43]. Sometimes, asymptotic results are established as if the marginal distributions are known. In our paper, we focus on the original, rank-based empirical angular measure and we make no assumptions on the unknown marginal cumulative distribution functions except for their continuity.

Concentration inequalities. It is the main goal of this paper to carry out a non-asymptotic analysis of the empirical angular measure $\widehat{\Phi}_p$, which, to the best knowledge, is the first of its kind. The concentration inequality established in our main result, Theorem 3.3.1, states that with a certain large probability $1-\delta$ the estimation error is not larger than a certain small quantity depending, among other things, on δ and on the number, k, of sample points used in the definition of the estimator. The bound concerns the worst-case estimation error

$$\sup_{A\in\mathcal{A}} \left|\widehat{\Phi}_p(A) - \Phi_p(A)\right| \tag{3.1}$$

over classes $\mathcal{A}$ of Borel subsets A of the unit sphere satisfying certain properties. In particular, the complexity of $\mathcal{A}$ comes into play via the Vapnik–Chervonenkis dimension of a collection of sets constructed from $\mathcal{A}$.

The result relies on two tools: first, the use of framing sets to capture certain random sets, as in [42] and [47], although the framing sets are defined in a different manner here (see Section 3.3.1); and second, a general concentration inequality for empirical processes indexed by rare events given in Theorem 3.A.1. The latter result is inspired by a similar inequality in [60] but the difference is that now the constant in the bound is explicit.

Although the angular measure Φ can be studied with respect to any norm on $\mathbb{R}^d$, this study is limited to the ones most frequently used in practice, the L_p-norms for $1 \le p \le \infty$, i.e., for $x \in \mathbb{R}^d$,

$$\|x\|_p \overset{\text{def}}{=} \begin{cases} (|x_1|^p + \cdots + |x_d|^p)^{1/p} & \text{if } 1 \le p < \infty, \\ \max(|x_1|, \ldots, |x_d|) & \text{if } p = \infty. \end{cases} \tag{3.2}$$

Applications to statistical learning. The concentration bounds for the empirical angular measure are leveraged to study two statistical learning problems: minimum-volume set estimation and binary classification in extreme regions. These two problems and the methods proposed to solve them were introduced in the conference papers [120] and [76], respectively. In both cases, the method was based upon the empirical angular measure. However, the technical difficulties inherent to the rank transformation were ignored and the theoretical analysis was performed as if the marginal distributions are known. Here, we apply concentration inequalities for the empirical angular measure to obtain finite-sample performance guarantees for the rank-based methods.

We briefly describe the two learning problems and the role of the angular measure. For unsupervised anomaly detection, the learning task can be formulated as minimum-volume set estimation on the sphere, that is, the statistical recovery of a Borel set of the sphere with minimum volume but containing a given, large fraction of the total mass of the angular measure [120]. Any large observation whose angle lies outside the minimum-volume set is considered as a potential anomaly. Concentration inequalities for the uniform estimation error over a class of candidate sets enable us in Section 3.4.1 to get statistical guarantees on the minimum-volume set selected within the class on the basis of an estimator of the angular measure.

The second problem concerns binary classification in extreme regions. Empirical risk minimization, which is the main paradigm of statistical learning theory, tends to ignore the predictive performance of candidate classifiers in low-density regions of the input space. Instead, we focus on the probability of classification error in such extreme regions. By controlling 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. In Theorem 3.4.1, we state a bound on the supremum of the empirical risk over a class of candidate classifiers.

Outline. The paper is organized as follows. In Section 3.2, the relevant notions pertaining to multivariate extreme value analysis are briefly recalled. The main result related to the non-asymptotic analysis of the empirical angular measure is formulated in Section 3.3. Section 3.4 illustrates the application of the concentration inequalities to two statistical learning problems, minimum-volume set estimation and binary classification in extreme regions. Numerical experiments are presented in Section 3.5. A discussion in Section 3.6 concludes the paper. Proofs and some auxiliary results, in particular a general concentration inequality for rare event probabilities, are deferred to the Appendices.

3.2 Upper tail dependence

We set out some basics of multivariate extreme value theory (Sections 3.2.1 and 3.2.2) and recall a rank-based standardization procedure (Section 3.2.3). Equipped with these notions, we describe the empirical angular measure (Section 3.2.4). For background, we refer to monographs such as [106, 13, 34].

3.2.1 Regular variation and exponent measure

Let $X = (X_1, \dots, X_d)$ be a random vector with distribution P and continuous marginal cumulative distribution functions $F_j(u) = \mathsf{P}(X_j \leq u)$ for $u \in \mathbb{R}$. We standardize each component of X to unit-Pareto margins by a combination of the probability integral transform and the quantile transform through $V_j \overset{\text{def}}{=} 1/(1 - F_j(X_j))$ for $j = 1, \dots, d$.

The working hypothesis in this paper is that the resulting random vector $V = (V_1, \dots, V_d)$ is multivariate regularly varying; this formalizes the assumption in point (ii) in the introduction. Specifically, we assume that there exists a non-zero Borel measure ν on the punctured orthant $\mathbb{E} = [0, \infty)^d \setminus \{(0, \dots, 0)\}$ which is finite on Borel sets bounded away from the origin and such that

$$\lim_{t \to \infty} t \, \mathsf{P}\left(t^{-1} V \in B\right) = \nu(B) \tag{3.3}$$

for all Borel sets B of $\mathbb{E}$ bounded away from the origin and such that $\nu(\partial B) = 0$, with ∂B the topological boundary of B. Convergence in (3.3) for all such B is equivalent to measure convergence $t\mathsf{P}(t^{-1}V \in \cdot) \to \nu(\cdot)$ as $t \to \infty$ in the space $\mathcal{M}_0$ of Borel measures on $\mathbb{E}$ that are finite on Borel sets bounded away from the origin [74]. Specifically, we have $\lim_{t\to\infty} t\mathsf{E}[f(t^{-1}V)] = \int_E f \mathrm{d}\nu$ for every bounded and continuous real function f on $\mathbb{E}$ vanishing in a neighbourhood of the origin. Alternatively, multivariate regular variation can be described in the language of vague convergence of Radon measures on the compactified, punctured orthant $[0, \infty]^d \setminus \{(0, \dots, 0)\}$ [106, 107].

The measure ν is referred to as the exponent measure because of its appearance in the exponent of the expression of the multivariate extreme value distribution to which the distribution of V is attracted [34, Definition 6.1.7]. The exponent measure is homogeneous: writing $\lambda A = \{\lambda x : x \in A\}$ for $\lambda > 0$ and $A \subseteq \mathbb{R}^d$, we have

$$\forall \lambda > 0, \qquad \nu(\lambda \cdot) = \lambda^{-1} \nu(\cdot). \tag{3.4}$$

Its margins are standardized: by (3.3) and since each component V_j is unit-Pareto distributed, we have

$$\forall y \in (0, \infty), \ \forall j \in \{1, \dots, d\}, \qquad \nu(\{x \in \mathbb{E} : x_j \geq y\}) = y^{-1}. \tag{3.5}$$

3.2.2 Angular measure

Recall $\|x\|_p$ in (3.2) and let

$$\mathbb{S}_p \stackrel{\text{def}}{=} \{x \in [0,\infty)^d : \|x\|_p = 1\}, \tag{3.6}$$

Consider the map $\theta_p : \mathbb{E} \to \mathbb{S}_p$ that assigns to any vector $x \in E$ its "angle" $\theta_p(x) \stackrel{\text{def}}{=} x/\|x\|_p$.

The angular measure Φ_p is defined as the push-forward measure by θ_p of the restriction of ν to $\{x \in \mathbb{E} : \|x\|_p \geq 1\}$: for any Borel set $A \subseteq \mathbb{S}_p$, we have

$$\Phi_p(A) = \nu(C_A) \quad \text{where} \quad C_A \stackrel{\text{def}}{=} \{x \in \mathbb{E} : \|x\|_p \geq 1,\ \theta_p(x) \in A\}. \tag{3.7}$$

In view of the marginal standardization (3.5) and the ensuing identity (3.9) below, the total mass $\Phi_p(\mathbb{S}_p) = \int_{\mathbb{S}_p} \|\theta\|_p \, \mathrm{d}\Phi_p(\theta)$ of the angular measure is finite and we have $1 \leq \Phi_p(\mathbb{S}_p) \leq d$.

The exponent measure ν is determined by the angular measure: by homogeneity (3.4),

$$\nu(\{x \in \mathbb{E} : \|x\|_p \geq u, \theta_p(x) \in A\}) = u^{-1}\Phi_p(A)$$

for every $u > 0$ and every Borel set $A \subseteq \mathbb{S}_p$. More generally, for any non-negative Borel measurable function f on $\mathbb{E}$, we have

$$\int_{\mathbb{E}} f \, \mathrm{d}\nu = \int_{\mathbb{S}_p} \int_0^\infty f(r\theta) \, \frac{\mathrm{d}r}{r^2} \, \mathrm{d}\Phi_p(\theta). \tag{3.8}$$

The combination of the marginal standardization (3.5) with the change-of-variable formula in (3.8) implies the identities

$$\forall j = 1, \ldots, d, \qquad \int_{\mathbb{S}_p} \theta_j \, \mathrm{d}\Phi_p(\theta) = 1. \tag{3.9}$$

In fact, any non-negative Borel measure Φ_p on $\mathbb{S}_p$ satisfying the moment constraints (3.9) is the angular measure of some random vector X. Indeed, from such a measure Φ_p on $\mathbb{S}_p$, we can define a measure ν on $\mathbb{E}$ through (3.8) and then consider the max-stable distribution with ν as exponent measure as in [36].

3.2.3 Data standardization and ranks

The component-wise transformation of X to a vector V with unit-Pareto margins is formalized by the map $v : \mathbb{R}^d \to [1,\ \infty]^d$, with

$$\forall x \in \mathbb{R}^d, \qquad v(x) \stackrel{\text{def}}{=} \left(\frac{1}{1 - F_1(x_1)}, \ldots, \frac{1}{1 - F_d(x_d)} \right), \tag{3.10}$$

with the convention $1/0 = \infty$. In this notation, we have $V = v(X)$.

Let $X_1, \ldots, X_n$ be an independent random sample from the distribution P of X, with $X_i = (X_{i1}, \ldots, X_{id})$. Since P is unknown in practice, it is replaced by its empirical version $P_n(\cdot) = n^{-1} \sum_{i=1}^n \mathbb{I}\{X_i \in \cdot\}$, where $\mathbb{I}\{\mathcal{E}\}$ denotes the indicator variable of the event $\mathcal{E}$. In particular, the marginal cumulative distribution functions F_j are substituted with their empirical counterparts

$$\widehat{F}_j(t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_{ij} \le t\}, \text{ for } t \in \mathbb{R} \text{ and } 1 \le j \le d,$$

in order to mimic the aforementioned standardization.

The empirically standardized sample points are

$$\widehat{V}_i \stackrel{\text{def}}{=} \hat{v}(X_i) = \big(\hat{v}_1(X_{i1}), \ldots, \hat{v}_d(X_{id})\big), \qquad i = 1, \ldots, n \tag{3.11}$$

where $\hat{v}(x) = (\hat{v}_1(x_1), \ldots, \hat{v}_d(x_d))$ for all $x \in \mathbb{R}^d$ and

$$\hat{v}_j(x_j) \stackrel{\text{def}}{=} \frac{1}{1 - \frac{n}{n+1}\widehat{F}_j(x_j)}, \tag{3.12}$$

for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. The factor $\frac{n}{n+1}$ in (3.12) serves to avoid division by zero in case $x_j \ge \max(X_{1j}, \ldots, X_{nj})$.

We have $\hat{v}_j(X_{ij}) = 1/(1 - R_{ij}/(n+1))$, where $R_{ij} = n\widehat{F}_j(X_{ij})$ is the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$. Any statistic that is a function $\widehat{V}_1, \ldots, \widehat{V}_n$ depends on the data $X_1, \ldots, X_n$ only through the component-wise ranks. This will be the case for the empirical angular measure.

3.2.4 The empirical angular measure

Let δ_x denote a point mass at x and put

$$\widehat{P}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \delta_{\widehat{V}_i}, \tag{3.13}$$

the empirical distribution of the pseudo-observations $\widehat{V}_1, \ldots, \widehat{V}_n$ in (3.11). The random measure $\widehat{P}_n$ can be legitimately considered as an estimator of the distribution of V.

The definition (3.3) of the exponent measure ν involves a limit as $t \to \infty$. Setting $t = n/k$ for $k \in \{1, \ldots, n\}$ such that both k and n/k are large yields the empirical exponent measure

$$\widehat{\nu}(B) \stackrel{\text{def}}{=} \tfrac{n}{k}\widehat{P}_n(\tfrac{n}{k}B) = \frac{1}{k} \sum_{i=1}^n \mathbb{I}\left\{\widehat{V}_i \in \tfrac{n}{k}B\right\} \tag{3.14}$$

for Borel sets $B \subseteq \mathbb{E}$. Consider a Borel set $A \subseteq \mathbb{S}_p$ and recall the angular measure Φ_p and the cone C_A in (3.7). In view of (3.14), *the empirical angular measure* [42] is defined as

$$\widehat{\Phi}_p(A) = \widehat{\nu}(C_A) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\left\{\widehat{V}_i \in (n/k)C_A\right\}$$
$$= \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\left\{\|\widehat{V}_i\|_p \geq n/k,\ \theta_p(\widehat{V}_i) \in A\right\}. \tag{3.15}$$

It is the empirical version of the pre-asymptotic angular measure

$$t\mathrm{P}\left(t^{-1}V \in C_A\right)$$

at level $t = n/k$.

In the bivariate case, the asymptotic distribution of the empirical angular measure has been investigated in case the sequence $k = k(n)$ satisfies $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. The max-norm was considered in [42], while the L_p-norm for $p \in [1, \infty]$ was studied in [47] together with empirical likelihood methods to exploit the moment constraints (3.9).

3.3 Concentration inequalities

Recall the empirical angular measure $\widehat{\Phi}_p$ in (3.15). Our main result provides a concentration inequality for the uniform deviations

$$\sup_{A \in \mathcal{A}} \left|\widehat{\Phi}_p(A) - \Phi_p(A)\right|. \tag{3.16}$$

The supremum is taken over classes $\mathcal{A}$ of Borel sets A of $\mathbb{S}_p$ in (3.6) satisfying appropriate assumptions. To bound the supremum, the empirical measure is rewritten in terms of random sets which are subsequently framed in between deterministic sets. We first explain the idea behind the framing approach (Section 3.3.1) before stating our main theorem on concentration bound for the empirical angular measure (Section 3.3.2). We conclude with examples of collections $\mathcal{A}$ (Section 3.3.3).

3.3.1 Framing random sets

Since the estimators depend on the data only through the ranks, they are invariant under increasing component-wise transformations. So although the marginal distributions $F_1, \ldots, F_d$ are unknown, we can and will nevertheless assume that they are unit-Pareto already. In that case, we have $v(x) = x$ for $x \in [1, \infty]^d$ in (3.10), so that $V = v(X) = X$ and $V_i = v(X_i) = X_i$ for $i = 1, \ldots, n$. In particular, the distribution of $V = X$ is P.

Recall $\widehat{P}_n$ in (3.13) and let $P_n \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n \delta_{V_i}$ denote the empirical distribution of $V_1, \dots, V_n$. Since $\widehat{V}_i = \hat{v}(X_i) = \hat{v}(V_i)$, we have $\widehat{P}_n = P_n \circ \hat{v}^{-1}$, that is, $\widehat{P}_n$ is the push-forward measure of P_n by $\hat{v}$. Up to a scaling factor, the value of the empirical angular measure $\widehat{\Phi}_p$ in a Borel set A of $\mathbb{S}_p$ can thus be expressed as P_n evaluated in the random set

$$\widehat{\Gamma}_A \stackrel{\text{def}}{=} \hat{v}^{-1}(\tfrac{n}{k}C_A) = \left\{x \in \mathbb{E} : \ \|\hat{v}(x)\|_p \geq \tfrac{n}{k},\ \theta_p(\hat{v}(x)) \in A\right\}, \tag{3.17}$$

where, as before, $\theta_p(y) = y/\|y\|_p$ for $y \in \mathbb{E}$. Indeed, we have

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

Let $\mathcal{A}$ be a class of Borel sets of $\mathbb{S}_p$. Following in the footsteps of [42] and [47], we construct for any $A \in \mathcal{A}$ two nested deterministic sets $\Gamma_A^- \subseteq \Gamma_A^+$ framing the cone C_A, i.e., such that $\Gamma_A^- \subseteq C_A \subseteq \Gamma_A^+$, and such that on an event that occurs with high probability, we have

$$\forall A \in \mathcal{A}, \qquad \tfrac{n}{k}\Gamma_A^- \subseteq \widehat{\Gamma}_A \subseteq \tfrac{n}{k}\Gamma_A^+. \tag{3.18}$$

On that event, the signed error can then be bounded from above by

$$\begin{aligned}
\widehat{\Phi}_p(A) - \Phi_p(A) &= \tfrac{n}{k}P_n(\widehat{\Gamma}_A) - \nu(C_A) \\
&\leq \tfrac{n}{k}P_n(\tfrac{n}{k}\Gamma_A^+) - \nu(\Gamma_A^-) \\
&\leq \tfrac{n}{k}\left|P_n(\tfrac{n}{k}\Gamma_A^+) - P(\tfrac{n}{k}\Gamma_A^+)\right| + \left|\tfrac{n}{k}P(\tfrac{n}{k}\Gamma_A^+) - \nu(\Gamma_A^+)\right| + \nu(\Gamma_A^+ \setminus \Gamma_A^-).
\end{aligned}$$

A lower bound can be derived in a similar way, yielding, on the same event,

$$\begin{aligned}
\left|\widehat{\Phi}_p(A) - \Phi_p(A)\right| \leq & \max_{B \in \{\Gamma_A^+, \Gamma_A^-\}} \left|\tfrac{n}{k}P(\tfrac{n}{k}B) - \nu(B)\right| && \text{(bias term)} \\
& + \max_{B \in \{\Gamma_A^+, \Gamma_A^-\}} \tfrac{n}{k}\left|P_n(\tfrac{n}{k}B) - P(\tfrac{n}{k}B)\right| && \text{(stochastic error)} \\
& + \nu(\Gamma_A^+ \setminus \Gamma_A^-) && \text{(framing gap).}
\end{aligned} \tag{3.19}$$

Next we will introduce assumptions to enable the construction of the framing sets Γ_A^- and Γ_A^+ together with a high-probability event on which (3.18) holds. The task is then to control the three terms on the right-hand side of (3.19) uniformly over $A \in \mathcal{A}$. The bias term will be left as such; controlling it is an entirely different subject requiring higher-order multivariate regular variation [12, 55]; however, see Remark 3.3.3. Under appropriate complexity assumptions on the class $\mathcal{A}$ and the collection of framing sets, the stochastic error term can be uniformly bounded by means of the concentration inequality for tail empirical processes established in Theorem 3.A.1. Finally, the framing gap can be controlled by ensuring that the set $\Gamma_A^+ \setminus \Gamma_A^-$ is small.

3.3.2 Concentration bounds for the empirical angular measure

The main result of this paper, Theorem 3.3.1, is stated in this subsection. Let $\mathbb{B} \stackrel{\text{def}}{=} [-1,+1]^d$ denote the closed unit ball in $\mathbb{R}^d$ with respect to the sup-norm $\|\cdot\|_\infty$. For sets $A, B \subseteq \mathbb{R}^d$ write $A + B \stackrel{\text{def}}{=} \{a + b : a \in A, b \in B\}$. For $\varepsilon > 0$ and $A \subseteq \mathbb{R}^d$, we thus have $A + \varepsilon\mathbb{B} = \{x \in \mathbb{R}^d : \exists a \in A, \|x - a\|_\infty \leq \varepsilon\}$.

Assumption 3.3.1 (Subsets of the sphere)**.** *The class $\mathcal{A}$ is a collection of non-empty Borel sets of $\mathbb{S}_p$ with the following properties:*

(i) There exists a countable collection $\mathcal{A}_0 \subseteq \mathcal{A}$ such that for every $A \in \mathcal{A}$ there is a sequence $A_n \in \mathcal{A}_0$ such that $\lim_{n\to\infty} \mathbb{I}\{x \in A_n\} = \mathbb{I}\{x \in A\}$ for every $x \in \mathbb{S}_p$.

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

$$\forall A \in \mathcal{A}, \qquad A \subseteq \{x \in \mathbb{S}_p : \min(x) > \tau\} \stackrel{\text{def}}{=} \mathbb{S}_p^\tau. \tag{3.20}$$

(iii) There is a constant $c > 0$ such that for any $A \in \mathcal{A}$ and $\varepsilon > 0$, there exist Borel subsets $A_+(\varepsilon)$ and $A_-(\varepsilon)$ of $\mathbb{S}_p$ satisfying

$$\Phi_p\big(A_+(\varepsilon) \setminus A_-(\varepsilon)\big) \leq c\,\varepsilon \tag{3.21}$$

together with the inclusions

$$\big(A_-(\varepsilon) + \varepsilon\mathbb{B}\big) \cap \mathbb{S}_p \subseteq A \qquad \text{and} \qquad (A + \varepsilon\mathbb{B}) \cap \mathbb{S}_p \subseteq A_+(\varepsilon). \tag{3.22}$$

Condition (i) amounts to the pointwise measurability of the indicators $\{\mathbb{I}\{\cdot \in A\} : A \in \mathcal{A}\}$ [122, Example 2.3.4] and ensures that for any $A \in \mathcal{A}$ we can find $A_n \in \mathcal{A}_0$ such that for any finite Borel measure ν on $\mathbb{S}_p$ we have $\nu(A) = \lim_{n\to\infty} \nu(A_n)$. The suprema over $\mathcal{A}$ in Eq. (3.16) are therefore equal to those over $\mathcal{A}_0$ and are thus measurable.

Condition (ii) stipulates that the elements of the class $\mathcal{A}$ are bounded away from the $2^d - 1$ faces of the sphere. Though it may be considered as restrictive at first glance, we point out 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 angular measure for the corresponding components of X.

The crucial point in Condition (iii) is inequality (3.21), bounding the measure of the difference between the inner and outer ε-hulls of $A \in \mathcal{A}$. The inequality is satisfied if it holds with Φ_p replaced by the $(d-1)$-dimensional Lebesgue measure on $\mathbb{S}_p$ and Φ_p has a bounded density on $\mathbb{S}_p$ with respect to this measure.

In order to deal with the estimation error stemming from the use of the marginal empirical distribution functions in (3.12), we frame the cones C_A in Eq. (3.7) between slightly smaller and larger sets built from the inner and outer hulls. For $A \in \mathcal{A}$, $\sigma \in \{-,+\}$ and $r, h > 0$, define

$$\Gamma_A^\sigma(r,h) \stackrel{\text{def}}{=} \left\{ x \in [0,\infty)^d : \ \|x\|_p \geq \tfrac{1}{r},\ \theta_p(x) \in A_\sigma(h\|x\|_p) \right\}.$$

For all $A \in \mathcal{A}$ and $h > 0$, we have $\Gamma_A^-(r,h) \subseteq C_A$ for $0 < r \leq 1$ and $C_A \subseteq \Gamma_A^+(r,h)$ for $r \geq 1$. The upper confidence bound at level $1-\delta$ for the maximal deviation (3.1) stated in Theorem 3.3.1 below is derived from the decomposition (3.19) with framing sets $\Gamma_A^+(r_+,h)$ and $\Gamma_A^-(r_-,h)$ for specific choices of r_+, r_- and h.

To control the stochastic error, we need a handle on the complexity of the collection of framing sets. The Vapnik–Chervonenkis (VC) dimension of a collection $\mathcal{F}$ of subsets of some set $\mathcal{X}$ is the supremum (possibly infinite) of the set of positive integers n with the property that there exists a subset $\{x_1,\dots,x_n\}$ of $\mathcal{X}$ with cardinality n such that all 2^n subsets of $\{x_1,\dots,x_n\}$ can be written in the form $\{x_1,\dots,x_n\} \cap F$ for some $F \in \mathcal{F}$. The VC-dimension is a central quantity in statistical learning and empirical process theory as it lies at the basis of many concentration inequalities for empirical distributions, see for instance [122] and [20].

Assumption 3.3.2. *For any $r, h > 0$, the collections $\{\Gamma_A^-(r,h) : A \in \mathcal{A}\}$ and $\{\Gamma_A^+(r,h) : A \in \mathcal{A}\}$ of subsets of $\mathbb{E}$ have finite VC-dimension.*

In the framing sets $\Gamma_A^\sigma(r,h)$, the tolerance ε for the angle $\theta_p(x)$ in the inner and outer hulls of a set A depends on the norm $\|x\|_p$. Therefore, in Assumption 3.3.2 it is not sufficient to assume that the collection $\mathcal{A}$ or even the collection of inner and outer hulls has finite VC-dimension. In Section 3.3.3 we provide a realistic example of an angular class $\mathcal{A}$—namely a class defined by linear inequality constraints—where Assumptions 3.3.1 and 3.3.2 are satisfied.

Theorem 3.3.1. *Let $X_1,\dots,X_n$ be an independent random sample from a distribution P on $\mathbb{R}^d$ with continuous margins and let P_V denote the distribution of $V_1 = v(X_1)$ where v is defined in* (3.10)*. Let ν be an exponent measure having angular measure Φ_p with respect to $\|\cdot\|_p$ for some $p \in [1,\infty]$. Let $\mathcal{A}$ be a collection of Borel sets of $\mathbb{S}_p$ such that Assumptions 3.3.1 and 3.3.2 are fulfilled. Let n, k, ρ be such that $\tau n > k > (3 \vee 6c)$ and $k/n < \rho < \tau$, where the constant $c > 0$ comes from point (iii) of Assumption 3.3.1. Let $\delta \in (0,1)$. Put*

$$\Delta_1 = \Delta_1(k,\delta,\rho) \stackrel{\text{def}}{=} \frac{1}{\sqrt{k\rho}} \left(60 + 2\sqrt{\log((d+1)/\delta)} \right) + \frac{2}{3k} \log((d+1)/\delta).$$

For $\sigma \in \{-,+\}$ and $A \in \mathcal{A}$, abbreviate $\Gamma_A^\sigma = \Gamma_A^\sigma(r_\sigma, 3\Delta_1)$ where $r_\pm = 1 \pm \Delta_2$ with $\Delta_2 = 4(\Delta_1 + 1/k)$. Then, with probability at least $1-\delta$, the empirical angular

measure satisfies

$$\sup_{A\in\mathcal{A}}\left|\widehat{\Phi}_p(A)-\Phi_p(A)\right| \le \sup_{A\in\mathcal{A},\,\sigma\in\{+,-\}}\left|\tfrac{n}{k}P_V(\tfrac{n}{k}\Gamma_A^\sigma)-\nu(\Gamma_A^\sigma)\right| + \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}}+2\sqrt{\log((d+1)/\delta)}\right)+\frac{2\log\left(\frac{d+1}{\delta}\right)}{3k} + 2d\Delta_2+3c\big(\log(d/3c)-\log(\Delta_1)+1\big)\Delta_1, \tag{3.23}$$

provided k is sufficiently large so that $(2/k)\le\Delta_1<(1/4-1/k)\wedge(1/(3c))$ and $\rho/(1-\Delta_1\rho)\le\tau$ and where $V_{\mathcal{F}}$ is the VC-dimension of the collection

$$\left\{\Gamma_A^-:A\in\mathcal{A}\right\}\cup\left\{\Gamma_A^+:A\in\mathcal{A}\right\}. \tag{3.24}$$

The proof is given in Appendix 3.C in the Supplement.

Note that $r_-\le 1\le r_+$, so that $\Gamma_A^-(r_-,h)\subseteq C_A\subseteq\Gamma_A^+(r_+,h)$ for all $h>0$. In the decomposition (3.19), the framing gap is bounded by $2d\Delta_2+3c\big(\log(d/3c)-\log(\Delta_1)+1\big)\Delta_1$ while the estimation error is bounded by

$$\sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}}+2\sqrt{\log((d+1)/\delta)}\right)+\frac{2}{3k}\log((d+1)/\delta)$$

with probability larger than $1-\delta$. This term is the one appearing in the concentration inequality for empirical processes over collections of sets of extreme values proved in Theorem 3.A.1 applied to the collection $\{(n/k)\Gamma_A^\sigma(r_\sigma,3\Delta_1):\sigma\in\{+,-\},\ A\in\mathcal{A}\}$. By Assumption 3.3.2, the collection in (3.24) has a finite VC-dimension: For two collections $\mathcal{F}_1$ and $\mathcal{F}_2$ of subsets of a set $\mathcal{X}$ with finite VC-dimensions d_1 and d_2, respectively, the VC dimension of $\mathcal{F}_1\cup\mathcal{F}_2$ is bounded by d_1+d_2+1 [93, Exercise 3.24].

Note that the bound is a decreasing function of the norm index p. It is related to the shape of the associated sphere $\mathbb{S}_p$, which is easier to work with if more aligned with the axes. The faces of the unit sphere induced by the supremum norm are parallel to the coordinate axes, a property that links up well with the use of component-wise ranks, and has the advantage that its value is not affected by the precise values of the smaller coordinates of x. Extensions to other p-norms can be performed through the use of the equivalences of norms

$$\forall x\in\mathbb{R}^d,\qquad \|x\|_\infty\le\|x\|_p\le d^{1/p}\|x\|_\infty. \tag{3.25}$$

For fixed ρ and δ, when ignoring the bias term, the bound on the right-hand side of (3.23) is of the order

$$\Delta_1\log(1/\Delta_1)=\mathrm{O}\left(\frac{\log k}{\sqrt{k}}\right),\qquad \text{as } k\to\infty. \tag{3.26}$$

Even though Δ_1 and Δ_2 are of the optimal order $\mathrm{O}(1/\sqrt{k})$, the factor $\log k$ shows up: its presence is caused by the framing gap, the third line in both (3.19) and (3.23). Asymptotic theory for the empirical angular measure in the bivariate case [42, 47] suggests a learning rate of the order $1/\sqrt{k}$; whether our logarithmic term is an artifact of our analysis or rather a genuine property of the estimator remains an open problem.

Remark 3.3.1 (Other concentration inequalities). The constant 60 appearing in the error stems comes from the use of chaining techniques as in [87, Theorems 1.16–17]. Relying instead on concentration inequalities for rare events derived recently in [83], it is possible to reduce the constant at the price of an additional logarithmic factor $\sqrt{\log k}$, which, for realistic values of k, may well be smaller than the constant involved in our bound. However, the bound would become more complicated and less accurate from an asymptotic point of view.

The term Δ_1 in Theorem 3.3.1 comes from the application of Theorem 3.A.1 to the tails of the empirical margins $\widehat{F}_j$. As observed by an anonymous Referee, other concentration inequalities exist for such one-dimensional tail empirical processes, see for instance Inequality 1 on page 446 in [116]. For finite samples, some of these inequalities may be sharper than the one we used; nevertheless, the convergence rate as a function of k would not improve since chaining is already optimal in that respect.

Remark 3.3.2 (Bias term and penultimate angular measure). As Theorem 3.3.1 is not concerned with asymptotics, we did not actually have to assume that Φ_p is the angular measure associated to P_V. The link between P_V and Φ_p is quantified instead by the bias term $\sup_{A,\sigma} \left| \frac{n}{k} P_V(\frac{n}{k}\Gamma_A^\sigma) - \nu(\Gamma_A^\sigma) \right|$. Even if P_V has an angular measure of its own, it may be different from the one appearing in the theorem. This flexibility allows for viewing the empirical angular measure as an estimator of a penultimate angular measure, that is, the one for which the induced bias term is minimal.

Remark 3.3.3 (Controlling the bias term). For absolutely continuous models, a primitive condition on the probability density function allows to control the bias term in Theorem 3.3.1. Let P_U denote the distribution of $U = (1 - F_1(X_1), \ldots, 1 - F_d(X_d)) = \iota(V)$ on $[0,1]^d$ where $\iota : (0,\infty)^d \to (0,\infty)^d$ is defined by $\iota(x) \stackrel{\text{def}}{=} (x_1^{-1}, \ldots, x_d^{-1})$. Assume that P_U is absolutely continuous with density p_U and that the measure $\Lambda = \nu \circ \iota^{-1}$ (the push-forward of ν by ι) is absolutely continuous with Lebesgue density λ on $(0,\infty)^d$. Then

$$\sup_{A\in\mathcal{A},\sigma\in\{+,-\}} \left| \tfrac{n}{k} P_V(\tfrac{n}{k}\Gamma_A^\sigma) - \nu(\Gamma_A^\sigma) \right| \leq \int_{(0,\infty)^d} 1\left\{ \min(y) \leq d^{1/p} r_+ \right\} \left| \left(\tfrac{k}{n}\right)^{d-1} p_U(\tfrac{k}{n} y) - \lambda(y) \right| \mathrm{d}y. \quad (3.27)$$

By way of example, consider the multivariate Cauchy density restricted to $(0,\infty)^d$ with probability density function

$$f(x) \stackrel{\text{def}}{=} 2^d \Gamma(\tfrac{d+1}{2}) \pi^{-(d+1)/2} \left(1 + x_1^2 + \cdots + x_d^2\right)^{-(d+1)/2}$$

for $x \in (0,\infty)^d$, with limit density

$$\lambda(x) = 2^{d-1} \Gamma(\tfrac{d+1}{2}) \pi^{-(d-1)/2} x_1^{-2} \cdots x_d^{-2} \left(x_1^{-2} + \cdots + x_d^{-2}\right)^{-(1+d)/2}.$$

In this case, the bound (3.27) is of the order $\mathrm{O}(k/n)$ as $k = k_n \to \infty$ in such a way that $k/n \to 0$. Detailed calculations are given in Appendix 3.D in the Supplement.

3.3.3 Examples of collections of subsets of the sphere

This sections aims at providing examples of classes $\mathcal{A}$ related to a wide range of statistical machine learning algorithms (such as logistic regression, classification and regression trees or linear discriminant analysis) for which Assumptions 3.3.1 and 3.3.2 are satisfied. Proofs are deferred to Appendix E in the Supplement.

Recall $\mathbb{S}_p^\tau$ in (3.20). The scalar product and the Euclidean norm on $\mathbb{R}^d$ are denoted by $\langle x, y\rangle$ and $\|x\|_2 = \sqrt{\langle x, x\rangle}$, respectively.

Example 3.3.1 (Linear restrictions). Fix $\tau \in (0,1)$ and consider the collection

$$\mathcal{A} = \left\{ A_{a,\beta,\tau} : \ (a,\beta) \in \mathbb{R}^d \times \mathbb{R}, \ \|a\|_2 = 1 \right\}$$

of Borel subsets of $\mathbb{S}_p^\tau$ defined via

$$A_{a,\beta} \stackrel{\text{def}}{=} \{x \in \mathbb{S}_p : \ \langle a, x\rangle \le \beta\}, \qquad A_{a,\beta,\tau} \stackrel{\text{def}}{=} A_{a,\beta} \cap \mathbb{S}_p^\tau.$$

Then $\mathcal{A}$ satisfies Assumption 3.3.2, and, if Φ_p has a bounded Lebesgue density on $\mathbb{S}_p$, also Assumption 3.3.1.

Example 3.3.2 (Stability under intersections and unions). Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two collections of Borel subsets of $\mathbb{S}_p$ that satisfy Assumptions 3.3.1 and 3.3.2. Then the same is true for the collection of intersections

$$\mathcal{A}_1 \sqcap \mathcal{A}_2 \stackrel{\text{def}}{=} \{A_1 \cap A_2 : \ A_1 \in \mathcal{A}_1, \ A_2 \in \mathcal{A}_2\}$$

and for the collection of unions

$$\mathcal{A}_1 \sqcup \mathcal{A}_2 \stackrel{\text{def}}{=} \{A_1 \cup A_2 : \ A_1 \in \mathcal{A}_1, \ A_2 \in \mathcal{A}_2\}.$$

In combination with Example 3.3.1, this property covers classifiers built from decision trees with a given depth. The leaves of such a tree correspond to unions of rectangles, the maximum number of rectangles in the union being determined by the tree depth.

3.4 Applications to statistical learning

We illustrate how a concentration inequality such as the one in Theorem 3.3.1 is useful to establish sound non-asymptotic guarantees for the validity of certain statistical learning procedures relying on the empirical angular measure and recently introduced in the literature. In Section 3.4.1, after introducing some minimal background about anomaly detection in extreme regions via minimum-volume set estimation, we show how our main results bring immediate guarantees on this matter. In Section 3.4.2, we recall the whys and wherefores of classification in extreme regions and leverage the techniques developed in Section 3.3 to control the excess risk of a specific empirical risk minimizer targeting the tail region of the covariate space.

3.4.1 Minimum-volume set estimation

To illustrate the usefulness of bounds on the supremum in (3.1), consider the problem of estimating a *minimum-volume set* [46], such sets extending the notion of univariate quantiles. A minimum-volume set at level α is a subset of the sample space of minimum (Lebesgue) volume, constrained to contain a probability mass of at least α. There are fruitful connections between minimum-volume sets and semi-supervised anomaly detection, where data are available from the majority class only and the goal is to construct a decision function delimitating the extent of the normal region. In a Neyman–Pearson framework, an optimal anomaly detection procedure at a certain level $0 < 1 - \alpha \ll 1$ would declare abnormal any new point such that no minimum-volume set of level α contains it [16].

In the context of anomaly detection, the tail of the random vector under scrutiny is of particular interest because many anomalies correspond to unusually large values of at least one component. However, it may not be appropriate to declare as abnormal all such points and a finer analysis of the tails can improve the overall performance of an anomaly detection algorithm. For instance, consider a complex infrastructure monitored by several physical variables. Raising an alert at each extreme value of one of its physical variables can lead to high false alarm rates. A way to reduce this false alarm rate is to study the multivariate distribution of the set of observations such that at least one of their variables is large. This framework can be useful in a wide variety of applications (e.g., fraud detection, safety in aeronautics), where the control of the false alarm rate is crucial (given the cost of safety inspections), thus making the detection of anomalies among extremes undeniably relevant.

In this section, we follow in the footsteps of [120] who consider the problem of constructing sets of relatively high probability in regions of the kind $\{x \in \mathbb{R}^d : \|x\| > t\}$ for large values of t, under regular variation assumptions. Their

statistical analysis is limited to the ideal case where the marginal distributions are known. We extend their guarantees in order to encompass the influence of the rank transformation.

In the context of the angular measure for multivariate upper extremes, the question is to find a Borel set Ω of the unit sphere $\mathbb{S}_p$ in $[0,\infty)^d$ with minimal volume $\lambda(\Omega)$—with λ some reference measure such as the $(d-1)$-dimensional Hausdorff measure—but still having content $\Phi_p(\Omega)$ not smaller than some pre-specified lower limit $\alpha \in (0, \Phi_p(\mathbb{S}_p))$. In [120], such a set is used for the purpose of (unsupervised) anomaly detection in extreme regions. As Ω is supposed to contain a large fraction $\Phi_p(\Omega)/\Phi_p(\mathbb{S}_p)$ of the possible directions of extreme points, a new such point is deemed to be a potential anomaly if it lies in a direction outside Ω. The fact that Ω has minimal volume $\lambda(\Omega)$ means that the critical region $\mathbb{S}_p \setminus \Omega$ to detect suspicious points is as large as possible.

As Φ_p is unknown, Ω needs to be learned from a training sample. Although Ω may be characterized as a certain super-level set of the density of Φ_p with respect to the reference measure λ, a more practical approach than estimating this density, especially in high dimensions, is to limit the search to an algorithmically manageable collection $\mathcal{A}$ of Borel subsets of $\mathbb{S}_p$. Let $\widehat{\Phi}_p$ be any estimator of Φ_p, not necessarily the empirical angular measure. Following the logic in [115], let $\widehat{A}$ solve the empirical angular minimum-volume set problem

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

where $\psi \in (0, \alpha)$ is a tolerance parameter. The price to pay for having to estimate the angular measure is that the minimal required content α has been relaxed to $\alpha - \psi$.

Bounds on the largest estimation error of $\widehat{\Phi}_p$ over $\mathcal{A}$ are helpful to provide probabilistic guarantees for $\widehat{A}$. Suppose that, on some event $\mathcal{E}$, we have

$$\sup_{A \in \mathcal{A}} \left|\widehat{\Phi}_p(A) - \Phi_p(A)\right| \leq \psi. \tag{3.28}$$

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

$$\Phi_p(\widehat{A}) \geq \widehat{\Phi}_p(\widehat{A}) - \psi \geq \alpha - 2\psi \tag{3.29}$$

as well as

$$\lambda(\widehat{A}) \leq \inf\left\{\lambda(A) : A \in \mathcal{A},\ \Phi_p(A) \geq \alpha\right\}. \tag{3.30}$$

Indeed, on $\mathcal{E}$, the collection $\{A \in \mathcal{A} : \Phi_p(A) \geq \alpha\}$ is contained in $\{A \in \mathcal{A} : \widehat{\Phi}_p(A) \geq \alpha - \psi\}$ so that the infimum of $\lambda(A)$ for A in the latter collection must be the smaller one. By (3.29), the empirical solution $\widehat{A}$ is guaranteed to have at least content $\alpha - 2\psi$ under Φ_p, while according to (3.30), the volume of $\widehat{A}$ is smaller than the one of the actual minimum-volume set under Φ_p.

Concentration inequalities for the supremum of $|\widehat{\Phi}_p(A) - \Phi_p(A)|$ over $A \in \mathcal{A}$ provide the existence, for a given $\delta > 0$, of a scalar $\psi \equiv \psi(\delta)$ such that (3.28) holds on an event with probability at least $1 - \delta$. It follows that, with high probability, the empirical minimum-volume set $\widehat{A}$ satisfies (3.29) and (3.30). Provided ψ and δ are both small, the combination of both properties justifies the use of $\widehat{A}$ as an approximation to the true but unknown minimum-volume set under Φ_p. For the empirical angular measure, a valid choice for the tolerance parameter $\psi(\delta)$ is given by the upper bound in Theorem 3.3.1.

3.4.2 Classification in extreme regions

We apply Theorem 3.3.1 to binary classification in extreme regions. Classification is arguably one of the most studied problems in the statistical learning literature. Most existing guarantees are formulated in terms of a risk which is an integrated version of the loss function over the whole covariate space. However, the local performance of a global classifier is not necessarily guaranteed in low probability regions of the covariate space, typically in regions corresponding to an exceedance by one component of a high quantile—where few training points are available—or outside the convex envelope of the training set. In other words, the risk of an error conditional to the norm of the input being large is not adequately controlled in the classical setting. Nevertheless, in a wide variety of applications, ranging from finance/insurance to environmental sciences through teletraffic data analysis for instance, extreme observations of the covariates are of crucial importance.

In this section, we adopt the framework originally proposed in [76], who formalize this argument and propose a risk minimization strategy aiming at improving performance of classification algorithms on such regions. In [76], the marginal distributions of the predictor variables were assumed to be known, so that a standardized vector with exact unit-Pareto margins is observable. Here, we rather assume that the margins are unknown and employ a rank-based standardization instead.

First we recall the set-up of [76]. Consider a random pair (V, Y) where Y in $\{-1, 1\}$ is the label to be predicted and V in $[0, \infty)^d$ is the vector of predictors (features). The goal is to learn a classifier $g : [0, \infty)^d \to \{-1, 1\}$ such that the classification risk for feature vectors V far away from the origin is small. Let $\varrho = \mathrm{P}(Y = 1)$ and assume $0 < \varrho < 1$.

The starting point in [76] is to assume a conditional version of the regular variation condition (3.3): there exist non-zero Borel measures ν_+ and ν_- on $\mathbb{E}$ that are finite on Borel sets bounded away from the origin and such that

$$\lim_{t \to \infty} t \, \mathrm{P}\left(t^{-1} V \in B \mid Y = \sigma 1\right) = \nu_\sigma(B) \tag{3.31}$$

for $\sigma \in \{-, +\}$ and Borel sets $B \subseteq \mathbb{E}$ bounded away from the origin satisfying

$\nu_\sigma(\partial B) = 0$. The angular measure associated to the L_p-norm of ν_σ is

$$\Phi_p^\sigma(A) = \nu_\sigma(C_A), \qquad \text{for } \sigma \in \{-,+\} \text{ and Borel sets } A \subseteq \mathbb{S}_p,$$

with C_A as in (3.7). By (3.31), the unconditional distribution of V is regularly varying as in (3.3) with limit measure $\nu = \varrho\nu_+ + (1-\varrho)\nu_-$ and angular measure $\Phi_p = \varrho\Phi_p^+ + (1-\varrho)\Phi_p^-$.

In [76] it was assumed that an independent random sample $\{(V_i, Y_i)\}_{i=1}^n$ from the distribution of (V, Y) is given. Here, instead, the set-up is that we observe an independent random sample $\{(X_i, Y_i)\}_{i=1}^n$ from the distribution of (X, Y) and that (3.31) holds with $V = v(X)$, where v is defined via the probability integral transform in (3.10). This means that the classifier will have to be learned from the pairs $(\widehat{V}_i, Y_i)$ where $\widehat{V}_i = \hat{v}(X_i)$ in (3.11) is based on the rank transform.

We emphasize that the marginal distribution functions F_j in the definition of v are not conditioned upon Y. Indeed, for a new observation, marginal standardization will have to be carried out without the knowledge of the label to be predicted.

In [76] the focus is on the conditional classification risk above level $t > 0$ of a classifier $g : [0,\infty)^d \setminus \{0\} \to \{-1, 1\}$ defined on the standardized input V

$$L_t^{\mathrm{cond}}(g) \stackrel{\mathrm{def}}{=} \mathsf{P}\left(Y \neq g(V) \mid \|V\|_p > t\right), \tag{3.32}$$

Let $\mathcal{G}$ be a pre-defined family of classifiers g and let

$$L_t^{\mathrm{cond}^*} \stackrel{\mathrm{def}}{=} \inf_{g\in\mathcal{G}} L_t^{\mathrm{cond}}(g)$$

be the smallest conditional classification risk for classifiers in $\mathcal{G}$. The purpose is to learn from the training sample a classifier $\hat{g} \in \mathcal{G}$ such that for large t the excess risk $L_t^{\mathrm{cond}}(\hat{g}) - L_t^{\mathrm{cond}*}$ is small.

In [76] it is argued that, asymptotically, attention can be restricted to *angular classifiers* g, that is, for which $g(x) = g(\theta_p(x))$ with $\theta_p(x) = x/\|x\|_p$ for $x \in \mathbb{E}$. Their analysis involves a random pair (V_∞, Y_∞) whose distribution is the weak limit as $t \to \infty$ of $(t^{-1}V, Y)$ conditionally on $\|V\|_p > t$, a limit which exists thanks to (3.31).

Let

$$\eta(x) \stackrel{\mathrm{def}}{=} \mathsf{P}(Y = 1 \mid V = x)$$

denote the regression function of (V, Y) and let

$$\eta_\infty(x) = \mathsf{P}\left(Y_\infty = 1 \mid V_\infty = x\right)$$

be the one of (V_∞, Y_∞). The respective Bayes classifiers are

$$\begin{aligned} g^*(x) &\stackrel{\mathrm{def}}{=} \mathbb{I}\{\eta(x) \geq 1/2\}, \\ g_\infty^*(x) &\stackrel{\mathrm{def}}{=} \mathbb{I}\{\eta_\infty(x) \geq 1/2\}. \end{aligned} \tag{3.33}$$

Note that g^* minimizes the conditional risk L_t^{cond} for any $t > 0$. Assume that when $\|x\|_p$ is large, $\eta(x)$ and $\eta_\infty(x)$ are uniformly close, i.e.,

$$\sup_{x \in [0,\infty)^d : \|x\|_p \geq t} |\eta(x) - \eta_\infty(x)| \to 0, \qquad \text{as } t \to \infty, \tag{3.34}$$

then by Theorem 1 in [76],

(i) the asymptotic Bayes classifier g_∞^* is angular;

(ii) the conditional excess risks of g^* and g_∞^* are asymptotically the same

$$L_t^{\mathrm{cond}}(g_\infty^*) - L_t^{\mathrm{cond}}(g^*) \to 0, \qquad \text{as } t \to \infty.$$

These properties motivate restricting the search to a class $\mathcal{G}$ of candidate classifiers g depending on the angle only.

Theorem 2 in [76] provides a concentration inequality for the excess risk of the empirical risk minimizer $\hat{g}_k \in \mathcal{G}$ learned from a sample $\{(V_i, Y_i)\}_{i=1}^n$, using only those points for which $\|V_i\|_p$ belongs to a top fraction among those observed. Here, we intend to do the same but for the rank-based transformed feature vectors $\widehat{V}_i = \hat{v}(X_i)$ in (3.11).

Classification risk and angular measure. For g in a class of angular classifiers $\mathcal{G}$, recall the conditional classification risk $L_t^{\mathrm{cond}}(g)$ above level t in (3.32) and define its unconditional version by

$$L_t(g) \stackrel{\mathrm{def}}{=} t\mathrm{P}\left(\|V\|_p \geq t\right) L_t^{\mathrm{cond}}(g) = t\mathrm{P}\left(g(V) \neq Y,\ \|V\|_p \geq t\right). \tag{3.35}$$

The multiplicative factor $t\mathrm{P}(\|V\|_p \geq t)$ converges to $\Phi_p(\mathbb{S}_p)$ and does not change the minimizer in the class $\mathcal{G}$. Working with the unconditional version $L_t(g)$ rather than with $L_t^{\mathrm{cond}}(g)$ simplifies the analysis that follows.

In view of Assumption 3.3.1 required in Section 3.3, we exclude from our empirical risk minimization (ERM) strategy those feature vectors whose angle (after standardization) is too close to the boundary of the unit sphere. Let $\tau \in (0,1)$ and recall $\mathbb{S}_p^\tau$ in (3.20). We have

$$L_t(g) = L_t^{>\tau}(g) + L_t^{\leq\tau}(g) \tag{3.36}$$

with

$$L_t^{>\tau}(g) \stackrel{\mathrm{def}}{=} t\mathrm{P}\left(g(V) \neq Y,\ \theta_p(V) \in \mathbb{S}_p^\tau,\ \|V\|_p \geq t\right),$$
$$L_t^{\leq\tau}(g) \stackrel{\mathrm{def}}{=} t\mathrm{P}\left(g(V) \neq Y,\ \theta_p(V) \notin \mathbb{S}_p^\tau,\ \|V\|_p \geq t\right).$$

The regions of the sphere $\mathbb{S}_p$ labeled $+1$ and -1 by $g \in \mathcal{G}$ are denoted by

$$\mathbb{S}_p^\sigma(g) \stackrel{\mathrm{def}}{=} \{x \in \mathbb{S}_p : g(x) = \sigma 1\}, \qquad \text{for } \sigma \in \{-,+\}.$$

We work hereafter under the following smoothness assumption.

Assumption 3.4.1 (Smoothness). *The scalar $\tau \in (0,1)$ is such that $\Phi_p(\partial \mathbb{S}_p^\tau) = 0$ and the class $\mathcal{G}$ is such that $\Phi_p(\partial \mathbb{S}_p^+(g)) = \Phi_p(\partial \mathbb{S}_p^-(g)) = 0$ for all $g \in \mathcal{G}$.*

Lemma 3.4.1. *If the conditional regular variation property* (3.31) *and Assumption 3.4.1 hold, then for any angular classifier $g \in \mathcal{G}$,*

$$\lim_{t\to\infty} L_t^{>\tau}(g) = L_\infty^{>\tau}(g) \stackrel{\text{def}}{=} \varrho \Phi_p^+(\mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau) + (1-\varrho)\Phi_p^-(\mathbb{S}_p^+(g) \cap \mathbb{S}_p^\tau),$$
$$\lim_{t\to\infty} L_t^{\leq\tau}(g) = L_\infty^{\leq\tau}(g) \stackrel{\text{def}}{=} \varrho \Phi_p^+(\mathbb{S}_p^-(g) \setminus \mathbb{S}_p^\tau) + (1-\varrho)\Phi_p^-(\mathbb{S}_p^+(g) \setminus \mathbb{S}_p^\tau),$$

and thus

$$\lim_{t\to\infty} L_t(g) = L_\infty(g) \stackrel{\text{def}}{=} \varrho \Phi_p^+(\mathbb{S}_p^-(g)) + (1-\varrho)\Phi_p^-(\mathbb{S}_p^+(g)).$$

The proof is deferred to appendix 3.F. The idea of the decomposition (3.36) is to discard points with angle outside $\mathbb{S}_p^\tau$. If Φ_p is concentrated on the interior of $\mathbb{S}_p$, the corresponding loss term $L_t^{\leq\tau}(g)$ can be expected to be small for τ close zero, since

$$\sup_{g\in\mathcal{G}} L_t^{\leq\tau}(g) \leq t\mathrm{P}\left(\theta_p(V) \notin \mathbb{S}_p^\tau,\ \|V\|_p > t\right) \to \Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau), \qquad \text{as } t \to \infty.$$

ERM classifier and decomposition of the excess risk. Given $0 < \tau < 1$ and integers $1 < k \leq n$, define the empirical risk of a classifier $g \in \mathcal{G}$ by

$$\widehat{L}^\tau(g) \stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=1}^n \mathbb{I}\left\{g(\widehat{V}_i) \neq Y_i,\ \theta_p(\widehat{V}_i) \in \mathbb{S}_p^\tau,\ \|\widehat{V}_i\|_p \geq n/k\right\}. \tag{3.37}$$

Assuming a minimizer exists, the ERM classifier is defined as

$$\hat{g}_k^\tau \in \operatorname*{argmin}_{g\in\mathcal{G}} \widehat{L}^\tau(g).$$

Otherwise, introduce a tolerance parameter and consider an approximate minimizer instead, i.e., an argument where the value of the objective function is close to the infimum.

Recall L_∞ in Lemma 3.4.1. A consequence of Theorem 1 in [76] is that if (3.34) holds, the Bayes classifier g_∞^* in (3.33) minimizes L_∞ over all measurable classifiers. One way to measure the performance of the ERM classifier $\hat{g}_k^\tau$ is via the asymptotic excess risk

$$L_\infty(\hat{g}_k^\tau) - \inf_{g\in\mathcal{G}} L_\infty(g).$$

The latter can be bounded in terms of the supremum deviation of the empirical and asymptotic risks over $\mathcal{G}$. Since $\widehat{L}^\tau(\hat{g}_k^\tau)$ is equal to the infimum of $\widehat{L}^\tau(g)$ over $g \in \mathcal{G}$, we have

$$L_\infty(\hat{g}_k^\tau) - \inf_{g\in\mathcal{G}} L_\infty(g) \le 2 \sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty(g)\right|. \tag{3.38}$$

Our main purpose is therefore to obtain a concentration inequality for the supremum on the right-hand side of this inequality.

In our context, the supremum deviation itself decomposes further, since for all $g \in \mathcal{G}$, we have $L_\infty^{\le\tau}(g) \le \Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau)$ and thus

$$\sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty(g)\right| \le \sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)\right| + \Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau). \tag{3.39}$$

The term $\Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau)$ may be viewed as an additional bias term which vanishes as $\tau \to 0$ provided Φ_p is concentrated on the interior of $\mathbb{S}_p$. On the other hand, the upper bounds in Theorem 3.3.1 and Theorem 3.4.1 grow roughly as $1/\sqrt{\tau}$ as $\tau \to 0$. The choice of τ thus constitutes an additional bias-variance compromise.

Lemma 3.F.1 in Appendix 3.F parallels Lemma 3.4.1 by relating the empirical risk $\widehat{L}^\tau(g)$ to the empirical angular measures of the positive and negative instances, i.e., for $A \subseteq \mathbb{S}_p$ and $\sigma \in \{-,+\}$,

$$\widehat{\Phi}_p^\sigma(A) \overset{\text{def}}{=} \frac{1}{k_\sigma} \sum_{i=1}^n \mathbb{I}\{Y_i = \sigma 1\} \cdot \mathbb{I}\left\{\theta_p(\widehat{V}_i) \in A,\ \|\widehat{V}_i\|_p \ge n/k\right\}, \tag{3.40}$$

where $k_\sigma \overset{\text{def}}{=} k n_\sigma / n$ and $n_\sigma \overset{\text{def}}{=} \sum_{i=1}^n \mathbb{I}\{Y_i = \sigma 1\}$ is the number of points such that $Y_i = \sigma 1$.

In view of the error decomposition (3.38)–(3.39), we state our main result in terms of the maximum deviation $\sup_{g\in\mathcal{G}} |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)|$, following the techniques from Section 3.3.

Theorem 3.4.1 (Deviations of the empirical tail risk)**.** *Let $\mathcal{G}$ be a class of angular classifiers. Consider the collection*

$$\mathcal{A} \overset{\text{def}}{=} \{\mathbb{S}_p^+(g) \cap \mathbb{S}_p^\tau : g \in \mathcal{G}\} \cup \{\mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau : g \in \mathcal{G}\}$$

If Assumptions 3.3.1 and 3.3.2 relative to the class $\mathcal{A}$ and the unconditional angular measure Φ_p are satisfied and if the conditional regular variation assumption (3.31) *and Assumption 3.4.1 hold, then, with probability at least $1-\delta$,*

$$\sup_{g\in\mathcal{G}} |\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)| \le 2(\text{error} + \text{bias} + \text{gap})$$

where error *and* gap *are nearly the same as in Theorem 3.3.1 (δ has been halved in the* error *term), that is,*

$$\text{error} \stackrel{\text{def}}{=} \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}} + 2\sqrt{\log(2(d+1)/\delta)}\right) + \frac{2}{3k}\log(2(d+1)/\delta),$$
$$\text{gap} \stackrel{\text{def}}{=} 2d\Delta_2 + 3c\big(\log(d/3c) - \log(\Delta_1) + 1\big)\Delta_1,$$

with Δ_1, Δ_2 and $V_{\mathcal{F}}$ as defined in Theorem 3.3.1, and

$$\text{bias} \stackrel{\text{def}}{=} \sup\Big\{\Big|\tfrac{n}{k}\mathrm{P}(V \in \tfrac{n}{k}B, Y = \sigma 1) - \mathrm{P}(Y = \sigma 1)\nu_\sigma(B)\Big| : B = \Gamma_A^+ \text{ or } B = \Gamma_A^- \text{ for some } A \in \mathcal{A} \text{ and } \sigma \in \{-,+\}\Big\}.$$

The proof relies on the relationships between the (empirical) classification risks and the (empirical) angular measure pointed out in Lemmata 3.4.1 and 3.F.1, which imply that

$$\begin{aligned}|\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)| \le{}& \left|\frac{n_+}{n}\widehat{\Phi}_p^+(\mathbb{S}_p^-(g)\cap\mathbb{S}_p^\tau) - \varrho\Phi_p^+(\mathbb{S}_p^-(g)\cap\mathbb{S}_p^\tau)\right| \\ &+ \left|\frac{n_-}{n}\widehat{\Phi}_p^-(\mathbb{S}_p^+(g)\cap\mathbb{S}_p^\tau) - (1-\varrho)\Phi_p^-(\mathbb{S}_p^+(g)\cap\mathbb{S}_p^\tau)\right|\end{aligned}$$

for $g \in \mathcal{G}$. The right-hand side of the latter display is then bounded uniformly in $g \in \mathcal{G}$ by adapting the proof of Theorem 3.3.1; see Appendix 3.F in the Supplement for details.

3.5 Simulation experiments

Our experiments aim at illustrating the influence of the threshold τ introduced in Equation (3.20) on the supremum error $\sup_{A\in\mathcal{A}_p}|\widehat{\Phi}_p(A) - \Phi_p(A)|$. For a simple class of sets $\mathcal{A}_p$, we will demonstrate empirically that the supremum error increases as τ decreases. This finding suggests that this additional parameter is not a mere artifact from our proof.

We report the results of Monte Carlo experiments on simulated data. The setting is such that $\Phi_p(A)$ can be approximated with arbitrary precision by Monte Carlo sampling, the standardization v to unit-Pareto margins can be computed analytically, and the bias term in the upper bounds of Theorems 3.3.1 is zero. The estimation error then only stems from the stochastic error and framing gap in (3.19) leading to the terms on the second and third lines in (3.23).

The experiments were implemented in Python 3 using the packages `numpy`, `scipy`, `matplotlib`, and `scikit-learn`. The computer code to reproduce the experiments is publicly available online.[1]

[1] `https://github.com/Hamid-Jalalzai/`

Experimental setting. We consider angular measures with respect to the L_p-norm for $p \in \{1, 2, \infty\}$ and we limit ourselves to dimensions $d \in \{2, \dots, 5\}$, higher dimensions requiring much more computational effort to evaluate the supremum error on the class of sets described below.

The different classes $\mathcal{A}_p$ for different values of p are all obtained by projection of a single class $\mathcal{A}$ defined on the L_∞-sphere. Namely, for a class $\mathcal{A}$ of sets on $\mathbb{S}_\infty^\tau$, we define $\mathcal{A}_p = \{\theta_p(A) \cap \mathbb{S}_p^\tau : A \in \mathcal{A}\}$ for $p \in [1, \infty]$.

We choose the class $\mathcal{A}$ on $\mathbb{S}_\infty^\tau$ as the finite collection of hyper-rectangles forming a regular grid on $\mathbb{S}_\infty^\tau$ with side length $h = (1-\tau)/S$ with $S = 10$. Each set $A \in \mathcal{A}$ is of the form $A = A_1 \times \cdots \times A_d$, where, for some $j_0 \in \{1, \dots, d\}$ and some integer vector $(i_j)_{j \in \{1,\dots,d\} \neq j_0} \in \{0, 1, \dots, S-1\}^{d-1}$, we have

$$A_j = \begin{cases} \{1\} & \text{if } j = j_0, \\ (\tau + i_j h, \tau + (i_j + 1)h) & \text{if } j \neq j_0. \end{cases}$$

We consider an independent random sample $X_1, \dots, X_n$ drawn from the distribution of a random vector $X = R\Theta$ where R and Θ are independent, the random variable R follows a unit-Pareto distribution on $[1, \infty)$ and Θ follows a symmetric Dirichlet distribution on the unit simplex $\mathbb{S}_1 \overset{\text{def}}{=} \{x \in [0,1]^d : x_1 + \cdots + x_d = 1\}$ with parameter $(\nu, \dots, \nu)$ for some concentration parameter $\nu > 0$. The larger ν, the stronger Θ is concentrated around the barycenter $(1/d, \dots, 1/d)$. As detailed in Lemma 3.G.1 in the Supplement, the angular measure Φ_p for $p \in [1, \infty]$ is $\Phi_p(A) = d\mathrm{P}(X \in C_A)$ for Borel sets $A \subseteq \mathbb{S}_p$. If $p = 1$, then this simplifies to $\Phi_1(A) = d\mathrm{P}(\Theta \in A)$.

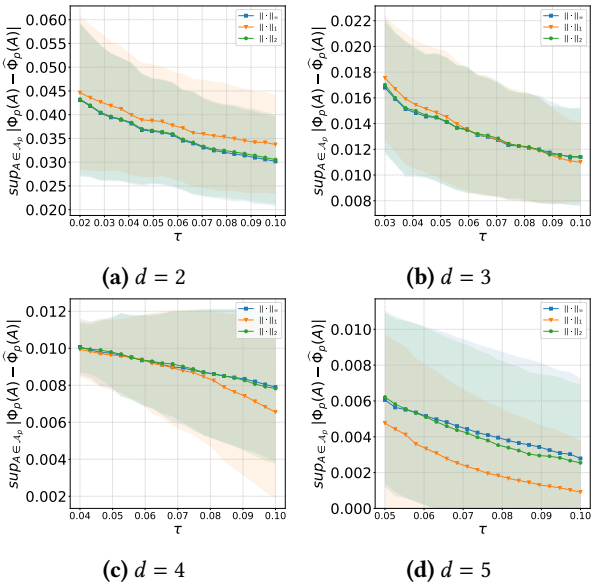
In this setting, all statistics involved in our analysis are easily computable. The following facts are an immediate consequence of Lemma 3.G.1 in the Supplement. Let $A \subseteq \mathbb{S}_p^\tau$ be a Borel set with $\tau \in (0, 1)$ and recall $\mathbb{S}_p^\tau$ in (3.20).

