JMLR: Workshop and Conference Proceedings vol 40:1–18, 2015

Learning the dependence structure of rare events: a non-asymptotic study

Nicolas Goix GOIX@TELECOM-PARISTECH.FR
Anne Sabourin SABOURIN@TELECOM-PARISTECH.FR
Stéphan Clémençon CLEMENCON@TELECOM-PARISTECH.FR
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

Abstract

Assessing the probability of occurrence of extreme events is a crucial issue in various fields like finance, insurance, telecommunication or environmental sciences. In a multivariate framework, the tail dependence is characterized by the so-called *stable tail dependence function* (STDF). Learning this structure is the keystone of multivariate extremes. Although extensive studies have proved consistency and asymptotic normality for the empirical version of the STDF, non-asymptotic bounds are still missing. The main purpose of this paper is to fill this gap. Taking advantage of adapted VC-type concentration inequalities, upper bounds are derived with expected rate of convergence in $O(k^{-1/2})$. The concentration tools involved in this analysis rely on a more general study of maximal deviations in low probability regions, and thus directly apply to the classification of extreme data.

KEYWORDS: VC theory, multivariate extremes, stable tail dependence function, concentration inequalities, extreme data classification.

1. Introduction

Extreme Value Theory (EVT) develops models for learning the unusual rather than the usual. These models are widely used in fields involving risk management like finance, insurance, telecommunication or environmental sciences. One major application of EVT is to provide a reasonable assessment of the probability of occurrence of rare events. To illustrate this point, suppose we want to manage the risk of a portfolio containing d different assets, $\mathbf{X} = (X_1, \dots, X_d)$. A fairly general purpose is then to evaluate the probability of events of the kind $\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\}$, for large multivariate thresholds $\mathbf{x} = (x_1, \dots, x_d)$. Under not too stringent conditions on the regularity of $\mathbf{X}$'s distribution, EVT shows that for large enough thresholds, (see Section 2 for details)

$$\mathbb{P}\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\} \simeq l(p_1, \dots, p_d),$$

where l is the *stable tail dependence function* and the p_j's are the marginal exceedance probabilities, $p_j = \mathbb{P}(X_j \geq x_j)$. Thus, the functional l characterizes the *dependence* among extremes. The *joint* distribution (over large thresholds) can thus be recovered from the knowledge of the marginal distributions together with the STDF l. In practice, l can be learned from 'moderately extreme' data, typically the k 'largest' ones among a sample of size n, with $k \ll n$. Recovering the p_j's can be done following a well paved way: in the univariate case, EVT essentially consists in modeling the distribution of the maxima (*resp.* the upper tail) as a generalized extreme value distribution, namely an element of the Gumbel, Fréchet or Weibull parametric families (*resp.* by a generalized Pareto distribution).

In contrast, in the multivariate case, there is no finite-dimensional parametrization of the dependence structure. The latter is characterized by the so-called *stable tail dependence function* (STDF). Estimating this functional is thus one of the main issues in multivariate EVT. Asymptotic properties of the empirical STDF have been widely studied, see Huang (1992), Drees and Huang (1998), Embrechts et al. (2000) and de Haan and Ferreira (2006) for the bivariate case, and Qi (1997), Einmahl et al. (2012) for the general multivariate case under smoothness assumptions.

However, to the best of our knowledge, no bounds exist on the finite sample error. It is precisely the purpose of this paper to derive such non-asymptotic bounds. Our results do not require any assumption other than the existence of the STDF. The main idea is as follows. The empirical estimator is based on the empirical measure of 'extreme' regions, which are hit only with low probability. It is thus enough to bound maximal deviations on such low probability regions. The key consists in choosing an adaptive VC class, which only covers the latter regions, and on the other hand, to derive VC-type inequalities that incorporate p, the probability of hitting the class at all.

The structure of the paper is as follows. The whys and wherefores of EVT and the STDF are explained in Section 2. In Section 3, concentration tools which rely on the general study of maximal deviations in low probability regions are introduced, with an immediate application to the framework of classification (Remark 5). The main result of the paper, a non-asymptotic bound on the convergence of the empirical STDF, is derived in Section 4. Section 5 concludes.

2. Background in extreme value theory

A useful setting to understand the use of EVT and to give intuition about the STDF concept is that of risk monitoring. In the univariate case, it is natural to consider the $(1-p)^{th}$ quantile of the distribution F of a random variable X, for a given exceedance probability p, that is $x_p = \inf\{x \in \mathbb{R},\ \mathbb{P}(X > x) \leq p\}$. For moderate values of p, a natural empirical estimate is $x_{p,n} = \inf\{x \in \mathbb{R},\ 1/n \sum_{i=1}^n \mathbb{1}_{X_i > x} \leq p\}$. However, if p is very small, the finite sample $X_1, \ldots, X_n$ contains insufficient information and $x_{p,n}$ becomes irrelevant. That is where EVT comes into play by providing parametric estimates of large quantiles: whereas statistical inference often involves sample means and the central limit theorem, EVT handles phenomena whose behavior is not ruled by an 'averaging effect'. The focus is on the sample maximum rather than the mean. The primal assumption is the existence of two sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, the a_n's being positive, and a non-degenerate distribution function G such that

$$\lim_{n\to\infty} n\,\mathbb{P}\left(\frac{X - b_n}{a_n}\ \geq\ x\right) = -\log G(x) \tag{1}$$

for all continuity points $x \in \mathbb{R}$ of G. If this assumption is fulfilled – it is the case for most textbook distributions – then F is said to be in the *domain of attraction* of G, denoted $F \in DA(G)$. The tail behavior of F is then essentially characterized by G, which is proved to be – up to rescaling – of the type $G(x) = \exp(-(1+\gamma x)^{-1/\gamma})$ for $1+\gamma x > 0$, $\gamma \in \mathbb{R}$, setting by convention $(1+\gamma x)^{-1/\gamma} = e^{-x}$ for $\gamma = 0$. The sign of γ controls the shape of the tail and various estimators of the rescaling sequence as well as γ have been studied in great detail, see *e.g.* Dekkers et al. (1989), Einmahl et al. (2009), Hill (1975), Smith (1987), Beirlant et al. (1996).

In the multivariate case, it is mathematically very convenient to decompose the joint distribution of $\mathbf{X} = (X^1, \ldots, X^d)$ into the margins on the one hand, and the dependence structure on the other hand. In particular, handling uniform margins is very helpful when it comes to establishing upper bounds on the deviations between empirical and mean measures. Define thus standardized variables $U^j = 1 - F_j(X^j)$, where F_j is the marginal distribution function of X^j, and $\mathbf{U} = (U^1, \ldots, U^d)$. Knowledge of the F_j's and of the joint distribution of $\mathbf{U}$ allows to recover that of $\mathbf{X}$, since $\mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = \mathbb{P}(U^1 \geq 1 - F_1(x_1), \ldots, U^d \geq 1 - F_d(x_d))$. With these notations, under a fairly general assumption similar to (1) (namely, standard multivariate regular variation of standardized variables, see *e.g.* Resnick (2007), chap. 6), there exists a limit measure Λ on $[0,\infty]^d \setminus \{\infty\}$ (called the *exponent measure*) such that

$$\lim_{t \to 0} t^{-1} \mathbb{P}\left[U^1 \leq t\, x_1 \text{ or } \ldots \text{ or } U^d \leq t\, x_d\right] = \Lambda[\mathbf{x}, \infty]^c := l(\mathbf{x})\,. \qquad (x_j \in [0,\infty], \mathbf{x} \neq \infty) \quad (2)$$

Notice that no assumption is made about the marginal distributions, so that our framework allows non-standard regular variation, or even no regular variation at all of the original data $\mathbf{X}$ (for more details see *e.g.* Resnick (2007), th. 6.5 or Resnick (1987), prop. 5.10.). The functional l in the limit in (2) is called the *stable tail dependence function*. In the remainder of this paper, the only assumption is the existence of a limit in (2), *i.e.*, the existence of the STDF.

