Orlicz norms and concentration inequalities for β-heavy tailed random variables

Linda Chamakh, Emmanuel Gobet, Wenjun Liu

Orlicz norms and concentration inequalities for β-heavy tailed random variables

LINDA CHAMAKH[1,2]
and EMMANUEL GOBET[1]
and WENJUN LIU[1]

[1] *CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France*
E-mail: emmanuel.gobet@polytechnique.edu; wenjun.liu@polytechnique.edu
[2] *Global Markets Quantitative Research – BNP Paribas, France*
E-mail: linda.chamakh@bnpparibas.com

We establish a new concentration-of-measure inequality for the sum of independent random variables with β-heavy tail. This includes exponential of Gaussian distributions (a.k.a. log-normal distributions), or exponential of Weibull distributions, among others. These distributions have finite polynomial moments at any order but may not have finite α-exponential moments. We exhibit a Orlicz norm adapted to this setting of β-heavy tails, we prove a new Talagrand inequality for the sum and a new maximal inequality. As consequence, a bound on the deviation probability of the sum from its mean is obtained, as well as a bound on uniform deviation probability.

MSC2020 subject classifications: Primary 60E15; secondary 60F10
Keywords: heavy tails; deviation inequality; Orlicz norm; Talagrand inequality; maximal inequality; empirical process

1. Introduction

1.1. Concentration inequalities

Understanding how sample statistical fluctuations impact prediction errors is crucial in learning algorithms. Typically, we are interested in bounding the probability that a sum of random variables exceeds a certain threshold, essentially in quantifying the deviation of the sum from its expectation. In other words, we aim at analyzing how fast the sum concentrates around its expectation. Take the notation $[M]$ for all integers from 1 to M included. For independent and centered random variables $(Y_m)_{m\in[M]}$ taking value in a Banach space $(B, \|.\|_B)$, the quantity of interest takes the form $\mathbb{P}\left(\left\|\sum_{m\in[M]} Y_m\right\|_B > \varepsilon\right) \leq f(\varepsilon, M)$ for the most explicit and tightest possible function f. The bounded, sub-Gaussian or the sub-exponential random variables have been largely covered by the literature (for example, via Bennett and Bernstein inequalities - see [BLM13] for an extensive review of main concentration inequalities techniques), as well as the case of alpha-exponential tails [CGS20] (random variables Y s.t. there exists $\alpha > 0, c > 0$, such that $\mathbb{E}\left[\exp(\frac{\|Y\|_B^\alpha}{c})\right] < \infty$). The fat-tailed case, for which the moment generating function does not exist but some polynomial moments exist, can be tackled for example via Burkholder or Fuk-Nagaev type of inequalities [Rio17, Mar17]. These inequalities are based on the existence and on the bounding of polynomial moments of the random variables. In this work, we focus on the heavy-tailed random variables case, in the limit case when no α-exponential moment is finite but every polynomial moment exist.

1.2. Orlicz norm

Orlicz norm [KR61] provides a nice tool to study the statistical fluctuations of an estimator for a given family of distributions. Consider an Orlicz function $\Psi : \mathbb{R}^+ \to \mathbb{R}^+$, that is a continuous non-decreasing function, vanishing in zero and with $\lim_{x\to+\infty} \Psi(x) = +\infty$, and define the Ψ-Orlicz norm of the B-valued random variable Y by

$$\|Y\|_\Psi := \inf\left\{c > 0 : \mathbb{E}\left[\Psi\left(\frac{\|Y\|_B}{c}\right)\right] \leq 1\right\}. \tag{1.1}$$

With the additional property that Ψ is convex, Orlicz functions are commonly referred to as "Young functions" (or "N-functions" as in [KR61]). Van de Geer and Lederer [vdGL13] exhibit in their work a "Bernstein-Orlicz" norm (the "(L)-Bernstein-Orlicz" norm) adapted to sub-Gaussian and sub-exponential tails and provide deviation inequalities for suprema of functions of random variables [vdGL13, Theorem 8]. The (L)-Bernstein-Orlicz norm is the Ψ_L-Orlicz norm with

$$\Psi_L(z) = \exp\left[\frac{\sqrt{1+2Lz}-1}{L}\right]^2 - 1.$$

Clearly $\|Y\|_{\Psi_L} < \infty$ implies the existence of exponential moment. As shown in Wellner [Wel17], it is possible to generalize these results to any Orlicz function $\Psi(x) = e^{h(x)} - 1$ with h convex. It requires again the existence of exponential moment which is not our framework. We would like to go beyond and do not assume any α-exponential moment.

As a new Orlicz function able to handle heavy-tail situations, we will consider:

$$\Psi_\beta^{\mathtt{HT}}(x) := \exp((\ln(x+1))^\beta) - 1, \qquad x \geq 0, \tag{1.2}$$

for a parameter $\beta > 1$. We say that Y is *β-heavy tailed* if there exists a $c > 0$ s.t.

$$\mathbb{E}\left[\Psi_\beta^{\mathtt{HT}}\left(\frac{\|Y\|_B}{c}\right)\right] < \infty.$$

Typically, we aim at encompassing situations like $Y = \exp(|G|^{\frac{2}{\beta}})$ where G is a Gaussian random variable; the case $\beta = 2$ corresponds to log-normal tails. See Section 2.2 for various examples.

Observe that when (1.1) is finite with $\Psi = \Psi_\beta^{\mathtt{HT}}$, Y has finite polynomial moment of order p for any $p > 0$, but may not have α-exponential moments. Besides, our β-heavy tailed setting is closely related to *long-tail* modelling[1], which is used for instance in queuing applications [Asm03, Chapter 10].

1.3. Deviation inequalities for sum via Talagrand and Markov inequalities

What we call Talagrand inequality is an inequality of type:

$$\left\|\sum_{m\in[M]} Y_m\right\|_\Psi \leq C_\Psi\left(\left\|\sum_{m\in[M]} Y_m\right\|_{L_1(B)} + \left\|\max_{m\in[M]} \|Y_m\|_B\right\|_\Psi\right). \tag{1.3}$$

[1] typically $S(x) := \mathbb{P}\left(\|Y\|_B > x\right) = \exp(-(\ln(1+x))^\beta)$ for which $\lim_{x\to+\infty} S(x+t)/S(x) = 1$ for any $t > 0$.

Talagrand [Tal89, Theorem 3] showed that this inequality is satisfied with $\Psi_\alpha(x) := e^{x^\alpha} - 1$ ($\alpha \in (0,1]$). For the sake of presentation, let us consider i.i.d. $(Y_m)_{m\in[M]}$. The first term is then $\left\|\sum_{m\in[M]} Y_m\right\|_{L_1(B)} \leq \mathcal{O}(\sqrt{M})$ by Bukholder inequality when $B \subseteq \mathbb{R}$ or the more general inequality of [Pis16, Proposition 4.35] when B is a Hilbert space or a Banach space of type 2.

When the maximal inequality is satisfied, that is under the form of [vdVW96, Lemma 2.2.2.], the second term is bounded by

$$\left\| \max_{m\in[M]} \|Y_m\|_B \right\|_\Psi \leq K_\Psi \Psi^{-1}(M) \max_{m\in[M]} \|Y_m\|_\Psi .$$

Hence, for any $\varepsilon > 0$, denoting $X := \frac{1}{M}\left\|\sum_{m\in[M]} Y_m\right\|_B$, thanks to the Markov inequality, the Talagrand inequality and the two previous norm controls, we get

$$\mathbb{P}(X \geq \varepsilon) \underset{Sect.\ 2.1-(iii)}{\leq} \frac{2}{1+\Psi(\varepsilon/\|X\|_\Psi)} = 2\left(1+\Psi\left(\frac{\varepsilon M}{\left\|\sum_{m\in[M]} Y_m\right\|_\Psi}\right)\right)^{-1} \tag{1.4}$$

$$\leq 2\left(1+\Psi\left(\frac{\varepsilon M}{C'_\Psi(\Psi^{-1}(M)+\sqrt{M})}\right)\right)^{-1}. \tag{1.5}$$

In particular, for $\Psi = \Psi_\alpha$, the above inequality simplifies to:

$$\mathbb{P}(X \geq \varepsilon) \leq 2\exp\left(-C'_\alpha\left(\varepsilon\sqrt{M}\right)^\alpha\right).$$

It is then possible to extend this type of inequality to suprema of functions as done in [CGS20], in the spirit of [Ada08]. In any case, a key element to derive these concentration inequalities is the Talagrand inequality (1.3).

1.4. Our contribution

The purpose of this work is mainly to establish the Talagrand inequality for $\Psi = \Psi_\beta^{\mathtt{HT}}$, to tackle β-heavy tailed random variables as a difference with previous contributions available in the literature, and to derive some ready-to-use consequences. Note that this particular Orlicz function (1.2) is not at all part of the general result established by Talagrand [Tal89, Proposition 12], which states that the inequality (1.3) holds for Orlicz function of the form $\Psi(x) := e^{x\zeta(x)}$ with ζ *non-decreasing* for x large enough and satisfying $\limsup_{u\to+\infty} \frac{\zeta(e^u)}{\zeta(u)} < +\infty$; indeed, in our setting, one easily checks that $\zeta(x) = x^{-1}\ln(\Psi_\beta^{\mathtt{HT}}(x)) = x^{-1}(\ln(\exp((\ln(x+1))^\beta)-1))$ is *decreasing* for x large.