- The true $\Phi_p(A)$ may be approached with arbitrary precision by the Monte Carlo estimator

$$\begin{aligned} \Phi_{p,\mathrm{MC}}(A) &= \frac{d}{N} \sum_{i=1}^{N} \mathbb{I}\left\{X_i' \in C_A\right\} \\ &= \frac{d}{N} \sum_{i=1}^{N} \mathbb{I}\left\{\Theta_i'/\|\Theta_i'\|_p \in A,\ R_i'\|\Theta_i'\|_p \geq 1\right\}, \end{aligned} \tag{3.41}$$

  where $X_i' = (R_i', \Theta_i')$ for $i \in \{1, \dots, N\}$ is another independent random sample from the distribution of (R, Θ). The Monte Carlo estimator is unbiased and has variance bounded by $d^2/(4N)$.

- We have $\Phi_p(A) = \frac{n}{k}\mathrm{P}(V \in \frac{n}{k}C_A)$ as soon as $n/k > d/\tau$, so that the bias term in Theorem 3.3.1 is null under the latter condition.



(a) $d = 2$ **(b)** $d = 3$

(c) $d = 4$ **(d)** $d = 5$

Figure 3.1: Average errors $\sup_{A\in\mathcal{A}} |\widehat{\Phi}_p(A) - \Phi_p(A)|$ **with** $p \in \{1, 2, \infty\}$**, in dimension** $d = 2$ **(top left panel),** $d = 3$ **(top right panel), dimension** $d = 4$ **(bottom left panel) and** $d = 5$ **(bottom right panel), as a function of** τ**. Colored intervals represent standard deviations of errors over** 500 **independent replications.**

Results. We choose a sample size of $n = 10^4$ and we let $k = 100$ in dimension $d \in \{2, \dots, 5\}$. The Monte Carlo parameter N in (3.41) is set to 10^7. The Dirichlet concentration parameter is chosen as $\nu = 1/10$ so that the angular measure is concentrated near the boundaries of the positive orthant.

Figure 3.1 displays the supremum errors $\sup_{A\in\mathcal{A}_p} |\widehat{\Phi}_p(A) - \Phi_p(A)|$ averaged over 500 independent replications as a function of the parameter τ. The latter varies in the range $[dk/n, 0.1]$ in line with Lemma 3.G.1, for the reason explained above. The average errors of all three estimators $\widehat{\Phi}_p$ decrease for larger values of τ in line with our theoretical findings. Note that the errors also decrease when the dimension increases. This is a direct consequence of the definition of the class $\mathcal{A}$, forming a rectangular grid on the sphere $\mathbb{S}_\infty^\tau$ consisting of $d \times 10^{d-1}$ subsets (each face requires 10^{d-1} rectangles to be covered) so that the grid becomes finer as the dimension increases and the Φ_p-mass of the sets A considered in the supremum error is decreasing.

Of particular interest is the behaviour of the different norms. The error associated to the L_2-norm and the L_∞-norm behave similarly while the one linked to the L_1-norm seems quite different. Moreover, the nature of the difference changes with the dimension. An explanation of these phenomena is depicted in Figure 3.2. Since the estimation error of $\widehat{\Phi}_p$ is measured with a supremum and increases while τ decreases, it suggests that this supremum is

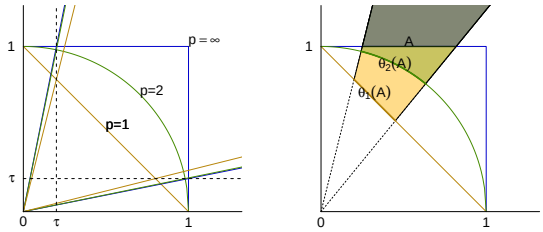


Figure 3.2: Restriction to coordinates larger than τ has an heavier impact on the L_1-sphere than on the L_2-sphere and the L_∞-sphere, which touch each other near the coordinate axes (left panel). The cone $\{x \in [0,\infty)^p : \|x\|_p \geq 1, x \in \theta_p(A)\}$ generated by the projection $\theta_p(A)$ of a set $A \subseteq \mathbb{S}_\infty^\tau$ onto $\mathbb{S}_p^\tau$ is largest when the L_1-norm is considered (right panel).

realized near the boundary of the considered subset $\mathbb{S}_p^\tau$ of the sphere, i.e., near the intersections with the coordinate axes. In this region, the L_2-sphere and the L_∞-sphere are close to each other, the latter being even tangent to the former in dimension $d = 2$. The L_1-sphere is quite different there since it forms a 45° angle with the L_∞-sphere. This explains why the error behaves so differently when the L_1-norm is considered. The fact that it becomes smaller in higher dimension can be understood as follows. Two forces play an opposite role in making the L_1-sphere different from the others: on the one hand, as illustrated in the left panel of Figure 3.2, censuring coordinates smaller than τ has a higher impact on the L_1-sphere since it looses more volume than the others, on the other hand, as illustrated in the right panel, the cones C_A generated by the projection of a set $A \subseteq \mathbb{S}_\infty^\tau$ are larger for $p = 1$ than for $p = 2$ or $p = \infty$. The first force tends to reduce the Φ_1-mass of the L_1-sphere while the second tends to increase it. Following the results of Figure 3.1, the latter is dominated by the former when the dimension increases.

3.6 Conclusion

We derived non-asymptotic guarantees in the form of concentration inequalities for the rank-based empirical angular measure with respect to any L_p-norm, $p \in [1,\infty]$. The bounds are valid in any dimension and concern the supremum error over certain collections of subsets of the L_p-sphere. Apart from a logarithmic term, the bounds match the convergence rate known in the bivariate case [42, 47]. Two applications to statistical learning based on observations in extreme regions were worked out: minimum-volume set estimation and binary classification.

It would be interesting to be able to complement our upper bounds with

lower bounds on the estimation error. Sharp bounds would provide guidance on the choice of the threshold parameter k, ideally in an adaptive manner as in [18].

The results are limited to subsets of the relative interior of the unit sphere to avoid non-extreme components, which are difficult to manage. Allowing for sets touching the boundary constitutes an important but challenging avenue left for further research. A numerical experiment has been provided to illustrate the influence of this restriction on the estimation error, for different norms and dimensions.

The application to classification was limited to two balanced classes. Extensions to multiple classes and unbalanced situations can be developed in the same way. Of an altogether different nature, however, is the prediction of a continuous response from observations in extreme regions. The latter would require concentration inequalities for the empirical angular measure evaluated on collections of functions more general than set indicators. This topic has not yet even been broached in the bivariate case.

3.A Concentration inequality for rare events

The main concentration tool that we use in the proof of Theorem 3.3.1 is the following. Since the result may be of independent interest, we provide a detailed proof.

Theorem 3.A.1. *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 $\mathcal{X}$. Let $\mathcal{A}$ be a VC-class of measurable subsets of $\mathcal{X}$ with VC-dimension $V_{\mathcal{A}}$. Let B be a measurable subset of $\mathcal{X}$ containing $\bigcup_{A\in\mathcal{A}} A$ and write $\kappa = P(B)$. Then, for all $\delta \in (0,1)$, there exists an event with probability at least $1-\delta$ on which we have*

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \le \sqrt{\frac{\kappa}{n}} \left(60\sqrt{V_{\mathcal{A}}} + 2\sqrt{\log(1/\delta)} \right) + \frac{2}{3n} \log(1/\delta).$$

In [60, Theorem 1], a similar inequality is derived, but with a generic constant $C > 0$ that is not made explicit. Just before their Lemma 14, there is a reference to [79] providing bounds on the expectation of a symmetrized supremum, but from that source, the value of the constant looks nearly impossible to trace. We follow an alternative route to obtain that value via Theorems 1.16–17 in [87], 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 in terms of its VC-dimension. In this way, the constant C can be made explicit. Its value is most likely not optimal but, should a sharp value for the constant be found in the future, it can be substituted in our result.

We now give the proof of Theorem 3.A.1. It is based on a McDiarmid's concentration inequality for the supremum around its expected value in [90], extending Bernstein's classical inequality and recalled in [60, Proposition 11]. We rewrite it in a form which is more convenient for us [83, Proposition 5].

Proposition 3.A.1. *Under the assumptions of Theorem 3.A.1, for all $\delta \in (0,1)$, there exists an event with probability at least $1-\delta$ on which we have*

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq 2\sqrt{\frac{\kappa}{n}\log(1/\delta)} + \frac{2}{3n}\log(1/\delta) + \mathsf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right].$$

Proof of Theorem 3.A.1. Apply Proposition 3.A.1. To bound the remaining expectation, we make use of Theorem 1.16 in [87], which relies on a chaining argument to ensure that

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leq \frac{24}{\sqrt{n}} \max_{x_1,\ldots,x_n\in\mathcal{X}} \int_0^1 \sqrt{\log\left(2\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right)\right)}\,\mathrm{d}r, \quad (3.42)$$

where $\mathcal{N}(r, \mathcal{A}(x_1^n))$ is the *covering number* (the smallest number of balls of radius r needed to cover a set) of the set

$$\mathcal{A}(x_1^n) \stackrel{\text{def}}{=} \left\{(\mathbb{I}_A(x_i))_{i=1}^n : A \in \mathcal{A}\right\} \subseteq \{0,1\}^n$$

with respect to the metric $\rho(b,c) \stackrel{\text{def}}{=} n^{-1/2}\|b-c\|_2$ for $b, c \in \{0,1\}^n$, see [87, page 29] for definitions and notation. Haussler [66] showed that, if $\mathcal{A}$ has a finite VC-dimension, then the associated covering number is bounded in the following way: for any $0 < r \leq 1$,

$$\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right) \leq e(V_{\mathcal{A}}+1)\left(\frac{2e}{r^2}\right)^{V_{\mathcal{A}}}. \quad (3.43)$$

Using (3.43) in (3.42) we can bound the integral as follows: for any points $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\begin{aligned}\int_0^1 \sqrt{\log\left(2\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right)\right)}\,\mathrm{d}r &\leq \int_0^1 \sqrt{\log\left(2e(V_{\mathcal{A}}+1)\right) + V_{\mathcal{A}}\log\left(2e/r^2\right)}\,\mathrm{d}r \\ &\leq \sqrt{V_{\mathcal{A}}}\int_0^1 \sqrt{3\log 2 + 2 - 2\log r}\,\mathrm{d}r \\ &\leq 2.44\sqrt{V_{\mathcal{A}}},\end{aligned}$$

by numerical integration or by expressing the integral in terms of the standard normal cumulative distribution function. Combining this result and (3.42), we get

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leq 59\sqrt{\frac{V_{\mathcal{A}}}{n}}. \quad (3.44)$$

This bound is of the right order but does not use the extreme nature of the events in the class $\mathcal{A}$. To this end, we use the *conditioning trick* in [83, Lemma 2], which states that if $K \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbb{I}\{X_i \in B\}$, with $B \supseteq \bigcup_{A\in\mathcal{A}} A$ as in the statement of the theorem, we have the distributional equality

$$\left[(P_n(A))_{A\in\mathcal{A}} \,\middle|\, K = k\right] \stackrel{\mathcal{L}}{=} \left(\frac{k}{n} P_k^Y(A)\right)_{A\in\mathcal{A}},$$

where P_k^Y, for $k \in \{1,\dots,n\}$, is the empirical measure associated to an independent random sample $Y_1,\dots,Y_k$ from the conditional distribution $P(\,\cdot \mid B)$. It easily follows that [83, Lemma 6]

$$\mathsf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \le \mathsf{E}\left[\frac{K}{n}\mathsf{E}\left[\sup_{A\in\mathcal{A}} \left|P_K^Y(A) - P(A \mid B)\right| \,\middle|\, K\right]\right] + \sqrt{\frac{\kappa}{n}}.$$

The conditional expectation on the right-hand side is bounded by means of (3.44), giving

$$\begin{aligned}\mathsf{E}\left[\frac{K}{n}\mathsf{E}\left[\sup_{A\in\mathcal{A}} \left|P_K^Y(A) - P(A \mid B)\right| \,\middle|\, K\right]\right] &\le \mathsf{E}\left[\frac{K}{n}\cdot 59\sqrt{\frac{V_{\mathcal{A}}}{K}}\right] \\ &\le \frac{59}{n}\sqrt{V_{\mathcal{A}}}\sqrt{\mathsf{E}[K]} \le 59\sqrt{\frac{\kappa}{n}V_{\mathcal{A}}}\end{aligned}$$

in view of Jensen's inequality and $\mathsf{E}[K] = n\kappa$. Combining everything and using the fact that $V_{\mathcal{A}} \ge 1$, we get the result. □

3.B Auxiliary results

We will use the following minor results in the proof of Theorem 3.3.1.

Lemma 3.B.1. *Let $x, v \in [1,\infty)^d$, $p \in [1,\infty]$ and $h \in [0,1/2)$. If*

$$\forall j \in \{1,\dots,d\}, \qquad \left|\frac{x_j}{v_j} - 1\right| \le h,$$

then

$$\left|\frac{\|v\|_p}{\|x\|_p} - 1\right| \le \frac{h}{1-h} \stackrel{\text{def}}{=} \Delta < 1,$$

which implies

$$\forall r > 0 : \|x\|_p \ge \frac{r}{1-\Delta} \implies \|v\|_p \ge r \implies \|x\|_p \ge \frac{r}{1+\Delta}.$$

Proof. By hypothesis, for every $j \in \{1, \dots, d\}$, we have

$$1 - h \leq \frac{x_j}{v_j} \leq 1 + h,$$

hence

$$\frac{1}{1+h} \leq \frac{v_j}{x_j} \leq \frac{1}{1-h}$$

and we deduce

$$\left| \frac{v_j}{x_j} - 1 \right| \leq \frac{h}{1-h}.$$

In particular,

$$\forall j \in \{1, \dots, d\} : \qquad |v_j - x_j| \leq \frac{h}{1-h} x_j.$$

Taking the p-th power on each side, summing over j and taking the p-th square root leads to

$$\left| \|v\|_p - \|x\|_p \right| \leq \|v - x\|_p \leq \frac{h}{1-h} \|x\|_p.$$

This shows the first part of the lemma. To obtain the second part, just note that the first part guarantees that

$$\|x\|_p (1 - \Delta) \leq \|v\|_p \leq \|x\|_p (1 + \Delta),$$

Those inequalities imply the second statement. □

Lemma 3.B.2. *For every $j \in \{1, \dots, d\}$ and $x_j > 0$, we have*

$$\left| \frac{x_j}{\hat{v}_j(x_j)} - 1 \right| \leq \left| x_j P_{n,j}((x_j, \infty)) - 1 \right| + \frac{|x_j - 1|}{n}.$$

Proof. By definition of the transform $\hat{v}$, we have for every $j \in \{1, \dots, d\}$ and $x_j > 0$,

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

Therefore, by the triangle inequality,

$$\begin{aligned}
\left| \frac{x_j}{\hat{v}_j(x_j)} - 1 \right| &= \left| \frac{x_j \left(n P_{n,j}((x_j, \infty)) + 1 \right)}{n+1} - 1 \right| \\
&= \left| \frac{x_j \left(n P_{n,j}((x_j, \infty)) + 1 \right) - (n+1)}{n+1} \right| \\
&= \left| \frac{n \left(x_j P_{n,j}((x_j, \infty)) - 1 \right) + (x_j - 1)}{n+1} \right| \\
&\leq \left| x_j P_{n,j}((x_j, \infty)) - 1 \right| + \frac{|x_j - 1|}{n}.
\end{aligned}$$

□

Lemma 3.B.3. *Let $\|\cdot\|$ be a norm on a real vector space and write $\theta(z) = z/\|z\|$ for non-zero vector z. For non-zero vectors x and y, we have*

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

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 $\|x\| = 1 \ge \|y\| > 0$, in which case we need to show that

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

By the triangle inequality, we have

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

and

$$\|y - \theta(y)\| = \left|1 - \frac{1}{\|y\|}\right| \cdot \|y\| = |\|y\| - 1| = |\|y\| - \|x\|| \le \|y - x\|. \qquad \square$$

3.C Proof of Theorem 3.3.1

Proof of Theorem 3.3.1. The empirical exponent and angular measures $\widehat{\nu}$ and $\widehat{\Phi}_p$ only depend on the sample $X_1, \ldots, X_n$ through the transformed data

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

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 the marginal cumulative distribution function 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 $\widehat{\nu}$ and $\widehat{\Phi}_p$ are rank statistics implies that for the sake of the proof, we may and will henceforth assume that the margins $F_1, \ldots, F_d$ are unit-Pareto.

Abbreviate $\Gamma_A^{\pm} = \Gamma_A^{\pm}(r_{\pm}, 3\Delta_1)$ with $r_{\pm} = 1 \pm \Delta_2$ and $0 < \Delta_1, \Delta_2 < 1$ are as in the statement of the theorem. Since $r_- \le 1 \le r_+$, we have

$$\forall A \in \mathcal{A}, \qquad \Gamma_A^- \subseteq C_A \subseteq \Gamma_A^+.$$

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

$$\forall A \in \mathcal{A}, \qquad \tfrac{n}{k}\Gamma_A^- \subseteq \widehat{\Gamma}_A \subseteq \tfrac{n}{k}\Gamma_A^+. \tag{3.45}$$

Then, on the event $\mathcal{E}_1$, the decomposition (3.19) 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.

Bias term. Taking the supremum over $A \in \mathcal{A}$ immediately yields the first term on the right-hand side of the bound (3.23).

Stochastic error. We apply Theorem 3.A.1 to the collection

$$\mathcal{F} \stackrel{\text{def}}{=} \left\{ \tfrac{n}{k}\Gamma_A^\sigma : \ \sigma \in \{-,+\},\ A \in \mathcal{A} \right\}, \tag{3.46}$$

which has finite VC dimension $V_{\mathcal{F}}$ in view of Assumption 3.3.2 and the paragraph following Theorem 3.3.1; the rescaling by the factor n/k obviously does not change the VC dimension. For every $A \in \mathcal{A}$, we have

$$\tfrac{n}{k}\Gamma_A^- \subseteq \tfrac{n}{k}\Gamma_A^+ \subseteq \left\{ x \in [0,\infty)^d : \ \|x\|_p \geq \tfrac{n}{k}\tfrac{1}{1+\Delta_2} \right\}.$$

By the equivalence of norms (3.25), since the margins of P are unit-Pareto, it follows that the probability κ appearing in Theorem 3.A.1 applied to $\mathcal{F}$ is bounded by

$$P\left[\bigcup_{j=1}^d \left\{ x \in [0,\infty)^d : x_j \geq \tfrac{n}{k}\tfrac{1}{d^{1/p}(1+\Delta_2)} \right\} \right] \leq d^{1+1/p}(1+\Delta_2)\tfrac{k}{n}. \tag{3.47}$$

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

$$\begin{aligned} &\sup_{A\in\mathcal{A},\sigma\in\{-,+\}} \tfrac{n}{k}\left| P_n\left(\tfrac{n}{k}\Gamma_A^\sigma\right) - P\left(\tfrac{n}{k}\Gamma_A^\sigma\right)\right| \\ &\leq \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}} + 2\sqrt{\log((d+1)/\delta)}\right) + \frac{2}{3k}\log((d+1)/\delta). \end{aligned}$$

This is the second term on the right-hand side of the bound (3.23). 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 $1 - \delta$.

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

$$\begin{aligned} \nu(\Gamma_A^+ \setminus \Gamma_A^-) &\leq \nu\left(\left\{ x \in [0,\infty)^d : \ (1+\Delta_2)^{-1} \leq \|x\|_p \leq (1-\Delta_2)^{-1} \right\}\right) \\ &+ \nu\left(\left\{ x \in [0,\infty)^d : \ \|x\|_p \geq 1,\ \theta_p(x) \in A_+(3\Delta_1\|x\|_p) \setminus A_-(3\Delta_1\|x\|_p) \right\}\right). \end{aligned} \tag{3.48}$$

The first term on the right-hand side in (3.48) is equal to

$$((1+\Delta_2) - (1-\Delta_2))\ \nu(\{x \in [0,\infty)^d : \|x\|_p \geq 1\}) = 2\Delta_2\Phi_p(\mathbb{S}_p).$$

The second term on the right-hand side in (3.48) can be computed via the product representation (3.8) of ν in polar coordinates: in view of (3.21) in Assumption 3.3.1, the result is

$$\int_1^\infty \Phi_p\big(A_+(3\Delta_1 r) \setminus A_-(3\Delta_1 r)\big)\frac{\mathrm{d}r}{r^2} \le \int_1^\infty \min\left\{\Phi_p(\mathbb{S}_p), 3c\Delta_1 r\right\}\frac{\mathrm{d}r}{r^2} = 3c\Delta_1\left(1 + \log\Phi_p(\mathbb{S}_p) - \log(3c\Delta_1)\right),$$

since $\int_1^\infty \min(b, ar)\frac{\mathrm{d}r}{r^2} = a(\log(b/a)+1)$ for $a \in (0, b]$ and $3c\Delta_1 \le 1 \le \Phi_p(\mathbb{S}_p)$ by assumption.

The resulting upper bound in (3.48) does not depend on $A \in \mathcal{A}$. Bounding $\Phi_p(\mathbb{S}_p)$ by d, we obtain the third term on the right-hand side of the bound (3.23).

Construction of the event $\mathcal{E}_1$. We still need to construct an event $\mathcal{E}_1$ with probability at least $1 - d\delta/(d+1)$ on which the inclusions (3.45) hold. To do this, we apply Theorem 3.A.1 to each of the collections

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

Fix $j = 1, \dots, d$ and let $P_{n,j} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n \delta_{X_{ij}}$. 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 $\kappa = \frac{k}{n}\rho^{-1}$. The class $\mathcal{F}_j$ has VC dimension 1. By Theorem 3.A.1, there exists an event $\mathcal{E}_{1,j}$ with probability at least $1 - \delta/(d+1)$ on which

$$\sup_{x_j \ge \frac{n}{k}\rho} \left|P_{n,j}((x_j,\infty)) - x_j^{-1}\right| \le \frac{k}{n}\left\{\sqrt{\frac{1}{k\rho}}\left(60 + 2\sqrt{\log((d+1)/\delta)}\right) + \frac{2}{3k}\log((d+1)/\delta)\right\} = \frac{k}{n}\Delta_1. \quad (3.49)$$

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

$$\forall x_j \ge \tfrac{n}{k}\rho, \qquad \frac{n+1}{nx_j^{-1} + k\Delta_1 + 1} \le \hat{v}_j(x_j) \le \frac{n+1}{(nx_j^{-1} - k\Delta_1)_+ + 1}. \quad (3.50)$$

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

$$\forall x_j \le \tfrac{n}{k}\rho, \qquad \hat{v}_j(x_j) \le \hat{v}_j(\tfrac{n}{k}\rho) \le \frac{n+1}{k(\rho^{-1} - \Delta_1) + 1} \le \frac{n}{k}\frac{\rho}{1 - \rho\Delta_1} \le \frac{n}{k}\tau. \quad (3.51)$$

Let $\mathcal{E}_1 \overset{\text{def}}{=} \bigcap_{j=1}^d \mathcal{E}_{1,j}$, 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 (3.45) 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 $\|\hat{v}(x)\|_p \geq \frac{n}{k}$ but there exists $j = 1, \dots, d$ with $x_j \leq \frac{n}{k}\rho$, then, by (3.51), we have on $\mathcal{E}_1$ the bound

$$\theta_{p,j}(\hat{v}(x)) = \frac{\hat{v}_j(x_j)}{\|\hat{v}(x)\|_p} \leq \tau$$

and thus, by Assumption 3.3.1, necessarily $\theta_p(\hat{v}(x)) \notin A$ for all $A \in \mathcal{A}$. Hence, on $\mathcal{E}_1$, we have

$$\widehat{\Gamma}_A = \left\{ x \in (\tfrac{n}{k}\rho, \infty)^d : \ \|\hat{v}(x)\|_p \geq \tfrac{n}{k},\ \theta_p(\hat{v}(x)) \in A \right\}. \tag{3.52}$$

Step 2: Radial framing

We consider the following decomposition of $\widehat{\Gamma}_A$

$$\left(\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_\infty \leq 2\tfrac{n}{k}\}\right) \cup \left(\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_\infty > 2\tfrac{n}{k}\}\right),$$

and we frame each set separately.

For points x with $\|x\|_\infty \leq 2\frac{n}{k}$, we seek to apply Lemma 3.B.1. To this end, we first construct $h \in [0, 1/2)$ such that, on $\mathcal{E}_1$, we have

$$\forall j \in \{1, \dots, d\} : \qquad \left| \frac{x_j}{\hat{v}_j(x_j)} - 1 \right| \leq h,$$

for every $x_j > \frac{n}{k}\rho$ (which is larger than 1, by assumption that $\rho > k/n$). We simply apply Lemma 3.B.2 combined with (3.49), giving

$$\left| \frac{x_j}{\hat{v}_j(x_j)} - 1 \right| \leq \tfrac{k}{n}\Delta_1 x_j + \tfrac{|x_j - 1|}{n} \leq \tfrac{k}{n} x_j \left(\Delta_1 + \tfrac{1}{k}\right) \leq 2\left(\Delta_1 + \tfrac{1}{k}\right) \overset{\text{def}}{=} h,$$

since $x_j \leq 2\frac{n}{k}$. Hence, we may apply Lemma 3.B.1 with $r = n/k$ and $\Delta = \Delta_2$ since

$$\frac{h}{1-h} \leq 4(\Delta_1 + 1/k) = \Delta_2.$$

This leads to the inclusions

$$\begin{aligned} &\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_\infty \leq 2\tfrac{n}{k}\} \\ &\qquad \subseteq \left\{ x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_p \geq \tfrac{n}{k} \tfrac{1}{1+\Delta_2} \right\} \cap \{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_\infty \leq 2\tfrac{n}{k}\} \end{aligned}$$

and

$$\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty \le 2\tfrac{n}{k}\} \supseteq \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_p \ge \tfrac{n}{k}\tfrac{1}{1-\Delta_2}\right\} \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty \le 2\tfrac{n}{k}\}.$$

For points x with $\|x\|_\infty > 2\frac{n}{k}$, similar inclusions hold trivially. Indeed, on this set, by definition of the L_∞-norm, there exists $j \in \{1,\dots,d\}$ such that $x_j > 2\frac{n}{k}$. Therefore, by our assumption that k is large enough such that $\Delta_1 + 1/k \le 1/4$, we have by (3.50),

$$\hat{v}_j(x_j) \ge \frac{n+1}{\frac{k}{2}+k\Delta_1+1} \ge \frac{n}{k}\frac{1}{\frac{1}{2}+(\Delta_1+\frac{1}{k})} \ge \frac{n}{k}.$$

Hence, we have $\|\hat{v}(x)\|_\infty \ge n/k$, and, by equivalence of norms (3.25), also $\|\hat{v}(x)\|_p \ge n/k$ for every $p \in [1,\infty]$. Furthermore, $\|x\|_p \ge \|x\|_\infty \ge 2\frac{n}{k} \ge \frac{n}{k}\frac{1}{1+\Delta_2}$ for every $0 < \Delta_2 < 1$. These inequalities imply the inclusions

$$\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty > 2\tfrac{n}{k}\} \subseteq \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_p \ge \tfrac{n}{k}\tfrac{1}{1+\Delta_2}\right\} \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty > 2\tfrac{n}{k}\}$$

and

$$\widehat{\Gamma}_A \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty > 2\tfrac{n}{k}\} \supseteq \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_p \ge \tfrac{n}{k}\tfrac{1}{1-\Delta_2}\right\} \cap \{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_\infty > 2\tfrac{n}{k}\}.$$

Combining the two cases, we get radial framing, i.e.,

$$\begin{aligned} \widehat{\Gamma}_A &\subseteq \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_p \ge \tfrac{n}{k}\tfrac{1}{1+\Delta_2}\right\}, \\ \widehat{\Gamma}_A &\supseteq \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \|x\|_p \ge \tfrac{n}{k}\tfrac{1}{1-\Delta_2}\right\}. \end{aligned} \tag{3.53}$$

Step 3: Angular framing

We will show that, on $\mathcal{E}_1$, for any $x \in (\frac{n}{k}\rho,\infty)^d$ and $A \in \mathcal{A}$,

$$\theta_p(x) \in A_-(3\tfrac{k}{n}\Delta_1\|x\|_p) \implies \theta_p(\hat{v}(x)) \in A \implies \theta_p(x) \in A_+(3\tfrac{k}{n}\Delta_1\|x\|_p). \tag{3.54}$$

In combination with (3.53), this will show the inclusions (3.45) and thus finish the proof.

Fix such x and A. Put $\varepsilon \stackrel{\text{def}}{=} 3\frac{k}{n}\Delta_1\|x\|_p$. We consider two cases: $\varepsilon < 1$ and $\varepsilon \ge 1$.

If $\varepsilon \geq 1$, the two implications in (3.54) are trivially fulfilled: Since the $\|\cdot\|_p$-diameter of $\mathbb{S}_p$ is equal to 1, we have $A_-(\varepsilon) = \emptyset$ (as $\mathbb{S}_p \setminus A$ is not-empty) while $A_+(\varepsilon) = \mathbb{S}_p$.

The interesting case is thus $\varepsilon < 1$. By Lemma 3.B.3, we have

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

Since $\varepsilon < 1$, we have $\|x\|_\infty \leq \|x\|_p \leq (n/k)/(3\Delta_1)$. Hence, for all $j = 1, \dots, d$, we have $nx_j^{-1} - k\Delta_1 \geq n \cdot 3\frac{k}{n}\Delta_1 - k\Delta_1 > 0$. By (3.50), we deduce

$$\begin{aligned}
|\hat{v}_j(x_j) - x_j| &\leq \max_{\sigma\in\{-1,+1\}} \left| \frac{n+1}{nx_j^{-1} + \sigma k\Delta_1 + 1} - x_j \right| \\
&= x_j \max_{\sigma\in\{-1,+1\}} \left| \frac{n+1}{n + (\sigma k\Delta_1 + 1)x_j} - 1 \right| \\
&= x_j \max_{\sigma\in\{-1,+1\}} \frac{|1 - (\sigma k\Delta_1 + 1)x_j|}{n + (\sigma k\Delta_1 + 1)x_j}.
\end{aligned}$$

Recall that $x_j \geq \frac{n}{k}\rho \geq 1$ and $k\Delta_1 \geq 2$. For he case $\sigma = +1$, we have

$$\frac{|1 - (+k\Delta_1 + 1)x_j|}{n + (+k\Delta_1 + 1)x_j} \leq \frac{(k\Delta_1 + 1)x_j}{n + (k\Delta_1 + 1)x_j} \leq \frac{3}{2}\frac{k}{n}\Delta_1 x_j.$$

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

$$\frac{|1 - (-k\Delta_1 + 1)x_j|}{n + (-k\Delta_1 + 1)x_j} = \frac{(k\Delta_1 - 1)x_j - 1}{n + (-k\Delta_1 + 1)x_j} \leq \frac{k\Delta_1 x_j}{n - k\Delta_1 x_j} \leq \frac{3}{2}\frac{k}{n}\Delta_1 x_j,$$

since $\varepsilon < 1$ implies $n - k\Delta_1 x_j \geq n - k\Delta_1\|x\|_p \geq n - n/3 = 2n/3$. We deduce that

$$\forall j \in \{1, \dots, d\} : \qquad |\hat{v}_j(x_j) - x_j| \leq \frac{3}{2}\frac{k}{n}\Delta_1 x_j^2.$$

Consequently,

$$\|\hat{v}(x) - x\|_p \leq \frac{3}{2}\frac{k}{n}\Delta_1\|x\|_p^2,$$

where we use the (easy to prove) inequality $\|(x_1^2, \dots, x_d^2)\|_p \leq \|(x_1, \dots, x_d)\|_p^2$, and thus

$$\|\theta_p(\hat{v}(x)) - \theta_p(x)\|_p \leq 3\frac{k}{n}\Delta_1\|x\|_p = \varepsilon.$$

The implications (3.54) now follow by definition of $A_-(\varepsilon)$ and $A_+(\varepsilon)$.

We conclude that, on $\mathcal{E}_1$, the inclusions (3.45) hold, as required. The proof Theorem 3.3.1 is complete. □

3.D Proofs of Remark 3.3.3

As $\|x\|_p \geq \rho$ implies $\|x\|_\infty \geq d^{-1/p}\rho$ for $x \in \mathbb{R}^d$ and $\rho > 0$, the bias term appearing in Theorem 3.3.1 is bounded by the total variation distance between the measures $\frac{n}{k}P_V(\frac{n}{k}\cdot)$ and ν restricted to

$$\mathbb{E}_c \stackrel{\text{def}}{=} \left\{x \in (0,\infty)^d : \max(x) \geq c\right\},$$

for $c \stackrel{\text{def}}{=} d^{-1/p} r_+^{-1}$. (This c is not the same as the one in the statement of Theorem 3.3.1.) More precisely, we have

$$\sup_{A\in\mathcal{A},\, \sigma\in\{+,-\}} \left|\tfrac{n}{k}P_V(\tfrac{n}{k}\Gamma_A^\sigma) - \nu(\Gamma_A^\sigma)\right| \leq \sup_{B\in\mathcal{B}(\mathbb{E}_c)} \left|\tfrac{n}{k}P_V(\tfrac{n}{k}B) - \nu(B)\right|,$$

where $\mathcal{B}(\mathbb{E}_c)$ denotes the Borel σ-field on $\mathbb{E}_c$. Recall p_U and λ in Remark 3.3.3 and define

$$\mathcal{D}_T(s) \stackrel{\text{def}}{=} \int_{L_T} \left|s^{d-1}p_U(sy) - \lambda(y)\right| \,\mathrm{d}y,$$

with

$$L_T \stackrel{\text{def}}{=} \{y \in (0,T]^d : \min(y) \leq 1\}.$$

The following proposition provides a refined version of the bound (3.27) in Remark 3.3.3 in the paper; the bound (3.27) itself appears in the course of the proof of the proposition.

Proposition 3.D.1. *For $c > 0$, let $\mathcal{B}(\mathbb{E}_c)$ denote the Borel σ-field on $\mathbb{E}_c$. Then, for $t \geq 1/c$,*

$$\sup_{B\in\mathcal{B}(\mathbb{E}_c)} |tP_V(tB) - \nu(B)| \leq \tfrac{1}{c}\mathcal{D}_{ct}(\tfrac{1}{ct}) + \nu\left(\left\{x \in (0,\infty)^d : \max(x) \geq c, \min(x) \leq 1/t\right\}\right).$$

Proof. Writing $u = t^{-1}$, we get

$$\begin{aligned}\sup_{B\in\mathcal{B}(\mathbb{E}_c)} |t\,P_V(tB) - \nu(B)| &= \sup_{B\in\mathcal{B}(\mathbb{E}_c)} \left|t\,P_U(t^{-1}\iota(B)) - \Lambda(\iota(B))\right| \\ &= \sup_{B\in\mathcal{B}(\iota(\mathbb{E}_c))} \left|u^{-1}\,P_U(uB) - \Lambda(B)\right|.\end{aligned}$$

For any Borel set $B \subseteq (0,\infty)^d$, we have, by a change of variables,

$$u^{-1}\,P_U(uB) = u^{-1}\int_{uB} p_U(z)\,\mathrm{d}z = u^{d-1}\int_B p_U(uy)\,\mathrm{d}y.$$

It follows that

$$\begin{aligned}\sup_{B\in\mathcal{B}(\mathbb{E}_c)} |t\,P_V(tB) - \nu(B)| &= \sup_{B\in\mathcal{B}(\iota(\mathbb{E}_c))} \left| \int_B \left(u^{d-1} p_U(uy) - \lambda(y) \right) \mathrm{d}y \right| \\ &\le \int_{\iota(\mathbb{E}_c)} \left| u^{d-1} p_U(uy) - \lambda(y) \right| \mathrm{d}y \\ &= \int_{0<\min(y)\le 1/c} \left| u^{d-1} p_U(uy) - \lambda(y) \right| \mathrm{d}y,\end{aligned}$$

which is (3.27) in the paper with $u = k/n$ and $c = d^{-1/p} r_+^{-1}$. Since the copula density p_U vanishes outside $[0,1]^d$, we get

$$\begin{aligned}\sup_{B\in\mathcal{B}(\mathbb{E}_c)} |t\,P_V(tB) - \nu(B)| = &\int_{\substack{0<\min(y)\le 1/c \\ \max(y)\le 1/u}} \left| u^{d-1} p_U(uy) - \lambda(y) \right| \mathrm{d}y \\ &+ \int_{\substack{0<\min(y)\le 1/c \\ \max(y)>1/u}} \lambda(y)\,\mathrm{d}y.\end{aligned}$$

The first integral on the right-hand side is denoted by $\mathcal{I}(u,c)$ and is analysed below. The second integral on the right-hand side is

$$\begin{aligned}&\Lambda\left(\left\{ y \in (0,\infty)^d : \min(y) \le 1/c,\ \max(y) > 1/u \right\} \right) \\ &\qquad = \nu\left(\left\{ x \in (0,\infty)^d : \max(x) \ge c,\ \min(x) < u \right\} \right).\end{aligned}$$

By a change of variables $y = c^{-1}x$ (component-wise), we find

$$\mathcal{I}(u,c) = c^{-d} \int_{\substack{0<\min(x)\le 1 \\ \max(x)\le c/u}} \left| u^{d-1} p_U(uc^{-1}x) - \lambda(c^{-1}x) \right| \mathrm{d}y.$$

The density λ is homogeneous of order $1-d$, i.e., $\lambda(c^{-1}x) = c^{d-1}x$. Writing $c^{-1}u = s$, we find

$$\begin{aligned}\mathcal{I}(u,c) &= c^{-d} \int_{\substack{0<\min(x)\le 1 \\ \max(x)\le c/u}} \left| c^{d-1} s^{d-1} p_U(sx) - c^{d-1}\lambda(x) \right| \mathrm{d}x \\ &= c^{-1} \int_{\substack{0<\min(x)\le 1 \\ \max(x)\le 1/s}} \left| s^{d-1} p_U(sx) - \lambda(x) \right| \mathrm{d}x \\ &= c^{-1} \mathcal{D}_{1/s}(s).\end{aligned}$$

Substituting $s = c^{-1}u = (ct)^{-1}$ yields the stated bound. □

Example 3.D.1 (The multivariate Cauchy distribution). Let us assume that X follows a multivariate Cauchy distribution on the positive orthant whose density is given by

$$f(x) \overset{\text{def}}{=} \frac{2^d \Gamma(\frac{1+d}{2})}{\pi^{\frac{1+d}{2}}} \frac{1}{(1+\|x\|_2^2)^{\frac{1+d}{2}}}$$

for $x \geq 0$ and $f(x) = 0$ otherwise. We will show that the bound in Proposition 3.D.1 is $\mathrm{O}(1/t)$ as $t \to \infty$. This implies that the bias term is $\mathrm{O}(k/n)$ as $k = k_n \to \infty$ in such a way that $k/n \to 0$.

For simplicity of notation, let $p = p_U$ denote the probability density function of $U = (1 - F_1(X_1), \ldots, 1 - F_d(X_d))$. Some computations related to the univariate Cauchy distribution lead to

$$p(x) = \pi^{\frac{d-1}{2}} \Gamma(\tfrac{1+d}{2}) \frac{\prod_{i=1}^d \left(1 + \tan^2(\frac{\pi}{2}(1-x_i))\right)}{\left(1 + \sum_{i=1}^d \tan^2(\frac{\pi}{2}(1-x_i))\right)^{\frac{1+d}{2}}}.$$

Asymptotic considerations on the tangent function permit to show that

$$\lim_{s \to 0} s^{d-1} p(sx) = \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}} \frac{x_1^{-2} \cdots x_d^{-2}}{\left(x_1^{-2} + \cdots + x_d^{-2}\right)^{\frac{1+d}{2}}} \overset{\text{def}}{=} \lambda(x).$$

We start with the first term in the upper bound of Proposition 3.D.1. We have the decomposition

$$\mathcal{D}_{1/s}(s) = \int_{L_{1/s}} \left(s^{d-1} p(sx) - \lambda(x)\right)_+ \mathrm{d}x \tag{3.55}$$

$$+ \int_{L_{1/s}} \left(\lambda(x) - s^{d-1} p(sx)\right)_+ \mathrm{d}x. \tag{3.56}$$

We start by studying the integral (3.55). We observe that for $z \in (0, 1]$

$$\tan^2\left(\tfrac{\pi}{2}(1-z)\right) = \cot^2\left(\tfrac{\pi}{2} z\right) = \frac{1}{\sin^2\left(\frac{\pi}{2} z\right)} - 1 = \frac{2}{1 - \cos(\pi z)} - 1.$$

If $x_i \leq 1$ for every $i \in \{1, \ldots, d\}$, we may apply (3.57) in Lemma 3.D.1 below

to get

$$\begin{aligned}
p(x) &= \pi^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\prod_{i=1}^{d}\left(1+\tan^2(\frac{\pi}{2}(1-x_i))\right)}{\left(1+\sum_{i=1}^{d}\tan^2(\frac{\pi}{2}(1-x_i))\right)^{\frac{1+d}{2}}} \\
&= \pi^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\frac{2}{1-\cos(\pi x_1)}\cdots\frac{2}{1-\cos(\pi x_d)}}{\left(1-d+\frac{2}{1-\cos(\pi x_1)}+\cdots+\frac{2}{1-\cos(\pi x_d)}\right)^{\frac{1+d}{2}}} \\
&= (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\frac{1}{1-\cos(\pi x_1)}\cdots\frac{1}{1-\cos(\pi x_d)}}{\left(\frac{1-d}{2}+\frac{1}{1-\cos(\pi x_1)}+\cdots+\frac{1}{1-\cos(\pi x_d)}\right)^{\frac{1+d}{2}}} \\
&\le (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\left(\frac{2}{\pi^2x_1^2}+\frac{1}{3}\right)\cdots\left(\frac{2}{\pi^2x_d^2}+\frac{1}{3}\right)}{\left(\frac{1-d}{2}+\frac{2}{\pi^2x_1^2}+\frac{1}{6}+\cdots+\frac{2}{\pi^2x_d^2}+\frac{1}{6}\right)^{\frac{1+d}{2}}}.
\end{aligned}$$

Hence, for $0 < s \le 1$ and $x \in L_{1/s}$, we have the upper bound

$$\begin{aligned}
s^{d-1}p(sx) &\le (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{s^{d-1}\left(\frac{2}{\pi^2s^2x_1^2}+\frac{1}{3}\right)\cdots\left(\frac{2}{\pi^2s^2x_d^2}+\frac{1}{3}\right)}{\left(\frac{1-d}{2}+\frac{2}{\pi^2s^2x_1^2}+\frac{1}{6}+\cdots+\frac{2}{\pi^2s^2x_d^2}+\frac{1}{6}\right)^{\frac{1+d}{2}}} \\
&= \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{\left(\frac{1}{x_1^2}+\frac{\pi^2s^2}{6}\right)\cdots\left(\frac{1}{x_d^2}+\frac{\pi^2s^2}{6}\right)}{\left(\frac{1}{x_1^2}+\cdots+\frac{1}{x_d^2}-\frac{\pi^2s^2}{4}(\frac{2d}{3}-1)\right)^{\frac{1+d}{2}}}.
\end{aligned}$$

Note that the upper bound converges to $\lambda(x)$ as $s \to 0$.

Now observe that

$$\begin{aligned}
&\frac{1}{\left(\frac{1}{x_1^2}+\cdots+\frac{1}{x_d^2}-\frac{\pi^2s^2}{4}(\frac{2d}{3}-1)\right)^{\frac{1+d}{2}}} \\
&\qquad = \frac{1}{(x_1^{-2}+\cdots+x_d^{-2})^{\frac{1+d}{2}}}\frac{1}{\left(1-\frac{\pi^2}{4}(\frac{2d}{3}-1)\frac{s^2}{x_1^{-2}+\cdots+x_d^{-2}}\right)^{\frac{1+d}{2}}},
\end{aligned}$$

with

$$\frac{\pi^2}{4}(\tfrac{2d}{3}-1)\frac{s^2}{x_1^{-2}+\cdots+x_d^{-2}} \le \frac{\pi^2}{4}(\tfrac{2d}{3}-1)s^2.$$

Since we are interested in the limit $s \to 0$, we may assume that $s^2 \le \frac{2}{\pi^2(\frac{2d}{3}-1)}$

so that we may apply the upper bound in (3.60) in Lemma 3.D.2 below. We get

$$\begin{aligned} s^{d-1}p(sx) &\le \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}} \frac{\left(\frac{1}{x_1^2}+\frac{\pi^2 s^2}{6}\right)\cdots\left(\frac{1}{x_d^2}+\frac{\pi^2 s^2}{6}\right)}{\left(\frac{1}{x_1^2}+\cdots+\frac{1}{x_d^2}-\frac{\pi^2 s^2}{4}(\frac{2d}{3}-1)\right)^{\frac{1+d}{2}}} \\ &\le \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}} \frac{\left(x_1^{-2}+\frac{\pi^2 s^2}{6}\right)\cdots\left(x_d^{-2}+\frac{\pi^2 s^2}{6}\right)}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{1+d}{2}}} \\ &\quad \cdot\left(1+s^2\frac{2\left(2^{\frac{1+d}{2}}-1\right)\frac{\pi^2}{4}(\frac{2d}{3}-1)}{x_1^{-2}+\cdots+x_d^{-2}}\right). \end{aligned}$$

Defining $C_1, C_2 > 0$ by

$$C_1 \stackrel{\text{def}}{=} \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}} \qquad \text{and} \qquad C_2 \stackrel{\text{def}}{=} \tfrac{\pi^2}{2}\left(2^{\frac{1+d}{2}}-1\right)(\tfrac{2d}{3}-1),$$

we thus have the bound

$$\begin{aligned} (s^{d-1}p(sx)-\lambda(x))_+ \le s^2 C_1 C_2 &\frac{x_1^{-2}\cdots x_d^{-2}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{3+d}{2}}} \\ &+ C_1 \frac{\sum_{k=0}^{d-1}\sum_{I\subset\{1,\dots,d\},|I|=k} x_I^{-2}(\frac{\pi^2 s^2}{6})^{d-k}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{1+d}{2}}} \\ &+ s^2 C_1 C_2 \frac{\sum_{k=0}^{d-1}\sum_{I\subset\{1,\dots,d\},|I|=k} x_I^{-2}(\frac{\pi^2 s^2}{6})^{d-k}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{3+d}{2}}}, \end{aligned}$$

where x_I denotes the sub-vector of x with coordinates in I and the square is to be understood coordinate-wise.

The first term is the easiest to handle as it is bounded by a multiple of $\lambda(x)$, it is integrable over L_∞ and its contribution to the bias term is at most $O(s^2)$. The third term is bounded by a multiple of the second term times s^2 and is therefore negligible in front of the second term. Now note that that the domain of integration $L_{1/s}$ can be rewritten as

$$L_{1/s} = \bigcup_{\emptyset\neq J\subset\{1,\dots,d\}} B_{J,1/s}$$

where

$$B_{J,1/s} \stackrel{\text{def}}{=} \left\{ x \in (0,1/s]^d : \forall j \in J, x_j \leq 1; \forall j \in J^c : x_j > 1 \right\}.$$

Hence, integrating the second term over $L_{1/s}$ implies that we will need to bound integrals of the form

$$s^{2(d-|I|)} \int_{B_{J,1/s}} \frac{\prod_{i\in I} x_i^{-2}}{\left(x_1^{-2} + \cdots + x_d^{-2}\right)^{\frac{1+d}{2}}} \mathrm{d}x,$$

for subsets $I, J \subset \{1, \ldots, d\}$ with $I^c \neq \emptyset$ and $J \neq \emptyset$. Without loss of generality we may assume that $J = \{1, \ldots, j\}$ for some $j \in \{1, \ldots, d\}$ and therefore, a change of variable $x_\ell = y_\ell^{-1}$ for all $\ell \in \{1, \ldots, d\}$ permits to rewrite the integral as

$$\begin{aligned} &s^{2(d-|I|)} \int_{x_1=0}^{1} \cdots \int_{x_j=0}^{1} \int_{x_{j+1}=1}^{1/s} \cdots \int_{x_d=1}^{1/s} \left(x_1^{-2} + \cdots + x_d^{-2}\right)^{-(1+d)/2} \prod_{i\in I} x_i^{-2} \mathrm{d}x \\ &\quad = s^{2|I^c|} \int_{y_1=1}^{\infty} \cdots \int_{y_j=1}^{\infty} \int_{y_{j+1}=s}^{1} \cdots \int_{y_d=s}^{1} \left(y_1^2 + \cdots + y_d^2\right)^{-(1+d)/2} \prod_{i\in I^c} y_i^{-2} \mathrm{d}y. \end{aligned}$$

If $j = d$, the integral is finite and hence the contribution to the bias is at least of the order $O(s^2)$ as $|I^c| \geq 1$. Thus, we shall assume $j < d$. Observe that for every $i \in I^c$, if $i \leq j$, then we may upper bound $y_i^{-1} \leq 1$. Consequently, the worst case will happen whence $I^c \cap J = \emptyset$. In this case, bounding $\left(y_1^2 + \cdots + y_d^2\right)^{-(1+d)/2}$ by $\left(y_1^2 + \cdots + y_j^2\right)^{-(1+d)/2}$ permits to upper bound the latter integral by

$$s^{2|I^c|} \int_{y_1=1}^{\infty} \cdots \int_{y_j=1}^{\infty} (y_1^2 + \cdots + y_j^2)^{-(1+d)/2} \cdot \mathrm{d}y_1 \cdots \mathrm{d}y_j \cdot \prod_{i\in I^c} \int_{y_i=s}^{1} y_i^{-2} \, \mathrm{d}y_i,$$

which is of order $O(s^{|I^c|})$ as the integral over $(y_1, \ldots, y_j)$ is finite. Since $|I^c| \geq 1$, we find that the order of the first integral in the bias decomposition (3.55) is of the order $O(s)$.

We now claim that second term in the decomposition (3.56) is always equal to zero for sufficiently small $s > 0$ and therefore conclude that $\mathcal{D}_{1/s}(s) = O(s)$ as $s \longrightarrow 0$. Previous computations and bounds (3.57) show that for every x such

that $x_i \le 1$ for $i \in \{1, \dots, d\}$,

$$p(x) = (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\frac{1}{1-\cos(\pi x_1)}\cdots\frac{1}{1-\cos(\pi x_d)}}{\left(\frac{1-d}{2}+\frac{1}{1-\cos(\pi x_1)}+\cdots+\frac{1}{1-\cos(\pi x_d)}\right)^{\frac{1+d}{2}}}$$
$$\ge (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{\left(\frac{2}{\pi^2 x_1^2}+\frac{1}{6}\right)\cdots\left(\frac{2}{\pi^2 x_d^2}+\frac{1}{6}\right)}{\left(\frac{1-d}{2}+\frac{2}{\pi^2 x_1^2}+\frac{1}{3}+\cdots+\frac{2}{\pi^2 x_d^2}+\frac{1}{3}\right)^{\frac{1+d}{2}}}.$$

Hence, for $x \in L_{1/s}$, we have

$$s^{d-1}p(sx) \ge (2\pi)^{\frac{d-1}{2}}\Gamma(\tfrac{1+d}{2})\frac{s^{d-1}\left(\frac{2}{\pi^2 s^2 x_1^2}+\frac{1}{6}\right)\cdots\left(\frac{2}{\pi^2 s^2 x_d^2}+\frac{1}{6}\right)}{\left(\frac{2}{\pi^2 s^2 x_1^2}+\cdots+\frac{2}{\pi^2 s^2 x_d^2}-(\frac{d}{6}-1/2)\right)^{\frac{1+d}{2}}}$$
$$= \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{(x_1^{-2}+\frac{\pi^2 s^2}{12})\cdots(x_d^{-2}+\frac{\pi^2 s^2}{12})}{\left(x_1^{-2}+\cdots+x_d^{-2}-s^2\pi^2(\frac{d}{3}-1)\right)^{\frac{1+d}{2}}}.$$

Note that the lower bound converges to $\lambda(x)$ as $s \to 0$. Since we are interested in $s \to 0$, we may assume that $s^2 \le (2\pi^2(d/3-1))^{-1}$ and apply the lower bound in (3.60) in Lemma 3.D.2 below to get that for every $x \in L_{1/s}$,

$$s^{d-1}p(sx) \ge \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{(x_1^{-2}+\frac{\pi^2 s^2}{12})\cdots(x_d^{-2}+\frac{\pi^2 s^2}{12})}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{1+d}{2}}}\left(1+s^2\frac{\pi^2(\frac{d}{3}-1)(\frac{1+d}{2})}{x_1^{-2}+\cdots x_d^{-2}}\right).$$

This shows in particular that $s^{d-1}p(sx) - \lambda(x) \ge 0$ for every value of $x \in L_{1/s}$ and $s \le (2\pi^2(d/3-1))^{-1/2}$ so that the integral

$$\int_{L_{1/s}} (\lambda(x) - s^{d-1}p(sx))_+ \mathrm{d}x = 0,$$

for those values of s and we get that the bias $\mathcal{D}_{1/s}(s) = O(s)$ as $s \to 0$.

To deal with the second term in the upper bound of Proposition 3.D.1, we note that since,

$$\frac{\mathrm{d}\nu}{\mathrm{d}x}(x) = \lambda(\iota(x))x_1^{-2}\cdots x_d^{-2} = C_1\|x\|_2^{-1-d},$$

the density associated to Φ_2 is constant, say φ, so that when $t \to \infty$,

$$\nu\left(\left\{x \in (0,\infty)^d : \max(x) \ge c, \min(x) \le 1/t\right\}\right)$$
$$\le c^{-1}\Phi_2\left(\{\theta \in \mathbb{S}_2 : \min(\theta) \le \tfrac{1}{ct}\}\right)$$
$$= c^{-1}\varphi\,\mathrm{leb}_{d-1}\left(\{\theta \in \mathbb{S}_2 : \min(\theta) \le \tfrac{1}{ct}\}\right)$$
$$= O(1/t).$$

This concludes the analysis of the bias term for the multivariate Cauchy distribution.

Lemma 3.D.1. *We have*

$$\frac{1}{6} \leq \frac{1}{1-\cos(t)} - \frac{2}{t^2} \leq \frac{1}{3} \qquad \text{for } 0 < t \leq \pi. \tag{3.57}$$

Proof. Note that we may rewrite the inequality as follows: for $0 \leq t \leq \pi$, we must have

$$\begin{aligned}
\frac{1}{6} \leq \frac{1}{1-\cos t} - \frac{2}{t^2} \leq \frac{1}{3} &\iff \frac{1}{6} + \frac{2}{t^2} \leq \frac{1}{1-\cos t} \leq \frac{1}{3} + \frac{2}{t^2} \\
&\iff \frac{12+t^2}{6t^2} \leq \frac{1}{1-\cos t} \leq \frac{6+t^2}{3t^2} \\
&\iff \frac{3t^2}{6+t^2} \leq 1-\cos t \leq \frac{6t^2}{12+t^2},
\end{aligned}$$

which gives the inequalities

$$\cos t \leq 1 - \frac{3t^2}{6+t^2} = \frac{6-2t^2}{6+t^2} \tag{3.58}$$

$$\cos t \geq 1 - \frac{6t^2}{12+t^2} = \frac{12-5t^2}{12+t^2}. \tag{3.59}$$

We start by showing (3.58). Taylor's theorem implies that

$$\cos t \leq 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \frac{t^8}{8!}.$$

Consequently,

$$\begin{aligned}
(6+t^2)\cos t &\leq (6+t^2) \cdot \left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \frac{t^8}{8!}\right) \\
&= 6 - 3t^2 + \frac{t^4}{4} - \frac{t^6}{120} + \frac{t^8}{6720} + t^2 - \frac{t^4}{2} + \frac{t^6}{24} - \frac{t^8}{720} + \frac{t^{10}}{40320} \\
&= 6 - 2t^2 + \frac{t^4}{40320}\left(t^6 - 50t^4 + 1344t^2 - 10080\right).
\end{aligned}$$

Hence, if we show that the polynomial $R(s) \stackrel{\text{def}}{=} s^3 - 50s^2 + 1344s - 10080$ is strictly negative on $0 \leq s \leq \pi^2$, we are done proving (3.58). It is sufficient to observe that $R(\pi^2) < 0$ and $R'(s) \geq 0$ for the concerned values of s since

$$R'(s) = 3s^2 - 100s + 1344$$

satisfies $\Delta_R = 1000 - 12 \cdot 1344 < 0$ and $R'(0) = 1344 > 0$.

To prove (3.59), we follow the same path: Taylor's theorem implies that

$$\cos t \geq 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \frac{t^6}{6!}.$$

Consequently,

$$\begin{aligned}(12+t^2)\cos t &\geq (12+t^2)\cdot\left(1-\frac{t^2}{2}+\frac{t^4}{4!}-\frac{t^6}{6!}\right)\\ &= 12 - 6t^2 + \frac{t^4}{2} - \frac{t^6}{60} + t^2 - \frac{t^4}{2} + \frac{t^6}{24} - \frac{t^8}{720}\\ &= 12 - 5t^2 + \frac{t^6}{720}\left(-t^2+18\right).\end{aligned}$$

Since $-t^2 + 18 \geq 0$ for $0 \leq t \leq \sqrt{18}$ and $\sqrt{18} > \pi$, we have proven (3.59). □

Lemma 3.D.2. *For $a \in [0, 1/2]$, we have*

$$1 + \left(\frac{1+d}{2}\right)a \leq (1-a)^{-\frac{1+d}{2}} \leq 1 + 2\left(2^{\frac{1+d}{2}} - 1\right)a. \tag{3.60}$$

Proof. To prove the upper bound, we will use a convexity argument. Note that the function

$$\phi(a) \stackrel{\text{def}}{=} (1-a)^{-\frac{1+d}{2}}$$

is a convex function of its argument. Since it is a C^2 function, we may prove that fact by differentiating it two times and show that $\phi''(a) > 0$ for $0 \leq a \leq 1/2$. We compute that

$$\phi''(a) = \tfrac{d+1}{2} \cdot \tfrac{d+3}{2} \cdot (1-a)^{-\frac{d+5}{2}},$$

so that the convexity on the domain of interest is proven. This implies that for any $\lambda \in [0,1]$ and any pair of points $a_1, a_2 \in [0, 1/2]$,

$$\phi(\lambda a_1 + (1-\lambda)a_2) \leq \lambda\phi(a_1) + (1-\lambda)\phi(a_2).$$

In particular, the graph of ϕ is anywhere bounded by the straight line joining the points $\phi(0) = 1$ and $\phi(1/2) = 2^{\frac{1+d}{2}}$ on the domain of interest. This affine function has analytic expression

$$L(a) \stackrel{\text{def}}{=} 1 + 2\left(2^{\frac{1+d}{2}} - 1\right)a$$

and corresponds to the upper bound in the statement.

To prove the lower bound, we use Taylor's theorem which ensures that for every a in the region of interest

$$\phi(a) = \phi(0) + \phi'(0)a + R_0^1\phi(a) \geq \phi(0) + \phi'(0)a,$$

since, by Lagrange's formula, the rest $R_0^1\phi(a) = \frac{\phi''(c)}{2}a^2$ for some $c \in (0, a)$ and the second derivative is positive at this point. Observing that $\phi(0) = 1$ and $\phi'(0) = (1+d)/2$ permits to conclude. □

3.E Proofs of examples

Proof of Example 3.3.1. We first consider Assumption 3.3.1. Restricting a and β to have rational coordinates yields a countable collection $\mathcal{A}_0 \subset \mathcal{A}$ satisfying item (i) in Assumption 3.3.1. Item (ii) is satisfied by definition. For $A_{a,\beta,\tau} \in \mathcal{A}$ and $\varepsilon > 0$, define the inner and outer hulls

$$A_{a,\beta,\tau;-}(\varepsilon) \stackrel{\text{def}}{=} A_{a,\beta-\sqrt{d}\varepsilon,\tau+\varepsilon},$$
$$A_{a,\beta,\tau;+}(\varepsilon) \stackrel{\text{def}}{=} A_{a,\beta+\sqrt{d}\varepsilon,\tau-\varepsilon},$$

with the convention that $\mathbb{S}_p^v = \mathbb{S}_p$ if $v < 0$ (a case which arises if $\varepsilon > \tau$). We show that item (iii) in Assumption 3.3.1 is satisfied, provided the angular measure Φ_p has a bounded density on $\mathbb{S}_p^\tau$ with respect to $(d-1)$-dimensional Lebesgue measure. First, we show the inclusions (3.22).

- For $x \in A_{a,\beta,\tau;-}(\varepsilon)$ and for $y \in \mathbb{S}_p$ such that $\|y - x\|_\infty \le \varepsilon$, we have $y \in A_{a,\beta,\tau}$: indeed,

  $$y_1 \wedge \cdots \wedge y_d \ge x_1 \wedge \cdots \wedge x_d - \varepsilon > \tau + \varepsilon - \varepsilon = \tau$$

  as well as (by equivalence of norms (3.25))

  $$\begin{aligned}\langle a, y\rangle = \langle a, x\rangle + \langle a, y - x\rangle &\le \beta - \sqrt{d}\varepsilon + \|a\|_2\|y - x\|_2 \\ &\le \beta - \sqrt{d}\varepsilon + \sqrt{d}\|y - x\|_\infty \\ &\le \beta.\end{aligned}$$

  This is the first inclusion in (3.22).

- For $x \in A_{a,\beta,\tau}$ and for $y \in \mathbb{S}_p$ such that $\|y - x\|_\infty \le \varepsilon$, we have

  $$y_1 \wedge \cdots \wedge y_d \ge x_1 \wedge \cdots \wedge x_d - \varepsilon > \tau - \varepsilon$$

  so that $y \in \mathbb{S}_p^{\tau-\varepsilon}$ (if $\varepsilon > \tau$ this trivial since $\mathbb{S}_p^v = \mathbb{S}_p$ for $v < 0$) together with

  $$\begin{aligned}\langle a, y\rangle = \langle a, x\rangle + \langle a, y - x\rangle &\le \beta + \|a\|_2\|y - x\|_2 \\ &\le \beta + \sqrt{d}\|y - x\|_\infty \\ &\le \beta + \sqrt{d}\varepsilon.\end{aligned}$$

  This implies the second inclusion in (3.22).

The difference between the inner and outer hulls is

$$\begin{aligned}&A_{a,\beta,\tau;+}(\varepsilon) \setminus A_{a,\beta,\tau;-}(\varepsilon) \\ &\qquad \subseteq \left\{x \in \mathbb{S}_p : \ \beta - \sqrt{d}\varepsilon < \langle a, x\rangle \le \beta + \sqrt{d}\varepsilon\right\} \cup (\mathbb{S}_p^{\tau-\varepsilon} \setminus \mathbb{S}_p^{\tau+\varepsilon}).\end{aligned}$$

Since $\|a\|_2 = 1$, the $(d-1)$-dimensional Lebesgue measure of the set on the right-hand side is bounded by a constant multiple of ε.