We emphasize that the knowledge of both l and the margins gives access to the probability of hitting 'extreme' regions of the kind $[\mathbf{0}, \mathbf{x}]^c$, for 'large' thresholds $\mathbf{x} = (x_1, \ldots, x_d)$ (*i.e.* such that for some $j \leq d$, $1 - F_j(x_j)$ is a $O(t)$ for some small t). Indeed, in such a case,

$$\begin{aligned}
\mathbb{P}(X^1 > x_1 \text{ or } \ldots \text{ or } X^d > x_d) &= \mathbb{P}\left(\bigcup_{j=1}^{d} (1 - F_j)(X^j) \leq (1 - F_j)(x_j)\right) \\
&= t\left\{\frac{1}{t}\mathbb{P}\left(\bigcup_{j=1}^{d} U^j \leq t\left[\frac{(1 - F_j)(x_j)}{t}\right]\right)\right\} \\
&\underset{t \to 0}{\sim} \; t\, l\Big(t^{-1}\,(1 - F_1)(x_1),\ \ldots,\ t^{-1}\,(1 - F_d)(x_d)\Big) \\
&= \; l\Big((1 - F_1)(x_1),\ \ldots,\ (1 - F_d)(x_d)\Big)
\end{aligned}$$

where the last equality follows from the homogeneity of l. This underlines the utmost importance of estimating the STDF and by extension stating non-asymptotic bounds on this convergence.

Any stable tail dependence function $l(.)$ is in fact a norm, (see Falk et al. (1994), p179) and satisfies

$$\max\{x_1, \ldots, x_n\} \;\leq\; l(\mathbf{x}) \;\leq\; x_1 + \ldots + x_d,$$

where the lower bound is attained if $\mathbf{X}$ is perfectly tail dependent (extremes of univariate marginals always occur simultaneously), and the upper bound in case of tail independence or asymptotic independence (extremes of univariate marginals never occur simultaneously). We refer to Falk et al. (1994) for more details and properties on the STDF.

3. A VC-type inequality adapted to the study of low probability regions

Classical VC inequalities aim at bounding the deviation of empirical from theoretical quantities on relatively simple classes of sets, called VC classes. These classes typically cover the support of the underlying distribution. However, when dealing with rare events, it is of great interest to have such bounds on a class of sets which only covers a small probability region and thus contains (very) few observations. This yields sharper bounds, since only differences between very small quantities are involved. The starting point of this analysis is the following VC-inequality stated below.

Theorem 1 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ $i.i.d.$ realizations of a* r.v. $\mathbf{X}$, *a VC-class $\mathcal{A}$ with VC-dimension $V_\mathcal{A}$ and shattering coefficient (or growth function) $S_\mathcal{A}(n)$. Consider the class union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and let $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Then there is an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}\big[\mathbf{X} \in A\big] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;\leq\; C\left[\sqrt{p}\sqrt{\frac{V_\mathcal{A}}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right]. \tag{3}$$

Proof (sketch of) Details of the proof are deferred to the appendix section. We use a Bernstein-type concentration inequality (McDiarmid (1998)) that we apply to the general functional

$$f(\mathbf{X}_{1:n}) = \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right|,$$

where $\mathbf{X}_{1:n}$ denotes the sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. The inequality in McDiarmid (1998) involves the variance of the *r.v.* $f(\mathbf{X}_1, \ldots, \mathbf{X}_k, x_{k+1}, \ldots, x_n) - f(\mathbf{X}_1, \ldots, \mathbf{X}_{k-1}, x_k, \ldots, x_n)$, which can easily be bounded in our setting. We obtain

$$\mathbb{P}\left[f(\mathbf{X}_{1:n}) - \mathbb{E}f(\mathbf{X}_{1:n}) \;\geq\; t\right] \;\leq\; e^{-\frac{nt^2}{2q + \frac{2t}{3}}}, \tag{4}$$

where the quantity $q \;=\; \mathbb{E}\left(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|\right)$ (with $\mathbf{X}'$ an independent copy of $\mathbf{X}$) is a measure of the complexity of the class $\mathcal{A}$ with respect to the distribution of $\mathbf{X}$. It leads to high probability bounds on $f(\mathbf{X}_{1:n})$ of the form $\mathbb{E}f(\mathbf{X}_{1:n}) + \frac{1}{n}\log(1/\delta) + \sqrt{\frac{2q}{n}\log(1/\delta)}$ instead of the standard Hoeffding-type bound $\mathbb{E}f(\mathbf{X}_{1:n}) + \sqrt{\frac{1}{n}\log(1/\delta)}$. It is then easy to see that $q \leq 2\sup_{A \in \mathcal{A}} \mathbb{P}(\mathbf{X} \in A) \leq 2p$. Finally, an upper bound on $\mathbb{E}f(\mathbf{X}_{1:n})$ is obtained by introducing re-normalized Rademacher averages

$$\mathcal{R}_{n,p} = \mathbb{E}\sup_{A \in \mathcal{A}} \frac{1}{np}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i \in A}\right|.$$

which are then proved to be of order $O(\sqrt{\frac{V_\mathcal{A}}{pn}})$, so that $\mathbb{E}(f(\mathbf{X}_{1:n})) \leq C\sqrt{\frac{V_\mathcal{A}}{pn}}$. ■

Remark 2 *(*COMPARISON WITH EXISTING BOUNDS*) The following re-normalized VC-inequality due to Vapnik and Chervonenkis (see Vapnik and Chervonenkis (1974), Anthony and Shawe-Taylor (1993) or Bousquet et al. (2004), Thm 7),*

$$\sup_{A \in \mathcal{A}} \left| \frac{\mathbb{P}(\mathbf{X} \in A) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A}}{\sqrt{\mathbb{P}(\mathbf{X} \in A)}} \right| \;\leq\; 2\sqrt{\frac{\log S_\mathcal{A}(2n) + \log\frac{4}{\delta}}{n}}, \tag{5}$$

which holds under the same conditions as Theorem 1, allows to derive a bound similar to (3), but with an additional $\log n$ *factor. Indeed, it is known as Sauer's Lemma (see Bousquet et al. (2004)-lemma 1 for instance) that for* $n \geq V_{\mathcal{A}}$, $S_{\mathcal{A}}(n) \leq (\frac{en}{V_{\mathcal{A}}})^{V_{\mathcal{A}}}$*. It is then easy to see from (5) that:*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq 2 \sqrt{\sup_{A \in \mathcal{A}} \mathbb{P}(\mathbf{X} \in A)} \sqrt{\frac{V_{\mathcal{A}} \log \frac{2en}{V_{\mathcal{A}}} + \log \frac{4}{\delta}}{n}} .$$

Introduce the union $\mathbb{A}$ *of all sets in the considered VC class,* $\mathbb{A} = \cup_{A \in \mathcal{A}} A$*, and let* $p = \mathbb{P}\left(\mathbf{X} \in \mathbb{A}\right)$*. Then, the previous bound immediately yields*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq 2\sqrt{p} \sqrt{\frac{V_{\mathcal{A}} \log \frac{2en}{V_{\mathcal{A}}} + \log \frac{4}{\delta}}{n}} .$$

