Concentration bounds for the empirical angular measure with statistical learning applications

Stéphan Clémençon, Hamid Jalalzai, Stéphane Lhaut, Anne Sabourin, Johan Segers

Concentration bounds for the empirical angular measure with statistical learning applications

STÉPHAN CLÉMENÇON[1,a], HAMID JALALZAI[1,b], STÉPHANE LHAUT[2,c], ANNE SABOURIN[3,e] and JOHAN SEGERS[2,d]

[1] *LTCI, Télécom Paris, Palaiseau , France.*
[a] *stephan.clemencon@telecom-paris.fr,* [b] *hamid.jalalzai@telecom-paris.fr*
[2] *LIDAM/ISBA, UCLouvain, Louvain-la-Neuve, Belgium.*
[c] *stephane.lhaut@uclouvain.be,* [d] *johan.segers@uclouvain.be*
[3] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.*
[e] *anne.sabourin@u-paris.fr*

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.

MSC2020 subject classifications: Primary 62G05; 62G30; 62G32; secondary 62H30
Keywords: angular measure; classification; concentration inequality; extreme value analysis; minimum-volume sets; ranks

1. Learning from multivariate extremes

Estimation and prediction problems regarding the extremal behaviour of a d-dimensional random vector X 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 [19], describes the first-order dependence structure of joint extremes of X and is rooted in the theory of multivariate regular variation [46]. 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 [14,54] 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 [17,20]. 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 [25]. It serves to model the action of a covariate on the extremal dependence of a baseline distribution through a density ratio model [8,15]. The angular density is also at the basis of an estimator of bivariate tail quantile regions [24]. Bounds for probabilities of joint excesses over high thresholds that are robust against misspecification of the angular measure are derived in [29]. The angular measure underlies techniques to find groups of variables exhibiting extremal dependence, with dimension reduction and sparse representations as objectives [9,40,44]; see [30] for a review. In [38], 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 [13,21]. Anomalous data can be detected from unusual combinations of variables being large simultaneously [33] 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 [52]. Binary classification in extreme regions can be performed on the basis of the differences between the intra-class angular measures of the explanatory variables [37].

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 [11,12] and new ones continue to be invented, with a growing emphasis on the use of flexible models almost of a semi-parametric nature [4,48,49].

Our focus is on non-parametric estimation and inference via the empirical angular measure, $\widehat{\Phi}_p$, introduced in [26] and generalized to every L_p-norms in [28], although already alluded to some years earlier [17,20]. 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 [26,28]. In higher dimensions, only its consistency has been established [38, 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 [7,23]. 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 [16,28] as well as a folding technique [34]. 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 [27]. 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 3.1, 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

$$\sup_{A\in\mathcal{A}}\left|\widehat{\Phi}_p(A)-\Phi_p(A)\right| \tag{1}$$

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 [26] and [28], although the framing sets are defined in a different manner here (see Section 3.1); and second, a general concentration inequality for empirical processes indexed by rare events given in Theorem A.1. The latter result is inspired by a similar inequality in [32] 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 $1\leqslant p\leqslant\infty$, i.e., for $x\in\mathbb{R}^d$,

$$\|x\|_p=\begin{cases}(|x_1|^p+\cdots+|x_d|^p)^{1/p} & \text{if } 1\leqslant p<\infty,\\ \max(|x_1|,\ldots,|x_d|) & \text{if } p=\infty.\end{cases} \tag{2}$$

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 [52] and [37], 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 [52]. 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 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 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 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. Section 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 5. A discussion in Section 6 concludes the paper. Proofs and some auxiliary results, in particular a general concentration inequality for rare event probabilities, are deferred to the Appendices and the Supplement.

2. Upper tail dependence

We set out some basics of multivariate extreme value theory (Sections 2.1 and 2.2) and recall a rank-based standardization procedure (Section 2.3). Equipped with these notions, we describe the empirical angular measure (Section 2.4). For background, we refer to monographs such as [2,18,46].

2.1. Regular variation and exponent measure

Let $X = (X_1, \ldots, X_d)$ be a random vector with distribution P and continuous marginal cumulative distribution functions $F_j(u) = \mathbb{P}[X_j \leqslant u]$ 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 $V_j = 1/(1 - F_j(X_j))$ for $j = 1, \ldots, d$.

The working hypothesis in this paper is that the resulting random vector $V = (V_1, \ldots, V_d)$ 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 μ on the punctured orthant $E = [0, \infty)^d \setminus \{(0, \ldots, 0)\}$ which is finite on Borel sets bounded away from the origin and such that

$$\lim_{t \to \infty} t\, \mathbb{P}[t^{-1} V \in B] = \mu(B) \tag{3}$$

for all Borel sets B of E bounded away from the origin and such that $\mu(\partial B) = 0$, with ∂B the topological boundary of B. Convergence in (3) for all such B is equivalent to measure convergence $t\, \mathbb{P}[t^{-1} V \in \cdot\,] \to \mu(\cdot)$ as $t \to \infty$ in the space $\mathcal{M}_0$ of Borel measures on E that are finite on Borel sets bounded away from the origin [36]. Specifically, we have $\lim_{t \to \infty} t\, \mathbb{E}[f(t^{-1} V)] = \int_E f \mathrm{d}\mu$ for every bounded and continuous real function f on 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 $[0, \infty]^d \setminus \{(0, \ldots, 0)\}$ [46,47].

The measure μ 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 [18,

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

$$\forall \lambda > 0, \qquad \mu(\lambda \cdot) = \lambda^{-1} \mu(\cdot). \tag{4}$$

Its margins are standardized: by (3) and since each component V_j is unit-Pareto distributed, we have

$$\forall y \in (0, \infty),\ \forall j \in \{1, \dots, d\}, \qquad \mu(\{x \in E : x_j \geqslant y\}) = y^{-1}. \tag{5}$$

2.2. Angular measure

Recall $\|x\|_p$ in (2) and let

$$\mathbb{S}_p = \{x \in [0, \infty)^d : \|x\|_p = 1\}, \tag{6}$$

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

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

$$\Phi_p(A) = \mu(C_A) \quad \text{where} \quad C_A = \{x \in E : \|x\|_p \geqslant 1,\ \theta_p(x) \in A\}. \tag{7}$$

In view of the marginal standardization (5) and the ensuing identity (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 $1 \leqslant \Phi_p(\mathbb{S}_p) \leqslant d$.

The exponent measure μ is determined by the angular measure: by homogeneity (4),

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

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

$$\int_E f \, \mathrm{d}\mu = \int_{\mathbb{S}_p} \int_0^\infty f(r\theta) \, \frac{\mathrm{d}r}{r^2} \, \mathrm{d}\Phi_p(\theta). \tag{8}$$

The combination of the marginal standardization (5) with the change-of-variable formula in (8) implies the identities

$$\forall j = 1, \dots, d, \qquad \int_{\mathbb{S}_p} \theta_j \, \mathrm{d}\Phi_p(\theta) = 1. \tag{9}$$

In fact, any non-negative Borel measure Φ_p on $\mathbb{S}_p$ satisfying the moment constraints (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 μ on E through (8) and then consider the max-stable distribution with μ as exponent measure as in [19].

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

$$\forall x \in \mathbb{R}^d, \qquad v(x) = \left(\frac{1}{1 - F_1(x_1)}, \dots, \frac{1}{1 - F_d(x_d)} \right), \tag{10}$$

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 $P_n(\cdot) = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i \in \cdot\}$, where $\mathbb{1}\{\mathcal{E}\}$ denotes the indicator variable of the event $\mathcal{E}$. In particular, the marginal cumulative distribution functions F_j are substituted with their empirical counterparts

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

in order to mimic the aforementioned standardization.

The empirically standardized sample points are

$$\widehat{V}_i = \hat{v}(X_i) = \big(\hat{v}_1(X_{i1}), \ldots, \hat{v}_d(X_{id})\big), \qquad i = 1, \ldots, n \tag{11}$$

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

$$\hat{v}_j(x_j) = \frac{1}{1 - \frac{n}{n+1} \widehat{F}_j(x_j)}, \tag{12}$$

for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. The factor $\frac{n}{n+1}$ in (12) serves to avoid division by zero in case $x_j \geqslant \max(X_{1j}, \ldots, X_{nj})$.

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 componentwise ranks. This will be the case for the empirical angular measure.

2.4. The empirical angular measure

Let δ_x denote a point mass at x and put

$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\widehat{V}_i}, \tag{13}$$

the empirical distribution of the pseudo-observations $\widehat{V}_1, \ldots, \widehat{V}_n$ in (11). The random measure $\widehat{P}_n$ can be legitimately considered as an estimator of the distribution of V.

The definition (3) of the exponent measure μ 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

$$\widehat{\mu}(B) = \tfrac{n}{k} \widehat{P}_n(\tfrac{n}{k} B) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\left\{\widehat{V}_i \in \tfrac{n}{k} B\right\} \tag{14}$$

for Borel sets $B \subseteq E$. Consider a Borel set $A \subseteq \mathbb{S}_p$ and recall the angular measure Φ_p and the cone C_A in (7). In view of (14), *the empirical angular measure* [26] is defined as

$$\widehat{\Phi}_p(A) = \widehat{\mu}(C_A) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\left\{\widehat{V}_i \in (n/k) C_A\right\} = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\left\{\|\widehat{V}_i\|_p \geqslant n/k,\, \theta_p(\widehat{V}_i) \in A\right\}. \tag{15}$$

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

$$t\, \mathbb{P}[t^{-1} V \in C_A]$$

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 [26], while the L_p-norm for $p \in [1, \infty]$ was studied in [28] together with empirical likelihood methods to exploit the moment constraints (9).

3. Concentration inequalities

Recall the empirical angular measure $\widehat{\Phi}_p$ in (15). Our main result provides a concentration inequality for the uniform deviations

$$\sup_{A \in \mathcal{A}} \left| \widehat{\Phi}_p(A) - \Phi_p(A) \right|. \tag{16}$$

The supremum is taken over classes $\mathcal{A}$ of Borel sets A of $\mathbb{S}_p$ in (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 3.1) before stating our main theorem on concentration bound for the empirical angular measure (Section 3.2). We conclude with examples of collections $\mathcal{A}$ (Section 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 (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 (13) and let $P_n = n^{-1} \sum_{i=1}^n \delta_{V_i}$ denote the empirical distribution of $V_1, \ldots, V_n$. 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

$$\widehat{\Gamma}_A = \hat{v}^{-1}(\tfrac{n}{k} C_A) = \left\{ x \in E : \ \|\hat{v}(x)\|_p \geqslant \tfrac{n}{k}, \, \theta_p(\hat{v}(x)) \in A \right\}, \tag{17}$$

where, as before, $\theta_p(y) = y/\|y\|_p$ for $y \in E$. 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 [26] and [28], 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

$$\forall A \in \mathcal{A}, \qquad \tfrac{n}{k} \Gamma_A^- \subseteq \widehat{\Gamma}_A \subseteq \tfrac{n}{k} \Gamma_A^+. \tag{18}$$

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

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

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

$$\left.\begin{aligned}\left|\widehat{\Phi}_p(A)-\Phi_p(A)\right| \leqslant & \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}}\left|\tfrac{n}{k}P(\tfrac{n}{k}B)-\mu(B)\right| && \text{(bias term)}\\ & + \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}}\tfrac{n}{k}\left|P_n(\tfrac{n}{k}B)-P(\tfrac{n}{k}B)\right| && \text{(stochastic error)}\\ & + \mu(\Gamma_A^+\setminus\Gamma_A^-) && \text{(framing gap).}\end{aligned}\right\} \tag{19}$$

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 (18) holds. The task is then to control the three terms on the right-hand side of (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 [1,31]; however, see Remark 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 A.1. Finally, the framing gap can be controlled by ensuring that the set $\Gamma_A^+\setminus\Gamma_A^-$ is small.