Next we show that Assumption 3.3.2 is satisfied. The VC-dimension is preserved under bijections. We identify $\mathbb{E} = [0,\infty)^d \setminus \{(0,\dots,0)\}$ with the product set $(0,\infty) \times \mathbb{S}_p$ via $x \mapsto (\|x\|_p, x/\|x\|_p)$. In the latter space, we need to establish the finiteness of the VC-dimensions of the collections $\{\Upsilon_A^-(r,h) : A \in \mathcal{A}\}$ and $\{\Upsilon_A^+(r,h) : A \in \mathcal{A}\}$ for fixed $r, h > 0$, where

$$\Upsilon_A^\sigma(r,h) \stackrel{\text{def}}{=} \left\{(u,\theta) \in (\tfrac{1}{r},\infty) \times \mathbb{S}_p : \ \theta \in A_\sigma(hu)\right\}$$

for $\sigma \in \{-,+\}$ and $A \in \mathcal{A}$. For $A = A_{a,\beta,\tau}$ we have

$$\theta \in A_-(hu) \iff \left[\theta_1 \wedge \cdots \wedge \theta_d \geq \tau + hu \text{ and } \langle a,\theta\rangle \leq \beta - \sqrt{d}hu\right].$$

This corresponds to $d+1$ linear inequality constraints on (u,θ) as (a,β) ranges over $\mathbb{R}^d \times \mathbb{R}$. The VC dimension of $\{\Upsilon_A^-(r,h) : A \in \mathcal{A}\}$ as a collection of subsets of $(0,\infty) \times \mathbb{S}_p$ is therefore finite [122, Lemma 2.6.17 and Exercise 14 on page 152], and the same is then true for $\{\Gamma_A^-(r,h) : A \in \mathcal{A}\}$ as a collection of subsets of $\mathbb{E}$. The argument for the outer hulls is similar. □

Proof of Example 3.3.2. We first consider the collection of intersections. In Assumption 3.3.1, item (i) follows by considering the countable collection of intersections $\mathcal{A}_{1,0} \sqcap \mathcal{A}_{2,0}$, where $\mathcal{A}_{j,0}$ is the countable subset of $\mathcal{A}_j$ for $j \in \{1,2\}$. Item (ii) is trivially satisfied. Regarding (iii): given $A_j \in \mathcal{A}_j$ with inner and outer hulls $A_{j,\sigma}(\varepsilon)$, the inner and outer hulls of $A_1 \cap A_2$ can chosen to be $A_{1,-}(\varepsilon) \cap A_{2,-}(\varepsilon)$ and $A_{1,+}(\varepsilon) \cap A_{2,+}(\varepsilon)$. The two inclusions (3.22) are straightforward to verify. The measure of the set difference between the inner and outer hulls can be controlled via

$$\big(A_{1,+}(\varepsilon) \cap A_{2,+}(\varepsilon)\big) \setminus \big(A_{1,-}(\varepsilon) \cap A_{2,-}(\varepsilon)\big) \subseteq \big(A_{1,+}(\varepsilon) \setminus A_{1,-}(\varepsilon)\big) \cup \big(A_{2,+}(\varepsilon) \setminus A_{2,-}(\varepsilon)\big).$$

If c_1 and c_2 denote the constants on the right-hand of (3.21) for $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, then $c_1 + c_2$ is a valid constant for $\mathcal{A}_1 \sqcap \mathcal{A}_2$.

To show that the collections of framing sets derived from the inner and outer hulls thus constructed have a finite VC-dimension (Assumption 3.3.2), it suffices to observe that, by definition,

$$\Gamma_{A_1 \cap A_2}^\sigma(r,h) = \Gamma_{A_1}^\sigma(r,h) \cap \Gamma_{A_2}^\sigma(r,h)$$

and thus

$$\{\Gamma_A^\sigma(r,h) : A \in \mathcal{A}_1 \sqcap \mathcal{A}_2\} = \{\Gamma_{A_1}^\sigma(r,h) : A_1 \in \mathcal{A}_1\} \sqcap \{\Gamma_{A_2}^\sigma(r,h) : A_2 \in \mathcal{A}_2\}.$$

The collection on the left-hand side has a finite VC-dimension since, by assumption, each of the two collections on the right-hand side has a finite VC-dimension; see for instance [122, Lemma 2.6.17].

For the collection of unions, the verification of Assumptions 3.3.1 and 3.3.2 is completely similar. The inner and outer hulls of a union $A_1 \cup A_2$ are now defined as the unions of the inner and outer hulls of A_1 and A_2. □

3.F Proofs of classification application

In the proofs, we find it sometimes convenient to take explicit note of the map $T : \mathbb{R}^d \to [0,\infty)^d$ used to transform the marginal distributions of the predictor variable; typically, $T = v$ or $T = \hat{v}$. In line with the theoretical and empirical classification risks $L_t(g)$ and $\widehat{L}^\tau(g)$ in (3.35) and (3.37) in the paper, define

$$L_t(g,T) \stackrel{\text{def}}{=} t\,\mathsf{P}\left(g(T(X)) \neq Y,\ \|T(X)\|_p \geq t\right),$$
$$\widehat{L}^\tau(g,T) \stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\{g(T(X_i)) \neq Y_i,\ \theta_p(T(X_i)) \in \mathbb{S}_p^\tau,\ \|T(X_i)\|_p \geq n/k\}.$$

Then $L_t(g)$ in (3.35) corresponds to $L_t(g,v)$ here and $\widehat{L}^\tau(g)$ in (3.37) to $\widehat{L}^\tau(g,\hat{v})$.

Proof of Lemma 3.4.1. We prove the first statement only. 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 and right hand sides of the first two identities.

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

$$L_{t,\pm}^{>\tau}(g,T) \stackrel{\text{def}}{=} t\mathsf{P}\left(g(T(X)) \neq Y,\ Y = \pm 1,\ \theta_p(T(X)) \in \mathbb{S}_p^\tau,\ \|T(X)\|_p \geq t\right).$$

Consider the cones generated by regions $\mathbb{S}_p^\pm(g) \stackrel{\text{def}}{=} \{x \in \mathbb{S}_p : g(x) = \pm 1\}$, that is, $R_p^\pm(g) \stackrel{\text{def}}{=} \{tx : x \in \mathbb{S}_p^\pm(g),\ t \geq 1\}$. Equipped with this notation we may write

$$L_{t,+}^{>\tau}(g,T) = t\mathsf{P}\left(T(X) \in R_p^-(g),\ Y = +1,\ \theta_p(T(X)) \in \mathbb{S}_p^\tau,\ \|T(X)\|_p \geq t\right),$$
$$L_{t,-}^{>\tau}(g,T) = t\mathsf{P}\left(T(X) \in R_p^+(g),\ Y = -1,\ \theta_p(T(X)) \in \mathbb{S}_p^\tau,\ \|T(X)\|_p \geq t\right).$$

Setting the standardization function T to v yields $T(X) = V$ and thus

$$\begin{aligned} L_{t,+}^{>\tau}(g,v) &= t\mathsf{P}\left(V \in R_p^-(g),\ \theta_p(V) \in \mathbb{S}_p^\tau,\ \|V\|_p \geq t,\ Y = +1\right) \\ &= \varrho t\mathsf{P}\left(\theta_p(V) \in \mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau,\ \|V\|_p \geq t \mid Y = +1\right) \\ &\to \varrho\Phi_p^+(\mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau), \qquad t \to \infty, \end{aligned}$$

where the last convergence occurs because of Assumption 3.4.1 and the fact that Φ_p^+ is dominated by Φ_p, so that $\mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau$ is a Φ_p^+-continuity set. Proceeding similarly with $L_{t,-}^{>\tau}(g,v)$, the result is obtained. □

The next result parallels Lemma 3.4.1 by relating the empirical risk with the empirical angular measure of the positive and negative classes.

Consider the type-I and type-II empirical errors,

$$\widehat{L}^{\tau}_{\pm}(g,T) \stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{n} \mathbb{I}\Big\{g(T(X_i)) \neq Y_i,\ Y_i = \pm 1,\ \theta_p(T(X_i)) \in \mathbb{S}^{\tau}_p,\ \|T(X_i)\|_p \geq n/k\Big\}.$$

Setting $T = \hat{v}$ as in (3.12) yields $T(X_i) = \widehat{V}_i$ and thus

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

Recall from (3.40) the empirical angular measures $\widehat{\Phi}^{\sigma}_p$ of the positive and negative instances. A similar treatment of $\widehat{L}^{\tau}_{-}(g,\hat{v})$ yields the following result.

Lemma 3.F.1. *In the case where T is the rank transformation $\hat{v}$,*

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

Proof of Theorem 3.4.1. Recall from the proof of Lemma 3.4.1 the decomposition of $L^{>\tau}_{\infty}$ into type-I and type-II errors, $L^{>\tau}_{\infty} = L^{>\tau}_{\infty,+} + L^{>\tau}_{\infty,-}$ with $L^{>\tau}_{\infty,+}(g) = \varrho\Phi^{+}_p(\mathbb{S}^{-}_p(g) \cap \mathbb{S}^{\tau}_p)$ and $L^{>\tau}_{\infty,-}(g) = (1-\varrho)\Phi^{-}_p(\mathbb{S}^{+}_p(g) \cap \mathbb{S}^{\tau}_p)$. Recall also the identities for their empirical counterparts $\widehat{L}^{\tau}_{\pm}$ in Lemma 3.F.1. 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,

$$\begin{aligned}\left|\widehat{L}^{\tau}(g,\hat{v}) - L^{>\tau}_{\infty}(g)\right| &= \left|\widehat{L}^{\tau}_{+}(g,\hat{v}) - L^{>\tau}_{\infty,+}(g) + \widehat{L}^{\tau}_{-}(g,\hat{v}) - L^{>\tau}_{\infty,-}(g)\right| \\ &\leq \left|\widehat{L}^{\tau}_{+}(g,\hat{v}) - L^{>\tau}_{\infty,+}(g)\right| + \left|\widehat{L}^{\tau}_{-}(g,\hat{v}) - L^{>\tau}_{\infty,-}(g)\right|. \qquad (3.61)\end{aligned}$$

Let us focus on the first term of the sum. The treatment of the second term is entirely similar and this accounts for the factor two on the right-hand side of the bound in the theorem. From Lemma 3.F.1 and the definition of $L^{>\tau}_{\infty,+}(g)$ recalled above, we have

$$|\widehat{L}^{\tau}_{+}(g,\hat{v}) - L^{>\tau}_{\infty,+}(g)| = \left|\frac{k_+}{k}\widehat{\Phi}^{+}_p(\mathbb{S}^{-}_p(g) \cap \mathbb{S}^{\tau}_p) - \varrho\Phi^{+}_p(\mathbb{S}^{-}_p(g) \cap \mathbb{S}^{\tau}_p)\right|,$$

which suggests extending the concentration results concerning the empirical measure $\widehat{\Phi}_p$ to its conditional version $\widehat{\Phi}^{+}_p$. To do so we shall work on the product space $\mathbb{R}^d \times \{-1,1\}$. We first introduce some notation. 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, i.e., $Q_n \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$. As in Section 3.3 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

$$\widehat{Q}_n = \frac{1}{n} \sum_{i=1}^n \delta_{(\hat{V}_i, Y_i)} = Q_n \circ (\hat{v}, \mathrm{id})^{-1}$$

where id is the identity function mapping (on $\{-1, 1\}$).

Define the limit measure on $\mathbb{E} \times \{-1, 1\}$: for $\sigma \in \{-, +\}$ and Borel sets $B \subseteq \mathbb{E}$ bounded away from the origin with $\nu_\sigma(\partial B) = 0$, put

$$\nu(B \times \{\sigma 1\}) \stackrel{\text{def}}{=} \lim_{t \to \infty} t \mathrm{P}\left(V \in tB, Y = \sigma 1\right) = \mathrm{P}\left(Y = \sigma 1\right)\ \nu_\sigma(B),$$

a limit which exists by the conditional regular variation assumption (3.31). Notice that ν is homogeneous of order -1 w.r.t. the first component. With this notation and the one borrowed from the proof of Theorem 3.3.1 (see (3.52) for the definition of $\widehat{\Gamma}_A$), we have, for $A \subseteq \mathbb{S}_p$,

$$\begin{aligned} \tfrac{k_+}{k} \widehat{\Phi}_p^+(A) &= \tfrac{n}{k} \widehat{Q}_n \left(\tfrac{n}{k} C_A \times \{+1\}\right) \\ &= \tfrac{n}{k} Q_n \left(\hat{v}^{-1}(\tfrac{n}{k} C_A) \times \{+1\}\right) \\ &= \tfrac{n}{k} Q_n(\widehat{\Gamma}_A \times \{+1\}) \end{aligned}$$

and

$$\varrho \Phi_p^+(A) = \nu(C_A \times \{+1\}).$$

It is shown in the proof of Theorem 3.3.1 that under the assumptions of the statement, there exists an event $\mathcal{E}_1$ of probability at least $1 - d\delta/(d+1)$ on which $\frac{n}{k}\Gamma_A^- \subseteq \widehat{\Gamma}_A \subseteq \frac{n}{k}\Gamma_A^+$, see (3.45). In addition recall that

$$\forall A \in \mathcal{A}, \qquad \Gamma_A^- \subseteq C_A \subseteq \Gamma_A^+.$$

Thus on the event $\mathcal{E}_1$, we can decompose the error as in the proof of Theorem 3.3.1 into

$$\begin{aligned} \tfrac{k_+}{k} \widehat{\Phi}_p^+(A) - \varrho \Phi_p^+(A) &= \tfrac{n}{k} Q_n(\widehat{\Gamma}_A \times \{+1\}) - \nu(C_A \times \{+1\}) \\ &\le \tfrac{n}{k} Q_n(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - \nu(\Gamma_A^- \times \{+1\}) \\ &\le \tfrac{n}{k} \left| Q_n(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - Q(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) \right| \\ &\quad + \left| \tfrac{n}{k} Q(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - \nu(\Gamma_A^+ \times \{+1\}) \right| \\ &\quad + \nu\left((\Gamma_A^+ \setminus \Gamma_A^-) \times \{+1\}\right). \end{aligned}$$

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

$$\begin{aligned}
&\left|\tfrac{k_+}{k}\widehat{\Phi}_p^+(A) - \varrho\Phi_p^+(A)\right| \\
&\le \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q_n(\tfrac{n}{k}B\times\{+1\}) - Q(\tfrac{n}{k}B\times\{+1\})\right| \qquad \text{(stochastic error)} \\
&\quad + \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q(\tfrac{n}{k}B\times\{+1\}) - \nu(B\times\{+1\})\right| \qquad \text{(bias term)} \\
&\quad + \nu\left((\Gamma_A^+ \setminus \Gamma_A^-)\times\{+1\}\right) \qquad \text{(framing gap).}
\end{aligned}$$

We treat the three terms separately, following closely the proof of Theorem 3.3.1.

Bias term. Taking the supremum over $A \in \mathcal{A}$ immediately yields the bias term in the statement of Theorem 3.4.1.

Stochastic error. Since $Q(B\times\{+1\}) \le \mathrm{P}(X \in B)$, the Q-probability of sets in the class $\mathcal{F}' \stackrel{\text{def}}{=} \mathcal{F}\times\{+1\}$ (with $\mathcal{F}$ defined in (3.46)) is less than $d^{1+1/p}(1+\Delta_2)\frac{k}{n}$, see (3.47). The class $\mathcal{F}'$ has the same VC-dimension $V_{\mathcal{F}}$ as $\mathcal{F}$. Thus on an event $\mathcal{E}_2^+$ of probability $1-\delta/(2(d+1))$ we have, by Theorem 3.A.1,

$$\begin{aligned}
&\sup_{A\in\mathcal{A}} \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q_n(\tfrac{n}{k}B\times\{+1\}) - Q(\tfrac{n}{k}B\times\{+1\})\right| \\
&\le \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}} + 2\sqrt{\log(2(d+1)/\delta)}\right) + \frac{2}{3k}\log(2(d+1)/\delta),
\end{aligned}$$

which is the term labelled "error" in the statement.

Framing gap. As in the proof of Theorem 3.3.1, the framing gap in the product space satisfies

$$\begin{aligned}
&\nu\big((\Gamma_A^+ \setminus \Gamma_A^-)\times\{+1\}\big) \\
&\le \nu\left(\left\{x\in[0,\infty)^d : (1+\Delta_2)^{-1} \le \|x\|_p < (1-\Delta_2)^{-1}\right\}\times\{+1\}\right) \\
&+ \nu\left(\left\{x\in[0,\infty)^d : \|x\|_p \ge 1,\ \theta_p(x) \in A_+(3\Delta_1\|x\|_p) \setminus A_-(3\Delta_1\|x\|_p)\right\}\times\{+1\}\right).
\end{aligned} \tag{3.62}$$

The first term on the right-hand side of (3.62) is equal to

$$2\Delta_2\nu(\{x\in[0,\infty)^d : \|x\|_p \ge 1\}\times\{+1\}) = 2\Delta_2\varrho\Phi_p^+(\mathbb{S}_p) \le 2\Delta_2\Phi_p(\mathbb{S}_p),$$

where the latter inequality comes from the decomposition $\Phi_p = \varrho\Phi_p^+ + (1-\varrho)\Phi_p^-$. The second term on the right-hand side in (3.62) can be bounded using the polar decomposition of ν_+ (and thus ν), yielding the bound

$$\int_1^\infty \varrho\Phi_p^+\big(A_+(3\Delta_1 r) \setminus A_-(3\Delta_1 r)\big)\frac{\mathrm{d}r}{r^2} \le \int_1^\infty \Phi_p\big(A_+(3\Delta_1 r) \setminus A_-(3\Delta_1 r)\big)\frac{\mathrm{d}r}{r^2} \le 3c\Delta_1(1 + \log\Phi_p(\mathbb{S}_p) - \log(3c\Delta_1)).$$

by the proof of Theorem 3.3.1. We thus obtain the same gap term as in Theorem 3.3.1.

So far we have only treated one of the two terms of the error decomposition (3.61). The second one is treated in the same way and the associated upper bound is identical, which yields the factor two in the statement of Theorem 3.4.1. The decomposition of $|\frac{k_-}{k}\widehat{\Phi}_p^-(A) - (1-\varrho)\Phi_p^-(A)|$ into a stochastic error, a bias term and framing gap holds true on the same event $\mathcal{E}_1$. The bound on the stochastic error is valid on an event $\mathcal{E}_2^-$ of probability at least $1 - \delta/(2(d+1))$. Thus the upper bound in the statement of Theorem 3.4.1 holds true on the intersection $\mathcal{E}_1 \cap \mathcal{E}_2^+ \cap \mathcal{E}_2^-$, which has probability at least $1 - \delta$. □

3.G Proofs of simulations experiments

The following lemma describes convenient properties relative to the simulation setting described in Section 3.5. Recall $\mathbb{S}_p^\tau = \{x \in \mathbb{S}_p : \min(x) > \tau\}$ for $\tau \in (0,1)$ as well as the cones C_A in (3.7).

Lemma 3.G.1. *Let R be a unit-Pareto distributed random variable and let Θ be a random vector independent from R, with support in $\mathbb{S}_1 = \{x \in [0,1]^d : x_1 + \cdots + x_d = 1\}$ and satisfying $\mathsf{E}(\Theta_j) = 1/d$ for $j \in \{1,\dots,d\}$. Let $X \overset{\text{def}}{=} R\Theta$ and write $V = v(X)$ with $v(x) = (1/(1-F_1(x_1)),\dots,1/(1-F_d(x_d)))$ for $x \in \mathbb{R}^d$, where $F_j(x_j) = \mathsf{P}(X_j \le x_j)$.*

(i) We have $v(x) = dx$ for all $x \in [1,\infty)^d$. Conversely, if $x \in \mathbb{R}^d$ satisfies $v(x) \in (d,\infty)^d$, then $x \in (1,\infty)^d$ and thus $v(x) = dx$.

(ii) The distribution of V is multivariate regularly varying and its angular measure Φ_p with respect to the L_p-norm, for $p \in [1,\infty]$, is given by

$$\Phi_p(A) = d\mathsf{P}(X \in C_A)$$

for Borel sets $A \subseteq \mathbb{S}_p$. Moreover, if $A \subseteq \mathbb{S}_p^\tau$ for some $\tau \in (0,1)$ and if $t > d/\tau$, then also

$$\Phi_p(A) = t\mathsf{P}(t^{-1}V \in C_A).$$

Proof of Lemma 3.G.1. We prove the two items (i) and (ii).

Proof of (i). For any $j \in \{1, \dots, d\}$ and any $x_j > 0$,

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

Since the support of Θ is contained in the unit simplex, we have $\Theta_j/x_j \le 1$ almost surely whenever $x_j \ge 1$. In addition $\mathsf{E}(\Theta_j) = 1/d$ by assumption. It follows that

$$1/(1 - F_j(x_j)) = \begin{cases} dx_j & \text{if } x_j \ge 1, \\ 1/\mathsf{E}[\min(1, \Theta_j/x_j)] & \text{if } 0 < x_j < 1. \end{cases}$$

This proves that $v(x) = dx \in [d, \infty)^d$ for $x \in [1, \infty)^d$. The converse statement follows from $1 - F_j(1) = 1/d$ and monotonicity.

Proof of (ii). Let $x \in (0, \infty)^d$ and let $t > 0$ be sufficiently large so that $tx_j \ge d$ for all $j \in \{1, \dots, d\}$. Then

$$\begin{aligned} 1 - \mathsf{P}\left(V_1 \le tx_1, \dots, V_d \le tx_d\right) &= \mathsf{P}\left(\exists j = 1, \dots, d : dR\Theta_j > tx_j\right) \\ &= \mathsf{P}\left[R > (t/d) \min_{j=1,\dots,d} (x_j/\Theta_j)\right] \\ &= (d/t)\mathsf{E}\left[\max_{j=1,\dots,d} (\Theta_j/x_j)\right]. \end{aligned}$$

In the last step, we conditioned on Θ and we used the fact that $(t/d)(x_j/\Theta_j) \ge 1$ a.s., by the assumptions on t and Θ. Recall $\mathbb{E} = [0, \infty)^d \setminus \{(0, \dots, 0)\}$. It follows that for all $x \in (0, \infty)^d$,

$$\forall t > \frac{d}{\min_{j=1,\dots,d} x_j}, \qquad t\mathsf{P}\left(t^{-1}V \in \mathbb{E} \setminus [0, x]\right) = d\mathsf{E}\left[\max_{j=1,\dots,d} (\Theta_j/x_j)\right]. \quad (3.63)$$

For fixed $x \in (0, \infty)^d$, the limit as $t \to \infty$ trivially exists. This implies that V is multivariate regularly varying as in Section 3.2.1 in the paper; see for instance [107, Lemma 6.1]. More precisely, $t\mathsf{P}(t^{-1}V \in \cdot\,) \to \nu(\,\cdot\,)$ as $t \to \infty$ in the space $\mathcal{M}_0$, where ν is determined by the right-hand side of the previous display via the values of $\nu(\mathbb{E} \setminus [0, x])$ for $x \in (0, \infty)^d$.

The angular measure Φ_p determines the exponent measure ν uniquely via the relation (3.8) in the paper. It follows by a standard calculation involving Fubini's theorem that

$$\nu(\mathbb{E} \setminus [0, x]) = \int_{\mathbb{S}_p} \max_{j=1,\dots,d} (\theta_j/x_j)\, \mathrm{d}\Phi_p(\theta).$$

Define a Borel measure Φ'_p on $\mathbb{S}_p$ by

$$\begin{aligned}\Phi'_p(A) &\stackrel{\text{def}}{=} d\mathsf{P}(X \in C_A)\\ &= d\mathsf{P}(R\|\Theta\|_p > 1,\ \Theta/\|\Theta\|_p \in A)\\ &= d\mathsf{E}[\|\Theta\|_p \mathbb{I}\{\Theta/\|\Theta\|_p \in A\}]\end{aligned}$$

for Borel sets $A \subseteq \mathbb{S}_p$. Recall that Φ_p is considered with respect to a general L_p-norm but Θ is a random vector in the unit simplex, so $0 < \|\Theta\|_p \leq 1$ almost surely since $p \geq 1$. The last equality follows by conditioning on Θ, the independence of R and Θ, and the assumption that the distribution of R is unit-Pareto. For Borel measurable, nonnegative functions g on $\mathbb{S}_p$, we get

$$\int_{\mathbb{S}_p} g(\theta)\, \mathrm{d}\Phi'_p(\theta) = d\mathsf{E}[\|\Theta\|_p\, g(\Theta/\|\Theta\|_p)].$$

Applying the latter identity to the function g defined by $g(\theta) = \max_{j=1,\dots,d}(\theta_j/x_j)$ for some fixed $x \in (0,\infty)^d$ yields

$$\int_{\mathbb{S}_p} \max_{j=1,\dots,d}(\theta_j/x_j)\, \mathrm{d}\Phi'_p(\theta) = d\mathsf{E}\left[\max_{j=1,\dots,d}(\Theta_j/x_j)\right] = \nu(\mathbb{E} \setminus [0,x]).$$

We conclude that Φ_p is equal to Φ'_p, as required.

Finally, let $0 < \tau < 1$ and let $A \subseteq \mathbb{S}_p^\tau$ be a Borel set. The cone $C_A = \{x \in \mathbb{E} : \|x\|_p \geq 1, x/\|x\|_p \in A\}$ is a subset of $(\tau,\infty)^d$. The equality (3.63) together with the inclusion–exclusion formula and the fact that rectangles form a measure-determining class imply that the measures $t\mathsf{P}(t^{-1}V \in \cdot\,)$ and ν correspond on $(\tau,\infty)^d$. It follows that

$$t\mathsf{P}(t^{-1}V \in C_A) = \nu(C_A) = \Phi_p(A). \qquad \square$$

Bivariate Asymptotic Theory and Testing for the Angular Measure

4

This chapter is based on
Stéphane Lhaut and Johan Segers
An asymptotic expansion of the empirical angular measure for bivariate extremal dependence
In Matteo Barigozzi, Siegfried Hörmann, and Davy Paindaveine, editors, Recent Advances in Econometrics and Statistics. Springer Cham, November 2024
and
Stéphane Lhaut and Johan Segers
Testing parametric models for the angular measure for bivariate extremes
arXiv preprint arXiv:2411.12673 (submitted to Extremes).

Abstract

In this chapter, we consider the angular measure in dimension $d = 2$. A first contribution consists of a functional asymptotic expansion for the empirical angular measure based on the theory of weak convergence in the space of bounded functions. From the expansion, not only can the known asymptotic distribution of the empirical angular measure be recovered, it also enables to find expansions and weak limits for other statistics based on the associated empirical process. As an application, we propose a procedure to test the goodness-of-fit of a parametric model to the extremal dependence structure of a bivariate random sample. The test statistic consists of a weighted L_1-Wasserstein distance between a nonparametric, rank-based estimator of the true angular measure obtained by maximizing a Euclidean likelihood on the one hand, and a parametric estimator of the angular measure on the other hand. The asymptotic distribution of the test statistic under the null hypothesis is derived and used to obtain critical values for the testing procedure via a parametric bootstrap. Consistency of the bootstrap algorithm is proved. A simulation study illustrates the finite-sample performance of the test for the logistic and Hüsler–Reiss models. We apply the method to test for the Hüsler–Reiss model in the context of river discharge data.

4.1 Introduction

In various domains, such as meteorology, finance or engineering, multivariate extreme events may induce important perturbations of the system of interest, so that the occurrence probabilities of such events are of main concern. Dangerous combinations of wave heights and still water levels could lead to the collapse of a dike at a certain point at the coast [33]. The financial portfolio selection problem involves maximizing the return on investment subject to risk constraints; if those constraints are expressed in terms of quantities such as the Value-at-Risk or the Expected Shortfall, tail quantiles of the portfolio need to be estimated accurately [108]. Floating systems in an offshore environment are submitted to natural forces such as wind, waves and currents, and may be sensible to certain combinations of large values of these forces, the probabilities of which need to be assessed [95].

Extreme value theory [34, 106] provides a convenient and solid mathematical framework to answer those questions. In the multivariate case, the classical assumption is that the random vector of interest $X = (X_1, \ldots, X_d)$, with values in $\mathbb{R}^d$, belongs to the *maximum domain of attraction* of a multivariate extreme value distribution. This hypothesis comprises two parts:

1. the marginal distributions of X belong to the maximal domains of attraction of some univariate extreme value distributions;

2. after marginal transformation, through the probability integral transform or a variation thereof, the joint distribution of the transformed vector belongs to the maximal domain of attraction of a multivariate extreme value distribution with pre-specified margins.

Under the side assumption that the marginal cumulative distribution functions of X are continuous, which we make in this paper, point 2 above only involves the copula of X in view of Sklar's theorem. Point 2 can be imposed on its own, that is, independently of the assumptions on the margins in point 1, and this is what we will do in this paper.

Angular measure. The multivariate extreme value distribution to which the distribution in point 2 is attracted is fully determined by an *angular measure*, denoted here by Φ, which is a measure on the intersection between the positive orthant $[0, \infty)^d$ with the unit sphere with respect to a given norm. This measure, originally called spectral measure in [36], describes the first-order dependence structure of joint extremes of X and is rooted in the theory of multivariate regular variation [106, Chapter 5].

Inference on the angular measure plays a major role in various problems related to extremes. Without being exhaustive, we mention the following

examples: estimating the probability of a failure set [33, 37], estimating extreme quantile regions [40], anomaly detection [59], approximating conditional densities of an element of a random vector given that the others are large [28], and binary classification in extreme regions [76].

For a given dimension d, the collection of all angular measures does not form a parametric family. Hence, a natural way to reduce the complexity of the inference problem is to *assume* a parametric form for Φ. Many models have been proposed, sometimes expressed in terms of other objects, such as the stable tail dependence function: the original bivariate logistic model [64], the asymmetric bivariate logistic model [119], the tilted Dirichlet model [26], the pairwise beta model [27], the Hüsler–Reiss model [70, 52], among others. For conclusions within the proposed models to be reliable, the parametric assumption has to be tested itself.

Empirical angular measure. Due to the marginal transformation explained in point 2. above, the empirical estimation of the angular measure leads to a rank based procedure [42]. Statistical guarantees on the resulting estimator are therefore much more complicated to obtain. Finite sample results have derived only recently in [25], see also Chapter 3 of this work. Asymptotically, the functional asymptotic distribution has only been obtained for $d = 2$ and every L_p-norm for $p \in [1, \infty]$ in [47].

In this paper, we propose a slight extension of the asymptotic result in the case $d = 2$ by providing an asymptotic functional expansion of the process associated with the estimator. This result is very helpful in obtaining the asymptotic distribution of any statistic based on this process, among which, the test statistic associated with the goodness-of-fit procedure introduced below. Moreover, the tools used to obtain the expansion are not exactly the same as the ones used in the original proof in [47]. In particular, we avoided complex probabilistic constructions involving the Skorokhod's representation theorem on complex probability spaces but rather directly rely on the weak convergence theory exposed, for example, in [122].

Testing parametric models for extremal dependence. Goodness-of-fit tests for parametric models of the tail dependence structure have already been proposed in the literature. A test based on a weighted L_2-distance between the tail empirical copula and the estimated copula under the postulated model with estimation of the parameter performed by censored likelihood maximization and critical values computed on the basis of parametric bootstrap is studied in [35]. In a similar way, the limiting distribution of a test statistic consisting of a L_2-distance between the empirical stable tail dependence function and a parametric estimator with parameter estimation performed by the method of moments is derived in [44]; computation of critical values is not considered,

however. Another L_2-distance type of statistic based on the tail copula is considered in [10] with parameter under the null estimated via a minimum distance method and critical values computed on the basis of a multiplier bootstrap. A different approach is proposed in [22] where authors develop asymptotically distribution-free tests based on various distances between a semi-parametric estimator of the tail copula and the estimated tail copula under the postulated model. Critical values are computed by simulating directly from limit law of the test statistic, which is distribution-free, that is, does not depend on unknown quantities. The computation of the test statistic itself is more involved.

In this paper, we propose a new test based on the angular measure directly. Our nonparametric estimator is a variation on the empirical angular measure, on which certain moment constraints are enforced by means of a Euclidean empirical likelihood in the spirit of [32], an idea going back to [99]. The discrepancy between this nonparametric estimator and the one estimated within the parametric model is quantified by a weighted version of the L_1-*Wasserstein distance* coming from optimal transport [126, 127]. The Wasserstein distance has a long history in hypothesis testing in the non-extreme setting, especially in the one-dimensional situation where it admits an explicit representation, see, e.g., the survey paper [100] for a review of its main statistical applications. Working with bivariate extremes, the angular measure is defined on a one-dimensional space, so that we may use the explicit formula to compute our test statistic. We derive its limiting distribution under the null hypothesis and propose a consistent method to estimate the associated critical values.

Extension to higher dimensions is far from trivial as the limit distribution of the empirical angular measure is not known in higher dimension. However, in the application in Section 4.6, we apply the test to several pairs of variables at once, correcting the p-values in order to account for the multiple hypothesis problem.

Outline. In Section 4.2, we describe the underlying mathematical framework and introduce the maximum empirical Euclidean likelihood estimator of the angular measure. The asymptotic expansion of the empirical angular measure (and the one of the maximum empirical Euclidean likelihood estimator) is provided in Section 4.3, which also contains results on (weighted) univariate and multivariate tail empirical processes (Propositions 4.3.1 and 4.3.2). The goodness-of-fit test is formalized in Section 4.4: Theorem 4.4.2 states the asymptotic distribution of the test statistic under general conditions on the estimator of the unknown model parameters, and Theorem 4.4.3 ensures that the critical values can be estimated consistently from the asymptotic distribution of the test statistic at the estimated parameter values. The finite-sample performance of the novel procedure is evaluated in Section 4.5 for

the logistic and the Hüsler–Reiss models, with favorable comparisons to the methods proposed in [35] and [22]. Finally, the goodness-of-fit of the Hüsler–Reiss model is tested for pairwise extremes of river discharge data in the Danube network in Section 4.6. Section 4.7 concludes the paper. All proofs are relegated to the appendices.

4.2 Regular variation and the angular measure

4.2.1 Regular variation

Let F_1 and F_2 denote the marginal cumulative distribution functions of X_1 and X_2, respectively, assumed to be continuous. As in [42, 47], our working hypothesis is that the distribution of the standardized random vector

$$U = (U_1, U_2) \stackrel{\text{def}}{=} (1 - F_1(X_1), 1 - F_2(X_2))$$

is *regularly varying* in $\mathcal{M}_\infty(\mathbb{E}^{-1})$, where $\mathbb{E} \stackrel{\text{def}}{=} [0, \infty)^2$ and $\mathbb{E}^{-1} \stackrel{\text{def}}{=} \{(x_1^{-1}, x_2^{-1}) : (x_1, x_2) \in \mathbb{E}\}$, in the sense that there exists a measure $\Lambda \in \mathcal{M}_\infty(\mathbb{E}^{-1})$ such that

$$t^{-1}\, \mathrm{P}\,(U \in t\cdot) \xrightarrow{\mathcal{M}_\infty(\mathbb{E}^{-1})} \Lambda, \qquad \text{as } t \to 0. \tag{4.1}$$

The space $\mathcal{M}_\infty(\mathbb{E}^{-1})$ consists of measures on (the Borel sets of) $\mathbb{E}^{-1} \setminus \{(\infty, \infty)\}$ which are bounded for Borel set bounded away from the point (∞, ∞). This is equivalent to

$$\lim_{t \to 0} t^{-1}\, \mathrm{P}\,(U \in tB) = \Lambda(B), \tag{4.2}$$

for any Borel set $B \subset \mathbb{E}^{-1}$ bounded away from (∞, ∞) and with $\Lambda(\partial B) = 0$ [74].

Remark 4.2.1 (Link with the exponent measure). Usually, multivariate regular variation is expressed through the standardized vector

$$V = (V_1, V_2) \stackrel{\text{def}}{=} \left(\frac{1}{1 - F_1(X_1)}, \frac{1}{1 - F_2(X_2)} \right)$$

with Pareto margins instead of uniform ones, and the convergence takes place in $\mathcal{M}_0(\mathbb{E})$ instead. The limiting measure obtained through this standardization is often denoted by ν and is referred to as the *exponent measure*. Here, we work with Λ for convenience, but it is clear that both measures are linked by the one-to-one relation

$$\nu = \Lambda \circ \iota^{-1} \qquad \text{and} \qquad \Lambda = \nu \circ \iota, \tag{4.3}$$

where $\iota : (x_1, x_2) \in \mathbb{E} \mapsto \iota(x_1, x_2) \stackrel{\text{def}}{=} (x_1^{-1}, x_2^{-1}) \in \mathbb{E}^{-1}$ is the inverse map.

Below, we list some properties of the measure Λ which will be useful later:

- Λ has Lebesgue margins: for any $0 \leq u < \infty$,

$$\Lambda([0,u] \times [0,\infty]) = \Lambda([0,\infty] \times [0,u]) = u. \tag{4.4}$$

- Λ is homogeneous: for any $c > 0$ and any Borel set $B \subset \mathbb{E}$,

$$\Lambda(cB) = c\Lambda(B). \tag{4.5}$$

4.2.2 Angular measure

The angular measure is a finite measure derived from the exponent measure and with support contained in the intersection of the unit sphere and the positive orthant $[0,\infty)^2$. Even though the angular measure can be defined with respect to any norm, we will only consider L_p-norms, the ones most used in practice, which are defined, for $p \in [1,\infty]$ and $x = (x_1, x_2) \in \mathbb{R}^2$, by

$$\|x\|_p \stackrel{\text{def}}{=} \begin{cases} (|x_1|^p + |x_2|^p)^{1/p} & \text{if } 1 \leq p < \infty, \\ \max(x_1, x_2) & \text{if } p = \infty. \end{cases} \tag{4.6}$$

For $p \in [1,\infty]$, we will let Φ_p denote the *angular measure* associated to the measure Λ in (4.1), defined on any Borel set $A \subset [0, \pi/2]$ by

$$\Phi_p(A) \stackrel{\text{def}}{=} \Lambda\left(\{(x_1,x_2) \in \mathbb{E}^{-1} : \|(x_1^{-1}, x_2^{-1})\|_p \geq 1, \arctan(x_2/x_1) \in A\}\right). \tag{4.7}$$

The inverses appearing in the radial component come from the fact that Φ_p is usually defined in terms of ν in (4.3) rather than Λ. If, for $\theta \in [0, \pi/2]$, we introduce the sets

$$C_{p,\theta} \stackrel{\text{def}}{=} \begin{cases} ([0,\infty] \times \{0\}) \cup (\{\infty\} \times [0,1]) & \text{if } \theta = 0, \\ \left\{(x,y) : 0 \leq x \leq \infty, 0 \leq y \leq \min\{x \tan\theta, y_p(x)\}\right\} & \text{if } 0 < \theta < \pi/2, \\ \left\{(x,y) : 0 \leq x \leq \infty, 0 \leq y \leq y_p(x)\right\} & \text{if } \theta = \pi/2, \end{cases} \tag{4.8}$$

where

$$y_p(x) \stackrel{\text{def}}{=} \begin{cases} \infty & \text{if } x \in [0,1), \\ \left(1 + \frac{1}{x^p - 1}\right)^{1/p} & \text{if } x \in [1,\infty] \text{ and } p \in [1,\infty), \\ 1 & \text{if } x \in [1,\infty] \text{ and } p = \infty, \end{cases}$$

then it follows that

$$\Phi_p(\theta) \stackrel{\text{def}}{=} \Phi_p([0,\theta]) = \Lambda(C_{p,\theta}).$$

For $x \geq 1$, $y_p(x)$ is the smallest value of $y \geq 1$ solving the equation $\|(x^{-1}, y^{-1})\|_p = 1$, while $x \tan\theta < y_p(x)$ if and only if $x < x_p(\theta) \stackrel{\text{def}}{=} \|(1, \cot\theta)\|_p$. We illustrate and relate these quantities in Figure 4.1 in case $p = 1$ and $0 < \theta < \pi/2$.

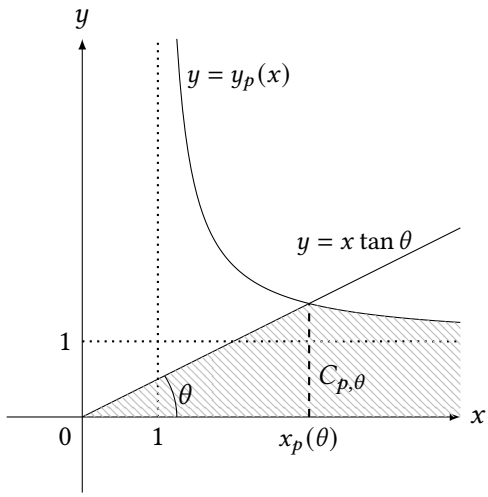


Figure 4.1: The set $C_{p,\theta}$ in (4.8) and related quantities for $p = 1$ and $0 < \theta < \pi/2$.

The Lebesgue margins of Λ in (4.4) imply constraints on Φ_p: necessarily, we must have

$$\int_0^{\pi/2} \frac{\sin\theta}{\|(\sin\theta,\cos\theta)\|_p} \mathrm{d}\Phi_p(\theta) = 1 = \int_0^{\pi/2} \frac{\cos\theta}{\|(\sin\theta,\cos\theta)\|_p} \mathrm{d}\Phi_p(\theta). \quad (4.9)$$

It will be convenient to work with the *angular probability measure* Q_p defined for Borel sets $A \subset [0,\pi/2]$ by

$$Q_p(A) \stackrel{\text{def}}{=} \frac{\Phi_p(A)}{\Phi_p([0,\pi/2])}, \quad (4.10)$$

for which the marginal constraints (4.9) become

$$\int_0^{\pi/2} \frac{\sin\theta}{\|(\sin\theta,\cos\theta)\|_p} \mathrm{d}Q_p(\theta) = \int_0^{\pi/2} \frac{\cos\theta}{\|(\sin\theta,\cos\theta)\|_p} \mathrm{d}Q_p(\theta), \quad (4.11)$$

the common value being equal to $m(Q_p) \stackrel{\text{def}}{=} \Phi_p([0,\pi/2])^{-1}$.

4.2.3 Maximum empirical Euclidean likelihood estimator

Given is an independent random sample $\{(X_{i1}, X_{i2}) : i = 1,\dots,n\}$ with the same distribution as (X_1, X_2). We would like to obtain a nonparametric estimator of Φ_p or, equivalently, Q_p.

One way to proceed is to observe that, by (4.2) and (4.7), for any Borel set $A \subset [0,\pi/2]$ whose topological boundary is a Φ_p-null set, we have, writing

$\overline{F}_j \stackrel{\text{def}}{=} 1 - F_j$,

$$\Phi_p(A) = \lim_{t\to 0} t^{-1}\mathrm{P}\left[\left\|\left(\frac{1}{\overline{F}_1(X_1)}, \frac{1}{\overline{F}_2(X_2)}\right)\right\|_p \geq t^{-1},\ \arctan\left\{\frac{\overline{F}_2(X_2)}{\overline{F}_1(X_1)}\right\} \in A\right].$$

Let $k = k(n)$ be such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Replacing t by k/n and the law of (X_1, X_2) by its empirical counterpart on the sample, and estimating the marginal distribution functions by $\widehat{F}_j(x_j) \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n \mathbb{I}\{X_{ij} < x_j\}$ with survival functions $\widehat{\overline{F}}_j \stackrel{\text{def}}{=} 1 - \widehat{F}_j$ for $j = 1, 2$, leads to the *empirical angular measure* with distribution function

$$\begin{aligned}\widehat{\Phi}_p(\theta) &\stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=1}^n \mathbb{I}\left\{\left\|\left(\frac{1}{\widehat{\overline{F}}_1(X_{i1})}, \frac{1}{\widehat{\overline{F}}_2(X_{i2})}\right)\right\|_p \geq \frac{n}{k},\ \arctan\left\{\frac{\widehat{\overline{F}}_2(X_{i2})}{\widehat{\overline{F}}_1(X_{i1})}\right\} \leq \theta\right\} \\ &= \frac{1}{k}\sum_{i=1}^n \mathbb{I}\Big\{(n+1-R_{i1})^{-p} + (n+1-R_{i2})^{-p} \geq k^{-p}, \\ &\qquad\qquad \frac{n+1-R_{i2}}{n+1-R_{i1}} \leq \tan\theta\Big\}, \qquad (4.12)\end{aligned}$$

for $\theta \in [0, \pi/2]$; here, R_{ij} denotes the rank of X_{ij} in the marginal sample $X_{1j}, \ldots, X_{nj}$.

The estimator $\widehat{\Phi}_p$ is not guaranteed to satisfy the marginal constraints (4.9) and is therefore not a *true* angular measure itself. One solution is to enforce the constraints by maximizing an empirical Euclidean likelihood as done in [32] for the case $p = 1$. In particular, they show that this procedure improves finite-sample performance of the estimator. Here, we extend their approach to the general case $p \in [1, \infty]$ to make use of the resulting estimator in our test.

Let

$$K \stackrel{\text{def}}{=} \sum_{i=1}^n \mathbb{I}\left\{\left\|\left(\frac{1}{\widehat{\overline{F}}_1(X_{i1})}, \frac{1}{\widehat{\overline{F}}_2(X_{i2})}\right)\right\|_p \geq \frac{n}{k}\right\}$$

be the (random) number of data points exceeding n/k on the Pareto scale and denote these points by $(X_{i_j,1}, X_{i_j,2})$ for $j \in \{1, \ldots, K\}$. The collection of angles associated to those points is

$$\widehat{\theta}_j \stackrel{\text{def}}{=} \arctan\left\{\frac{\widehat{\overline{F}}_2(X_{i_j,2})}{\widehat{\overline{F}}_1(X_{i_j,1})}\right\}, \qquad j \in \{1, \ldots, K\}. \qquad (4.13)$$

It is then easily seen from (4.12) that for any $\theta \in [0, \pi/2]$,

$$\widehat{Q}_p(\theta) \stackrel{\text{def}}{=} \frac{\widehat{\Phi}_p(\theta)}{\widehat{\Phi}_p(\pi/2)} = \sum_{j=1}^K K^{-1}\mathbb{I}\{\widehat{\theta}_j \leq \theta\}.$$

Following the footsteps of [32], we suggest to modify this estimator by introducing the *maximum (empirical) Euclidean likelihood estimator*

$$\widetilde{Q}_p(\theta) \stackrel{\text{def}}{=} \sum_{j=1}^{K} \widehat{p}_j \mathbb{I}\left\{\widehat{\theta}_j \le \theta\right\}, \qquad \theta \in [0, \pi/2], \tag{4.14}$$

with $\widehat{\theta}_j$ as in (4.13) and where $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_K)$ is the solution of the optimization problem

$$\min_{p \in \mathbb{R}^K} \frac{1}{2} \sum_{j=1}^{K} (Kp_j - 1)^2 \tag{4.15}$$

under the constraints

$$\begin{cases} \sum_{j=1}^{K} p_j & = 1, \\ \sum_{j=1}^{K} p_j f(\widehat{\theta}_j) & = 0, \end{cases}$$

where we wrote

$$f(\theta) \stackrel{\text{def}}{=} \frac{\sin\theta - \cos\theta}{\|(\sin\theta, \cos\theta)\|_p}, \qquad \theta \in [0, \pi/2]. \tag{4.16}$$

The constraints ensure that the estimated measure has total mass equal to one and that (4.11) is satisfied. The objective function in (4.15) is a second-order approximation to a negative empirical log-likelihood known as an empirical Euclidean log-likelihood [99], in view of the use of the Euclidean distance between the probability mass vector $(p_1, \ldots, p_K)$ and the uniform distribution $(1/K, \ldots, 1/K)$. The above optimization problem can be solved by the method of Lagrange multipliers, with solution

$$\widehat{p}_j \stackrel{\text{def}}{=} \frac{1}{K}\left\{1 - \frac{\overline{f_{\widehat{\theta}}}}{\sigma^2_{f_{\widehat{\theta}}}}\left(f(\widehat{\theta}_j) - \overline{f_{\widehat{\theta}}}\right)\right\}, \qquad j \in \{1, \ldots, K\},$$

where we used the notation

$$\overline{f_{\widehat{\theta}}} \stackrel{\text{def}}{=} \frac{1}{K}\sum_{j=1}^{K} f(\widehat{\theta}_j) \quad \text{and} \quad \sigma^2_{f_{\widehat{\theta}}} \stackrel{\text{def}}{=} \frac{1}{K}\sum_{j=1}^{K}\left(f(\widehat{\theta}_j) - \overline{f_{\widehat{\theta}}}\right)^2.$$

Note that the weights $\widehat{p}_j$ may be negative, but that the set on which this happens has probability tending to zero, see [32, page 5]. This follows from the facts that

$$\overline{f_{\widehat{\theta}}} \xrightarrow{\text{P}} \mu_{Q_p}(f) = 0 \quad \text{and} \quad \sigma^2_{f_{\widehat{\theta}}} \xrightarrow{\text{P}} \sigma^2_{Q_p}(f) > 0$$

as $n \to \infty$, where we used the notation in (4.30), together with $\left|f(\widehat{\theta}_j) - \overline{f_{\widehat{\theta}}}\right| \le 2$.

4.3 Asymptotic functional expansion of the empirical angular measure

The asymptotic distribution of the empirical process

$$\left\{\sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_p(\theta)\right) : \theta \in [0, \pi/2]\right\} \tag{4.17}$$

associated to the maximum Euclidean likelihood estimator derived in the previous section can be obtained on the basis of the one of the process $\sqrt{k}(\widehat{\Phi}_p - \Phi_p)$ associated to the empirical angular measure. The latter was analysed in [42] for the case $p = \infty$ and in [47] for general $p \in [1, \infty]$.

Here, as we are interested in using the estimator $\widetilde{Q}_p$ in a goodness-of-fit problem, our aim is to obtain a slightly stronger result than the asymptotic distribution and consider a functional asymptotic expansion instead. We start by stating two general weak convergence results which may be of independent interest and then state our main results of this section.

4.3.1 Weak convergence of tail empirical processes

Here and below, weak convergence of probability measures in metric spaces will be denoted by $\xrightarrow{\mathcal{L}}$. Let $\ell^\infty(\mathcal{F})$ denote the Banach space of bounded functions from a set $\mathcal{F}$ to R, equipped with the supremum norm.

Proposition 4.3.1 (Weak convergence of the tail empirical process)**.** *Let $\mathcal{A}$ be a suitably measurable [122, Definition 2.3.3] Vapnik–Chervonenkis (VC) class [123] of Borel sets on $[0, \infty)^d$, for some integer $d \geq 1$, such that there exists $M > 0$ with*

$$A \subseteq L_M \stackrel{\text{def}}{=} \{x \in [0, \infty)^d : \min(x_1, \ldots, x_d) \leq M\}, \qquad \forall A \in \mathcal{A}.$$

Let P be a Borel probability measure on $[0, 1]^d$ with uniform margins such that there exists a measure Λ in $\mathcal{M}_\infty\left((0, \infty]^d \setminus \{(\infty, \ldots, \infty)\}\right)$ satisfying

$$\lim_{t \to 0} \left| t^{-1} P(tB) - \Lambda(B) \right| = 0,$$

for any Borel set bounded away from infinity with $\Lambda(\partial B) = 0$, and

$$\lim_{t \to 0} \sup_{A, A' \in \mathcal{A}} \left| t^{-1} P(t(A \triangle A')) - \Lambda(A \triangle A') \right| = 0. \tag{4.18}$$

Let P_n be the empirical measure of an i.i.d. random sample of size $n \in \mathbb{N}$ from P. Let $k = k(n) \in \mathbb{N}$ be such that $k \to \infty$ and $k = o(n)$ as $n \to \infty$.

Then, in $\ell^\infty(\mathcal{A})$,

$$\left(W_{n,k}(A) \stackrel{\text{def}}{=} \sqrt{k}\left\{\tfrac{n}{k} P_n\left(\tfrac{k}{n} A\right) - \tfrac{n}{k} P\left(\tfrac{k}{n} A\right)\right\}\right)_{A \in \mathcal{A}} \xrightarrow{\mathcal{L}} (W_\Lambda(A))_{A \in \mathcal{A}}, \tag{4.19}$$

where W_Λ is a tight, centred Gaussian process on $\mathcal{A}$ with covariance function $\mathsf{E}[W_\Lambda(A)\, W_\Lambda(A')] = \Lambda(A \cap A')$.

Moreover, the process $W_{n,k}$ is asymptotically equicontinuous in probability with respect to the semi-metric $\rho(A, A') \overset{\text{def}}{=} \Lambda(A \triangle A')$, i.e., for all $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathsf{P}^* \left[\sup_{A, A' \in \mathcal{A}, \rho(A, A') < \delta} \left| W_{n,k}(A) - W_{n,k}(A') \right| > \epsilon \right] = 0. \tag{4.20}$$

Proposition 4.3.2 (Weighted weak convergence of tail marginal empirical processes)**.** *Let $U_1, \ldots, U_n$ denote uniformly distributed random variables on* $[0, 1]$. *Let*

$$\Gamma_n(u) \overset{\text{def}}{=} \begin{cases} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{U_i \leq u\} & \text{if } u \in [0, 1] \\ u & \text{if } u > 1 \end{cases}$$

denote the associated (modified) empirical cumulative distribution function and let

$$Q_n(u) \overset{\text{def}}{=} \begin{cases} \inf\{x \geq 0 : \Gamma_n(x) \geq u\} & \text{if } u > (2n)^{-1}, \\ 0 & \text{if } 0 \leq u \leq (2n)^{-1}. \end{cases}$$

be the (modified) left-continuous inverse of Γ_n. For $x \geq 0$, consider the processes

$$w_n(x) \overset{\text{def}}{=} \sqrt{k} \left(\tfrac{n}{k} \Gamma_n(\tfrac{k}{n} x) - x \right) \qquad \text{and} \qquad v_n(x) \overset{\text{def}}{=} \sqrt{k} \left(\tfrac{n}{k} Q_n(\tfrac{k}{n} x) - x \right).$$

Let $w : x \in (0, \infty) \mapsto w(x) \overset{\text{def}}{=} 1/(x^\delta \vee x^\eta)$ with $0 < \delta < 1/2 < \eta$ and set $w(0) \overset{\text{def}}{=} 0$. In $\ell^\infty([0, \infty))$, we have

$$\{w_n(x) w(x) : x \in [0, \infty)\} \xrightarrow{\mathcal{L}} \left\{ W_j(x) w(x) : x \in [0, \infty) \right\}, \tag{4.21}$$

as $n \to \infty$. Furthermore, for any $M > 1$,

$$\sup_{x \in [0, M]} \frac{|v_n(x) + w_n(x)|}{x^\delta} = o_{\mathsf{P}}(1). \tag{4.22}$$

The proofs of these results are to be found in Appendix 4.A.

4.3.2 Asymptotic expansions

Propositions 4.3.1 and 4.3.2 will form the technical basis to deliver the expansion of the process in (4.17). We need some technical assumptions to apply these results in our context.

Assumption 4.3.1. *The measure Λ is absolutely continuous with respect to the Lebesgue measure on $(0, \infty)^2$ with density λ which admits a continuous extension on $[0, \infty)^2 \setminus \{(0, 0)\}$. Furthermore, $\Lambda(\{\infty\} \times [0, 1]) = \Lambda([0, 1] \times \{\infty\}) = 0$.*

Assumption 4.3.1 implies that Φ_p is concentrated on $(0, \pi/2)$, thereby excluding asymptotic independence. Indeed, a calculation shows that $\Phi_p(\{0\}) = \Lambda(\{\infty\} \times [0,1])$ and $\Phi_p(\{\pi/2\}) = \Lambda([0,1] \times \{\infty\})$, both of which are zero by Assumption 4.3.1. It also implies that the map $\theta \in (0, \pi/2) \mapsto \Phi_p(\theta)$ has a continuous derivative φ_p on $(0, \pi/2)$. In particular, λ and φ_p are related through the following expression: for every $x, y > 0$,

$$\lambda(x,y) = \frac{xy}{x^2+y^2} \|(x,y)\|_p^{-1} \varphi_p(\arctan(y/x)). \tag{4.23}$$

The proof of this identity can be found in Appendix 4.B. Note that relation (4.23) implies

$$\lambda(ax, ay) = a^{-1}\lambda(x,y), \qquad a > 0,\ (x,y) \in [0,\infty)^2 \setminus \{(0,0)\}. \tag{4.24}$$

Evaluating expression (4.23) at $(x,y) = (\cos\theta, \sin\theta)$ for $\theta \in (0/\pi/2)$ leads to the useful formula

$$\varphi_p(\theta) = \frac{\|(\cos\theta, \sin\theta)\|_p}{\cos\theta \sin\theta} \lambda(\cos\theta, \sin\theta). \tag{4.25}$$

Assumption 4.3.2. *If c denotes the density of $U = (1 - F_1(X_1), 1 - F_2(X_2))$, the quantity*

$$\mathcal{D}_T(t) \overset{\text{def}}{=} \iint_{\mathcal{L}_T} |tc(tu_1, tu_2) - \lambda(u_1, u_2)| \, \mathrm{d}u_1 \mathrm{d}u_2, \quad 1 \le T < \infty,\ t > 0,$$

where $\mathcal{L}_T \overset{\text{def}}{=} \{(u_1,u_2) \in [0,T]^2 : u_1 \wedge u_2 \le 1\}$, satisfies $\mathcal{D}_{1/t}(t) = \mathrm{O}(t^{\alpha_1})$ and for any $p \in [1,\infty]$,

$$\Phi_p(t) = \mathrm{O}(t^{\alpha_2}) \qquad \text{and} \qquad \Phi_p\left(\tfrac{\pi}{2}\right) - \Phi_p\left(\tfrac{\pi}{2} - t\right) = \mathrm{O}(t^{\alpha_3})$$

for some $\alpha_1, \alpha_2, \alpha_3 > 0$ as $t \to 0$ (it is easy to show that they can be taken independently of p). Furthermore,

$$k = \mathrm{o}\left(n^{\frac{2\alpha}{2\alpha+1}}\right), \qquad \text{where } \alpha \overset{\text{def}}{=} \min\{\alpha_1, \alpha_2, \alpha_3\}.$$

Note in particular that under Assumption 4.3.2,

$$\left.\begin{aligned} &\lim_{n\to\infty} \sqrt{k}\, \mathcal{D}_{n/k}(k/n) \\ &\lim_{n\to\infty} \sqrt{k}\, \Phi_p\left(\tfrac{k}{n}\right) \\ &\lim_{n\to\infty} \sqrt{k}\, \left\{\Phi_p\left(\tfrac{\pi}{2}\right) - \Phi_p\left(\tfrac{\pi}{2} - \tfrac{k}{n}\right)\right\} \end{aligned}\right\} = 0.$$

Assumption 4.3.2 will be key to control the difference between $\Phi_p(\theta) = \Lambda(C_{p,\theta})$ and its preasymptotic version $sP(s^{-1}C_{p,\theta})$ for $s > 0$ large. In particular, note that the sets of interest $C_{p,\theta}$ are not included in $\mathcal{L}_T$ even for large T. However, thanks to some homogeneity property of λ in (4.24), we will be able to cover $C_{p,\theta}$ by a set on which the bias vanishes asymptotically thanks to the first part of the assumption and a tail set whose measure will be small by the second part of the assumption. Details are provided in the proof of our asymptotic expansion.

We are now ready to formulate Theorem 2 in [84], and, consequently, Theorem 3.1 in [47].

Theorem 4.3.1. *Let P be the law of $U = (1 - F_1(X_1), 1 - F_2(X_2))$ and let P_n denote the empirical measure of an independent sample $U_1, \ldots, U_n$ from the law of U. For $j \in \{1, 2\}$ and $u \in [0, 1]$, let the associated marginal empirical cumulative distribution functions be denoted by $\Gamma_{jn}(u) \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \mathbb{I}\{U_{ij} \le u\}$. We set $\Gamma_{jn}(u) \overset{\text{def}}{=} u$ for $u > 1$. Let $Q_{jn} \overset{\text{def}}{=} \inf\{x \ge 0 : \Gamma_{jn}(x) \ge u\}$ denote the quantile function associated to Γ_{jn} and set $Q_{jn}(y) \overset{\text{def}}{=} 0$ for $0 \le y \le (2n)^{-1}$. Define the tail marginal empirical processes*

$$w_{jn}(x) \overset{\text{def}}{=} \sqrt{k}\left\{\tfrac{n}{k}\Gamma_{jn}(\tfrac{k}{n}x) - x\right\} \quad \text{and} \quad v_{jn}(x) \overset{\text{def}}{=} \sqrt{k}\left\{\tfrac{n}{k}Q_{jn}(\tfrac{k}{n}x) - x\right\},$$

for $j \in \{1, 2\}$ and $x \ge 0$. Under Assumptions 4.3.1 and 4.3.2, we have for any $p \in [1, \infty]$

$$\sup_{\theta \in [0, \pi/2]} \left| \sqrt{k}\left(\widehat{\Phi}_p(\theta) - \Phi_p(\theta)\right) - E_{n,p}(\theta) \right| = \mathrm{o_P}(1),$$

as $n \to \infty$, where

$$\begin{aligned} E_{n,p}(\theta) &\overset{\text{def}}{=} \sqrt{k}\left\{\tfrac{n}{k}P_n\left(\tfrac{k}{n}C_{p,\theta}\right) - \tfrac{n}{k}P\left(\tfrac{k}{n}C_{p,\theta}\right)\right\} \\ &\quad + \int_0^{x_p(\theta)} \lambda(x, x\tan\theta)\left\{w_{1n}(x)\tan\theta - w_{2n}\left(x\tan\theta\right)\right\}\mathrm{d}x \\ &\quad + \begin{cases} \int_{x_p(\theta)}^{\infty} \lambda(x, y_p(x))\left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\}\mathrm{d}x & \text{if } p < \infty, \\ -w_{1n}(1)\int_1^{1\vee\tan\theta} \lambda(1, y)\mathrm{d}y - w_{2n}(1)\int_{1\vee\cot\theta}^{\infty} \lambda(x, 1)\mathrm{d}x & \text{if } p = \infty. \end{cases} \end{aligned}$$

The proof is deferred to Appendix 4.C.