1.5. Outline

In Section 2, we recall the motivating example and define the adapted Orlicz function. Then we state our main results: Talagrand inequality (Theorem 2.1), maximal inequality (Theorem 2.2), pointwise and uniform deviation estimates (Corollary 2.3 and Theorem 2.4). Derivations of finite-sample confidence intervals and of excess risk bounds [Kol11] in regression are given, as direct applications of our results. Section 3 is devoted to the proofs. In all these results, some universal constants appear: we do not investigate the question of having the best possible constants.

2. Motivating examples and main results

2.1. Orlicz norm properties

Although $\|.\|_\Psi$ defined in (1.1) may not satisfy in general the triangle inequality, we keep calling it Orlicz norm for the sake of simplicity. For a given Banach space $(B, \|.\|_B)$ over the field $\mathbb{R}$, we denote $L_\Psi(B) := \{Y : \Omega \to B \text{ s.t. } \|Y\|_\Psi < +\infty\}$ the set of B-valued random variables with finite Ψ-Orlicz norm. For self-containedness we summarize a few well-known properties of the $\|\cdot\|_\Psi$ norm, for a given Orlicz function Ψ, which hold independently of the convexity of Ψ (unless explicitly required). See [KR61], or more recently [CGS20, Section 4].

(i) Normalization: If $Y \in L_\Psi(B)$ then $\mathbb{E}\left[\Psi\left(\frac{\|Y\|_B}{\|Y\|_\Psi}\right)\right] \leq 1$.
(ii) Homogeneity: If $Y \in L_\Psi(B)$ and $c \in \mathbb{R}$ then $cY \in L_\Psi(B)$ and $\|cY\|_\Psi = |c| \|Y\|_\Psi$.
(iii) Deviation inequality: If $Y \in L_\Psi(B)$ then $\mathbb{P}(\|Y\|_B \geq c) \leq \frac{2}{\Psi(c/\|Y\|_\Psi)+1}$ for any $c \geq 0$.
(iv) If Ψ is convex, $\|.\|_\Psi$ satisfies to the triangle inequality.

2.2. Motivating examples of heavy-tailed distributions and adapted Orlicz norm

2.2.1. Log-normal distribution

Let Y be a scalar random variable with log-normal distribution, i.e.

$$\ln(Y) \overset{d}{=} \mathcal{N}\left(\mu, \sigma^2\right),$$

with $\sigma > 0$. The distribution of Y admits the density

$$f_Y(y; \mu, \sigma) := \frac{1}{\sigma\sqrt{2\pi}y} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} \mathbf{1}_{y>0}.$$

Let us investigate what kind of Orlicz function Ψ can be used to have $\|Y\|_\Psi < \infty$. In particular, we search for $\Psi(x) = \exp(\xi(x)) - 1$ such that ξ is non-decreasing, $\xi(0) = 0$ and $\lim_{x\to+\infty} \xi(x) = +\infty$ in order to ensure that $\Psi(0) = 0$ and $\lim_{x\to+\infty} \Psi(x) = +\infty$. Let $c > 0$, observe that

$$\mathbb{E}\left[\exp\left(\left|\xi\left(\frac{|Y|}{c}\right)\right|\right)\right] < \infty \implies \liminf_{x\to\infty} \xi\left(\frac{x}{c}\right) - \frac{(\ln x)^2}{2\sigma^2} = -\infty. \tag{2.1}$$

Consider the following functions for $\beta > 0$:

1. $\xi_\beta(x) = (\ln(x+1))^\beta$, $x \geq 0$. Note that the case $\beta \leq 1$ is not much interesting in our setting since it quantifies tails with finite expectation at most (fat tail cases).
2. $\xi_\beta(x) = (\ln(x+1))^\beta (\ln(\ln(x+1)+1))^\alpha$, $x \geq 0$, $\alpha \in \mathbb{R}$. This second case is a scale refinement of the first case. It is not studied here.

These functions satisfy the necessary condition (2.1) if $\beta < 2$ and for a large c [2]. Furthermore, since for any $c > 0$, $\xi_\beta\left(\frac{x}{c}\right) < \varepsilon(\ln x - \mu)^2$ for any $\varepsilon > 0$ for x large enough, $\mathbb{E}\left[\exp\left(\left|\xi_\beta\left(\frac{|Y|}{c}\right)\right|\right)\right] < +\infty$.

[2] $\beta = 2$ is possible under restriction on σ: if $\sigma < \frac{1}{\sqrt{2}}$, then $\liminf_{x\to\infty} \xi_2(x) - \frac{(\ln x)^2}{2\sigma^2} = -\infty$.

2.2.2. *Other distributions satisfying* $\mathbb{E}\left[\exp\left(\xi_\beta\left(\frac{\|Y\|_B}{c}\right)\right)\right] < +\infty$

The associated Orlicz function $\Psi_\beta^{\mathtt{HT}}(x) := \exp(\xi_\beta(x)) - 1$ is adapted to other distributions than just the log-normal distribution. For any random variable X admitting finite α-exponential moment with $\alpha > 1$, then Y defined by $\ln(Y) = X$ will admit β-heavy tailed for any $1 < \beta < \alpha$. We refer the reader to [CGS20, Table 2] for an exhaustive list of distributions admitting α-exponential moments. Here are a few examples:

- The Generalized normal distribution with parameters $c \in \mathbb{R}$, $b > 0$, $\alpha > 0$ has a density $f(x) = c_f e^{-\frac{1}{2}\left(\frac{|x-c|}{b}\right)^\alpha}$ up to a positive normalization constant c_f: it clearly admits a finite α-exponential moment. Hence $Y = \exp(X)$ where X has density f hence admits β-heavy tails for $\beta < \alpha$.
- The Skew normal distribution with parameters $b \in \mathbb{R}$, $c \in \mathbb{R}$, $v > 0$ has a density $f(x) = c_f e^{-\frac{(x-c)^2}{2v}} \Phi\left(\frac{b(x-c)}{\sqrt{v}}\right)$, where Φ denotes the standard Gaussian cumulative distribution function and c_f is a positive normalization constant: it admits 2-exponential moment. If X has this density, then $Y = \exp(X)$ has β-heavy tails for $\beta < 2$.
- The Weibull distribution with parameters $\lambda > 0, k > 0$ has a density $f(x) = c_f x^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \mathbf{1}_{x \geq 0}$ up to a positive normalization constant c_f: it has finite k-exponential moment. Consequently, $Y = \exp(X)$ where X has the density as above, admits β-heavy tails for $\beta < k$. Such distributions are used, for instance, for earthquake magnitude modelling [HR99] .

2.3. $\Psi_\beta^{\mathtt{HT}}$-Orlicz norm: properties and inequalities

We state different properties of the Orlicz function to be used for β-heavy tailed distribution. The proof is postponed to Section 3.4.

Proposition 2.1. *For $\beta > 0$ define $\Psi_\beta^{\mathtt{HT}} : \mathbb{R}^+ \to \mathbb{R}^+$ by*

$$\Psi_\beta^{\mathtt{HT}}(x) := \exp(\xi_\beta(x)) - 1 \quad \textit{with} \quad \xi_\beta(x) := (\ln(1+x))^\beta, \qquad x \geq 0. \tag{2.2}$$

The following properties hold:

1. *The application $\beta \mapsto \Psi_\beta^{\mathtt{HT}}$ defines a group isomorphism between $((0,+\infty),\times)$ and $((\Psi_\beta^{\mathtt{HT}} : \beta > 0), \circ)$, and in particular, $(\Psi_\beta^{\mathtt{HT}})^{-1} = \Psi_{1/\beta}^{\mathtt{HT}}$.*
2. *For $\beta > 0$, $\Psi_\beta^{\mathtt{HT}}$ is an Orlicz function.*
3. *For $\beta > 1$, $\Psi_\beta^{\mathtt{HT}}$ is convex.*
4. *For $\beta > 1$ (resp. $\beta < 1$), the limit as $x \to +\infty$ of $\Psi_\beta^{\mathtt{HT}}(x)/x^k$ equals to $+\infty$ (resp. 0), for any $k > 0$.*

As a consequence, the associated $\Psi_\beta^{\mathtt{HT}}$-Orlicz norm satisfies to the triangle inequality for $\beta > 1$.

Hereafter, we mostly restrict the results to the more interesting case $\beta > 1$. Let us start with the Talagrand inequality (1.3) for the $\Psi_\beta^{\mathtt{HT}}$-Orlicz norm.