3.2. Concentration bounds for the empirical angular measure

The main result of this paper, Theorem 3.1, is stated in this subsection. Let $\mathbb{B}=[-1,+1]^d$ 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 $A+B=\{a+b: a\in A, b\in B\}$. For $\varepsilon>0$ and $A\subseteq\mathbb{R}^d$, we thus have $A+\varepsilon\mathbb{B}=\{x\in\mathbb{R}^d:\exists a\in A,\|x-a\|_\infty\leqslant\varepsilon\}$.

Assumption 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 $\lim_{n\to\infty}\mathbb{1}\{x\in A_n\}=\mathbb{1}\{x\in A\}$ for every $x\in\mathbb{S}_p$.

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

$$\forall A\in\mathcal{A},\qquad A\subseteq\{x\in\mathbb{S}_p:\ \min(x)>\tau\}=:\mathbb{S}_p^\tau. \tag{20}$$

(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

$$\Phi_p\big(A_+(\varepsilon)\setminus A_-(\varepsilon)\big)\leqslant c\,\varepsilon \tag{21}$$

together with the inclusions

$$\big(A_-(\varepsilon)+\varepsilon\mathbb{B}\big)\cap\mathbb{S}_p\subseteq A\qquad\text{and}\qquad(A+\varepsilon\mathbb{B})\cap\mathbb{S}_p\subseteq A_+(\varepsilon). \tag{22}$$

Condition (i) amounts to the pointwise measurability of the indicators $\{\mathbb{1}\{\cdot\in A\}:A\in\mathcal{A}\}$ [53, 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. (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 (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 (12), we frame the cones C_A in Eq. (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

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

For all $A \in \mathcal{A}$ and $h > 0$, we have $\Gamma_A^-(r,h) \subseteq C_A$ for $0 < r \leqslant 1$ and $C_A \subseteq \Gamma_A^+(r,h)$ for $r \geqslant 1$. The upper confidence bound at level $1-\delta$ for the maximal deviation (1) stated in Theorem 3.1 below is derived from the decomposition (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 [53] and [6].

Assumption 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 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 3.2 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 we provide a realistic example of an angular class $\mathcal{A}$—namely a class defined by linear inequality constraints—where Assumptions 3.1 and 3.2 are satisfied.

Theorem 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 (10). Let μ 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 3.1 and 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 3.1. Let $\delta \in (0,1)$. Put*

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

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

$$\begin{aligned}\sup_{A\in\mathcal{A}}\left|\widehat{\Phi}_p(A) - \Phi_p(A)\right| \leqslant{}& \sup_{A\in\mathcal{A},\ \sigma\in\{+,-\}}\left|\tfrac{n}{k}P_V\left(\tfrac{n}{k}\Gamma_A^\sigma\right) - \mu(\Gamma_A^\sigma)\right| \\ &+ \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(60\sqrt{V_{\mathcal{F}}} + 2\sqrt{\log((d+1)/\delta)}\right) + \frac{2\log\left(\frac{d+1}{\delta}\right)}{3k} \\ &+ 2d\Delta_2 + 3c\big(\log(d/3c) - \log(\Delta_1) + 1\big)\Delta_1,\end{aligned} \tag{23}$$

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

$$\left\{\Gamma_A^- : A \in \mathcal{A}\right\} \cup \left\{\Gamma_A^+ : A \in \mathcal{A}\right\}. \tag{24}$$

The proof is given in Appendix C in the Supplement.

Note that $r_- \leqslant 1 \leqslant r_+$, so that $\Gamma_A^-(r_-, h) \subseteq C_A \subseteq \Gamma_A^+(r_+, h)$ for all $h > 0$. In the decomposition (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 A.1 applied to the collection $\{(n/k)\Gamma_A^\sigma(r_\sigma, 3\Delta_1) : \sigma \in \{+,-\},\ A \in \mathcal{A}\}$. By Assumption 3.2, the collection in (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 $d_1 + d_2 + 1$ [45, 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

$$\forall x \in \mathbb{R}^d, \qquad \|x\|_\infty \leqslant \|x\|_p \leqslant d^{1/p}\|x\|_\infty. \tag{25}$$

For fixed ρ and δ, when ignoring the bias term, the bound on the right-hand side of (23) is of the order

$$\Delta_1 \log(1/\Delta_1) = \mathrm{O}\left(\frac{\log k}{\sqrt{k}}\right), \text{ as } k \to \infty. \tag{26}$$

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 (19) and (23). Asymptotic theory for the empirical angular measure in the bivariate case [26,28] 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 3.1 (Other concentration inequalities). The constant 56 appearing in the error stems comes from the use of chaining techniques as in [42, Theorems 1.16–17]. Relying instead on concentration inequalities for rare events derived recently in [41], 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 3.1 comes from the application of Theorem 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 [51]. 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 3.2 (Bias term and penultimate angular measure). As Theorem 3.1 is not concerned with asymptotics, we did not actually have to assume that Φ_p is the angular measure associated P_V. The link between P_V and Φ_p is quantified instead by the bias term $\sup_{A,\sigma} |\frac{n}{k} P_V(\frac{n}{k}\Gamma_A^\sigma) - \mu(\Gamma_A^\sigma)|$. 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 (Controlling the bias term). For absolutely continuous models, a primitive condition on the probability density function allows to control the bias term in Theorem 3.1. 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 $\iota(x) = (x_1^{-1}, \ldots, x_d^{-1})$. Assume that P_U is absolutely continuous with density p_U and that the measure $\Lambda = \mu \circ \iota^{-1}$ (the push-forward of μ by ι) is absolutely continuous with Lebesgue density λ on $(0,\infty)^d$. Then

$$\sup_{A\in\mathcal{A},\sigma\in\{+,-\}} \left|\tfrac{n}{k} P_V(\tfrac{n}{k}\Gamma_A^\sigma) - \mu(\Gamma_A^\sigma)\right| \leqslant \int_{(0,\infty)^d} \mathbb{1}\left\{\min(y) \leqslant d^{1/p} r_+\right\} \left|\left(\tfrac{k}{n}\right)^{d-1} p_U(\tfrac{k}{n}y) - \lambda(y)\right| \mathrm{d}y. \tag{27}$$

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

$$f(x) = 2^d \Gamma(\tfrac{d+1}{2}) \pi^{-(d+1)/2} \left(1 + x_1^2 + \cdots + x_d^2\right)^{-(d+1)/2}, \qquad x \in (0,\infty)^d,$$

with limit density

$$\lambda(x) = 2^{d-1} \Gamma(\tfrac{d+1}{2}) \pi^{-(d-1)/2} x_1^{-2} \cdots x_d^{-2} \left(x_1^{-2} + \cdots + x_d^{-2}\right)^{-(1+d)/2}, \qquad x \in (0,\infty)^d.$$

In this case, the bound (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 D in the Supplement.

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 3.1 and 3.2 are satisfied. Proofs are deferred to Appendix E in the Supplement.

Recall $\mathbb{S}_p^\tau$ in (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 3.1 (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

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

Then $\mathcal{A}$ satisfies Assumption 3.2, and, if Φ_p has a bounded Lebesgue density on $\mathbb{S}_p$, also Assumption 3.1.

Example 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 3.1 and 3.2. Then the same is true for the collection of intersections

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

and for the collection of unions

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

In combination with Example 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.

4. Applications to statistical learning

We illustrate how a concentration inequality such as the one in Theorem 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 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 4.2, we recall the whys and wherefores of classification in extreme regions and leverage the techniques developed in Section 3 to control the excess risk of a specific empirical risk minimizer targeting the tail region of the covariate space.

4.1. Minimum-volume set estimation

To illustrate the usefulness of bounds on the supremum in (1), consider the problem of estimating a *minimum-volume set* [22], 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 [3]. 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 [52] 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 [52], 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 $\hat{\Phi}_p$ be any estimator of Φ_p, not necessarily the empirical angular measure. Following the logic in [50], let $\hat{A}$ solve the empirical angular minimum-volume set problem

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

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 $\hat{\Phi}_p$ over $\mathcal{A}$ are helpful to provide probabilistic guarantees for $\hat{A}$. Suppose that, on some event $\mathcal{E}$, we have

$$\sup_{A \in \mathcal{A}} \left|\hat{\Phi}_p(A) - \Phi_p(A)\right| \leqslant \psi. \tag{28}$$

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

$$\Phi_p(\hat{A}) \geqslant \hat{\Phi}_p(\hat{A}) - \psi \geqslant \alpha - 2\psi \tag{29}$$

as well as

$$\lambda(\hat{A}) \leqslant \inf\left\{\lambda(A) :\ A \in \mathcal{A},\ \Phi_p(A) \geqslant \alpha\right\}. \tag{30}$$

Indeed, on $\mathcal{E}$, the collection $\{A \in \mathcal{A} : \Phi_p(A) \geqslant \alpha\}$ is contained in $\{A \in \mathcal{A} : \hat{\Phi}_p(A) \geqslant \alpha - \psi\}$ so that the infimum of $\lambda(A)$ for A in the latter collection must be the smaller one. By (29), the empirical solution $\hat{A}$ is guaranteed to have at least content $\alpha - 2\psi$ under Φ_p, while according to (30), the volume of $\hat{A}$ is smaller than the one of the actual minimum-volume set under Φ_p.

Concentration inequalities for the supremum of $|\hat{\Phi}_p(A) - \Phi_p(A)|$ over $A \in \mathcal{A}$ provide the existence, for a given $\delta > 0$, of a scalar $\psi \equiv \psi(\delta)$ such that (28) holds on an event with probability at least $1 - \delta$. It follows that, with high probability, the empirical minimum-volume set $\hat{A}$ satisfies (29) and (30). Provided ψ and δ are both small, the combination of both properties justifies the use of $\hat{A}$ 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 3.1.