Let W_Λ be a tight centered Wiener process indexed by Borel sets on $\mathbb{E}$ with covariance function $\mathbb{E}[W_\Lambda(C)W_\Lambda(C')] = \Lambda(C \cap C')$. As explained in the proof of Theorem 4.3.1, it is possible to obtain the joint weak convergence (in the space $\ell^\infty([0, \pi/2]) \times \ell^\infty([0, \infty))^2$)

$$\begin{aligned} &\left(\sqrt{k}\{\tfrac{n}{k}P_n(\tfrac{k}{n}C_{p,\cdot}) - \tfrac{n}{k}P(\tfrac{k}{n}C_{p,\cdot})\}, w_{1n}(\cdot)w(\cdot), w_{2n}(\cdot)w(\cdot)\right) \\ &\qquad \xrightarrow{\mathcal{L}} \left(W_\Lambda(C_{p,\cdot}), W_1(\cdot)w(\cdot), W_2(\cdot)w(\cdot)\right), \end{aligned}$$

from Propositions 4.3.1 and 4.3.2. Therefore, writing $w_{jn} = w_{jn} w \frac{1}{w}$ and noting that the integrals $\int_0^{x_p(\theta)} \lambda(x, x\tan\theta) \frac{1}{w(x)} \mathrm{d}x$ and so on are finite, the continuous mapping theorem then yields $E_{n,p} \xrightarrow{\mathcal{L}} \alpha_p$ in $\ell^\infty([0,\pi/2])$ where $\alpha_p(\theta) \stackrel{\text{def}}{=} W_\Lambda(C_{p,\theta}) + Z_p(\theta)$ for $\theta \in [0,\pi/2]$ with

$$Z_p(\theta) \stackrel{\text{def}}{=} \int_0^{x_p(\theta)} \lambda(x, x\tan\theta)\,[W_1(x)\tan\theta - W_2(x\tan\theta)]\;\mathrm{d}x + \begin{cases} \int_{x_p(\theta)}^\infty \lambda(x, y_p(x)) \left[W_1(x) y_p'(x) - W_2(y_p(x))\right] \mathrm{d}x & \text{if } p < \infty, \\ -W_1(1) \int_1^{1\vee\tan\theta} \lambda(1,y)\,\mathrm{d}y - W_2(1) \int_{1\vee\cot\theta}^\infty \lambda(x,1)\,\mathrm{d}x & \text{if } p = \infty, \end{cases}$$

with y_p' the derivative of y_p and with $W_1(x) \stackrel{\text{def}}{=} W_\Lambda((0,x]\times(0,\infty])$ and $W_2(y) \stackrel{\text{def}}{=} W_\Lambda((0,\infty]\times(0,y])$ the marginal processes. We therefore recover the main result of [47].

Based on Theorem 4.3.1, a simple application of Slutsky's Lemma permits to obtain an expansion of the process associated to the empirical angular probability measure. Some computations lead to

$$\sup_{\theta\in[0,\pi/2]} \left| \sqrt{k} \left(\widehat{Q}_p(\theta) - Q_p(\theta) \right) - E_{n,p}^Q(\theta) \right| = \mathrm{o_P}(1), \tag{4.26}$$

where

$$E_{n,p}^Q(\theta) \stackrel{\text{def}}{=} \frac{E_{n,p}(\theta)\Phi_p(\frac{\pi}{2}) - \Phi_p(\theta)E_{n,p}(\frac{\pi}{2})}{\Phi_p(\frac{\pi}{2})^2}, \qquad \theta \in [0,\pi/2]. \tag{4.27}$$

Consequently, we get weak convergence in $\ell^\infty([0,\pi/2])$,

$$\left\{ \sqrt{k} \left(\widehat{Q}_p(\theta) - Q_p(\theta) \right) \right\}_{\theta\in[0,\pi/2]} \xrightarrow{\mathcal{L}} \left\{ \beta_p(\theta) \stackrel{\text{def}}{=} \frac{\alpha_p(\theta)\Phi_p(\frac{\pi}{2}) - \Phi_p(\theta)\alpha_p(\frac{\pi}{2})}{\Phi_p(\frac{\pi}{2})^2} \right\}_{\theta\in[0,\pi/2]}. \tag{4.28}$$

Let $\Psi : D_\Psi \subset \ell^\infty([0,\pi/2]) \to \ell^\infty([0,\pi/2])$ be the map defined for each element $F \in D_\Psi$, the set of cumulative distribution functions of non-degenerate probability measures on $[0,\pi/2]$, by

$$(\Psi(F))(\theta) \stackrel{\text{def}}{=} F(\theta) - \frac{\mu_F(f)}{\sigma_F^2(f)} \int_0^\theta (f(\psi) - \mu_F(f))\,\mathrm{d}F(\psi), \tag{4.29}$$

where

$$\mu_F(f) \stackrel{\text{def}}{=} \int_0^{\pi/2} f(\psi)\mathrm{d}F(\psi) \quad \text{and} \quad \sigma_F^2(f) \stackrel{\text{def}}{=} \int_0^{\pi/2} (f(\psi) - \mu_F(f))^2\,\mathrm{d}F(\psi), \tag{4.30}$$

and f is defined as in (4.16). The Euclidean likelihood estimator $\widetilde{Q}_p$ satisfies

$$\widetilde{Q}_p(\theta) = (\Psi(\widehat{Q}_p))(\theta), \qquad \theta \in [0, \pi/2].$$

Hadamard differentiability of the map Ψ combined with the functional delta method [121, Theorem 20.8] permits to derive an asymptotic expansion for the maximum Euclidean likelihood estimator. This is formalized in the following result, the proof of which is to be found in Appendix 4.D.

Corollary 4.3.1. *The map $\Psi : D_\Psi \subset \ell^\infty([0, \pi/2]) \to \ell^\infty([0, \pi/2])$ defined in (4.29) is Hadamard differentiable at $Q_p \in D_\Psi$ tangentially to $C_0([0, \pi/2])$, the space of continuous functions on $[0, \pi/2]$ vanishing at the boundaries, with derivative given by*

$$(\Psi'_{Q_p}[h])(\theta) = h(\theta) + \frac{\int_0^{\pi/2} h(\psi) f'(\psi) \mathrm{d}\psi}{\sigma^2_{Q_p}(f)} \int_0^\theta f(\psi) \mathrm{d}Q_p(\psi)$$

for any $h \in C_0([0, \pi/2])$ and $\theta \in [0, \pi/2]$. Furthermore, Ψ'_{Q_p} is defined and continuous on the measurable elements in $\ell^\infty([0, \pi/2])$ so that in $\ell^\infty([0, \pi/2])$,

$$\sqrt{k}\left(\widetilde{Q}_p - Q_p\right) = \sqrt{k}\left(\Psi(\widehat{Q}_p) - \Psi(Q_p)\right) = \Psi'_{Q_p}\left[\sqrt{k}\left(\widehat{Q}_p - Q_p\right)\right] + \mathrm{o_P}(1).$$

Consequently, under the same assumptions as Theorem 4.3.1, we have, as $n \to \infty$,

$$\sup_{\theta \in [0,\pi/2]} \left| \sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_p(\theta)\right) - \left(E^Q_{n,p}(\theta) + \frac{\int_0^{\pi/2} E^Q_{n,p}(\psi) f'(\psi) \mathrm{d}\psi}{\sigma^2_{Q_p}(f)} \int_0^\theta f(\psi) \mathrm{d}Q_p(\psi)\right) \right| = \mathrm{o_P}(1),$$

and, in $\ell^\infty([0, \pi/2])$,

$$\left\{\sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_p(\theta)\right)\right\}_{\theta \in [0,\pi/2]} \xrightarrow{\mathcal{L}} \left\{\gamma_p(\theta) \stackrel{\text{def}}{=} \beta_p(\theta) + \frac{\int_0^{\pi/2} \beta_p(\psi) f'(\psi) \mathrm{d}\psi}{\sigma^2_{Q_p}(f)} \int_0^\theta f(\psi) \mathrm{d}Q_p(\psi)\right\}_{\theta \in [0,\pi/2]}.$$

The asymptotic process in Corollary 4.3.1 equals the one in Theorem 4.1 of [47], showing that, at least asymptotically, enforcing the marginal constraints (4.11) via maximization of the empirical likelihood or its Euclidean version makes no difference, as already discussed and illustrated in [32] for the case $p = 1$.

4.4 Goodness-of-fit tests for parametric models

4.4.1 Test statistic

Given a parametric model $\mathcal{M} \overset{\text{def}}{=} \{\Phi_{p,r} : r \in \mathcal{R} \subseteq \mathbb{R}^m\}$ for the angular measure Φ_p, our main goal is to test

$$H_0 : \Phi_p \in \mathcal{M} \qquad \text{vs} \qquad H_1 : \Phi_p \notin \mathcal{M}. \tag{4.31}$$

As a test statistic, we propose to use a weighted version of the L_1-*Wasserstein distance* from optimal transport theory [126] between $\widetilde{Q}_p$ in (4.14) and $Q_{p,\widehat{r}_n}$, where $\widehat{r}_n$ is some estimator of the true parameter r_0 under H_0, which can be seen as a nuisance parameter. More explicitly, our test statistic will be

$$T_n \overset{\text{def}}{=} \sqrt{k} \int_0^{\pi/2} \left| \widetilde{Q}_p(\theta) - Q_{p,\widehat{r}_n}(\theta) \right| q(\theta) \mathrm{d}\theta, \tag{4.32}$$

where $q : (0, \pi/2) \to (0, \infty)$ is some positive, integrable weight function. The special case $q \equiv 1$ reduces to the classical L_1-Wasserstein distance. In the simulation study, we will also consider $q(\theta) = (|\theta - \pi/4|)^{-1/2}$. Under differentiability assumptions on the model $\mathcal{M}$ and general conditions on the estimator $\widehat{r}_n$, we can derive the asymptotic distribution of T_n in Section 4.4.3. Before we do so, we develop an asymptotic expansion of $\widehat{r}_n$ in Section 4.4.2.

Below, we will use the notation $\varrho_p(\theta) \overset{\text{def}}{=} \varphi_p(\theta)/\Phi_p(\pi/2)$ to denote the density of Q_p at $\theta \in [0, \pi/2]$.

4.4.2 Parameter estimator and empirical stable tail dependence function

The condition on the parameter estimator will be formulated in terms of the *stable tail dependence function* $\ell : [0, \infty)^2 \to [0, \infty)$ defined by

$$\ell(x_1, x_2) \overset{\text{def}}{=} \lim_{t \to 0} t^{-1} \mathrm{P}\left[U_1 \le t x_1 \text{ or } U_2 \le t x_2\right], \tag{4.33}$$

which is known to exist due to (4.2). This function describes the extremal dependence structure of a random vector and has a close link with the angular measure, see [13, Section 8.6.2]. As for the angular measure, we may estimate ℓ nonparametrically by replacing the distribution functions in the expression by their empirical counterparts and by replacing t by k/n. This leads to the rank-based estimator

$$\widehat{\ell}_n(x_1, x_2) \overset{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{n} \mathbb{I}\left\{R_{i1} > n + \tfrac{1}{2} - k x_1 \text{ or } R_{i2} > n + \tfrac{1}{2} - k x_2\right\}, \tag{4.34}$$

where R_{ij} denotes the rank of X_{ij} in the marginal sample $X_{1j}, \dots, X_{nj}$. The constant $1/2$ is added in the indicators to improve finite-sample properties but will not play any role in our asymptotic considerations [45, Equation (7.1)].

Let $x \cdot y$ denote the scalar product of $x, y \in \mathbb{R}^m$ and let $|x|$ denote the Euclidean norm of $x \in \mathbb{R}^m$.

Assumption 4.4.1. *1. The point r_0 lies in the interior of $\mathcal{R}$ and the map $\mathfrak{R} : r \in \mathcal{R} \subseteq \mathbb{R}^m \mapsto Q_{p,r} \in \ell^\infty([0, \pi/2])$ is Hadamard differentiable with derivative $\mathfrak{R}'_r : \mathbb{R}^m \to \ell^\infty([0, \pi/2])$. For every $\theta \in [0, \pi/2]$, the map $r \in \mathcal{R} \mapsto \varrho_{p,r}(\theta)$ is differentiable and the derivative satisfies*

$$\int_0^{\pi/2} \left| \nabla_r \varrho_{p,r_0}(\theta) \right| \mathrm{d}\theta < \infty, \qquad \forall r_0 \in \mathcal{R}.$$

Differentiation with respect to r and integration with respect to θ can be exchanged while considering the map $(r, \theta) \in \mathcal{R} \times [0, \pi/2] \mapsto \varrho_{p,r}(\theta)$, so that

$$(\mathfrak{R}'_r[h])(\theta) = \int_0^\theta \nabla_r \varrho_{p,r}(x) \mathrm{d}x \cdot h.$$

2. The estimator $\widehat{r}_n$ satisfies the following expansion [72, Assumption 4.6]: there exists a finite Borel measure σ on $[0,1]^2$ and a function $g : [0,1]^2 \to \mathbb{R}^m$ in $\left(L_2([0,1]^2, \sigma)\right)^m$ such that, as $n \to \infty$,

$$\sqrt{k}(\widehat{r}_n - r_0) = \int_{[0,1]^2} \sqrt{k} \left(\widehat{\ell}_n(x_1, x_2) - \ell(x_1, x_2) \right) g(x_1, x_2) \, \mathrm{d}\sigma(x_1, x_2) + \mathrm{o_P}(1).$$

The second part of Assumption 4.4.1 will permit to derive an asymptotic expansion of $\sqrt{k}\,(\widehat{r}_n - r_0)$ on the basis of an expansion of the process $\sqrt{k}\,(\widehat{\ell}_n - \ell)$ and hence to derive the limit distribution of T_n under H_0 using the first part of the assumption. It is satisfied by many common estimators, e.g., M-estimators (and hence moment estimators) or weighted least-square estimators, see Examples 4.8–10 in [72].

Asymptotic theory for the process $\sqrt{k}\,(\widehat{\ell}_n - \ell)$ is studied in [45], in general dimension $d \geq 2$ (the definitions are analogous to the bivariate case). For any $T > 0$, the process is asymptotically Gaussian in the space $\ell^\infty([0,T]^d)$, and this under general conditions which are met in our bivariate setting. Under the same assumptions as in [45], we provide an asymptotic expansion of the process in general dimension too.

Assumption 4.4.2. *Let $(X_1, \dots, X_d)$ be a random vector in $\mathbb{R}^d$ with continuous distribution function F and marginal distribution functions $F_1, \dots, F_d$. Let $U \overset{\text{def}}{=} (1 - F_1(X_1), \dots, 1 - F_d(X_d))$ be its standardization to uniform margins, whose*

law will be denoted by P, for which the stable tail dependence function is assumed to exist, that is, for $x \in [0,\infty)^d$*, the limit*

$$\lim_{t\to 0} t^{-1}P\left[tA_x\right] = \lim_{t\to 0} t^{-1}\mathrm{P}\left[U_1 \le tx_1 \text{ or } \ldots \text{ or } U_d \le tx_d\right] \stackrel{\text{def}}{=} \ell(x_1,\ldots,x_d)$$

exists, with $A_x \stackrel{\text{def}}{=} \{u \in [0,\infty)^d : u_1 \le x_1 \text{ or } \ldots \text{ or } u_d \le x_d\}$*. Moreover,*

1. $|t^{-1}\mathrm{P}\left[U_1 \le tx_1 \text{ or } \ldots \text{ or } U_d \le tx_d\right] - \ell(x_1,\ldots,x_d)| = \mathrm{O}(t^\alpha)$ *uniformly in* $x \in [0,\infty)^d$ *such that* $x_1 + \cdots + x_d = 1$ *as* $t \to 0$*;*

2. $k = \mathrm{o}(n^{2\alpha/(1+2\alpha)})$ *and* $k \to \infty$ *as* $n \to \infty$*;*

3. *for every* $j \in \{1,\ldots,d\}$*, the first-order partial derivative of* ℓ *with respect to* x_j*, denoted by* $\dot{\ell}_j$*, exists and is continuous on the set of points* x *such that* $x_j > 0$*;*

where $\alpha > 0$ *is the same in points 1 and 2 above.*

Theorem 4.4.1 (Asymptotic expansion of the empirical stable tail dependence function)**.** *Under Assumption 4.4.2, let* P_n *denote the empirical distribution of an independent random sample from P of size n. Then, for every* $T > 0$*, as* $n \to \infty$*,*

$$\sup_{x\in[0,T]^d}\left|\sqrt{k}\left(\ell_n(x) - \ell(x)\right) - \left(v_n(x) - \sum_{j=1}^{d}\dot{\ell}_j(x)v_{nj}(x_j)\right)\right| = \mathrm{o_P}(1),$$

where, for every $x \in [0,\infty)^d$ *and* $j = 1,\ldots,d$*,*

$$v_n(x) \stackrel{\text{def}}{=} \sqrt{k}\left\{\tfrac{n}{k}P_n\left(\tfrac{k}{n}A_x\right) - \tfrac{n}{k}P\left(\tfrac{k}{n}A_x\right)\right\}$$
$$v_{nj}(x_j) \stackrel{\text{def}}{=} v_n(0,\ldots,0,x_j,0,\ldots,0),$$

and where $\dot{\ell}_j$ *is taken to be the right-hand partial derivative if* $x_l = 0$ *for some* $l = 1,\ldots,d$ *(the latter always exists, see [45, Equation (2.7)]).*

Furthermore, if $d = 2$*, Assumption 4.4.2 is satisfied with the same* $\alpha > 0$ *as in Assumption 4.3.2 provided Assumptions 4.3.1 and 4.3.2 are verified.*

The proof of Theorem 4.4.1 can be found in Appendix 4.E.

4.4.3 Limit distribution of test statistic and estimation of critical values

We are now ready to state the limit distribution of the test statistic T_n under H_0. The proof of Theorem 4.4.2 is provided in Appendix 4.F.

Theorem 4.4.2. *If Assumptions 4.3.1, 4.3.2 and 4.4.1 are satisfied and if $r_0 \in \mathcal{R}$ denotes the true value of the parameter under H_0, then under H_0, we have*

$$T_n \xrightarrow{\mathcal{L}} L_{r_0} \stackrel{\text{def}}{=} \int_0^{\pi/2} \left| \gamma_{p,r_0}(\theta) - \int_0^{\theta} \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot I_{g,\sigma,r_0} \right| q(\theta)\,\mathrm{d}\theta,$$

as $n \to \infty$, where we recall γ_p in Corollary 4.3.1 and where

$$I_{g,\sigma,r_0} \stackrel{\text{def}}{=} \int_{[0,1]^2} \left(W_{\Lambda_{r_0}}(A_{(x_1,x_2)}) - \sum_{j=1}^{2} \dot{\ell}_{r_0,j}(x_1,x_2) W_j(x_j) \right) g(x_1,x_2)\mathrm{d}\sigma(x_1,x_2).$$

Under an additional regularity condition on the model, which will be easily verified in practice, one may show that the critical values for the asymptotic distribution can be computed from an estimated version of the true unknown parameter under H_0. The critical values can thus be computed by Monte Carlo simulation from the limit distribution of T_n in Theorem 4.4.2 but at the estimated parameter value, as explained in Section 4.5.

Assumption 4.4.3. *Let $r_0 \in \mathcal{R}$ be an interior point and introduce the function*

$$\eta : \theta \in [0,\pi/2] \mapsto \eta(\theta) \stackrel{\text{def}}{=} \min\left\{\tan\theta, \tan(\tfrac{\pi}{2} - \theta)\right\} \in [0,1].$$

1. *We have*
$$\sup_{\theta\in[0,\pi/2]} \left\{\eta(\theta)\varphi_{p,r_0}(\theta)\right\} < \infty.$$

2. *There exists $\nu \in (0,1)$ such that*
$$\lim_{n\to\infty} \sup_{\theta\in[0,\pi/2]} \left\{\eta(\theta)^{\nu} \left|\varphi_{p,r_n}(\theta) - \varphi_{p,r_0}(\theta)\right|\right\} = 0,$$
and
$$\lim_{n\to\infty} \sup_{\theta\in[0,\pi/2]} \left\{\eta(\theta)^{\nu} \left|\nabla_r\varphi_{p,r_n}(\theta) - \nabla_r\varphi_{p,r_0}(\theta)\right|\right\} = 0,$$
for any sequence $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ such that $r_n \to r_0$ as $n \to \infty$.

Theorem 4.4.3. *For any $\alpha \in (0,1)$ and $r \in \mathcal{R}$, let $Q_r(\alpha)$ be the α-quantile of the random variable L_r in Theorem 4.4.2. Under the same assumptions as in Theorem 4.4.2 in combination with Assumption 4.4.3, we have*

$$Q_{\widehat{r}_n}(1-\alpha) \xrightarrow{\mathrm{P}} Q_{r_0}(1-\alpha), \qquad \alpha \in (0,1),$$

for any estimator $\widehat{r}_n$ with values in $\mathcal{R}$ which is consistent for r_0, the true parameter under H_0. The convergence also holds locally uniformly: for any $[\alpha_0,\alpha_1] \subset [0,1]$,

$$\sup_{\alpha\in[\alpha_0,\alpha_1]} \left|Q_{\widehat{r}_n}(1-\alpha) - Q_{r_0}(1-\alpha)\right| \xrightarrow{\mathrm{P}} 0.$$

The proof of Theorem 4.4.3 is involved and can be found in Appendix 4.G.

Remark 4.4.1. Point 1 of Assumption 4.4.1 is formulated in terms of the map $r \in \mathcal{R} \mapsto Q_{p,r}$ but is fully equivalent to its counterpart for the map $r \in \mathcal{R} \mapsto \Phi_{p,r}$. The same remark holds true for Assumption 4.4.3 which is equivalent to the present formulation where $\varphi_{p,r}$ would be replaced by $\varrho_{p,r}$. The normalizing constant permitting to move from one definition to another, even though it depends on r, can be shown to have no influence on our assumptions.

4.5 Simulations and finite-sample performance

Theorems 4.4.2 and 4.4.3 provide a natural way to perform the goodness-of-fit test proposed in Eq. (4.31), for any size $\alpha \in (0,1)$. Given a model to be tested and an estimator for the unknown parameter of the model, one can approximate the quantiles $Q_r(1-\alpha)$ for any $r \in \mathcal{R}$ by directly simulating the random variable L_r, the distribution of which only depends on the given model and the asymptotic expansion of the estimator as in point 2 of Assumption 4.4.1, and then taking empirical quantiles. This is typically done for a finite grid of values for $r \in \mathcal{R}$ and, by interpolation, leads to critical values for any value of the parameter. Once those values have been obtained, they can be used for any dataset for which the model needs to be tested, provided that the estimator of the model parameter satisfies the same expansion as the one used to obtain the critical values. The procedure is illustrated for the logistic (Section 4.5.1) and the Hüsler–Reiss (Section 4.5.2) models. Details underlying the simulation of L_r are to be found in Appendix 4.H.

4.5.1 Logistic model

Model. One of the most famous and oldest parametric families of bivariate extreme value dependence structures is the *logistic model* with the stable tail dependence function (4.33)

$$\ell_r(x,y) = \left(x^{1/r} + y^{1/r}\right)^r, \qquad r \in \mathcal{R} = (0,1].$$

The model ranges from asymptotic full dependence ($r \to 0$) to asymptotic independence ($r = 1$). Since $\ell_r(x,y) = x + y - \Lambda_r([0,x] \times [0,y])$, one easily obtains

$$\lambda_r(x,y) = \left(\frac{1}{r} - 1\right) \frac{(xy)^{\frac{1}{r}-1}}{(x^{1/r} + y^{1/r})^{2-r}}, \qquad x, y > 0.$$

For $r \in (0,1)$, the angular measure $\Phi_{p,r}$ on the L_p-sphere, for $1 \le p < \infty$, satisfies $\Phi_p(\{0, \pi/2\}) = 0$, while on the interior of the sphere $(0, \pi/2)$, it is absolutely continuous with a density which can be computed using relation (4.25)

and is given, for $\theta \in (0, \pi/2)$, by

$$\varphi_{p,r}(\theta) = \left(\frac{1}{r} - 1\right) \|(\sin\theta, \cos\theta)\|_p \frac{(\sin\theta\cos\theta)^{\frac{1}{r}-2}}{\left((\sin\theta)^{1/r} + (\cos\theta)^{1/r}\right)^{2-r}}.$$

In particular, Assumption 4.3.1 is satisfied for this model. Point 1 of Assumption 4.4.1 and Assumption 4.4.3 are also satisfied for $r_0 = 0.5$.

Parameter estimation. Given a random sample $X_1, \ldots, X_n$, the unknown parameter $r \in \mathcal{R}$ of the model will be estimated by inverting the *extremal coefficient* [13, Chapter 8]

$$\chi \stackrel{\text{def}}{=} \ell(1,1) \in [1,2],$$

which equals $\chi = 2^r$ for the logistic model. In the limiting case $\chi \to 1$, we approach asymptotic full dependence while $\chi \to 2$ means we reach asymptotic independence. Our estimator simply consists of writing $r = \log_2(\chi)$ and replacing χ by its empirical version leading to

$$\widehat{r}_n = \log_2(\widehat{\ell}_n(1,1)), \tag{4.35}$$

where $\widehat{\ell}_n$ is defined in (4.34). It follows from the delta method that $\widehat{r}_n$ satisfies the asymptotic expansion

$$\sqrt{k}\,(\widehat{r}_n - r) = \sqrt{k}\left(\widehat{\ell}_n(1,1) - \ell(1,1)\right) \cdot \frac{1}{2^r \log(2)} + o_P(1),$$

as $n \to \infty$, meaning that point 2 of Assumption 4.4.1 is satisfied with

$$g \equiv \frac{1}{2^r \log(2)} \qquad \text{and} \qquad \sigma = \delta_{(1,1)}.$$

Data generating processes. We generate data from the copula mixture model

$$C_\lambda(u,v) \stackrel{\text{def}}{=} (1-\lambda)C_0(u,v) + \lambda C_a(u,v), \qquad (u,v) \in [0,1]^2, \tag{4.36}$$

for $\lambda \in \{0, 0.1, 0.2, \ldots, 0.8\}$ where

$$C_0(u,v) \stackrel{\text{def}}{=} \exp\left\{-\left((-\log u)^2 + (-\log v)^2\right)^{0.5}\right\} \tag{4.37}$$

is the *Gumbel* copula with parameter equal to 2 and with two choices for the second component $C_a \in \{C_1, C_2\}$:

- In scenario 1, $C_1(u,v) \stackrel{\text{def}}{=} \min(u,v)$ is the copula of the random vector (U,U) where U is uniformly distributed on $[0,1]$.

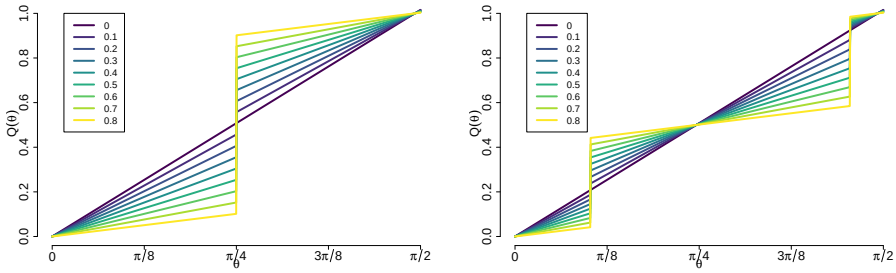
- In scenario 2,

$$C_2(u,v) \overset{\text{def}}{=} \exp\Big\{ -\max(-0.7\log u, -0.1\log v) \\ -\max(-0.3\log u, -0.9\log v)\Big\}$$

is the copula of the max-linear factor model

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 0.7Z_1 \vee 0.3Z_2 \\ 0.1Z_1 \vee 0.9Z_2 \end{pmatrix}$$

where Z_1 and Z_2 are independent Fréchet(1) random factors.



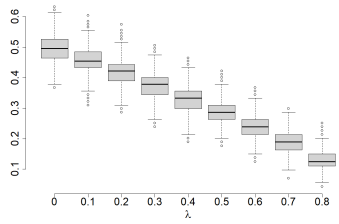
(a) Scenario 1, $C_a = C_1$ **(b) Scenario 2,** $C_a = C_2$

Figure 4.2: Angular distribution functions $\theta \mapsto Q_p(\theta)$ for $p = 2$ of the copula mixture model (4.36) **with Gumbel copula C_0 in** (4.37) **and under two scenarios for C_a.**

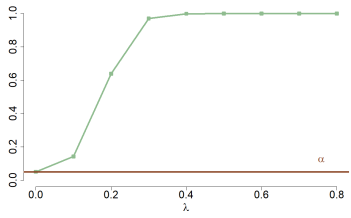
In both cases, when $\lambda = 0$, the null hypothesis of the test (4.31) is satisfied for the logistic model with parameter $r_0 = 0.5$ and Assumption 4.3.2 holds true if we choose $\alpha = 1$, while for $0 < \lambda < 1$, the null hypothesis is no longer satisfied. The asymptotic dependence structures corresponding to those mixtures are represented on Figure 4.2 by their angular distribution function Q_p for $p = 2$. Scenario 1 corresponds exactly to the simulation setup in [35], while scenario 2 is similar to the one in [22], to facilitate comparison.

Results. The results for scenario 1 are presented in Figure 4.3. On the top-left panel, estimates of the logistic model parameter are shown as a function of the mixture weight λ. Results are consistent with the choice $r_0 = 0.5$ and the fact that stronger dependence corresponds to smaller logistic parameter values. The three other panels show the empirical power (green line) of the test for different values of the sample size n and effective sample size k. The weight function is $q(\theta) = 1/\sqrt{|\theta - \pi/4|}$, emphasizing the points close to

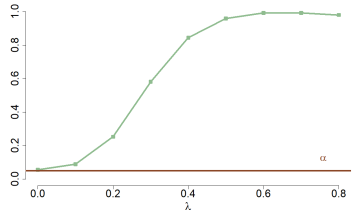
$\theta = \pi/4$. Regarding Figure 4.2a, this choice is expected to help in scenario 1, but, perhaps more surprisingly, it is also helpful in scenario 2 where the largest deviation from the null hypothesis is no longer located at $\theta = \pi/4$. Overall, such a choice produces better results than the classical L_1-Wasserstein distance with $q \equiv 1$. Both the power and the asymptotic quantile are estimated on the basis of 2000 replications of the corresponding quantities. The size α of the test is 5% in all cases and is shown as a horizontal brown line. The graphs confirm the good performance of the test. On the bottom-right panel, the additional orange dots visualize the empirical power of the method in [35] for the same scenario and sample size n. Their test is a bit more conservative under the null, while its power under the alternative is less than the one of our method. The slight power decrease as $\lambda \longrightarrow 1$ is unsurprising since H_0 holds true for the copula C_1 with $r = 0$.



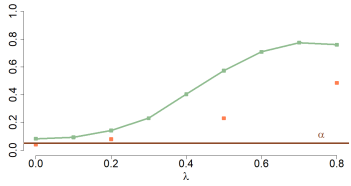
(a) Parameter estimates, $n = 10\,000$ and $k = 100$



(b) Power curve, $n = 10\,000$ and $k = 100$



(c) Power curve, $n = 3\,000$ and $k = 50$

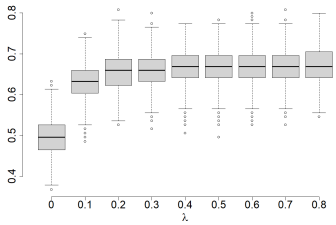


(d) Power curve, $n = 500$ and $k = 25$; orange dots show power of test in [35]

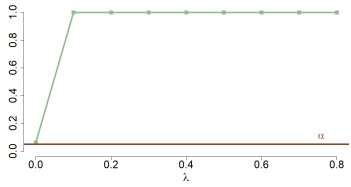
Figure 4.3: Logistic model: parameter estimates (a) and empirical power curves (b–d) of weighted L_1-Wasserstein goodness-of-fit test in scenario 1 of copula mixture model (4.36)

The corresponding results for scenario 2 are presented in Figure 4.4. The empirical power curves are obtained in the same way as for scenario 1. The weighted L_1-Wasserstein test easily detects the asymmetric alternative. The same scenario is considered in [22] (with a slightly different factor model for the alternative, yielding different locations of the atoms of the angular

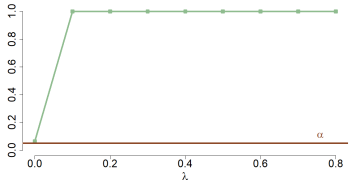
measure), where an empirical power of 90% to 95% is attained depending on the method. In the bottom-right panel, the empirical power almost attains 1, even for $\lambda = 0.8$ and $n = 500$. The method in [22] has the advantage that the critical values do not depend on the model and thus have to be computed only once; the computation of the test statistic is more involved, however.



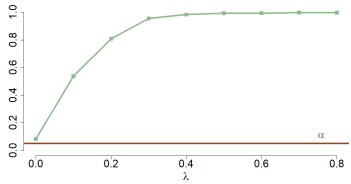
(a) Parameter estimates, $n = 10\,000$ **and** $k = 100$



(b) Power curve, $n = 10\,000$ **and** $k = 100$



(c) Power curve, $n = 3\,000$ **and** $k = 50$



(d) Power curve, $n = 500$ **and** $k = 25$

Figure 4.4: Logistic model: parameter estimates (a) and empirical power curves (b–d) of weighted L_1-Wasserstein goodness-of-fit test in scenario 2 of copula mixture model (4.36)

4.5.2 Hüsler–Reiss model

Model. Also highly popular in extreme value analysis is the *Hüsler–Reiss* model [70], with stable tail dependence function (4.33)

$$\ell_r(x, y) = x\,\Phi\left(r + (2r)^{-1}\log(x/y)\right) + y\,\Phi\left(r + (2r)^{-1}\log(y/x)\right), \quad r \in (0, \infty),$$

where Φ is the standard normal distribution function, not to be confounded with the angular measure Φ. Again, the model covers the full range from asymptotic full dependence ($r \to 0$) to asymptotic independence ($r \to \infty$).

Straightforward computations give

$$\lambda_r(x,y) = \frac{1}{2ry}\phi\left(r + (2r)^{-1}\log(x/y)\right)\left(0.5 - (4r^2)^{-1}\log(x/y)\right) + \frac{1}{2rx}\phi\left(r + (2r)^{-1}\log(y/x)\right)\left(0.5 - (4r^2)^{-1}\log(y/x)\right), \qquad x, y > 0,$$

where ϕ is the standard normal density. The L_p-sphere angular measure $\Phi_{p,r}$, for $1 \le p < \infty$, again satisfies $\Phi_p(\{0, \pi/2\}) = 0$, while on the interior of the sphere, its density can be computed via relation (4.25). In particular, Assumption 4.3.1 is satisfied for this model. Point 1 of Assumption 4.4.1 and Assumption 4.4.3 are also satisfied for $r_0 = 1$.

Parameter estimation. Given a random sample $X_1, \ldots, X_n$, the unknown model parameter $r \in \mathcal{R}$ is again estimated by inverting the extremal coefficient $\chi = \ell(1,1)$ which equals $\chi = 2\Phi(r)$ in this case, yielding

$$\widehat{r}_n = \Phi^{-1}\left(\frac{\widehat{\ell}_n(1,1)}{2}\right). \tag{4.38}$$

By the delta method, we deduce

$$\sqrt{k}\,(\widehat{r}_n - r) = \sqrt{k}\left(\widehat{\ell}_n(1,1) - \ell(1,1)\right) \cdot \frac{1}{2\phi(r)} + \mathrm{o_P}(1),$$

as $n \to \infty$, implying that point 2 of Assumption 4.4.1 is satisfied with

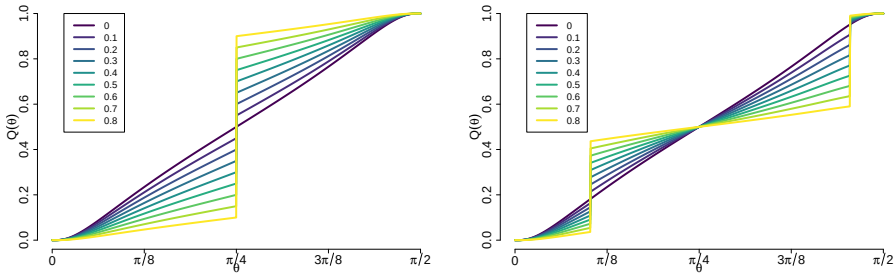
$$g \equiv \frac{1}{2\phi(r)} \qquad \text{and} \qquad \sigma = \delta_{(1,1)}.$$

Data generating process. As for the logistic model, we generate data from the copula mixture model (4.36) with mixture weight $\lambda \in \{0, 0.1, 0.2, \ldots, 0.8\}$ and with alternative copula component $C_a \in \{C_1, C_2\}$ as in the same two scenarios, while

$$C_0(u,v) \stackrel{\text{def}}{=} \exp\left[\Phi\left\{0.5 + \log\left(\frac{\log v}{\log u}\right)\right\}\log u + \Phi\left\{0.5 + \log\left(\frac{\log u}{\log v}\right)\right\}\log v\right] \tag{4.39}$$

is the *Hüsler–Reiss* copula with parameter $r_0 = 1$.

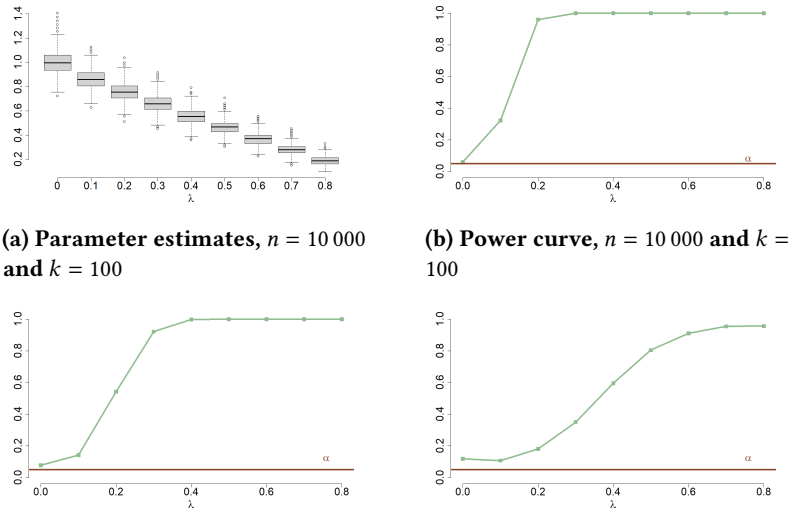
The null hypothesis of the test (4.31) is satisfied for the Hüsler–Reiss model with parameter $r_0 = 1$ when $\lambda = 0$ and Assumption 4.3.2 holds true if we choose $\alpha = 1$, while for $0 < \lambda < 1$, the null hypothesis is not satisfied anymore. For the two scenarios, the angular distribution functions $\theta \mapsto Q_p(\theta)$ at $p = 2$ of the copula mixture model are shown in Figure 4.5.



(a) Scenario 1, $C_a = C_1$ **(b) Scenario 2,** $C_a = C_2$

Figure 4.5: Angular distribution functions $\theta \mapsto Q_p(\theta)$ **for** $p = 2$ **of the copula mixture model** (4.36) **with Hüsler–Reiss copula** C_0 **in** (4.39) **and under two scenarios for** C_a**.**

Results. The results for scenarios 1 and 2 are shown in Figures 4.6 and 4.7, respectively.



(a) Parameter estimates, $n = 10\,000$ **and** $k = 100$

(b) Power curve, $n = 10\,000$ **and** $k = 100$

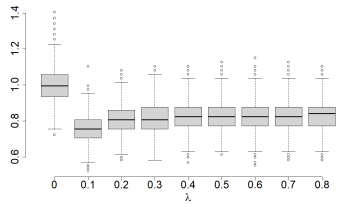
(c) Power curve at $n = 3\,000$ **and** $k = 50$

(d) Power curve at $n = 500$ **and** $k = 25$

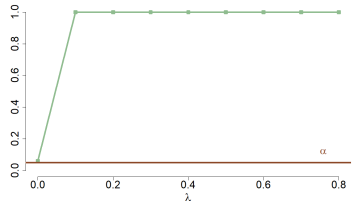
Figure 4.6: Hüsler–Reiss model: parameter estimates (a) and empirical power curves (b–d) of weighted L_1**-Wasserstein goodness-of-fit test in scenario 1 of copula mixture model** (4.36)

The settings are the same as those for the logistic model in Section 4.5.1. The results are similar, even though for $\lambda = 0$ and $n = 500$, the test rejects the

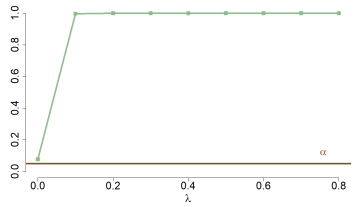
null hypothesis too often. In scenario 2, the estimated power remains higher than in scenario 1, despite the fact that the distribution function of the angular measure for the mixing component C_2 resembles that of the Hüsler–Reiss model with $r_0 = 1$ more.



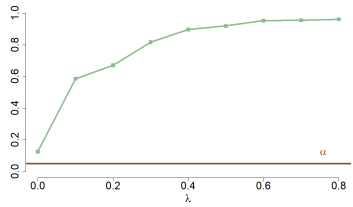
(a) Parameter estimates, $n = 10\,000$ and $k = 100$



(b) Power curve, $n = 10\,000$ and $k = 100$



(c) Power curve, $n = 3000$ and $k = 50$



(d) Power curve, $n = 500$ and $k = 25$

Figure 4.7: Hüsler–Reiss model: parameter estimates (a) and empirical power curves (b–d) of weighted L_1-Wasserstein goodness-of-fit test in scenario 1 of copula mixture model (4.36)

4.6 Testing the Hüsler–Reiss model for river discharge data

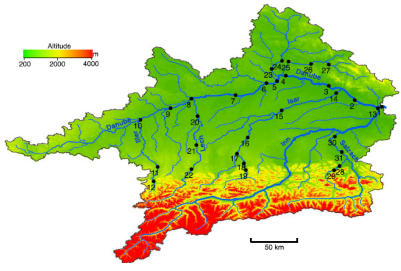
We consider the `danube` dataset from the `R` package `graphicalExtremes` [49]. Average daily discharges are recorded at 31 gauging stations in the upper Danube basin covering parts of Germany, Austria and Switzerland which are often affected by flooding. To not have to deal with seasonality, we only consider the daily discharges in the summer months (June, July and August), and we use the declustered time series from the `data_clustered` dataset. The same data have already been analysed by several authors, e.g. [72], who estimate bivariate dependence structures using the Hüsler–Reiss model for connected pairs of stations after estimating a Markov tree on the full dataset. The topographic map of the upper Danube basin which can be found in Figure 1 of [8] is displayed in Figure 4.8 together with the associated flow chart of its river network.

Following the approach of [72], our goal is to assess the goodness-of-fit

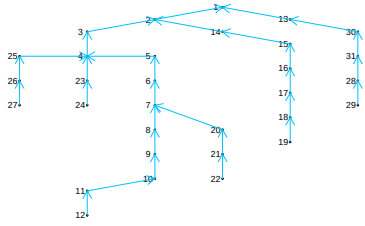
of the Hüsler–Reiss model to all $m = 30$ adjacent pairs in the flow chart of Figure 4.8b. Naturally, this leads to multiple testing challenges due to the many dependent tests involved. The p-values can be adjusted using either the classical Bonferroni correction or the False Discovery Rate (FDR) paradigm of Benjamini and Hochberg [14], adapted for dependent tests in [15]. The sample size is $n = 428$, and we set $k = \sqrt{n}$ as we observed in Section 4.5.2 that this typically leads to better results for the nonparametric estimation of Q_p, while still providing a reasonable estimate of the nuisance parameter r, with respect to which the quantile values $Q_r(1-\alpha)$ do not vary too quickly. Ideally, we would like to use two different values of k for the nonparametric and parametric estimators in the test statistic T_n, as a larger k often improves the accuracy of the parameter estimators. This, however, is a topic for future research. We estimate the p-values P_j corresponding to each of the $j \in \{1, \dots, m\}$ tests by

$$\widehat{P}_j = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left\{ T_n > L_{\widehat{r}_n}^{(b)} \right\}$$

for $B = 4\,000$, where $L_{\widehat{r}_n}^{(b)}$ is drawn from the distribution of L_r in Theorem 4.4.2 at $r = \widehat{r}_n$, computed in the same way as in Section 4.5.2. Although alternative parameter estimators could be used, this would result in more complex expressions for $I_{g,\sigma,r}$ in the definition of T_n, and we aimed to keep the methodology simple. Results are shown in Table 4.1.



(a) Topographic map from [8]



(b) Flow chart

Figure 4.8: The upper Danube basin

With the Bonferroni correction, the Hüsler–Reiss model would only be rejected for edges 14–2 and 23–4, using the standard significance level of $\alpha = 5\%$. In fact, even at the individual test level, a large proportion of these p-values are significantly higher than the rejection threshold α. In Figure 4.9, we compare the nonparametric estimate of the angular distribution function $\theta \in [0, \pi/2] \mapsto Q_2(\theta)$ with its estimated counterpart under the null hypothesis of the Hüsler–Reiss model for edges 14–2 and 23–4. The asymmetry in the

edge	p-value	edge	p-value
2–1	0.02500	12–11	0.05425
13–1	0.60825	30–13	0.07450
3–2	0.75750	15–14	0.49675
14–2	0.00000	16–15	0.84725
4–3	0.91175	17–16	0.84875
5–4	0.34800	18–17	0.25175
23–4	0.00150	19–18	0.15525
25–4	0.43725	21–20	0.17775
6–5	0.67950	22–21	0.15525
7–6	0.64925	24–23	0.61300
8–7	0.40375	26–25	0.72550
20–7	0.76375	27–26	0.78075
9–8	0.51550	29–28	0.19350
10–9	0.85875	31–28	0.57775
11–10	0.16925	31–30	0.33550

Table 4.1: p-values of the weighted L_1-Wasserstein goodness-of-fit test of the bivariate Hüsler–Reiss model for the 30 pairs of adjacent gauging stations in the Danube data

angular distribution function observed in the data for these edges cannot be captured by the assumed Hüsler–Reiss model. This leads to the rejection of the null hypothesis and suggests that this parametric form may not be well-suited for these pairs of stations.

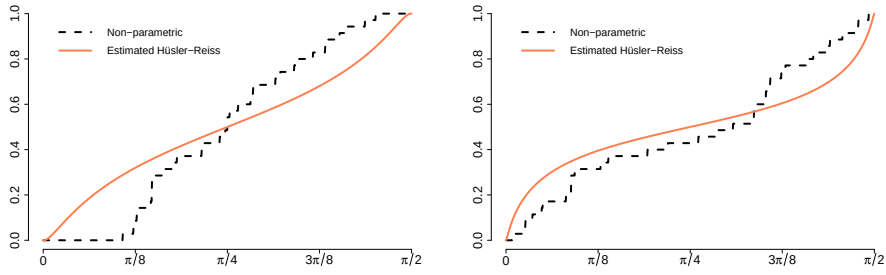


Figure 4.9: Danube data: estimated angular distributions functions $\theta \in [0, \pi/2] \mapsto Q_2(\theta)$ for edges 14–2 and 23–4 of gauging stations

4.7 Conclusion

We provide an extension of the result in [47] which consists of a functional asymptotic expansion of the empirical process associated with the empirical

angular measure. Relying on this expansion, we propose a new inference method based on the angular measure to evaluate whether the asymptotic dependence structure of a bivariate dataset aligns with a given parametric model. Unlike previous approaches, our method involves sampling directly from the asymptotic distribution of the test statistic under the estimated parameter value. Leveraging continuity of the asymptotic quantile function, this leads to a detailed proof of the method's consistency. The finite-sample performance is also compelling when compared to other techniques, and the ease of computing the test statistic makes it a strong candidate for practical applications. Although the method is currently limited to bivariate data, many multivariate extreme value methods rely on multiple bivariate analyses, and we demonstrate how this can be done on a real dataset.

Extending the procedure to the general multivariate case is certainly of interest but requires a deeper understanding of the asymptotic behavior of the empirical angular measure, about which little is currently known. A promising approach would be to focus on discrete angular measures, for which more concrete results might be achievable. Using the Wasserstein distance as a test statistic to quantify the discrepancy between $\widetilde{Q}_p$ and $Q_{p,\widehat{r}_n}$ could be effective, leveraging results from [118], where the Wasserstein distance is shown to be Hadamard differentiable with respect to its arguments for measures with discrete supports.

4.A Proof of weak convergence results

Both proofs rely on an application of Theorem 19.28 in [121] which we recall here for convenience.

Theorem 4.A.1. *Let $X_1, \ldots, X_n$ be a random sample from a probability distribution P on a measurable space $(S, \mathcal{S})$. Let P_n denote the associated empirical distribution. Let $\mathcal{F}_n = \{f_{n,t} : S \to \mathbb{R},\ t \in T\}$ be a class of measurable functions indexed by a totally bounded semimetric space (T, ρ) satisfying*

$$\lim_{n\to\infty} \sup_{\rho(s,t)<\delta_n} P\left((f_{n,s} - f_{n,t})^2\right) = 0, \qquad \forall \delta_n \downarrow 0,$$

and with envelope function F_n satisfying the Lindeberg condition

$$P\left(F_n^2\right) = \mathrm{O}(1) \qquad \textit{and} \qquad P\left(F_n^2 \mathbb{I}\{F_n > \epsilon\sqrt{n}\}\right) = \mathrm{o}(1), \quad \epsilon > 0.$$

If every class $\mathcal{F}_n$ is suitably measurable in the sense of Example 2.3.4 in [122] and if

$$\lim_{n\to\infty} \sup_{Q} \int_0^{\delta_n} \sqrt{\log N(\epsilon \|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q))}\, \mathrm{d}\epsilon = 0, \qquad \forall \delta_n \downarrow 0,$$

where the supremum is taken over all finitely discrete probability measures Q on $(S, \mathcal{S})$ for which $\mathcal{F}_n$ is not identically zero and N is the covering number as defined in [121, page 274], then the sequence of empirical processes $\{G_n(f_{n,t}) \stackrel{\text{def}}{=} \sqrt{n}(P_n(f_{n,t}) - P(f_{n,t})) : t \in T\}$ converges in distribution to a tight Gaussian process, provided the sequence of covariance functions $P(f_{n,s}f_{n,t}) - P(f_{n,s})P(f_{n,t})$ converges pointwise on $T \times T$.

Proof of Proposition 4.3.1. For any $n \in \mathbb{N}$ and $A \in \mathcal{A}$, let $\mathcal{F}_n$ be the class of functions $f_{n,A} : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f_{n,A}(x) \stackrel{\text{def}}{=} \sqrt{\tfrac{n}{k}}\mathbb{I}\left\{x \in \tfrac{k}{n}A\right\}.$$

These classes satisfy the measurability assumptions given the assumptions on the class $\mathcal{A}$. Note that for $A \in \mathcal{A}$,

$$G_n\left(f_{n,A}\right) = \sqrt{k}\left(\tfrac{n}{k}P_n(\tfrac{k}{n}A) - \tfrac{n}{k}P(\tfrac{k}{n}A)\right).$$

Let us verify that the technical conditions underlying Theorem 19.28 in [121] are satisfied.

We first show that the metric space $(\mathcal{A}, \rho)$ is totally bounded. Indeed, noting that $\Lambda(L_M) \leq d \times M$, since each element of $\mathcal{A}$ is included in L_M, we may construct the standard empirical process with respect to the probability measure $P_M(\cdot) \stackrel{\text{def}}{=} \Lambda(\cdot \cap L_M)/\Lambda(L_M)$ on L_M, indexed by functions $f_A \stackrel{\text{def}}{=} \mathbb{I}_A$ for $A \in \mathcal{A}$, and use the fact that $\mathcal{A}$ is a VC-class with $|f_A| \leq F_M \stackrel{\text{def}}{=} \mathbb{I}_{L_M}$ pointwise for each $A \in \mathcal{A}$. Apply Theorem 19.14 in [121] to conclude that the considered process converges weakly in $\ell^\infty(\mathcal{A})$, where we identify $\mathcal{A}$ with its class of indicator functions. In view of Theorem 18.14 and Lemma 18.15 in [121], the metric space $(\mathcal{A}, \rho_M)$ with $\rho_M \stackrel{\text{def}}{=} \rho/\Lambda(L_M)$ must be totally bounded, which also guarantees that $(\mathcal{A}, \rho)$ is totally bounded.

We now show that

$$\lim_{n\to\infty} \sup_{A,A' \in \mathcal{A}:\rho(A,A')<\delta_n} P\left((f_{n,A} - f_{n,A'})^2\right) = 0 \tag{4.40}$$

whenever $\delta_n \downarrow 0$ as $n \to \infty$. Indeed,

$$\begin{aligned}\sup_{A,A':\rho(A,A')<\delta_n} P\left((f_{n,A} - f_{n,A'})^2\right) &= \sup_{A,A':\rho(A,A')<\delta_n} \tfrac{n}{k}P\left((\mathbb{I}_{(k/n)A} - \mathbb{I}_{(k/n)A'})^2\right) \\ &= \sup_{A,A':\rho(A,A')<\delta_n} \tfrac{n}{k}P\left(\tfrac{k}{n}(A \triangle A')\right).\end{aligned}$$

Fix $\epsilon > 0$. It follows from our assumptions that there exists $N_1(\epsilon) \in \mathbb{N}$ such that for $n \geq N_1(\epsilon)$,

$$\sup_{A,A' \in \mathcal{A}:\rho(A,A')<\delta_n} \left|\tfrac{n}{k}P\left(\tfrac{k}{n}(A \triangle A')\right) - \Lambda(A \triangle A')\right| \leq \epsilon/2,$$

while, since $\delta_n \to 0$ as $n \to \infty$, we may find $N_2(\epsilon) \in \mathbb{N}$ such that for $n \geq N_2(\epsilon)$,

$$\sup_{A,A' \in \mathcal{A}: \rho(A,A') < \delta_n} \Lambda(A \triangle A') \leq \epsilon/2.$$

It follows that for $n \geq \max\{N_1(\epsilon), N_2(\epsilon)\}$, we have

$$\sup_{A,A' \in \mathcal{A}: \rho(A,A') < \delta_n} P\left((f_{n,A} - f_{n,A'})^2\right) \leq \epsilon,$$

which yields (4.40).

We now verify that the classes $\mathcal{F}_n$ possess envelope functions F_n satisfying the Lindeberg condition. It suffices to note that for every $n \in \mathbb{N}$ and $A \in \mathcal{A}$,

$$f_{n,A}(x) = \sqrt{\tfrac{n}{k}}\mathbb{I}\left\{x \in \tfrac{k}{n}A\right\} \leq \sqrt{\tfrac{n}{k}}\mathbb{I}\left\{x \in \tfrac{k}{n}L_M\right\} \stackrel{\text{def}}{=} F_n(x).$$

The first part of the Lindeberg condition clearly holds since

$$\lim_{n\to\infty} P\left(F_n^2\right) = \lim_{n\to\infty} \tfrac{n}{k} P\left(\tfrac{k}{n}L_M\right) = \Lambda\left(L_M\right) \leq d \cdot M.$$

The second part of the Lindeberg condition is immediate since for any $\epsilon > 0$, the indicator $\mathbb{I}\{F_n > \epsilon\sqrt{n}\}$ is zero as soon as $1/\sqrt{k} \leq \epsilon$, which is the case for n large enough, since, by assumption, $k = k(n) \to \infty$.

The last condition to verify is the one associated to the uniform entropy integral. Here, since the VC dimension of any $\mathcal{F}_n$ equals the one of $\mathcal{A}$, which is finite by assumption, we may simply apply [122, Example 2.11.24].

Consequently, Theorem 19.28 in [122] applies and the covariance of the limiting Gaussian process is obtained by computing the limit

$$\lim_{n\to\infty} P\left(f_{n,A} f_{n,A'}\right) - P\left(f_{n,A}\right) P\left(f_{n,A'}\right) = \Lambda(A \cap A'),$$

for $A, A' \in \mathcal{A}$, which follows from our assumption.

Finally, we show (4.20). By Example 1.5.10 in [122] the processes $W_{n,k}$ are asymptotically equicontinuous in probability with respect to the standard deviation semi-metric $\rho_2(A, A') \stackrel{\text{def}}{=} (\mathbb{E}[|W_\Lambda(A) - W_\Lambda(A')|^2])^{1/2}$. Since

$$\mathbb{E}[|W_\Lambda(A) - W_\Lambda(A')|^2] = \Lambda(A) + \Lambda(A') - 2\Lambda(A \cap A') = \Lambda(A \triangle A') = \rho(A, A'),$$

the said equicontinuity property holds with respect to ρ as well. □

Proof of Proposition 4.3.2. For any $n \in \mathbb{N}$ and $t \in T \stackrel{\text{def}}{=} [0, \infty)$, let $\mathcal{F}_n$ be the class of functions $f_{n,t} : [0, \infty) \to \mathbb{R}$ defined by

$$f_{n,t}(x) \stackrel{\text{def}}{=} \sqrt{\tfrac{n}{k}}\mathbb{I}\left(x \leq \tfrac{k}{n}t\right) w(t) \mathbb{I}\left(t \leq \tfrac{n}{k}\right).$$

These classes trivially satisfy the measurability assumptions underlying Theorem 19.28 in [121] since it suffices to consider positive rational coordinates t. For each $t \in T$,

$$G_n(f_{n,t}) = \sqrt{n}\left(P_n(f_{n,t}) - P(f_{n,t})\right) = w_n(t)w(t),$$

where P_n is the empirical distribution of the i.i.d. sample $U_1, \ldots, U_n$ whose common law is P, the uniform distribution on $[0,1]$.

Our aim is to apply Theorem 19.28 in [121]. We first need to find a metric ρ on T such that the metric space (T, ρ) is totally bounded and

$$\lim_{n\to\infty} \sup_{\rho(s,t)<\delta_n} P\left((f_{n,s} - f_{n,t})^2\right) = 0, \tag{4.41}$$

for any $(\delta_n)_{n\in\mathbb{N}}$ such that $\delta_n \downarrow 0$. Since the weak limit of the process is expected to be centered Gaussian process, we may choose ρ to be the "standard deviation metric" as in [121, Lemma 18.15]. After some computations, we find that it is given by

$$\rho(s,t) \stackrel{\text{def}}{=} \sqrt{w(t)^2 t + w(s)^2 s - 2w(t)w(s)(s \wedge t)}, \qquad s,t \in T.$$

With this metric, it is easily checked that (4.41) holds and that the metric space (T, ρ) is totally bounded.

We now verify that the classes $\mathcal{F}_n$ possess envelope functions F_n satisfying the Lindeberg condition. It suffices to note that for every $n \in \mathbb{N}$ and $t \in T$,

$$f_{n,t}(x) \le \sqrt{\tfrac{n}{k}}\, w\left(\tfrac{n}{k}x\right) \mathbb{I}(x \le 1) \stackrel{\text{def}}{=} F_n(x).$$

We have

$$\begin{aligned} P\left(F_n^2\right) &= \frac{n}{k}\int_0^1 w\left(\frac{n}{k}x\right)^2 \mathrm{d}x \\ &= \frac{n}{k}\left\{\int_0^{k/n} w\left(\frac{n}{k}x\right)^2 \mathrm{d}x + \int_{k/n}^1 w\left(\frac{n}{k}x\right)^2 \mathrm{d}x\right\} \\ &= \frac{1}{1-2\delta} + \frac{1 - \left(\frac{k}{n}\right)^{2\eta-1}}{2\eta - 1} \to \frac{1}{1-2\delta} + \frac{1}{2\eta-1} \end{aligned}$$

as $n \to \infty$, so that the first part of the Lindeberg condition holds. To verify the second part, we note that for any $\epsilon > 0$,

$$\begin{aligned} P\left(F_n^2 \mathbb{I}(F_n > \epsilon\sqrt{n})\right) &= \tfrac{n}{k}\int_0^1 w\left(\tfrac{n}{k}x\right)^2 \mathbb{I}\left(w\left(\tfrac{n}{k}x\right) > \epsilon\sqrt{k}\right) \mathrm{d}x \\ &\le \int_0^\infty w(y)^2 \mathbb{I}\left(w(y) > \epsilon\sqrt{k}\right) \mathrm{d}y. \end{aligned}$$

Since $y \in (0,\infty) \mapsto w(y)^2$ is integrable and the integrand, which is bounded pointwise by this function, converges pointwise to zero, the dominated convergence theorem permits to conclude.

The last condition to verify is the one associated to the uniform entropy integral. We propose to use Example 2.11.24 in [122]. To this end, we need to verify that the classes $\mathcal{F}_n$ have finite VC dimensions which can be bounded independently of $n \in \mathbb{N}$. The VC dimension of $\mathcal{F}_n$ is, by definition, the VC dimension of the class of sets

$$\mathcal{S}_n \stackrel{\text{def}}{=} \left\{ S_{n,t} \subseteq \mathbb{R}^2 : t \in T \right\},$$

where $S_{n,t}$ is the sub-graph of $f_{n,t}$ given by

$$S_{n,t} \stackrel{\text{def}}{=} \left\{ (x,h) : f_{n,t}(x) > h \right\}.$$

We compute that

$$S_{n,t} = \left(\left[0, \tfrac{k}{n} t\right] \times \left[0, \sqrt{\tfrac{n}{k}} w(t)\right] \right) \cup \left([0,\infty) \times (-\infty, 0) \right).$$

Hence, each element of the class $\mathcal{S}_n$ is formed from the union of two rectangles where the second rectangle does not depend on t. Hence, the VC dimension of $\mathcal{S}_n$ is bounded by 4, which corresponds to the VC dimension of the class of all rectangles on $\mathbb{R}^2$. The uniform entropy condition is then verified.

Theorem 19.28 in [121] applies and the computation of the covariance function of the limiting Gaussian process is an easy exercise.