Remark 3 *(SIMPLER BOUND) If we assume furthermore that* $\delta \geq e^{-np}$*, then we have:*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq C\sqrt{p} \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} .$$

Remark 4 *(INTERPRETATION) Inequality (3) can be seen as an interpolation between the best case (small* p*) where the rate of convergence is* $O(1/n)$*, and the worst case (large* p*) where the rate is* $O(1/\sqrt{n})$*. An alternative interpretation is as follows: divide both sides of (3) by* p*, so that the left hand side becomes a supremum of conditional probabilities upon belonging to the union class* $\mathbb{A}$, $\{\mathbb{P}(\mathbf{X} \in A \big| \mathbf{X} \in \mathbb{A})\}_{A \in \mathbb{A}}$*. Then the upper bound is proportional to* $\epsilon(np, \delta)$ *where* $\epsilon(n, \delta) := \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta}$ *is a classical VC-bound;* np *is in fact the expected number of observations involved in (3), and can thus be viewed as the effective sample size.*

Remark 5 *(CLASSIFICATION OF EXTREMES) A key issue in the prediction framework is to find upper bounds for the maximal deviation* $\sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$*, where* $L(g) = \mathbb{P}(g(\mathbf{X}) \neq Y)$ *is the risk of the classifier* $g : \mathcal{X} \to \{-1, 1\}$*, associated with the* r.v. $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{-1, 1\}$. $L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(\mathbf{X}_i) \neq Y_i\}$ *is the empirical risk based on a training dataset* $\{(\mathbf{X}_1, Y_1),\ \ldots,\ (\mathbf{X}_n, Y_n)\}$*. Strong upper bounds on* $\sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$ *ensure the accuracy of the empirical risk minimizer* $g_n := \operatorname{argmin}_{g \in \mathcal{G}} L_n(g)$*.*

In a wide variety of applications (e.g. Finance, Insurance, Networks), it is of crucial importance to predict the system response Y *when the input variable* $\mathbf{X}$ *takes extreme values, corresponding to shocks on the underlying mechanism. In such a case, the risk of a prediction rule* $g(\mathbf{X})$ *should be defined by integrating the loss function* $L(g)$ *with respect to the conditional joint distribution of the pair* $(\mathbf{X}, Y)$ *given* $\mathbf{X}$ *is extreme. For instance, consider the event* $\{\|\mathbf{X}\| \geq t_\alpha\}$ *where* t_α *is the* $(1-\alpha)^{th}$ *quantile of* $\|\mathbf{X}\|$ *for a small* α*. To investigate the accuracy of a classifier* g *given* $\{\|\mathbf{X}\| \geq t_\alpha\}$*, introduce*

$$L_\alpha(g) : = \frac{1}{\alpha} \mathbb{P}\left(Y \neq g(\mathbf{X}),\ \|\mathbf{X}\| > t_\alpha\right) = \mathbb{P}\left(Y \neq g(\mathbf{X}) \;\middle|\; \|\mathbf{X}\| \geq t_\alpha\right) ,$$

and its empirical counterpart

$$L_{\alpha,n}(g) : \;\; = \frac{1}{n\alpha}\sum_{i=1}^{n}\mathbb{I}_{\{Y_i \neq g(\mathbf{X}_i),\ \|\mathbf{X}_i\| > \|\mathbf{X}_{(\lfloor n\alpha \rfloor)}\|\}} \ ,$$

where $\|\mathbf{X}_{(1)}\| \geq \ldots \geq \|\mathbf{X}_{(n)}\|$ *are the order statistics of* $\|\mathbf{X}\|$*. Then as an application of Theorem 1 with* $\mathcal{A} = \{(\mathbf{x}, y), g(\mathbf{x}) \neq y, \|\mathbf{x}\| > t_\alpha\},\ g \in \mathcal{G}$*, we have :*

$$\sup_{g\in\mathcal{G}}\left|\widehat{L}_{\alpha,n}(g) - L_\alpha(g)\right| \leq C\left[\sqrt{\frac{V_{\mathcal{G}}}{n\alpha}\log\frac{1}{\delta}} + \frac{1}{n\alpha}\log\frac{1}{\delta}\right] . \tag{6}$$

We refer to the appendix for more details. Again the obtained rate by empirical risk minimization meets our expectations (see remark 4), insofar as α *is the fraction of the dataset involved in the empirical risk* $L_{\alpha,n}$*. We point out that* α *may typically depend on* n*,* $\alpha = \alpha_n \to 0$*. In this context a direct use of the standard version of the* VC *inequality would lead to a rate of order* $1/(\alpha_n\sqrt{n})$*, which may not vanish as* $n \to +\infty$ *and even go to infinity if* α_n *decays to* 0 *faster than* $1/\sqrt{n}$ *.*

Let us point out that rare events may be chosen more general than $\{\|\mathbf{X}\| > t_\alpha\}$*, say* $\{\mathbf{X} \in Q\}$ *with unknown probability* $q = \mathbb{P}(\{\mathbf{X} \in Q\})$*. The previous result still applies with* $\widetilde{L}_Q(g) := \mathbb{P}\left(Y \neq g(\mathbf{X}), \mathbf{X} \in Q\right)$ *and* $\widetilde{L}_{Q,n}(g) := \mathbb{P}_n\left(Y \neq g(\mathbf{X}), \mathbf{X} \in Q\right)$*; then the obtained upper bound on* $\sup_{g\in\mathcal{G}}\frac{1}{q}\left|\widetilde{L}_Q(g) - \widetilde{L}_{Q,n}(g)\right|$ *is of order* $O(1/\sqrt{qn})$*.*

Similar results can be established for the problem of distribution-free regression, when the error of any predictive rule $f(\mathbf{x})$ *is measured by the conditional mean squared error* $\mathbb{E}[(Z - f(\mathbf{X}))^2 \mid Z > q_{\alpha_n}]$*, denoting by* Z *the real-valued output variable to be predicted from* $\mathbf{X}$ *and by* q_α *its quantile at level* $1 - \alpha$*.*

4. A bound on the STDF

Let us place ourselves in the multivariate extreme framework introduced in Section 1: Consider a random variable $\mathbf{X} = (X^1, \ldots X^d)$ in $\mathbb{R}^d$ with distribution function F and marginal distribution functions $F_1, \ldots, F_d$. Let $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ be an $i.i.d.$ sample distributed as $\mathbf{X}$. In the subsequent analysis, the only assumption is the existence of the STDF defined in (2) and the margins F_j are supposed to be unknown. The definition of l may be recast as

$$l(\mathbf{x}) := \lim_{t\to 0} t^{-1}\tilde{F}(t\mathbf{x}) \tag{7}$$

with $\tilde{F}(\mathbf{x}) = (1 - F)\big((1 - F_1)^{\leftarrow}(x_1), \ldots, (1 - F_d)^{\leftarrow}(x_d)\big)$. Here the notation $(1 - F_j)^{\leftarrow}(x_j)$ denotes the quantity $\sup\{y \,:\, 1 - F_j(y) \geq x_j\}$. Notice that, in terms of standardized variables U^j, $\tilde{F}(\mathbf{x}) = \mathbb{P}\Big(\bigcup_{j=1}^d \{U^j \leq x_j\}\Big) = \mathbb{P}(\mathbf{U} \in [\mathbf{x}, \infty[^c)$.