4.2. Classification in extreme regions

We apply Theorem 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 [37], who formalize this argument and propose a risk minimization strategy aiming at improving performance of classification algorithms on such regions. In [37], 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 [37]. 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 $\varrho = \mathbb{P}[Y = 1]$ and assume $0 < \varrho < 1$.

The starting point in [37] is to assume a conditional version of the regular variation condition (3): there exist non-zero Borel measures μ_+ and μ_- on $E = [0, \infty)^d \setminus \{0\}$ that are finite on Borel sets bounded away from the origin and such that

$$\lim_{t \to \infty} t\, \mathbb{P}[t^{-1} V \in B \mid Y = \sigma\, 1] = \mu_\sigma(B) \tag{31}$$

for $\sigma \in \{-, +\}$ and Borel sets $B \subseteq E$ bounded away from the origin satisfying $\mu_\sigma(\partial B) = 0$. The angular measure associated to the L_p-norm of μ_σ is

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

with C_A as in (7). By (31), the unconditional distribution of V is regularly varying as in (3) with limit measure $\mu = \varrho\mu_+ + (1 - \varrho)\mu_-$ and angular measure $\Phi_p = \varrho\Phi_p^+ + (1 - \varrho)\Phi_p^-$.

In [37] 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 (31) holds with $V = v(X)$, where v is defined via the probability integral transform in (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 (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 [37] 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:

$$L_t^{\mathrm{cond}}(g) = \mathbb{P}[Y \neq g(V) \mid \|V\|_p > t], \tag{32}$$

Let $\mathcal{G}$ be a pre-defined family of classifiers g and let $L_t^{\mathrm{cond}^*} = \inf_{g \in \mathcal{G}} L_t^{\mathrm{cond}}(g)$ 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 [37] 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 $x \in E$. 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 (31). Let $\eta(x) = \mathbb{P}[Y = 1 \mid V = x]$ denote the regression function of (V, Y) and let $\eta_\infty(x) = \mathbb{P}[Y_\infty = 1 \mid V_\infty = x]$ be the one of (V_∞, Y_∞). The respective Bayes classifiers are

$$\begin{aligned} g^*(x) &= \mathbb{1}\{\eta(x) \geqslant 1/2\}, \\ g^*_\infty(x) &= \mathbb{1}\{\eta_\infty(x) \geqslant 1/2\}. \end{aligned} \tag{33}$$

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:

$$\sup_{x \in [0,\infty)^d : \|x\|_p \geqslant t} |\eta(x) - \eta_\infty(x)| \to 0 \quad \text{as } t \to \infty. \tag{34}$$

Then by Theorem 1 in [37], (i) the asymptotic Bayes classifier g^*_∞ is angular, and (ii) the latter's excess conditional risk over the actual Bayes classifier g^* vanishes in the limit, that is, $L_t^{\mathrm{cond}}(g^*_\infty) - L_t^{\mathrm{cond}}(g^*) \to 0$ 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 [37] 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 (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 (32) and define its unconditional version by

$$L_t(g) = t\,\mathbb{P}[\|V\|_p \geqslant t] L_t^{\mathrm{cond}}(g) = t\,\mathbb{P}[g(V) \neq Y,\ \|V\|_p \geqslant t]. \tag{35}$$

The multiplicative factor $t\,\mathbb{P}[\|V\|_p \geqslant t]$ 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 3.1 required in Section 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 (20). We have

$$L_t(g) = L_t^{>\tau}(g) + L_t^{\leqslant\tau}(g) \tag{36}$$

with

$$\begin{aligned} L_t^{>\tau}(g) &= t\,\mathbb{P}[g(V) \neq Y,\ \theta_p(V) \in \mathbb{S}_p^\tau,\ \|V\|_p \geqslant t], \\ L_t^{\leqslant\tau}(g) &= t\,\mathbb{P}[g(V) \neq Y,\ \theta_p(V) \notin \mathbb{S}_p^\tau,\ \|V\|_p \geqslant t]. \end{aligned}$$

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

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

We work hereafter under the following smoothness assumption. Let ∂A denote the boundary of set A.

Assumption 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 4.1. *If the conditional regular variation property* (31) *and Assumption 4.1 hold, then for any angular classifier* $g \in \mathcal{G}$*,*

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

$$\lim_{t\to\infty} L_t^{\leqslant\tau}(g) = L_\infty^{\leqslant\tau}(g) := \varrho\Phi_p^+(\mathbb{S}_p^-(g) \setminus \mathbb{S}_p^\tau) + (1-\varrho)\Phi_p^-(\mathbb{S}_p^+(g) \setminus \mathbb{S}_p^\tau),$$

and thus

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

The proof is deferred to Appendix F. The idea of the decomposition (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 $L_t^{\leqslant\tau}(g)$ can be expected to be small for τ close zero, since

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

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

$$\widehat{L}^\tau(g) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\left\{g(\widehat{V}_i) \neq Y_i,\ \theta_p(\widehat{V}_i) \in \mathbb{S}_p^\tau,\ \|\widehat{V}_i\|_p \geqslant n/k\right\}. \tag{37}$$

Assuming a minimizer exists, the ERM classifier is defined as

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

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 4.1. A consequence of Theorem 1 in [37] is that if (34) holds, the Bayes classifier g_∞^* in (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

$$L_\infty(\hat{g}_k^\tau) - \inf_{g\in\mathcal{G}} L_\infty(g) \leqslant 2\sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty(g)\right|. \tag{38}$$

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 $L_\infty^{\leqslant\tau}(g) \leqslant \Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau)$ and thus

$$\sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty(g)\right| \leqslant \sup_{g\in\mathcal{G}} \left|\widehat{L}^\tau(g) - L_\infty^{>\tau}(g)\right| + \Phi_p(\mathbb{S}_p \setminus \mathbb{S}_p^\tau). \tag{39}$$

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 3.1 and Theorem 4.1 grow roughly as $1/\sqrt{\tau}$ as $\tau \to 0$. The choice of τ thus constitutes an additional bias-variance compromise.

Lemma F.1 in Appendix F parallels Lemma 4.1 by relating the empirical risk $\widehat{L}^\tau(g)$ to the empirical angular measures of the positive and negative instances,

$$\widehat{\Phi}_p^\sigma(A) = \frac{1}{k_\sigma} \sum_{i=1}^{n} \mathbb{1}\{Y_i = \sigma 1\} \cdot \mathbb{1}\left\{\theta_p(\widehat{V}_i) \in A,\ \|\widehat{V}_i\|_p \geqslant n/k\right\}, \qquad A \subseteq \mathbb{S}_p, \sigma \in \{-,+\}, \tag{40}$$

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

In view of the error decomposition (38)–(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.

Theorem 4.1 (Deviations of the empirical tail risk). *Let $\mathcal{G}$ be a class of angular classifiers. Consider the collection $\mathcal{A} = \{\mathbb{S}_p^+(g) \cap \mathbb{S}_p^\tau : g \in \mathcal{G}\} \cup \{\mathbb{S}_p^-(g) \cap \mathbb{S}_p^\tau : g \in \mathcal{G}\}$. If Assumptions 3.1 and 3.2 relative to the class $\mathcal{A}$ and the unconditional angular measure Φ_p are satisfied and if the conditional regular variation assumption (31) and Assumption 4.1 hold, then, with probability at least $1-\delta$,*

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

where error *and* gap *are nearly the same as in Theorem 3.1 (δ has been halved in the* error *term), that is,*

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

with Δ_1, Δ_2 and $V_{\mathcal{F}}$ as defined in Theorem 3.1, and

$$\text{bias II} = \sup\Big\{\Big|\tfrac{n}{k}\,\mathbb{P}[V \in \tfrac{n}{k}B, Y = \sigma 1] - \mathbb{P}[Y = \sigma 1]\mu_\sigma(B)\Big| : B = \Gamma_A^+ \text{ or } B = \Gamma_A^- \text{ for some } A \in \mathcal{A} \text{ and } \sigma \in \{-,+\}\Big\}.$$

The proof relies on the relationships between the (empirical) classification risks and the (empirical) angular measure pointed out in Lemmata 4.1 and F.1, which imply that

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

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 3.1; see Appendix F in the Supplement for details.

5. Simulation experiments

Our experiments aim at illustrating the influence of the threshold τ introduced in Equation (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 3.1 is zero. The estimation error then only stems from the stochastic error and framing gap in (19) leading to the terms on the second and third lines in (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]

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, let $\theta_p(x) = \|x\|_p^{-1} x$ for $x \in \mathbb{R}^d \setminus \{0\}$. Recall $\mathbb{S}_p^\tau = \mathbb{S}_p \cap (\tau, \infty)^d$. Fix $0 < \tau < 1$. 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 $\mathbb{S}_1 = \{x \in [0, 1]^d : x_1 + \cdots + x_d = 1\}$ 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 G.1 in the Supplement, the angular measure Φ_p for $p \in [1, \infty]$ is $\Phi_p(A) = d\,\mathbb{P}[X \in C_A]$ for Borel sets $A \subseteq \mathbb{S}_p$. If $p = 1$, then this simplifies to $\Phi_1(A) = d\,\mathbb{P}[\Theta \in A]$.

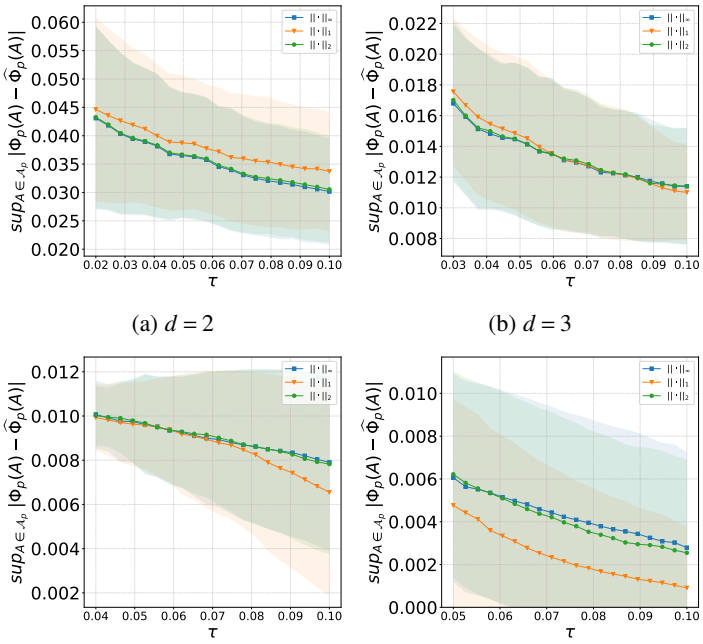
In this setting, all statistics involved in our analysis are easily computable. The following facts are an immediate consequence of Lemma G.1 in the Supplement. Let $A \subseteq \mathbb{S}_p^\tau$ be a Borel set with $\tau \in (0, 1)$ and recall $\mathbb{S}_p^\tau$ in (20).