To prove the last assertion, we need, given any $\epsilon, \eta > 0$, to find $N = N(\epsilon,\eta) \in \mathbb{N}$ such that for any $n \geq N$,

$$\mathsf{P}\left[\sup_{x \in [0,M]} \frac{|v_n(x) + w_n(x)|}{x^\delta} > \eta \right] < \epsilon.$$

To do so, observe that for any $0 < a < 1$,

$$\begin{aligned} \mathsf{P}\left[\sup_{x \in [0,M]} \frac{|v_n(x) + w_n(x)|}{x^\delta} > \eta \right] &\leq \mathsf{P}\left[\sup_{x \in [a,M]} |v_n(x) + w_n(x)| > \eta a^\delta \right] \\ &\quad + \mathsf{P}\left[\sup_{x \in [0,a]} \frac{|w_n(x)|}{x^\delta} > \frac{\eta}{2} \right] + \mathsf{P}\left[\sup_{x \in [0,a]} \frac{|v_n(x)|}{x^\delta} > \frac{\eta}{2} \right]. \end{aligned} \tag{4.42}$$

By the functional delta method, we may find $N_1 \in \mathbb{N}$ such that for $n \geq N_1$, the first term on the right hand side of (4.42) satisfies

$$\mathsf{P}\left[\sup_{x \in [a,M]} |v_n(x) + w_n(x)| > \eta a^\delta \right] < \frac{\epsilon}{3}.$$

For the second term on the right hand side of (4.42), note that by the first part of the proposition, there exists $N_2 \in \mathbb{N}$ such that for $n \geq N_2$,

$$\mathrm{P}\left[\sup_{x\in[0,a]} \frac{|w_n(x)|}{x^\delta} > \frac{\eta}{2}\right] \leq \mathrm{P}\left[\sup_{x\in[0,a]} \frac{|W_j(x)|}{x^\delta} > \frac{\eta}{2}\right] + \frac{\epsilon}{6}$$

and we may find $a_1 \in (0,1)$ small enough such that for $0 < a < a_1$,

$$\mathrm{P}\left[\sup_{x\in[0,a]} \frac{|W_j(x)|}{x^\delta} > \frac{\eta}{2}\right] < \frac{\epsilon}{6}.$$

For the second term on the right hand side of (4.42), note that for $x \in [0, (2k)^{-1}]$, we have $v_n(x) = -x\sqrt{k}$. Hence,

$$\sup_{x\in[0,(2k)^{-1}]} \frac{|v_n(x)|}{x^\delta} = \sqrt{k} \sup_{x\in[0,(2k)^{-1}]} x^{1-\delta} \leq k^{1/2-1+\delta} \to 0, \qquad n \to \infty.$$

Therefore, we may restrict the supremum to $x \in ((2k)^{-1}, a]$. On that region, we have

$$\forall x > (2k)^{-1}, \forall u \geq 0, \qquad Q_n(\tfrac{k}{n}x) > u \iff \Gamma_n(u) < \tfrac{k}{n}x. \tag{4.43}$$

By definition of the absolute value and by the union bound, for every $\eta > 0$,

$$\begin{aligned}\mathrm{P}&\left[\sup_{x\in((2k)^{-1},a]} \frac{|v_n(x)|}{x^\delta} > \eta\right] \\ &\leq \mathrm{P}\left[\sup_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} > \eta\right] + \mathrm{P}\left[\inf_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} < -\eta\right].\end{aligned} \tag{4.44}$$

We will treat each of the two terms on the right-hand side of (4.44) separately. Recall $w_n(x) = \sqrt{k}\left(\frac{n}{k}\Gamma_n(\frac{k}{n}x) - x\right)$ for $x \geq 0$. First, by (4.43), for $x > (2k)^{-1}$,

$$\begin{aligned}\frac{v_n(x)}{x^\delta} > \eta &\iff Q_n(\tfrac{k}{n}x) > \tfrac{k}{n}(x + \eta x^\delta/\sqrt{k}) \\ &\iff \Gamma_n\left(\tfrac{k}{n}(x + \eta x^\delta/\sqrt{k})\right) < \tfrac{k}{n}x \\ &\iff \sqrt{k}\left\{\tfrac{n}{k}\Gamma_n\left(\tfrac{k}{n}(x + \eta x^\delta/\sqrt{k})\right) - (x + \eta x^\delta/\sqrt{k})\right\} < -\eta x^\delta \\ &\iff -\frac{w_n(x + \eta x^\delta/\sqrt{k})}{x^\delta} > \eta.\end{aligned}$$

Therefore,

$$\mathrm{P}\left[\sup_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} > \eta\right] = \mathrm{P}\left[\sup_{x\in((2k)^{-1},a]} \frac{-w_n(x + \eta x^\delta/\sqrt{k})}{x^\delta} > \eta\right].$$

Let $\gamma = \delta/(\delta + 1/2)$ and note that $\delta < \gamma < 1/2$ since $0 < \delta < 1/2$. We have

$$\frac{w_n(x + \eta x^\delta/\sqrt{k})}{x^\delta} = \frac{w_n(x + \eta x^\delta/\sqrt{k})}{(x + \eta x^\delta/\sqrt{k})^\gamma} \cdot \frac{(x + \eta x^\delta/\sqrt{k})^\gamma}{x^\delta}.$$

Since $\delta/\gamma = \delta + 1/2$, the second factor on the right-hand side satisfies

$$\begin{aligned}\frac{(x + \eta x^\delta/\sqrt{k})^\gamma}{x^\delta} &= \left(x \cdot x^{-\delta/\gamma} + \eta x^\delta \cdot x^{-\delta/\gamma}/\sqrt{k}\right)^\gamma \\ &= \left(x^{1-\delta-1/2} + \eta x^{-1/2}/\sqrt{k}\right)^\gamma \le 2, \qquad \forall x \in ((2k)^{-1}, 1].\end{aligned}$$

We find

$$\begin{aligned}\mathsf{P}\left[\sup_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} > \eta\right] &\le \mathsf{P}\left[\sup_{x\in((2k)^{-1},a]} \frac{w_n(x + \eta x^\delta/\sqrt{k})}{(x + \eta x^\delta/\sqrt{k})^\gamma} > \eta/2\right] \\ &= \mathsf{P}\left[\sup_{z\in[0,a+\eta a^\delta/\sqrt{k}]} \frac{w_n(z)}{z^\gamma} > \eta/2\right] \\ &\le \mathsf{P}\left[\sup_{z\in[0,2a]} \frac{w_n(z)}{z^\gamma} > \eta/2\right]\end{aligned}$$

for k sufficiently large. As $0 < \gamma < 1/2$, it follows from previous considerations that we may find $N_3 \in \mathbb{N}$ and $a_2 \in (0, 1)$ such that for $n \ge \mathbb{N}_3$ and $0 < a < a_2$,

$$\mathsf{P}\left[\sup_{z\in[0,2a]} \frac{w_n(z)}{z^\gamma} > \eta/2\right] < \frac{\epsilon}{6}.$$

Let us now consider the second term in (4.44). By (4.43), for $x > (2k)^{-1}$ such that also $x \ge \eta x^\delta \sqrt{k}$,

$$\begin{aligned}\frac{v_n(x)}{x^\delta} \le -\eta &\iff Q_n(\tfrac{k}{n}x) \le \tfrac{k}{n}(x - \eta x^\delta/\sqrt{k}) \\ &\iff \Gamma_n\left(\tfrac{k}{n}(x - \eta x^\delta/\sqrt{k})\right) \ge \tfrac{k}{n}x \\ &\iff \sqrt{k}\left\{\tfrac{n}{k}\Gamma_n\left(\tfrac{k}{n}(x - \eta x^\delta/\sqrt{k})\right) - (x - \eta x^\delta/\sqrt{k})\right\} \ge \eta x^\delta \\ &\iff \frac{w_n(x - \eta x^\delta/\sqrt{k})}{x^\delta} \ge \eta.\end{aligned}$$

If $x < \eta x^\delta \sqrt{k}$, then the inequality $v_n(x)/x^\delta \le -\eta$ is impossible, since $Q_n \ge 0$

and thus $v_n(x) \geq -x\sqrt{k}$. It follows that

$$\begin{aligned}\mathsf{P}\left[\inf_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} < -\eta\right] &\leq \mathsf{P}\left[\inf_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} \leq -\eta\right] \\ &= \mathsf{P}\left[\sup_{\substack{x\in((2k)^{-1},a] \\ x\geq \eta x^\delta\sqrt{k}}} \frac{w_n(x-\eta x^\delta/\sqrt{k})}{x^\delta} \geq \eta\right].\end{aligned}$$

Since $x - \eta x^\delta/\sqrt{k} \leq x$, the weight $1/x^\delta$ can only increase if we replace it by $1/(x-\eta x^\delta/\sqrt{k})^\delta$. But then

$$\sup_{\substack{x\in((2k)^{-1},a] \\ x\geq \eta x^\delta\sqrt{k}}} \frac{w_n(x-\eta x^\delta/\sqrt{k})}{x^\delta} \leq \sup_{z\in[0,a]} \frac{w_n(z)}{z^\delta}.$$

Consequently, it follows again from our results on the weighted tail empirical process that we may find $N_4 \in \mathbb{N}$ and $a_3 \in (0,1)$ such that for $n \geq \mathbb{N}_4$ and $0 < a < a_3$,

$$\mathsf{P}\left[\inf_{x\in((2k)^{-1},a]} \frac{v_n(x)}{x^\delta} < -\eta\right] < \frac{\epsilon}{6}.$$

Combining everything, we have for $n \geq \max\{N_1,\ldots,N_4\}$ and $0 < a < \min\{a_1,a_2,a_3\}$ in (4.42),

$$\mathsf{P}\left[\sup_{x\in[0,M]} \frac{|v_n(x)+w_n(x)|}{x^\delta} > \eta\right] < \epsilon.$$

This concludes the proof. □

4.B Link between densities of Λ and Φ_p

For any $x, y > 0$, we have

$$\lambda(x,y) = \frac{\partial^2}{\partial x\partial y}\Lambda\left((0,x]\times(0,y]\right) = \frac{\partial^2}{\partial x\partial y}\nu\left([x^{-1},\infty)\times[y^{-1},\infty)\right),$$

where ν factorizes in polar coordinates under

$$T : x \in \mathbb{E} \mapsto T(x) \stackrel{\text{def}}{=} \left(\|x\|_p, \arctan(x_1/x_2)\right) \in (0,\infty)\times[0,\pi/2],$$

i.e.,

$$\mathrm{d}(\nu\circ T^{-1})(r,\theta) = \mathrm{d}\nu_{-1}(r)\otimes \mathrm{d}\Phi_p(\theta),$$

where $\mathrm{d}\nu_{-1}(r) \stackrel{\text{def}}{=} \mathrm{d}r/r^2$ for any $r > 0$. Hence,

$$\begin{aligned}
&\nu\left([x^{-1},\infty)\times[y^{-1},\infty)\right)\\
&= \left(\nu_{-1}\otimes\Phi_p\right)\left(\left\{T(u,v) : u \geq x^{-1}, v \geq y^{-1}\right\}\right)\\
&= \left(\nu_{-1}\otimes\Phi_p\right)\left(\left\{(r,\theta) : \frac{r\sin\theta}{\|(\sin\theta,\cos\theta)\|_p} \geq x^{-1}, \frac{r\cos\theta}{\|(\sin\theta,\cos\theta)\|_p} \geq y^{-1}\right\}\right)\\
&= \int_0^{\pi/2}\left(\frac{x\sin\theta}{\|(\sin\theta,\cos\theta)\|_p} \wedge \frac{y\cos\theta}{\|(\sin\theta,\cos\theta)\|_p}\right)\varphi_p(\theta)\mathrm{d}\theta\\
&= x\int_0^{\arctan(y/x)}\frac{\sin\theta}{\|(\sin\theta,\cos\theta)\|_p}\varphi_p(\theta)\mathrm{d}\theta\\
&\qquad + y\int_{\arctan(y/x)}^{\pi/2}\frac{\cos\theta}{\|(\sin\theta,\cos\theta)\|_p}\varphi_p(\theta)\mathrm{d}\theta.
\end{aligned}$$

Differentiating this expression with respect to x and y using the Leibniz differentiation rule yields the desired formula (4.23) for λ, showing that its continuity on $(0,\infty)^2$ is equivalent to the one of φ_p on $(0,\pi/2)$.

4.C Proof of Theorem 4.3.1

The key idea behind the proof is the following decomposition that can be found in [42, 47] or more recently in [25]. For each $\theta \in [0,\pi/2]$,

$$\begin{aligned}
\sqrt{k}\left(\widehat{\Phi}_p(\theta) - \Phi_p(\theta)\right) &= \sqrt{k}\left(\tfrac{n}{k}P_n\left(\tfrac{k}{n}\hat{C}_{p,\theta}\right) - \tfrac{n}{k}P\left(\tfrac{k}{n}\hat{C}_{p,\theta}\right)\right)\\
&\quad + \sqrt{k}\left(\tfrac{n}{k}P\left(\tfrac{k}{n}\hat{C}_{p,\theta}\right) - \Lambda\left(\hat{C}_{p,\theta}\right)\right)\\
&\quad + \sqrt{k}\left(\Lambda\left(\hat{C}_{p,\theta}\right) - \Lambda\left(C_{p,\theta}\right)\right)\\
&\stackrel{\text{def}}{=} V_{n,p}(\theta) + r_{n,p}(\theta) + Z_{n,p}(\theta),
\end{aligned} \tag{4.45}$$

where the set $C_{p,\theta}$ is estimated by

$$\begin{aligned}
\hat{C}_{p,\theta} \stackrel{\text{def}}{=} \Big\{(x,y) : 0 \leq x \leq \infty, 0 \leq &y \leq \tfrac{n}{k}Q_{2n}\left(\Gamma_{1n}\left(\tfrac{k}{n}x\right)\tan\theta\right),\\
&y \leq \tfrac{n}{k}Q_{2n}\left(\tfrac{k}{n}y_p\left(\tfrac{n}{k}\Gamma_{1n}\left(\tfrac{k}{n}x\right)\right)\right)\Big\}.
\end{aligned} \tag{4.46}$$

The three terms in the decomposition (4.45) will be called the *stochastic term*, *bias term*, and *random set term*, respectively.

Outline of the proof. In Section 4.C.2, we state the main underlying weak convergence result needed for the proof, based on Proposition 4.3.1. Unlike [42, 47], we prove our asymptotic expansion without requiring a Skorokhod construction but work with the original sample space and random

variables directly. Technically, it implies that we have to work with weak convergence instead of the almost sure representations arising from the Skorokhod construction, which makes some continuity arguments a bit more delicate, but the proof becomes more direct. In Section 4.C.3, we treat each term of (4.45) separately. Before starting the main proof, we argue that we can reduce the problem by only showing the result for $\theta \in [0, \pi/4]$ in Section 4.C.1.

4.C.1 Reduction to $\theta \in [0, \pi/4]$

Let $S : (x, y) \in \mathbb{R}^2 \mapsto (y, x) \in \mathbb{R}^2$ be the symmetry map. By extension, we also define the set-valued map $S(A) \stackrel{\text{def}}{=} \{S(x, y) : (x, y) \in A\}$ for $A \subseteq \mathbb{R}^2$. Clearly, the measures $P_n \circ S^{-1}$ and $P \circ S^{-1}$ underlying the statement of the theorem satisfy Assumptions 4.3.1 and 4.3.2 with respect to the measure $\Lambda_S \stackrel{\text{def}}{=} \Lambda \circ S^{-1}$, which has density $\lambda_S \stackrel{\text{def}}{=} \lambda \circ S$. The smoothness assumption is trivial. For the bias assumption, letting $c_S \stackrel{\text{def}}{=} c \circ S$ denote the density of $P \circ S^{-1}$, it suffices to note that $S(\mathcal{L}_T) = \mathcal{L}_T$ for any $T \geq 1$ and thus

$$\iint_{\mathcal{L}_T} |t c_S(tu_1, tu_2) - \lambda_S(u_1, u_2)| \, \mathrm{d}u_1 \mathrm{d}u_2 = \mathcal{D}_T(t),$$

for any $t > 0$. We may thus apply our result for $P_n \circ S^{-1}$ and $P \circ S^{-1}$ for $\theta \in [0, \pi/4]$ once it has been proved for P_n and P.

Now consider $\theta \in [\pi/4, \pi/2]$. Then, up to sets of Hausdorff dimension 1, sets which are asymptotically negligible for our purposes, we have

$$C_{p,\theta} = C_{p,\pi/4} \cup \left(S(C_{p,\pi/4}) \setminus S(C_{p,\pi/2-\theta}) \right). \tag{4.47}$$

Consequently, letting $\widehat{\Phi}_p^S$ denote the empirical angular measure associated with $P_n \circ S^{-1}$ and Φ_p^S the angular measure associated with $P \circ S^{-1}$, we have for such $\theta \in [\pi/4, \pi/2]$ that asymptotically, by continuity of the limiting process,

$$\begin{aligned}\sqrt{k}\left(\widehat{\Phi}_p(\theta) - \Phi_p(\theta)\right) = \sqrt{k}\left(\widehat{\Phi}_p(\pi/4) - \Phi_p(\pi/4)\right) + \sqrt{k}\left(\widehat{\Phi}_p^S(\pi/4) - \Phi_p^S(\pi/4)\right) \\ - \sqrt{k}\left(\widehat{\Phi}_p^S(\pi/2-\theta) - \Phi_p^S(\pi/2-\theta)\right) + \mathrm{o_P}(1),\end{aligned}$$

where the $\mathrm{o_P}(1)$ term is uniform in $\theta \in [\pi/4, \pi/2]$. Assuming that the theorem has been shown to hold for $\theta \in [0, \pi/4]$, if $E_{n,p}^S(\theta)$ denotes the asymptotic expansion obtained on the basis of the measures $P_n \circ S^{-1}$ and $P \circ S^{-1}$, we have for $\theta \in [\pi/4, \pi/2]$,

$$\sqrt{k}\left(\widehat{\Phi}_p(\theta) - \Phi_p(\theta)\right) = E_{n,p}(\pi/4) + E_{n,p}^S(\pi/4) - E_{n,p}^S(\pi/2-\theta) + \mathrm{o_P}(1),$$

where the $o_P(1)$ term is uniform in $\theta \in [\pi/4, \pi/2]$. It is now sufficient to show that

$$E_{n,p}(\pi/4) + E^S_{n,p}(\pi/4) - E^S_{n,p}(\pi/2-\theta) = E_{n,p}(\theta), \qquad \theta \in [\pi/4, \pi/2]. \quad (4.48)$$

To do so, we expand $E_{n,p}$ and $E^S_{n,p}$ into three terms as in Theorem 4.3.1. The expansions of $E_{n,p}(\pi/4)$ and $E_{n,p}(\theta)$ are provided in the theorem. We compute the expansions of $E^S_{n,p}(\pi/4)$ and $E^S_{n,p}(\pi/2-\theta)$.

For $E^S_{n,p}(\pi/4)$, the first term in the expansion is trivially given by

$$\sqrt{k}\left(\tfrac{n}{k}P_n\left(\tfrac{k}{n}S(C_{p,\pi/4})\right) - \tfrac{n}{k}P\left(\tfrac{k}{n}S(C_{p,\pi/4})\right)\right).$$

For the second term in the expansion, the integral over $x < x_p(\pi/4)$ becomes

$$\int_0^{2^{1/p}} \lambda_S(x,x)\{w_{2n}(x) - w_{1n}(x)\}\mathrm{d}x = -\int_0^{2^{1/p}} \lambda(x,x)\{w_{1n}(x) - w_{2n}(x)\}\mathrm{d}x,$$

since $\tan(\pi/4) = \cot(\pi/4) = 1$ and the measure $S_\# P_n$ has first marginal tail empirical process w_{2n} and second marginal tail empirical process w_{1n}. We see that it equals the opposite of the integral over $x < x_p(\pi/4)$ in the expansion of $E_{n,p}(\pi/4)$ so that they will cancel out. The third term in the expansion is an integral over $x \geq x_p(\pi/4)$, for which we only consider the more complicated case $p < \infty$. It equals

$$\int_{2^{1/p}}^{\infty} \lambda_S(x, y_p(x)) \left\{w_{2n}(x)y_p'(x) - w_{1n}(y_p(x))\right\} \mathrm{d}x$$
$$= \int_1^{2^{1/p}} \lambda(z, y_p(z)) \left\{w_{1n}(z)y_p'(z) - w_{2n}(y_p(z))\right\} \mathrm{d}z,$$

where the second line follows from the change of variable $x = y_p(z)$ which is an involution and satisfies $y_p'(y_p(z))y_p'(z) = 1$.

For $E^S_{n,p}(\pi/2-\theta)$, the reasoning is similar. The first term in the expansion is

$$\sqrt{k}\left(\tfrac{n}{k}P_n\left(\tfrac{k}{n}S\big(C_{p,\pi/2-\theta}\big)\right) - \tfrac{n}{k}P\left(\tfrac{k}{n}S\big(C_{p,\pi/2-\theta}\big)\right)\right).$$

Since $\tan(\pi/2-\theta) = \cot(\theta)$, the second term in the expansion is

$$\int_0^{x_p(\pi/2-\theta)} \lambda_S(x, x\tan(\pi/2-\theta))\{w_{2n}(x)\tan(\pi/2-\theta) - w_{1n}(x\tan(\pi/2-\theta))\}\mathrm{d}x$$
$$= -\int_0^{x_p(\theta)} \lambda(y, y\tan\theta)\left\{w_{1n}(y)\tan\theta - w_{2n}(y\tan\theta)\right\}\mathrm{d}y.$$

The third term is an integral over $x \geq x_p(\pi/2 - \theta)$. Again assuming $p < \infty$, using the same change of variable as before and the fact that $y_p(x_p(\pi/2-\theta)) = x_p(\theta)$, we get

$$\int_1^{x_p(\theta)} \lambda(z, y_p(z)) \left\{ w_{1n}(z) y_p'(z) - w_{2n}(y_p(z)) \right\} \mathrm{d}z.$$

Combining everything, we find (4.48). Consequently, once the asymptotic expansion stated in the theorem has been shown to hold on the region $\theta \in [0, \pi/4]$, it holds on $\theta \in [0, \pi/2]$.

4.C.2 Weak convergence of a tail empirical process

As in [47], the main idea of the proof is to approximate the random sets $\hat{C}_{p,\theta}$ by those in a carefully constructed VC class of sets built from a covering of the positive real line by (small) intervals on which the random perturbations of the original sets $C_{p,\theta}$ are manageable. Fix any $\Delta \in \{1, 1/2, 1/3, \ldots\}$ and $M > 1$ and let $\mathcal{A} = \mathcal{A}(\Delta, M)$ be the following VC class of sets, already introduced in [47]. For $m \in \{0, 1, 2, \ldots, 1/\Delta - 1\}$, define the intervals

$$I_\Delta(m, \theta) \stackrel{\text{def}}{=} \begin{cases} [m\Delta x_p(\theta), (m+1)\Delta x_p(\theta)] & \text{if } \theta \in (0, \pi/4], \\ [0, \infty) & \text{if } \theta = 0 \text{ and } m = 0, \\ \emptyset & \text{if } \theta = 0 \text{ and } m > 0. \end{cases} \tag{4.49}$$

and

$$J_\Delta(m) \stackrel{\text{def}}{=} [y_p(1 + (2^{1/p} - 1)(m+1)\Delta), y_p(1 + (2^{1/p} - 1)m\Delta)]. \tag{4.50}$$

Set $\tilde{\mathcal{A}}$ to be the class of sets containing the following sets:

- $\bigcup_{m=0}^{1/\Delta-1} \left\{(x, y) : x \in I_\Delta(m, \theta), 0 \leq y \leq x \tan\theta + B_m (x \tan\theta)^{1/16}\right\}$ for some $\theta \in [0, \pi/4]$ and $B_0, \ldots, B_{1/\Delta-1} \in [-1, 1]$.
- $\bigcup_{m=0}^{1/\Delta-1} \left\{(x, y) : x \in J_\Delta(m), x \geq x_p(\theta), 0 \leq y \leq y_p(x)(1 + K_m)\right\}$ for some $\theta \in [0, \pi/4]$ and $K_0, \ldots, K_{1/\Delta-1} \in [-1/2, 1/2]$.
- $\{(x, y) : x \leq a\}$, $\{(x, y) : y \leq a\}$ and $\{(x, y) : x \leq a \text{ or } y \leq a\}$ for some $a \in [0, M]$.

Next, define $\tilde{\mathcal{A}}_s \stackrel{\text{def}}{=} \{A_s : A \in \tilde{\mathcal{A}}\}$ where for $A \in \tilde{\mathcal{A}}$, $A_s \stackrel{\text{def}}{=} \{(x, y) : (y, x) \in A\}$. Finally, let $\mathcal{A} = \tilde{\mathcal{A}} \cup \tilde{\mathcal{A}}_s$.

As required in Proposition 4.3.1, every set in $\mathcal{A}$ is contained in $L_M = \{x \in [0, \infty)^d : \min(x_1, \ldots, x_d) \leq M\}$, provided $M > 1$ is picked large enough—which is not restrictive since we will consider it large later. Furthermore, the

other conditions underlying Proposition 4.3.1 are easily shown to hold in view of the assumptions. In particular, the uniform regular variation condition (4.18) is satisfied thanks to the convergence in total variation in Assumption 4.3.2. Consequently, we can conclude that (4.19) and (4.20) hold in this setting.

4.C.3 Asymptotic expansions of the three terms

The stochastic term $V_{n,p}(\theta)$. Recall $W_{n,k}$ in Proposition 4.3.1 and note that $V_{n,p}(\theta) = W_{n,k}(\hat{C}_{p,\theta})$. Our aim is to show that, as $n \to \infty$,

$$\sup_{\theta\in[0,\pi/4]} \left|W_{n,k}(\hat{C}_{p,\theta}) - W_{n,k}(C_{p,\theta})\right| = \mathrm{o_P}(1). \tag{4.51}$$

We will consider separately $V_{n,p,1}$ and $V_{n,p,2}$, where $V_{n,p,i}(\theta) \overset{\text{def}}{=} W_{n,k}(\hat{C}_{p,\theta,i})$ with $\hat{C}_{p,\theta,1} \overset{\text{def}}{=} \hat{C}_{p,\theta} \cap \{(x,y) : x < x_p(\theta)\}$ and $\hat{C}_{p,\theta,2} \overset{\text{def}}{=} \hat{C}_{p,\theta} \cap \{(x,y) : x \geq x_p(\theta)\}$. In particular, it would follow from the triangle inequality that the expansion (4.51) is satisfied, provided that

$$\sup_{\theta\in[0,\pi/4]} \left|V_{n,p,i}(\theta) - W_{n,k}(C_{p,\theta,i})\right| = \mathrm{o_P}(1), \qquad i \in \{1,2\}, \tag{4.52}$$

where $C_{p,\theta,1} \overset{\text{def}}{=} C_{p,\theta} \cap \{(x,y) : x < x_p(\theta)\}$ and $C_{p,\theta,2} \overset{\text{def}}{=} C_{p,\theta} \cap \{(x,y) : x \geq x_p(\theta)\}$.

Let us first consider $i = 1$. Recall equation (4.46) and observe that, since $\frac{n}{k}\Gamma_{jn}(\frac{k}{n}x) = x + w_{jn}(x)/\sqrt{k}$ and $\frac{n}{k}Q_{jn}(\frac{k}{n}x) = x + v_{jn}(x)/\sqrt{k}$,

$$\frac{n}{k}Q_{2n}\left(\Gamma_{1n}\left(\frac{k}{n}x\right)\tan\theta\right) = x\tan\theta + \frac{z_{n,\theta}(x)}{\sqrt{k}},$$
$$\frac{n}{k}Q_{2n}\left(\frac{k}{n}y_p\left(\frac{n}{k}\Gamma_{1n}\left(\frac{k}{n}x\right)\right)\right) = y_p(x) + \frac{s_{n,p}(x)}{\sqrt{k}},$$

where

$$z_{n,\theta}(x) \overset{\text{def}}{=} w_{1n}(x)\tan\theta + v_{2n}\left(x\tan\theta + \frac{w_{1n}(x)\tan\theta}{\sqrt{k}}\right), \tag{4.53}$$

$$s_{n,p}(x) \overset{\text{def}}{=} \sqrt{k}\left\{y_p\left(x + \frac{w_{1n}(x)}{\sqrt{k}}\right) - y_p(x)\right\} + v_{2n}\left(y_p\left(x + \frac{w_{1n}(x)}{\sqrt{k}}\right)\right). \tag{4.54}$$

Now set, for any $m \in \{0,1,2,\ldots,1/\Delta - 1\}$, with the convention that $0/0 = 0$,

$$V^{+}_{m,\Delta,\theta} \overset{\text{def}}{=} \sup_{x\in I_\Delta(m,\theta)} \frac{z_{n,\theta}(x) \wedge \left(s_{n,p}(x) + \sqrt{k}(y_p(x) - x\tan\theta)\right)}{(x\tan\theta)^{1/16}},$$

$$V^{-}_{m,\Delta,\theta} \overset{\text{def}}{=} \inf_{x\in I_\Delta(m,\theta)} \frac{z_{n,\theta}(x) \wedge \left(s_{n,p}(x) + \sqrt{k}(y_p(x) - x\tan\theta)\right)}{(x\tan\theta)^{1/16}},$$

where the intervals $I_\Delta(m,\theta)$ were defined in (4.49). For fixed θ, note that these intervals cover $[0, x_p(\theta)]$, on which $y_p(x) \geq x\tan\theta$. It follows that on an event whose probability tends to one, $V^{\pm}_{m,\Delta,\theta}$ is equal to the same expression as in its definition but with the numerator simplified to $z_{n,\theta}(x)$. In the following, we work with these simplified definitions. Moreover, the weak convergence

$$\left\{\frac{|z_{n,\theta}(x)|}{(x\tan\theta)^{1/16}} : \theta\in[0,\pi/4], x\in[0,x_p(\theta)]\right\} \xrightarrow{\mathcal{L}} \left\{\frac{W_1(x)\tan\theta - W_2(x\tan\theta)}{(x\tan\theta)^{1/16}} : \theta\in[0,\pi/4], x\in[0,x_p(\theta)]\right\} \quad (4.55)$$

holds in the Banach space of bounded functions on the indicated domain. To see (4.55), recall the definition of $z_{n,\theta}$ in (4.53) and note that for any $\theta\in[0,\pi/4]$,

$$\frac{z_{n,\theta}(x)}{(x\tan\theta)^{1/16}} = \frac{w_{1n}(x)}{x^{1/16}}(\tan\theta)^{15/16} + \frac{v_{2n}\left(x\tan\theta + \frac{w_{1n}(x)\tan\theta}{\sqrt{k}}\right)}{\left(x\tan\theta + \frac{w_{1n}(x)\tan\theta}{\sqrt{k}}\right)^{1/4}}\left((x\tan\theta)^{3/4} + \frac{(\tan\theta)^{3/4}}{\sqrt{k}}\frac{w_{1n}(x)}{x^{1/4}}\right)^{1/4}.$$

Using equations (4.21) and (4.22) in Proposition 4.3.2 together with the fact that $x\tan\theta \leq 2^{1/p}$ on the said domain yields (4.55). In particular, we have

$$\sup_{\theta\in[0,\pi/4]}\sup_{x\in[0,x_p(\theta)]}\frac{|z_{n,\theta}(x)|}{(x\tan\theta)^{1/16}} = O_P(1), \qquad n\to\infty. \quad (4.56)$$

Defining

$$M^{\pm}_{\Delta,\theta} \stackrel{\text{def}}{=} \bigcup_{m=0}^{\Delta-1}\left\{(x,y) : x\in I_\Delta(m,\theta), 0\leq y\leq x\tan\theta + \frac{V^{\pm}_{m,\Delta,\theta}}{\sqrt{k}}(x\tan\theta)^{1/16}\right\},$$

it is easy to see that for any $\theta\in[0,\pi/4]$ we have $M^{-}_{\Delta,\theta}\subseteq\hat{C}_{\theta,p,1}\subseteq M^{+}_{\Delta,\theta}$ and thus

$$V^{-}_{n,\Delta,1}(\theta) - R_{n,\Delta,1}(\theta) \leq V_{n,p,1}(\theta) \leq V^{+}_{n,\Delta,1}(\theta) + R_{n,\Delta,1}(\theta), \quad (4.57)$$

where

$$V^{\pm}_{n,\Delta,1}(\theta) \stackrel{\text{def}}{=} W_{n,k}(M^{\pm}_{\Delta,\theta}), \quad (4.58)$$

$$R_{n,\Delta,1}(\theta) \stackrel{\text{def}}{=} \sqrt{k}\tfrac{n}{k}P\left(\tfrac{k}{n}(M^{+}_{\Delta,\theta}\setminus M^{-}_{\Delta,\theta})\right). \quad (4.59)$$

Observe that, due to Assumption 4.3.2 and the behavior of the marginal tail empirical and quantile processes,

$$\sup_{\theta\in[0,\pi/4]}\left|R_{n,\Delta,1}(\theta) - \sqrt{k}\Lambda(M^{+}_{\Delta,\theta}\setminus M^{-}_{\Delta,\theta})\right| = o_P(1).$$

Similar computations as in [42, Section 3.1.1] show that

$$\sqrt{k}\Lambda(M^+_{\Delta,\theta} \setminus M^-_{\Delta,\theta}) \leq 16 \times 2^{1/p} \sup_{y\geq 0} \lambda(1,y) \max_{m=0,\ldots,\Delta-1} (V^+_{m,\Delta,\theta} - V^-_{m,\Delta,\theta}),$$

where $\sup_{y\geq 0} \lambda(1,y) < \infty$ since λ is continuous on $[0,\infty)^2 \setminus \{(0,0)\}$ thanks to Assumption 4.3.1 and $\lim_{y\to\infty} \lambda(1,y) = 0$ (the latter limit following from the homogeneity of λ in (4.24)). Since $V^+_{m,\Delta,0} = V^-_{m,\Delta,0} = 0$ by convention for any value of m and Δ, we deduce from (4.55) that, for every $\epsilon > 0$,

$$\lim_{\Delta\to 0} \limsup_{n\to\infty} \mathrm{P}^* \left[\sup_{\theta\in[0,\pi/4]} \max_{m=0,\ldots,\Delta-1} (V^+_{m,\Delta,\theta} - V^-_{m,\Delta,\theta}) > \epsilon \right] = 0.$$

Collecting the latter three equations, we conclude that, for every $\epsilon > 0$,

$$\lim_{\Delta\to 0} \limsup_{n\to\infty} \mathrm{P}^* \left[\sup_{\theta\in[0,\pi/4]} R_{n,\Delta,1}(\theta) > \epsilon \right] = 0. \tag{4.60}$$

Furthermore, since (4.56) implies that on an event with probability tending to one we have $M^\pm_{\Delta,\theta} \in \mathcal{A}$ for all $\theta \in [0,\pi/4]$, it follows that for all $\epsilon, \delta > 0$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{P}^* \left[\max_{\sigma\in\{-,+\}} \sup_{\theta\in[0,\pi/4]} \left| W_{n,k}(M^\sigma_{\Delta,\theta}) - W_{n,k}(C_{p,\theta,1}) \right| > \epsilon \right] \\
&\qquad \leq \limsup_{n\to\infty} \mathrm{P}^* \left[\max_{\sigma\in\{-,+\}} \sup_{\theta\in[0,\pi/4]} \rho(M^\sigma_{\Delta,\theta}, C_{p,\theta,1}) \geq \delta \right] \\
&\qquad + \limsup_{n\to\infty} \mathrm{P}^* \left[\sup_{A,A'\in\mathcal{A},\rho(A,A')<\delta} \left| W_{n,k}(A) - W_{n,k}(A') \right| > \epsilon \right].
\end{aligned} \tag{4.61}$$

Standard computations lead to

$$\begin{aligned}
&\sup_{\theta\in[0,\pi/4]} \rho(M^\pm_{\Delta,\theta}, C_{p,\theta,1}) \\
&\qquad \leq \frac{16 \times 2^{1/p}}{\sqrt{k}} \sup_{y\geq 0} \lambda(1,y) \sup_{\theta\in[0,\pi/4]} \max_{m=0,\ldots,\Delta-1} |V^\pm_{m,\Delta,\theta}| = \mathrm{o_P}(1),
\end{aligned}$$

so that the first limsup on the right hand side of (4.61) is zero. The second limsup can be made arbitrary close to zero as $\delta \downarrow 0$ by (4.20).

Combining everything leads to (4.52) for $i = 1$. Indeed, for $\epsilon > 0$, we have

$$\mathrm{P}^*\left[\sup_{\theta\in[0,\pi/4]}\left|V_{n,p,1}(\theta) - W_{n,k}(C_{p,\theta,1})\right| > \epsilon\right] \leq \sum_{\sigma\in\{-,+\}} \mathrm{P}^*\left[\sup_{\theta\in[0,\pi/4]}\left|V^{\sigma}_{n,\Delta,1}(\theta) - W_{n,k}(C_{p,\theta,1})\right| > \epsilon/2\right] + 2\mathrm{P}^*\left[\sup_{\theta\in[0,\pi/4]} R_{n,\Delta,1}(\theta) > \epsilon/2\right].$$

The first term on the right-hand side vanishes as $n \to \infty$, while the limsup over $n \to \infty$ of the second term can be made arbitrarily small as $\Delta \to 0$.

Let us now consider the case $i = 2$ in (4.52). Similarly to the case $i = 1$, it will be useful to note that for any $\theta \in [0, \pi/4]$,

$$\hat{C}_{p,\theta,2} = \left\{(x,y) : x \geq x_p(\theta), 0 \leq y \leq \left(x\tan\theta + \frac{z_{n,\theta(x)}}{\sqrt{k}}\right) \wedge \left(y_p(x) + \frac{s_{n,p}(x)}{\sqrt{k}}\right)\right\},$$

with $z_{n,\theta}$ and $s_{n,p}$ the processes defined in (4.53) and (4.54) respectively. For each $m \in \{0, 1, 2, \ldots, 1/\Delta - 1\}$, we let

$$S^{+}_{m,\Delta,\theta} \overset{\text{def}}{=} \sup_{x\in J_\Delta(m)} \frac{s_{n,p}(x) \wedge \left(z_{n,\theta}(x) + \sqrt{k}(x\tan\theta - y_p(x))\right)}{y_p(x)},$$

$$S^{-}_{m,\Delta,\theta} \overset{\text{def}}{=} \inf_{x\in J_\Delta(m)} \frac{s_{n,p}(x) \wedge \left(z_{n,\theta}(x) + \sqrt{k}(x\tan\theta - y_p(x))\right)}{y_p(x)},$$

where the intervals $J_\Delta(m)$ were defined in (4.50). These intervals cover $[2^{1/p}, \infty]$, on which $x\tan\theta \geq x_p(\theta)$ for any $\theta \in [0, \pi/4]$. It follows that on an event whose probability tends to one, $S^{\pm}_{m,\Delta,\theta}$ is equal to the same expression as in its definition but with the numerator simplified to $s_{n,p}(x)$. In the following, we work with these simplified definitions. Note that, by (4.21) and (4.22), the mean value theorem and the extended continuous mapping theorem [122, Theorem 1.11.1], the weak convergence

$$\left\{s_{n,p}(x) : x \in [2^{1/p}, \infty)\right\} \xrightarrow{\mathcal{L}} \left\{W_1(x)y_p'(x) - W_2(y_p(x)) : x \in [2^{1/p}, \infty)\right\} \tag{4.62}$$

holds in $\ell^\infty([2^{1/p}, \infty))$. In particular, since $y_p(x) \geq 1$ when $x \geq 2^{1/p}$, we have

$$\sup_{x\in[2^{1/p},\infty]} \frac{|s_{n,p}(x)|}{y_p(x)} = \mathrm{O_P}(1). \tag{4.63}$$

Defining

$$N^{\pm}_{\Delta,\theta} \stackrel{\text{def}}{=} \bigcup_{m=0}^{\Delta-1} \left\{ (x,y) : x \geq x_p(\theta), x \in J_\Delta(m), 0 \leq y \leq y_p(x) \left(1 + \frac{S^{\pm}_{m,\Delta,\theta}}{\sqrt{k}}\right) \right\},$$

we see that for any $\theta \in [0, \pi/4]$ we have $N^{-}_{\Delta,\theta} \subseteq \hat{C}_{\theta,p,2} \subseteq N^{+}_{\Delta,\theta}$ and thus

$$V^{-}_{n,\Delta,2}(\theta) - R_{n,\Delta,2}(\theta) \leq V_{n,p,2}(\theta) \leq V^{+}_{n,\Delta,2}(\theta) + R_{n,\Delta,2}(\theta), \tag{4.64}$$

where

$$V^{\pm}_{n,\Delta,2}(\theta) \stackrel{\text{def}}{=} W_{n,k}(N^{\pm}_{\Delta,\theta}), \tag{4.65}$$

$$R_{n,\Delta,2}(\theta) \stackrel{\text{def}}{=} \sqrt{k}\tfrac{n}{k} P\left(\tfrac{k}{n}(N^{+}_{\Delta,\theta} \setminus N^{-}_{\Delta,\theta})\right). \tag{4.66}$$

Due to Assumption 4.3.2 and the behavior of the marginal tail empirical and quantile processes,

$$\sup_{\theta\in[0,\pi/4]} \left| R_{n,\Delta,2}(\theta) - \sqrt{k}\Lambda(N^{+}_{\Delta,\theta} \setminus N^{-}_{\Delta,\theta}) \right| = \mathrm{o_P}(1).$$

Computations in [47, Paragraph A.1] show that, with probability tending to one,

$$\sup_{\theta\in[0,\pi/4]} \sqrt{k}\Lambda(N^{+}_{\Delta,\theta} \setminus N^{-}_{\Delta,\theta}) \leq 3 \sup_{\theta\in[0,\pi/4]} \max_{m=0,\ldots,\Delta-1} (S^{+}_{m,\Delta,\theta} - S^{-}_{m,\Delta,\theta}).$$

In view of (4.62), we have, for any $\epsilon > 0$,

$$\lim_{\Delta\to 0} \limsup_{n\to\infty} \mathrm{P}^* \left[\sup_{\theta\in[0,\pi/4]} \max_{m=0,\ldots,\Delta-1} (S^{+}_{m,\Delta,\theta} - S^{-}_{m,\Delta,\theta}) > \epsilon \right] = 0,$$

which implies (4.60) with $R_{n,\Delta,1}$ replaced by $R_{n,\Delta,2}$.

By a similar argument as for $i = 1$, we have

$$\sup_{\theta\in[0,\pi/4]} \rho(N^{\pm}_{\Delta,\theta}, C_{p,\theta,2}) \leq \frac{3}{\sqrt{k}} \sup_{\theta\in[0,\pi/4]} \max_{m=0,\ldots,\Delta-1} |S^{\pm}_{m,\Delta,\theta}| = \mathrm{o_P}(1),$$

and thus (4.52) holds for $i = 2$.

The bias term $r_{n,p}(\theta)$. Note that all sets $\hat{C}_{p,\theta}$ for $\theta \in [0, \pi/4]$ are contained in the set $\hat{C}_{p,\pi/4}$ and, with probability tending to one, the latter is contained

in the set $\{(x,y) : 0 \le x \wedge y \le M\}$ for $M = M(p) = 2^{1/p}$. It follows that, with probability tending to one,

$$\begin{aligned}
\sup_{\theta\in[0,\pi/4]} \left|r_{n,p}(\theta)\right| &\le \sqrt{k} \iint_{0\le x\wedge y\le M} \left|\frac{k}{n} c\left(\frac{k}{n}x, \frac{k}{n}y\right) - \lambda(x,y)\right| \mathrm{d}x\mathrm{d}y \\
&= \sqrt{k}M \iint_{0\le u\wedge v\le 1} \left|\frac{k}{n} Mc\left(\frac{k}{n}Mu, \frac{k}{n}Mv\right) - \lambda(u,v)\right| \mathrm{d}u\mathrm{d}v \\
&= \sqrt{k}M\mathcal{D}_{n/k}\left(\frac{k}{n}M\right) + \sqrt{k}M \iint_{\substack{0\le u\wedge v\le 1 \\ u\vee v > \frac{n}{k}}} \frac{k}{n} Mc\left(\frac{k}{n}Mu, \frac{k}{n}Mv\right) \mathrm{d}u\mathrm{d}v \\
&\quad + \sqrt{k}M \iint_{\substack{0\le u\wedge v\le 1 \\ u\vee v > \frac{n}{k}}} \lambda(u,v)\mathrm{d}u\mathrm{d}v.
\end{aligned}$$

By the first part of Assumption 4.3.2, the first term in the right-hand side vanishes asymptotically as $n \to \infty$. The second integral actually equals zero for any value of $n \in \mathbb{N}$ since the copula density c vanishes outside $[0,1]^2$. For the last integral, assuming n large enough so that $n/k > 1$, the domain of integration is the union of two distinct parts given by $\{(u,v) : 0 \le u \le 1, v > n/k\}$ and $\{(u,v) : u > n/k, 0 \le v \le 1\}$. For the first part, we have

$$M\sqrt{k} \iint_{\substack{0\le u\le 1 \\ v > n/k}} \lambda(u,v)\mathrm{d}u\mathrm{d}v \le M\sqrt{k}\left(\Phi_p(\tfrac{\pi}{2}) - \Phi_p(\tfrac{\pi}{2} - \tfrac{k}{n})\right),$$

while for the second part, we have

$$M\sqrt{k} \iint_{\substack{u > n/k \\ 0\le v\le 1}} \lambda(u,v)\mathrm{d}u\mathrm{d}v \le M\sqrt{k}\Phi_p(\tfrac{k}{n}).$$

By the second part of Assumption 4.3.2, those two quantities converge to zero as $n \to \infty$.

The random set term $Z_{n,p}(\theta)$. Similarly as for the stochastic term, we work on

$$Z_{n,p,i}(\theta) = \sqrt{k}\left(\Lambda\left(\hat{C}_{p,\theta,i}\right) - \Lambda\left(C_{p,\theta,i}\right)\right)$$

separately for $i = 1, 2$. In particular, our aim is to show that, as $n \to \infty$,

$$\sup_{\theta\in[0,\pi/4]} \left|Z_{n,p,1}(\theta) - \int_0^{x_p(\theta)} \lambda(x, x\tan\theta)\left\{w_{1n}(x)\tan\theta - w_{2n}(x\tan\theta)\right\}\mathrm{d}x\right| = o_P(1) \quad (4.67)$$

and, for $p < \infty$ (since this is the most involved case),

$$\sup_{\theta\in[0,\pi/4]}\left|Z_{n,p,2}(\theta) - \int_{x_p(\theta)}^{\infty} \lambda(x, y_p(x))\left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\}\mathrm{d}x\right| = o_P(1). \quad (4.68)$$

We start with (4.67). Since, as pointed out earlier, $y_p(x) \geq x\tan\theta$ on $[0, x_p(\theta)]$, we have, on an event with probability tending to one,

$$Z_{n,p,1}(\theta) = \sqrt{k}\int_0^{x_p(\theta)}\int_{x\tan\theta}^{x\tan\theta+\frac{z_{n,\theta}(x)}{\sqrt{k}}} \lambda(x, y)\mathrm{d}y\mathrm{d}x,$$

where $z_{n,\theta}$ was defined in (4.53). Changing variables $y(z) = x\tan\theta + z/\sqrt{k}$ and making use of the mean value theorem leads to

$$Z_{n,p,1}(\theta) = \int_0^{x_p(\theta)} \lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) z_{n,\theta}(x)\mathrm{d}x,$$

where $y(x,\theta)$ is between 0 and $z_{n,\theta}(x)$. Observing that

$$\begin{aligned}
&\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) z_{n,\theta}(x) - \lambda(x, x\tan\theta)\left\{w_{1n}(x)\tan\theta - w_{2n}\left(x\tan\theta\right)\right\} \\
&\quad = \left\{\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right\} z_{n,\theta}(x) \\
&\qquad + \left[z_{n,\theta}(x) - \left\{w_{1n}(x)\tan\theta - w_{2n}\left(x\tan\theta\right)\right\}\right]\lambda(x, x\tan\theta),
\end{aligned}$$

we see that sufficient conditions for relation (4.67) to hold are provided by

$$\sup_{\theta\in[0,\pi/4]}\int_0^{x_p(\theta)}\left|\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right||z_{n,\theta}(x)|\mathrm{d}x = o_P(1) \quad (4.69)$$

$$\sup_{\theta\in[0,\pi/4]}\int_0^{x_p(\theta)}\lambda(x, x\tan\theta)\left|z_{n,\theta}(x) - \left\{w_{1n}(x)\tan\theta - w_{2n}\left(x\tan\theta\right)\right\}\right|\mathrm{d}x = o_P(1). \quad (4.70)$$

Let us first consider (4.69). Since we will need uniform continuity of λ, we will need to work on a compact domain. Hence, for some $M > 1$, we consider separately the case $\theta \in [0, \arctan(1/M)]$, on which we have no control on the value of $x_p(\theta)$ which can be arbitrary large, and the case $\theta \in [\arctan(1/M), \pi/4]$ on which it is easy to see that $x_p(\theta) \leq 1 + M$. It will be useful to note that as $n \to \infty$, by (4.56),

$$\|z_n\| \stackrel{\text{def}}{=} \sup_{\theta\in[0,\pi/4]}\sup_{x\in[0,x_p(\theta)]}|z_{n,\theta}(x)| = O_P(1). \quad (4.71)$$

Consider the supremum of (4.69) on the region $\theta \in [\arctan(1/M), \pi/4]$ first, for which the domain of integration is included in $x \in [0, 1+M]$. We split this domain further on $x \in [0, \delta]$ and $x \in [\delta, 1+M]$ for some $\delta > 0$. The latter case will be considered first. Observe that, by (4.71),

$$\sup_{\substack{\theta\in[\arctan(1/M),\pi/4]\\ x\in[\delta,1+M]}} \left\| \left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - (x, x\tan\theta)\right\|_1 \leq \frac{\|z_n\|}{\sqrt{k}} = o_P(1),$$

which, by uniform continuity of λ on $[\delta, 1+M] \times [\delta/(2M), 2+M]$ implies

$$\sup_{\theta\in[\arctan(1/M),\pi/4]} \sup_{x\in[\delta,1+M]} \left|\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right| = o_P(1).$$

Consequently,

$$\sup_{\theta\in[\arctan(1/M),\pi/4]} \int_{\delta}^{x_p(\theta)} \left|\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right| |z_{n,\theta}(x)| \mathrm{d}x$$

is $o_P(1)$ too. The next step is to consider the same region for θ but $x \in [0, \delta]$. The idea is to bound the integral as follows: by homogeneity of λ in (4.24), we have

$$\begin{aligned}\sup_{\theta\in[\arctan(1/M),\pi/4]} &\int_{0}^{\delta} \left|\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right| |z_{n,\theta}(x)| \mathrm{d}x \\ &\leq 32 \sup_{y\geq 0} \lambda(1,y) \sup_{\theta\in[\arctan(1/M),\pi/4]} \sup_{x\in[0,\delta]} \frac{|z_{n,\theta}(x)|}{(x\tan\theta)^{1/16}} (\tan\theta)^{1/16} \delta^{1/16}.\end{aligned}$$

Using (4.56) and picking δ small enough permits to conclude and to show that (4.69) is verified on the region $\theta \in [\arctan(1/M), \pi/4]$.

Consider now the supremum of (4.69) on the region $\theta \in [0, \arctan(1/M)]$, we split again the domain of integration in two parts : $x \in [0,1]$ and $x \in [1, x_p(\theta)]$. We start with the region near zero. Similar computations as done previously lead to

$$\begin{aligned}\sup_{\theta\in[0,\arctan(1/M)]} &\int_{0}^{1} \left|\lambda\left(x, x\tan\theta + \frac{y(x,\theta)}{\sqrt{k}}\right) - \lambda(x, x\tan\theta)\right| |z_{n,\theta}(x)| \mathrm{d}x \\ &\leq 32 \sup_{y\geq 0} \lambda(1,y) \sup_{\theta\in[0,\arctan(1/M)]} \sup_{x\in[0,1]} \frac{|z_{n,\theta}(x)|}{(x\tan\theta)^{1/16}} \times (\tan\theta)^{1/16}.\end{aligned}$$

Picking $M > 1$ large enough permits to conclude for this region. For the region $x \in [1, x_p(\theta)]$, using the homogeneity of λ in (4.24) and factoring out

$(x\tan\theta)^{1/16}$, we can show that

$$\sup_{\theta\in[0,\arctan(1/M)]}\int_1^{x_p(\theta)}\left|\lambda\left(x,x\tan\theta+\frac{y(x,\theta)}{\sqrt{k}}\right)-\lambda(x,x\tan\theta)\right||z_{n,\theta}(x)|\mathrm{d}x$$
$$\leq 64\sup_{\substack{\theta\in[0,\arctan(1/M)]\\ x\in[1,x_p(\theta)]}}\frac{|z_{n,\theta}(x)|}{(x\tan\theta)^{1/16}}\times\sup_{\substack{\theta\in[0,\arctan(1/M)]\\ |z|\leq\|z_n\|}}\lambda(1,\tan\theta+z/\sqrt{k}).$$

By continuity of λ, we have $\lim_{y\to 0}\lambda(1,y)=\lambda(1,0)=0$: indeed, we have

$$\infty>\Lambda([1,\infty)\times[0,1])=\iint_{x\geq 1\geq y\geq 0}\lambda(x,y)\,\mathrm{d}y\,\mathrm{d}x$$
$$=\iint_{x\geq 1\geq y\geq 0}x^{-1}\lambda(1,y/x)\,\mathrm{d}y\,\mathrm{d}x=\int_{x\geq 1}\int_{z=0}^{1/x}\lambda(1,z)\,\mathrm{d}z\,\mathrm{d}x,$$

so that if $\lambda(1,0)>0$, the inner integral would of the order $\lambda(1,0)/x$ as $x\to\infty$, making the outer integral diverge, yielding a contradiction. Again using (4.71) permits to conclude for the region $\theta\in[0,\arctan(1/M)]$ by picking $M>1$ large enough. Combining everything, we have proven (4.69).

We now turn to (4.70). By arguments similar to those leading up to (4.55), we have

$$\sup_{\theta\in[0,\pi/4]}\sup_{x\in[0,x_p(\theta)]}\frac{\left|z_{n,\theta}(x)-\{w_{1n}(x)\tan\theta-w_{2n}(x\tan\theta)\}\right|}{(x\tan\theta)^{1/16}}$$
$$=\sup_{\theta\in[0,\pi/4]}\sup_{x\in[0,x_p(\theta)]}\frac{\left|v_{2n}\left(x\tan\theta+\frac{w_{1n}(x)\tan\theta}{\sqrt{k}}\right)+w_{2n}(x\tan\theta)\right|}{(x\tan\theta)^{1/16}}=\mathrm{o_P}(1).$$

It follows from the homogeneity (4.24) and a computation similar to the previous ones that (4.70) holds. As already explained, (4.69) and (4.70) imply (4.67).

Finally, we show (4.68). Since $y_p(x)\leq x\tan\theta$ on $[x_p(\theta),\infty]$, we have, on an event with probability tending to one,

$$Z_{n,p,2}(\theta)=\sqrt{k}\int_{x_p(\theta)}^{\infty}\int_{y_p(x)}^{y_p(x)+s_{n,p}(x)/\sqrt{k}}\lambda(x,y)\mathrm{d}y\mathrm{d}x,$$

where $s_{n,p}$ was defined in (4.54). Changing variables $y(s)=y_p(x)+s/\sqrt{k}$ and making use of the mean value theorem leads to

$$Z_{n,p,2}(\theta)=\int_{x_p(\theta)}^{\infty}\lambda\left(x,y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right)s_{n,p}(x)\mathrm{d}x,$$

where $s(x,p)$ is between 0 and $s_{n,p}(x)$. Observing that

$$\begin{aligned}\lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) s_{n,p}(x) - \lambda(x, y_p(x))\left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\} \\ = \left\{\lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) - \lambda(x, y_p(x))\right\} s_{n,p}(x) \\ + \left[s_{n,p}(x) - \left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\}\right] \lambda(x, y_p(x)),\end{aligned}$$

and that for any $\theta \in [0, \pi/4]$ and $1 \le p < \infty$ we have $x_p(\theta) \ge 2^{1/p}$, we see that sufficient conditions for (4.68) to hold are provided by

$$\int_{2^{1/p}}^{\infty} \left|\lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) - \lambda(x, y_p(x))\right| |s_{n,p}(x)| \mathrm{d}x = \mathrm{o_P}(1), \quad (4.72)$$

$$\int_{2^{1/p}}^{\infty} \lambda(x, y_p(x)) \left|s_{n,p}(x) - \left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\}\right| \mathrm{d}x = \mathrm{o_P}(1). \quad (4.73)$$

Let us first consider (4.72). Since we will need uniform continuity of λ again, we want to work on a compact domain. Hence, for some $M > 1$, we consider separately the cases $x \in [2^{1/p}, M]$ and $x \in [M, \infty)$. Note that by (4.62) in Appendix F,

$$\|s_{n,p}\| \stackrel{\text{def}}{=} \sup_{x \in [2^{1/p}, \infty)} |s_{n,p}(x)| = \mathrm{O_P}(1) \quad (4.74)$$

as $n \to \infty$, and, by the mean value theorem, for each $x \in [2^{1/p}, \infty)$,

$$s_{n,p}(x) = y_p'(c_n(x))w_{1n}(x) + v_{2n}\left(y_p\left(x + \frac{w_{1n}(x)}{\sqrt{k}}\right)\right), \quad (4.75)$$

where $c_n(x)$ is between x and $x + w_{1n}(x)/\sqrt{k}$ and $y_p'(x) = -1/(x^p - 1)^{1+1/p}$. Since

$$\sup_{x \in [2^{1/p}, \infty)} \left\|\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) - (x, y_p(x))\right\|_1 \le \frac{\|s_{n,p}\|}{\sqrt{k}} = \mathrm{o_P}(1),$$

it follows from the uniform continuity of λ on compact sets included in $[0,\infty)^2 \setminus \{(0,0)\}$ that for any $M > 1$,

$$\sup_{x \in [2^{1/p}, M]} \left|\lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) - \lambda(x, y_p(x))\right| = \mathrm{o_P}(1)$$

and (4.72) follows easily for the region $x \in [2^{1/p}, M]$. For the region $x \in [M, \infty]$, just observe that by (4.74),

$$\begin{aligned}\int_M^{\infty} \left|\lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) - \lambda(x, y_p(x))\right| |s_{n,p}(x)| \mathrm{d}x \\ \le \|s_{n,p}\| \left\{\int_M^{\infty} \lambda\left(x, y_p(x)+\frac{s(x,p)}{\sqrt{k}}\right) \mathrm{d}x + \int_M^{\infty} \lambda(x, y_p(x)) \mathrm{d}x\right\}. \quad (4.76)\end{aligned}$$

We first consider the second integral which is easier to deal with. By homogeneity (4.24) and a change of variables,

$$\int_M^\infty \lambda(x, y_p(x))\mathrm{d}x = \int_M^\infty \lambda((x^p-1)^{1/p}, 1)(y_p(x))^{-1}\mathrm{d}x$$
$$\leq \int_{(M^p-1)^{1/p}}^\infty \lambda(u,1)\left(\frac{u^p}{u^p+1}\right)^{1-1/p}\mathrm{d}u \leq \int_{(M^p-1)^{1/p}}^\infty \lambda(u,1)\mathrm{d}u.$$

Since $\int_0^\infty \lambda(u,1)\mathrm{d}u = 1$, we may apply the dominated convergence theorem to ensure that the above integral can be made as small as desired provided we pick $M > 1$ large enough. Now consider the second integral on the right-hand side of (4.76). Since $0 < s(x,p) < s_{n,p}(x)$ or $0 > s(x,p) > s_{n,p}(x)$, we have, by (4.62), for all $M > 2^{1/p}$,

$$\int_M^\infty \lambda\left(x, y_p(x) + \frac{s(x,p)}{\sqrt{k}}\right)\mathrm{d}x = \int_M^\infty \lambda(x, y_p(x))\mathrm{d}x + \mathrm{o_P}(1),$$

and since the integral on the right-hand side vanishes as $M \to \infty$, we have shown (4.72).

Let us now consider (4.73). By Proposition 4.3.2, the continuity of y_p and its derivative y_p', we observe that

$$\sup_{x\in[2^{1/p},\infty)} \left|s_{n,p}(x) - \left\{y_p'(x)w_{1n}(x) - w_{2n}(y_p(x))\right\}\right|$$
$$\leq \sup_{x\in[2^{1/p},\infty)} \left|\left(y_p'(c_n(x)) - y_p'(x)\right)w_{1n}(x) + \left(v_{2n}\left(y_p\left(x+\tfrac{w_{1n}(x)}{\sqrt{k}}\right)\right) + w_{2n}(y_p(x))\right)\right|,$$

which is $\mathrm{o_P}(1)$. By integrability of the function $x \in [2^{1/p}, \infty) \mapsto \lambda(x, y_p(x))$, it follows that

$$\int_{2^{1/p}}^\infty \lambda(x, y_p(x))\left|s_{n,p}(x) - \left\{y_p'(x)w_{1n}(x) - w_{2n}\left(y_p(x)\right)\right\}\right|\mathrm{d}x = \mathrm{o_P}(1),$$

and (4.73) holds.

Conclusion of the proof of Theorem 4.3.1. Adding the asymptotic expansions of the three terms in (4.45) permits to conclude the proof for $\theta \in [0, \pi/4]$. By the symmetry argument in Section 4.C.1, we can extend the conclusion to all $\theta \in [0, \pi/2]$.

4.D Proof of Corollary 4.3.1

The second part is just an application of the functional delta method, i.e., Theorem 20.8 in [121]. The only thing we need to show is Hadamard differentiability of the map Ψ, that is, let $h \in C_0([0, \pi/2])$ and $(h_t)_{t>0}$ be any sequence

such that $\sup_{\theta\in[0,\pi/2]} |h_t(\theta) - h(\theta)| \to 0$ as $t \to 0$, then

$$\sup_{\theta\in[0,\pi/2]} \left| \frac{(\Psi(Q_p + th_t))(\theta) - (\Psi(Q_p))(\theta)}{t} - (\Psi'_{Q_p}[h])(\theta) \right| \to 0, \qquad t \to 0. \tag{4.77}$$

Noting the useful relation

$$\int_0^\theta f(\psi)\mathrm{d}F(\psi) = F(\theta)f(\theta) - \int_0^\theta F(\varphi)f'(\varphi)\mathrm{d}\varphi,$$

the candidate for the derivative is obtained by computing

$$(\Psi'_{Q_p}[h])(\theta) = \frac{\mathrm{d}}{\mathrm{d}t}(\Psi(Q_p + th))(\theta)\Big|_{t=0},$$

which leads to the indicated formula. Showing that (4.77) holds with this candidate is an easy exercise and completes the proof.

4.E Proof of Theorem 4.4.1

To facilitate the notation later in the proof, we introduce the following functions: for $j \in \{1, \ldots, d\}$,

$$\begin{aligned} Q_{nj}(u_j) &\stackrel{\text{def}}{=} U_{\lceil nu_j\rceil:n,j}, \qquad u \in [0,1]^d, \\ S_{nj}(x_j) &\stackrel{\text{def}}{=} \frac{n}{k} Q_{nj}\left(\frac{k}{n}x_j\right), \qquad x_j > 0, \\ S_n(x) &\stackrel{\text{def}}{=} (S_{n1}(x_1), \ldots, S_{nd}(x_d)), \qquad x \in [0,\infty)^d, \end{aligned}$$

where $U_{1:n,j} \le U_{2:n,j} \le \ldots \le U_{n:n,j}$ are the order statistics of the marginal sample $U_{1j}, \ldots, U_{nj}$ and $\lceil a\rceil$ is the smallest integer not smaller than $a \in \mathbb{R}$. We also write for $x \in [0,\infty)^d$,

$$\begin{aligned} V_n(x) &\stackrel{\text{def}}{=} \frac{n}{k}\mathrm{P}\left[U_1 \le \frac{k}{n}x_1 \text{ or } \ldots \text{ or } U_d \le \frac{k}{n}x_d\right]; \\ T_n(x) &\stackrel{\text{def}}{=} \frac{n}{k}\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{U_{i1} \le \frac{k}{n}x_1 \text{ or } \ldots \text{ or } U_{id} \le \frac{k}{n}x_d\right\}; \\ \hat{L}_n(x) &\stackrel{\text{def}}{=} \frac{n}{k}\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{U_{i1} \le \frac{k}{n}S_{n1}(x_1) \text{ or } \ldots \text{ or } U_{id} \le \frac{k}{n}S_{nd}(x_d)\right\} \\ &= \frac{1}{k}\sum_{i=1}^n \mathbb{I}\left\{R_{i1} > n + 1 - kx_1 \text{ or } \ldots \text{ or } R_{id} > n + 1 - kx_d\right\}. \end{aligned}$$

With this notation, we note that $v_n(x) = \sqrt{k}\{T_n(x) - V_n(x)\}$ and $\hat{L}_n(x) = T_n(S_n(x))$. As claimed in [45, Equation (7.1)], it is easily seen that the asymptotic properties of ℓ_n and $\hat{L}_n$ are the same so that we will consider the process $\sqrt{k}(\hat{L}_n - \ell)$ instead.

We consider the following decomposition: for $x \in [0,\infty)^d$,

$$\begin{aligned}\sqrt{k}(\hat{L}_n(x) - \ell(x)) = &\sqrt{k}\,(T_n(S_n(x)) - V_n(S_n(x))) \\ &+ \sqrt{k}\,(V_n(S_n(x)) - \ell(S_n(x))) \\ &+ \sqrt{k}\,(\ell(S_n(x)) - \ell(x))\,,\end{aligned}$$

where we call the first term of the decomposition the *stochastic term* (and denote it by $D_1(x)$), the second term will be called the *bias term* (and will be denoted by $D_2(x)$) and the last term will be called the *random set term* (and will be denoted by $D_3(x)$).

Skorokhod construction. As in [45], we will work on a Skorokhod construction, without changing the notation to keep the proof readable.