Theorem 2.1 (Talagrand type inequality)**.** *Let $\beta \in (1,+\infty)$. Then there is a universal constant $\mathcal{K}_{\beta,(2.3)}$ s.t. for all independent, mean zero, random variables sequence $(Y_m)_{m\in[M]}$ with $Y_m \in$*

$L_{\Psi_\beta^{\mathtt{HT}}}(B)$ *for all* $m \in [M]$*, we have*

$$\left\| \sum_{m\in[M]} Y_m \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq \mathcal{K}_{\beta,(2.3)} \left(\left\| \sum_{m\in[M]} Y_m \right\|_{L_1(B)} + \left\| \max_{m\in[M]} \|Y_m\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} \right). \tag{2.3}$$

We also establish that the general maximal inequality [vdVW96, Lemma 2.2.2.] (recalled in Lemma 3.6) holds for the $\Psi_\beta^{\mathtt{HT}}$ function:

Theorem 2.2 (A $\Psi_\beta^{\mathtt{HT}}$ maximal inequality)**.** *Let* $\beta \in (1, +\infty)$*. Then there exists a universal constant* $\mathcal{C}_{\beta,(2.4)}$ *s.t. for any random variables* $Y_1, \ldots, Y_M$ *in* $L_{\Psi_\beta^{\mathtt{HT}}}(B)$*,*

$$\left\| \max_{m\in[M]} \|Y_m\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq \mathcal{C}_{\beta,(2.4)} (\Psi_\beta^{\mathtt{HT}})^{-1}(M) \max_{m\in[M]} \|Y_m\|_{\Psi_\beta^{\mathtt{HT}}}. \tag{2.4}$$

Recall that $(\Psi_\beta^{\mathtt{HT}})^{-1}(M) = \Psi_{1/\beta}^{\mathtt{HT}}(M)$. As a consequence of the Talagrand inequality (2.3) and the maximal inequality (2.4), by following the same steps as described in (1.5), we can derive the following concentration inequality:

Corollary 2.3 (A concentration inequality for sum of independent β-heavy tailed random variables)**.** *Let* $\beta \in (1, +\infty)$*. Assume that* B *is an Hilbert space or a Banach space of type* 2*. Then for any* $Y_1, \ldots, Y_M$ *independent and centered random variables in* $L_{\Psi_\beta^{\mathtt{HT}}}(B)$*, for any* $\varepsilon > 0$*,*

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{M} \left\| \sum_{m\in[M]} Y_m \right\|_B \geq \varepsilon \right) \\ &\leq 2\exp\left(- \left(\ln\left(1 + \frac{\varepsilon M}{\mathcal{K}_{\beta,(2.3)} \left(C(2)^{1/2} \mu_2 \sqrt{M} + \mathcal{C}_{\beta,(2.4)} \mu_{\Psi_\beta^{\mathtt{HT}}} \Psi_{1/\beta}^{\mathtt{HT}}(M) \right)} \right) \right)^\beta \right), \end{aligned} \tag{2.5}$$

where $\mu_{\Psi_\beta^{\mathtt{HT}}} := \max_{m\in[M]} \|Y_m\|_{\Psi_\beta^{\mathtt{HT}}}$ *and* $\mu_2 := \max_{m\in[M]} \|Y_m\|_{L_2(B)}$*,* $C(2)$ *denotes the universal constant in the Pisier inequality [Pis16, Proposition 4.35],* $\mathcal{K}_{\beta,(2.3)}$ *the Talagrand constant in* (2.3) *and* $\mathcal{C}_{\beta,(2.4)}$ *the maximal inequality constant in* (2.4)*.*

Recall that $\Psi_{1/\beta}^{\mathtt{HT}}(M)$ goes to infinity slowlier than M^k (for $\beta > 1$, $k > 0$) (Proposition 2.1-(4)). Thus, when $Y_1, \ldots, Y_M$ are i.i.d. – implying that $\mu_{\Psi_\beta^{\mathtt{HT}}}$ and μ_2 do not depend on M – the above upper bound takes the simple form

$$2\exp\left(- \left(\ln(1 + K\varepsilon\sqrt{M}) \right)^\beta \right),$$

for some universal constant $K > 0$ (depending on $\mu_{\Psi_\beta^{\mathtt{HT}}}$ and μ_2).

One may wonder about the sharpness of the bound (2.5) with respect to M and ε; there is no reason for which constants in (2.5) are optimal. For $M = 1$, one retrieves a bound of the form $2\exp\left(-(\ln(1+K\varepsilon))^{\beta}\right)$ which optimality (w.r.t. ε) is somehow equivalent (up to constant) to that of the Markov inequality combined with Orlicz norm in the left-hand side inequality (1.4): hence, possibilities of improvement are quite limited in general. For $M > 1$, investigating non-sharpness (w.r.t. ε) would require, for instance, to identify the distribution of $\sum_{m\in[M]} Y_m$ in some cases and doing so, showing a large gap between the left and right-hand sides of (2.5); unfortunately, to the best of our knowledge, we do not know such a situation where the distribution of the sum (under β-heavy tail conditions) has a tractable expression.

Besides, Corollary 2.3 can potentially be used to construct nonasymptotic confidence intervals for the mean of $f(X)$ using i.i.d. observations, under the assumption of β heavy-tails. For this, set $Y := f(X) - \mathbb{E}[f(X)]$. In that case, renormalizing the deviation by $\sqrt{M}$ as for asymptotic confidence intervals based on the central limit theorem (CLT), Corollary 2.3 writes

$$\mathbb{P}\left(\sqrt{M}\left\|\frac{1}{M}\sum_{m\in[M]} f(X_m) - \mathbb{E}[f(X)]\right\|_B \geq \varepsilon\right) \leq 2\exp\left(-(\ln(1+K\varepsilon))^{\beta}\right),$$

where K depends explicitly on constants of Corollary 2.3. Had these constants been known, we would obtain a tractable nonasymptotic confidence intervals. The rate is $\sqrt{M}$ as in usual CLT, but the dependence in ε in the tails is different because of the β-heavy tail setting. Alternatively, using CLT-based confidence intervals would reduce to asymptotic Gaussian bounds, which lighter tails would result in narrower confidence intervals: the latter might be quite incorrect for finite samples and it highlights the interest of Corollary 2.3 for deriving nonasymptotic confidence intervals.

In addition, the pointwise estimate from Corollary 2.3 can be turned into a uniform deviation estimate. On the technical side, the strategy consists in splitting the deviation between truncated functions and their residuals. The residuals are handled using Hoffman-Jorgensen inequality [LT13, Proposition 6.8], following an initial idea from [Ada08] and the recent analysis of [CGS20]. The "truncated part" can be handled using Klein-Rio concentration bounds, together with the Dudley entropy integral bounds. For the latter which is related to the complexity of the space of functions and their related covering numbers, we choose to describe it using its Vapnik-Chervonenkis (VC) dimension (see [GKKW02, Theorem 9.4]). For alternative descriptions, see [vdG00, Sections 2.3 and 2.4] and [NP07]; adaptation of the following result to these other complexity descriptions is somehow direct and left to the reader.

Theorem 2.4 (A uniform concentration inequality for β-heavy tailed random variables)**.** *Let $\beta \in (1,+\infty)$. Let $(X_1,\dots,X_M)$ be independent random variables taking values in $\mathbb{R}^d$ and let $\mathcal{F}$ be a countably-generated class of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ with envelope $F(x) := \sup_{f\in\mathcal{F}} |f(x)|$, such that $F(X_m) \in L_{\Psi_\beta^{\mathtt{HT}}}(\mathbb{R})$ for any $m \in [M]$. Set*

$$\begin{aligned}
\mu_{\Psi_\beta^{\mathtt{HT}}} &:= \max_{m\in[M], f\in\mathcal{F}} \|f(X_m)\|_{\Psi_\beta^{\mathtt{HT}}}, \\
\bar{\mu}_{\Psi_\beta^{\mathtt{HT}}} &:= \max_{m\in[M]} \|F(X_m)\|_{\Psi_\beta^{\mathtt{HT}}}, \\
\mu_2 &:= \max_{m\in[M], f\in\mathcal{F}} \|f(X_m)\|_{L_2}.
\end{aligned} \tag{2.6}$$

Assume that the Vapnik-Chervonenkis dimension $\mathrm{V}_{\mathcal{F}^+}$ *of* $\mathcal{F}^+ := \{\{(x,t) \in \mathbb{R}^d \times \mathbb{R}, t \leq f(x)\}; f \in \mathcal{F}\}$ *is finite. Then, there exist two universal constants* K_1, K_2 *(depending only on* β*) such that for any* $\varepsilon > 0$ *satisfying the constraint*

$$\varepsilon \geq K_1 c \sqrt{\frac{\mathrm{V}_{\mathcal{F}^+}}{M}} \tag{2.7}$$

with

$$c := \left(K_1 \Psi^{\mathtt{HT}}_{1/\beta}(M) \bar{\mu}_{\Psi^{\mathtt{HT}}_\beta}\right) \vee \left(\mu_{\Psi^{\mathtt{HT}}_\beta}\left(\exp\left[\left(2\ln_+\left(K_1 \mu_{\Psi^{\mathtt{HT}}_\beta}/\varepsilon\right)\right)^{1/\beta}\right] - 1\right)\right), \tag{2.8}$$

$$\ln_+(x) := \max(\ln(x), 0),$$

we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{M} \sum_{m\in[M]} \left(f(X_m) - \mathbb{E}\left[f(X_m)\right]\right) \geq \varepsilon\right)$$

$$\leq 2\exp\left(-\left(\ln\left(1 + \frac{M\varepsilon}{K_2 \bar{\mu}_{\Psi^{\mathtt{HT}}_\beta} \Psi^{\mathtt{HT}}_{1/\beta}(M)}\right)\right)^\beta\right) + \exp\left(-\frac{M\varepsilon^2}{K_2(\mu_2^2 + c\varepsilon)}\right). \tag{2.9}$$

A similar bound holds for lower deviations, i.e. replacing the sup and $\geq \varepsilon$ by inf and $\leq -\varepsilon$: it is obtained by changing $\mathcal{F}$ into $-\mathcal{F}$ in the bounds.