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

$$\Phi_{p,\mathrm{MC}}(A) = \frac{d}{N} \sum_{i=1}^{N} \mathbb{1}\left\{X_i' \in C_A\right\} = \frac{d}{N} \sum_{i=1}^{N} \mathbb{1}\left\{\Theta_i'/\|\Theta_i'\|_p \in A,\ R_i'\|\Theta_i'\|_p \geqslant 1\right\}, \tag{41}$$

  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 $\Phi_p(A) = \frac{n}{k}\,\mathbb{P}[V \in \frac{n}{k} C_A]$ as soon as $n/k > d/\tau$, so that the bias term in Theorem 3.1 is null under the latter condition.

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



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

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

Figure 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 (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 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 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 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 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

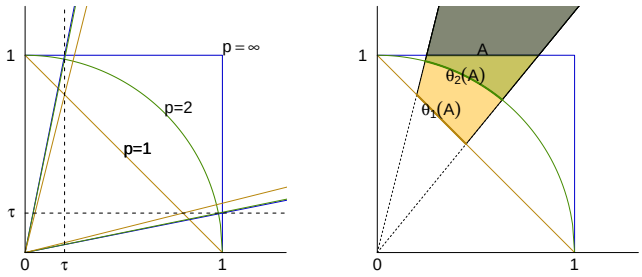


Figure 2: Restriction to coordinates larger than τ has an heavier impact on the L_1-sphere than on the L_2-sphere and the L_∞-sphere, which touch each other near the coordinate axes (left panel). The cone $\{x \in [0,\infty)^p : \|x\|_p \geqslant 1, x \in \theta_p(A)\}$ generated by the projection $\theta_p(A)$ 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).

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 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 1, the latter is dominated by the former when the dimension increases.

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 [26,28]. 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 [5].

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.

Appendix A: Concentration inequality for rare events

The main concentration tool that we use in the proof of Theorem 3.1 is the following. Since the result may be of independent interest, we provide a detailed proof.

Theorem 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*

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

In [32, Theorem 1], a similar inequality is derived, but with a generic constant C that is not made explicit. Just before their Lemma 14, there is a reference to [39] 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 [42], 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 A.1. It is based on a McDiarmid's concentration inequality for the supremum around its expected value in [43], extending Bernstein's classical inequality and recalled in [32, Proposition 11]. We rewrite it in a form which is more convenient for us [41, Proposition 5].

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

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

Proof of Theorem A.1. Apply Proposition A.1. To bound the remaining expectation, we make use of Theorem 1.16 in [42], which relies on a chaining argument to ensure that

$$\mathbb{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leqslant \frac{24}{\sqrt{n}} \max_{x_1,\ldots,x_n\in\mathcal{X}} \int_0^1 \sqrt{\log\left(2\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right)\right)}\,\mathrm{d}r, \tag{42}$$

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 $\mathcal{A}(x_1^n) := \left\{(\mathbb{1}_A(x_i))_{i=1}^n : A \in \mathcal{A}\right\} \subseteq \{0,1\}^n$ with respect to the metric $\rho(b,c) = n^{-1/2}\|b-c\|_2$ for $b, c \in \{0,1\}^n$, see [42, page 29] for definitions and notation. Haussler [35] showed that, if $\mathcal{A}$ has a finite VC-dimension, then the associated covering number is bounded in the following way: for any $0 < r \leqslant 1$,

$$\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right) \leqslant e(V_{\mathcal{A}}+1)\left(\frac{2e}{r^2}\right)^{V_{\mathcal{A}}}. \tag{43}$$

Using (43) in (42) we can bound the integral as follows: for any points $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\int_0^1 \sqrt{\log\left(2\mathcal{N}\left(r, \mathcal{A}(x_1^n)\right)\right)}\,\mathrm{d}r \leqslant \int_0^1 \sqrt{\log\left(2e(V_{\mathcal{A}}+1)\right) + V_{\mathcal{A}}\log\left(2e/r^2\right)}\,\mathrm{d}r$$

$$\leqslant \sqrt{V_{\mathcal{A}}} \int_0^1 \sqrt{3\log 2 + 2 - 2\log r}\, \mathrm{d}r \leqslant 2.44\sqrt{V_{\mathcal{A}}},$$

by numerical integration or by expressing the integral in terms of the standard normal cumulative distribution function. Combining this result and (42), we get

$$\mathbb{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leqslant 59\sqrt{\frac{V_{\mathcal{A}}}{n}}. \tag{44}$$

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 [41, Lemma 2], which states that if $K = \sum_{i=1}^n \mathbb{1}\{X_i \in B\}$, with $B \supseteq \bigcup_{A\in\mathcal{A}} A$ as in the statement of the theorem, we have the distributional equality

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

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 [41, Lemma 6]

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

The conditional expectation on the right-hand side is bounded by means of (44), giving

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

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

Supplementary Material

Supplement to "Concentration bounds for the empirical angular measure with statistical learning applications"
The supplement contains some auxiliary lemmas and proofs of all results developed in the paper: main theorem, examples, binary classification application and simulations.

Acknowledgments

The authors would like to thank two anonymous Referees for helpful comments and suggestions on an earlier version of the paper. A significant part of the work done by Anne Sabourin on this project was accomplished while she was an associate professor at Télécom Paris.

References

[1] Beirlant, J., Escobar-Bach, M., Goegebeur, Y. and Guillou, A. (2016). Bias-corrected estimation of stable tail dependence function. *J. Multivariate Anal.* **143** 453–466.

[2] BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J. (2005). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics.
[3] BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009.
[4] BOLDI, M. O. and DAVISON, A. C. (2007). A mixture model for multivariate extremes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 217–229.
[5] BOUCHERON, S. and THOMAS, M. (2015). Tail index estimation, concentration and adaptivity. *Electron. J. Stat.* **9** 2751–2792.
[6] BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2004). Introduction to Statistical Learning Theory. In *Advanced Lectures on Machine Learning*, (O. Bousquet, U. von Luxburg and G. Rätsch, eds.). *Lecture Notes in Artificial Intelligence* **3176** 169–207. Springer, Berlin.
[7] BÜCHER, A., SEGERS, J. and VOLGUSHEV, S. (2014). When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *Ann. Statist.* **42** 1598–1634.
[8] CASTRO CAMILO, D. and DE CARVALHO, M. (2017). Spectral density regression for bivariate extremes. *Stoch. Environ. Res. Risk Assess.* **31** 1603–1613.
[9] CHAUTRU, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electron. J. Stat.* **9** 383–418.
[10] CLÉMENÇON, S., JALALZAI, H., LHAUT, S., SABOURIN, A. and SEGERS, J. (2022). Concentration bounds for the empirical angular measure with statistical learning applications. Submitted to *Bernoulli*.
[11] COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **53** 377–392.
[12] COOLEY, D., DAVIS, R. A. and NAVEAU, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *J. Multivariate Anal.* **101** 2103–2117.
[13] COOLEY, D. and THIBAUD, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika* **106** 587–604.
[14] DAS, B., MITRA, A. and RESNICK, S. (2013). Living on the multidimensional edge: seeking hidden risks using regular variation. *Adv. in Appl. Probab.* **45** 139–163.
[15] DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776.
[16] DE CARVALHO, M., OUMOW, B., SEGERS, J. and WARCHOŁ, M. (2013). A Euclidean likelihood estimator for bivariate tail dependence. *Comm. Statist. Theory Methods* **42** 1176–1192.
[17] DE HAAN, L. and DE RONDE, J. (1998). Sea and wind: multivariate extremes at work. *Extremes* **1** 7–45.
[18] DE HAAN, L. and FERREIRA, A. (2007). *Extreme Value Theory: An Introduction*. Springer Science & Business Media.
[19] DE HAAN, L. and RESNICK, S. I. (1977). Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **40** 317–337.
[20] DE HAAN, L. and SINHA, A. K. (1999). Estimating the probability of a rare event. *Ann. Statist.* **27** 732–759.
[21] DREES, H. and SABOURIN, A. (2021). Principal component analysis for multivariate extremes. *Electron. J. Stat.* **15** 908–943.
[22] EINMAHL, J. H. J. and MASON, D. M. (1992). Generalized quantile process. *Ann. Statist.* **20** 1062–1078.
[23] EINMAHL, J. H., KRAJINA, A. and SEGERS, J. (2012). An M-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.* **40** 1764–1793.
[24] EINMAHL, J. H. J., DE HAAN, L. and KRAJINA, A. (2013). Estimating extreme bivariate quantile regions. *Extremes* **16** 121–145.
[25] EINMAHL, J. H. J., DE HAAN, L. and LI, D. (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *Ann. Statist.* **34** 1987–2014.
[26] EINMAHL, J. H. J., DE HAAN, L. and PITERBARG, V. I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Statist.* **29** 1401–1423.
[27] EINMAHL, J. H. J., DE HAAN, L. and SINHA, A. K. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Process. Appl.* **70** 143–171.
[28] EINMAHL, J. H. J. and SEGERS, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.* **37** 2953–2989.

[29] ENGELKE, S. and IVANOVS, J. (2017). Robust bounds in multivariate extremes. *Ann. Appl. Probab.* **27** 3706–3734.
[30] ENGELKE, S. and IVANOVS, J. (2021). Sparse structures for multivariate extremes. *Annu. Rev. Stat. Appl.* **8** 241–270.
[31] FOUGÈRES, A.-L., DE HAAN, L. and MERCADIER, C. (2015). Bias correction in multivariate extremes. *Ann. Statist.* **43** 903–934.
[32] GOIX, N., SABOURIN, A. and CLÉMENÇON, S. (2015). Learning the Dependence Structure of Rare Events: a Nonasymptotic Study. In *Proceedings of the International Conference on Learning Theory, COLT'15*.
[33] GOIX, N., SABOURIN, A. and CLÉMENÇON, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *J. Multivariate Anal.* **161** 12–31.
[34] GUILLOU, A., NAVEAU, P. and YOU, A. (2015). A folding methodology for multivariate extremes: estimation of the spectral probability measure and actuarial applications. *Scand. Actuar. J.* **2015** 549–572.
[35] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217–232.
[36] HULT, H. and LINDSKOG, F. (2006). Regular variation for measures on metric spaces. *Publ. Inst. Math. (Beograd) (N.S.)* **94** 121–140.
[37] JALALZAI, H., CLÉMENÇON, S. and SABOURIN, A. (2018). On Binary Classification in Extreme Regions. In *Advances in Neural Information Processing Systems* 3092–3100.
[38] JANSSEN, A. and WAN, P. (2020). k-means clustering of extremes. *Electron. J. Stat.* **14** 1211–1233.
[39] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *Ann. Statist.* **34** 2593–2706.
[40] LEHTOMAA, J. and RESNICK, S. I. (2020). Asymptotic independence and support detection techniques for heavy-tailed multivariate data. *Insurance Math. Econom.* **93** 262–277.
[41] LHAUT, S., SABOURIN, A. and SEGERS, J. (2022). Uniform concentration bounds for frequencies of rare events. *Statistics and Probability Letters* **189** 109610.
[42] LUGOSI, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001). CISM Courses and Lect.* **434** 1–56. Springer, Vienna.
[43] MCDIARMID, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics. Algorithms Combin.* **16** 195–248. Springer, Berlin.
[44] MEYER, N. and WINTENBERGER, O. (2021). Sparse regular variation. *Adv. in Appl. Probab.* **53** 1115–1148.
[45] MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2018). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.
[46] RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes* **4**. Springer-Verlag, New York.
[47] RESNICK, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media.
[48] SABOURIN, A. (2015). Semi-parametric modeling of excesses above high multivariate thresholds with censored data. *J. Multivariate Anal.* **136** 126–146.
[49] SABOURIN, A. and NAVEAU, P. (2014). Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Comput. Statist. Data Anal.* **71** 542–567.
[50] SCOTT, C. and NOWAK, R. (2006). Learning minimum volume sets. *J. Mach. Learn. Res.* **7** 665–704.
[51] SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics.
[52] THOMAS, A., CLEMENCON, S., GRAMFORT, A. and SABOURIN, A. (2017). Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. SINGH and J. ZHU, eds.). *Proceedings of Machine Learning Research* **54** 1011–1019. PMLR, Fort Lauderdale, FL, USA.
[53] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Process. With Applications to Statistics*. Springer-Verlag, New York.
[54] WADSWORTH, J. L., TAWN, J. A., DAVISON, A. C. and ELTON, D. M. (2017). Modelling across extremal dependence classes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79** 149–175.