Let $k = k(n)$ be a sequence of positive integers such that $k \to \infty$ and $k = o(n)$ as $n \to \infty$. A natural estimator of l is its empirical version defined as follows, see Huang (1992), Qi (1997), Drees and Huang (1998), Einmahl et al. (2006):

$$l_n(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n}\mathbb{1}_{\{X_i^1 \geq X^1_{(n-\lfloor kx_1\rfloor+1)} \text{ or } \ldots \text{ or } X_i^d \geq X^d_{(n-\lfloor kx_d\rfloor+1)}\}} \ , \tag{8}$$

The expression is indeed suggested by the definition of l in (7), with all distribution functions and univariate quantiles replaced by their empirical counterparts, and with t replaced by k/n. Extensive studies have proved consistency and asymptotic normality of this nonparametric estimator of l, see Huang (1992), Drees and Huang (1998) and de Haan and Ferreira (2006) for the asymptotic normality in dimension 2, Qi (1997) for consistency in arbitrary dimension, and Einmahl et al. (2012) for asymptotic normality in arbitrary dimension under differentiability conditions on l.

To our best knowledge, there is no established non-asymptotic bound on the maximal deviation $\sup_{0\le\mathbf{x}\le T}|l_n(\mathbf{x}) - l(\mathbf{x})|$. It is the purpose of the remainder of this section to derive such a bound, without any smoothness condition on l.

First, Theorem 1 needs adaptation to a particular setting: introduce a random vector $\mathbf{Z} = (Z^1, \dots, Z^d)$ with uniform margins, *i.e.*, for every $j = 1, \dots, d$, the variable Z^j is uniform on $[0, 1]$. Consider the class

$$\mathcal{A} = \left\{ \left[\frac{k}{n}\mathbf{x}, \infty\right[^{\,c} \;:\quad \mathbf{x}\in\mathbb{R}_+^d, \quad 0\le x_j\le T\ (1\le j\le d) \right\}$$

This is a VC-class of VC-dimension d, as proved in Devroye et al. (1996), Theorem 13.8, for its complementary class $\big\{[\mathbf{x}, \infty[,\ \mathbf{x} > 0\big\}$. In this context, the union class $\mathbb{A}$ has mass $p \le dT\frac{k}{n}$ since

$$\mathbb{P}(\mathbf{Z}\in\mathbb{A}) = \mathbb{P}\left[\mathbf{Z}\in\left(\left[\frac{k}{n}T,\infty\right[^{d}\right)^{c}\right] = \mathbb{P}\left[\bigcup_{j=1..d}\mathbf{Z}^j < \frac{k}{n}T\right] \le \sum_{j=1}^{d}\mathbb{P}\left[\mathbf{Z}^j < \frac{k}{n}T\right]$$

Consider the measures $C_n(\,\cdot\,) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i\in\,\cdot\,\}}$ and $C(\mathbf{x}) = \mathbb{P}(Z\in\,\cdot\,)$. As a direct consequence of Theorem 1 the following inequality holds true with probability at least $1-\delta$,

$$\sup_{0\le\mathbf{x}\le T}\frac{n}{k}\left|C_n(\frac{k}{n}[\mathbf{x},\infty[^c) - C(\frac{k}{n}[\mathbf{x},\infty[^c)\right| \;\le\; Cd\left(\sqrt{\frac{T}{k}\log\frac{1}{\delta}} \;+\; \frac{1}{k}\log\frac{1}{\delta}\right)\,.$$

If we assume furthermore that $\delta\ge e^{-k}$, then we have

$$\sup_{0\le\mathbf{x}\le T}\frac{n}{k}\left|C_n(\frac{k}{n}[\mathbf{x},\infty[^c) - C(\frac{k}{n}[\mathbf{x},\infty[^c)\right| \;\le\; Cd\sqrt{\frac{T}{k}\log\frac{1}{\delta}}\,. \tag{9}$$

Inequality (9) is the cornerstone of the following theorem, which is the main result of the paper. In the sequel, we consider a sequence $k(n)$ of integers such that $k = o(n)$ and $k(n)\to\infty$. For notational convenience, we often drop the dependence in n and simply write k instead of $k(n)$.

Theorem 6 *Let T be a positive number such that $T\ge\frac{7}{2}(\frac{\log d}{k}+1)$, and δ such that $\delta\ge e^{-k}$. Then there is an absolute constant C such that for each $n>0$, with probability at least $1-\delta$:*

$$\sup_{0\le\mathbf{x}\le T}|l_n(\mathbf{x}) - l(\mathbf{x})| \;\le\; Cd\sqrt{\frac{T}{k}\log\frac{d+3}{\delta}} \;+\; \sup_{0\le\mathbf{x}\le 2T}\left|\frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) - l(\mathbf{x})\right| \tag{10}$$

The second term on the right hand side of (10) is a bias term which depends on the discrepancy between the left hand side and the limit in (2) or (7) at level $t = k/n$. The value k can be interpreted as the effective number of observations used in the empirical estimate, *i.e.* the effective sample size for tail estimation. Considering classical inequalities in empirical process theory such as VC-bounds, it is thus no surprise to obtain one in $O(1/\sqrt{k})$. Too large values of k tend to yield a large bias, whereas too small values of k yield a large variance. For a more detailed discussion on the choice of k we recommend Einmahl et al. (2009).

The proof of Theorem 6 follows the same lines as in Qi (1997). For unidimensional random variables $Y_1, \dots, Y_n$, let us denote by $Y_{(1)} \leq \ldots \leq Y_{(n)}$ their order statistics. Define then the empirical version $\tilde{F}_n$ of $\tilde{F}$ (introduced in (7)) as

$$\tilde{F}_n(\mathbf{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^1 \leq x_1 \text{ or } \ldots \text{ or } U_i^d \leq x_d\}} \,,$$

so that $\frac{n}{k}\tilde{F}_n(\frac{k}{n}\mathbf{x}) \;=\; \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^1 \leq \frac{k}{n}x_1 \text{ or } \ldots \text{ or } U_i^d \leq \frac{k}{n}x_d\}}$. Notice that the U_i^j's are not observable (since F_j is unknown). In fact, $\tilde{F}_n$ will be used as a substitute for l_n allowing to handle uniform variables. The following lemmas make this point explicit.

Lemma 7 (Link between l_n and $\tilde{F}_n$) *The empirical version of $\tilde{F}$ and that of l are related* via

$$l_n(\mathbf{x}) \;=\; \frac{n}{k}\tilde{F}_n(U^1_{(\lfloor kx_1 \rfloor)}, \;\ldots\;, U^d_{(\lfloor kx_d \rfloor)}).$$

Proof Consider the definition of l_n in (8), and note that for $j = 1, \ldots, d$,

$$\begin{aligned} X_i^j \geq X^j_{(n-\lfloor kx_i \rfloor+1)} &\Leftrightarrow rank(X_i^j) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow rank(F_j(X_i^j)) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow rank(1 - F_j(X_i^j)) \leq \lfloor kx_j \rfloor \\ &\Leftrightarrow U_i^j \leq U^j_{(\lfloor kx_j \rfloor)}, \end{aligned}$$

so that $l_n(\mathbf{x}) \;=\; \frac{1}{k}\;\sum_{j=1}^{n} \mathbb{1}_{\{U_j^1 \leq U^1_{(\lfloor kx_1 \rfloor)} \text{ or } \ldots \text{ or } U_j^d \leq U^d_{(\lfloor kx_d \rfloor)}\}}$. ■