If $\mathcal{F}$ is a finite-dimensional vector space, $\mathrm{V}_{\mathcal{F}^+} \leq \dim(\mathcal{F}) + 1$ [GKKW02, Theorem 9.5].

For i.i.d. $(X_m)_m$, i.e. the μ-parameters (2.6) do not depend on M, both the condition (2.8) and the bound (2.9) take simple forms in terms of M (without focusing much on the best constants), which makes Theorem 2.4 even more easily applicable.

- The bound (2.9) becomes $2\exp\left(-\left(\ln(1 + \frac{M\varepsilon}{K\Psi^{\mathtt{HT}}_{1/\beta}(M)})\right)^\beta\right) + \exp\left(-\frac{M\varepsilon^2}{K(1+c\varepsilon)}\right)$ for a positive constant K depending on β and the μ-parameters.
- The equation (2.8) becomes simply

$$c := K_1 \Psi^{\mathtt{HT}}_{1/\beta}(M), \tag{2.10}$$

  with a new constant K_1, depending on β and the μ-parameters. Indeed, from the first term in the definition (2.8) of c, one gets that $c \geq \inf_{M\geq 1}\left(K_1 \Psi^{\mathtt{HT}}_{1/\beta}(M) \bar{\mu}_{\Psi^{\mathtt{HT}}_\beta}\right) =: c_0 > 0$, which, from (2.7), yields the rough lower bound $\varepsilon \geq K_1 c_0/\sqrt{M}$. This implies in turn (after tedious computations) that the second term in the definition (2.8) of c cannot be (up to constant) larger than the first term, hence the equality (2.10).
- Furthermore, the probability upper bound (2.9) can be re-parametrized under the form $\mathbb{P}(\cdots \geq \varepsilon(t)) \leq e^{-t}$ for any given $t \geq 0$ with some appropriate $\varepsilon(t)$. Consider again the i.i.d. case to simplify the exposure, leveraging the above simplifications.
  - The inequality $2\exp\left(-\left(\ln\left(1 + \frac{M\varepsilon}{K\Psi^{\mathtt{HT}}_{1/\beta}(M)}\right)\right)^\beta\right) \leq e^{-t}/2$ holds when $\varepsilon \geq \varepsilon_1(t) := K\frac{\Psi^{\mathtt{HT}}_{1/\beta}(M)}{M}\Psi^{\mathtt{HT}}_{1/\beta}(4e^t - 1)$.

– Because $\exp\left(-\frac{M\varepsilon^2}{K(1+c\varepsilon)}\right) \le \exp\left(-\frac{M\varepsilon^2}{2K}\right) + \exp\left(-\frac{M\varepsilon}{2Kc}\right)$, the inequality $\exp\left(-\frac{M\varepsilon^2}{K(1+c\varepsilon)}\right) \le e^{-t}/2$ holds as soon as $\varepsilon \ge \varepsilon_2(t) := \sqrt{\frac{2K(t+\ln(4))}{M}}$ and $\varepsilon \ge \varepsilon_3(t) := \frac{2KK_1\Psi^{\mathtt{HT}}_{1/\beta}(M)(t+\ln(4))}{M}$.

– Last, (2.7) remains, i.e. $\varepsilon \ge \varepsilon_0 := K_1^2\Psi^{\mathtt{HT}}_{1/\beta}(M)\sqrt{\frac{\mathrm{V}_{\mathcal{F}^+}}{M}}$.

All in all, (2.9) can be rewritten as

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{1}{M}\sum_{m\in[M]}(f(X_m)-\mathbb{E}\left[f(X_m)\right]) \ge \max(\varepsilon_0,\varepsilon_1(t),\varepsilon_2(t),\varepsilon_3(t))\right) \le e^{-t}, \qquad t\ge 0.$$

Observe that as $t\to+\infty$ (all other parameters being fixed), the terms $\varepsilon_2(t),\varepsilon_3(t)$ grows like t and $\sqrt{t}$, like in the usual cases of sub-exponential and sub-gaussian tails. The effect of β-heavy tails appears in $\varepsilon_1(t)$ which grows like $e^{t^{1/\beta}}$.

Theorem 2.4 is an instrumental result in statistical learning theory, in particular for Empirical Risk Minimization. We refer to the lectures [Kol11] for a broad exposition. Let us exemplify in a regression problem, by deriving data-dependent bounds on the excess risk in this setting of β-heavy tailed random variables. Results for bounded functions can be found in [Kol11, Chapter 5]. Under the assumptions of Theorem 2.4, set

$$f^\star(x) := \mathbb{E}\left[Y \mid X = x\right] \quad \text{and} \quad f_M(x) := \arg\inf_{f\in\mathcal{F}}\frac{1}{M}\sum_{m=1}^{M}(Y_m - f(X_m))^2,$$

corresponding to the empirical regression function with quadratic loss. For a given measurable function $g:\mathbb{R}^d\times\mathbb{R}\mapsto\mathbb{R}$, define its squared L_2 norm under the true and the empirical measures by

$$\|g(X,Y)\|_2^2 := \int_{\mathbb{R}^d}\int_{\mathbb{R}} g^2(x,y)P(\mathrm{d}x,\mathrm{d}y), \qquad \|g(X,Y)\|_{2,M}^2 := \frac{1}{M}\sum_{m=1}^{M} g^2(X_m,Y_m).$$

We claim that the excess risk is bounded as follows

$$\begin{aligned}&\|f_M(X)-f^\star(X)\|_2^2 - \inf_{f\in\mathcal{F}}\|f(X)-f^\star(X)\|_2^2\\ &\le \sup_{f\in\mathcal{F}}\left(\|Y-f(X)\|_2^2 - \|Y-f(X)\|_{2,M}^2\right) + \sup_{f\in\mathcal{F}}\left(\|Y-f(X)\|_{2,M}^2 - \|Y-f(X)\|_2^2\right), \quad (2.11)\end{aligned}$$

which implies a bound on the probability of excess risk deviating more than ε:

$$\begin{aligned}&\mathbb{P}\left(\|f_M(X)-f^\star(X)\|_2^2 - \inf_{f\in\mathcal{F}}\|f(X)-f^\star(X)\|_2^2 \ge \varepsilon\right)\\ &\le \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left(\|Y-f(X)\|_2^2 - \|Y-f(X)\|_{2,M}^2\right) \ge \varepsilon/2\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left(\|Y-f(X)\|_{2,M}^2 - \|Y-f(X)\|_2^2\right) \ge \varepsilon/2\right).\end{aligned}$$

The above probabilities are estimated by two applications of Theorem 2.4 with $\tilde{X} = (X,Y)$ and $\tilde{\mathcal{F}}_\pm = \{(x,y)\mapsto \pm(y-f(x))^2, f\in\mathcal{F}\}$, leading to an explicit bound for the probability of large excess risk. To prove (2.11), observe that for any $f\in\mathcal{F}$, $\|Y-f(X)\|_2^2 = \|Y-f^\star(X)\|_2^2 + \|f(X)-f^\star(X)\|_2^2$

and

$$\|Y - f_M(X)\|_{2,M}^2 - \inf_{f\in\mathcal{F}} \|Y - f(X)\|_{2,M}^2 = 0. \tag{2.12}$$

This readily gives

$$\begin{aligned}
&\|f_M(X) - f^\star(X)\|_2^2 - \inf_{f\in\mathcal{F}} \|f(X) - f^\star(X)\|_2^2 \\
&= \|Y - f_M(X)\|_2^2 - \inf_{f\in\mathcal{F}} \|Y - f(X)\|_2^2 - \|Y - f_M(X)\|_{2,M}^2 + \inf_{f\in\mathcal{F}} \|Y - f(X)\|_{2,M}^2 \\
&\leq \sup_{f\in\mathcal{F}} \left(\|Y - f(X)\|_2^2 - \|Y - f(X)\|_{2,M}^2 \right) + \sup_{f\in\mathcal{F}} \left(\|Y - f(X)\|_{2,M}^2 - \|Y - f(X)\|_2^2 \right),
\end{aligned}$$

as claimed.

3. Proofs

3.1. Proof of Theorem 2.1

3.1.1. Preliminary results

Here, we recall Lemmas 8 and 9 of [Tal89], as well as the "Basic Estimate", which will enable us to prove Theorem 2.1. In addition to the independent B-valued random variables $(Y_m)_{m\in[M]}$, we will need to consider extra independent Rademacher random variables. Everything is defined as follows. Let $\left(\Omega^M \times \Omega', \sum^M \otimes \sum', \mathbb{P}\right)$ the basic probability space, where $\mathbb{P} = \mathrm{P} \otimes \mathrm{P}'$ such that the variables Y_m are defined on Ω^M and for $\omega = (\omega_m)_{m\in[M]}$, $Y_m(\omega)$ depends only on ω_m. Let $(\varepsilon_m)_{m\in[M]}$ be a set of random variables defined on Ω' with a Rademacher distribution independent of $(Y_m)_{m\in[M]}$. The following inequalities can be proven independently apart from the context of Orlicz norms.