Supplement

This supplement contains mathematical details related to in the paper [10]. Appendix B contains auxiliary results. The proof of the main result in the paper, Theorem 3.1, is given in Appendix C. Appendix D provides a practical criterion to control the bias term in Theorem 3.1 and applies it to the multivariate Cauchy distribution. In Appendix E, it is verified that the examples of collections of sets $\mathcal{A}$ in Section 3.3 in the paper indeed satisfy Assumptions 3.1 and 3.2. The proofs of the results in Section 4.2 on classification in extreme regions are given in Appendix F, while Appendix G contains an additional result related to the simulation experiments in Section 5 in the paper.

Appendix B: Auxiliary results

We will use the following minor results in the proof of Theorem 3.1.

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

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

then

$$\left|\frac{\|v\|_p}{\|x\|_p} - 1\right| \leqslant \frac{h}{1-h} =: \Delta < 1,$$

which implies

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

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

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

hence

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

and we deduce

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

In particular,

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

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

$$\big|\|v\|_p - \|x\|_p\big| \leqslant \|v - x\|_p \leqslant \frac{h}{1-h}\|x\|_p.$$

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

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

Those inequalities imply the second statement. □

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

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

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

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

Therefore, by the triangle inequality,

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

□

Lemma 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*

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

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

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

By the triangle inequality, we have

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

and

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

□

Appendix C: Proof of Theorem 3.1

Proof of Theorem 3.1. The empirical exponent and angular measures $\widehat{\mu}$ and $\widehat{\Phi}_p$ only depend on the sample $X_1, \dots, X_n$ through the transformed data

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

for $i = 1, \dots, n$ and $j = 1, \dots, d$. The sum of indicators is equal to the rank of X_{ij} within $X_{1j}, \dots, X_{nj}$. As the marginal cumulative distribution function F_j is continuous, the ranks of $X_{1j}, \dots, X_{nj}$ are with probability one equal to those of $V_{1j}, \dots, V_{nj}$, where $V_{ij} = 1/(1 - F_j(X_{ij}))$. Hence, even though the margins $F_1, \dots, F_d$ are unknown, the fact that $\widehat{\mu}$ and $\widehat{\Phi}_p$ are rank statistics implies that for the sake of the proof, we may and will henceforth assume that the margins $F_1, \dots, F_d$ 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 $r_- \leqslant 1 \leqslant r_+$, 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

$$\forall A \in \mathcal{A}, \qquad \tfrac{n}{k}\Gamma_A^- \subseteq \widehat{\Gamma}_A \subseteq \tfrac{n}{k}\Gamma_A^+. \tag{45}$$

Then, on the event $\mathcal{E}_1$, the decomposition (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 (23).

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

$$\mathcal{F} = \left\{ \tfrac{n}{k}\Gamma_A^{\sigma} : \sigma \in \{-,+\},\ A \in \mathcal{A} \right\}, \tag{46}$$

which has finite VC dimension $V_{\mathcal{F}}$ in view of Assumption 3.2 and the paragraph following Theorem 3.1; the rescaling by the factor n/k obviously does not change the VC dimension. For every $A \in \mathcal{A}$, we have

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

By the equivalence of norms (25), since the margins of P are unit-Pareto, it follows that the probability κ appearing in Theorem A.1 applied to $\mathcal{F}$ is bounded by

$$P\left[\cup_{j=1}^{d} \left\{ x \in [0,\infty)^d : x_j \geqslant \tfrac{n}{k} \tfrac{1}{d^{1/p}(1+\Delta_2)} \right\} \right] \leqslant d^{1+1/p}(1+\Delta_2)\tfrac{k}{n}. \tag{47}$$

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

$$\sup_{A \in \mathcal{A}, \sigma \in \{-,+\}} \tfrac{n}{k} \left| P_n\left(\tfrac{n}{k}\Gamma_A^{\sigma}\right) - P\left(\tfrac{n}{k}\Gamma_A^{\sigma}\right) \right|$$

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

This is the second term on the right-hand side of the bound (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,

$$\begin{aligned}\mu(\Gamma_A^+ \setminus \Gamma_A^-) \leqslant \mu\left(\left\{x \in [0,\infty)^d :\ (1+\Delta_2)^{-1} \leqslant \|x\|_p \leqslant (1-\Delta_2)^{-1}\right\}\right) \\ + \mu\left(\left\{x \in [0,\infty)^d :\ \|x\|_p \geqslant 1,\ \theta_p(x) \in A_+(3\Delta_1\|x\|_p) \setminus A_-(3\Delta_1\|x\|_p)\right\}\right).\end{aligned} \tag{48}$$

The first term on the right-hand side in (48) is equal to

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

The second term on the right-hand side in (48) can be computed via the product representation (8) of μ in polar coordinates: in view of (21) in Assumption 3.1, the result is

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

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

The resulting upper bound in (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 (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 (45) hold. To do this, we apply Theorem A.1 to each of the collections

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

Fix $j = 1,\dots,d$ and let $P_{n,j} = n^{-1}\sum_{i=1}^n \delta_{X_{ij}}$. 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 A.1, there exists an event $\mathcal{E}_{1,j}$ with probability at least $1 - \delta/(d+1)$ on which

$$\begin{aligned}\sup_{x_j \geqslant \frac{n}{k}\rho}\left|P_{n,j}((x_j,\infty)) - x_j^{-1}\right| &\leqslant \frac{k}{n}\left\{\sqrt{\frac{1}{k\rho}}\left(56 + 2\sqrt{\log((d+1)/\delta)}\right) + \frac{2}{3k}\log((d+1)/\delta)\right\} \\ &= \frac{k}{n}\Delta_1.\end{aligned} \tag{49}$$

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

$$\forall x_j \geqslant \tfrac{n}{k}\rho, \qquad \frac{n+1}{nx_j^{-1} + k\Delta_1 + 1} \leqslant \hat{v}_j(x_j) \leqslant \frac{n+1}{(nx_j^{-1} - k\Delta_1)_+ + 1}. \tag{50}$$

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

$$\forall x_j \leqslant \tfrac{n}{k}\rho, \qquad \hat{v}_j(x_j) \leqslant \hat{v}_j(\tfrac{n}{k}\rho) \leqslant \frac{n+1}{k(\rho^{-1}-\Delta_1)+1} \leqslant \frac{n}{k}\frac{\rho}{1-\rho\Delta_1} \leqslant \frac{n}{k}\tau. \tag{51}$$

Let $\mathcal{E}_1 = \bigcap_{j=1}^d \mathcal{E}_{1,j}$, the probability of which is at least $1 - d\delta/(d+1)$, as required. We need to show that on $\mathcal{E}_1$, the inclusions (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 $\|\hat{v}(x)\|_p \geqslant \frac{n}{k}$ but there exists $j = 1,\dots,d$ with $x_j \leqslant \frac{n}{k}\rho$, then, by (51), we have on $\mathcal{E}_1$ the bound

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

and thus, by Assumption 3.1, necessarily $\theta_p(\hat{v}(x)) \notin A$ for all $A \in \mathcal{A}$. Hence, on $\mathcal{E}_1$, we have

$$\widehat{\Gamma}_A = \left\{x \in (\tfrac{n}{k}\rho,\infty)^d : \ \|\hat{v}(x)\|_p \geqslant \tfrac{n}{k},\ \theta_p(\hat{v}(x)) \in A\right\}. \tag{52}$$

Step 2: Radial framing. — We consider the following decomposition of $\widehat{\Gamma}_A$:

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

and we frame each set separately.

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

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

for every $x_j > \frac{n}{k}\rho$ (which is larger than 1, by assumption that $\rho > k/n$). We simply apply Lemma B.2 combined with (49), giving

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

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

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

This leads to the inclusions

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

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 $\Delta_1 + 1/k \leqslant 1/4$, we have by (50),

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

Hence, we have $\|\hat{v}(x)\|_\infty \geqslant n/k$, and, by equivalence of norms (25), also $\|\hat{v}(x)\|_p \geqslant n/k$ for every $p \in [1, \infty]$. Furthermore, $\|x\|_p \geqslant \|x\|_\infty \geqslant 2\frac{n}{k} \geqslant \frac{n}{k} \frac{1}{1+\Delta_2}$ for every $0 < \Delta_2 < 1$. These inequalities imply the inclusions

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

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

$$\begin{aligned}
\widehat{\Gamma}_A &\subseteq \left\{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_p \geqslant \tfrac{n}{k} \tfrac{1}{1+\Delta_2}\right\} \\
\widehat{\Gamma}_A &\subseteq \left\{x \in (\tfrac{n}{k}\rho, \infty)^d : \|x\|_p \geqslant \tfrac{n}{k} \tfrac{1}{1-\Delta_2}\right\}.
\end{aligned} \tag{53}$$

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

$$\theta_p(x) \in A_-(3\tfrac{k}{n}\Delta_1 \|x\|_p) \implies \theta_p(\hat{v}(x)) \in A \implies \theta_p(x) \in A_+(3\tfrac{k}{n}\Delta_1 \|x\|_p). \tag{54}$$

In combination with (53), this will show the inclusions (45) and thus finish the proof.

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

If $\varepsilon \geqslant 1$, the two implications in (54) are trivially fulfilled: Since the $\|\cdot\|_p$-diameter of $\mathbb{S}_p$ is equal to 1, we have $A_-(\varepsilon) = \varnothing$ (as $\mathbb{S}_p \setminus A$ is not-empty) while $A_+(\varepsilon) = \mathbb{S}_p$.