For fixed $T > 0$, Theorem 3.1 in [48] ensures that on a Skorokhod construction (depending on T), there exists a sequence of processes $(v_n)_{n\in\mathbb{N}}$ and a Wiener process $W_l(x) \overset{\text{def}}{=} W_\Lambda(A_x)$, where W_Λ is the d-dimensional analog to the one appearing in Theorem 4.3.1 as in Proposition 4.3.1, such that as $n \to \infty$,

$$\sup_{x\in[0,2T]^d} |v_n(x) - W_l(x)| \xrightarrow{a.s.} 0. \tag{4.78}$$

In particular, this implies the marginal convergences

$$\sup_{x_j\in[0,2T]} \left|v_{nj}(x_j) - W_j(x_j)\right| \xrightarrow{a.s.} 0, \qquad j \in \{1,\dots,d\}, \tag{4.79}$$

where $W_j(x_j) \overset{\text{def}}{=} W_l(0,\dots,0,x_j,0,\dots,0)$. By Vervaat's Lemma [34, Lemma A.0.2], we also have the convergence of the marginal quantile processes

$$\sup_{x_j\in[0,2T]} \left|\sqrt{k}\left(S_{nj}(x_j) - x_j\right) + W_j(x_j)\right| \xrightarrow{a.s.} 0, \qquad j \in \{1,\dots,d\}. \tag{4.80}$$

Asymptotic expansions of the bias term. We show that D_2 is asymptotically negligible for our purposes, that is, $\sup_{x\in[0,T]^d} |D_2(x)| = \mathrm{o_P}(1)$ as $n \to \infty$. It follows from (4.80) that $S_{nj}(x_j) = x_j + \mathrm{O_P}(1/\sqrt{k})$ where the $\mathrm{O_P}$ term is uniform in $x_j \in [0,T]$ so that, with probability tending to one, we have

$$\begin{aligned}\sup_{x\in[0,T]^d} |D_2(x)| &\le \sup_{y\in[0,2T]^d} \sqrt{k}\,|V_n(y) - \ell(y)| \\ &\le 2T\sqrt{k} \sup_{w\in[0,2T]^d,\,\|w\|_1=1} |V_n(w) - \ell(w)|\,.\end{aligned}$$

By points 1 and 2 of Assumption 4.4.2, the latter expression is of the order

$$\sqrt{k}\,\mathrm{O}\left((k/n)^{\alpha}\right) = \mathrm{O}\left(\left(\frac{k}{n^{2\alpha/(1+2\alpha)}}\right)^{1/2+\alpha}\right) = \mathrm{o}(1)$$

as $n \to \infty$.

Asymptotic expansions of the stochastic term. Note that for any $x \in [0,T]^d$, $D_1(x) = v_n(S_n(x))$. Our aim is to show that

$$\sup_{x\in[0,T]^d} |v_n(S_n(x)) - v_n(x)| = \mathrm{o_P}(1),$$

as $n \to \infty$. By the triangle inequality, we have

$$\begin{aligned}\sup_{x\in[0,T]^d} |v_n(S_n(x)) - v_n(x)| \leq & \sup_{x\in[0,T]^d} |v_n(S_n(x)) - W_l(S_n(x))| \\ & + \sup_{x\in[0,T]^d} |W_l(S_n(x)) - W_l(x)| \\ & + \sup_{x\in[0,T]^d} |W_l(x) - v_n(x)|\,.\end{aligned}$$

Arguing as for the bias term, by (4.80), we see that the first term in this decomposition is bounded with probability tending to one by $\sup_{y\in[0,2T]^d} |v_n(y) - W_l(y)|$ which is $\mathrm{o_P}(1)$ as $n \to \infty$, by (4.78). The second term of this decomposition is $\mathrm{o_P}(1)$ as $n \to \infty$ by uniform continuity of the limiting process W_l. Finally, the last term of this decomposition is immediately $\mathrm{o_P}(1)$ as $n \to \infty$ by (4.78).

Asymptotic expansions of the random set term. It follows from point 3 in Assumption 4.4.2 and the mean value theorem that for $x \in [0,T]^d$,

$$D_3(x) = \sqrt{k}\left(\ell(S_n(x)) - \ell(x)\right) = \sum_{j=1}^{d} \sqrt{k}\left(S_{nj}(x_j) - x_j\right)\dot{\ell}_j(\xi_n),$$

for some ξ_n such that $\xi_{nj} \in (x_j, S_{nj}(x_j))$ for any $j = 1,\ldots,d$. Hence, by the triangle inequality, we are done proving our expansion if we show

$$\sup_{x\in[0,T]^d} \sum_{j=1}^{d} \left|\sqrt{k}\left(S_{nj}(x_j) - x_j\right)\dot{\ell}_j(\xi_n) + v_{nj}(x_j)\dot{\ell}_j(x)\right| = \mathrm{o_P}(1).$$

Every term in this sum can be dealt with in the exact same way so that we only consider $j = 1$.

Fix $0 < \delta < T$. By the triangle inequality and (4.80),

$$\begin{aligned}
&\sup_{x\in[0,T]^d} \left|\sqrt{k}\,(S_{n1}(x_1) - x_1)\,\dot{\ell}_1(\xi_n) + v_{n1}(x_1)\dot{\ell}_1(x)\right| \\
&\leq \sup_{y\in[0,2T]^d} \left|\dot{\ell}_1(y)\right| \cdot \sup_{x_1\in[0,T]} \left|\sqrt{k}\,(S_{n1}(x_1) - x_1) + v_{n1}(x_1)\right| \\
&\quad + \sup_{x_1\in[0,T]} |v_{n1}(x_1)| \cdot \sup_{x\in[\delta,T]\times[0,T]^{d-1}} \left|\dot{\ell}_1(\xi_n) - \dot{\ell}_1(x)\right| \\
&\quad + 2 \sup_{y\in[0,2T]^d} \left|\dot{\ell}_1(y)\right| \cdot \sup_{x_1\in[0,\delta]} |v_{n1}(x_1)|\,.
\end{aligned}$$

In each term of the latter upper bound, it is easy to see that the first factor in the product is bounded (in probability) and that the second factor converges to zero (in probability). For the first term, we simply use that $0 \leq \dot{\ell}_1 \leq 1$ and we combine (4.79) and (4.80). For the second term, we use (4.79) which shows that v_{n1} is uniformly bounded in probability on $[0,T]^d$ by Prohorov's theorem and the fact that $\dot{\ell}_1$ is continuous on the compact set $[\delta,T]\times[0,T]^{d-1}$, hence, uniformly continuous on the same set. For the third term, we again use that $0 \leq \dot{\ell}_1 \leq 1$ and (4.79) which ensures that by picking n large enough, $\sup_{x_1\in[0,\delta]} |v_{n1}(x_1)|$ is arbitrary close to $\sup_{x_1\in[0,\delta]} |W_1(x_1)|$ in probability and, by picking δ small enough, the latter supremum will be arbitrary close to zero. This completes the proof of the first part of the theorem.

For the second part, note that

$$\begin{aligned}
&\sup_{\substack{x_1,x_2>0\\ x_1+x_2=1}} \left|t^{-1}\mathrm{P}\left[U_1 \leq tx_1, U_2 \leq tx_2\right] - \ell(x_1,x_2)\right| \\
&\qquad\qquad \leq \sup_{\substack{x_1,x_2>0\\ x_1+x_2=1}} \iint_{A_{(x_1,x_2)}} |tc(tu_1,tu_2) - \lambda(u_1,u_2)|\,\mathrm{d}u_1\mathrm{d}u_2,
\end{aligned}$$

where we recall $A_{(x_1,x_2)} = \{(u_1,u_2) \in [0,\infty)^2 : u_1 \leq x_1 \text{ or } u_2 \leq x_2\}$. One observes that the latter supremum can be bounded, independently of (x_1,x_2) by

$$\mathcal{D}_{1/t}(t) + \iint_{\substack{u_1\wedge u_2\leq 1\\ u_1\vee u_2>1/t}} |tc(tu_1,tu_2) - \lambda(u_1,u_2)|\,\mathrm{d}u_1\mathrm{d}u_2.$$

Under Assumption 4.3.2, the first term satisfies $\mathcal{D}_{1/t}(t) = \mathrm{O}(t^\alpha)$ as $t \to \infty$. For the second term, assuming $t < 1$, it equals

$$\iint_{\substack{u_1\wedge u_2\leq 1\\ u_1\vee u_2>1/t}} \lambda(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2,$$

since the copula density $c \equiv 0$ outside of the unit square $[0,1]^2$. Moving to polar coordinates, one shows that the latter integral is bounded by

$$\Phi_p(t) + \left\{\Phi_p\left(\tfrac{\pi}{2}\right) - \Phi_p\left(\tfrac{\pi}{2} - t\right)\right\} = O(t^\alpha),$$

where the equality also follows from Assumption 4.3.2. This shows that condition 1 of Assumption 4.4.2 is verified in this setting. Condition 2 is immediate from Assumption 4.3.2. Condition 3 follows from Assumption 4.3.1 since for $x = (x_1, x_2) \in [0, \infty)^2$, using the fact that Λ has Lebesgue margins,

$$\begin{aligned}\ell(x) = \Lambda(A_x) &= \Lambda\left([0, x_1] \times [0, \infty)\right) + \Lambda\left([0, \infty) \times [0, x_2]\right) - \Lambda\left([0, x_1] \times [0, x_2]\right) \\ &= x_1 + x_2 - \Lambda\left([0, x_1] \times [0, x_2]\right),\end{aligned}$$

which is continuously differentiable under Assumption 4.3.1.

4.F Proof of Theorem 4.4.2

We assume that we are working under H_0 and that $r_0 \in \mathcal{R}$ denotes the true parameter underlying the data.

First observe that

$$T_n = F\left[\sqrt{k}(\widetilde{Q}_p - Q_{p,\widehat{r}_n})\right],$$

where $F : \ell^\infty([0, \pi/2]) \to (0, \infty)$ is defined for any $z \in \ell^\infty([0, \pi/2])$ by

$$F(z) \stackrel{\text{def}}{=} \int_0^{\pi/2} |z(\theta)| q(\theta) \mathrm{d}\theta.$$

The map F is continuous since if $(z_n)_{n\in\mathbb{N}} \subset \ell^\infty([0, \pi/2])$ is such that $z_n \to z \in \ell^\infty([0, \pi/2])$,

$$\begin{aligned}|F(z_n) - F(z)| &\leq \int_0^{\pi/2} \big||z_n(\theta)| - |z(\theta)|\big| \, q(\theta)\mathrm{d}\theta \\ &\leq \|z_n - z\|_{\ell^\infty([0,\pi/2])} \int_0^{\pi/2} q(\theta)\mathrm{d}\theta \to 0,\end{aligned}$$

as $n \to \infty$ since q is assumed to be integrable. Consequently, in view of the continuous mapping theorem, it suffices to determine the asymptotic distribution of the process

$$\left\{\sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_{p,\widehat{r}_n}(\theta)\right) : \theta \in [0, \pi/2]\right\},$$

in $\ell^\infty([0, \pi/2])$ to determine the one of our test statistic.

For any $\theta \in [0, \pi/2]$, we have

$$\sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_{p,\widehat{r}_n}(\theta)\right) = \sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_{p,r_0}(\theta)\right) - \sqrt{k}\left(Q_{p,\widehat{r}_n}(\theta) - Q_{p,r_0}(\theta)\right).$$

In view of Corollary 4.3.1, the first term in the right hand side satisfies

$$\sqrt{k}\left(\widetilde{Q}_p(\theta) - Q_{p,r_0}(\theta)\right) = E_{n,p}^{Q_{p,r_0}}(\theta) + \frac{\int_0^{\pi/2} E_{n,p}^{Q_{p,r_0}}(\psi) f'(\psi) \mathrm{d}\psi}{\sigma^2_{Q_{p,r_0}}(f)} \int_0^\theta f(\psi) \mathrm{d}Q_{p,r_0}(\psi) + \mathrm{o_P}(1),$$

as $n \to \infty$ where the reminder is uniform in $\theta \in [0, \pi/2]$. In view of Assumption 4.4.1 and Theorem 4.4.1, the second term satisfies,

$$\begin{aligned}
&\sqrt{k}\left(Q_{p,\widehat{r}_n}(\theta) - Q_{p,r_0}(\theta)\right) \\
&= \sqrt{k}\left(\mathfrak{R}(\widehat{r}_n) - \mathfrak{R}(r_0)\right) \\
&= \int_0^\theta \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot \sqrt{k}(\widehat{r}_n - r_0) + \mathrm{o_P}(1) \\
&= \int_0^\theta \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot \\
&\qquad \int_{[0,1]^2} \sqrt{k}\left(\widehat{\ell}_n(x_1,x_2) - \ell_{r_0}(x_1,x_2)\right) g(x_1,x_2)\, \mathrm{d}\sigma(x_1,x_2) + \mathrm{o_P}(1) \\
&= \int_0^\theta \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot \\
&\qquad \int_{[0,1]^2} \left(v_n(x_1,x_2) - \sum_{j=1}^2 \dot{\ell}_{r_0,j}(x_1,x_2) v_{nj}(x_j)\right) g(x_1,x_2)\, \mathrm{d}\sigma(x_1,x_2) \\
&\qquad\qquad + \mathrm{o_P}(1),
\end{aligned}$$

where the reminder is uniform in $\theta \in [0, \pi/2]$.

Summarizing, we have shown that, up to asymptotically negligible terms, all the terms appearing in the decomposition of the process $\sqrt{k}(\widetilde{Q}_p - Q_{p,\widehat{r}_n})$ are asymptotically equivalent to continuous functionals of the tail empirical process

$$W_{n,k}(C) = \sqrt{k}\left\{\frac{n}{k} P_n\left(\frac{k}{n}C\right) - \frac{n}{k} P\left(\frac{k}{n}C\right)\right\},$$

evaluated on sets belonging to a collection $\mathcal{C}$ which is of finite VC dimension. Combining Propositions 4.3.1 and 4.3.2 with a suitable application of the continuous mapping theorem [122, Theorem 1.3.6] permits to conclude the proof.

4.G Proof of Theorem 4.4.3

The proof of Theorem 4.4.3 relies on intermediate results that we formulate and prove below.

Consider again the Gaussian process W_Λ indexed by Borel sets of $\mathbb{E}$ introduced below Theorem 4.3.1. Given a simple function

$$f(x) \stackrel{\text{def}}{=} \sum_{i=1}^{N} c_i \mathbb{I}_{C_i}(x),$$

with constants $c_1, \dots, c_N \in \mathbb{R}$ and Borel sets $C_1, \dots, C_N$ in $\mathbb{E}$, it will be convenient to introduce the notation

$$W_\Lambda(f) \stackrel{\text{def}}{=} \sum_{i=1}^{N} c_i W_\Lambda(C_i).$$

It is easy to see that for any such functions $f_1, \dots, f_k$, the random vector $\big(W_\Lambda(f_1), \dots, W_\Lambda(f_k)\big)$ remains normally distributed with zero mean and the covariance between any pair equals

$$\mathbb{E}[W_\Lambda(f_i) W_\Lambda(f_j)] = \Lambda(f_i f_j) = \int_{\mathbb{E}} f_i(x) f_j(x) \mathrm{d}\Lambda(x).$$

Consequently, we may view W_Λ as a Gaussian process indexed by simple functions on $\mathbb{E}$.

Lemma 4.G.1. *Consider the class of functions*

$$\begin{aligned}\mathcal{F} &\stackrel{\text{def}}{=} \left\{\mathbb{I}_{C_{p,\theta}} : \theta \in [0, \pi/2]\right\} \cup \left\{\mathbb{I}_{[0,x]\times[0,\infty]} w(x) : x \in [0,\infty)\right\} \\ &\qquad \cup \left\{\mathbb{I}_{[0,\infty]\times[0,y]} w(y) : y \in [0,\infty)\right\} \cup \left\{\mathbb{I}_{A_{(x,y)}} : (x,y) \in [0,1]^2\right\} \\ &\stackrel{\text{def}}{=} \mathcal{F}_1 \cup \mathcal{F}_2 \cup \mathcal{F}_3 \cup \mathcal{F}_4,\end{aligned}$$

where

$$w(x) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{x^\eta \vee x^\gamma}, & \text{if } x > 0 \\ 0, & \text{if } x = 0, \end{cases}$$

with $0 \le \eta < 1/2 < \gamma < 1$. Let $r_0 \in \mathcal{R}$. If $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ is such that $r_n \to r_0$ as $n \to \infty$, we have in $\ell^\infty(\mathcal{F})$ the weak convergence

$$\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}\right\} \stackrel{\mathcal{L}}{\to} \left\{W_{\Lambda_{r_0}}(f) : f \in \mathcal{F}\right\}, \qquad n \to \infty,$$

provided the maps $\theta \in [0, \pi/2] \mapsto \Phi_{p,r_0}(\theta)$ and $r \in \mathcal{R} \mapsto \Phi_{p,r}(\theta)$ (for fixed $\theta \in [0, \pi/2]$) are continuous, an assumption that is verified under Assumption 4.3.1 and point 1 of Assumption 4.4.1.

Proof. By Theorem 1.5.4 in [122], the desired weak convergence holds provided

1. For any $k \in \mathbb{N}$ and any $f_1, \ldots, f_k \in \mathcal{F}$,
$$\left(W_{\Lambda_{r_n}}(f_1), \ldots, W_{\Lambda_{r_n}}(f_k)\right) \xrightarrow{\mathcal{L}} \left(W_{\Lambda_{r_0}}(f_1), \ldots, W_{\Lambda_{r_0}}(f_k)\right)$$
as random vectors on $\mathbb{R}^k$.

2. The sequence
$$\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}\right\}_{n\in\mathbb{N}}$$
is asymptotically tight.

Point 1 follows from the fact that each random vector $\left(W_{\Lambda_{r_n}}(f_1), \ldots, W_{\Lambda_{r_n}}(f_k)\right)$ is Gaussian with zero mean and covariance matrix $\Sigma_k(r_n)$ having elements
$$\lim_{n\to\infty} \Sigma_{k,ij}(r_n) = \Sigma_{k,ij}(r_0), \qquad i, j \in \{1, \ldots, k\},$$
where $\Sigma_k(r_0)$ is the covariance matrix of $\left(W_{\Lambda_{r_0}}(f_1), \ldots, W_{\Lambda_{r_0}}(f_k)\right)$. Indeed, each function $f \in \mathcal{F}$ is an (possibly weighted) indicator function of a certain set in $\mathbb{E}$ so that element-wise convergence of the covariance matrix $\Sigma_k(r_n)$ to $\Sigma_k(r_0)$ follows from the continuity of the map $r \in \mathcal{R} \mapsto \Phi_{p,r}(\theta)$ for fixed $\theta \in [0, \pi/2]$. This is in particular guaranteed by point 1 of Assumption 4.4.1.

Point 2 will follow from establishing the asymptotic tightness of each sequence
$$\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}_j\right\}_{n\in\mathbb{N}}, \qquad j \in \{1, \ldots, 4\}$$
separately. By Theorem 1.5.7 in [122], it is equivalent to show that for each $j = 1, \ldots, 4$, there exists a semi-metric ρ_j on $\mathcal{F}_j$ such that $(\mathcal{F}_j, \rho_j)$ is totally bounded and for any $\epsilon, \eta > 0$, there exists $\delta = \delta(\epsilon, \eta) > 0$ and $n_0 = n_0(\delta) \in \mathbb{N}$ such that
$$\mathrm{P}\left[\sup_{f,g\in\mathcal{F}_j, \rho_j(f,g)<\delta} \left|W_{\Lambda_{r_n}}(f) - W_{\Lambda_{r_n}}(g)\right| > \epsilon\right] < \eta, \qquad n \geq n_0.$$

Case $j = 1$. Each function $f \in \mathcal{F}_1$ is in bijection with a certain $\theta \in [0, \pi/2]$. Hence, we may view the sequence $\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}_1\right\}_{n\in\mathbb{N}}$ as a sequence of processes indexed by $T_1 \overset{\text{def}}{=} [0, \pi/2]$. Trivially, each element of $\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}_1\right\}_{n\in\mathbb{N}}$ is a Gaussian process and, hence, sub-Gaussian as defined on page 101 of [122] with respect to the standard deviation semi-metric
$$\rho_{1,n}(\theta_1, \theta_2) \overset{\text{def}}{=} \sqrt{\left|\Phi_{p,r_n}(\theta_1) - \Phi_{p,r_n}(\theta_2)\right|} \overset{\text{def}}{=} \sqrt{d_{1,n}(\theta_1, \theta_2)}.$$
By Corollary 2.2.8 and page 98 of [122], it follows that there exists a universal constant $K > 0$ such that for any $\delta > 0$ and any $n \in \mathbb{N}$, we have
$$\mathbb{E}\left[\sup_{\rho_{1,n}(\theta_1,\theta_2)<\delta} \left|W_{\Lambda_{r_n}}(\theta_1) - W_{\Lambda_{r_n}}(\theta_2)\right|\right] \leq K \int_0^\delta \sqrt{\log N(\tfrac{t}{2}, T_1, \rho_{1,n})}\mathrm{d}t, \quad (4.81)$$

where $N(t, T_1, \rho_{1,n})$ is the minimal number of $\rho_{1,n}$-balls of radius $t > 0$ needed to cover T_1. First observe that for any $t > 0$, we have

$$N(t, T_1, \rho_{1,n}) = N(t^2, T_1, d_{1,n}).$$

Furthermore, the map

$$\theta \in (T_1, d_{1,n}) \mapsto \Phi_{p,r_n}(\theta) \in ([0, \Phi_{p,r_n}(\pi/2)], e(x,y) \stackrel{\text{def}}{=} |x-y|)$$

is an isometry so that we can compute the covering number in the latter space. Since $[0, \Phi_{p,r_n}(\pi/2)] \subseteq I \stackrel{\text{def}}{=} [0,2]$ for any $n \in \mathbb{N}$, we have

$$N(t, T_1, d_{1,n}) \leq N(t, I, e), \qquad \forall t > 0.$$

Note that for any $x \in I$,

$$B(x,t,e) \stackrel{\text{def}}{=} \{y \in I : e(x,y) \leq t\} = [x-t, x+t] \cap I,$$

so that if $t \geq 1$, we have $N(t, I, e) = 1$ and for $0 < t < 1$,

$$N(t, I, e) \leq t^{-1}.$$

We deduce that for $t \geq 1$, we have $N(t, T_1, \rho_{1,n}) = 1$, while for $0 < t < 1$,

$$N(t, T_1, \rho_{1,n}) = N(t^2, T_1, d_{1,n}) \leq N(t^2, I, e) \leq t^{-2}.$$

Since this holds for any $n \in \mathbb{N}$, note in particular that $N(t, T_1, \rho_1)$ where

$$\rho_1(\theta_1, \theta_2) \stackrel{\text{def}}{=} \lim_{n \to \infty} \rho_{1,n}(\theta_1, \theta_2), \qquad (\theta_1, \theta_2) \in [0, \pi/2]^2$$

satisfies the same bound and the space (T_1, ρ_1) is totally bounded. Now let $\delta \in (0,2)$. Then for any $t \in [0, \delta]$, we have by previous computations,

$$\sqrt{\log N(\tfrac{t}{2}, T_1, \rho_{1,n})} \leq \sqrt{\log(4/t^2)} = \sqrt{2\log(2/t)}.$$

thus, using a standard bound on the complementary error function erfc,

$$\begin{aligned}\int_0^\delta \sqrt{\log N(\tfrac{t}{2}, T_1, \rho_{1,n})}\mathrm{d}t &\leq \sqrt{2}\left\{2(\delta/2)\sqrt{\log(2/\delta)} + \sqrt{\pi}\,\mathrm{erfc}\left(\sqrt{\log(2/\delta)}\right)\right\} \\ &\leq \sqrt{2}\delta\left\{\sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)})\right\}.\end{aligned}$$

It follows from (4.81) and Markov's inequality that for any $\epsilon, \delta > 0$ and $n \in \mathbb{N}$,

$$\begin{aligned}&\mathrm{P}\left[\sup_{\rho_{1,n}(\theta_1,\theta_2)<\delta} \left|W_{\Lambda_{r_n}}(\theta_1) - W_{\Lambda_{r_n}}(\theta_2)\right| > \epsilon\right] \\ &\qquad \leq K\sqrt{2}\epsilon^{-1}\delta\left\{\sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)})\right\}.\end{aligned}$$

Now fix $\epsilon, \eta > 0$. Pick $\delta = \delta(\epsilon, \eta) \in (0, 2)$ such that the inequality

$$K 2^{3/2} \varepsilon^{-1} \delta \left\{ \sqrt{\log(1/\delta)} + 1/(2\sqrt{\log(1/\delta)}) \right\} < \eta$$

holds true and pick $n_0 = n_0(\delta) \in \mathbb{N}$ such that for $n \geq n_0$,

$$\left\{ (\theta_1, \theta_2) \in [0, \pi/2]^2 : \rho_1(\theta_1, \theta_2) < \delta \right\} \subset \left\{ (\theta_1, \theta_2) \in [0, \pi/2]^2 : \rho_{1,n}(\theta_1, \theta_2) < 2\delta \right\}$$

(such n_0 exists by uniform convergence of $\rho_{1,n}$ to ρ_1 which itself follows from the uniform convergence $\Phi_{p,r_n} \to \Phi_{p,r_0}$ which is a consequence of the pointwise convergence and the continuity of Φ_{p,r_0}) then, for $n \geq n_0$ we have

$$\begin{aligned} &\mathsf{P}\left[\sup_{\rho_1(\theta_1,\theta_2)<\delta} \left| W_{\Lambda_{r_n}}(\theta_1) - W_{\Lambda_{r_n}}(\theta_2) \right| > \epsilon \right] \\ &\qquad \leq \mathsf{P}\left[\sup_{\rho_{1,n}(\theta_1,\theta_2)<2\delta} \left| W_{\Lambda_{r_n}}(\theta_1) - W_{\Lambda_{r_n}}(\theta_2) \right| > \epsilon \right] < \eta, \end{aligned}$$

and the asymptotic tightness condition is satisfied for $j = 1$.

Case $j = 2$ and $j = 3$. We only consider $j = 2$ since the other situation is perfectly symmetric and can be handled with exactly the same tools. Here, each function is in bijection with a certain $x \in [0, \infty)$ and we will view the sequence $\left\{ W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}_2 \right\}_{n \in \mathbb{N}}$ as a sequence of processes indexed by $T_2 \stackrel{\text{def}}{=} [0, \infty)$. For any $n \in \mathbb{N}$, the standard deviation semi-metric of the Gaussian process $\left\{ W_{\Lambda_{r_n}}(x) : x \in [0, \infty) \right\}$ equals

$$\begin{aligned} \rho_2(x_1, x_2) &\stackrel{\text{def}}{=} \sqrt{w(x_1)^2 x_1 + w(x_2)^2 x_2 - 2w(x_1)w(x_2)\min(x_1, x_2)} \\ &\stackrel{\text{def}}{=} \sqrt{d_2(x_1, x_2)}. \end{aligned}$$

Consequently, there exists a universal constant $K > 0$ such that for any $n \in \mathbb{N}$ and any $\delta > 0$,

$$\mathbb{E}\left[\sup_{\rho_2(x_1,x_2)<\delta} \left| W_{\Lambda_{r_n}}(x_1) - W_{\Lambda_{r_n}}(x_2) \right| \right] \leq K \int_0^\delta \sqrt{\log N(\tfrac{t}{2}, T_2, \rho_2)} \mathrm{d}t, \quad (4.82)$$

with $N(t, T_2, \rho_2)$ the minimal number of ρ_2-balls of radius $t > 0$ needed to cover T_2.

Observe that for any $t > 0$,

$$N(t, T_2, \rho_2) = N(t^2, T_2, d_2) \leq N(t^2, T_{2,1}, d_2) + N(t^2, T_{2,2}, d_2), \quad (4.83)$$

where $T_{2,1} \overset{\text{def}}{=} [0,1]$ and $T_{2,2} \overset{\text{def}}{=} [1,\infty]$. We start by dealing with the quantity $N(t, T_{2,1}, d_2)$. Note that for any $x_1, x_2 \in T_{2,1}$, we have

$$d_2(x_1,x_2) = x_1^{1-2\eta} + x_2^{1-2\eta} - 2(x_1 \wedge x_2)(x_1x_2)^{-\eta}. \tag{4.84}$$

In particular, for any $x \in T_{2,1}$,

$$d_2(0,x) = x^{1-2\eta}.$$

This shows that $N(t, T_{2,1}, d_2) = 1$ for $t \geq 1$. Allowing d_2 to be defined on $[0,\infty)$ in the exact same way as in (4.84), we observe that for any $t > 0$ and $n \in \mathbb{N}$,

$$d_2(nt,(n+1)t) = t^{1-2\eta} d_2(n,n+1) = d_2(0,t)d_2(n,n+1).$$

Our aim is to show that $d_2(n,n+1) \leq 1$ for any $n \in \mathbb{N}$. To this end, take any $n \in \mathbb{N}$ and note that

$$\begin{aligned} d_2(n,n+1) \leq 1 \iff g(\eta) &\overset{\text{def}}{=} d_2(n,n+1) \\ &= n^{1-2\eta} + (n+1)^{1-2\eta} - 2n^{1-\eta}(n+1)^{-\eta} \\ &\leq g(0) = 1, \end{aligned}$$

so that it suffices to show that the map $\eta \in (0,1/2) \mapsto g(\eta)$ is decreasing. This follows from writing

$$g(\eta) = \left((n+1)^{\frac{1}{2}-\eta} - n^{\frac{1}{2}-\eta} \right)^2 + 2\sqrt{n}[n(n+1)]^{-\eta} \left[\sqrt{n+1} - \sqrt{n} \right],$$

which shows that $g(\eta)$ is a sum of two terms decreasing in $\eta \in (0,1/2)$. This is obvious for the second term while for the first it follows from the fact that the application $\alpha \in (0,1/2) \mapsto (n+1)^\alpha - n^\alpha$ is increasing as shown by the direct computation

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \{(n+1)^\alpha - n^\alpha\} = \log(n+1)(n+1)^\alpha - \log(n) n^\alpha \geq 0.$$

For any $0 < t < 1$, the previous arguments guarantee that $d_2(t^{\frac{1}{1-2\eta}}, 2t^{\frac{1}{1-2\eta}}) \leq d_2(0, t^{\frac{1}{1-2\eta}}) = t$, and, similarly, using the notation $B(x,t,d_2) \overset{\text{def}}{=} \{y \in T_{2,1} : d_2(y,x) \leq t\}$, we see that

$$\begin{aligned} B(t^{\frac{1}{1-2\eta}}, t, d_2) &\subseteq [0, 2t^{\frac{1}{1-2\eta}}] \\ B(3t^{\frac{1}{1-2\eta}}, t, d_2) &\subseteq [2t^{\frac{1}{1-2\eta}}, 4t^{\frac{1}{1-2\eta}}] \\ &\vdots \end{aligned}$$

hence showing that

$$T_{2,1} \subseteq \bigcup_{n=}^{N(t)} B((n+1)t^{\frac{1}{1-2\eta}}, t, d_2),$$

where

$$N(t, T_{2,1}, d_2) \leq N(t) \leq \frac{1}{2} t^{-\frac{1}{1-2\eta}}.$$

Regarding $N(t, T_{2,2}, d_2)$, simply observe that the map

$$x \in (T_2, d_2) \mapsto x^{-1} \in (T_2, d_2)$$

is an isometry for $\eta = 1 - \gamma$. We deduce that $N(t, T_{2,2}, d_2) = 1$ for any $t \geq 1$ and

$$N(t, T_{2,2}, d_2) \leq \frac{1}{2} t^{-\frac{1}{2\gamma-1}}, \qquad 0 < t < 1,$$

so that in view of equation (4.83), for any $t \in (0, 1)$,

$$N(t, T_2, \rho_2) \leq \frac{1}{2} \left(t^{-\frac{2}{1-2\eta}} + t^{-\frac{2}{2\gamma-1}} \right).$$

In particular, the space (T_2, ρ_2) is totally bounded. Pick $\delta \in (0, 2)$. Then by previous computations, we have that for any $t \in [0, \delta]$,

$$\sqrt{\log N(\tfrac{t}{2}, T_2, \rho_2)} \leq \sqrt{\frac{2}{1-2\alpha}} \sqrt{\log(2/t)},$$

where $\alpha \stackrel{\text{def}}{=} \min\{\eta, 1-\gamma\} \in [0, 1/2)$. Computations similar to the case $j = 1$ then leads to

$$\begin{aligned} \int_0^\delta \sqrt{\log N(\tfrac{t}{2}, T_2, \rho_2)} \mathrm{d}t &\leq \sqrt{\frac{2}{1-2\alpha}} \left\{ 2(\delta/2)\sqrt{\log(2/\delta)} + \sqrt{\pi}\, \mathrm{erfc}\left(\sqrt{\log(2/\delta)} \right) \right\} \\ &\leq \sqrt{\frac{2}{1-2\alpha}} \delta \left\{ \sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)}) \right\}. \end{aligned}$$

Hence, equation (4.82) and Markov's inequality imply that for any $\epsilon, \delta > 0$,

$$\begin{aligned} &\mathrm{P}\left[\sup_{\rho_2(x_1,x_2)<\delta} \left| W_{\Lambda_{r_n}}(x_1) - W_{\Lambda_{r_n}}(x_2) \right| > \varepsilon \right] \\ &\qquad \leq \sqrt{\frac{2}{1-2\alpha}} K \varepsilon^{-1} \delta \left\{ \sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)}) \right\}. \end{aligned}$$

Fixing $\epsilon, \eta > 0$ and picking $\delta = \delta(\epsilon, \eta) \in (0, 2)$ sufficiently small such that

$$\sqrt{\frac{2}{1-2\alpha}} K \varepsilon^{-1} \delta \left\{ \sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)}) \right\} < \eta,$$

show the asymptotic tightness condition for $j = 2$.

Case $j = 4$. As for the previous cases, each function $f \in \mathcal{F}_4$ is in a one-to-one relation with a certain $(x, y) \in [0,1]^2$ and we will view the sequence $\left\{W_{\Lambda_{r_n}}(f) : f \in \mathcal{F}_4\right\}_{n\in\mathbb{N}}$ as a sequence of processes indexed by $T_4 \stackrel{\text{def}}{=} [0,1]^2$. Observe that for any $n \in \mathbb{N}$, the standard deviation semi-metric associated with the Gaussian process $\left\{W_{\Lambda_{r_n}}(x, y) : (x, y) \in [0,1]^2\right\}$ satisfies

$$\begin{aligned}\rho_{4,n}\left((x_1, y_1), (x_2, y_2)\right) &\stackrel{\text{def}}{=} \sqrt{\Lambda_{r_n}\left(A_{(x_1,y_1)} \triangle A_{(x_2,y_2)}\right)} \\ &\leq \sqrt{|x_1 - x_2| + |y_1 - y_2|} \stackrel{\text{def}}{=} \rho_4\left((x_1, y_1), (x_2, y_2)\right).\end{aligned}$$

Therefore, the process $\left\{W_{\Lambda_{r_n}}(x, y) : (x, y) \in [0,1]^2\right\}$ is sub-Gaussian with respect to ρ_4 for any $n \in \mathbb{N}$ and satisfies for any $\delta > 0$,

$$\mathbb{E}\left[\sup_{\rho_4((x_1,y_1),(x_2,y_2))<\delta} \left|W_{\Lambda_{r_n}}(\theta_1) - W_{\Lambda_{r_n}}(\theta_2)\right|\right] \leq K \int_0^\delta \sqrt{\log N(\tfrac{t}{2}, T_4, \rho_4)}\mathrm{d}t, \tag{4.85}$$

for some universal constant $K > 0$, where $N(t, T_4, \rho_4)$ is the minimal number of ρ_4-balls of radius $t > 0$ needed to cover T_4. Defining $d_4 \stackrel{\text{def}}{=} \rho_4^2$, we see that for any $t > 0$,

$$N(t, T_4, \rho_4) = N(t^2, T_4, d_4).$$

Define for any $t > 0$ and $(x, y) \in T_4$,

$$B((x, y), t, d_4) \stackrel{\text{def}}{=} \{(u, v) \in T_4 : |x - u| + |y - v| \leq t\}.$$

Pythagoras' theorem shows that for $t \geq 1$,

$$T_4 \subseteq B((0.5, 0.5), t, d_4),$$

so that $N(t, T_4, d_4) = 1$ for such values of t. For $0 < t < 1$, note that

$$B((x, y), t, d_4) \supseteq \{(u, v) \in T_4 : \max\{|u - x|, |v - y|\} \leq t/2\}.$$

The latter set corresponds to a square centered at (x, y) aligned with the axes with area t^2. We need can cover T_4 with t^{-2} such squares so that for $0 < t < 1$,

$$N(t, T_4, \rho_4) = N(t^2, T_4, d_4) \leq t^{-4}.$$

In particular that the space (T_4, ρ_4) is totally bounded. Pick $\delta \in (0, 2)$. Then by previous computations, we have that for any $t \in [0, \delta]$,

$$\sqrt{\log N(\tfrac{t}{2}, T_4, \rho_4)} \leq 2\sqrt{\log(2/t)}.$$

Computations similar to the previous cases leads to

$$\int_0^{\delta} \sqrt{\log N(\tfrac{t}{2}, T_4, \rho_4)}\mathrm{d}t \leq 2\left\{2(\delta/2)\sqrt{\log(2/\delta)} + \sqrt{\pi}\,\mathrm{erfc}\left(\sqrt{\log(2/\delta)}\right)\right\}$$
$$\leq 2\delta\left\{\sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)})\right\}.$$

Equation (4.85) and Markov's inequality imply that for any $\varepsilon, \delta > 0$,

$$\mathsf{P}\left[\sup_{\rho_4((x_1,y_1),(x_2,y_2))<\delta} \left|W_{\Lambda_{r_n}}(x_1, y_1) - W_{\Lambda_{r_n}}(x_1, y_1)\right| > \varepsilon\right]$$
$$\leq 2K\varepsilon^{-1}\delta\left\{\sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)})\right\}.$$

Given $\epsilon, \eta > 0$, by picking $\delta = \delta(\epsilon, \eta) > 0$ small enough such that

$$2K\varepsilon^{-1}\delta\left\{\sqrt{\log(2/\delta)} + 1/(2\sqrt{\log(2/\delta)})\right\} < \eta,$$

we verify the asymptotic tightness condition for this case. □

For any $r \in \mathcal{R}$, let P_r denote the law of L_r on $\mathbb{R}$.

Lemma 4.G.2. *Under the same assumptions as Lemma 4.G.1 and Assumption 4.4.3, if $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ is such that $r_n \to r_0$ as $n \to \infty$ for an interior point $r_0 \in \mathcal{R}$, we have $\mathsf{P}_{r_n} \xrightarrow{\mathcal{L}} \mathsf{P}_{r_0}$ on $\mathbb{R}$.*

Proof. Let $r_0 \in \mathcal{R}$ be an interior point. Recall that

$$L_{r_0} = \int_0^{\pi/2} \left|X_{r_0}(\theta)\right| q(\theta)\mathrm{d}\theta,$$

where for $\theta \in [0, \pi/2]$,

$$X_{r_0}(\theta) \overset{\text{def}}{=} \gamma_{p,r_0}(\theta) - \int_0^{\theta} \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot I_{g,\sigma,r_0}.$$

The idea is to write

$$X_{r_0} = T_{r_0}(W_{\Lambda_{r_0}}), \tag{4.86}$$

for some linear operator $T_{r_0} : \ell^\infty(\mathcal{F}) \mapsto \ell^\infty([0, \pi/2])$ satisfying the property that if $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ is such that $r_n \to r_0$ in $\mathbb{R}^m$ and $(z_n)_{n\in\mathbb{N}} \subset \ell^\infty(\mathcal{F})$ is such that $z_n \to z$ in $\ell^\infty(\mathcal{F})$, then also

$$T_{r_n}(z_n) \to T_{r_0}(z) \qquad \text{in } \ell^\infty([0, \pi/2]). \tag{4.87}$$

The result would then follow from Lemma 4.G.1 combined with the extended continuous mapping theorem [122, Theorem 1.11.1].

A sufficient condition to ensure (4.87) is to show that the operator T_{r_0} is bounded, i.e.,

$$\|T_{r_0}\| \stackrel{\text{def}}{=} \sup\left\{\|T_{r_0}(z)\|_{\ell^\infty([0,\pi/2])} : \|z\|_{\ell^\infty(\mathcal{F})} \leq 1\right\} < \infty$$

together with

$$\lim_{n\to\infty} \|T_{r_n} - T_{r_0}\| = 0. \tag{4.88}$$

In practice, if we are able to decompose T_{r_0} as $T_r = \sum_{i=1}^k T_{r_0,i}$ for bounded linear operators $T_{r_0,i} : \ell^\infty(\mathcal{F}) \mapsto \ell^\infty([0,\pi/2])$ satisfying (4.88), then we are done in view of the triangle inequality. The same reasoning also applies to composition of such operators: if $T_n, T : E \to G$ are bounded linear operators between two normed spaces such that

$$T_n = V_n \circ U_n \qquad \text{and} \qquad T = V \circ U,$$

for some bounded linear operators $U_n, U : E \to F$ and $V_n, V : F \to G$ between normed spaces satisfying

$$\lim_{n\to\infty} \|U_n - U\|_{E,F} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \|V_n - V\|_{F,G} = 0,$$

then also

$$\lim_{n\to\infty} \|T_n - T\|_{E,G} = 0.$$

Indeed, for any $e \in E$, we have

$$\begin{aligned}\|T_n(e) - T(e)\|_G &= \|V_n(U_n(e)) - V(U(e))\|_G \\ &= \|V_n(U_n(e)) - V(U_n(e)) + V(U_n(e)) - V(U(e))\|_G \\ &\leq \|V_n - V\|_{F,G}\|U_n\|_{E,F}\|e\|_E + \|V\|_{F,G}\|U_n - U\|_{E,F}\|e\|_E.\end{aligned}$$

Since $\limsup_{n\to\infty} \|U_n\|_{E,F} < \infty$, we can conclude.

Process α_{p,r_0}. Let us first consider the linear operator $T_{\alpha,r_0} : \ell^\infty(\mathcal{F}) \mapsto \ell^\infty([0,\pi/2])$ which is such that

$$T_{\alpha,r_0}(W_{\Lambda_{r_0}}) = \alpha_{p,r_0}. \tag{4.89}$$

Our aim is to show that T_{α,r_0} is bounded and satisfies (4.88). We will assume $p < \infty$ since this corresponds to the most involved case. Note that T_{α,r_0} can be trivially decomposed as

$$T_{\alpha,r_0} = \sum_{i=1}^3 T_{\alpha,r_0,i},$$

where for $z \in \ell^\infty(\mathcal{F})$ and $\theta \in [0,\pi/2]$ we defined

$$\left(T_{\alpha,r_0,1}(z)\right)(\theta) \stackrel{\text{def}}{=} z(\mathbb{I}_{C_{p,\theta}}), \tag{4.90}$$

and

$$\left(T_{\alpha,r_0,2}(z)\right)(\theta) \stackrel{\text{def}}{=} \lambda_{r_0}(1,\tan\theta) \int_0^{x_p(\theta)} x^{-1}\Big[z\left(\mathbb{I}_{[0,x]\times[0,\infty]}w(x)\right)w(x)^{-1}\tan\theta - z\left(\mathbb{I}_{[0,\infty]\times[0,x\tan\theta]}w(x\tan\theta)\right)w(x\tan\theta)^{-1}\Big]\mathrm{d}x, \quad (4.91)$$

and

$$\left(T_{\alpha,r_0,3}(z)\right)(\theta) \stackrel{\text{def}}{=} \int_{x_p(\theta)}^{\infty} x^{-1}\lambda_{r_0}(1,(x^p-1)^{-1/p}) \Big[z\left(\mathbb{I}_{[0,x]\times[0,\infty]}w(x)\right)w(x)^{-1}y_p'(x) - z\left(\mathbb{I}_{[0,\infty]\times[0,y_p(x)]}w(y_p(x))\right)w(y_p(x))^{-1}\Big]\mathrm{d}x. \quad (4.92)$$

In view of our previous comments, we can analyse each of those operators separately.

Let $z \in \ell^\infty(\mathcal{F})$ be such that $\|z\|_{\ell^\infty(\mathcal{F})} \leq 1$. Then we immediately see that

$$\|T_{\alpha,r_0,1}(z)\|_{\ell^\infty([0,\pi/2])} \leq 1.$$

Therefore, $T_{\alpha,r_0,1}$ is bounded. Furthermore, if $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ is a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$, condition (4.88) is trivially satisfied since

$$\|T_{\alpha,r_n,1} - T_{\alpha,r_0,1}\| = 0 \qquad \text{for any } n \in \mathbb{N}.$$

Let us now turn to the second operator. Let $z \in \ell^\infty(\mathcal{F})$ be such that $\|z\|_{\ell^\infty(\mathcal{F})} \leq 1$. Then we see that

$$\|T_{\alpha,r_0,2}(z)\|_{\ell^\infty([0,\pi/2])} \leq \sup_{\theta\in[0,\pi/2]}\left\{\lambda_{r_0}(1,\tan\theta)\int_0^{x_p(\theta)} x^{-1}\left[w(x)^{-1}\tan\theta + w(x\tan\theta)^{-1}\right]\mathrm{d}x\right\},$$

so that if we show the latter supremum is finite, the operator will be bounded. First observe that by (4.23), we have for any $\theta \in [0,\pi/2]$,

$$\lambda_{r_0}(1,\tan\theta) \leq \min\left\{\tan\theta, (\tan\theta)^{-2}\right\}\varphi_{p,r_0}(\theta).$$

Furthermore, one may compute that for any $\theta \in [0,\pi/2]$,

$$\int_0^{x_p(\theta)} x^{-1}w(x)^{-1}\tan\theta\mathrm{d}x = \tan\theta\left\{\eta^{-1} + \gamma^{-1}\left(x_p(\theta)^\gamma - 1\right)\right\},$$

and

$$\int_0^{x_p(\theta)} x^{-1} w(x\tan\theta)^{-1}\mathrm{d}x = \eta^{-1} + (\tan\theta)^{\gamma}(x_p(\theta)^{\gamma} - (\tan\theta)^{-\gamma})/\gamma.$$

One may show that for any $\theta \in [0, \pi/4]$,

$$\int_0^{x_p(\theta)} x^{-1}\left[w(x)^{-1}\tan\theta + w(x\tan\theta)^{-1}\right]\mathrm{d}x \le 2\eta^{-1} + \frac{2^{\gamma/p}}{\gamma},$$

while for $\theta \in [\pi/4, \pi/2]$,

$$\frac{1}{\tan\theta}\int_0^{x_p(\theta)} x^{-1}\left[w(x)^{-1}\tan\theta + w(x\tan\theta)^{-1}\right]\mathrm{d}x \le 2\eta^{-1} + \frac{2^{\gamma/p}}{\gamma}.$$

Consequently, observing that $(\tan\theta)^{-1} = \tan(\frac{\pi}{2} - \theta)$, we find the bound

$$\sup_{\theta\in[0,\pi/2]}\left\{\lambda_{r_0}(1,\tan\theta)\int_0^{x_p(\theta)} x^{-1}\left[w(x)^{-1}\tan\theta + w(x\tan\theta)^{-1}\right]\mathrm{d}x\right\}$$
$$\le \left(2\eta^{-1} + \frac{2^{\gamma/p}}{\gamma}\right)\sup_{\theta\in[0,\pi/2]}\left\{\min\left\{\tan\theta, \tan(\tfrac{\pi}{2} - \theta)\right\}\varphi_{p,r_0}(\theta)\right\}.$$

The latter quantity is bounded in view of point 1 in Assumption 4.4.3. We deduce that the operator $T_{\alpha,r_0,2}$ is bounded. Now, let $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ be a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$. Similar computations as before show that for any $z \in \ell^\infty(\mathcal{F})$ with $\|z\|_{\ell^\infty(\mathcal{F})} \le 1$ and any $\theta \in [0, \pi/2]$, we have

$$\left|\left(T_{\alpha,r_n,2}(z)\right)(\theta) - \left(T_{\alpha,r_0,2}(z)\right)(\theta)\right|$$
$$\le \min\{\tan\theta, \tan(\tfrac{\pi}{2} - \theta)\}\left|\varphi_{p,r_n}(\theta) - \varphi_{p,r_0}(\theta)\right|\left(2\eta^{-1} + \frac{2^{\gamma/p}}{\gamma}\right).$$

The latter quantity converges to zero in view of point 2 in Assumption 4.4.3 so that condition (4.88) is verified for $T_{\alpha,r_0,2}$.

Let us finally turn to the third operator $T_{\alpha,r_0,3}$ defined by (4.92). We start by showing that it is bounded. Let $z \in \ell^\infty(\mathcal{F})$ be such that $\|z\|_{\ell^\infty(\mathcal{F})} \le 1$. One computes that

$$\|T_{\alpha,r_0,3}(z)\|_{\ell^\infty([0,\pi/2])} \le \int_1^\infty x^{-1}\lambda_{r_0}(1,(x^p-1)^{-1/p})\left[x^{\gamma}|y_p'(x)| + y_p(x)^{\gamma}\right]\mathrm{d}x,$$

so that we just need to show that the latter integral is finite. Formula (4.23) permits to show that

$$\lambda_{r_0}(1,(x^p-1)^{-1/p}) = \frac{1}{x}\cdot\frac{1}{1+(x^p-1)^{-2/p}}\cdot\varphi_{p,r_0}\left(\arctan((x^p-1)^{-1/p})\right).$$

Note also that $|y_p'(x)| = (x^p - 1)^{-(1+1/p)}$ and recall that $y_p(x) = x/(x^p-1)^{1/p}$. Considering the change of variable $u = \arctan((x^p-1)^{-1/p})$ with inverse $x = x_p(u) = \|(1, \cot u)\|_p$ and differential $\mathrm{d}x = -x_p(u)^{1-p}(\tan u)^{-(1+p)}(\cos u)^{-2}\mathrm{d}u$ and noting that $(x^p-1)^{-1/p} = \tan u$, tedious computations show that the latter integral equals

$$\int_0^{\pi/2} \varphi_{p,r_0}(u)\left(\|(1,\cot u)\|_p^{\gamma-(1+p)} + \|(1,\tan u)\|_p^{\gamma-(1+p)}\right)\mathrm{d}u.$$

The term in parenthesis is always smaller than 1 so the integrand is bounded by φ_{p,r_0}, which is integrable on $[0,\pi/2]$ and the operator $T_{\alpha,r_0,3}$ is bounded. Now, let $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ be a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$. Similar computations as before show that for any $z \in \ell^\infty(\mathcal{F})$ with $\|z\|_{\ell^\infty(\mathcal{F})} \le 1$,

$$\begin{aligned}&\|T_{\alpha,r_n,3}(z) - T_{\alpha,r_0,3}(z)\|_{\ell^\infty([0,\pi/2])}\\ &\le \int_0^{\pi/2} \left|\varphi_{p,r_n}(u) - \varphi_{p,r_0}(u)\right|\left(\|(1,\cot u)\|_p^{\gamma-(1+p)} + \|(1,\tan u)\|_p^{\gamma-(1+p)}\right)\mathrm{d}u\end{aligned}$$

By point 2 of Assumption 4.4.3 and integrability of the map $\theta \in [0,\pi/2] \mapsto (\eta(\theta))^{-\nu}$ for any $\nu \in (0,1)$, the latter quantity converges to zero, showing that condition (4.88) is verified for $T_{\alpha,r_0,3}$.

Process $\boldsymbol{\beta_{p,r_0}}$. Let us now consider the linear operator $T_{\beta,r_0} : \ell^\infty([0,\pi/2]) \mapsto \ell^\infty([0,\pi/2])$ which is such that

$$T_{\beta,r_0}(\alpha_{p,r_0}) = \beta_{p,r_0}. \tag{4.93}$$

Our aim is to show that T_{β,r_0} is bounded and satisfies (4.88).

Let $z \in \ell^\infty([0,\pi/2])$. The operator T_{β,r_0} is defined for any $\theta \in [0,\pi/2]$ by

$$\left(T_{\beta,r_0}(z)\right)(\theta) \stackrel{\text{def}}{=} \frac{z(\theta)\Phi_{p,r_0}(\frac{\pi}{2}) - \Phi_{p,r_0}(\theta)z(\frac{\pi}{2})}{\Phi_{p,r_0}(\frac{\pi}{2})^2}. \tag{4.94}$$

It is clear that if $\|z\|_{\ell^\infty([0,\pi/2])} \le 1$, since the angular distribution function satisfies $1 \le \Phi_{p,r_0}(\pi/2) \le 2$ for any $r \in \mathcal{R}$,

$$\|T_{\beta,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \le 2,$$

so that T_{β,r_0} is bounded. Now, let $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ be a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$. Similar computations show that if $\|z\|_{\ell^\infty([0,\pi/2])} \le 1$,

$$\|T_{\beta,r_n}(z) - T_{\beta,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \le 6\|\Phi_{p,r_n} - \Phi_{p,r_0}\|_{\ell^\infty([0,\pi/2])}.$$

The latter quantity converges to zero as $n \to \infty$ since $\Phi_{p,r_n} \to \Phi_{p,r_0}$ in $\ell^\infty([0,\pi/2])$ if $r_n \to r$ in $\mathbb{R}^m$, as explained in the proof of Lemma 4.G.1. We deduce that T_{β,r_0} satisfies (4.88).

Process γ_{p,r_0}. Let us finally consider the linear operator $T_{\gamma,r_0} : \ell^\infty([0,\pi/2]) \mapsto \ell^\infty([0,\pi/2])$ which is such that

$$T_{\gamma,r_0}(\beta_{p,r_0}) = \gamma_{p,r_0}. \tag{4.95}$$

Our aim is to show that T_{γ,r_0} is bounded and satisfies (4.88).

Let $z \in \ell^\infty([0,\pi/2])$. The operator T_{γ,r_0} is defined for any $\theta \in [0,\pi/2]$ by

$$\left(T_{\gamma,r_0}(z)\right)(\theta) \stackrel{\text{def}}{=} z(\theta) + \frac{\int_0^{\pi/2} z(\psi) f'(\psi) \mathrm{d}\psi}{\sigma^2_{Q_{p,r_0}}(f)} \int_0^\theta f(\psi) \mathrm{d}Q_{p,r_0}(\psi), \tag{4.96}$$

where we recall that the functions f and f' are defined by

$$f(\psi) = \frac{\sin\psi - \cos\psi}{\|(\sin\psi, \cos\psi)\|_p} \qquad \text{and}$$
$$f'(\psi) = \frac{(\sin\psi)^{p-1} + (\cos\psi)^{p-1}}{\|(\sin\psi, \cos\psi)\|_p^{1+p}}$$

satisfy the bounds $|f| \le 1$ and $0 \le f' \le 2$. It readily follows that if $\|z\|_{\ell^\infty([0,\pi/2])} \le 1$, then

$$\|T_{\gamma,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \le 1 + \frac{\pi}{\sigma^2_{Q_{p,r_0}}(f)} < \infty,$$

so T_{γ,r_0} is a bounded operator. Now, let $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ be a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$. For $\|z\|_{\ell^\infty([0,\pi/2])} \le 1$, we compute

$$\begin{aligned}
&\|T_{\gamma,r_n}(z) - T_{\gamma,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \\
&\le \pi \left\| \frac{\int_0^\theta f(\psi)\mathrm{d}Q_{p,r_n}(\psi)}{\sigma^2_{Q_{p,r_n}}(f)} - \frac{\int_0^\theta f(\psi)\mathrm{d}Q_{p,r_0}(\psi)}{\sigma^2_{Q_{p,r_0}}(f)} \right\|_{\ell^\infty([0,\pi/2])}.
\end{aligned}$$

Fubini's theorem leads to

$$\begin{aligned}
&\left\| \int_0^\theta f(\psi)\mathrm{d}Q_{p,r_n}(\psi) - \int_0^\theta f(\psi)\mathrm{d}Q_{p,r_0}(\psi) \right\|_{\ell^\infty([0,\pi/2])} \\
&\qquad\qquad \le (1+\pi)\|Q_{p,r_n} - Q_{p,r_0}\|_{\ell^\infty([0,\pi/2])},
\end{aligned}$$

and the latter quantity converges to zero as $n \to 0$ since $Q_{p,r_n} \to Q_{p,r_0}$ in $\ell^\infty([0,\pi/2])$. The same argument implies

$$\sigma^2_{Q_{p,r_n}}(f) \to \sigma^2_{Q_{p,r_0}}(f) > 0, \qquad \text{as } n \to \infty,$$

and this suffices to verify (4.88) for the operator T_{γ,r_0}.

Process associated with the derivative. The last operator to be considered is the linear operator $T_{\nabla,r_0} : \ell^\infty(\mathcal{F}) \to \ell^\infty([0,\pi/2])$ which is such that for $\theta \in [0,\pi/2]$,

$$\left(T_{\nabla,r_0}\left(W_{\Lambda,r_0}\right)\right)(\theta) = \int_0^\theta \nabla_r \varrho_{p,r_0}(x)\mathrm{d}x \cdot I_{g,\sigma,r_0}, \tag{4.97}$$

where we recall that

$$I_{g,\sigma,r_0} = \int_{[0,1]^2} \left(W_{\Lambda_{r_0}}(A_{(x_1,x_2)}) - \sum_{j=1}^2 \dot{\ell}_{r,j}(x_1,x_2)W_j(x_j)\right) g(x_1,x_2)\mathrm{d}\sigma(x_1,x_2),$$

for some finite Borel measure σ on $[0,1]^2$ and a function $g : [0,1]^2 \to \mathbb{R}^m$ in $\left(L_2([0,1]^2,\sigma)\right)^m$. The latter is naturally decomposed as $T_{\nabla,r_0} = \sum_{k=1}^m T_{\nabla,k,r_0}$ where the operators $T_{\nabla,k,r_0} : \ell^\infty(\mathcal{F}) \to \ell^\infty([0,\pi/2])$ are defined by

$$\left(T_{\nabla,K,r_0}\left(W_{\Lambda,r_0}\right)\right)(\theta) \stackrel{\text{def}}{=} \left(I_{g,\sigma,r_0}\right)_k \int_0^\theta \partial_{r_k}\varrho_{p,r_0}(x)\mathrm{d}x, \qquad k \in \{1,\dots,m\}.$$

Fix $k \in \{1,\dots,m\}$ and let $z \in \ell^\infty(\mathcal{F})$ be such that $\|z\|_{\ell^\infty(\mathcal{F})} \leq 1$. Using the fact that $\dot{\ell}_{r_0,j} \leq 1$ for any $j \in \{1,2\}$, we readily see that

$$\begin{aligned}
&\|T_{\nabla,K,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \\
&\qquad \leq 3\int_{[0,1]^2} g_k(x_1,x_2)\mathrm{d}\sigma(x_1,x_2)\int_0^{\pi/2}\left|\partial_{r_k}\varrho_{p,r_0}(x)\right|\mathrm{d}x < \infty,
\end{aligned}$$

so that T_{∇,K,r_0} is bounded since the partial derivative $\partial_{r_k}\varrho_{p,r_0}$ is assumed to be integrable on $[0,\pi/2]$ by point 1 of Assumption 4.4.1. Let us now consider a sequence $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ which is such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$. We compute

$$\begin{aligned}
&\|T_{\nabla,k,r_n}(z) - T_{\nabla,k,r_0}(z)\|_{\ell^\infty([0,\pi/2])} \\
&\leq \int_0^{\pi/2}\left|\partial_{r_k}\varrho_{p,r_n}(x)\right|\mathrm{d}x \sum_{j=1}^2 \int_{[0,1]^2}\left|\dot{\ell}_{r_n,j}(x_1,x_2) - \dot{\ell}_{r_0,j}(x_1,x_2)\right| g(x_1,x_2)\mathrm{d}\sigma(x_1,x_2) \\
&\qquad + 3\int_{[0,1]^2} g_k(x_1,x_2)\mathrm{d}\sigma(x_1,x_2)\int_0^{\pi/2}\left|\partial_{r_k}\varrho_{p,r_n}(x) - \partial_{r_k}\varrho_{p,r_0}(x)\right|\mathrm{d}x.
\end{aligned}$$

Using the fact that for any $r \in \mathcal{R}$

$$\begin{aligned}
\ell_r(x_1,x_2) &= x_1 + x_2 - \Lambda_r([0,x_1]\times[0,x_2]) \\
&= x_1 + x_2 - \iint_{[0,x_1]\times[0,x_2]} \lambda_r(x,y)\mathrm{d}x\mathrm{d}y \\
&= x_1 + x_2 - \iint_{[0,x_1]\times[0,x_2]} \frac{xy}{x^2+y^2}\|(x,y)\|_p^{-1}\varphi_{p,r}(\arctan(y/x))\mathrm{d}x\mathrm{d}y,
\end{aligned}$$

where the last equality makes use of (4.23). A change of variable then leads to

$$\dot{\ell}_{r,1}(x_1,x_2) = 1 - \int_0^{\arctan(\frac{x_2}{x_1})} \frac{\tan\theta}{(1+(\tan\theta)^p)^{1/p}} \varphi_{p,r}(\theta)\mathrm{d}\theta \quad \text{and}$$
$$\dot{\ell}_{r,2}(x_1,x_2) = 1 - \int_{\arctan(\frac{x_2}{x_1})}^{\pi/2} \frac{\tan(\frac{\pi}{2}-\theta)}{(1+(\tan(\frac{\pi}{2}-\theta)^p)^{1/p}} \varphi_{p,r}(\theta)\mathrm{d}\theta.$$

Point 2 of Assumption 4.4.3 then leads to

$$\dot{\ell}_{r_n,j}(x_1,x_2) \to \dot{\ell}_{r_0,j}(x_1,x_2), \qquad \forall (x_1,x_2) \in [0,1]^2,\ j \in \{1,2\}$$

as $n \to \infty$. The dominated convergence theorem permits to conclude that the first term in the upper bound converges to zero. Thanks to point 1 of Assumption 4.4.1 the second term also converges to zero by point 2 of Assumption 4.4.3, showing that condition (4.88) is satisfied for T_{∇,r_0}.

Combining all the above arguments permit to conclude the proof of Lemma 4.G.2. □

If $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$ is a sequence such that $r_n \to r_0$ in $\mathbb{R}^m$ as $n \to \infty$ for some interior point $r_0 \in \mathcal{R}$, Lemma 4.G.2 implies pointwise convergence of the cumulative distribution associated to L_{r_n} to the one of L_{r_0} which is continuous. Since it is also strictly increasing, we get the pointwise convergence $Q_{r_n}(1-\alpha) \to Q_{r_0}(1-\alpha)$ for any $\alpha \in (0,1)$. The continuous mapping then permits to conclude that

$$Q_{\widehat{r}_n}(1-\alpha) \xrightarrow{\mathrm{P}} Q_{r_0}(1-\alpha), \qquad \alpha \in (0,1),$$

for any consistent estimator $\widehat{r}_n$ of r_0. The local uniformity follows from the continuity and monotonicity of the map $\alpha \in (0,1) \mapsto Q_{r_0}(1-\alpha)$.

4.H Sampling from the asymptotic distribution of the test statistic

In this section, we briefly describe how to generate samples from the asymptotic distribution of T_n in Theorem 4.4.2, as it could be of independent interest. The underlying principle is not new and has been used, e.g., in [42] to draw confidence bounds for the empirical angular measure based on its asymptotic distribution. However, no detailed procedure could be found in the literature and our aim is to fill in this gap.