Lemma 8 (Uniform bound on $\tilde{F}_n$'s deviations) *For any finite $T > 0$, and $\delta \geq e^{-k}$, with probability at least $1 - \delta$, the deviation of $\tilde{F}_n$ from $\tilde{F}$ is uniformly bounded:*

$$\sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k}\tilde{F}_n(\frac{k}{n}\mathbf{x}) - \frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) \right| \leq Cd\sqrt{\frac{T}{k}\log\frac{1}{\delta}}$$

Proof Notice that

$$\sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k}\tilde{F}_n(\frac{k}{n}\mathbf{x}) - \frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) \right| = \frac{n}{k}\left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{U}_i \in \frac{k}{n}]\mathbf{x}, \infty]^c\}} - \mathbb{P}\left[\mathbf{U} \in \frac{k}{n}]\mathbf{x}, \infty]^c\right] \right|,$$

and apply inequality (9). ■

Lemma 9 (Bound on the order statistics of U) *Let $\delta \geq e^{-k}$. For any finite positive number $T > 0$ such that $T \geq 7/2((\log d)/k + 1)$, we have with probability greater than $1 - \delta$,*

$$\forall\, 1 \leq j \leq d, \qquad \frac{n}{k} U^j_{(\lfloor kT \rfloor)} \;\leq\; 2T\,, \tag{11}$$

and with probability greater than $1 - (d+1)\delta$,

$$\max_{1 \leq j \leq d} \; \sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U^j_{(\lfloor kx_j \rfloor)} \right| \;\leq\; C\sqrt{\frac{T}{k} \log \frac{1}{\delta}}\,.$$

Proof Notice that $\sup_{[0,T]} \frac{n}{k} U^j_{(\lfloor k\,\cdot\, \rfloor)} = \frac{n}{k} U^j_{(\lfloor kT \rfloor)}$ and let $\Gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq t\}}$. It then straightforward to see that

$$\frac{n}{k} U^j_{(\lfloor kT \rfloor)} \leq 2T \;\;\Leftrightarrow\;\; \Gamma_n\Big(\frac{k}{n} 2T\Big) \geq \frac{\lfloor kT \rfloor}{n}$$

so that

$$\mathbb{P}\left(\frac{n}{k} U^j_{(\lfloor kT \rfloor)} > 2T\right) \;\leq\; \mathbb{P}\left(\sup_{\frac{2kT}{n} \leq t \leq 1} \frac{t}{\Gamma_n(t)} > 2\right).$$

Using Wellner (1978), Lemma 1-(ii) (we use the fact that, with the notations of this reference, $h(1/2) \geq 1/7$), we obtain

$$\mathbb{P}\left(\frac{n}{k} U^j_{(\lfloor kT \rfloor)} > 2T\right) \leq e^{-\frac{2kT}{7}},$$

and thus

$$\mathbb{P}\left(\exists j,\; \frac{n}{k} U^j_{(\lfloor kT \rfloor)} > 2T\right) \leq d e^{-\frac{2kT}{7}} \leq e^{-k} \leq \delta$$

as required in (11). Yet,

$$\begin{aligned}
\sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U^j_{(\lfloor kx_j \rfloor)} \right| &= \sup_{0 \leq x_j \leq T} \left| \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq U^j_{(\lfloor kx_j \rfloor)}\}} - \frac{n}{k} U^j_{(\lfloor kx_j \rfloor)} \right| \\
&= \frac{n}{k} \sup_{0 \leq x_j \leq T} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq U^j_{(\lfloor kx_j \rfloor)}\}} - \mathbb{P}\left[U_1^j \leq U^j_{(\lfloor kx_j \rfloor)}\right] \right| \\
&= \sup_{0 \leq x_j \leq T} \Theta_j\Big(\frac{n}{k} U^j_{(\lfloor kx_j \rfloor)}\Big),
\end{aligned}$$

where $\Theta_j(y) = \frac{n}{k} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq \frac{k}{n} y\}} - \mathbb{P}\left[U_1^j \leq \frac{k}{n} y\right] \right|$. Then, by (11), with probability greater than $1 - \delta$,

$$\max_{1 \leq j \leq d} \sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U^j_{(\lfloor kx_j \rfloor)} \right| \;\leq\; \max_{1 \leq j \leq d} \sup_{0 \leq y \leq 2T} \Theta_j(y)$$

and from (9), each term $\sup_{0 \leq y \leq 2T} \Theta_j(y)$ is bounded by $C\sqrt{\frac{T}{k} \log \frac{1}{\delta}}$ (with probability $1 - \delta$). In the end, with probability greater than $1 - (d+1)\delta$:

$$\max_{1 \leq j \leq d} \sup_{0 \leq y \leq 2T} \Theta_j(y) \;\leq\; C\sqrt{\frac{T}{k} \log \frac{1}{\delta}}\,,$$

which is the desired inequality ■

We may now proceed with the proof of Theorem 6. First of all, noticing that $\tilde{F}(t\mathbf{x})$ is non-decreasing in x_j for every l and that $l(\mathbf{x})$ is non-decreasing and continuous (thus uniformly continuous on $[0,T]^d$), from (7) it is easy to prove by subdivising $[0,T]^d$ (see Qi (1997) p.174 for details) that

$$\sup_{0\leq \mathbf{x}\leq T}\left|\frac{1}{t}\tilde{F}(t\mathbf{x}) - l(\mathbf{x})\right| \to 0 \quad \text{as} \quad \mathrm{t}\to 0\,. \tag{12}$$

Using Lemma 7, we can write :

$$\begin{aligned}
\sup_{0\leq \mathbf{x}\leq T}|l_n(\mathbf{x})-l(\mathbf{x})| \;&=\; \sup_{0\leq \mathbf{x}\leq T}\left|\frac{n}{k}\tilde{F}_n\left(U^1_{(\lfloor kx_1\rfloor)},\dots,U^d_{(\lfloor kx_d\rfloor)}\right) - l(\mathbf{x})\right| \\
&\leq\; \sup_{0\leq \mathbf{x}\leq T}\left|\frac{n}{k}\tilde{F}_n\left(U^1_{(\lfloor kx_1\rfloor)},\dots,U^d_{(\lfloor kx_d\rfloor)}\right) - \frac{n}{k}\tilde{F}\left(U^1_{(\lfloor kx_1\rfloor)},\dots,U^d_{(\lfloor kx_d\rfloor)}\right)\right| \\
&\quad+ \sup_{0\leq \mathbf{x}\leq T}\left|\frac{n}{k}\tilde{F}\left(U^1_{(\lfloor kx_1\rfloor)},\dots,U^d_{(\lfloor kx_d\rfloor)}\right) - l\left(\frac{n}{k}U^1_{(\lfloor kx_1\rfloor)},\dots,\frac{n}{k}U^d_{(\lfloor kx_d\rfloor)}\right)\right| \\
&\quad+ \sup_{0\leq \mathbf{x}\leq T}\left|l\left(\frac{n}{k}U^1_{(\lfloor kx_1\rfloor)},\dots,\frac{n}{k}U^d_{(\lfloor kx_d\rfloor)}\right) - l(\mathbf{x})\right| \\
&=:\; \Lambda(n) \;+\; \Xi(n) \;+\; \Upsilon(n)\,.
\end{aligned}$$