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

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

Since $\varepsilon < 1$, we have $\|x\|_\infty \leqslant \|x\|_p \leqslant (n/k)/(3\Delta_1)$. Hence, for all $j = 1, \dots, d$, we have $nx_j^{-1} - k\Delta_1 \geqslant n \cdot 3\frac{k}{n}\Delta_1 - k\Delta_1 > 0$. By (50), we deduce

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

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

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

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

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

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

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

Consequently,

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

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

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

The implications (54) now follow by definition of $A_-(\varepsilon)$ and $A_+(\varepsilon)$.

We conclude that, on $\mathcal{E}_1$, the inclusions (45) hold, as required. The proof Theorem 3.1 is complete. □

Appendix D: Proofs of Remark 3.3

As $\|x\|_p \geqslant \rho$ implies $\|x\|_\infty \geqslant d^{-1/p}\rho$ for $x \in \mathbb{R}^d$ and $\rho > 0$, the bias term appearing in Theorem 3.1 is bounded by the total variation distance between the measures $\frac{n}{k}P_V(\frac{n}{k}\cdot)$ and μ restricted to

$$E_c = \left\{x \in (0,\infty)^d : \max(x) \geqslant c\right\},$$

for $c = d^{-1/p}r_+^{-1}$. (This c is not the same as the one in the statement of Theorem 3.1.) More precisely, we have

$$\sup_{A\in\mathcal{A},\ \sigma\in\{+,-\}} \left|\tfrac{n}{k}P_V(\tfrac{n}{k}\Gamma_A^\sigma) - \mu(\Gamma_A^\sigma)\right| \leqslant \sup_{B\in\mathcal{B}(E_c)} \left|\tfrac{n}{k}P_V(\tfrac{n}{k}B) - \mu(B)\right|,$$

where $\mathcal{B}(E_c)$ denotes the Borel σ-field on E_c. Recall p_U and λ in Remark 3.3 and define

$$\mathcal{D}_T(s) = \int_{L_T} \left|s^{d-1}p_U(sy) - \lambda(y)\right| \mathrm{d}y \text{ with } L_T = \{y \in (0,T]^d : \min(y) \leqslant 1\}.$$

The following proposition provides a refined version of the bound (27) in Remark 3.3 in the paper; the bound (27) itself appears in the course of the proof of the proposition.

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

$$\sup_{B\in\mathcal{B}(E_c)} |tP_V(tB) - \mu(B)| \leqslant \tfrac{1}{c}\mathcal{D}_{ct}(\tfrac{1}{ct}) + \mu\left(\left\{x \in (0,\infty)^d : \max(x) \geqslant c, \min(x) \leqslant 1/t\right\}\right).$$

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

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

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

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

which is (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

$$\sup_{B\in\mathcal{B}(E_c)} |t\,P_V(tB) - \mu(B)| = \int_{\substack{0<\min(y)\leqslant 1/c\\ \max(y)\leqslant 1/u}} \left|u^{d-1}p_U(uy) - \lambda(y)\right|\mathrm{d}y + \int_{\substack{0<\min(y)\leqslant 1/c\\ \max(y)>1/u}} \lambda(y)\,\mathrm{d}y.$$

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

$$\Lambda\left(\left\{y\in(0,\infty)^d : \min(y)\leqslant 1/c,\ \max(y) > 1/u\right\}\right) = \mu\left(\left\{x\in(0,\infty)^d : \max(x)\geqslant c,\ \min(x) < u\right\}\right).$$

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

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

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

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

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

Example 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

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

for $x \geqslant 0$ and $f(x) = 0$ otherwise. We will show that the bound in Proposition 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

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

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

$$\mathcal{D}_{1/s}(s) = \int_{L_{1/s}} \left(s^{d-1} p(sx) - \lambda(x)\right)_+ \mathrm{d}x \tag{55}$$

$$+ \int_{L_{1/s}} \left(\lambda(x) - s^{d-1} p(sx)\right)_+ \mathrm{d}x. \tag{56}$$

We start by studying the integral (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 $x_i \leqslant 1$ for every $i \in \{1, \ldots, d\}$, we may apply (57) in Lemma D.1 below to get

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

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

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

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

$$= \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{\left(\frac{1}{x_1^2}+\frac{\pi^2s^2}{6}\right)\cdots\left(\frac{1}{x_d^2}+\frac{\pi^2s^2}{6}\right)}{\left(\frac{1}{x_1^2}+\cdots+\frac{1}{x_d^2}-\frac{\pi^2s^2}{4}(\frac{2d}{3}-1)\right)^{\frac{1+d}{2}}}.$$

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

Now observe that

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

with

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

Since we are interested in the limit $s \to 0$, we may assume that $s^2 \leqslant \frac{2}{\pi^2(\frac{2d}{3}-1)}$ so that we may apply the upper bound in (60) in Lemma D.2 below. We get

$$s^{d-1}p(sx) \leqslant \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{\left(\frac{1}{x_1^2}+\frac{\pi^2s^2}{6}\right)\cdots\left(\frac{1}{x_d^2}+\frac{\pi^2s^2}{6}\right)}{\left(\frac{1}{x_1^2}+\cdots+\frac{1}{x_d^2}-\frac{\pi^2s^2}{4}(\frac{2d}{3}-1)\right)^{\frac{1+d}{2}}}$$

$$\leqslant \frac{2^{d-1}\Gamma(\frac{1+d}{2})}{\pi^{\frac{d-1}{2}}}\frac{\left(x_1^{-2}+\frac{\pi^2s^2}{6}\right)\cdots\left(x_d^{-2}+\frac{\pi^2s^2}{6}\right)}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{1+d}{2}}}\cdot\left(1+s^2\frac{2\left(2^{\frac{1+d}{2}}-1\right)\frac{\pi^2}{4}(\frac{2d}{3}-1)}{x_1^{-2}+\cdots+x_d^{-2}}\right).$$

Defining $C_1, C_2 > 0$ by

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

we thus have the bound

$$(s^{d-1}p(sx)-\lambda(x))_+ \leqslant s^2C_1C_2\frac{x_1^{-2}\cdots x_d^{-2}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{3+d}{2}}}+C_1\frac{\sum_{k=0}^{d-1}\sum_{I\subset\{1,\dots,d\},|I|=k}x_I^{-2}(\frac{\pi^2s^2}{6})^{d-k}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{1+d}{2}}}$$
$$+s^2C_1C_2\frac{\sum_{k=0}^{d-1}\sum_{I\subset\{1,\dots,d\},|I|=k}x_I^{-2}(\frac{\pi^2s^2}{6})^{d-k}}{\left(x_1^{-2}+\cdots x_d^{-2}\right)^{\frac{3+d}{2}}},$$

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

$$L_{1/s}=\bigcup_{\emptyset\neq J\subset\{1,\dots,d\}}B_{J,1/s}\text{ where }B_{J,1/s}=\left\{x\in(0,1/s]^d:\forall j\in J,x_j\leqslant 1;\forall j\in J^c:x_j>1\right\}.$$

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

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

If $j=d$, the integral is finite and hence the contribution to the bias is at least of the order $O(s^2)$ as $|I^c|\geqslant 1$. Thus, we shall assume $j<d$. Observe that for every $i\in I^c$, if $i\leqslant j$, then we may upper bound $y_i^{-1}\leqslant 1$. 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

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

which is of order $O(s^{|I^c|})$ as the integral over $(y_1,\dots,y_j)$ is finite. Since $|I^c|\geqslant 1$, we find that the order of the first integral in the bias decomposition (55) is of the order $O(s)$.

We now claim that second term in the decomposition (56) is always equal to zero for sufficiently small $s > 0$ and therefore conclude that $\mathcal{D}_{1/s}(s) = O(s)$ as $s \to 0$. Previous computations and bounds (57) show that for every x such that $x_i \leqslant 1$ for $i \in \{1, \dots, d\}$,

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

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

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

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 $s^2 \leqslant (2\pi^2(d/3-1))^{-1}$ and apply the lower bound in (60) in Lemma D.2 below to get that for every $x \in L_{1/s}$,

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

This shows in particular that $s^{d-1}p(sx) - \lambda(x) \geqslant 0$ for every value of $x \in L_{1/s}$ and $s \leqslant (2\pi^2(d/3-1))^{-1/2}$ 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 D.1, we note that since,

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

the density associated to Φ_2 is constant, say φ, so that

$$\mu\left(\left\{x \in (0,\infty)^d : \max(x) \geqslant c, \min(x) \leqslant 1/t\right\}\right) \leqslant c^{-1}\Phi_2\left(\{\theta \in \mathbb{S}_2 : \min(\theta) \leqslant \tfrac{1}{ct}\}\right)$$

$$= c^{-1}\varphi \operatorname{leb}_{d-1}\left(\{\theta \in \mathbb{S}_2 : \min(\theta) \leqslant \tfrac{1}{ct}\}\right) = O(1/t),$$

when $t \to \infty$. This concludes the analysis of the bias term for the multivariate Cauchy distribution.

Lemma D.1. *We have*

$$\frac{1}{6} \leqslant \frac{1}{1-\cos(t)} - \frac{2}{t^2} \leqslant \frac{1}{3} \qquad \text{for } 0 < t \leqslant \pi. \tag{57}$$

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

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

which gives the inequalities

$$\cos t \leqslant 1 - \frac{3t^2}{6+t^2} = \frac{6-2t^2}{6+t^2} \tag{58}$$

$$\cos t \geqslant 1 - \frac{6t^2}{12+t^2} = \frac{12-5t^2}{12+t^2}. \tag{59}$$

We start by showing (58). Taylor's theorem implies that

$$\cos t \leqslant 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \frac{t^8}{8!}.$$

Consequently,

$$\begin{aligned}
(6+t^2)\cos t &\leqslant (6+t^2)\cdot\left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \frac{t^8}{8!}\right) \\
&= 6 - 3t^2 + \frac{t^4}{4} - \frac{t^6}{120} + \frac{t^8}{6720} + t^2 - \frac{t^4}{2} + \frac{t^6}{24} - \frac{t^8}{720} + \frac{t^10}{40320} \\
&= 6 - 2t^2 + \frac{t^4}{40320}\left(t^6 - 50t^4 + 1344t^2 - 10080\right).
\end{aligned}$$

Hence, if we show that the polynomial $R(s) := s^3 - 50s^2 + 1344s - 10080$ is strictly negative on $0 \leqslant s \leqslant \pi^2$, we are done proving (58). It is sufficient to observe that $R(\pi^2) < 0$ and $R'(s) \geqslant 0$ 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 (59), we follow the same path: Taylor's theorem implies that

$$\cos t \geqslant 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \frac{t^6}{6!}.$$