As in the proof of Lemma 4.G.2 in Appendix 4.G, the key observation is that the random variable L_{r_0} appearing as the limit in distribution of T_n can be written as

$$L_{r_0} = \int_0^{\pi/2} \left|X_{r_0}(\theta)\right| \, q(\theta)\, \mathrm{d}\theta,$$

for some process $X_{r_0} \in \ell^\infty([0, \pi/2])$ which is a linear transformation of the Gaussian process $W_{\Lambda_{r_0}}$ introduced before Theorem 4.3.1. From now on, we will omit the subscript r_0 to alleviate the notation. The random variable L will be approximated by the Riemann sum

$$L \approx \frac{1}{N} \sum_{i=1}^{N} |X(\theta_i)|\, q(\theta_i)$$

for a large equidistant grid $0 < \theta_1 < \ldots < \theta_N < \pi/2$. Typically, we choose $N = 1\,000$. Hence, to obtain a large number B of simulated values for L, it suffices to generate B trajectories of the process X on the chosen grid. In Section 4.5, we choose $B = 2\,000$ while the p-value computations in Section 4.6 are based on $B = 4\,000$ replications. We give an overview of the procedure below.

The first step is to obtain a realization of the underlying process W_Λ by considering a sufficiently large rectangular grid of points $\{(x_i, y_j) : i, j = 1, \ldots, M\}$ on $\mathbb{E}$ of the form

$$x_i = (i-1)h \qquad \text{and} \qquad y_j = (j-1)h,$$

for a small step size $h > 0$. Typically, we choose $h = 0.05$ and $M = 1000$. These points are used to define the cells

$$C_{ij} \stackrel{\text{def}}{=} \big\{(x, y) \in \mathbb{E} : x \in [x_i, x_{i+1}) \text{ and } y \in [y_j, y_{j+1})\big\}, \qquad i, j \in \{1, \ldots, M-1\},$$

which are considered as an approximate disjoint covering of the space $\mathbb{E}$. The fact that the cells do not cover points $(x, y) \in \mathbb{E}$ for which $x \vee y > (M-1)h$ is not supposed to affect the results too much as the variance of W_Λ is negligible on this part of the space.

Associated with those cells, we generate a matrix $W = (W_{ij})_{i,j=1,\ldots,M-1}$ of independent centered Gaussian random variables with $\mathrm{Var}(W_{ij}) = \Lambda(C_{ij})$. The property $\mathbb{E}[W_\Lambda(C) W_\Lambda(C')] = 0$ for disjoint Borel sets $C, C' \subset \mathbb{E}$ justifies the approximation

$$\widetilde{W}_\Lambda(C) \stackrel{\text{def}}{=} \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} W_{ij}\, 1\{(x_i, y_j) \in C\} \tag{4.98}$$

for any Borel set $C \subset \mathbb{E}$ to obtain a realization of the process W_Λ.

Of particular interest are sets of the form $C = C_{p,\theta}$ for some $1 \le p \le \infty$

and $\theta \in [0, \pi/2]$, for which formula (4.98) simplifies to

$$\begin{aligned}\widetilde{W}_\Lambda(C_{p,\theta}) &= \sum_{i=1}^{M-1}\sum_{j=1}^{M-1} W_{ij}\, 1\{y_j \le \min(y_p(x_i), x_i \tan\theta)\} \\ &= \sum_{i=1}^{M-1} \sum_{j=1}^{\left[\frac{y_p(x_i)}{h}+1\right]\wedge(M-1)} W_{ij}\, 1\left\{(j-1) \le (i-1)\tan\theta\right\},\end{aligned}$$

where $[x]$ denotes the integer part of $x \in \mathbb{R}$. Similar approximations based on (4.98) can be found for the marginal processes W_1 and W_2, leading to

$$\widetilde{W}_1(x) = \sum_{i=1}^{\left[\frac{x}{h}+1\right]\wedge(M-1)} W_{i\bullet} \quad \text{and} \quad \widetilde{W}_2(y) = \sum_{j=1}^{\left[\frac{y}{h}+1\right]\wedge(M-1)} W_{\bullet j},$$

where $W_{i\bullet} \overset{\text{def}}{=} \sum_{j=1}^{M-1} W_{ij}$ and $W_{\bullet j} \overset{\text{def}}{=} \sum_{i=1}^{M-1} W_{ij}$ for $(i,j) \in \{1,\dots,M-1\}^2$. For $C = A_{(x,y)}$ as introduced in Assumption 4.4.2 and appearing in the formula for $X(\theta)$, we get

$$\widetilde{W}_\Lambda\left(A_{(x,y)}\right) = \widetilde{W}_1(x) + \widetilde{W}_2(y) - \sum_{i=1}^{\left[\frac{x}{h}+1\right]\wedge(M-1)} \sum_{j=1}^{\left[\frac{y}{h}+1\right]\wedge(M-1)} W_{ij}.$$

As the random variable $X(\theta)$ consists of a linear transformation of the process W_Λ evaluated on sets of the above forms, these formulas will lead to a convenient approximation for the random vector $(X(\theta_1),\dots,X(\theta_N))$ as scalar products between W and weights, possibly depending on θ, which can be computed independently of the generation of W, before the evaluation of the scalar products, which can be done in an efficient way using parallel computing.

Wasserstein–Aitchison Generative Adversarial Networks for angular measures

5

This chapter is based on
Stéphane Lhaut, Holger Rootzén and Johan Segers
Wasserstein-Aitchison GAN for angular measures of multivariate extremes
arXiv preprint arXiv:2504.21438 (submitted to Extremes).

Abstract

Economically responsible mitigation of multivariate extreme risks – extreme rainfall in a large area, huge variations of many stock prices, widespread breakdowns in transportation systems – requires estimates of the probabilities that such risks will materialize in the future. This paper develops a new method, Wasserstein–Aitchison Generative Adversarial Networks (WA-GAN), which provides simulated values of future d-dimensional multivariate extreme events and which can be used to give estimates of such probabilities. The method is a combination of standard extreme value analysis modeling of the tails of the marginal distributions with nonparametric GAN modeling of the angular distribution. For the latter, the angular values are transformed to Aitchison coordinates in a full $(d-1)$-dimensional linear space, and a Wasserstein GAN is trained on these coordinates and used to generate new values. A reverse transformation is then applied to these values and gives simulated values on the original data scale. The method shows good performance compared to other existing methods in the literature, both in terms of capturing the dependence structure of the extremes in the data, as well as in generating accurate new extremes of the data distribution. The comparison is performed on simulated multivariate extremes from a logistic model in dimensions up to 50 and on a 30-dimensional financial data set.

5.1 Introduction

Dependence between extreme events is often of crucial importance. Extreme rainfall will be even more dangerous if it occurs at the same time at many spatial locations; costs associated with breakdowns in transportation systems are more harmful if they happen in many parts of the system; and extreme fluctuations in financial markets can cause much more damage if they occur in many parts of the economy. During the last decades the statistical research on dependent, or multivariate, extremes has grown rapidly and now contains many useful methods to model dependence between extreme values.

These methods typically build on parametric statistical models. However in the last three years papers using neural networks to simulate new multivariate extremes have started to appear. Often, but not always, these papers aim at nonparametric modeling of the dependence structure of the multivariate extreme values. The end result in the papers is an algorithm which can produce simulated values from the neural network "model" for the observations. A common approach is to transform the marginal distributions to some standard scale, then to combine nonparametric neural network modeling of the dependence structure with a parametric model from extreme value analysis (EVA) to simulate values on the standard scale, and finally to apply the inverse of the marginal transform to get simulated values on the real scale. Section 5.2.1 below gives a brief survey of some of this literature. We refer to [4] for a more extensive theoretical and simulation based overview of the area.

Here we develop a new method, WA-GAN, which combines parametric modeling of the marginal tails with nonparametric modeling of dependence in multivariate extremes. Our main assumptions are that the tails of the marginal distributions asymptotically follow a one-dimensional Generalized Pareto (GP) distribution and that after transformation to the unit-Pareto scale, the observations are multivariate regularly varying. Our method proceeds by the following main steps.

1. The d-dimensional observations are transformed to the unit-Pareto scale using the empirical marginal distribution functions.

2. Using the L_1-distance and a suitably large radius r, the transformed observations with radius larger than r are separated into a one-dimensional radial value and an angular value on the L_1-sphere.

3. Using Aitchison coordinates from compositional data, the angular values are transformed to values on the full $\mathbb{R}^{d-1}$ space.

4. A Wasserstein GAN is trained to use the Aitchison coordinates to produce simulated values on the L_1-sphere.

5. A result on multivariate GP distributions is used to transform the simulated values back to multivariate GP values on the original scale.

The next section contains the background needed for this together with a result (Proposition 5.2.1) providing the theoretical tool to transform the angular data of step 4 above to the real GP scale as described in step 5. Section 5.3 describes the details of WA-GAN, and provides the algorithms which are used by it. In Section 5.4.2 WA-GAN is compared to HTGAN [57] and GPGAN [85] on simulated data with a logistic dependence structure in dimension $d \in \{10, 50\}$ and, in Section 5.4.3, on financial data consisting of daily returns for $d = 30$ industry portfolios. Section 5.5 sums up the results of this paper.

5.2 Background

This section starts with a brief survey of the literature on neural network modeling of multivariate extremes. Section 5.2.2 gives background on regular variation and EVA and develops some new theory. Sections 5.2.3 and 5.2.4 contain brief recapitulations of Wasserstein GANs and of Aitchison coordinates. These are the building blocks which we use in Section 5.3 to construct the WA-GAN method.

5.2.1 A brief literature survey

For the terminology used in this survey see Section 5.2.2 below and see monographs on extreme value theory such as [13], [34] and [107] or more modern treatments such as [105] or [92].

In [3] a one-layer Re-LU network is combined with a new parametrization of a GAN to generate multivariate extreme data from heavy-tailed distributions. The method is called EV-GAN and is shown to be superior to a standard GAN.

ExtVAE [80] is based on the assumption that the observations are multivariate regularly varying and that the latent variable in the variational autoencoder (VAE) follows an inverse Gamma distribution and, applying some further constraints to insure regular variation, simulates values of the radial component. It then uses another VAE to produce simulated angles on the unit L_1-sphere using a VAE with Dirichlet distributed latent variables. Simulated observations are then obtained by multiplying the radial variable with the the angular value.

Under the assumption that estimates of the true marginal tail indices are are available, HTGAN [57] transforms multivariate observations to the unit-Pareto scale, trains a GAN with heavy-tailed latent variables and transforms simulated variables back to original scale. It generates samples from the entire data set, not just from the extreme part.

GP-GAN [85] adaptively selects exceedance levels, fits univariate GP distributions to the excesses of these levels, uses a standard GAN to propose a

spectral value on the exponential scale, and then uses the fitted univariate GP distributions and the spectral value to generate a multidimensional value on the original scale. This value is then sent to the discriminator for optimization of the spectral GAN.

The evtGAN [19] is aimed at multivariate block maxima observations. It uses one-dimensional Generalized Extreme Value (GEV) distributions to transform marginals to approximately uniform $[0, 1]$ distributions, then trains a convolutional GAN on the uniform scale, and finally applies the inverse transform of the marginal distribution to obtain simulated values on the scale of the observations. The structure of a multivariate GEV distribution is not used in the simulation of the uniform variables.

The paper [65] uses the GEV model to transform observations of block maxima to the Pickands dependence function scale and then uses either the spectral measure or the Pickands measure restrictions to train a GAN to simulate from the dependence function. The simulated observations are then transformed back to original scale. The authors argue that this approach also is useful for efficient simulation from parametric models.

Finally, DeepGauge [75] applies a different framework than the references above derived from the theory of geometric extremes. It uses a multi-layer perceptron with ReLu activation function to first model radial thresholds and then another multi-layer perceptron to find a gauge function which describes the boundary of the deterministic limiting shape of scaled sample clouds.

Parallel works. During the development of this work, several other generative methods for (multivariate) extremes were also proposed in the literature, many of which were submitted to the same special issue of the journal Extremes. Although a formal comparison with these methods is beyond the scope of this work for obvious reasons, we highlight several that are particularly relevant to our topic. Approaches based on normalizing flows have been explored in the context of geometric extremes [94] and within the standard multivariate threshold exceedance framework [71]. The ExcessGAN model [2] employs a GAN-based architecture with a specifically designed network structure aimed at reducing bias. A broader comparison of many existing generative methods is provided in [130].

5.2.2 Regular variation and extreme value analysis

Multivariate regular variation and angular measure. Let $X = (X_j)_{j=1}^d$ be the random vector of interest with values in $\mathbb{R}^d$. Based on a random sample of the (unknown) distribution of X, we wish to generate new samples with the same tail behavior as X, even at levels beyond those encountered in the sample. Such extrapolation is possible only if we are willing to make some

regularity assumptions. Our core assumptions to obtain accurate modeling of the tail of X are founded in the theory of multivariate regular variation [106]. There are two approaches to use this hypothesis.

1. Either we postulate it on the random vector X itself. This implies that all univariate tail indices are same.

2. Or we postulate it on a transformed vector $V = v(X)$, where v is a coordinatewise transformation of X so that the d components of V all have the same distribution. This implies that the multivariate assumption only depends on the dependence structure of X and not on the margins, which can be assessed separately.

Here we choose the second approach and transform to unit-Pareto margins using the transformation

$$v : x = \left(x_j\right)_{j=1}^{d} \mapsto v(x) \stackrel{\text{def}}{=} \left(\frac{1}{1 - F_j(x_j)}\right)_{j=1}^{d},$$

where F_j denotes the distribution function of X_j. We assume throughout that these marginal distribution functions are continuous. It is straightforward to see that after this transformation the margins of $V = v(X)$ have unit-Pareto distributions, $\mathrm{P}(V_j > y) = 1/y$ for $y \geq 1$. Our first assumption describes the asymptotic dependence structure of V.

Assumption 5.2.1 (Multivariate regular variation)**.** *There exists a non-zero Borel measure ν on $\mathbb{E} \stackrel{\text{def}}{=} [0, \infty)^d \setminus \{(0, \ldots, 0)\}$ which is finite on Borel sets bounded away from 0 and such that*

$$\lim_{t \to \infty} t \, \mathrm{P}\left(t^{-1} V \in B\right) = \nu(B)$$

for all Borel sets B of $\mathbb{E}$ bounded away from 0 with $\nu(\partial B) = 0$, with ∂B the topological boundary of B.

Intuitively, the measure ν helps in describing the tail behaviour of V as for large $t > 0$, we have $\mathrm{P}(V \in tB) \approx \nu(B)/t$ where $tB \stackrel{\text{def}}{=} \{tx : x \in B\}$. The measure ν is often referred to as the *exponent measure* of V, because of its appearance in the exponent of the expression of the multivariate extreme value distribution to which the law of V is attracted [34, Definition 6.1.7].

Our procedure below will benefit from an equivalent formulation of Assumption 5.2.1 in polar coordinates with respect to the L_1-norm $\|x\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^{d} |x_j|$ for $x \in \mathbb{R}^d$. More concretely, it can be shown [107, Theorem 6.1] that due to a convenient homogeneity property of ν [34, Theorem 6.1.9], if we put

$$R \stackrel{\text{def}}{=} \|V\|_1 \qquad \text{and} \qquad W \stackrel{\text{def}}{=} R^{-1} V, \tag{5.1}$$

where W takes values in the unit simplex $\Delta_{d-1} \overset{\text{def}}{=} \{x \in [0,1]^d : \|x\|_1 = 1\}$, there exists a probability measure Q on Δ_{d-1} such that

$$\lim_{t\to\infty} \mathsf{P}\left(W \in A \mid R \geq t\right) = Q(A) \tag{5.2}$$

for any Borel set $A \subseteq \Delta_{d-1}$ such that $Q(\partial A) = 0$.

The measure Q is often referred to as the *angular (probability) measure* of V as it describes how its "angular" component W behaves when its radial part R becomes large. The set of possible probability measures that can be obtained in such a way forms a nonparametric class as the only constraint they should satisfy—due to the unit-Pareto margins of V—are

$$\int_{\Delta_{d-1}} w_j \, \mathrm{d}Q(w) = \frac{1}{d}, \qquad j \in \{1, \ldots, d\}, \tag{5.3}$$

winch in turn implies

$$\mathsf{P}(R > t) \sim \frac{d}{t}, \qquad \text{as } t \to \infty. \tag{5.4}$$

It is possible for Q to have positive mass on one of the faces of the simplex Δ_{d-1}, that is, sets for which at least one of the coordinates is zero. In terms of extremes, it corresponds to scenarios where some components of X are large while others are small. These are called *extreme directions* and have to be treated by specific methods which our methodology is not able to handle, see, e.g., [96]. We exclude this situation here.

Assumption 5.2.2. *The measure Q in* (5.2) *is concentrated on the open simplex* $\Delta_{d-1}^{\mathrm{o}} \overset{\text{def}}{=} \{x \in (0,1)^d : \|x\|_1 = 1\}$, *that is, if* $\Theta \sim Q$, *then*

$$\mathsf{P}\left(\Theta \in \Delta_{d-1}^{\mathrm{o}}\right) = 1.$$

Univariate generalized Pareto distributions. Assumption 5.2.1 is not sufficient by itself to model the tail behaviour of X as it only concerns the dependence structure. Our second main assumption concerns the tails of the margins X_j and allows for heterogeneous tails. Here and below, u_{*j} denotes the (possibly infinite) upper endpoint of X_j.

Assumption 5.2.3 (Generalized Pareto limit of marginal tails)**.** *For every* $j \in \{1, \ldots, d\}$, *there exist* $\xi_j \in \mathbb{R}$ *and a function* $\sigma_j(\cdot) : (-\infty, u_{*j}) \to (0, \infty)$ *such that for all* $y > 0$,

$$\lim_{u \nearrow u_{*j}} \mathsf{P}\left(\frac{X_j - u}{\sigma_j(u)} > y \,\middle|\, X_j > u\right) = \left(1 + \xi_j y\right)_+^{-1/\xi_j},$$

where $a_+ \overset{\text{def}}{=} \max(a, 0)$ *for* $a \in \mathbb{R}$; *if* $\xi_j = 0$, *the limit is to be understood as* $\exp(-y)$.

By the Pickands–Balkema–de Haan theorem [9, 103], Assumption 5.2.3 is equivalent to the assumption that X_j is in the maximum domain of attraction of a GEV distribution. The limit in Assumption 5.2.3 holds uniformly in $y > 0$. The assumption justifies the following approximation for $y > 0$ such that $1 + \xi_j y/\sigma_j > 0$:

$$\mathrm{P}\left(X_j - u > y \,\middle|\, X_j > u\right) \approx \left(1 + \frac{\xi_j y}{\sigma_j}\right)^{-1/\xi_j}, \tag{5.5}$$

where $\sigma_j > 0$ is a scale parameter that accounts for the unknown function $\sigma_j(\cdot)$. Estimation of the bivariate parameter (σ_j, ξ_j) is typically performed by the method of maximum likelihood. This is the well-known *Peaks-Over-Threshold (PoT)* approach for modeling the tails of real-valued data, see for instance [117] and [13, Section 5.3]. A random variable with distribution function

$$H_{\sigma,\xi}(y) \stackrel{\text{def}}{=} 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \qquad y > 0,$$

for some $(\sigma, \xi) \in (0, \infty) \times \mathbb{R}$ is said to have a *generalized Pareto (GP) distribution.*

Remark 5.2.1 (Equivalent formulations of Assumption 5.2.3). As already explained, Assumption 5.2.3 is equivalent to the hypothesis that X_j is in the maximum domain of attraction of a GEV distribution. As a consequence, X_j satisfies all the equivalent formulations of this condition as presented, e.g., in Theorem 1.1.6 of [34]. As some of them will be needed below, we recall them here. Let $b_j(t) \stackrel{\text{def}}{=} F_j^{-1}(1 - 1/t) = (1/(1 - F_j))^{-1}(t)$, for $t > 1$, be the tail quantile function of F_j. Then, the following statements are equivalent to Assumption 5.2.3.

(i) There exist $\xi_j \in \mathbb{R}$ and a scaling function $a_j(\cdot) : (1, \infty) \to (0, \infty)$ such that

$$\lim_{t\to\infty} \mathrm{P}\left(\frac{X_j - b_j(t)}{a_j(t)} > y \mid X_j > b_j(t)\right) = \left(1 + \xi_j y\right)_+^{-1/\xi_j}, \qquad y > 0.$$

(ii) There exist $\xi_j \in \mathbb{R}$ and a scaling function $a_j(\cdot) : (1, \infty) \to (0, \infty)$ such that

$$\lim_{t\to\infty} \frac{b_j(ty) - b_j(t)}{a_j(t)} = \frac{y^{\xi_j} - 1}{\xi_j}, \qquad y > 0,$$

where the right-hand side is to be interpreted as $\log y$ if $\xi_j = 0$. The convergence also holds locally uniformly in $y \in (0, \infty)$, see Section 1.2 in [34].

The scalar $\xi_j \in \mathbb{R}$ in (i) and (ii) is equal to the one in Assumption 5.2.3, while the scaling function $a_j(\cdot)$ in (i) and (ii) is related to $\sigma_j(\cdot)$ in Assumption 5.2.3 via $\sigma_j(u) = a_j(1/(1 - F_j(u)))$ for $u < u_{*j}$.

Multivariate generalized Pareto distributions. What do our Assumptions tell us about the tail behaviour of X? The answer can be conveniently formulated in terms of multivariate generalized Pareto (MGP) distributions, which extend Assumption 5.2.3 to the multivariate case. These distributions have been introduced in [111] and are reviewed in [97]. Below, we provide an introduction to their usage in modeling multivariate extremes along with a theoretical result (Proposition 5.2.1) which will be useful for sampling from a MGP distribution.

In arbitrary dimension, it is not clear how to define an "extreme" point. Some authors are interested in modeling the vector of componentwise maxima of the data, leading to the multivariate GEV distributions, studied extensively in the literature after the pioneering work of [36]. Recently, see e.g. [110], more attention has been given to an extension of the standard PoT procedure in a multivariate context. The approach relies on the fact that the excess of X above a large threshold vector u asymptotically follows an MGP distribution, with an appropriate definition of when an exceedance over a multivariate threshold takes place.

More concretely, given a vector $u \in \mathbb{R}^d$ of "high thresholds", i.e., below but close to $u_* = (u_{*j})_{j=1}^d$, we consider the random vector of excesses

$$Y_u \overset{\text{def}}{=} X - u \mid X \not\leq u, \tag{5.6}$$

where $X - u \overset{\text{def}}{=} (X_j - u_j)_{j=1}^d$ and where $\not\leq$ means that, for at least one $j \in \{1, \dots, d\}$, we have $X_j > u_j$. Here and below, operations on real numbers are extended to vectors in $\mathbb{R}^d$ in a coordinate-wise manner. Under the aforementioned assumptions, we can establish the asymptotic distribution of Y_u as $u \to u_*$, in a particular way made precise in the statement of the result. Weak convergence in $\mathbb{R}^d$ is denoted by $\xrightarrow{\mathcal{L}}$.

Proposition 5.2.1 (Weak convergence of excesses to the MGP distribution)**.** *Under Assumptions 5.2.1–5.2.3, we have*

$$\frac{X - b(t)}{a(t)} \,\Big|\, X \not\leq b(t) \xrightarrow{\mathcal{L}} \frac{Y^\xi - 1}{\xi}, \qquad t \to \infty, \tag{5.7}$$

where Y is a multivariate Pareto *random vector, with distribution*

$$Y \overset{\mathcal{L}}{=} (R\Theta \mid R\Theta \not\leq 1) \tag{5.8}$$

with R a unit-Pareto random variable independent of $\Theta \sim Q$, and where $\xi = (\xi_j)_{j=1}^d$ is the vector of marginal coefficients as in Assumption 5.2.3; $a(\cdot) = (a_j(\cdot))_{j=1}^d$ and $b(\cdot) = (b_j(\cdot))_{j=1}^d$ are the same functions as the one introduced in

Remark 5.2.1. Consequently, along the curve $u : t \in (1, \infty) \mapsto u(t) \overset{\text{def}}{=} b(t)$ *we have,*

$$\frac{Y_{u(t)}}{\sigma(u(t))} \overset{\mathcal{L}}{\to} \frac{Y^{\xi} - 1}{\xi}, \qquad t \to \infty, \tag{5.9}$$

where $\sigma(\cdot) = (\sigma_j(\cdot))_{j=1}^{d}$ *contains the scaling functions in Assumption 5.2.3.*

The proof of Proposition 5.2.1 is provided in Appendix 5.A.

Typically, as for the univariate case, the scaling function in (5.9) is viewed as a scaling parameter and Proposition 5.2.1 leads to the approximation in distribution for

$$Y_{u(t)} \overset{\mathcal{L}}{\approx} H \overset{\text{def}}{=} \sigma \frac{Y^{\xi} - 1}{\xi}, \qquad t \to \infty \tag{5.10}$$

with $\sigma \in (0, \infty)^d$ and $\xi \in \mathbb{R}^d$ containing, respectively, the scale and shape parameters from the marginal GP distributions in Assumption 5.2.3 and where Y is Multivariate Pareto distributed as in (5.8).

Remark 5.2.2 (Standard MGP random vectors). The random vector H in (5.10) is MGP(σ, ξ, S) distributed, in the sense of [97, Definition 2.2], with *spectral generator* S having distribution

$$\mathsf{P}(S \le x) = \frac{\mathsf{E}\left[\exp(Q)\, \mathbb{I}\{U \le x + Q\}\right]}{\mathsf{E}[\exp(Q)]}, \qquad x \in \mathbb{R}^d, \tag{5.11}$$

where $U \overset{\text{def}}{=} \log(\Theta)$, with $\Theta \sim Q$, and $Q \overset{\text{def}}{=} \max(U)$. In other words, the random vector U is a U-generator of H in the sense of Proposition 9 of [109]; see also equation (12) in [97]. Note in particular that the moment condition $0 < \mathsf{E}[\exp(U_j)] < \infty$ for all $j \in \{1, \ldots, d\}$ is satisfied as we have, from (5.3),

$$\mathsf{E}[\exp(U_j)] = \mathsf{E}[\Theta_j] = \int_{\Delta_{d-1}} w_j \, \mathrm{d}Q(w) = \frac{1}{d}, \qquad j \in \{1, \ldots, d\}.$$

Here, since we work mainly with the angular measure Q in our procedure, we decided to consider the multivariate Pareto random vector Y, as in (5.8), to be the "standard" MGP, that is with $\xi = 1$ and $\sigma = 1$ instead of the choice $\xi = 0$ made in [109, 97, 110], for example. This approach has also been considered in [53]. Of course, it suffices to take $Z = \log(Y)$ to obtain an MGP random vector with parameter $\xi = 0$, as can be seen from comparing (5.10) to Equation (4) in [97].

Tail modeling and rare event probabilities. Thanks to Proposition 5.2.1, we have now a clear path to modeling the tail of X. Suppose we are interested in estimating a probability $\mathsf{P}(X \in C)$ for some $C \subseteq \{x \in \mathbb{R}^d : x \not\le u(t)\}$ for some large $t > 1$, with the same $u(t)$ as in (5.9), so that the approximation of

$Y_{u(t)} = X - u(t) \mid X \not\leq u(t)$ by the MGP H as in (5.10) is accurate. In such a region, there may be no or few observations available, so that an empirical estimate is not reliable. Instead, we write the probability of interest as

$$\begin{aligned} \mathsf{P}(X \in C) &= \mathsf{P}(X \not\leq u(t))\ \mathsf{P}\left(Y_{u(t)} \in C - u(t)\right) \\ &\approx \mathsf{P}(X \not\leq u(t))\ \mathsf{P}\left(H \in C - u(t)\right). \end{aligned} \tag{5.12}$$

The first probability can be estimated using an empirical evaluation on the sample while the second one will be computed using an estimate of the MGP distribution of H.

As can be seen from (5.10), the MGP random vector H is a simple transformation of the standard MGP random vector Y introduced in (5.8), involving only marginal parameters for which estimation is well developed mathematically and numerically. Therefore, the main challenge to obtain informative samples to evaluate tail probabilities related to X is to obtain realizations of Y. By Proposition 5.2.1, this is possible if we are able to simulate from Q—at least approximately since it is a limit distribution by nature, so that no direct observation of it will ever be available.

We propose to use a specific GAN structure to achieve this goal. The main principles underlying this framework are recalled in the following section.

5.2.3 Wasserstein Generative Adversial Networks

GANs have been introduced by Ian Goodfellow and co-authors in [61]. Initially, they have been proposed for synthetic image generation but since then, they have been applied to any kind of data, in particular tabular data which are of particular interest to statisticians [5].

Suppose we are interested in estimating the unknown distribution P of some, potentially complex, data on a large vector space $\mathcal{X}$. For instance, think of the angular measure Q in Section 5.2.3 supported on the simplex Δ_{d-1} for large d. Then, the standard parametric approaches based on densities may not be satisfactory as they are often too restrictive to take into account all the possible characteristics of a complex angular measure in large dimension. An alternative direction is to start from a simple random variable $Z \in \mathcal{Z}$ in some "latent" space–typically, of lower dimension than $\mathcal{X}$–whose distribution is known and to look for a parametric transformation $g_\theta : \mathcal{Z} \to \mathcal{X}$ such that $g_\theta(Z)$ has a distribution sufficiently close to P. Practically, this has several advantages. We get more flexibility as the map g_θ can be virtually anything if we consider a sufficiently rich family of parametric functions and the latent space/distribution can also be selected freely. It is easier to simulate from P as it suffices to obtain a sample of Z (easy) and transform it based on g_θ rather than more costly algorithms based on the knowledge of the density values.

A Wasserstein GAN (WGAN) architecture [7] follows this path to estimate P by considering g_θ to be a neural network and by measuring the difference in distribution between the real data distribution P and the law of $g_\theta(Z)$ via the *Earth Mover's Distance (EMD)*, which is a particular instance of the Wasserstein distance in optimal transport [126].

A WGAN consists of two neural networks: a so-called "Generator" $G = G_\theta : \mathcal{Z} \mapsto \mathcal{X}$ and a "Discriminator" $D = D_w : \mathcal{X} \mapsto \mathbb{R}$ (also sometimes called "Critic" depending on the authors). These two networks play an opposite role: G_θ aims to produce samples as realistic as possible and close to the target distribution and D_w aims at efficiently discriminating samples and determining which ones are generated by G_θ and which ones are real. This "game" can be formulated as a min-max optimization problem in the EMD setup that we recall know.

The EMD is simply a distance between probability measures with finite first moment defined on a metric space $(\mathcal{X}, \rho)$. If P and Q are two such distributions, it is equivalently defined as

$$W_1(\mathsf{P}, \mathsf{Q}) \overset{\text{def}}{=} \inf_{\pi \in \Pi(\mathsf{P},\mathsf{Q})} \int_{\mathcal{X}\times\mathcal{X}} \rho(x, y)\, \mathrm{d}\pi(x, y) \tag{5.13}$$

$$= \sup_{f \in \mathrm{Lip}_1} \left\{ \int_{\mathcal{X}} f(x)\, \mathrm{dP}(x) - \int_{\mathcal{X}} f(x)\, \mathrm{dQ}(x) \right\}, \tag{5.14}$$

where $\Pi(\mathsf{P}, \mathsf{Q})$ is the set of probability measures on $\mathcal{X} \times \mathcal{X}$ with P and Q as first and second margins and Lip_1 is the set of functions $f : \mathcal{X} \to \mathbb{R}$ that are 1-Lipschitz continuous, that is, functions that satisfy

$$\sup_{x,y\in\mathcal{X},x\neq y} \frac{|f(x) - f(y)|}{\rho(x, y)} \leq 1.$$

If $\mathcal{X}$ is an open subset of $\mathbb{R}^N$ for some $N \in \mathbb{N}$, this implies

$$\|\nabla f(x)\|_2 \leq 1 \qquad \text{for almost every } x \in \mathcal{X}, \tag{5.15}$$

that is, for every $x \in \mathcal{X}$ except for a subset of Lebesgue measure zero. The equivalence between (5.13) and (5.14) is known as the Kantorovich–Rubinstein duality theorem, the proof of which is to be found in, e.g., [126, Section 1.2].

In formulation (5.14) of the EMD, the discriminator D_w will play the role of the "separating" function f and Q will correspond to the measure generated by the generator G_θ, that is, the distribution of $G_\theta(Z)$, while P will again be our target measure from which we observe a random sample. This leads to the following min-max procedure:

$$\min_{\theta\in\mathcal{S}_G} \max_{w\in\mathcal{S}_D} \left\{ \int_{\mathcal{X}} D_w(x)\, \mathrm{dP}(x) - \int_{\mathcal{Z}} D_w\left(G_\theta(z)\right)\, \mathrm{dP}_Z(z) \right\}, \tag{5.16}$$

where P_Z denotes the law of the latent vector Z and $\mathcal{S}_G$ and $\mathcal{S}_D$ denote the (finite-dimensional) parameter spaces of the generator and the discriminator respectively. Optimization is performed in the respective parameter spaces of the generator and the discriminator, which contain all the weights in their respective network structures. Typically, the expectations are replaced by averages over batches of real data with distribution P and fake data with distribution P_Z. The joint minimization/maximization of the objective function is performed by gradient descent based on the standard backpropagation algorithm. The whole GAN procedure is summarized in Figure 5.1.

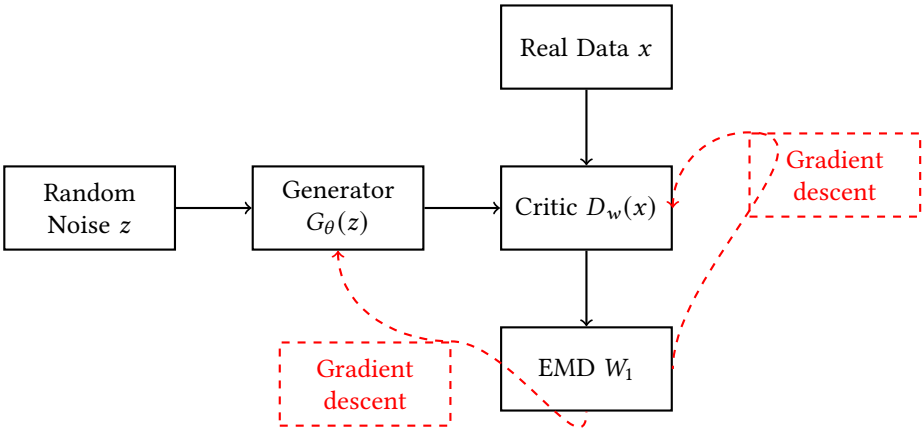


Figure 5.1: Illustration of a Wasserstein GAN with gradient-based optimization

In practice there is no need for the neural network D_w to be Lipschitz continuous as required in definition (5.14) of the EMD. Initially, in [7], authors proposed to clip the weights, that is, forcing them to lie in a fixed compact space, typically $[-c, c]^{|\mathcal{S}_D|}$ for some small $c > 0$. One may show that if the discriminator consists of a multilayer perceptron, which will always be the case here, its Lipschitz norm will be bounded by a finite constant depending only on c. Even though this constant may not be equal to one, replacing Lip_1 by Lip_C for $C > 0$ in (5.14) only amounts to multiplying the value of the resulting distance by $C > 0$ and it does not play a role in the optimization of the generator.

Since then, it has been shown that in some cases it is preferable to enforce the Lipschitz constraint in a soft manner, using (5.15). Of course, controlling the gradient ∇D_w everywhere on $\mathcal{X}$ is not possible, so what is usually done is to add a penalty term to the objective function of the form

$$\lambda \int_{\mathcal{X}} \left(\|\nabla D_w(x)\|_2 - 1\right)^2 \, \mathrm{d}\mu(x), \qquad \lambda > 0, \tag{5.17}$$

where μ is a probability measure on $\mathcal{X}$. The standard approach in [63] is to take μ as the law of the mixture $U \cdot X_r + (1-U) \cdot G_\theta(Z)$ where $X_r \sim \mathrm{P}$ follows the data distribution, $Z \sim \mathrm{P}_Z$ is random noise and U is uniformly distributed on $[0,1]$ and independent of the rest. This choice is motivated by Proposition 1 in [63] and we follow this approach here.

As explained at the end of Section 5.2.2 and in the beginning of this section, our aim is to use this architecture to obtain accurate samples from the angular measure Q on Δ_{d-1}. Since the simplex structure of such observations makes their distribution a bit non-standard and could prevent the generator from producing high-quality output, we propose to apply the procedure on a transformation of the angular data to coordinates in an orthonormal basis of the simplex and to transform the coordinates back to angles afterwards. The next section and Appendix 5.B provide more details about this transformation.

5.2.4 Aitchison coordinates

A core idea of compositional data analysis [1, 39], mainly developed by John Aitchison (1926–2016), consists of transforming raw simplex data in $\mathbb{R}^d$ to coordinates with respect to an orthonormal basis of a certain $(d-1)$-dimensional subspace of $\mathbb{R}^d$. This makes interpretation more complicated, but having data on a full linear $(d-1)$-dimensional sample space instead of on the unit simplex in $\mathbb{R}^d$ can improve statistical and neural network modeling substantially. Since we are not all that concerned with interpretation in the WGAN architecture described in Section 5.2.3 but instead with obtaining the highest possible efficiency in producing accurate new multivariate extreme samples, we will apply this idea in our context.

Infinitely many orthogonal bases exist for the open simplex $\Delta^{\mathrm{o}}_{d-1}$ equipped with its inner product space structure. Here, we work with the one proposed in Proposition 2 of [39]. Details of the underlying construction and more information on the Aitchison simplex, i.e., the unit simplex equipped with a certain vector space structure and an inner product, can be found in Appendix 5.B. The essence of the matter is to map $\Delta^{\mathrm{o}}_{d-1}$ to $\mathbb{H} = \{x \in \mathbb{R}^d : \sum_{j=1}^d x_j = 0\}$ by means of the centered logratio function (clr) in (5.29). The set $\mathbb{H}$ is a $(d-1)$-dimensional linear space equipped with the standard Euclidean inner product, the structure of which carries over to $\Delta^{\mathrm{o}}_{d-1}$ through the function clr.

Proposition 5.2.2 (An orthonormal basis for the Aitchison simplex)**.** *The vectors $\boldsymbol{e}^*_1, \ldots, \boldsymbol{e}^*_{d-1}$ defined by*

$$\boldsymbol{e}^*_i \overset{\text{def}}{=} \operatorname{softmax}(e_i) \in \Delta^{\mathrm{o}}_{d-1} \text{ with } e_i \overset{\text{def}}{=} \sqrt{\frac{i}{i+1}} \left(\underbrace{i^{-1}, \ldots, i^{-1}}_{i \text{ times}}, -1, 0 \ldots, 0 \right) \in \mathbb{R}^d$$

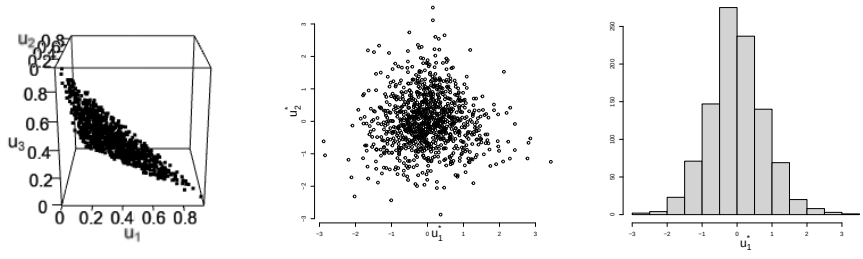
for $i \in \{1, \ldots, d-1\}$ form an orthonormal basis of the Aitchison simplex.

Now that we have constructed an orthonormal basis of the Aitchison simplex in Proposition 5.2.2, we are able to decompose any vector $u \in \Delta^{\mathrm{o}}_{d-1}$ in this basis as

$$u = \bigoplus_{i=1}^{d-1} \langle u, \boldsymbol{e}_i^* \rangle_A \odot \boldsymbol{e}_i^* = \bigoplus_{i=1}^{d-1} \langle \mathrm{clr}(u), \mathrm{clr}(\boldsymbol{e}_i^*) \rangle \odot \boldsymbol{e}_i^* = \bigoplus_{i=1}^{d-1} \langle \mathrm{clr}(u), \boldsymbol{e}_i \rangle \odot \boldsymbol{e}_i^*.$$

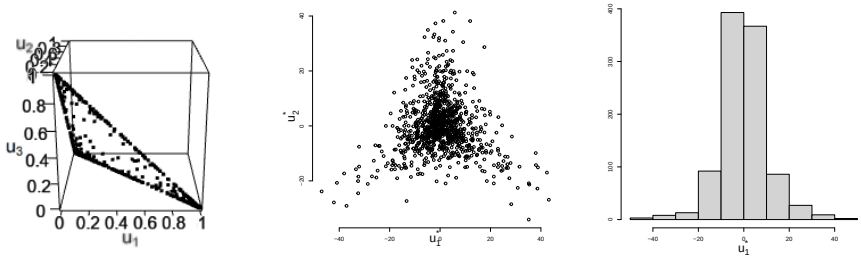
the notation being defined in Appendix 5.B. The coordinates $u_i^* \overset{\mathrm{def}}{=} \langle \mathrm{clr}(u), \boldsymbol{e}_i \rangle$ for $i \in \{1, \ldots, d-1\}$ are sufficient to completely describe the vector u. These coordinates will be the quantities entering the WGAN structure described in Section 5.2.3 as they typically have a nicer distribution than the original simplex random vectors of interest and hence will be easier to "learn", which helps facing the difficulty that sample sizes in extreme value analysis are often not very large.

One possible drawback of this method is that, if the simplex data are concentrated around the axes, corresponding to a situation with less tail dependence, the corresponding Aitchison coordinates will exhibit more heaviness and this may not catch up well with the Gaussian random vector we use as "prior" for the Generator in the Wasserstein GAN procedure described in Section 5.2.3. This is illustrated below with $d = 3$ and $n = 1000$ data points sampled from a Dirichlet distribution with parameter vector $\alpha = (2, 2, 2)$ (Figure 5.2) and $\alpha = (0.1, 0.1, 0.1)$ (Figure 5.3). In each situation, we display the raw simplex data, the corresponding coordinates in the orthonormal basis introduced in Proposition 5.2.2 and the frequency histogram of the first coordinate values.



(a) Raw simplex data **(b) Aitchison coordinates** **(c) Histogram of the first coordinate values**

Figure 5.2: The Aitchison transform for Dirichlet distributed data in $d = 3$ with parameter $\alpha = (2, 2, 2)$.



(a) Raw simplex data **(b) Aitchison coordinates** **(c) Histogram of the first coordinate values**

Figure 5.3: The Aitchison transform for Dirichlet distributed data in $d = 3$ with parameter $\alpha = (0.1, 0.1, 0.1)$.

5.3 Combining GAN, Aitchison coordinates and Extremes

Recall that our main objective is to provide an algorithm which is able to accurately sample from the tail of a random vector $X \in \mathbb{R}^d$. We do this by combining Algorithm 1 and Algorithm 2 described below.

- Algorithm 1 develops a Wasserstein GAN to simulate from the dependence structure of the unit-Pareto transformed observations.
- Algorithm 2 uses estimates of the marginal GPD parameters to transform these simulated values to the original scale.

The simulated values can then be used to obtain accurate estimates of probabilities of the form $\mathsf{P}(X \in C)$ where $C \subseteq \{x \in \mathbb{E} : x \not\leq u(t)\}$ for some "high" threshold vector $u(t) \in \mathbb{R}^d$ as in (5.9), close to u_*.

The main challenge lies in simulation from the angular measure Q of V in Assumption 5.2.1, since moving from Q to the MGP distribution is a matter of scaling and marginal tail estimation, for which theory and practice are well developed. Consequently, we propose to use the WGAN architecture of Section 5.2.3 to obtain accurate samples from Q. This is done via the Aitchison coordinates, Section 5.2.4, as detailed in Algorithm 1 below.

The input to Algorithm 1 is a set of training observations $X_1, \ldots, X_n$, with $X_i = (X_{i1}, \ldots, X_{id})$, which will be empirically converted to their unit-Pareto margins version using the empirical marginal cumulative distribution functions

$$\widehat{F}_j(x) \stackrel{\text{def}}{=} \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}\{X_{ij} \leq x\}, \qquad x \in \mathbb{R},\ j \in \{1, \ldots, d\}. \tag{5.18}$$

Division by $n+1$ instead of n is there to prevent division by zero in the standardization. The angles associated to "large" observations are supposed to provide

an approximate sample from Q in view of Assumption 5.2.1. The output of Algorithm 1 is a trained generator G_θ which is able to produce new coordinates in the Aitchison orthonormal basis, defined by the orthonormal basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\}$ of $\mathbb{H}$, that are hopefully close in distribution to the coordinates of angles associated to the large observations in the training set.

The WA-GAN procedure in Algorithm 1 only uses the observations for which $R = \|V\|_1$ is larger than $t = n/k_1$ so that, by (5.4), the number of observations used by the algorithm is asymptotically binomial with parameters n and $(dk_1)/n$. Here k_1 is a value to be chosen by the user, with larger k_1 meaning that more observations are used and hence the variance is smaller, and with smaller k_1 (hopefully) making the approximation in Assumption 5.2.1 more accurate. Finite sample bounds on the quality of this approximation are given in [25]. The relatively low effective sample size makes it important that the architectures of the generator and the discriminator are simple enough so that their parameters can be learned from a moderate number of examples. Those considerations are further explored in the numerical sections.

Algorithm 2 uses for each margin the k_2 largest observations, so that between k_2 and dk_2 of the multivariate observations are used. According to Remark 5.3.2, the value of k_2 should not be larger than k_1. A natural convenient choice is to take $k_1 = k_2$.

Algorithm 1 is essentially the standard WGAN algorithm with gradient penalty of [63] applied to the coordinates of the large angles in an orthonormal basis of the Aitchison simplex. The ADAM optimizer which is used to update the weights of the generator and the discriminator at each step is a popular gradient descent algorithm with adaptive learning rate introduced in [78]. The function `Adam()` in the algorithm is one such gradient step and is the default optimizer proposed for the WGAN architecture with gradient penalty. Compared to the standard algorithm, one additional regularization term

$$\rho \left\| \frac{1}{m} \sum_{i=1}^{m} \operatorname{softmax}\left(V G_\theta(z_i)\right) - \frac{1_d}{d} \right\|_2 \tag{5.19}$$

is added to enforce the marginal constraint (5.3) in the generator in a soft way and thus to enforce sampling from a genuine angular measure.

Algorithm 1 focuses solely on the extremal dependence of X and is independent of any marginal assumptions besides continuity of the marginal distribution functions. Therefore, if one is interested in computing a quantity which only depends on Q in Assumption 5.2.1, such as the coefficients θ_J introduced in the beginning of Section 5.4, this can be done independently of Algorithm 2. This is not the case for methods such as [85] which directly evaluate sample quality on the scale of the MGP distribution. Note also that even though Algorithm 1 only considers the angular measure Q with respect

Algorithm 1 WA-GAN for tail dependence estimation

Require: observations $X_1, \ldots, X_n$, orthonormal basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\}$ of $\mathbb{H}$, and $k_1 \in \{1, \ldots, n\}$

Require: gradient penalty coefficient $\lambda > 0$, marginal penalty coefficient $\rho > 0$, the number of discriminator iterations per generator iteration n_D, the batch size m, Adam hyperparameters $\alpha > 0$ (learning rate), $\beta_1, \beta_2 \in [0, 1)$, the latent space dimension $\ell \in \mathbb{N}$, the number of epochs n_E

Require: discriminator initial parameters $w_0 \in \mathcal{S}_D$, generator initial parameters $\theta_0 \in \mathcal{S}_G$

$t \leftarrow n/k_1$

for $i = 1, \ldots, n$ **do**

$\widehat{V}_i \leftarrow \left(1/(1 - \widehat{F}_j(X_{ij}))\right)_{j=1}^{d}$

$R_i \leftarrow \|\widehat{V}_i\|_1$ and $W_i \leftarrow \widehat{V}_i/R_i$

if $R_i \geq t$ **then**

$W_i^* \leftarrow \langle \mathrm{clr}(W_i), \boldsymbol{e}_i \rangle$

end if

end for

$K \leftarrow \sum_{i=1}^{n} \mathbb{I}\{R_i \geq t\}$

$V \leftarrow (\boldsymbol{e}_1 \mid \ldots \mid \boldsymbol{e}_{d-1}) \in \mathbb{R}^{d \times (d-1)}$

$\theta \leftarrow \theta_0$ and $w \leftarrow w_0$

for epoch $e = 1, \ldots, n_E$ **do**

for $t = 1, \ldots, n_D$ **do**

for $i = 1, \ldots, m$ **do**

sample w_i^* from $\{W_1^*, \ldots, W_K^*\}$, sample z_i from a standard multivariate normal distribution on $\mathbb{R}^\ell$, sample u_i from an uniform distribution on $[0, 1]$

$\widetilde{w}_i^* \leftarrow G_\theta(z_i)$

$\widehat{w}_i^* \leftarrow u_i\, w_i^* + (1 - u_i)\, \widetilde{w}_i^*$

end for

$L_D \leftarrow \frac{1}{m} \sum_{i=1}^{m} \left\{ D_w(\widetilde{w}_i^*) - D_w(w_i^*) + \lambda \left(\|\nabla_{w^*} D_w(\widehat{w}_i^*)\|_2 - 1 \right)^2 \right\}$

$w \leftarrow \mathrm{Adam}(L_D, w, \alpha, \beta_1, \beta_2)$

end for

for $i = 1, \ldots, m$ **do**

sample z_i from a standard multivariate normal distribution on $\mathbb{R}^\ell$

end for

$L_G \leftarrow -\frac{1}{m} \sum_{i=1}^{m} D_w\left(G_\theta(z_i)\right) + \rho \left\| \frac{1}{m} \sum_{i=1}^{m} \mathrm{softmax}\left(V G_\theta(z_i)\right) - \frac{1_d}{d} \right\|_2$

$\theta \leftarrow \mathrm{Adam}(L_G, \theta, \alpha, \beta_1, \beta_2)$

end for

Output: Trained generator G_θ

to the L_1-norm, a simple post-processing step allows for any norm on $\mathbb{R}^d$, as explained in Appendix 5.C.

Remark 5.3.1 (Standardization of the margins to unit-Pareto). The procedure in Algorithm 1 is based on data standardized to unit-Pareto margins, in accordance with Assumption 5.2.1. As in (5.18), we work with an empirical standardization $\widehat{F}_j$ for each margin $j = 1, \ldots, d$, resulting in a procedure using only the marginal ranks of the original data and solely focusing on dependence, not requiring any assumption on the tail of F_j as in Assumption 5.2.3. The same approach was followed in [42, 47] and, more recently, in [25], for example. An alternative would be to use the empirical distribution function $\widehat{F}_j$ only up to a threshold $u_j < u_{*j}$ and fit a univariate GP distribution above this threshold, relying on Assumption 5.2.3. Doing so could improve the quality of the transformation to unit-Pareto margins, but at the price of making Algorithm 1 also require marginal assumptions. Furthermore, this route has already been explored in [43] and the same authors advocate in [42] for the pure nonparametric approach. Therefore, we stick to the rank-based procedure.

Some care needs is needed when sampling from $X \mid X \not\leq u(t)$ for $t > 1$ large, where $u(t)$ is the vector of thresholds such that for each component separately, the excess probability is $1/t$. For example, a common situation is the case where X represents some positive multivariate risks, that is, X takes values in $[0, \infty)^d$. If we directly make use of the formula $u(t) + H$, with estimated quantities, as suggested by formula (5.10), we risk generating "extreme" points for which some coordinates are negative. Our approach avoids this issue.

Algorithm 1 permits to sample from the multivariate Pareto distribution Y associated to X using (5.8). Let Y^* denote such a generated quantity. From (5.26), we have

$$Y^* \overset{\mathcal{L}}{\approx} \left(\frac{k}{n} V \mid V \not\leq \frac{n}{k} \right),$$

where $\overset{\mathcal{L}}{\approx}$ means an approximation in distribution. Equivalently,

$$V^* \overset{\text{def}}{=} \frac{n}{k} Y^* \overset{\mathcal{L}}{\approx} \left(V \mid V \not\leq \frac{n}{k} \right).$$

Now, to pass from V to X one needs to apply the transformation b described in Proposition 5.2.1, which is based on the marginal quantile functions. Let $\widehat{b}$ denote an estimator of this quantity. The final output of our next algorithm will be

$$X^* = \widehat{b}(V^*) = \widehat{b}\left(\frac{n}{k} Y^* \right) \overset{\mathcal{L}}{\approx} \left(X \mid V \not\leq \frac{n}{k} \right) = \left(X \mid X \not\leq b\left(\frac{n}{k} \right) \right),$$

by the continuity of margins. The estimator $\widehat{b}_j((n/k)y)$ for $y > 0$ depends on whether $y \leq 1$ or $y > 1$. For $y \leq 1$ we simply use the estimator based on

the empirical cumulative distribution function $\widehat{F}_j$. For $y > 1$, the estimator is obtained through Assumption 5.2.3, more precisely its equivalent formulation (ii) in Remark 5.2.1. This leads to

$$\widehat{b}_j((n/k)y) \overset{\text{def}}{=} \begin{cases} X_{(\lceil n-k/y \rceil \vee 1):n,j}, & 0 < y \leq 1, \\ \widehat{b}_j(n/k) + \widehat{a}_j(n/k) \cdot \frac{y^{\widehat{\xi}_j} - 1}{\widehat{\xi}_j}, & y > 1, \end{cases} \tag{5.20}$$

where $X_{1:n,j} < \ldots < X_{n:n,j}$ are the ascending order statistics in the j-th sample, and where $\widehat{a}_j(n/k) \overset{\text{def}}{=} \widehat{\sigma}_j$ and $\widehat{\xi}_j$ are obtained by maximum likelihood for the GP distribution.

Remark 5.3.2 (Relation between thresholds in both algorithms). When combining Algorithm 2 with the output of Algorithm 1, one actually uses two different thresholds: on the one hand, the threshold $t_1 = n/k_1 > 0$ which has been used to fit the angular measure in Algorithm 1, and, on the other hand, the threshold vector $u(t_2) = b(n/k_2)$ for which we aim to sample from $X \mid X \not\leq u(t_2)$. As Algorithm 2 relies on the output of Algorithm 1, one should ensure that $u(t_2)$ is large enough so that the approximation of the angular measure Q in Algorithm 1 is valid. In terms of the random variable R in (5.1), it means that we must ensure $R = \|V\|_1 > n/k_1$ given that $X \not\leq u(t_2)$, which is equivalent to $V \not\leq n/k_2$, i.e., $\|V\|_\infty > n/k_2$. Since we have $\|V\|_1 \geq \|V\|_\infty$, it suffices to take $k_2 \leq k_1$. This is illustrated in Figure 5.4 for $d = 2$.

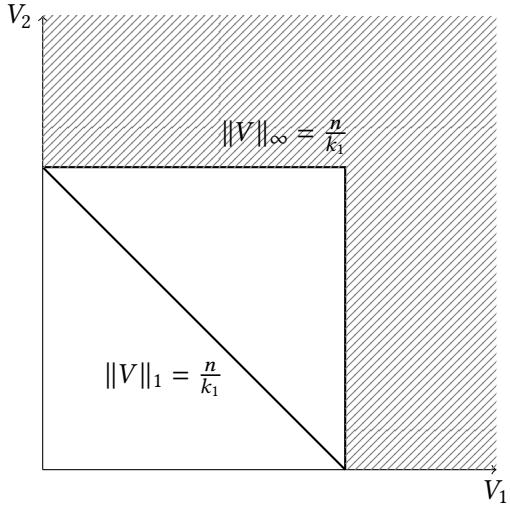


Figure 5.4: Relation between thresholds in the V space. The gray zone is where we can safely sample using Algorithm 2.

Algorithm 2 Sampling from the tail of X

Require: observations $X_1, \dots, X_n$, orthonormal basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_{d-1}\}$ of $\mathbb{H}$, trained generator G_θ, the size of the latent space on which G_θ has been trained ℓ, and $k_2 \in \{1, \dots, n\}$
Require: desired number of samples n^*
  for $j = 1, \dots, d$ **do**
    $u_j \leftarrow X_{n-k_2:n,j}$
    Compute the maximum likelihood estimator $(\widehat{\sigma}_j, \widehat{\xi}_j)$ of an univariate GPD
  end for
  $V \leftarrow (\boldsymbol{e}_1 \mid \dots \mid \boldsymbol{e}_{d-1}) \in \mathbb{R}^{d\times(d-1)}$
  $i \leftarrow 1$ and $\mathcal{G}$ an empty list of size n^*
  while $i \leq n^*$ **do**
    sample z from a standard multivariate normal distribution on $\mathbb{R}^\ell$
    $W \leftarrow \mathrm{clr}^{-1}(VG_\theta(z))$
    sample R from a unit-Pareto distribution
    $Y \leftarrow RW$
    if $\max(Y) > 1$ **then**
      for $j = 1, \dots, d$ **do**
        if $Y_j \leq 1$ **then**
          $(\mathcal{G}[i])_j \leftarrow X_{(\lceil n-k_2/Y_j \rceil \vee 1):n,j}$
        end if
        if $Y_j > 1$ **then**
          $(\mathcal{G}[i])_j \leftarrow u_j + \widehat{\sigma}_j(Y_j^{\widehat{\xi}_j} - 1)/\widehat{\xi}_j$
        end if
      end for
      $i \leftarrow i + 1$
    end if
  end while
  Output: List of n^* random variables $\mathcal{G}$ approximately distributed as $X \mid X \not\leq u$

5.4 Numerical experiments

We use both simulated and real data sets to explore the performance of our proposed method and to compare it to other existing methods. We start with providing the details of how the different methods are compared and how the the computations are made. The next section uses simulated data to compare WA-GAN with HTGAN and GPGAN. In the final section these methods are applied to a financial data set.

5.4.1 Performance and computation

Performance measures. To assess the quality of the generated sample $X^{\mathcal{G}}$, we will typically compare it with a sample from the data $X^{\mathcal{T}}$ (test set) which has not been used for training. We propose to use measures that are inspired by multivariate extreme value theory since our focus is on the tail of X.

A first set of measures that we propose is based on the *extremal coefficients* [13, Section 8.2.7]. For any non-empty subset $J \subseteq \{1, \dots, d\}$ with at least two elements $|J| \geq 2$, we define its extremal coefficient θ_J as

$$\theta_J \stackrel{\text{def}}{=} d \cdot \int_{\Delta_{d-1}} \bigvee_{j \in J} w_j \, \mathrm{d}Q(w) \in [1, |J|]. \tag{5.21}$$

The closer θ_J is to 1, the stronger the dependence in the tail of $(X_j)_{j \in J}$, while θ_J close to $|J|$ corresponds to weaker tail dependence. For a given coefficient order $k = |J|$, to compute a summary score of their difference in the test set $X^{\mathcal{T}}$ and the generated set $X^{\mathcal{G}}$, we compute their relative absolute difference as follows:

$$E_{\overline{\theta}}(k) \stackrel{\text{def}}{=} \frac{1}{\binom{d}{k}} \sum_{\substack{J \subseteq \{1,\dots,d\} \\ |J|=k}} \left| 1 - \frac{\theta_J^{\mathcal{G}}}{\theta_J^{\mathcal{T}}} \right|. \tag{5.22}$$

The coefficients are estimated for the test set using the empirical angular measure in (5.21) and for the generative methods using the sampled realizations of Q. For the test set, the radial threshold n/k is the same as for the training set. Typically, we will consider the bivariate coefficients, with $k = 2$, as in [19], and the trivariate coefficients, with $k = 3$. This measure focuses on the extremal dependence structure only and aims to measure the quality of Algorithm 1. As it is fast to compute, it will be used as the validating measure for model selection.

A second measure we consider to assess the quality of the generated angles by Algorithm 1 is based on a simple extension of the EMD in Section 5.2.3 which is known as the 2-Wasserstein distance [126, Chapter 7]. If $n_{\mathcal{G}}$ and $n_{\mathcal{T}}$ denote, respectively, the size of the test set and the generated set of angles on

the simplex, the measure can be expressed as

$$W_2(\Theta^{\mathcal{G}}, \Theta^{\mathcal{T}}) \overset{\text{def}}{=} \inf_{\pi \in \Pi\left(\frac{1_{n_{\mathcal{G}}}}{n_{\mathcal{G}}}, \frac{1_{n_{\mathcal{T}}}}{n_{\mathcal{T}}}\right)} \sqrt{\sum_{i=1}^{n_{\mathcal{G}}} \sum_{j=1}^{n_{\mathcal{T}}} \|\Theta_i^{\mathcal{G}} - \Theta_j^{\mathcal{T}}\|_2^2 \, \pi_{ij}} \tag{5.23}$$

where, for probability vectors $a \in \mathbb{R}^k$ and $b \in \mathbb{R}^\ell$, we use the notation

$$\Pi(a, b) \overset{\text{def}}{=} \left\{ \pi \in \mathbb{R}_+^{k \times \ell} : \pi 1_\ell = a \text{ and } \pi^T 1_k = b \right\}.$$

We can view π_{ij} as the amount of mass transferred from $\Theta_i^{\mathcal{G}}$ to $\Theta_j^{\mathcal{T}}$. In a similar way as in definition (5.13) of the EMD, $W_2(\Theta^{\mathcal{G}}, \Theta^{\mathcal{T}})$ measures, among all the possible ways to transport the empirical distribution of the vectors in $\Theta^{\mathcal{G}}$ to the one of the vectors in $\Theta^{\mathcal{T}}$, the cheapest way to do so if the cost of unitary transportation is equal to the squared Euclidean distance between points. In the case where we would have normally distributed quantities, this quantity would reduce to computing the well-known *Fréchet inception distance*, which is popularly used to assess the quality of generated images [68] using generative systems like the GAN, also used in [85] to assess the quality of the MGP output of their algorithm. As we compare quantities that are far from being normally distributed, we think it makes more sense to compute the 2-Wasserstein distance instead.

The last measure that we propose aims at evaluating the output of Algorithm 2. Let $X^{\mathcal{T}}$ denote a set of observations distributed like X, which typically have not been used for training, and let $X_u^{\mathcal{G}}$ denote a set of observations generated by Algorithm 2 based on the threshold vector $u = b(n/k_n) \in \mathbb{R}^d$. Then if the MGP approximation is accurate enough, we expect $X_u^{\mathcal{T}} \overset{\text{def}}{=} X^{\mathcal{T}} \,|\, X^{\mathcal{T}} \not\leq u$ to be close in distribution to $X_u^{\mathcal{G}}$. Hence, similarly as for the angles, we propose to compute their difference in distribution using

$$W_2(X_u^{\mathcal{G}}, X_u^{\mathcal{T}}) = \inf_{\pi \in \Pi\left(\frac{1_{n_{\mathcal{G}}}}{n_{\mathcal{G}}}, \frac{1_{n_{\mathcal{T}}}}{n_{\mathcal{T}}}\right)} \sqrt{\sum_{i=1}^{n_{\mathcal{G}}} \sum_{j=1}^{n_{\mathcal{T}}} \|(X_u^{\mathcal{G}})_i - (X_u^{\mathcal{T}})_j\|_2^2 \, \pi_{ij}}. \tag{5.24}$$

Architectures. For both the generator and the discriminator in Algorithm 1, we consider fully connected multilayer perceptrons with leaky rectified linear unit activation function [89] which applies the map

$$x \in \mathbb{R} \mapsto \begin{cases} x & \text{if } x \geq 0, \\ \alpha x & \text{else,} \end{cases} \tag{5.25}$$

to each component of the linear transformation in a given layer, where $\alpha \in [0, 1]$ is typically set to 0.01. Taking $\alpha = 0$ leads to the rectified linear unit activation function and $\alpha = 1$ leads to a simple linear layer. We refer to [89] for a comparison with other standard activation functions. Here, we consider $\alpha = 0.01$ for any hidden layer while we take $\alpha = 1$ for the final layer in each network. More recently, α has been considered as a learnable parameter in the optimization process [67] but this option is not considered in this work. Other architectures are possible if the data have some structure. For example, the authors of [19] consider convolutions in their network as they work with spatial data on a grid, and this could be adapted to our framework too, but here we only consider basic multivariate data and keep this extension for future work.