Now, by (11) we have with probability greater than $1-\delta$:

$$\Lambda(n) \;\leq\; \sup_{0\leq \mathbf{x}\leq 2T}\left|\frac{n}{k}\tilde{F}_n(\frac{k}{n}\mathbf{x}) - \frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x})\right|$$

and by Lemma 8,

$$\Lambda(n) \leq Cd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}}$$

with probability at least $1-2\delta$. Similarly,

$$\Xi(n) \;\leq\; \sup_{0\leq \mathbf{x}\leq 2T}\left|\frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) - \frac{n}{k}l(\frac{k}{n}\mathbf{x})\right| = \sup_{0\leq \mathbf{x}\leq 2T}\left|\frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) - l(\mathbf{x})\right| \;\to\; 0 \quad \text{(bias term)}$$

by virtue of (12). Concerning $\Upsilon(n)$, we have :

$$\begin{aligned}
\Upsilon(n) \;&\leq\; \sup_{0\leq \mathbf{x}\leq T}\left|l\left(\frac{n}{k}U^1_{(\lfloor kx_1\rfloor)},\dots,\frac{n}{k}U^d_{(\lfloor kx_d\rfloor)}\right) - l(\frac{\lfloor kx_1\rfloor}{k},\dots,\frac{\lfloor kx_d\rfloor}{k})\right| \\
&\quad+ \sup_{0\leq \mathbf{x}\leq T}\left|l(\frac{\lfloor kx_1\rfloor}{k},\dots,\frac{\lfloor kx_d\rfloor}{k}) - l(\mathbf{x})\right| \\
&=\; \Upsilon_1(n) \;+\; \Upsilon_2(n)
\end{aligned}$$

Recall that l is 1-Lipschitz on $[0,T]^d$ regarding to the $\|.\|_1$-norm, so that

$$\Upsilon_1(n) \;\leq\; \sup_{0\leq \mathbf{x}\leq T} \sum_{l=1}^{d} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U^j_{(\lfloor kx_j \rfloor)} \right|$$

so that by Lemma 9, with probability greater than $1-(d+1)\delta$:

$$\Upsilon_1(n) \;\leq\; Cd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}}\;.$$

On the other hand, $\Upsilon_2(n) \;\leq\; \sup_{0\leq \mathbf{x}\leq T} \sum_{l=1}^{d} \left| \frac{\lfloor kx_j \rfloor}{k} - x_j \right| \;\leq\; \frac{d}{k}$. Finally we get, for every $n>0$, with probability at least $1-(d+3)\delta$:

$$\begin{aligned}
\sup_{0\leq \mathbf{x}\leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| \;&\leq\; \Lambda(n) + \Upsilon_1(n) + \Upsilon_2(n) + \Xi(n) \\
&\leq\; Cd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}} \;+\; Cd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}} \;+\; \frac{d}{k} \;+\; \sup_{0\leq \mathbf{x}\leq 2T} \left| \tilde{F}(\mathbf{x}) - \frac{n}{k} l(\frac{k}{n}\mathbf{x}) \right| \\
&\leq\; C'd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}} \;+\; \sup_{0\leq \mathbf{x}\leq 2T} \left| \frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) - l(\mathbf{x}) \right|
\end{aligned}$$

5. Discussion

We provide a non-asymptotic bound of VC type controlling the error of the empirical version of the STDF. Our bound achieves the expected rate in $O(k^{-1/2}) + \text{bias}(k)$, where k is the number of (extreme) observations retained in the learning process. In practice the smaller k/n, the smaller the bias. Since no assumption is made on the underlying distribution, other than the existence of the STDF, it is not possible in our framework to control the bias explicitly. One option would be to make an additional hypothesis of 'second order regular variation' (see *e.g.* de Haan and Resnick, 1996). We made the choice of making as few assumptions as possible, however, since the bias term is separated from the 'variance' term, it is probably feasible to refine our result with more assumptions.

For the purpose of controlling the empirical STDF, we have adopted the more general framework of maximal deviations in low probability regions. The VC-type bounds adapted to low probability regions derived in Section 3 may directly be applied to a particular prediction context, namely where the objective is to learn a classifier (or a regressor) that has good properties on low probability regions. This may open the road to the study of classification of extremal observations, with immediate applications to the field of anomaly detection.

References

Martin Anthony and John Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47(3):207 – 217, 1993.

Jan Beirlant, Petra Vynckier, and Jozef L. Teugels. Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667, 1996.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer Berlin Heidelberg, 2004.

L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

Laurens de Haan and Sidney Resnick. Second-order regular variation and rates of convergence in extreme-value theory. *The Annals of Probability*, pages 97–124, 1996.

A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 17(4):1833–1855, 12 1989.

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

Holger Drees and Xin Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *J. Multivar. Anal.*, 64(1):25–47, January 1998.

John H. J. Einmahl, Laurens de Haan, and Deyuan Li. Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *Ann. Statist.*, 34(4): 1987–2014, 08 2006.

John H. J. Einmahl, Jun Li, and Regina Y. Liu. Thresholding events of extreme in simultaneous monitoring of multiple risks. *Journal of the American Statistical Association*, 104(487):982–992, 2009.

John H. J. Einmahl, Andrea Krajina, and Johan Segers. An m-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, 40(3):1764–1793, 06 2012.

Paul Embrechts, Laurens de Haan, and Xin Huang. Modelling multivariate extremes. *Extremes and Integrated Risk Management (Ed. P. Embrechts)*, RISK Books(59-67), 2000.

M. Falk, J. Huesler, and R. D. Reiss. *Laws of Small Numbers: Extremes and Rare Events*. Birkhauser, Boston, 1994.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174, 09 1975.

Xin Huang. Statistics of bivariate extreme values, 1992.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

Colin McDiarmid. Concentration. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 195–248. Springer Berlin Heidelberg, 1998.

Yongcheng Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica*, 13(2):167–175, 1997.

Sidney Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering, 1987.

Sidney Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

Richard L. Smith. Estimating tails of probability distributions. *Ann. Statist.*, 15(3):1174–1207, 09 1987.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

JonA. Wellner. Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 45(1):73–88, 1978.

Appendix A. Proof of Theorem 1

Theorem 1 is actually a short version of Theorem 10 below:

Theorem 10 (Maximal deviations) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ $i.i.d.$ realizations of a* r.v. $\mathbf{X}$ *valued in* $\mathbb{R}^d$*, a VC-class $\mathcal{A}$, and denote by $\mathcal{R}_{n,p}$ the associated relative Rademacher average defined by*

$$\mathcal{R}_{n,p} = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| . \tag{13}$$

Define the union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Fix $0 < \delta < 1$, then with probability at least $1 - \delta$,

$$\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;\leq\; 2\mathcal{R}_{n,p} \;+\; \frac{2}{3np} \log \frac{1}{\delta} \;+\; 2\sqrt{\frac{1}{np} \log \frac{1}{\delta}} \,,$$

and there is a constant C independent of n, p, δ such that with probability greater than $1 - \delta$,

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;\leq\; C \left(\sqrt{p} \sqrt{\frac{V_\mathcal{A}}{n} \log \frac{1}{\delta}} \;+\; \frac{1}{n} \log \frac{1}{\delta} \right) .$$

If we assume furthermore that $\delta \geq e^{-np}$, then we both have:

$$\begin{aligned}
\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;&\leq\; 2\mathcal{R}_{n,p} \;+\; 3\sqrt{\frac{1}{np} \log \frac{1}{\delta}} \\
\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;&\leq\; C\sqrt{\frac{V_\mathcal{A}}{np} \log \frac{1}{\delta}} \,.
\end{aligned}$$

In the following, $\mathbf{X}_{1:n}$ denotes an $i.i.d.$ sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ distributed as $\mathbf{X}$, a $\mathbb{R}^d$-valued random vector. The classical steps to prove VC inequalities consist in applying a concentration inequality to the function

$$f(\mathbf{X}_{1:n}) := \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right|, \tag{14}$$

and then establishing bounds on the expectation $\mathbb{E} f(\mathbf{X}_{1:n})$, using for instance Rademacher average. Here we follow the same lines, but applying a Bernstein type concentration inequality instead of the usual Hoeffding one, since the variance term in the bound involves the probability p to be in the union of the VC-class $\mathcal{A}$ considered. We then introduce relative Rademacher averages instead of the conventional ones, to take into account p for bounding $\mathbb{E} f(\mathbf{X}_{1:n})$.