Lemma 3.1 ([Tal89, Lemma 8]). *If* $\mathbb{P}\left(\max_{m\in[M]} \|Y_m\|_B \geq t\right) \leq \frac{1}{2}$, *then*

$$\sum_{m\in[M]} \mathbb{P}\left(\|Y_m\|_B \geq t\right) \leq 2\mathbb{P}\left(\max_{m\in[M]} \|Y_m\|_B \geq t\right). \tag{3.1}$$

Lemma 3.2 ([Tal89, Lemma 9]). *Set* $X^{(r)}$ *the* r*-th largest term of* $(\|Y_m\|_B)_{m\in[M]}$. *Then*

$$\mathbb{P}\left(X^{(r)} \geq t\right) \leq \frac{1}{r!} \left(\sum_{m\in[M]} \mathbb{P}\left(\|Y_m\|_B \geq t\right) \right)^r. \tag{3.2}$$

Set

$$\mu := \mathbb{E}\left[\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_B \right], \tag{3.3}$$

$\mu > 0$ because the Y_m's are not all zero random variables (to avoid trivial situations). We now recall a key inequality which, combined with the previous lemmas, will enable us to prove the announced theorem.

Theorem 3.1 ([Tal89, Equation (2.5)])**.** *For k, q positive integers s.t. $k \geq q$, $u > 0$ and $u' > 0$, we have*

$$\mathbb{P}\left(\left\|\sum_{i\in[M]} \varepsilon_i Y_i\right\|_B \geq 4q\mu + u + u'\right) \leq 4\exp\left(-\frac{u^2}{64q\mu^2}\right) + \left(\frac{K_0}{q}\right)^k + \mathbb{P}\left(\sum_{r\leq k} X^{(r)} > u'\right)$$

where the constant K_0 is a universal constant.

3.1.2. Symmetrisation argument for Ψ convex

In the next Subsection, because we rely on Theorem 3.1, we are going to prove the inequality (2.3) on symmetric random variables first (e.g. variables Y_m s.t. $\varepsilon_m Y_m \overset{d}{=} Y_m$). The extension to non-symmetric variables will be direct thanks to Lemma 3.4 which establishes an "equivalence in norms" relationship between the Orlicz norm of the sum of random variables and its associated Rademacher average.

Lemma 3.3. *Let Ψ be convex Orlicz function and $\|\cdot\|_\Psi$ the associate Orlicz norm. For any mean zero random variable $Z \in L_\Psi(B)$, we have $\|Z\|_\Psi \leq \left\|Z - Z'\right\|_\Psi$, with Z' any B-valued random variable such that $\mathbb{E}\left[Z'|Z\right] = 0$.*

Proof. Let $c > 0$,

$$\mathbb{E}\left[\Psi(\|Z\|_B/c)\right] \overset{(a)}{=} \mathbb{E}\left[\Psi\left(\frac{\left\|\mathbb{E}\left[Z - Z' \mid Z\right]\right\|_B}{c}\right)\right] \overset{(b)}{\leq} \mathbb{E}\left[\Psi\left(\frac{\mathbb{E}\left[\left\|Z - Z'\right\|_B \mid Z\right]}{c}\right)\right]$$
$$\overset{(c)}{\leq} \mathbb{E}\left[\mathbb{E}\left[\Psi\left(\frac{\left\|Z - Z'\right\|_B}{c}\right) \mid Z\right]\right] = \mathbb{E}\left[\Psi\left(\frac{\left\|Z - Z'\right\|_B}{c}\right)\right]$$

where in (a) we use Z' has a zero conditional mean, in (b) we use that Ψ is non decreasing and the triangular inequality holds for the $\|.\|_B$, in (c) we apply the Jensen inequality. Hence by taking $c = \left\|Z - Z'\right\|_\Psi > 0$, the right hand side is smaller than 1 (using Property (i) in Section 2.1), and therefore $\|Z\|_\Psi \leq c = \left\|Z - Z'\right\|_\Psi$. □

Lemma 3.4. *Let Ψ be as in Lemma 3.3. Let $(Y_m)_{m\in[M]}$ be a sequence of independent mean-zero random variables in $L_\Psi(B)$. Let $(\varepsilon_m)_{m\in[M]}$ be independent Rademacher random variables, and let $(Y'_m)_{m\in[M]}$ be an independent copy of the sequence $(Y_m)_{m\in[M]}$. Then*

$$\left\|\sum_{m\in[M]} Y_m\right\|_\Psi \leq \left\|\sum_{m\in[M]} Y_m - \sum_{m\in[M]} Y'_m\right\|_\Psi = \left\|\sum_{m\in[M]} \varepsilon_m(Y_m - Y'_m)\right\|_\Psi$$

$$\leq 2 \left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_\Psi \leq 4 \left\| \sum_{m\in[M]} Y_m \right\|_\Psi .$$

Later on, we will apply these inequalities with $\Psi = \Psi_\beta^{\mathtt{HT}}$ and $\Psi(x) = x$ (the associated Orlicz norm corresponds then to the L_1 norm).

Proof. The first inequality comes from the application of Lemma 3.3 with $Z = \sum_{m\in[M]} Y_m$ and $Z' = \sum_{m\in[M]} Y'_m$. Since ε_m takes values ± 1 independently of Z, Z', we have $Y_m - Y'_m \overset{d}{=} Y'_m - Y_m \overset{d}{=} \varepsilon_m(Y_m - Y'_m)$. Since the sequences are independent in m, we obtain the equality of Lemma 3.4. The second inequality is a consequence of the triangular inequality (iv) and the previous identities in distribution. The last inequality is a consequence of the application of Lemma 3.3 with $Z = \sum_{m\in[M]} \varepsilon_m Y_m$ and $Z' = \sum_{m\in[M]} \varepsilon_m Y'_m$ satisfying

$$\mathbb{E}\left[Z'|Z\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{m\in[M]} \varepsilon_m Y'_m \,|\varepsilon_m, Y_m, m\in[M]\right] \mid Z\right] = 0$$

and of the triangular inequality: $\|Z\|_\Psi \leq \left\|Z - Z'\right\|_\Psi = \left\|\sum_{m\in[M]}(Y_m - Y'_m)\right\|_\Psi \leq 2\left\|\sum_{m\in[M]} Y_m\right\|_\Psi$. □

3.1.3. *Completion of the proof of Theorem 2.1*

We will denote K a positive constant depending only on β, that may vary from line to line. We assume that at least one of the Y_m's is not zero *a.s.*, otherwise the announced inequality (2.3) is obvious.

In view of the inequalities of Lemma 3.4, it is enough to do the reasoning and show the inequality (2.3) with the variables $(\varepsilon_m Y_m, m\in[M])$ instead of $(Y_m, m\in[M])$.

▷ **Rescaling.** Note that (2.3) is invariant by homogeneous rescaling (see Property (ii) of Section 2.1), i.e. the inequality remains the same for the random variables $\tilde{Y}_m := \frac{\varepsilon_m Y_m}{C}$ for any $C > 0$. For the choice

$$C := \left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_{L_1(B)} + \left\| \max_{m\in[M]} \|Y_m\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} > 0,$$

observe that

$$\left\| \sum_{m\in[M]} \tilde{Y}_m \right\|_{L_1(B)} \leq 1 \qquad \text{and} \qquad \left\| \max_{m\in[M]} \left\|\tilde{Y}_m\right\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq 1, \tag{3.4}$$

therefore the inequality (2.3) writes

$$\left\| \sum_{m\in[M]} \tilde{Y}_m \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq 2K.$$

Conversely, if the above holds for some K (independent from the $\tilde{Y}_m$'s verifying (3.4)), then (2.3) holds for the Y_m's. All in all, it means that without loss of generality, we can assume

$$\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_{L_1(B)} \leq 1 \qquad \text{and} \qquad \left\| \max_{m\in[M]} \|Y_m\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq 1,$$

and then show, under these assumptions, the existence of $K \in \mathbb{R}$ (independent on Y_m's) such that

$$\mathbb{E}\left[\exp\left(\xi_\beta \left(\frac{\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_B}{K} \right) \right) \right] \leq 2.$$

▷ **Deviation bounds.** By Property (iii) of Section 2.1 and since we assumed $\left\| \max_{m\in[M]} \|Y_m\|_B \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq 1$,

$$\mathbb{P}\left(\max_{m\in[M]} \|Y_m\|_B \geq t \right) \leq 2\exp\left(-\xi_\beta(t)\right), \qquad t \geq 0.$$