Hyperparameter search. In addition to the learnable parameters of the generator and discriminator networks in Algorithm 1, many external parameters (hyperparameters) can be chosen by the user. We consider the following values for the hyperparameters:

- batch size $m \in \{\lfloor n/k \rfloor : k = 1, 2, 4, 8, 16\}$;
- dimension of the generator's latent space $\ell \in \{\lfloor (d-1)/k \rfloor : k = 0.25, 0.5, 0.75, 1, 2\}$;
- maximum number of neurons in a hidden layer in $\{32, 64, 128, 256, 512\}$;
- number of hidden layers in $\{1, 2, 4, 8\}$;
- learning rate for the ADAM optimizer in $\{0.01, 0.005, 0.001, 0.0005, 0.0001\}$ and coefficients $\beta = (\beta_1, \beta_2)$ used for computing running averages of gradient and its square in $\{(0.0, 0.9), (0.5, 0.9), (0.5, 0.99)\}$;
- the gradient penalty coefficient $\lambda \in \{0.1, 1, 3, 5, 7, 9\}$;
- the marginal penalty coefficient $\rho \in \{0.001, 0.01, 0.1, 1, 3\}$;
- the number of discriminator iterations per generator iterations $n_D \in \{1, 3, 5, 10\}$.

Inside the space of possible values for those hyperparameters, we perform a random search. Typically, we consider around 2000 different models in this space for a given dataset. We train them with $n_e = 5000$ epochs, since training them even longer did not significantly improve the performance, but greatly affects the computation time. Evaluation of the associated models is based on the extremal coefficient metric $E_{\overline{\theta}}(k)$ in (5.22) with $k = 2$ and $k = 3$. Validation and final performance assessment of the models are always performed on separate datasets.

Comparison with existing methods. As reviewed in Section 5.2.1, numerous methods already exist in the rapidly growing field that combines generative approaches with extreme value theory, each with its own advantages and limitations. Therefore, an exhaustive comparison with all existing methods is not feasible, even less so as the software code underlying the numerical results is often not available from the papers. Instead, we choose to compare our approach with versions of two recent methods: HTGAN from [57] and GPGAN from [85]. We discuss our specific implementations in Appendix 5.D.

Software. The code used to produce the results of this section was written in `Python`—in particular, the `PyTorch` [102] framework was used to train the generative models—and in `R` [104], which was mainly used for data preprocessing and computation of the performance measures associated with the generated data, notably via the `transport` [114] package for the computation of the Wasserstein distance W_2 in (5.24).

Hardware. The training of the various algorithms and the computation of performance metrics have been performed on the CPUs of the Lemaitre4 and the NIC5 computer clusters, which are managed by the Consortium of Équipements de Calcul Intensif (CÉCI). Details and more specifications related to these computers can be found on the CECI website `https://www.ceci-hpc.be/`.

5.4.2 Simulated data: logistic dependence structure

As an illustration of the method's performance in a controlled setting, we start by considering simulated data. To keep things simple, we consider the same experimental setup as in Section 3.2 of [57], with whom we are going to compare our method.

The random vector X is generated using a d-dimensional Gumbel copula with parameter $\theta \in [1, \infty)$ and Pareto distributed marginals with parameter $\alpha > 0$. It is easy to verify that this data-generating process satisfies Assumptions 5.2.1 to 5.2.3 in Section 5.2.2. In particular, the extremal coefficient introduced as a performance measure at the beginning of Section 5.4 is $\theta_J = |J|^{1/\theta}$. Hence, $\theta \to 1$ corresponds to less dependence in the tail X, while $\theta \to \infty$ corresponds to perfect tail dependence.

We consider various scenarios in which we assess the performance of WAGAN and compare it to HTGAN and GPGAN. For HTGAN we follow [57] and make the unrealistic assumption that α is known. We take $d \in \{10, 20, 50\}$ and $\theta \in \{4/3, 2, 4\}$ leading to a Kendall's tau of $\tau \in \{1/4, 1/2, 3/4\}$. For the margins, we consider $\alpha = 2$ in any scenario. The models are trained on $n_{\text{train}} = 10\,000$

observations and validated on $n_{\text{val}} = 5\,000$ observations. The performance assessment is done on $n_{\text{test}} = 20\,000$ points.

Results. For each method and each scenario, three scores are computed. On the one hand, two dependence scores: the score $\{E_{\overline{\theta}}(2) + E_{\overline{\theta}}(3)\}/2$, with $E_{\overline{\theta}}(k)$ as in (5.22), which is also the one used to validate the hyperparameters, and $W_2(\Theta^{\mathcal{G}}, \Theta^{\mathcal{T}})$ as in (5.23), the 2-Wasserstein distance between the generated and test angles. On the other hand, to assess the quality of generated extremes $X \mid X \not\leq u$, the 2-Wasserstein distance $W_2(X_u^{\mathcal{G}}, X_u^{\mathcal{T}})$ as defined in (5.24) is computed, where $u_j = \widehat{b}_j(n/k_2)$ with $\widehat{b}_j$ as in (5.20), for some $k_2 \in \{1, \dots, n\}$ which satisfies $k_2 \leq k_1$ where $k_1 \in \{1, \dots, n\}$ is used to train the generator in Algorithm 1. Here we take $n = n_{\text{train}}$ and $k_1 = k_2 = [\sqrt{n}]$.

The results are reported in Table 5.1 and 5.2 below.

		$\{E_{\overline{\theta}}(2) + E_{\overline{\theta}}(3)\}/2$			$W_2(\Theta^{\mathcal{G}}, \Theta^{\mathcal{T}})$		
τ	Model	$d = 10$	$d = 20$	$d = 50$	$d = 10$	$d = 20$	$d = 50$
$\tau = \frac{1}{4}$	WA-GAN	**0.0103**	**0.0074**	**0.0054**	**0.0518**	**0.0539**	**0.0429**
	HTGAN	0.0146	0.0185	0.0199	0.0625	0.0667	0.0521
	GPGAN	0.0262	0.0321	0.0326	0.0836	0.0819	0.0526
$\tau = \frac{1}{2}$	WA-GAN	**0.0176**	0.0207	**0.0097**	**0.0918**	**0.1032**	**0.0959**
	HTGAN	0.0247	**0.0179**	0.0126	0.1299	0.1255	0.1029
	GPGAN	0.0511	0.0671	0.0672	0.1956	0.2035	0.1582
$\tau = \frac{3}{4}$	WA-GAN	**0.0208**	**0.0266**	**0.0158**	**0.1063**	**0.1317**	**0.1483**
	HTGAN	0.0284	0.0338	0.055	0.1402	0.1794	0.1554
	GPGAN	0.0497	0.0726	0.0818	0.2491	0.2944	0.2443

Table 5.1: Tail dependence scores for different dependence τ and dimensions d. For each scenario and each score, the best score (i.e., the lowest) is indicated in bold.

These results show that WA-GAN is competitive with other methods, from the tail dependence perspective as well as from the quality of the generated extremes. The advantage of WA-GAN seems to increase as the dimension d increases. Large values of d is often at the center of interest when generative models for the dependence are used. Note also that it may be surprising at first glance that the scores become better when d increases, but this is probably due to the dependence structure of the data generating process being the same for all pairs of variables, making it easier to learn it when the dimension of the space increases.

Some additional comments can be made for each method. For WA-GAN, the penalization coefficient ρ in (5.19) is typically selected to be large among the possible values ($\rho = 1$ or $\rho = 3$) showing that it is a good idea to enforce

τ	Model	$W_2(X_u^{\mathcal{G}}, X_u^{\mathcal{T}})$ $d = 10$	$d = 20$	$d = 50$
$\tau = \frac{1}{4}$	WA-GAN	67.311	**62.905**	**42.825**
	HTGAN	1479.9	3106.7	1170.7
	GPGAN	**66.663**	65.429	52.133
$\tau = \frac{1}{2}$	WA-GAN	**66.734**	**52.832**	**32.468**
	HTGAN	2725.3	1447.9	2367.4
	GPGAN	67.841	55.308	46.996
$\tau = \frac{3}{4}$	WA-GAN	**59.548**	**44.602**	**31.808**
	HTGAN	891.62	612.44	440.45
	GPGAN	59.919	53.758	45.541

Table 5.2: Extremes scores for different dependence τ and dimensions d. For each scenario and each score, the best score (i.e., the lowest) is indicated in bold.

marginal constraints (5.3) in the generative model. For HTGAN, the obtained values of the hyperparameters match the discussion on pages 13–15 in the paper [57] proposing this method. In particular, we get big batch sizes, big latent space dimensions, and low values of the learning rate included in the range recommendend by the authors. For GPGAN, the training phase is quicker since, as illustrated on Figure 5.4, we have a lower training size in this case as this method is trained on points with $\|V\|_\infty > n/k$ while WA-GAN is trained on points with $\|V\|_1 > n/k$, which is less restrictive. This can be an asset if we have enough data, but can harm GPGAN's performance if d is large and n is not too big.

As an illustration of the performances of the different methods to accurately reproduce the tail dependence of the data, we plot the empirical bivariate and trivariate extremal coefficients ($|J| = 2$ and $|J| = 3$ in (5.21)) of the test set against the ones of the training set, and against those of all the generative methods for $d \in \{10, 50\}$ and $\tau = 0.5$. In this setting, we know that the true values of these coefficients are $\sqrt{|J|}$. This is displayed n Figure 5.5 for $d \in \{10, 50\}$. There is no clear visual difference between WA-GAN and HTGAN, both performing well at representing the coefficients of the test set. GPGAN, however, fails to do so, and even more when the dimension increases. When d gets larger, the empirical extremal coefficients are below their true values, even in the test set. The reason is that the empirical estimator of Q in (5.21) becomes less accurate as d increases [25].

Next, to illustrate the accuracy of generated extremes by each method, we plot the projection of each generated sample on the first two margins and compare it to the first two margins of the extremes in the test set. We

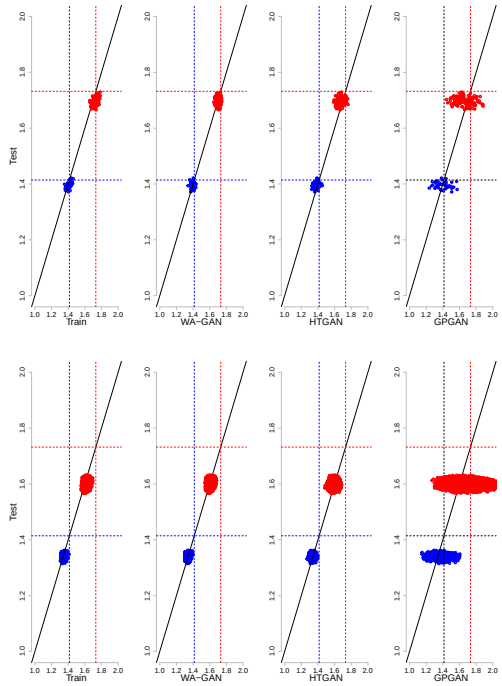


Figure 5.5: Top: $d = 10$; bottom: $d = 50$. – Extremal coefficients of order $k = 2$ (blue) and $k = 3$ (red). Solid black line is the diagonal. Dashed blue and red lines are the true values of the coefficients for $k = 2$ and $k = 3$ respectively.

take again $\tau = 0.5$. Results for are displayed on Figure 5.6. WA-GAN seems to perform well, whatever the dimension. The results of HTGAN are less satisfying as it is clear from the figure that the angular measure associated with this method is discrete (this corresponds to the "directions" that we observe in the generated extremes) and this does not match the true angular measure which is logistic. This is theoretically justified by Proposition 2.10 in [57] introducing this method, and already visible from their Figure 2 on page 14. GPGAN also performs well even though its performance decreases in comparison to WA-GAN as the dimension increases; see Table 5.2.

5.4.3 Financial data: daily returns of industry portfolios

We apply our methodology to analyze the tail behavior of the "value-averaged" daily returns of $d = 30$ industry portfolios compiled and posted as part of the Kenneth French Data Library. The data in consideration span between 1950 and 2015 with $n = 16\,694$ observations. This data set is very widely used; for analyses in an extreme value context see e.g. [77] and [29]. Details about

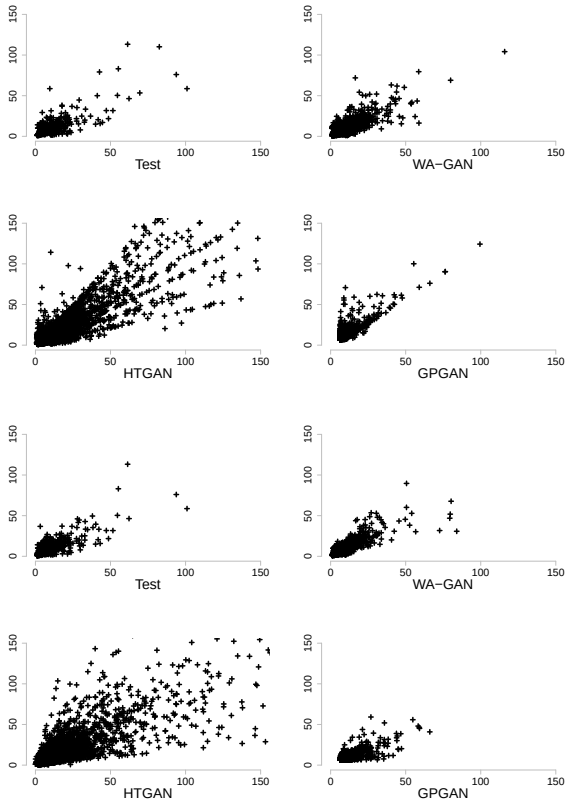


Figure 5.6: Top: $d = 10$**; bottom:** $d = 50$**. – First two margins of extremes in the test set and the different generative methods.**

the portfolio constructions can be found online.[1] Since we are interested in extreme losses we first multiply all returns by -1.

As illustrated by the Kendall's correlation matrix in Figure 5.7, the daily losses are positively related between the portfolios. We aim to investigate this dependence in the tail part of the distribution.

To sample extremes from the joint distribution, we first compute the maximum likelihood estimators of the univariate GPD parameters for each margin. Boxplots illustrating the values of those parameters are reported in Figure 5.8. The estimated shape values clearly indicates the presence of heavy tails in the marginal distributions. The GPD qq-plots of the marginal fits are to be found on Figures 5.11, 5.12 and 5.13 in Appendix 5.E. They indicate a relatively good fit, even though some extreme quantiles are larger than the

[1] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_30_ind_port.html`

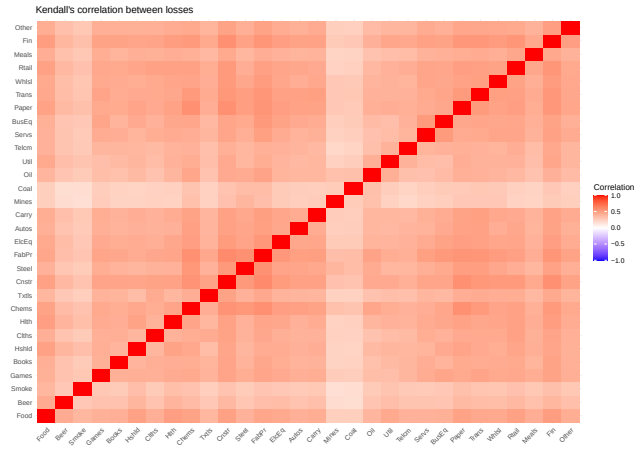


Figure 5.7: Kendall's correlation matrix between daily losses of the portfolios.

ones associated with the GPD model.

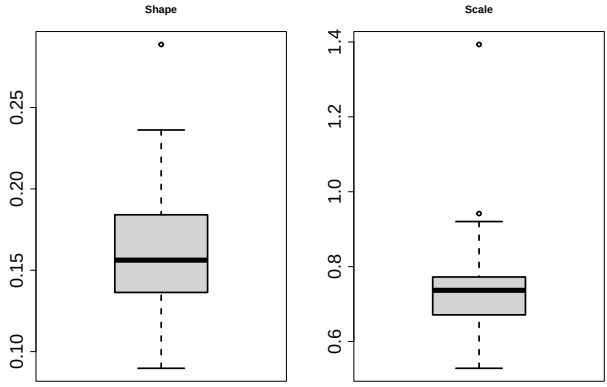


Figure 5.8: Boxplots of estimated marginal GPD parameters

For comparison, we start by computing the dependence score $\{E_{\overline{\theta}}(2) + E_{\overline{\theta}}(3)\}/2$ and the extreme score $W_2(X_u^{\mathcal{G}}, X_u^{\mathcal{T}})$ for WA-GAN, HTGAN and GP-GAN. The training and validation procedures are exactly the same as described in Section 5.4.2. Here, we train on $n_{\text{train}} = 7\,000$, we validate on $n_{\text{val}} = 3\,000$ and compute the scores on $n_{\text{test}} = 6\,694$ observations. We again consider $k = \sqrt{n_{\text{train}}}$ for WA-GAN. For HTGAN, we do not compute the scores on extremes because the marginal distributions are not known, see Appendix 5.D. Table 5.3 shows that WA-GAN seems to be the best at capturing the dependence between extremes in the data while GPGAN has a better 2-Wasserstein distance score.

Method	$\{E_{\overline{\theta}}(2) + E_{\overline{\theta}}(3)\}/2$	$W_2(X_u^{\mathcal{G}}, X_u^{\mathcal{T}})$
WA-GAN	**0.018**	11.93
HTGAN	0.026	/
GPGAN	0.043	**8.46**

Table 5.3: Test scores of each method on the $d = 30$ financial dataset. The best score (i.e., the lowest) is indicated in bold.

As already pointed out in the analysis of [77], the extreme losses of those portfolios can be clustered into different kind of industries. For example, the tobacco industry exhibits asymptotic independence to all other categories (e.g., to the textile industries) while the extreme losses of the business and IT-related industries are dependent. We illustrate the fact that WA-GAN also captures those phenomena in Figure 5.9 by displaying the two-dimensional projections of the estimated angular measure on those margins. We see that the generated "angles" associated with tobacco/textile portfolios are much more concentrated around the axes than for business/IT, showing that WA-GAN captures this complex tail dependence.

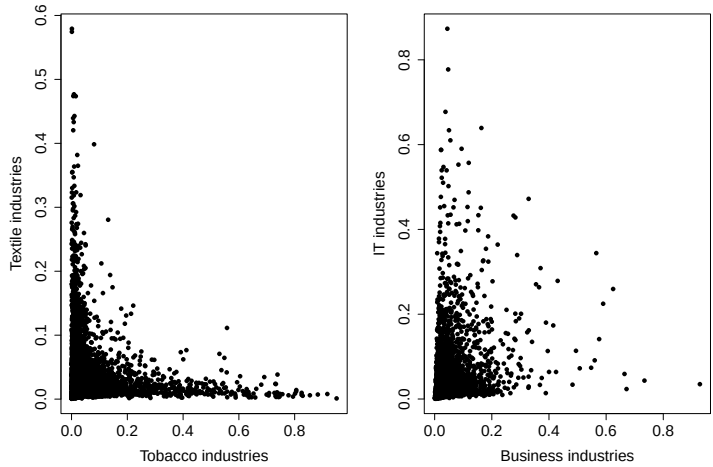


Figure 5.9: Two-dimensional projections of the generated "angles" from Q by WA-GAN.

As a further illustration we consider a portfolio obtained by combining the mines industry portfolio with the coal industry portfolio (with equal weight). It is intuitive and has been shown in [29, 77] that the mines and coal industries exhibit asymptotic dependence. If one wants to sample the losses associated to the portfolio we propose, this tail dependence should be taken into account.

Figure 5.10 represents the densities of simulated values of the extreme losses of the combined portfolio (i.e., both marginal losses exceeds a large threshold) based on WA-GAN (in black) and based on independent simulations from the marginal GPDs (in blue). Not taking the tail dependence into account leads to underestimation of the risk associated with this portfolio.

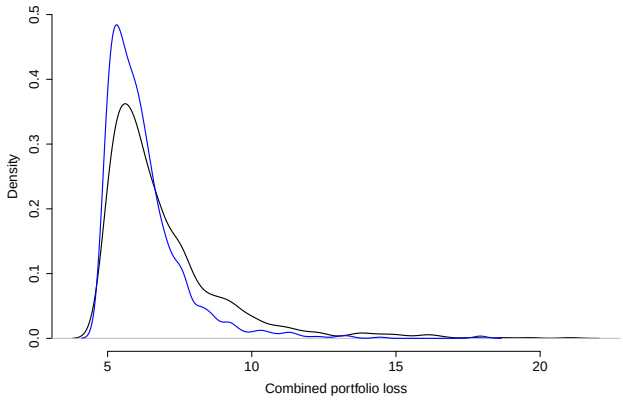


Figure 5.10: Densities of extreme losses in the combined portfolio of mines and coals industries based on WA-GAN (black) and independently simulated margins (blue).

5.5 Conclusions

This paper develops the WA-GAN method which is built on the L_1-norm and on the asymptotic independence of the angular and radial parts of a d-dimensional regularly varying random vector. An important component is the use of Aitchison coordinates which transforms values on the unit simplex in $\mathbb{R}^d$ to the entire linear space $\mathbb{R}^{d-1}$. The method provides simulated extreme values and can hence be used to estimate probabilities of extreme events, even more extreme than those already observed.

The method is tried out on simulated extreme value data sets of dimensions $d = 10$, 20 and 50 with 10 000 observations in the training set and 5 000 observations in the validation set. Plots and two performance measures which (i) score the dependence between extremes using extremal coefficients and (ii) score the quality of the simulated values on the real scale using the 2-Wasserstein distance show that WA-GAN performs well. Further, in Table **??** WA-GAN is shown to have better values of these performance scores than the HTGAN and GPGAN methods.

The methods are also applied to a financial data set from the Kenneth

French Data Library. Again WA-GAN performs well and underlines the importance of being able to handle dependence between different financial portfolios. In terms of performance measures, WA-GAN has the best dependence score, while GPGAN has a better 2-Wasserstein score than WA-GAN.

In this paper we do not quantify the uncertainty in the estimated probability of an extreme event. However, provided enough computational power is available, this could be done using a approach similar to a parametric bootstrap. For this one would repeat the following, say, 100 times: first use the fitted WA-GAN to simulate a new extreme data set of the same size as the original one, then fit a new WA-GAN to this simulated data set and use the new WA-GAN to estimate the probability of interest. This will provide 100 estimated probability values which can be used to find "confidence bounds" for the original estimated probability.

Code and data availability

The code is available from the first author on the Github repository `https://github.com/stephanelh98/extreme-WGAN`. The data underlying the analysis in Section 5.4.3 is publicly available from the Kenneth R. French data library.

Acknowledgments

The research of Stéphane Lhaut was supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium) within the framework of a FRIA grant (No. 1.E.114.23F). Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

5.A Proof of Proposition 5.2.1

Proof. We start by noting that $X \overset{\mathcal{L}}{=} b(V)$ and, by the continuity of the margins F_j, each tail quantile function b_j is strictly increasing, so that $\{\boldsymbol{X} \not\leq b(t)\} = \{V \not\leq t\}$ for any $t > 1$. Consequently,

$$\frac{X - b(t)}{a(t)} \,\Big|\, X \not\leq b(t) \overset{\mathcal{L}}{=} \frac{b(t(t^{-1}V)) - b(t)}{a(t)} \,\Big|\, V \not\leq t.$$

Furthermore, it follows easily from Assumption 5.2.1 that

$$\left(t^{-1}V \mid V \not\leq t\right) \xrightarrow{\mathcal{L}} Y, \qquad t \to \infty, \tag{5.26}$$

where, for Borel set $B \subseteq \mathbb{E}$, writing $\mathbb{L}_1 \stackrel{\text{def}}{=} \{x \in [0,\infty)^d : x \not\leq 1\}$, we have

$$\mathsf{P}(Y \in B) = \frac{\nu(B \cap \mathbb{L}_1)}{\nu(\mathbb{L}_1)}. \tag{5.27}$$

The exponent measure ν is connected to the angular measure Q introduced in (5.2). More precisely [13, Chapter 8], for any bounded, measurable function $f : \mathbb{E} \to \mathbb{R}$ that is zero in a neighborhood of the origin, we have

$$\nu(f) \stackrel{\text{def}}{=} \int_{\mathbb{E}} f(x)\,\mathrm{d}\nu(x) = d \int_{\Delta_{d-1}} \int_0^\infty f(yw) \frac{\mathrm{d}y}{y^2}\,\mathrm{d}Q(w). \tag{5.28}$$

In particular, for any for Borel set $B \subseteq \mathbb{E}$,

$$\begin{aligned} d^{-1}\nu(B \cap \mathbb{L}_1) &= \int_{\Delta_{d-1}} \int_0^\infty \mathbb{I}_{B\cap\mathbb{L}_1}(yw) \frac{\mathrm{d}y}{y^2}\,\mathrm{d}Q(w) \\ &= \int_{\Delta_{d-1}} \int_1^\infty \mathbb{I}_{B\cap\mathbb{L}_1}(yw) \frac{\mathrm{d}y}{y^2}\,\mathrm{d}Q(w) \\ &= \mathsf{P}(R\Theta \in B \cap \mathbb{L}_1), \end{aligned}$$

where R is unit-Pareto distributed and independent of $\Theta \sim Q$. Whence, by (5.27),

$$\mathsf{P}(Y \in B) = \frac{d\,\mathsf{P}(R\Theta \in B \cap \mathbb{L}_1)}{d\,\mathsf{P}(R\Theta \in \mathbb{L}_1)} = \mathsf{P}\left(R\Theta \in B \mid R\Theta \in \mathbb{L}_1\right),$$

From the local uniformity of the convergence in point (ii) of Remark 5.2.1, guaranteed by Assumption 5.2.3, and the fact that $\mathsf{P}(Y_j > 0) = 1$ for any $j \in \{1,\dots,d\}$ by the previous computations and Assumption 5.2.2, the extended continuous mapping theorem [122, Theorem 1.11.1] applies, yielding

$$\frac{b(t(t^{-1}V)) - b(t)}{a(t)} \,\Big|\, V \not\leq t \xrightarrow{\mathcal{L}} \frac{Y^\xi - 1}{\xi}, \qquad t \to \infty.$$

This concludes the proof of (5.7) and (5.8). Equation (5.9) is a direct consequence of (5.7) by noting that

$$\sigma_j(u_j(t)) = a_j\left(\frac{1}{1 - F_j(b_j(t))}\right) = a_j(t), \qquad t > 1$$

since $F_j(F_j^{-1}(p)) = p$ for all $0 < p < 1$ by continuity of F_j. □

5.B The Aitchison simplex

The open simplex Δ_{d-1}° is equipped with an inner product space structure with the following operations [1]: for $v = (v_1,\dots,v_d), w = (w_1,\dots,w_d) \in \Delta_{d-1}^{\circ}$ and $\alpha \in \mathbb{R}$, define the following operations:

- Addition:
$$v \oplus w \overset{\text{def}}{=} \left(\frac{v_j w_j}{\sum_{i=1}^d v_i w_i} \right)_{j=1}^d$$
- Scalar multiplication:
$$\alpha \odot v \overset{\text{def}}{=} \left(\frac{v_j^\alpha}{\sum_{i=1}^d v_i^\alpha} \right)_{j=1}^d$$
- (Aitchison) inner product:
$$\langle v, w \rangle_A \overset{\text{def}}{=} \sum_{i=1}^d \log\left(\frac{v_i}{g(v)}\right) \log\left(\frac{w_i}{g(w)}\right),$$
where $g : u \in \Delta_{d-1}^\circ \mapsto g(u) \overset{\text{def}}{=} (\prod_{i=1}^d u_i)^{1/d}$ is the geometric mean.

It is an easy exercise to verify that $(\Delta_{d-1}^\circ, \oplus, \odot)$ forms a real vector space with zero element $0_A = 1_d/d \in \Delta_{d-1}^\circ$, where $1_d = (1, \ldots, 1) \in \mathbb{R}^d$, and that $\langle \cdot, \cdot \rangle_A$ defines an inner product on it. The open simplex equipped with this structure is often called the *Aitchison simplex*.

A trivial but useful observation is that if one introduces the so-called *Centered LogRatio (CLR)* operator

$$\operatorname{clr} : u \in \Delta_{d-1}^\circ \mapsto \operatorname{clr}(u) \overset{\text{def}}{=} \left(\log\left(\frac{u_j}{g(u)}\right) \right)_{j=1}^d \in \mathbb{H} \subset \mathbb{R}^d, \tag{5.29}$$

where $\mathbb{H} \overset{\text{def}}{=} \{x \in \mathbb{R}^d : \langle x, 1_d \rangle = 0\}$ with $\langle \cdot, \cdot \rangle$ being the standard inner product in $\mathbb{R}^d$, then

$$\langle v, w \rangle_A = \langle \operatorname{clr}(v), \operatorname{clr}(w) \rangle, \qquad v, w \in \Delta_{d-1},$$

so that clr naturally defines an isometry between $(\Delta_{d-1}^\circ, \langle \cdot, \cdot \rangle_A)$ and $(\mathbb{H}, \langle \cdot, \cdot \rangle)$. The inverse transformation $\operatorname{clr}^{-1} : \mathbb{H} \mapsto \Delta_{d-1}^\circ$ is the well known *softmax* function

$$\operatorname{clr}^{-1}(x) = \left(\frac{\exp(x_j)}{\sum_{i=1}^d \exp(x_i)} \right)_{j=1}^d = \operatorname{softmax}(x), \qquad x \in \mathbb{H}.$$

The above considerations lead to a clear and easy way to construct an orthonormal basis of the Aitchison simplex:

1. Take any linearly independent family $\{x_1, \ldots, x_{d-1}\}$ in $\mathbb{H}$.

2. Apply the Gram–Schmidt procedure to produce an orthonormal basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{d-1}\}$ of $\mathbb{H}$ with respect to the standard inner product on $\mathbb{R}^d$.

3. Compute $\{\boldsymbol{e}_i^* \stackrel{\text{def}}{=} \operatorname{clr}^{-1}(\boldsymbol{e}_i) : i = 1, \ldots, d-1\}$, which forms an orthornormal basis of the Aitchison simplex.

Applying the above procedure to the free family

$$\left\{x_i \stackrel{\text{def}}{=} (0, \ldots, 0, 1, -1, 0, \ldots, 0) \in \mathbb{H} : i = 1, \ldots, d-1\right\},$$

where in x_i the 1 element lies at the i-th position, leads to the orthonormal basis of the Aitchison simplex presented in Proposition 5.2.2.

5.C Angular measure with respect to another norm

The angular measure Q introduced in (5.2) can be introduced with respect to an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, that is, if we let

$$\widetilde{R} = \|V\| \qquad \text{and} \qquad \widetilde{W} = \widetilde{R}^{-1}V,$$

one can consider the measure $\widetilde{Q}$ on $\widetilde{\Delta}_{d-1} \stackrel{\text{def}}{=} \{x \in [0,1]^d : \|x\| = 1\}$ obtained by taking, as in (5.2),

$$\widetilde{Q}(A) = \lim_{t\to\infty} \mathrm{P}\left(\widetilde{W} \in A \mid \widetilde{R} \geq t\right), \tag{5.30}$$

for any Borel set $A \subseteq \widetilde{\Delta}_{d-1}$ such that $\widetilde{Q}(\partial A) = 0$. Popular choices for $\|\cdot\|$ consist of the L_p norms $\|\cdot\|_p$ on $\mathbb{R}^d$ for $p \in [1, \infty]$, see, e.g., [42, 47, 25]. Even though we only consider the norm $\|\cdot\|_1$ in Algorithm 1, we show that a simple post-processing step permits to provide an estimate for $\widetilde{Q}$ also.

From (5.30) and (5.4), one may show that $\widetilde{Q}(A) = \widetilde{\Phi}(A)/\widetilde{\Phi}(\widetilde{\Delta}_{d-1})$ where $\widetilde{\Phi}$ is the finite Borel measure on $\widetilde{\Delta}_{d-1}$ obtained by taking

$$\widetilde{\Phi}(A) = \lim_{t\to\infty} t\, \mathrm{P}\left(\widetilde{W} \in A, \widetilde{R} \geq t\right)$$

provided $\widetilde{\Phi}(\partial A) = 0$. From (5.28), for any bounded, measurable $f : \mathbb{E} \to \mathbb{R}$ vanishing in a neighborhood of 0, we get

$$d \int_{\Delta_{d-1}} \int_0^\infty f(rw) \frac{\mathrm{d}r}{r^2}\, \mathrm{d}Q(w) = \int_{\widetilde{\Delta}_{d-1}} \int_0^\infty f(r\widetilde{w}) \frac{\mathrm{d}r}{r^2}\, \mathrm{d}\widetilde{\Phi}(\widetilde{w}),$$

Taking $f(x) = g(x/\|x\|)\mathbb{I}(\|x\| \geq 1)$ for $x \in \mathbb{E}$, where $g : \widetilde{\Delta}_{d-1} \to \mathbb{R}$ is some bounded function, we get

$$d \int_{\Delta_{d-1}} g\left(\frac{w}{\|w\|}\right) \|w\| \mathrm{d}Q(w) = \int_{\widetilde{\Delta}_{d-1}} g(\widetilde{w}) \mathrm{d}\widetilde{\Phi}(\widetilde{w}).$$

Picking $g(\widetilde{w}) = \mathbb{I}(\widetilde{w} \in \widetilde{\Delta}_{d-1})$ gives

$$\widetilde{\Phi}(\widetilde{\Delta}_{d-1}) = d \int_{\Delta_{d-1}} \|w\| \, \mathrm{d}Q(w),$$

so that for any bounded $g : \widetilde{\Delta}_{d-1} \to \mathbb{R}$ we have

$$\int_{\widetilde{\Delta}_{d-1}} g(\widetilde{w}) \, \mathrm{d}\widetilde{Q}(\widetilde{w}) = \frac{\int_{\Delta_{d-1}} g\left(\frac{w}{\|w\|}\right) \|w\| \, \mathrm{d}Q(w)}{\int_{\Delta_{d-1}} \|w\| \, \mathrm{d}Q(w)}. \tag{5.31}$$

In particular, $\widetilde{Q}$ is fully determined by Q.

Suppose that we observe $\Theta_1, \ldots, \Theta_n \sim Q$ as it would be (approximately) the case via the trained generator in Algorithm 1. They lead to the empirical estimator $\widehat{Q} = (1/n) \sum_{i=1}^n \delta_{\Theta_i}$ for Q. Relation (5.31) then justifies the approximation

$$\begin{aligned}\int_{\widetilde{\Delta}_{d-1}} g(\widetilde{w}) \, \mathrm{d}\widetilde{Q}(\widetilde{w}) &\approx \frac{\int_{\Delta_{d-1}} g\left(\frac{w}{\|w\|}\right) \|w\| \, \mathrm{d}\widehat{Q}(w)}{\int_{\Delta_{d-1}} \|w\| \, \mathrm{d}\widehat{Q}(w)} \\ &= \sum_{i=1}^n \Lambda_i g\left(\frac{\Theta_i}{\|\Theta_i\|}\right) = \int_{\widetilde{\Delta}_{d-1}} g(\widetilde{w}) \, \mathrm{d}\widehat{\widetilde{Q}}(\widetilde{w}),\end{aligned}$$

where

$$\widehat{\widetilde{Q}} \overset{\text{def}}{=} \sum_{i=1}^n \Lambda_i \delta_{\Theta_i/\|\Theta_i\|}, \qquad \Lambda_i \overset{\text{def}}{=} \frac{\|\Theta_i\|}{\sum_{i=1}^n \|\Theta_i\|}.$$

Said otherwise, a sample $\Theta_1, \ldots, \Theta_n \sim Q$ can be used to produce an estimator for $\widetilde{Q}$ by considering the weighted empirical distribution of the rescaled angles Θ_i on $\widetilde{\Delta}_{d-1}$, where the weights are proportional to the norm $\|\Theta\|_i$.

5.D Implementation of competing methods

HTGAN is follows a standard GAN architecture [61] with heavy-tailed latent variables that are Pareto(α) distributed for some $\alpha > 0$. It does not make use of (multivariate) extreme value theory at all and rests on the capacity of the GAN to accurately capture the dependence of the data, also in the tail. As advised in their paper, we base our implementation on the one in [86]. Their preprocessing step in their Algorithm 1 on page 10, equation (3.11), requires an estimate of $\gamma = 1/\alpha$ which is not provided in their methodology. They use the true value in their simulations, which is why we will also do so here, in Section 5.4.2. For real data applications, we will consider $\gamma = 1/1.5$ as recommended by the authors in their results on page 19. Note also that it is not clear to us how step (3.12) in their Algorithm 2 is supposed to work

as $\tilde{X}_j^{(i)}$ is not an element of $[0,1]$ so that it does not make sense to apply the inverse of the estimated marginal distribution function F_j. Since in this simulation setting the output of their generator is supposed to live in the same space as the original data, we do not apply any transform to compute the score (5.24). The score (5.22) is, of course, not affected by any marginal transformation. In real data applications, we do not compute the W_2 score and only consider the dependence score as done in the original paper. In terms of architectures and hyperparameters, we consider the same setting as for our method. This is very close to what the authors also do in their paper, besides the hyperparameter search which is performed through a Bayesian search in their numerical section. Both the generator and the discriminator are trained on 5000 epochs. The final layer of the discriminator has a sigmoid activation function, as it is commonly applied in GAN architectures.

The GPGAN is closer to our approach as it is also based on the MGP distribution to accurately sample from the tail of X, even though the approach in [85] is quite different. The authors also rely on a GAN architecture so that we also rely on [86] to implement their methodology. The main differences between their approach and ours are as follows:

- In [85], the dependence is captured by the spectral vector S as introduced in (5.11), while our techniques rely on sampling from Q after a transformation to the Aitchison coordinates which helps simplifying the distributions.

- The discriminator in the GPGAN procedure directly evaluates sample quality on the MGP scale (after evaluating the marginal parameters), while our technique separates the dependence structure, for which we rely on the WGAN architecture (Algorithm 1), and the marginal considerations that permit to transform the output back to the MGP scale (Algorithm 2). Our approach has the advantage that if one is only interested in evaluating a tail dependence measure, like the extremal coefficients in (5.22), it is possible to do so without putting any assumptions on the marginal tail behavior and to work only with Algorithm 1.

The adaptive procedure in [85] to pick the threshold u based on the "extremeness" function E was not entirely clear to us. Therefore, we decided to pick the same u as we do to compute the marginal parameters, so that we work with the same estimates of the parameters σ and ξ. Also, as in [85] the latent distribution is not indicated, we consider a multivariate standard normal distribution on the latent space. The architectures and hyperparameters are also treated in the same way as for our method and the HTGAN method. Both the generator and the discriminator are trained on 5000 epochs, based on the Adam optimizer, as done in the original paper [85]. The final layer of the

discriminator has a sigmoid activation function, as it is commonly applied in GAN architectures. Note that computing the score (5.24) poses no problem with this method as it outputs directly on the scale of $X \mid X \not\leq u(t)$; however, for the score (5.22), one needs to be more cautious. For this, we rely on the fact that if $X_1^*, \dots, X_n^*$ forms a sample from $X \mid X \not\leq u(t)$, then

$$V_i^* \overset{\text{def}}{=} b(X_i^*) \sim V \mid V \not\leq t = V \mid \|V\|_\infty > t, \qquad i \in \{1, \dots, n\}.$$

For $t > 1$ large enough, it means that

$$\Theta_i^* \overset{\text{def}}{=} \frac{V_i^*}{\|V_i^*\|_\infty} \overset{\mathcal{L}}{\approx} Q_\infty, \qquad i \in \{1, \dots, n\}$$

where Q_∞ the angular probability measure with respect to the L_∞-norm. Relying on the change of norm explained in Appendix 5.C, we get an estimate for Q by considering the discrete measure

$$\widehat{Q}^* \overset{\text{def}}{=} \sum_{i=1}^{n} \Lambda_i^* \delta_{\frac{\Theta_i^*}{\|\Theta_i^*\|_1}}, \qquad \Lambda_i^* \overset{\text{def}}{=} \frac{\|\Theta_i^*\|_1}{\sum_{i=1}^n \|\Theta_i^*\|_1}.$$

The estimator can be plugged in the formula of extremal coefficient (5.21) which is used in computing the score (5.22).

5.E Additional figures

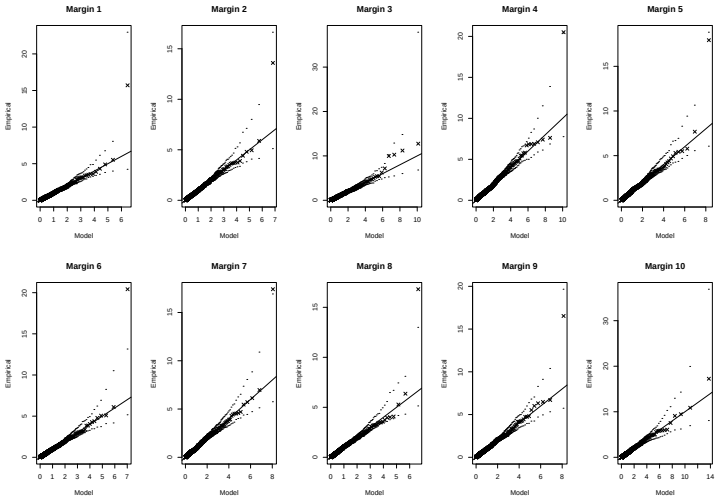


Figure 5.11: GPD-qqplots for margins 1 to 10 in the financial dataset.

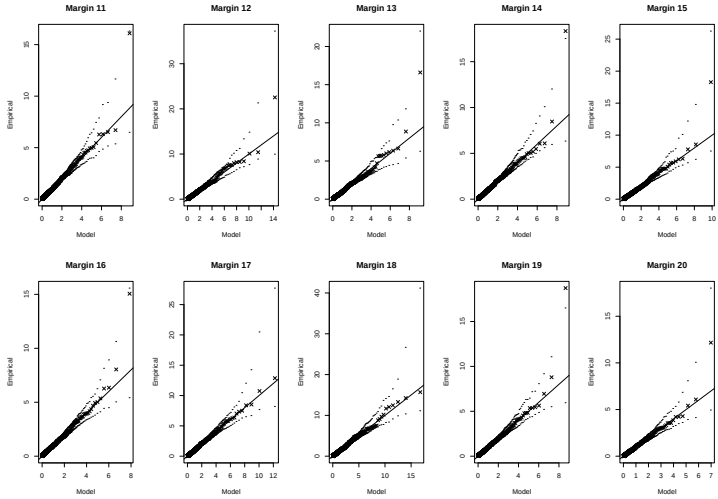


Figure 5.12: GPD-qqplots for margins 11 to 20 in the financial dataset.

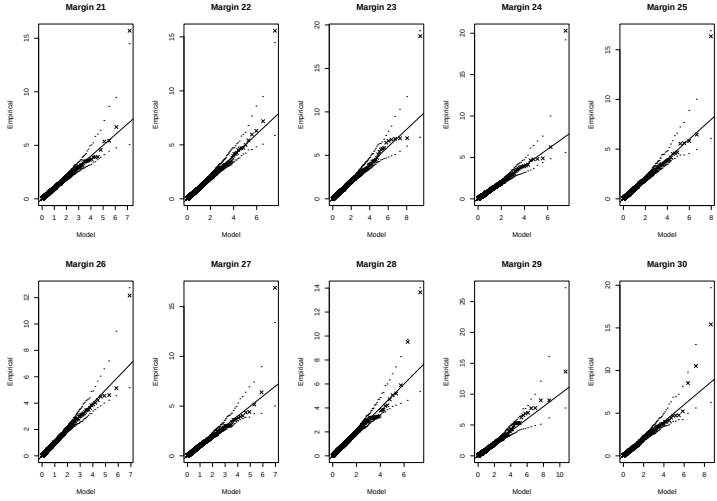


Figure 5.13: GPD-qqplots for margins 21 to 30 in the financial dataset.

Concluding Remarks and Perspectives

6

The angular measure plays a key role in assessing tail probabilities related to multivariate random phenomena. Statistical inference is involved, mainly in the nonparametric framework. This thesis attempts to provide a better understanding of this object by contributing to the probabilistic study of its sample version, from the finite and large sample perspective, and by developing statistical applications like classification, estimation, testing or the more modern generative approaches.

Despite those efforts, a lot remains to be covered regarding this complex object. A non-exhaustive list of ideas, in the theme of this dissertation, is provided below. Some of them are already in progress.

Generalised concentration bounds for the empirical angular measure. One possibility is to extend the concentration bounds for the process associated with the empirical angular measure $\widehat{\Phi}$ of Chapter 3 to a functional setting. More explicitly, we would study the finite sample behaviour of

$$\sqrt{k}\left\{\widehat{\Phi}(h) - \Phi(h)\right\}, \qquad h : \mathbb{S} \to \mathbb{R}, \quad h \in \mathcal{H}$$

in a functional sense, where $\mathcal{H}$ is a class of functions containing infinitely many functions of finite complexity (typically, a VC-class). This would lead to direct applications in regression settings where we are interested in constructing a prediction function $\widehat{f}$ which (approximately) minimizes a "tail least-square risk" defined as [73, Equation (3)]

$$R_\infty(f) \stackrel{\text{def}}{=} \lim_{t\to\infty} \mathsf{E}\left[(Y - f(X))^2 \mid \|X\| \geq t\right].$$

The limit exists under an appropriate regular variation assumption on the pair (X, Y) and minimization is performed using the Empirical Risk Minimization principle. The best possible performance with respect to the risk R_∞ is attained by an "angular" predictive rule, i.e., $f(x) = h(x/\|x\|)$ where $h : \mathbb{S} \to \mathbb{R}$, making the link between the angular measure and this regression problem. Such results would permit to extend the work in [73] where no preliminary empirical transformation to unit Pareto margins is considered.

Goodness-of-Fit tests in general dimension. Extending the goodness-of-fit procedure introduced in Chapter 4 to the general case of an angular (probability) measure Q in a space of general dimension $d > 2$ seems too ambitious given that no asymptotic theory exists in such case. As such, obtaining the asymptotic distribution of a test statistic of the form $\mathcal{W}(\widehat{Q}, Q_{\widehat{r}})$, the Wasserstein distance between a non-parametric estimate of Q and an estimation of its postulated parametric form, is impossible. What could be done instead is to restrict the class of possible angular measures to the ones supported on a finite number of points, an important case arising in max-linear or factor models [58]. Asymptotic theory in this framework can be developed as the margins are regularly varying with the same tail index and no preliminary transformation is needed so that the empirical estimator depends on the original data directly. Then, based on results of [118] which states directional Hadamard differentiability of the Wasserstein distance with respect to its arguments in the discrete case, the asymptotic distribution of the Wasserstein distance between the estimated and true angular measure can be derived. The additional challenge coming from the extreme value context is that, even if the true angular measure is discrete, the empirically observed extreme directions in the sample may not coincide, and even their number may be different. Clustering algorithms [77, 54] allowing to aggregate the atoms of the empirical angular measure into fewer atoms of larger mass are expected to be helpful in that respect.

Concentration of the empirical angular Wasserstein distance. Another idea would be to apply the bounds derived in Chapter 3 to obtain non-asymptotic guarantees on the Wasserstein distance between the empirical and the true angular measures. Technically, this can be done in two steps. The first step, the easiest one, is to show that the Wasserstein distance concentrates well around its expected value. This follows from a standard bounded difference argument [129, Section 6]. The second step is to control the expected value. This is done by the technique of dyadic transport [129, Proposition 1] which permits to bound the Wasserstein distance between two measures in terms of the differences of the masses they assign to the members of a nested sequence of partitions. Based on the bounds in Chapter 3, we may then bound this quantity and provide a bound on the expected Wasserstein distance. Combining the two steps lead to a finite-sample bound on the Wasserstein distance and will help providing guarantees in statistical learning procedures involving this quantity, like the celebrated k-means procedure [129, Section 7.2] also used for extremes [77].

Asymptotics for a Tail Pairwise Dependence Matrix. Applying the functional asymptotic expansion derived in Chapter 4 would permit to obtain

the asymptotic distribution of statistics derived from the pairwise angular measures of a multivariate random vector X. One such quantity could be an alternatively defined Tail Pairwise Dependence Matrix, as defined in [29], truly based on all the pairwise angular measures associated with pairs (X_j, X_k) for $j \neq k$. Estimation of this matrix could then be used to perform estimation in parametric models or dimension reduction to help visualizing the tail dependence structure of X.

Generative methods. Extending Chapter 5 by leveraging more recent generative modeling techniques beyond Wasserstein GANs would be particularly interesting. For example, diffusion models [69] have demonstrated remarkable performance in generating high-fidelity samples and exhibit improved training stability. Similarly, transformer based generative models [125, 101] offer a powerful alternative, especially in scenarios where capturing long-range dependencies is critical. These models could lead to more powerful and flexible generative approaches, improving performance across a wide range of tasks.

Bibliography

[1] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society serie B*, 44(2):139–177, 1982.

[2] Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. ExcessGAN: simulation above extreme thresholds using Generative Adversarial Networks. working paper or preprint, 2025.

[3] Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. Ev-gan: Simulation of extreme events with relu neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022.

[4] Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. On the simulation of extreme events with neural networks. *HAL preprint, hal-04416809v2*, pages 1–35, 2024.

[5] Hugues Annoye. *Some contributions to synthetic data creation.* PhD thesis, UCLouvain, 2024.

[6] Martin Anthony and John Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47(3):207–217, 1993.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[8] Peiman Asadi, Anthony C. Davison, and Sebastian Engelke. Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050, 2015.

[9] August A. Balkema and Laurens de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.

[10] Axel Bücher and Holger Dette. Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655—-1687, 2013.

[11] Axel Bücher, Johan Segers, and Stanislav Volgushev. When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *The Annals of Statistics*, 42(4):1598–1634, 2014.

[12] Jan Beirlant, Mikael Escobar-Bach, Yuri Goegebeur, and Armelle Guillou. Bias-corrected estimation of stable tail dependence function. *Journal of Multivariate Analysis*, 143:453–466, 2016.

[13] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, 2005.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1):289–300, 1995.

[15] Yoav Benjamini and Daniel Yejutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[16] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

[17] Marie-Odile Boldi and Anthony C. Davison. A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229, 2007.

[18] Stéphane Boucheron and Maud Thomas. Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics*, 9(2):2751–2792, 2015.

[19] Younes Boulaguiem, Jakob Zscheischler, Edoardo Vignotto, Karin van der Wiel, and Sebastian Engelke. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science*, 1:1–18, 2022.

[20] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Artificial Intelligence*, pages 169–207. Springer, Berlin, 2004.

[21] Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Springer, Berlin, Heidelberg, 2004.

[22] Sami Umut Can, John H. J. Einmahl, Estate V. Khmaladze, and Roger J. A. Laeven. Asymptotically distribution-free goodness-of-fit testing for tail copulas. *The Annals of Statistics*, 43(2):878–902, 2015.

[23] Daniela Castro Camilo and Miguel de Carvalho. Spectral density regression for bivariate extremes. *Stochastic Environmental Research and Risk Assessment*, 31:1603–1613, 2017.

[24] Emilie Chautru. Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383–418, 2015.

[25] Stéphan Clémençon, Hamid Jalalzai, Stéphane Lhaut, Anne Sabourin, and Johan Segers. Concentration bounds for the empirical angular measure with statistical learning applications. *Bernoulli*, 29(4):2797–2827, 2023.

[26] Stuart G. Coles and Jonathan A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B*, 53(2):377–392, 1991.

[27] Daniel Cooley, Richard A. Davis, and Philippe Naveau. The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117, 2010.

[28] Daniel Cooley, Richard A. Davis, and Philippe Naveau. Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *The Annals of Applied Statistics*, 6(4):1406–1429, 2012.

[29] Daniel Cooley and Emeric Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019.

[30] Bikramjit Das, Abhimanyu Mitra, and Sidney Resnick. Living on the multidimensional edge: seeking hidden risks using regular variation. *Advances in Applied Probability*, 45(1):139–163, 2013.

[31] Miguel de Carvalho and Anthony C. Davison. Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association*, 109(506):764–776, 2014.

[32] Miguel de Carvalho, Boris Oumow, Johan Segers, and Michal Warchol. A Euclidean likelihood estimator for bivariate tail dependence. *Communication in Statistics – Theory and Methods*, 42(7):1176–1192, 2013.

[33] Laurens de Haan and John de Ronde. Sea and wind: Multivariate extremes at work. *Extremes*, 1(1):7–45, 1998.

[34] Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction.* Springer Science & Business Media, 2007.

[35] Laurens de Haan, Cláudia Neves, and Liang Peng. Parametric tail copula estimation and model testing. *Journal of Multivariate Analysis*, 99(6):1260–1275, 2008.

[36] Laurens de Haan and Sidney Resnick. Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40:317–337, 1977.

[37] Laurens de Haan and Ashoke Kumar Sinha. Estimating the probability of a rare event. *The Annals of Statistics*, 27(2):732–759, 1999.

[38] Holger Drees and Anne Sabourin. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943, 2021.

[39] Juan J. Egozcue, Vera Pawlowsky-Glahn, Gloria Mateu-Figueras, and Carles Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[40] John H. J. Einmahl, Laurens de Haan, and Andrea Krajina. Estimating extreme bivariate quantile regions. *Extremes*, 16(2):121–145, 2013.

[41] John H. J. Einmahl, Laurens de Haan, and Deyuan Li. Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *The Annals of Statistics*, 34(4):1987–2014, 2006.

[42] John H. J. Einmahl, Laurens de Haan, and Vladimir I. Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423, 2001.

[43] John H. J. Einmahl, Laurens de Haan, and Ashoke Kumar Sinha. Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications*, 70(2):143–171, 1997.

[44] John H. J. Einmahl, Andrea Krajina, and Johan Segers. A method of moments estimator of tail dependence. *Bernoulli*, 14(4):1003–1026, 2008.

[45] John H. J. Einmahl, Andrea Krajina, and Johan Segers. An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793, 2012.

[46] John H. J. Einmahl and David M. Mason. Generalized quantile process. *The Annals of Statistics*, 20:1062–1078, 1992.

[47] John H. J. Einmahl and Johan Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989, 2009.

[48] Uwe Einmahl and David M. Mason. Gaussian approximation of local empirical processes indexed by functions. *Probability Theory and Related Fields*, 107:283–311, 1997.

[49] Sebastian Engelke, Adrien S. Hitz, Nicola Gnecco, and Manuel Hentschel. *graphicalExtremes: Statistical Methodology for Graphical Extreme Value Models*, 2024. R package version 0.3.2, `https://CRAN.R-project.org/package=graphicalExtremes`.

[50] Sebastian Engelke and Jevgenijs Ivanovs. Robust bounds in multivariate extremes. *The Annals of Applied Probability*, 27(6):3706–3734, 2017.

[51] Sebastian Engelke and Jevgenijs Ivanovs. Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8:241–270, 2021.

[52] Sebastian Engelke, Alexander Malinowski, Zakhar Kabluchko, and Martin Schlather. Estimation of hüsler—reiss distributions and brown–resnick processes. *Journal of the Royal Statistical Society. Series B*, 77(1):239–265, 2015.

[53] Ana Ferreira and Laurens de Haan. The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717 – 1737, 2014.

[54] Viacheslav Fomichov and Jevgenijs Ivanovs. Spherical clustering in detection of groups of concomitant extremes. *Biometrika*, 110(1):135–153, 2023.

[55] Anne-Laure Fougères, Laurens de Haan, and Cécile Mercadier. Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934, 2015.

[56] Evarist Giné and Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[57] Stéphane Girard, Emmanuel Gobet, and Jean Pachebat. Deep generative modeling of multivariate dependent extremes. *HAL preprint, hal-04700084v2*, pages 1–35, 2024.

[58] Nadine Gissibl and Claudia Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24:2693–2720, 2018.

[59] Nicolas Goix, Anne Sabourin, and Stephan Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31, 2017.

[60] Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. Learning the dependence structure of rare events: a non-asymptotic study. *Journal of Machine Learning Research*, 40:1–18, 2015.

[61] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[62] Armelle Guillou, Philippe Naveau, and Alexandre You. A folding methodology for multivariate extremes: estimation of the spectral probability measure and actuarial applications. *Scandinavian Actuarial Journal*, 2015(7):549–572, 2015.

[63] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[64] Emil Julius Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

[65] Ali Hasan, Khalil Elkhalil, Yuting Ng, João M. Pereira, Sina Farsiu, Jose H. Blanchet, and Vahid Tarokh. Modeling extremes with d-max-decreasing neural networks. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *Proceedings of Machine Learning Research*, pages 759–768. PMLR, 2022.

[66] David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[68] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[69] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[70] Jürg Hüsler and Rolf-Dieter Reiss. Maxima of normal random vectors: Between independence and complete dependence. *Statistics and Probability Letters*, 7(4):283–286, 1989.

[71] Chenglei Hu and Daniela Castro-Camilo. GPDFlow: Generative Multivariate Threshold Exceedance Modeling via Normalizing Flows, 2025.

[72] Shuang Hu, Zuoxiang Peng, and Johan Segers. Modelling multivariate extreme value distributions via markov trees. *Scandinavian Journal of Statistics*, 51(2):760–800, 2024.

[73] Nathan Huet, Stephan Clémençon, and Anne Sabourin. On regression in extreme regions. arXiv:2303.03084, 2024.

[74] Henrik Hult and Filip Lindskog. Regular variation for measures on metric spaces. *Publications de l'Institut Mathematique*, 80(94):121–140, 2006.

[75] Callum J. R. Murphy-Barltrop, R. Majumder, and J. Richards. Deep Learning of Multivariate Extremes via a Geometric Representation. *arXiv:2406.19936v2*, pages 1–66, 2024.

[76] Hamid Jalalzai, Stephan Clémençon, and Anne Sabourin. On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pages 3092–3100, 2018.

[77] Anja Janßen and Phyllis Wan. k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233, 2020.

[78] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[79] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

[80] Nicolas Lafon, Philippe Naveau, and Ronan Fablet. A vae approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*, 2023.

[81] Jaakko Lehtomaa and Sidney I. Resnick. Asymptotic independence and support detection techniques for heavy-tailed multivariate data. *Insurance: Mathematics and Economics*, 93:262–277, 2020.

[82] Stéphane Lhaut. Inégalités de concentration pour évènements rares. Master's thesis, Faculté des sciences, Université catholique de Louvain, 2021. Prom. : Segers, Johan.

[83] Stéphane Lhaut, Anne Sabourin, and Johan Segers. Uniform concentration bounds for frequencies of rare events. *Statistics and Probability Letters*, 189:109610, 2022.

[84] Stéphane Lhaut and Johan Segers. An asymptotic expansion of the empirical angular measure for bivariate extremal dependence. In Matteo Barigozzi, Siegfried Hörmann, and Davy Paindaveine, editors, *Recent Advances in Econometrics and Statistics*. Springer Cham, November 2024.

[85] Jiawei Li, Dong Li, Peng Li, and Gennady Samorodnitsky. Generalized pareto gan: Generating extremes of distributions. In *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[86] Erik Linder-Norén. PyTorch–GAN, 2025. Available on: `https://github.com/eriklindernoren/PyTorch-GAN`.

[87] G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lect.*, pages 1–56. Springer, Vienna, 2002.

[88] Gabor Lugosi. Pattern classification and learning theory. In László Györfi, editor, *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lect.*, pages 1–56. Springer, Vienna, 2002.

[89] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013. Workshop on Deep Learning for Audio, Speech and Language Processing.

[90] Colin McDiarmid. Concentration. In Michel et al. Habib, editor, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 195–248. Springer Berlin Heidelberg, 1998.

[91] Nicolas Meyer and Olivier Wintenberger. Sparse regular variation. *Advances in Applied Probability*, 53(4):1115–1148, 2021.

[92] Thomas Mikosch and Olivier Wintenberger. *Extreme Value Theory for Time Series: Models with Power-Law Tails*. Springer Series in Operations Research and Financial Engineering. Springer Cham, 2024.

[93] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2018.

[94] Lambert De Monte, Raphaël Huser, Ioannis Papastathopoulos, and Jordan Richards. Generative modelling of multivariate geometric extremes using normalising flows, 2025.

[95] Ian D. Morton and John Andrew Bowers. Extreme value analysis in a multivariate offshore environment. *Applied Ocean Research*, 18(6):303–317, 1996.

[96] Anas Mourahib, Anna Kiriliouk, and Johan Segers. Multivariate generalized Pareto distributions along extreme directions. *Extremes*, pages 10.1007/s10687–024–00501–4, 2024.

[97] Philippe Naveau and Johan Segers. Multivariate extreme value theory. *arXiv preprint*, pages 1–33, 2024.

[98] Serguei Novak. *Extreme value methods with applications to finance*. CRC Press, 2011.

[99] Art Owen. Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747, 1991.

[100] Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019.

[101] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[102] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style,

high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[103] James Pickands III. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[104] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[105] Sidney Resnick. *The Art of Finding Hidden Risks: Hidden Regular Variation in the 21st Century*. Springer Cham, 2024.

[106] Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 1987.

[107] Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media, 2007.

[108] Marco Rocco. Extreme value theory in finance: a survey. *Journal of Economic Surveys*, 28(1):82–108, 2014.

[109] Holger Rootzén, Johan Segers, and Jennifer L. Wadsworth. Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131, 2018.

[110] Holger Rootzén, Johan Segers, and Jennifer L. Wadsworth. Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145, 2018.

[111] Holger Rootzén and Nader Tajvidi. Multivariate generalized pareto distributions. *Bernoulli*, 12(5):917–930, 2006.

[112] Anne Sabourin. Semi-parametric modeling of excesses above high multivariate thresholds with censored data. *Journal of Multivariate Analysis*, 136:126–146, 2015.

[113] Anne Sabourin and Philippe Naveau. Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics and Data Analysis*, 71:542–567, 2014.

[114] Dominic Schuhmacher, Björn Bähre, Nicolas Bonneel, Carsten Gottschlich, Valentin Hartmann, Florian Heinemann, Bernhard Schmitzer, and Jörn Schrieber. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, 2024. R package version 0.15-4.

[115] Clayton Scott and Robert Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

[116] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 2009.

[117] Richard L Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207, 1987.

[118] Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

[119] Jonathan A. Tawn. Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415, 1988.

[120] Albert Thomas, Stéphan Clemencon, Alexandre Gramfort, and Anne Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1011–1019, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[121] Aad van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[122] Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag New York, 1996.

[123] Vladimir N. Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, XVI(2):264–280, 1971.

[124] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer New York, NY, 2000.

[125] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[126] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate studies in mathematics*. American Mathematical Society, 2003.

[127] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

[128] Jennifer L. Wadsworth, Jonathan A. Tawn, Anthony C. Davison, and Daniel M. Elton. Modelling across extremal dependence classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):149–175, 2017.

[129] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[130] Jakob Benjamin Wessel, Callum J. R. Murphy-Barltrop, and Emma S. Simpson. A comparison of generative deep learning methods for multivariate angular simulation, 2025.