We need first to control the variability of the random variable $f(\mathbf{X}_{1:n})$ when fixing all but one marginal $\mathbf{X}_i$. For that purpose introduce the functional

$$h(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \mathbb{E}\left[f(\mathbf{X}_{1:n}) | \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_k = \mathbf{x}_k\right] - \mathbb{E}\left[f(\mathbf{X}_{1:n}) | \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_{k-1} = \mathbf{x}_{k-1}\right]$$

The *positive deviation* of $h(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{X}_k)$ is defined by

$$dev^+(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \left\{ h(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}) \right\},$$

and maxdev$^+$, the maximum of all positive deviations, by

$$\text{maxdev}^+ = \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}} \max_k \; dev^+(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}) \,.$$

Finally, define $\hat{v}$, the *maximum sum of variances*, by

$$\hat{v} = \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_n} \sum_{k=1}^{n} \mathbf{Var}\; h(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{X}_k) \,.$$

We have now the tools to state an extension of the classical Bernstein inequality, which is proved in McDiarmid (1998).

Proposition 11 *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ as above, and f any function $(\mathbb{R}^d)^n \to \mathbb{R}$. Let maxdev$^+$ and $\hat{v}$ the maximum sum of variances, both of which we assume to be finite, and let μ be the mean of $f(\mathbf{X}_{1:n})$. Then for any $t \geq 0$,*

$$\mathbb{P}\big[f(\mathbf{X}_{1:n}) - \mu \geq t\big] \;\leq\; \exp\left(- \frac{t^2}{2\hat{v}(1 + \frac{maxdev^+ t}{3\hat{v}})} \right).$$

Note that the term $\frac{\text{maxdev}^+ t}{3\hat{v}}$ is view as an 'error term' and is often negligible. Let us apply this theorem to the specific function f defined in (14). Then the following lemma holds true:

Lemma 12 *In the situation of Proposition 11 with f as in (14), we have*

$$maxdev^+ \leq \frac{1}{n} \;\; and \;\; \hat{v} \leq \frac{q}{n},$$

where

$$q \;=\; \mathbb{E}\left(\sup_{A\in\mathcal{A}} |\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}|\right) \;\leq\; 2\mathbb{E}\left(\sup_{A\in\mathcal{A}} |\mathbb{1}_{\mathbf{X}'\in A}\mathbb{1}_{\mathbf{X}\notin A}|\right), \tag{15}$$

with $\mathbf{X}'$ *an independent copy of* $\mathbf{X}$.

Proof Considering the definition of f, we have:

$$\begin{aligned} h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{x}_k) =\; & \mathbb{E}\sup_{A\in\mathcal{A}}\left|\mathbb{P}(\mathbf{X}\in A) - \frac{1}{n}\sum_{i=1}^{k}\mathbb{1}_{\mathbf{x}_i\in A} - \frac{1}{n}\sum_{i=k+1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right| \\ & - \mathbb{E}\sup_{A\in\mathcal{A}}\left|\mathbb{P}(\mathbf{X}\in A) - \frac{1}{n}\sum_{i=1}^{k-1}\mathbb{1}_{\mathbf{x}_i\in A} - \frac{1}{n}\sum_{i=k}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right| . \end{aligned}$$

Using the fact that $\left|\sup_{A\in\mathcal{A}}|F(A)| - \sup_{A\in\mathcal{A}}|G(A)|\right| \leq \sup_{A\in\mathcal{A}}|F(A)-G(A)|$ for every function F and G of A, we obtain:

$$\left|h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{x}_k)\right| \;\leq\; \mathbb{E}\sup_{A\in\mathcal{A}}\frac{1}{n}|\mathbb{1}_{\mathbf{x}_k\in A} - \mathbb{1}_{\mathbf{X}_k\in A}| \;. \tag{16}$$

The term on the right hand side of (16) is less than $\frac{1}{n}$ so that $\mathrm{maxdev}^+ \leq \frac{1}{n}$. Moreover, if $\mathbf{X}'$ is an independent copy of $\mathbf{X}$, (16) yields

$$\left|h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{X}')\right| \;\leq\; \mathbb{E}\left[\sup_{A\in\mathcal{A}}\frac{1}{n}|\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}| \;\middle|\; \mathbf{X}'\right],$$

so that

$$\begin{aligned} \mathbb{E}\left[h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{X}')^2\right] \;&\leq\; \mathbb{E}\,\mathbb{E}\left[\sup_{A\in\mathcal{A}}\frac{1}{n}|\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}| \;\middle|\; \mathbf{X}'\right]^2 \\ &\leq\; \mathbb{E}\left[\sup_{A\in\mathcal{A}}\frac{1}{n^2}|\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}|^2\right] \\ &\leq\; \frac{1}{n^2}\mathbb{E}\left[\sup_{A\in\mathcal{A}}|\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}|\right] \end{aligned}$$

Thus $\mathbf{Var}(h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{X}_k)) \leq \mathbb{E}[h(\mathbf{x}_1,\ldots,\mathbf{x}_{k-1},\mathbf{X}_k)^2] \leq \frac{q}{n^2}$. Finally $\hat{v} \leq \frac{q}{n}$ as required. ■

As a consequence with Proposition 11 the following general inequality holds true:

$$\mathbb{P}\left[f(\mathbf{X}_{1:n}) - \mathbb{E}f(\mathbf{X}_{1:n}) \;\geq\; t\right] \;\leq\; e^{-\frac{nt^2}{2q+\frac{2t}{3}}} \tag{17}$$

where the quantity $q \;=\; \mathbb{E}\left(\sup_{A\in\mathcal{A}}|\mathbb{1}_{\mathbf{X}'\in A} - \mathbb{1}_{\mathbf{X}\in A}|\right)$ seems to be a central characteristic of the VC-class $\mathcal{A}$ given the distribution $\mathbf{X}$. It may be interpreted as a measure of the complexity of the class $\mathcal{A}$ with respect to the distribution of $\mathbf{X}$: how often the class $\mathcal{A}$ is able to separate two independent realizations of $\mathbf{X}$.

Recall that the union class $\mathbb{A}$ and its associated probability p are defined as $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Noting that for all $A \in \mathcal{A}$, $\mathbb{1}_{\{. \in A\}} \leq \mathbb{1}_{\{. \in \mathbb{A}\}}$, it is then straightforward from (15) that $q \leq 2p$. As a consequence (17) holds true when changing q by $2p$. Let us now explicit the link between the expectation of f and the Rademacher average

$$\mathcal{R}_n = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| ,$$

where $(\sigma_i)_{i \geq 1}$ is a Rademacher chaos independent of the $\mathbf{X}_i$'s.