The function $\xi_\beta(\cdot) = (\ln(1+\cdot))^\beta$ being continuously increasing from 0 to $+\infty$, there exists t_0 s.t. $\xi_\beta(t_0) = 2\ln 2$ and $\forall t \geq t_0,\ 2\exp\left(-\xi_\beta(t)\right) \leq 1/2$; for further use, notice the value $t_0 = e^{(2\ln 2)^{\frac{1}{\beta}}} - 1$. Then applying Lemma 3.1, for $t \geq t_0$, we have

$$\sum_{m\in[M]} \mathbb{P}\left(\|Y_m\|_B \geq t \right) \leq 2\mathbb{P}\left(\max_{m\in[M]} \|Y_m\|_B \geq t \right) \leq 4\exp\left(-\xi_\beta(t)\right).$$

Hence Lemma 3.2 yields for $r \in \mathbb{N}^*$, $t \geq t_0$

$$\mathbb{P}\left(X^{(r)} \geq t \right) \leq \frac{4^r \exp\left(-r\xi_\beta(t)\right)}{r!}. \tag{3.5}$$

Denote $\tilde{\beta} = \lfloor \beta \rfloor + 1 \geq 2$. Equation (3.5) yields for $t \geq (e^{\tilde{\beta}} - 1) r^{\frac{\tilde{\beta}}{\beta}}$ (notice that $t \geq e^2 - 1 \geq t_0$ as requested)

$$\mathbb{P}\left(X^{(r)} \geq t r^{-\frac{\tilde{\beta}}{\beta}} \right) \leq \frac{4^r \exp\left(-r[\ln(1 + t r^{-\frac{\tilde{\beta}}{\beta}})]^\beta \right)}{r!} =: f(r,t).$$

Since $\tilde{\beta}/\beta > 1$, the sequence $(r^{-\frac{\tilde{\beta}}{\beta}})_{r\geq 1}$ is summable. Set $S_\beta := \sum_{r\geq 1} r^{-\frac{\tilde{\beta}}{\beta}} < +\infty$ and $g(t) := (t/(e^{\tilde{\beta}} - 1))^{\beta/\tilde{\beta}}$. From the inclusion $\{\sum_{r\leq g(t)} X^{(r)} \geq t S_\beta\} \subset \bigcup_{r\leq g(t)} \{X^{(r)} \geq t r^{-\frac{\tilde{\beta}}{\beta}}\}$ and writing a union bound, we get

$$\mathbb{P}\left(\sum_{r\leq g(t)} X^{(r)} \geq t S_\beta \right) \leq \sum_{r\leq g(t)} f(r,t). \tag{3.6}$$

We claim that for all $1 \leq r \leq g(t)$

$$r^{1/\beta} \ln(1 + t r^{-\frac{\tilde{\beta}}{\beta}}) \geq \ln(1+t). \tag{3.7}$$

This is a consequence of the above lemma applied with $\rho = r^{\frac{1}{\beta}} \geq 1$ and $\tau = t r^{-\frac{\tilde{\beta}}{\beta}} \geq e^{\tilde{\beta}} - 1$.

Lemma 3.5. *For all $\rho \geq 1$ and $\tau \geq e^{\tilde{\beta}} - 1$, we have $\rho \ln(1+\tau) \geq \ln(1 + \tau \rho^{\tilde{\beta}})$.*

Proof. The function $f(\rho) := \rho \ln(1+\tau) - \ln(1 + \tau\rho^{\tilde{\beta}})$ vanishes at $\rho = 1$, let us prove that it is non-decreasing in ρ provided that $\tau \geq e^{\tilde{\beta}} - 1$. Indeed,

$$f'(\rho) = \ln(1+\tau) - \frac{\tilde{\beta}\rho^{\tilde{\beta}-1}\tau}{1+\rho^{\tilde{\beta}}\tau}.$$

Since $\rho^{\tilde{\beta}}\tau \geq 0$ and $\rho \geq 1$, we have $\frac{\tilde{\beta}\rho^{\tilde{\beta}-1}\tau}{1+\rho^{\tilde{\beta}}\tau} \leq \frac{\tilde{\beta}}{\rho} \leq \tilde{\beta}$. Hence, $f'(\rho) \geq \ln(1+\tau) - \tilde{\beta} \geq 0$. We are done. □

Plugging (3.7) into (3.6) yields

$$\mathbb{P}\left(\sum_{r \leq g(t)} X^{(r)} \geq t S_\beta \right) \leq \sum_{r \leq g(t)} \frac{4^r}{r!} \exp\left(-[\ln(1+t)]^\beta \right) \leq \exp(4) \exp(-\xi_\beta(t)).$$

Let us recall that $\mu = \left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_{L_1(B)} \leq 1$. We are now at the point to apply Theorem 3.1 with $q = \lceil eK_0 \rceil$, $u = t$, $u' = tS_\beta$, $2q\mu \leq t$ and $k = \lfloor g(t) \rfloor$:

$$\begin{aligned}
&\mathbb{P}\left(\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_B \geq t\left(S_\beta + 3\right) \right) \\
&\leq \mathbb{P}\left(\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_B \geq 4q\mu + u + u' \right) \\
&\leq 4\exp\left(-\frac{t^2}{64 q \mu^2} \right) + \exp\left(-\lfloor g(t) \rfloor \right) + \mathbb{P}\left(\sum_{r \leq g(t)} X^{(r)} \geq t S_\beta \right) \\
&\leq 4\exp\left(-\frac{t^2}{64 q \mu^2} \right) + \exp\left(-\lfloor g(t) \rfloor \right) + \exp\left(4 - \xi_\beta(t) \right).
\end{aligned}$$

The above inequality is valid for any $t \geq t_0 \vee (2\lceil eK_0 \rceil \mu)$. Besides, in the above upper bound, the last third is asymptotically the largest one, therefore there exists $K > 0$ such that

$$\mathbb{P}\left(\left\| \sum_{m\in[M]} \varepsilon_m Y_m \right\|_B \geq Kt \right) \leq K \exp\left(-\xi_\beta(t) \right), \qquad t \geq 0.$$

▷ **Orlicz norm bounds.** The estimate implies for all $c > 0$:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\xi_\beta\left(\left\|\sum_{m\in[M]}\varepsilon_m Y_m\right\|_B/(cK)\right)\right)\right]-1\\
&=\int_0^\infty \exp\left(\xi_\beta(t)\right)\xi_\beta'(t)\,\mathbb{P}\left(\frac{\left\|\sum_{m\in[M]}\varepsilon_m Y_m\right\|_B}{cK}\geq t\right)\mathrm{d}t\\
&\leq K\int_0^\infty \xi_\beta'(t)\exp\left(\xi_\beta(t)-\xi_\beta(ct)\right)\mathrm{d}t. \qquad (3.8)
\end{aligned}$$

Let us check that the above integral is finite for $c > 1$. Only the integrability at $t \to +\infty$ is questionable. Write

$$\begin{aligned}
\xi_\beta(t)-\xi_\beta(ct) &= (\ln(1+t))^\beta\left(1-\left[1+\frac{\ln(\frac{(1+ct)}{(1+t)})}{\ln(1+t)}\right]^\beta\right)\\
&\approx_{t\to+\infty} -\beta(\ln(1+t))^{\beta-1}\ln(c).
\end{aligned}$$

Therefore, the function to integrate is bounded for t large by (up to constant)

$$g(t) := \frac{(\ln(1+t))^{\beta-1}}{(1+t)}e^{-\frac{1}{2}\beta(\ln(1+t))^{\beta-1}\ln(c)}.$$

We easily check that $\int_0^{+\infty} g(t)\mathrm{d}t = \int_0^{+\infty} y^{\beta-1}e^{-\frac{1}{2}\beta y^{\beta-1}\ln(c)}\mathrm{d}y < +\infty$ since $\beta > 1$ and $c > 1$.

Furthermore, by monotone convergence theorem, the bound (3.8) converges to 0 as $c \to +\infty$, consequently

$$\mathbb{E}\left[\exp\left(\xi_\beta\left(\frac{\left\|\sum_{m\in[M]}\varepsilon_m Y_m\right\|_B}{cK}\right)\right)\right]\leq 2$$

for a $c = c_\beta$ large enough. We have proved that $\left\|\sum_{m\in[M]}\varepsilon_m Y_m\right\|_{\Psi_\beta^{\mathrm{HT}}} \leq c_\beta K$. □

3.2. Proof of Theorem 2.2

We start by recalling the general maximal inequality on which our proof is based.