Consequently,

$$\begin{aligned}(12+t^2)\cos t &\geqslant (12+t^2)\cdot\left(1-\frac{t^2}{2}+\frac{t^4}{4!}-\frac{t^6}{6!}\right)\\ &= 12-6t^2+\frac{t^4}{2}-\frac{t^6}{60}+t^2-\frac{t^4}{2}+\frac{t^6}{24}-\frac{t^8}{720}\\ &= 12-5t^2+\frac{t^6}{720}\left(-t^2+18\right).\end{aligned}$$

Since $-t^2+18\geqslant 0$ for $0\leqslant t\leqslant\sqrt{18}$ and $\sqrt{18}>\pi$, we have proven (59). □

Lemma D.2. *For $a\in[0,1/2]$, we have*

$$1+\left(\frac{1+d}{2}\right)a\leqslant(1-a)^{-\frac{1+d}{2}}\leqslant 1+2\left(2^{\frac{1+d}{2}}-1\right)a. \tag{60}$$

Proof. To prove the upper bound, we will use a convexity argument. Note that the function $\phi(a)=(1-a)^{-\frac{1+d}{2}}$ on the left hand side 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 $0\leqslant a\leqslant 1/2$. Since we compute that

$$\phi''(a)=\tfrac{d+1}{2}\cdot\tfrac{d+3}{2}\cdot(1-a)^{-\frac{d+5}{2}},$$

the convexity on the domain of interest is proven. This implies that for any $\lambda\in[0,1]$ and any pair of points $a_1,a_2\in[0,1/2]$,

$$\phi(\lambda a_1+(1-\lambda)a_2)\leqslant\lambda\phi(a_1)+(1-\lambda)\phi(a_2).$$

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

$$L(a)=1+2\left(2^{\frac{1+d}{2}}-1\right)a$$

and corresponds to the upper bound in the statement.

To prove the lower bound, define $\phi(a)=(1-a)^{-(d+1)/2}$. Then, we already showed that it satisfies $\phi''(a)\geqslant 0$ for $a\in[0,1/2]$. Consequently, Taylor's theorem ensures that for every a in the region of interest

$$\phi(a)=\phi(0)+\phi'(0)a+R_0^1\phi(a)\geqslant\phi(0)+\phi'(0)a,$$

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. □

Appendix E: Proofs of examples

Proof of Example 3.1. We first consider Assumption 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 3.1. Item (ii) is satisfied

by definition. For $A_{a,\beta,\tau} \in \mathcal{A}$ and $\varepsilon > 0$, define the inner and outer hulls

$$A_{a,\beta,\tau;-}(\varepsilon) = A_{a,\beta-\sqrt{d}\varepsilon,\tau+\varepsilon},$$
$$A_{a,\beta,\tau;+}(\varepsilon) = A_{a,\beta+\sqrt{d}\varepsilon,\tau-\varepsilon},$$

with the convention that $\mathbb{S}_p^\upsilon = \mathbb{S}_p$ if $\upsilon < 0$ (a case which arises if $\varepsilon > \tau$). We show that item (iii) in Assumption 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 (22).

- For $x \in A_{a,\beta,\tau;-}(\varepsilon)$ and for $y \in \mathbb{S}_p$ such that $\|y - x\|_\infty \leqslant \varepsilon$, we have $y \in A_{a,\beta,\tau}$: indeed,

  $$y_1 \wedge \cdots \wedge y_d \geqslant x_1 \wedge \cdots \wedge x_d - \varepsilon > \tau + \varepsilon - \varepsilon = \tau$$

  as well as (by equivalence of norms (25))

  $$\langle a, y\rangle = \langle a, x\rangle + \langle a, y - x\rangle \leqslant \beta - \sqrt{d}\varepsilon + \|a\|_2\|y - x\|_2 \leqslant \beta - \sqrt{d}\varepsilon + \sqrt{d}\|y - x\|_\infty \leqslant \beta.$$

  This is the first inclusion in (22).
- For $x \in A_{a,\beta,\tau}$ and for $y \in \mathbb{S}_p$ such that $\|y - x\|_\infty \leqslant \varepsilon$, we have

  $$y_1 \wedge \cdots \wedge y_d \geqslant x_1 \wedge \cdots \wedge x_d - \varepsilon > \tau - \varepsilon$$

  so that $y \in \mathbb{S}_p^{\tau-\varepsilon}$ (if $\varepsilon > \tau$ this trivial since $\mathbb{S}_p^\upsilon = \mathbb{S}_p$ for $\upsilon < 0$) together with

  $$\langle a, y\rangle = \langle a, x\rangle + \langle a, y - x\rangle \leqslant \beta + \|a\|_2\|y - x\|_2 \leqslant \beta + \sqrt{d}\|y - x\|_\infty \leqslant \beta + \sqrt{d}\varepsilon.$$

  This implies the second inclusion in (22).

The difference between the inner and outer hulls is

$$A_{a,\beta,\tau;+}(\varepsilon) \setminus A_{a,\beta,\tau;-}(\varepsilon) \subseteq \left\{x \in \mathbb{S}_p :\ \beta - \sqrt{d}\varepsilon < \langle a, x\rangle \leqslant \beta + \sqrt{d}\varepsilon\right\} \cup (\mathbb{S}_p^{\tau-\varepsilon} \setminus \mathbb{S}_p^{\tau+\varepsilon}).$$

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 3.2 is satisfied. The VC-dimension is preserved under bijections. We identify $E = [0,\infty)^d \setminus \{(0,\ldots,0)\}$ 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

$$\Upsilon_A^\sigma(r,h) = \left\{(u,\theta) \in (\tfrac{1}{r},\infty) \times \mathbb{S}_p :\ \theta \in A_\sigma(hu)\right\}$$

for $\sigma \in \{-,+\}$ and $A \in \mathcal{A}$. For $A = A_{a,\beta,\tau}$ we have

$$\theta \in A_-(hu) \iff \left[\theta_1 \wedge \cdots \wedge \theta_d \geqslant \tau + hu \text{ and } \langle a, \theta\rangle \leqslant \beta - \sqrt{d}hu\right].$$

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 [53, 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 E. The argument for the outer hulls is similar. □

Proof of Example 3.2. We first consider the collection of intersections. In Assumption 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 (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 (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 3.2), it suffices to observe that, by definition,

$$\Gamma^{\sigma}_{A_1 \cap A_2}(r,h) = \Gamma^{\sigma}_{A_1}(r,h) \cap \Gamma^{\sigma}_{A_2}(r,h)$$

and thus

$$\{\Gamma^{\sigma}_{A}(r,h) : A \in \mathcal{A}_1 \sqcap \mathcal{A}_2\} = \{\Gamma^{\sigma}_{A_1}(r,h) : A_1 \in \mathcal{A}_1\} \sqcap \{\Gamma^{\sigma}_{A_2}(r,h) : A_2 \in \mathcal{A}_2\}.$$

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 [53, Lemma 2.6.17].

For the collection of unions, the verification of Assumptions 3.1 and 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. □

Appendix 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 (35) and (37) in the paper, define

$$L_t(g,T) = t\,\mathbb{P}[g(T(X)) \neq Y,\ \|T(X)\|_p \geqslant t],$$

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

Then $L_t(g)$ in (35) corresponds to $L_t(g,v)$ here and $\widehat{L}^{\tau}(g)$ in (37) to $\widehat{L}^{\tau}(g,\hat{v})$.

Proof of Lemma 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

$$L_{t,\pm}^{>\tau}(g,T) = t\,\mathbb{P}[g(T(X)) \neq Y,\ Y = \pm 1,\ \theta_p(T(X)) \in \mathbb{S}_p^{\tau},\ \|T(X)\|_p \geqslant t].$$

Consider the cones generated by regions $\mathbb{S}_p^{\pm}(g) = \{x \in \mathbb{S}_p : g(x) = \pm 1\}$, that is, $R_p^{\pm}(g) = \{tx : x \in \mathbb{S}_p^{\pm}(g),\ t \geqslant 1\}$. Equipped with this notation we may write

$$L_{t,+}^{>\tau}(g,T) = t\,\mathbb{P}[T(X) \in R_p^{-}(g),\ Y = +1,\ \theta_p(T(X)) \in \mathbb{S}_p^{\tau},\ \|T(X)\|_p \geqslant t],$$

$$L^{>\tau}_{t,-}(g,T) = t\,\mathbb{P}[T(X) \in R^+_p(g),\, Y=-1,\, \theta_p(T(X)) \in \mathbb{S}^\tau_p,\, \|T(X)\|_p \geqslant t].$$

Setting the standardization function T to v yields $T(X) = V$ and thus

$$\begin{aligned} L^{>\tau}_{t,+}(g,v) &= t\,\mathbb{P}[V \in R^-_p(g),\, \theta_p(V) \in \mathbb{S}^\tau_p,\, \|V\|_p \geqslant t,\, Y=+1] \\ &= \varrho t\,\mathbb{P}[\theta_p(V) \in \mathbb{S}^-_p(g) \cap \mathbb{S}^\tau_p,\, \|V\|_p \geqslant t \mid Y=+1] \\ &\to \varrho\Phi^+_p(\mathbb{S}^-_p(g) \cap \mathbb{S}^\tau_p), \qquad t\to\infty, \end{aligned}$$

where the last convergence occurs because of Assumption 4.1 and the fact that Φ^+_p is dominated by Φ_p, so that $\mathbb{S}^-_p(g) \cap \mathbb{S}^\tau_p$ is a Φ^+_p-continuity set. Proceeding similarly with $L^{>\tau}_{t,-}(g,v)$, the result is obtained. □

The next result parallels Lemma 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,

$$\widehat{L}^\tau_\pm(g,T) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{g(T(X_i)) \neq Y_i,\, Y_i = \pm 1,\, \theta_p(T(X_i)) \in \mathbb{S}^\tau_p,\, \|T(X_i)\|_p \geqslant n/k\}.$$

Setting $T = \hat{v}$ as in (12) yields $T(X_i) = \widehat{V}_i$ and thus

$$\widehat{L}^\tau_+(g,\hat{v}) = \frac{1}{k}\sum_{i=1}^n \mathbb{1}\{\theta_p(\widehat{V}_i) \in \mathbb{S}^-_p(g) \cap \mathbb{S}^\tau_p,\, Y_i = +1,\, \|\widehat{V}_i\|_p \geqslant n/k\}.$$

Recall from (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 immediately:

Lemma 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 4.1. Recall from the proof of Lemma 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 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,

$$\begin{aligned} \left|\widehat{L}^\tau(g,\hat{v}) - L^{>\tau}_\infty(g)\right| &= \left|\widehat{L}^\tau_+(g,\hat{v}) - L^{>\tau}_{\infty,+}(g) + \widehat{L}^\tau_-(g,\hat{v}) - L^{>\tau}_{\infty,-}(g)\right| \\ &\leqslant \left|\widehat{L}^\tau_+(g,\hat{v}) - L^{>\tau}_{\infty,+}(g)\right| + \left|\widehat{L}^\tau_-(g,\hat{v}) - L^{>\tau}_{\infty,-}(g)\right|. \end{aligned} \tag{61}$$