Lemma 13 *With this notations the following inequality holds true:*

$$\mathbb{E} f(\mathbf{X}_{1:n}) \ \leq \ 2\mathcal{R}_n$$

Proof The proof of this lemma relies on classical arguments: Introducing a ghost sample $(\mathbf{X}_i^{'})_{1 \leq i \leq n}$ namely i.i.d independent copy of the $\mathbf{X}_i$'s, we may write:

$$\begin{aligned}
\mathbb{E} f(\mathbf{X}_{1:n}) &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \\
&= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i^{'} \in A} \right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \\
&\leq \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i^{'} \in A} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \\
&= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{\mathbf{X}_i^{'} \in A} - \mathbb{1}_{\mathbf{X}_i \in A} \right) \right| \\
&\leq \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i^{'} \in A} \right| \ + \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} -\sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| \\
&= 2\mathcal{R}_n
\end{aligned}$$

■

Combining (17) with Lemma 13 and the fact that $q \leq 2p$ gives:

$$\mathbb{P}\left[f(\mathbf{X}_{1:n}) - 2\mathcal{R}_n \geq t \right] \ \leq \ e^{-\frac{nt^2}{4p + \frac{2t}{3}}} . \tag{18}$$

Recall that the relative Rademacher average are defined in (13) as $\mathcal{R}_{n,p} = \mathcal{R}_n / p$. It is well-known that $\mathcal{R}_n$ is of order $\mathcal{O}((V_{\mathcal{A}}/n)^{1/2})$, see Koltchinskii (2006) for instance. However, we hope a stronger bound than just $\mathcal{R}_{n,p} = \mathcal{O}(p^{-1}(V_{\mathcal{A}}/n)^{1/2})$ since $\frac{1}{np} |\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i \in A}|$ with $\mathbb{P}(\mathbf{X}_i \in \mathbb{A}) = p$ is expected to be like $\frac{1}{np} |\sum_{i=1}^{np} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A}|$ with $\mathbf{Y}_i$ such that $\mathbb{P}(\mathbf{Y}_i \in \mathbb{A}) = 1$. The result below confirms this heuristic:

Lemma 14 *The relative Rademacher average $\mathcal{R}_{n,p}$ is of order $\mathcal{O}(\sqrt{\frac{V_{\mathcal{A}}}{pn}})$.*

Proof Let us defined $i.i.d.$ r.v. $\mathbf{Y}_i$ independent from $\mathbf{X}_i$ whose law is the law of $\mathbf{X}$ conditioned on the event $\mathbf{X} \in \mathbb{A}$. If $\overset{d}{=}$ means equal in distribution it is easy to show that $\sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \overset{d}{=} \sum_{i=1}^\kappa \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A}$, where $\kappa \sim Bin(n,p)$ independent of the $\mathbf{Y}_i$'s. Thus,

$$\begin{aligned}
\mathcal{R}_{n,p} &= \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^\kappa \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \\
&= \mathbb{E}\left[\mathbb{E}\left[\sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^\kappa \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \ \mid \kappa \right]\right] \\
&= \mathbb{E}\left[\Phi(\kappa)\right]
\end{aligned}$$

where

$$\phi(K) = \mathbb{E}\left[\sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^K \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \right] = \frac{K}{np} \mathcal{R}_K \leq \frac{K}{np} \frac{C\sqrt{V_{\mathcal{A}}}}{\sqrt{K}} \ .$$

Thus,

$$\mathcal{R}_{n,p} \ \leq \ \mathbb{E}\left[\frac{\sqrt{\kappa}}{np} C\sqrt{V_{\mathcal{A}}} \right] \ \leq \ \frac{\sqrt{\mathbb{E}[\kappa]}}{np} C\sqrt{V_{\mathcal{A}}} \ \leq \ \frac{C\sqrt{V_{\mathcal{A}}}}{\sqrt{np}} \ .$$

■

Finally we obtain from (18) and Lemma 14 the following bound:

$$\mathbb{P}\left[\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| - 2\mathcal{R}_{n,p} \ > \ t \right] \ \leq \ e^{-\frac{npt^2}{4 + \frac{2t}{3}}} \tag{19}$$

Solving $\exp\left[-\frac{npt^2}{4 + \frac{2}{3}t} \right] = \delta$ with $t > 0$ leads to

$$t \ = \ \frac{1}{3np} \log \frac{1}{\delta} + \sqrt{\left(\frac{1}{3np} \log \frac{1}{\delta} \right)^2 + \frac{4}{np} \log \frac{1}{\delta}} \ := \ h(\delta)$$

so that

$$\mathbb{P}\left[\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| - 2\mathcal{R}_{n,p} \ > h(\delta) \right] \ \leq \ \delta$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ if $a, b \geq 0$, we have $h(\delta) < \frac{2}{3np} \log \frac{1}{\delta} + 2\sqrt{\frac{1}{np} \log \frac{1}{\delta}}$. In the case of $\delta \geq e^{-np}$, $\frac{2}{3np} \log \frac{1}{\delta} \leq \frac{2}{3} \sqrt{\frac{1}{np} \log \frac{1}{\delta}}$ so that $h(\delta) < 3\sqrt{\frac{1}{np} \log \frac{1}{\delta}}$. This ends the proof.

Appendix B. Note on Remark 5

To obtain the bound in (6), the following easy to show inequality is needed before applying Theorem 1 :

$$\begin{aligned}\sup_{g\in\mathcal{G}} |L_{\alpha,n}(g) - L_\alpha(g)| \;\leq\; \frac{1}{\alpha}\Bigg[\sup_{g\in\mathcal{G}} \Bigg| \mathbb{P}\left(Y \neq g(\mathbf{X}),\ \|\mathbf{X}\| \ > t_\alpha\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i \neq g(\mathbf{X}_i),\ \|\mathbf{X}_i\| > t_\alpha\}} \Bigg| \\ + \left| \mathbb{P}\left(\|\mathbf{X}\| \ > t_\alpha\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\|\mathbf{X}_i\| > t_\alpha\}} \right| \ + \ \frac{1}{n} \Bigg].\end{aligned}$$

Note that the final objective would be to bound the quantity $\sup_{g\in\mathcal{G}} |L_\alpha(g) - L_\alpha(g^*_\alpha)|$, where g^*_α is a Bayes classifier for the problem at stake, *i.e.* a solution of the conditional risk minimization problem $\inf_{\{g \text{ measurable}\}} L_\alpha(g)$. Such a bound involves a bias term $\inf_{g\in\mathcal{G}} L_\alpha(g) - L_\alpha(g^*_\alpha)$, as in the classical setting. Further, it can be shown that the standard Bayes classifier $g^*(\mathbf{x}) := 2\mathbb{I}\{\eta(\mathbf{x}) > 1/2\} - 1$ (where $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$) is also a solution of the conditional risk minimization problem. Finally, the conditional bias $\inf_{g\in\mathcal{G}} L_\alpha(g) - L_\alpha(g^*_\alpha)$ can be expressed as $\frac{1}{\alpha}\inf_{g\in\mathcal{G}} \mathbb{E}\left[|2\eta(\mathbf{X}) - 1| \mathbb{1}_{g(\mathbf{X})\neq g^*(\mathbf{X})} \mathbb{1}_{\|\mathbf{X}\|\geq t_\alpha}\right]$, to be compared with the standard bias $\inf_{g\in\mathcal{G}} \mathbb{E}\left[|2\eta(\mathbf{X}) - 1| \mathbb{1}_{g(\mathbf{X})\neq g^*(\mathbf{X})}\right]$.