Lemma 3.6 ([vdVW96, Lemma 2.2.2]). *Let Ψ be a convex Orlicz function satisfying*

$$\limsup_{x,y\to+\infty}\Psi(x)\Psi(y)/\Psi(c_\Psi xy) < +\infty \qquad (3.9)$$

for some constant $c_\Psi > 0$. Then, there is a constant $K > 0$ such that for any B-valued random variables $Y_1, \ldots, Y_M$,

$$\left\|\max_{m\in[M]}\|Y_m\|_B\right\|_\Psi \leq K\Psi^{-1}(M)\max_{m\in[M]}\|Y_m\|_\Psi.$$

For $\beta > 1$, $\Psi_\beta^{\mathtt{HT}}$ is a convex Orlicz function, thus it remains to establish (3.9) to get Theorem 2.2. We prove that one can take $c_\Psi = 1$. Let $c \geq 1$[3] s.t. $\Psi_\beta^{\mathtt{HT}}(c^2) \geq 1$. Let x, y st. $x \geq c$ and $y \geq c$: then

$$\begin{aligned}\Psi_\beta^{\mathtt{HT}}(x)\Psi_\beta^{\mathtt{HT}}(y) &\leq e^{(\ln(1+x))^\beta} e^{(\ln(1+y))^\beta} \\ &\leq e^{(\ln(x))^\beta + (\ln(y))^\beta - (\ln(xy))^\beta} e^{(\ln(1+xy))^\beta} \underbrace{e^{2\sup_{z\geq c}(\ln(1+z))^\beta - (\ln z)^\beta}}_{:=C(c)}.\end{aligned}$$

- $C(c)$ is finite: indeed, by standard equivalents, we have that

$$(\ln(1+z))^\beta - (\ln z)^\beta = (\ln z)^\beta \left(\left[1 + \frac{\ln(1+z^{-1})}{\ln(z)} \right]^\beta - 1 \right) \sim_{z\to\infty} \beta \frac{(\ln z)^{\beta-1}}{z}$$

  which converges to 0 at infinity.
- Notice that $(\ln(xy))^\beta - (\ln(x))^\beta \geq (\ln(y))^\beta$ for any $x, y \geq 1$. Indeed, setting $u = \ln x \geq 0$, $v = \ln y \geq 0$,

$$(u+v)^\beta - u^\beta = \int_u^{u+v} \beta z^{\beta-1} \mathrm{d}z \geq \int_0^v \beta z^{\beta-1} \mathrm{d}z = v^\beta$$

  (because $z \mapsto \beta z^{\beta-1}$ is increasing since $\beta > 1$).
- Last, since $e^{(\ln(1+xy))^\beta} = \Psi_\beta^{\mathtt{HT}}(xy) + 1 \geq \Psi_\beta^{\mathtt{HT}}(c^2) + 1 \geq 2$, one has

$$e^{(\ln(1+xy))^\beta} = \frac{e^{(\ln(1+xy))^\beta}}{e^{(\ln(1+xy))^\beta} - 1} \Psi_\beta^{\mathtt{HT}}(xy) \leq 2\Psi_\beta^{\mathtt{HT}}(xy).$$

All in all, we conclude that $\Psi_\beta^{\mathtt{HT}}(x)\Psi_\beta^{\mathtt{HT}}(y) \leq 2C(c)\Psi_\beta^{\mathtt{HT}}(xy)$, for any $x, y \geq c$. We are done. □

3.3. Proof of Theorem 2.4

We follow the strategy of [CGS20] by truncating the unbounded functions f by a threshold c, whose impact is analyzed using the Hoffman-Jorgensen inequality [LT13, Proposition 6.8] and the Talagrand inequality of Theorem 2.1. The deviation probability related to the newly bounded random variables is quantified thanks to the Klein-Rio inequalities [KR05] together with the Dudley entropy integral bound.

Here are the notations used along this proof. We denote by K a positive constant that may change from line to line in the computations: this generic constant K may depend on universal constants and β, but it does not depend on the sample $X_1, \ldots, X_M$, its size M, nor the class of functions $\mathcal{F}$, neither ε. For ease of notations, we write $a \leq_K b$ when $a \leq Kb$.
For a given $c > 0$, set

$$\mathcal{R}_c f := f - \mathcal{T}_c f \qquad \text{where} \qquad \mathcal{T}_c f := -c \vee f \wedge c,$$
$$\mathcal{T}_c \mathcal{F} := \{\mathcal{T}_c f : f \in \mathcal{F}\},$$

[3] one can take $c = \sqrt{e^{(\ln 2)^{1/\beta}} - 1} \geq 1$ for which $\Psi_\beta^{\mathtt{HT}}(c^2) = 1$.

$$\mathcal{T}_c^m f(\cdot) := \mathcal{T}_c f(\cdot) - \mathbb{E}\left[\mathcal{T}_c f(X_m)\right],$$
$$Z_c := \sup_{f\in\mathcal{F}} \frac{1}{M} \sum_{m\in[M]} \mathcal{T}_c^m f(X_m).$$

Note that the function $\mathcal{T}_c^m f$ is centered w.r.t. the distribution of X_m, and bounded by $2c$. Assume that $c > 0$ and $\varepsilon > 0$ are such that

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\mathcal{R}_c f(X_m)\right] \right| \leq \varepsilon/4, \tag{3.10}$$
$$\mathbb{E}\left[Z_c\right] \leq \varepsilon/4. \tag{3.11}$$

By writing $f = \mathcal{R}_c f + \mathcal{T}_c f$ and using the sub-additivity of the supremum, we easily get

$$\sup_{f\in\mathcal{F}} \frac{1}{M} \sum_{m\in[M]} \left(f(X_m) - \mathbb{E}\left[f(X_m)\right]\right)$$
$$\leq Z_c - \mathbb{E}\left[Z_c\right] + \sup_{f\in\mathcal{F}} \left| \frac{1}{M} \sum_{m\in[M]} \mathcal{R}_c f(X_m) \right| + \varepsilon/2.$$

Hence, the probability of deviation in Theorem 2.4 is bounded by

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{M} \sum_{m\in[M]} \mathcal{R}_c f(X_m) \right| \geq \varepsilon/4 \right) + \mathbb{P}\left(Z_c - \mathbb{E}\left[Z_c\right] \geq \varepsilon/4\right) =: (\star) + (\star\star).$$

▷ **Term** $(\star)$**.** Owing to the deviation inequality (iii) from Section 2.1, it is bounded by

$$\mathbb{P}\left(\sum_{m\in[M]} \sup_{f\in\mathcal{F}} |\mathcal{R}_c f(X_m)| \geq M\varepsilon/4 \right)$$
$$\leq 2\exp\left(- \left(\ln\left(\frac{M\varepsilon/4}{\left\| \sum_{m\in[M]} \sup_{f\in\mathcal{F}} |\mathcal{R}_c f(X_m)| \right\|_{\Psi_\beta^{\mathtt{HT}}}} + 1 \right) \right)^\beta \right).$$

Using the Talagrand inequality of Theorem 2.1 and the Hoffman-Jorgensen inequality [LT13, Proposition 6.8], and following line by line the arguments of [CGS20, Section 5.5, Inequalities (38) and (39)], we can show that the above $\|.\|_{\Psi_\beta^{\mathtt{HT}}}$ norm is bounded by

$$K \left\| \max_{m\in[M]} F(X_m) \right\|_{\Psi_\beta^{\mathtt{HT}}},$$

provided that $c \geq 8\mathbb{E}\left[\max_{m\in[M]} F(X_m)\right]$. The above arguments are crucial to deal both with the truncation in c and the sup in f. Furthermore, the maximal inequality (2.4) with $Y_m := F(X_m)$ gives

that

$$\left\| \max_{m\in[M]} F(X_m) \right\|_{\Psi_\beta^{\mathtt{HT}}} \leq \mathcal{C}_{\beta,(2.4)} \Psi_{1/\beta}^{\mathtt{HT}}(M) \bar{\mu}_{\Psi_\beta^{\mathtt{HT}}}. \tag{3.12}$$

All in all, we have

$$c \geq 8\mathbb{E}\left[\max_{m\in[M]} F(X_m)\right] \implies (\star) \leq 2\exp\left(-\left(\ln\left(\frac{M\varepsilon}{K\bar{\mu}_{\Psi_\beta^{\mathtt{HT}}}\Psi_{1/\beta}^{\mathtt{HT}}(M)}+1\right)\right)^\beta\right).$$

The above condition on the left hand side is met as soon as

$$c \geq_K \Psi_{1/\beta}^{\mathtt{HT}}(M)\bar{\mu}_{\Psi_\beta^{\mathtt{HT}}},$$

where we have used $\mathbb{E}[\cdot] \leq_K \|\cdot\|_{\Psi_\beta^{\mathtt{HT}}}$ and (3.12).

▷ **Term** $(\star\star)$**.** Apply the Klein-Rio inequality [KR05, Theorem 1.1] (we shall use the form presented in [CGS20, Theorem 8] which directly fits our setting), it shows that $(\star\star)$ is bounded by

$$\exp\left(-\frac{M(\varepsilon/4)^2}{2(\sigma^2+4c\mathbb{E}(Z_c))+6c(\varepsilon/4)}\right)$$

where $\sigma^2 := \sup_{f\in\mathcal{F}} \frac{1}{M}\max_{m\in[M]} \mathbb{E}\left[(\mathcal{T}_c^m f)^2(X_m)\right]$. Observe that

$$\sigma^2 \leq \sup_{f\in\mathcal{F}}\max_{m\in[M]} \mathbb{V}\mathrm{ar}\left[\mathcal{T}_c^m f(X_m)\right] \leq \sup_{f\in\mathcal{F}}\max_{m\in[M]} \mathbb{E}\left[\mathcal{T}_c f^2(X_m)\right] \leq \mu_2^2.$$

Using in addition the bound (3.11) on $\mathbb{E}(Z_c)$, we get

$$(\star\star) \leq \exp\left(-\frac{M\varepsilon^2}{K(\mu_2^2+c\varepsilon)}\right)$$

where K is a universal constant.