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 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, $Q_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. As in Section 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 $E \times \{-1,1\}$: for $\sigma \in \{-,+\}$ and Borel sets $B \subseteq E$ bounded away from the origin with $\mu_\sigma(\partial B) = 0$, put

$$\nu(B \times \{\sigma 1\}) = \lim_{t\to\infty} t\,\mathbb{P}[V \in tB, Y = \sigma 1] = \mathbb{P}[Y = \sigma 1]\,\mu_\sigma(B),$$

a limit which exists by the conditional regular variation assumption (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 3.1 (see (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 3.1 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 (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 3.1 into

$$\begin{aligned}
\tfrac{k_+}{k}\widehat{\Phi}_p^+(A) - \varrho\Phi_p^+(A) &= \tfrac{n}{k}Q_n(\widehat{\Gamma}_A \times \{+1\}) - \nu(C_A \times \{+1\}) \\
&\leqslant \tfrac{n}{k}Q_n(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - \nu(\Gamma_A^- \times \{+1\}) \\
&\leqslant \tfrac{n}{k}\left|Q_n(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - Q(\tfrac{n}{k}\Gamma_A^+ \times \{+1\})\right| \\
&\quad + \left|\tfrac{n}{k}Q(\tfrac{n}{k}\Gamma_A^+ \times \{+1\}) - \nu(\Gamma_A^+ \times \{+1\})\right| \\
&\quad + \nu\left((\Gamma_A^+ \setminus \Gamma_A^-) \times \{+1\}\right).
\end{aligned}$$

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

$$\begin{aligned}
\left|\tfrac{k_+}{k}\widehat{\Phi}_p^+(A) - \varrho\Phi_p^+(A)\right| \leqslant & \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q_n(\tfrac{n}{k}B \times \{+1\}) - Q(\tfrac{n}{k}B \times \{+1\})\right| \quad \text{(stochastic error II)} \\
& + \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q(\tfrac{n}{k}B \times \{+1\}) - \nu(B \times \{+1\})\right| \quad \text{(bias term II)}
\end{aligned}$$

$$+\,\nu\left((\Gamma_A^+ \setminus \Gamma_A^-) \times \{+1\}\right) \qquad \text{(framing gap II)}.$$

We treat the three terms separately, following closely the proof of Theorem 3.1.

Bias term II. Taking the supremum over $A \in \mathcal{A}$ immediately yields the bias term in the statement of Theorem 4.1.

Stochastic error II. Since $Q(B \times \{+1\}) \leqslant \mathbb{P}[X \in B]$, the Q-probability of sets in the class $\mathcal{F}' = \mathcal{F} \times \{+1\}$ (with $\mathcal{F}$ defined in (46)) is less than $d^{1+1/p}(1+\Delta_2)\frac{k}{n}$, see (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 A.1,

$$\sup_{A\in\mathcal{A}} \max_{B\in\{\Gamma_A^+,\Gamma_A^-\}} \tfrac{n}{k}\left|Q_n(\tfrac{n}{k}B\times\{+1\}) - Q(\tfrac{n}{k}B\times\{+1\})\right| \\ \leqslant \sqrt{\frac{d^{1+1/p}(1+\Delta_2)}{k}}\left(56\sqrt{V_{\mathcal{F}}} + 2\sqrt{\log(2(d+1)/\delta)}\right) + \frac{2}{3k}\log(2(d+1)/\delta),$$

which is the term labelled 'error' in the statement.

Framing gap II. As in the proof of Theorem 3.1, the framing gap in the product space satisfies

$$\begin{aligned}&\nu\big((\Gamma_A^+ \setminus \Gamma_A^-) \times \{+1\}\big)\\ &\leqslant \nu\left(\left\{x \in [0,\infty)^d :\ (1+\Delta_2)^{-1} \leqslant \|x\|_p < (1-\Delta_2)^{-1}\right\} \times \{+1\}\right)\\ &+ \nu\left(\left\{x \in [0,\infty)^d :\ \|x\|_p \geqslant 1,\ \theta_p(x) \in A_+(3\Delta_1\|x\|_p) \setminus A_-(3\Delta_1\|x\|_p)\right\} \times \{+1\}\right).\end{aligned} \tag{62}$$

The first term on the right-hand side of (62) is equal to

$$2\Delta_2\nu(\{x \in [0,\infty)^d : \|x\|_p \geqslant 1\} \times \{+1\}) = 2\Delta_2\varrho\Phi_p^+(\mathbb{S}_p) \leqslant 2\Delta_2\Phi_p(\mathbb{S}_p),$$

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 (62) can be expressed using the polar decomposition of μ_+ (and thus ν), yielding

$$\nu\left(\left\{x \in [0,\infty)^d :\ \|x\|_p \geqslant 1,\ \theta_p(x) \in A_+(3\Delta_1\|x\|_p) \setminus A_-(3\Delta_1\|x\|_p)\right\} \times \{+1\}\right) \\ = \int_1^\infty \varrho\Phi_p^+\big(A_+(3\Delta_1 r) \setminus A_-(3\Delta_1 r)\big)\frac{\mathrm{d}r}{r^2} \leqslant \int_1^\infty \Phi_p\big(A_+(3\Delta_1 r) \setminus A_-(3\Delta_1 r)\big)\frac{\mathrm{d}r}{r^2}.$$

In the proof of Theorem 3.1, it has been shown that the bound in the latter display is less than

$$3c\Delta_1(1 + \log\Phi_p(\mathbb{S}_p) - \log(3c\Delta_1)).$$

We thus obtain the same gap term as in Theorem 3.1.

So far we have only treated one of the two terms of the error decomposition (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 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 4.1 holds true on the intersection $\mathcal{E}_1 \cap \mathcal{E}_2^+ \cap \mathcal{E}_2^-$, which has probability at least $1-\delta$. □

Appendix G: Proofs of simulations experiments

The following lemma describes convenient properties relative to the simulation setting described in Section 5 in the paper. 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 (7) in the paper.

Lemma 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 $\mathbb{E}(\Theta_j) = 1/d$ for $j \in \{1,\dots,d\}$. Let $X = R\Theta$ 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 $F_j(x_j) = \mathbb{P}[X_j \leqslant x_j]$.*

(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*

$$\Phi_p(A) = d\,\mathbb{P}[X \in C_A]$$

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

$$\Phi_p(A) = t\,\mathbb{P}[t^{-1}V \in C_A].$$

Proof of Lemma G.1. *(i)* For any $j \in \{1,\dots,d\}$ and any $x_j > 0$,

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

Since the support of Θ is contained in the unit simplex, we have $\Theta_j/x_j \leqslant 1$ almost surely whenever $x_j \geqslant 1$. In addition $\mathbb{E}(\Theta_j) = 1/d$ by assumption. It follows that

$$1/(1-F_j(x_j)) = \begin{cases} dx_j & \text{if } x_j \geqslant 1, \\ 1/\mathbb{E}\{\min(1,\Theta_j/x_j)\} & \text{if } 0 < x_j < 1. \end{cases}$$

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.

(ii) Let $x \in (0,\infty)^d$ and let $t > 0$ be sufficiently large so that $tx_j \geqslant d$ for all $j \in \{1,\dots,d\}$. Then

$$\begin{aligned} 1 - \mathbb{P}[V_1 \leqslant tx_1,\dots,V_d \leqslant tx_d] &= \mathbb{P}[\exists j = 1,\dots,d : dR\Theta_j > tx_j] \\ &= \mathbb{P}\left[R > (t/d) \min_{j=1,\dots,d}(x_j/\Theta_j)\right] \\ &= (d/t)\,\mathbb{E}\left[\max_{j=1,\dots,d}(\Theta_j/x_j)\right]. \end{aligned}$$

In the last step, we conditioned on Θ and we used the fact that $(t/d)(x_j/\Theta_j) \geqslant 1$ a.s., by the assumptions on t and Θ. Recall $E = [0,\infty)^d \setminus \{(0,\dots,0)\}$. It follows that

$$\forall x \in (0,\infty)^d, \quad \forall t > \frac{d}{\min_{j=1,\dots,d} x_j}, \qquad t\,\mathbb{P}[t^{-1}V \in E \setminus [0,x]] = d\,\mathbb{E}\left[\max_{j=1,\dots,d}(\Theta_j/x_j)\right]. \tag{63}$$

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 2.1 in the paper; see for instance [47, Lemma 6.1]. More precisely, $t\,\mathbb{P}[t^{-1}V \in \cdot\,] \to \mu(\cdot)$ as $t \to \infty$ in the space $\mathcal{M}_0$, where μ is determined by the right-hand side of the previous display via the values of $\mu(E \setminus [0,x])$ for $x \in (0,\infty)^d$.

The angular measure Φ_p determines the exponent measure μ uniquely via the relation (8) in the paper. In that identity, letting f be the indicator function of the set $E \setminus [0,x]$ with $x \in (0,\infty)^d$, it follows by a standard calculation involving Fubini's theorem that

$$\mu(E \setminus [0,x]) = \int_{\mathbb{S}_p} \max_{j=1,\dots,d} (\theta_j/x_j)\, \mathrm{d}\Phi_p(\theta).$$

Define a Borel measure Φ'_p on $\mathbb{S}_p$ by

$$\begin{aligned}\Phi'_p(A) &= d\,\mathbb{P}[X \in C_A] \\ &= d\,\mathbb{P}[R\|\Theta\|_p > 1,\, \Theta/\|\Theta\|_p \in A] \\ &= d\,\mathbb{E}[\|\Theta\|_p \mathbb{1}\{\Theta/\|\Theta\|_p \in A\}]\end{aligned}$$

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 $0 < \|\Theta\|_p \leqslant 1$ almost surely since $p \geqslant 1$. 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

$$\int_{\mathbb{S}_p} g(\theta)\, \mathrm{d}\Phi'_p(\theta) = d\,\mathbb{E}[\|\Theta\|_p\, g(\Theta/\|\Theta\|_p)].$$

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

$$\int_{\mathbb{S}_p} \max_{j=1,\dots,d} (\theta_j/x_j)\, \mathrm{d}\Phi'_p(\theta) = d\,\mathbb{E}\left[\max_{j=1,\dots,d} (\Theta_j/x_j)\right] = \mu(E \setminus [0,x]).$$

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 $C_A = \{x \in E : \|x\|_p \geqslant 1, x/\|x\|_p \in A\}$ is a subset of $(\tau,\infty)^d$. The equality (63) together with the inclusion–exclusion formula and the fact that rectangles form a measure-determining class imply that the measures $t\,\mathbb{P}[t^{-1}V \in \cdot\,]$ and μ correspond on $(\tau,\infty)^d$. It follows that

$$t\,\mathbb{P}[t^{-1}V \in C_A] = \mu(C_A) = \Phi_p(A). \qquad \square$$