▷ **Condition** (3.10)**.** From $\mathcal{R}_c f(x) = (f(x)-c)_+ - (f(x)+c)_-$, we easily get

$$\begin{aligned}
|\mathbb{E}[\mathcal{R}_c f(X_m)]| &\leq \int_c^{+\infty} \mathbb{P}(|f(X_m)| \geq z)\,\mathrm{d}z \\
&\leq 2\int_c^{+\infty} \exp(-(\ln(z/\lambda+1))^\beta)\mathrm{d}z \\
&= 2\lambda\int_{c/\lambda}^{+\infty} \exp(-(\ln(z+1))^\beta)\mathrm{d}z =: 2\lambda\mathcal{I}(c/\lambda)
\end{aligned}$$

where $\lambda := \mu_{\Psi_\beta^{\mathtt{HT}}}$. A standard calculus shows that

$$\mathcal{I}(y) \sim_{y\to+\infty} \frac{y}{\beta(\ln(y+1))^{\beta-1}}\exp(-(\ln(y+1))^\beta),$$

and thus

$$\mathcal{I}(y) \leq_K \exp(-(\ln(y+1))^\beta/2) =: \mathcal{J}(y), \qquad \forall y \geq 0.$$

This gives

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\mathcal{R}_c f(X_m)\right] \right| \leq_K \lambda \exp\Big(- (\ln(c/\lambda+1))^\beta/2 \Big).$$

Therefore, to ensure (3.10) it is enough to take

$$c \geq \mu_{\Psi_\beta^{\mathrm{HT}}} \left(\exp\left((2\ln_+(K\mu_{\Psi_\beta^{\mathrm{HT}}}/\varepsilon))^{1/\beta} \right) - 1 \right)$$

for some constant $K > 0$. Observe that the use of $\ln_+(.)$ guarantees that for a deviation ε large enough, the above lower bound is zero, meaning that any value of $c \geq 0$ ensures that (3.10) holds, as it is expected (for large ε).

▷ **Condition (3.11).** Deriving a bound on the expectation of the supremum follows a standard routine using Dudley entropy integral bound. For sake of brevity, we closely follow the arguments of [CGS20, p.20, term $\mathbb{E}\left[\bar{Z}^{\mathcal{T}_c}\right]$]. It gives that

$$\mathbb{E}\left[Z_c\right] \leq 2\mathbb{E}\left[\frac{C_D}{\sqrt{M}} \int_0^\infty \sqrt{\ln(\mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F}))}\mathrm{d}z \right] \tag{3.13}$$

where $\mathrm{d}_\mathcal{F}(f,g) := \left(\frac{1}{M}\sum_{m=1}^M |f(X_m) - g(X_m)|^2 \right)^{1/2}$ and $\mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F})$ is the covering number of $\mathcal{T}_c\mathcal{F}$ with respect to the distance $\mathrm{d}_\mathcal{F}$ with balls of radius z (see [GKKW02, Definition 9.3]). Actually, since functions in $\mathcal{T}_c\mathcal{F}$ are bounded by c, $\mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F}) = 1$ for $z \geq 2c$ and therefore, the above integral can be restricted to $[0, 2c]$ without modification. In addition, we have the following universal upper bound in terms of VC dimension:

$$0 < z < 2c/4 \implies \mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F}) \leq 3\left(2e\left(\frac{2c}{z}\right)^2 \ln\left(3e\left(\frac{2c}{z}\right)^2 \right) \right)^{\mathrm{V}_{\mathcal{F}^+}}. \tag{3.14}$$

Indeed, the above estimate follows from [GKKW02, Lemma 9.2, Theorem 9.4 with $B = 2c$ and $p = 2$, $\mathrm{V}_{(\mathcal{T}_c\mathcal{F})^+} \leq \mathrm{V}_{\mathcal{F}^+}$ in the proof of Theorem 9.6]. See [vdVW96, Theorem 2.6.7] for a variant of this upper bound. Since $\mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F})$ is non-decreasing in z, and since we do not pay much attention to universal constants, we can simply write

$$0 < z \leq 2c \implies \mathcal{N}_2(z, \mathrm{d}_\mathcal{F}, \mathcal{T}_c\mathcal{F}) \leq \left(\frac{Kc}{z} \right)^{3\mathrm{V}_{\mathcal{F}^+}},$$

for a universal constant K. Plugging this into (3.13) readily leads to

$$\mathbb{E}\left[Z_c\right] \leq Kc\frac{\sqrt{\mathrm{V}_{\mathcal{F}^+}}}{\sqrt{M}}.$$

▷ **Conclusion.** Gathering all the estimates and conditions leads to the statement of Theorem 2.4. □

3.4. Proof of Proposition 2.1

Item 1. Observe that $\Psi_1^{\mathtt{HT}}(x) = x$ and $\Psi_{\beta_1}^{\mathtt{HT}}(\Psi_{\beta_2}^{\mathtt{HT}}(x)) = \Psi_{\beta_1\beta_2}^{\mathtt{HT}}(x)$ for any $x \geq 0$; the property of group isomorphism readily follows.
Items 2 and 4 are straightforward to verify.
Item 3. $\Psi_\beta^{\mathtt{HT}}$ is a C^∞-function on $(0,\infty)$, with a second derivative equal to

$$\Psi_\beta^{\mathtt{HT}\prime\prime}(x) = \frac{\exp((\ln(1+x))^\beta)\,(\ln(1+x))^{\beta-2}}{(1+x)^2} \times \beta \times \underbrace{\Big(\beta\,(\ln(1+x))^\beta + (\beta-1) - \ln(1+x)\Big)}_{=:g(\ln(1+x))}.$$

The function g is continuously differentiable on $\mathbb{R}^+$, strictly positive at 0 ($g(0) = \beta - 1 > 0$) and goes to infinity at infinity (since $\beta > 1$); the critical points of g' are solutions to $\beta^2 y^{\beta-1} - 1 = 0$, therefore it is unique (equal to $y_\beta := \beta^{-\frac{2}{\beta-1}}$) and corresponds to the minimum of g. Let us evaluate the sign of g at the minimum:

$$g(y_\beta) = \beta y_\beta^\beta + (\beta-1) - y_\beta = \frac{y_\beta}{\beta} + (\beta-1) - y_\beta$$
$$= (\beta-1)\left(1 - \frac{y_\beta}{\beta}\right) = (\beta-1)\left(1 - \frac{1}{\beta^{\frac{\beta+1}{\beta-1}}}\right) > 0.$$

All in all, we have proved that $\Psi_\beta^{\mathtt{HT}\prime\prime}(x) > 0$ for any $x > 0$. □

4. Conclusion

To conclude, we have extended the Talagrand inequality for an Orlicz norm adapted to variables with β-heavy tails (Proposition 2.1 and Theorem 2.1). We have also shown that a maximal inequality holds (Theorem 2.2), which, in combination with the Talagrand inequality, allows for a concentration inequality for the sum of independent centered β-heavy tailed random variables (Corollary 2.3). Then we have extended this inequality to supremum of functions of random variables with β-heavy tails (Theorem 2.4), by combining previous results with the Hoffman-Jorgensen, Klein-Rio and Dudley entropy integral inequalities.

Acknowledgements

This research is supported by the *Chair Stress Test, RISK Management and Financial Steering of the Foundation Ecole Polytechnique* and by the *Association Nationale de la Recherche Technique*. This work is part of the first author's doctoral thesis (funded by BNP Paribas), under the supervision of the second author.

References

[Ada08] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

[Asm03] S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[CGS20] L. Chamakh, E. Gobet, and Z. Szabó. Orlicz random Fourier feature. *Journal of Machine Learning Research (JMLR)*, 21:1–37, 2020.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[HR99] T. Huillet and H.-F. Raynaud. Rare events in a log-Weibull scenario-Application to earthquake magnitude data. *The European Physical Journal B - Condensed Matter and Complex Systems*, 12(3):457–469, 1999.

[Kol11] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[KR61] M.A. Krasnoselskii and Ya.B. Rutickii. *Convex functions and Orlicz spaces*. Translated from the first Russian edition by Leo F. Boron. P. Noordhoff Ltd., Groningen, 1961.

[KR05] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[LT13] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[Mar17] A. Marchina. Concentration inequalities for suprema of unbounded empirical processes. *hal-01545101*, 2017.

[NP07] R. Nickl and B.M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. *J. Theoret. Probab.*, 20(2):177–199, 2007.

[Pis16] G. Pisier. *Martingales in Banach spaces*, volume 155 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2016.

[Rio17] E. Rio. About the constants in the Fuk-Nagaev inequalities. *Electronic Communications in Probability*, 22, 2017.

[Tal89] M. Talagrand. Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *The Annals of Probability*, 17(4):1546–1570, 1989.

[vdG00] S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[vdGL13] S. van de Geer and J. Lederer. The Bernstein-Orlicz norm and deviation inequalities. *Probab. Theory Related Fields*, 157(1-2):225–250, 2013.

[vdVW96] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[Wel17] J.A. Wellner. The Bennett-Orlicz norm. *Sankhya A. The Indian Journal of Statistics*, 79(2):355–383, 2017.